Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Systems Engineering

MASTER OF SCIENCE (M.Sc.) THESIS

Models for Decision Support Systems

Supervised by

Doc. RNDr. Helena Brožová, CSc.

Submitted by

Shiekh Samsuzzhan Alam

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

SHEIKH SAMSUZZHAN ALAM, MS

Informatics

Thesis title

Models for Decision Support Systems

Objectives of thesis

Main aim is to analyze suitability of selected Data Mining techniques and modern visualization for application in Decision Support Systems.
The second aim is to show some practical applications of the selected tools.

Methodology

Study of Decision Support System theory.

Study of selected Data Mining and modern visualization techniques.

Analysis of selected tools suitability for Decision Support Systems.

Practical applications of the selected tools.

The proposed extent of the thesis

70 pp.

Keywords

Decision Support Systems, Data Mining, visualization

Recommended information sources

BEALE, R. – JACKSON, T. *Neural Computing : An introduction.* Bristol: Institute of Physics Publishing, 1994. ISBN 0-85274-262-2.

BURSTEIN, F. – HOLSAPPLE, C W. *Handbook on decision support systems. Vol. 2, Variations.* Berlin: Springer, 2008. ISBN 978-3-540-48715-9.

KECMAN, V. *Learning and soft computing : support vector machines, neural networks, and fuzzy logic models.* Cambridge: MIT Press, 2001. ISBN 0-262-11255-8.

POWER, D J. *Decision support systems : concepts and resources for managers.* Westport: Quorumbooks, 2002. ISBN 156720497.

TURBAN, E. *Decision support and business intelligence systems.* Boston: Pearson , 2011. ISBN 978-0-13-245323-3.

TURBAN, E. – TURBAN, E. *Decision support and business intelligence systems.* Upper Saddle River, N.J.: Pearson Prentice Hall, 2007. ISBN 0131986600.

Expected date of thesis defence

2016/17 SS – FEM

The Diploma Thesis Supervisor

doc. RNDr. Helena Brožová, CSc.

Supervising department

Department of Systems Engineering

Electronic approval: 28. 3. 2017	Electronic approval: 28. 3. 2017
doc. Ing. Tomáš Šubrt, Ph.D.	**Ing. Martin Pelikán, Ph.D.**
Head of department	Dean

Prague on 09. 05. 2017

Declaration

I certify that the paper entitled "Models for Decision Support Systems" submitted as a partial requirement for the degree of Master of Science (M.Sc.) in Informatics is the result of my own research, except where otherwise acknowledged, and this project report in whole or in part has not been submitted for an award, including a higher degree to any other University or Institution.

Name : Sheikh Samsuzzhan Alam
Signed:
Date:

Acknowledgements

At first I express gratitude to the almighty for the blessing guidance and immeasurable grace for a successful completion of the thesis and submission. It is a great pleasure in expressing enormous gratefulness and sincere gratitude to my respected supervisor, Doc. RNDr. Helena Brožová, CSc., Department of Systems Engineering (FEM), Czech University of Life Sciences Prague (CULS), for her valuable advice and continuous help throughout the period of my research work.

I am extremely grateful to my erudite teacher and course coordinator Inz. Martin Kozák for study guidance, preparation of supporting documents on demand for training and scientific workshop.

I would like to express my heartfelt thanks to my father, mother and friends of CULS and specially my beloved friend Makina Rahaman, my band members of Pivo Prosim and Radial Symmetry, my beloved friend Joana Costa, my Teamlead Pavol Dravečky and other colleague of the company I am working, for their enthusiastic support.

Abstract

Enterprises deals with a highly dynamic environment that requires management to integrate available knowledge gained from the data and information for effective decision-making. The advancement of the web technology and cloud services make it possible to integrate the whole process. As a result, all most all of the computing work and analytics is delegated to the computer and the underlying systems. Improvement of Internet technologies and software systems provide a new means of decision support functionalities and delivering decision support within a very short time frame. As large enterprises generates high volume of internal and external data some Data Mining technique is also need to build a proper data and model driven Decision Support System (DSS) application. The process of Data Mining converts information to knowledge by using tools from the disciplines of computational statistics, database technologies, machine learning, signal processing, nonlinear dynamics, process modeling, simulation, and allied disciplines. Data Mining allows business problems to be analyzed from diverse perspectives, including dimensionality reduction, correlation and co-occurrence, clustering and classification, regression and forecasting, anomaly detection, and change analysis. The tools that enable the transformation of raw data to actionable predictive insights are collectively referred to as decision support tools.

This paper describes process of building a Data Mining based Decision Support System from scratch for enterprise decision making purpose Which offers basic descriptive analytics and advance predictive modeling as well. Through our study we were able to analyse the clothing retailing industry for a particular company expansion in Australia which is currently running operation in New South Wales. We user several different technology such as Alteryx Designer and R Language for data preparation and advance modeling of available data. For visualization Tableau Desktop and Tableau Server to create a interactive decision support application.

Keywords: Decisions Support System, Data Mining, Data Preparation and ETL, Data Visualization, Time Series Forecasting, Tableau, Alteryx, R, Interactive Visualization.

Contents

List of Tables

List of Figures

1 Introduction

The working environment in the enterprises, business and financial organizations, health care services, banking systems and almost all the private and public services are dynamic and it is becoming more and more complex day by day. Private and public enterprises are under pressure that force them to respond as early as possible to change request and to be innovative in the way they operate. So organizations require to be agile and to make frequent and quick strategic, tactical, and operational decisions, some of which are very complex and critical. To make such decision management of responsible person may require considerable amounts of relevant data, information, and knowledge. Processing data, in the framework of the needed decisions must be done quickly frequently in real time but with out loosing valuable information. Such processes usually requires full computerized support. Decision support can be provided in many different configuration. These forms depends on the decision situations and the specific technologies used for support. Decision Support System (DSS) is one of the best possible way to deal with decision making processes.

Moore and Chang defined the Decision Support System as “an extensible system, capable of ad-hoc analysis and decision modeling, reporting, focused on future planning and used at unplanned and irregular time stamp” (Watson et al., 1991). Also Muntean (2004) explains Decision Support Systems as being “interactive systems that help decedent makers use data and models in resolving unstructured and semi-structured economical problems”.

Turban et al. (2011) defines a Decision Support System (DSS) as “an interactive, flexible and adaptable system, exclusively designed to offer support in solving unstructured or semi-structured managerial problems, aiming to improve the decision processes. The system uses data (internal and external) and models, providing a simple and easy-to-use interface, thus, allowing the decision maker control over the decision process. The DDS offers support in all decision process's stages".

Decision support mostly focuses on developing systems to help decision-makers to solve unstructured or semi-structured problems. Decision support provides a selection of data analysis, simulation, visualization and modeling techniques, and software technologies such as Decision Support Systems, group decision support and mediation systems, expert systems, databases and data warehouses.

In this context, studies show that the process of defining a Decision Support System has started from the idea of how the objectives of a DDS can be achieved, how a DDSs components can be identified, the features that are provided to the end user and from the perception of what such a system is capable of doing (offering support in decision making processes, in solving structured and unstructured problems).

Considering all the definitions mentioned above, Bâra and Lungu (2012) proposed some of the most essential characteristics of the DDS are: uses data and models; enhances the learning process; grows the efficiency of the decision making process; offers support in the decision making process and allows the decision maker control over the entire process; offers support in all stages of the decision making process; offers support for decision makers in solving structured or unstructured problems; offers support for a user or for a group of users etc.

Data Mining is concerned with finding models and patterns from the available data and predict the future trend. Though Data Mining is very general term which can be split in to several different area includes different predictive algorithms and models. Such model can be used for prediction and classification. descriptive Data Mining algorithms for finding interesting patterns in the data, associations, clusters and subgroups, regression methods, survival analysis etc (Mladenic, 2003).

The ability to anticipate, react, and adapt to changes in the market and the capability to implement appropriate business strategies are two of the most important characteristics in which enterprises want to focus. Until recently, uncertainties inherent to forecasting models drove many companies to rely on decision making based on the gut feel of senior management rather than on verifiable data. As business conditions become more dynamic and active responses to subtle changes prove more valuable, continuous visualization and monitoring of the state of business through different tools and technologies. This is the reason why Data Mining (DM), Decision Support Systems (DSS) and business intelligence (BI) are becoming increasingly important. Correspondingly, the development of predictive modeling capabilities is becoming more and more popular for the support of decision making through the verifiable data presented in DM technologies (Lavrač et al., 2007). DM and DSS have generally been studied in relative isolation. Frameworks that bridge the gap between DM analysis and predictions to actions and decisions in decision support are required to ensure better integration of the two methodologies. In this paper, we discuss the use of DM technologies in the creation of DSS and combine both to crate a Data Mining and Decision Support System (DMDSS) application.

Data Mining and decision support can be integrated to better solve data analysis and decision support problems. In knowledge management, such integration is interesting for several reasons (Smith and Farquhar, 2000). For example, in Data Mining it is often unclear which algorithm is best suited for the problem. Here, we require some decision support for Data Mining. Another example is when there is a lack of data for the analysis. To ensure that appropriate data is recorded when the collection process begins it is useful to first build a decision model and use it as a basis for defining the attributes that will describe the data. These two examples show that Data Mining and decision support can complement each other, to achieve better results. Different aspects of Data Mining and decision support integration have been investigated in (Mladenic, 2003).

1.1 Problem Statement

Information by itself has no longer been perceived as an asset. Significantly large amount of business transactions are recorded in enterprise scale data warehouses every day and for knowledge discovery and decision making it needs to go through automated data discovery process. Recent advances in (remote or other) sensor technologies have led to the development of scientific data repositories. Database technologies, ranging from relational systems, to extensions like spatial, temporal, time series, text or media, big data ecosystem, no sequel data system as well as specialized tools like geographical information systems (GIS), have transformed the design of enterprise scale business or large scientific applications. The question increasingly faced by the data scientist or business decision-maker, is not how one can get more information or design better information systems, but what to make of the information and systems already in place. The challenge is to be able to utilize the available information, to gain a better understanding of the past, and predict or influence the future through better decision-making. Researchers in Data Mining technologies (DMT) and Decision Support Systems (DSS) are responding to this challenge. Broadly defined, Data Mining (DM) relies on scalable statistics, artificial intelligence, machine learning or knowledge discovery in databases (KDD). DSS utilize available information and DMT to provide a decision-making tool usually relying on human-computer interaction. Together, DMT and DSS represent the spectrum of analytical information technologies (AIT), and provide a unifying platform for an optimal combination of data dictated and human driven analytic.

Now a days in large enterprise management, managers have to make quick decision from a large amount of information. So building Decision Support System as a piece of software from scratch is not effective in that case. Because we need time, resources and significant amount money for that. Also improvement of web technology now enterprises are emphasizing on cloud based web application rather that static software. So that managers and employees can have access to the system any time from any where of the world.

To conduct quick data cleaning and mining task software and programming language such as R, Python, SAS etc. are used frequently. Also there is some new tools such as Alteryx, Talend, Keboola, Good Data, Pentaho for quick data preperation and ETL exist in the market which are more user friendly. User need no prior knowledge about programming logic because all of them are drag and drop user interface based software. For building Decision Support application and Visual report on top of mined data Tableau, Domo, Microsoft Power BI, Splunk are used frequently now a days in big enterprises. Combination of these tools consultants and management are able to make quick visual and Decision support in a minute.

Through our study we will describe the procedure of data preparation, blending and ETL by Alteryx and R, descriptive analysis and predictive analytics of data with R. Finally we will build user friendly and interactive Decision Support Application with Tableau Desktop and Tableau Server which can be viewed anytime and any where.

Followed by the process stated above, we will be studied on real world problem and we will develop sophisticated decision support application for the management to help them to discover more insights from the data. We will analyse the clothing retailing market for a particular company which is currently operation from New South Wales, Australia and wish to expand their operation to some other states in Australia. We will analyse net profit profit margin of several companies, monthly retail turnover and per capita sales to make a DSS application.

2 Objective and Methodology

2.1 Objective of the Study

Independently, Data Mining, decision support and visualization of summary information for Data Mining (which can be considered as knowledge discovery) are well-developed research areas. The main objective of this study is to ingrate all this this three processes through one model and build a user friendly and interactive **Decision Support System**. The main objective can be split into following parts:

- Introduce current technology and tool to create such web based Decision Support System or application.
- Identify the process of integrating the Decision Support System, Data Mining and Visualization and introduce a general model.
- Finally Create a Decision Support Application for a case studies which included Data Mining techniques and interactive visualization.

2.2 Methodology of the Study

We develope Decision Support System application based on DMDSS process model. Through the building of our model we partially used the process model for DMDSS described by Rupnik and Kukar (2007). But we create the whole application with respect to enterprise environment and equipped with the latest technology and services. To apply DMDSS process in our study we combined several different technology for different purpose as follows:

- Gather knowledge about Data Mining and Decision Support System (DMDSS) frame work for enterprise environment.

- Data preparation and Data ETL through Alteryx Designer and R language.
- Descriptive analysis for cross section and time series data with Tableau.
- General predictive modeling of time series with combination of R and Alteryx Designer.
- Combine all result and graphics through Tableau Desktop to create Decision Support Application and turn the application in to a interactive and cross platform compatible cloud server based application through Tableau Server.

3 Framework for Data Mining and Decision Support System

Computerized Decision Support Systems became practical with the development of minicomputers, timeshare operating systems, and distributed computing. The history of the implementation of such systems begins in the mid-1960s. In a technology field as diverse as decision support, chronicling history is neither neat nor linear. Different people perceive the field of Decision Support Systems from various vantage points and report different accounts of what happened and what was important (Power, 2008). As technology evolved new computerized decision support applications were developed and studied. Researchers used multiple frameworks to help build and understand these systems. Today, one can organize the history of DSSs into the five broad DSS categories explained in Power (2008) including:communications-driven, data-driven, document-driven, knowledge-driven and model-driven Decision Support Systems.

In the first section of this chapter traces decision support applications and research studies related to model and data-oriented systems, management expert systems, multidimensional data analysis, query and reporting tools, business intelligence, group DSSs, document management, spatial DSSs, and executive information systems as the technologies emerge, converge and diverge. All of these technologies have been used to support decision making. In the second section we discuss broadly about Several different data mining technique and their application in differet study area. In the later chapter we discuss how to combine the both DM and DSS process and use tool to visualize the output.

3.1 Decision Support System (DSS)

The early definition of a Decision Support System (DSS) identified it as a system intended to support managerial decision makers in semi structured and unstructured decision situation. DSS were meant to be adjuncts to decision makers, extending their capabilities but not replacing their judgement. They were aimed at decisions that required judgment or at decisions that could not be completely supported by algorithms. Not specifically stated but implied in the early definition was the notion that the system would be computer based would operate interactively on-line and preferably would have graphical output capabilities, now simplified via web servers and browsers (Power, 2008).

A structure of of computerised DSS explained by Turban in 2011. According to Turban et al. (2011), a DSS typically built to support the solution of a certain problem or to evaluate an opportunity. Formally, a DSS is an approach for supporting decision making. It uses an interactive, flexible adaptable computer based information system (CBIS) especially developed for supporting the solution to a specific non-structured management problem. It uses data, provides an easy user interface and can incorporate the decision maker's own insights. I addition, a DSS includes models and is developed through an iterative and iterative process. It supports all phases of decision making and may include a knowledge component. Finally , a DSS can be used by a single user on PC or can be web based for use by many people at several locations which is the main focus of our study.

It is beneficial first to deal with the characteristics and capabilities of DSS. Figure 3.1 shows a typical web based DSS architecture proposed by Turban et al. (2011). This DSS structure utilizes models in Business Intelligence work. Processing is distributed across several servers in solving large analytical problems. This architecture uses a Web browser to run programs on an application server. The servers accesses data to construct one or more models. Data may also be provided by a data server that optionally extracts data from a data warehouse or a legacy mainframe system. When the user requires that the model be optimized, the model, populated with the data, is transferred to an optimized server. The optimization server accesses additional data form from the data server, if needed, solves the the problem an provides the solution directly to the user's web browser. Generated solution reports, which the application server may massage to make

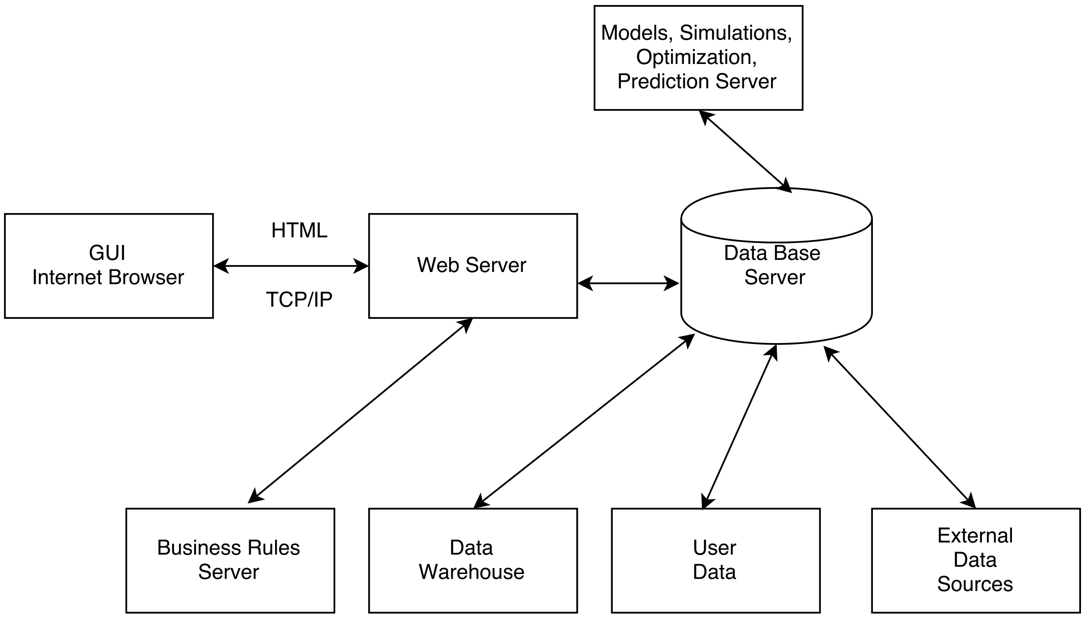


Figure 3.1: Architecture of a Web-Based Decision Support System

readable by managers may be sent directly to appropriate parties via e-mail or may be made available through another web portal as par to this enterprise information system. DSS also may run in stand-alone mode, usually through a spreadsheet or a modeling language.

3.1.1 Decision Support System Characteristics and Capabilities

Because there is no consensus on exactly what a DSS is, there is obviously no agreement on the standard characteristics and capabilities of DSS. The capabilities in Figure 3.2 constitute an idea set, some members of which are describe in the definition of DSS given by Turban et al. (2011).

The key characteristics and capabilities of DSS given by Turban et al. (2011) are:

1. Support for decision maker, mainly in semi-structured and unstructured situations, by bringing together human judgment. Such problem can not be solved by other computerized system or through use of standard quantitative methods or tools. Generally these problems gain structured as the DSS is developed. Even some structured problems has been solved by DSS.

2. Support for all managerial level from executive to line manager.

3. Support for individual as well as group. Less-structure problems often required

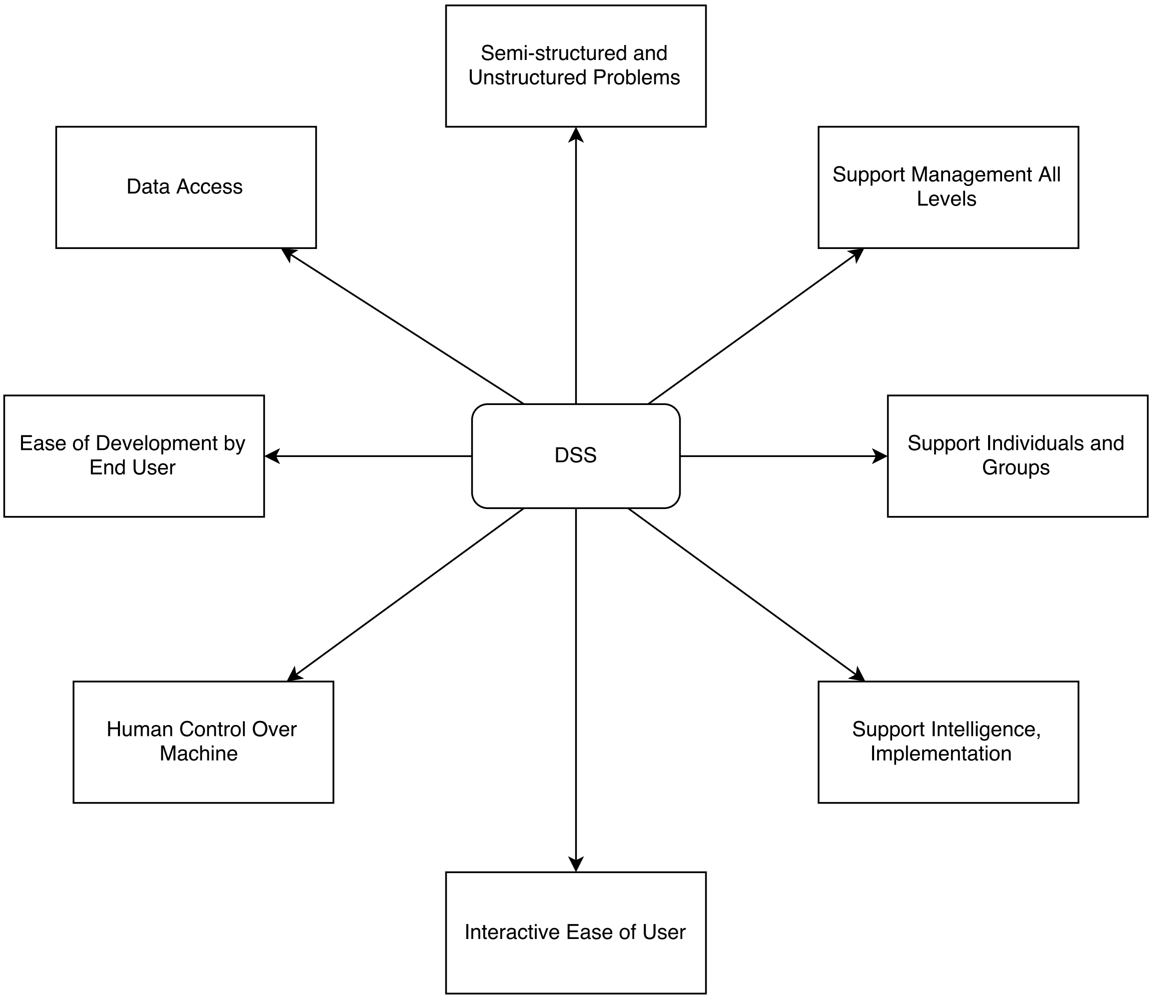


Figure 3.2: Different Key Characteristics and Capabilities of DSS

involvement of individuals from different department and organizational levels or even from different organizations. DSS support virtual term through collaborative web tools. DSS has been developed to support individual and group work as well as to support individual decision making and group of decision maker working some what independently.

4. Support for independent and sequential decisions. The decision may be made once several times and repeatedly.

5. Support in all phase of decision making processes: intelligence, design, choices and implementation.

6. Support for variety of decision making process and style.

7. The decision maker should be reactive, able to confront changing conditions quickly,

and able to adapt the DSS to meet these changes. DSS is flexible so user can add, delete, combine, change and rearrange basic elements. They are also flexible in that they can be readily modified to solve others, similar problems.

8. User-friendliness, strong graphical capabilities and a natural language interactive human machine interface can greatly increase the effective of DSS. Most new DSS applications use web-based interface.

9. Improvement or effectiveness of decision making. When DSS are deployed, decision making often takes longer, but the decision are better.

10. The decision maker has complete control over all steps of the decision making process in solving a problem. A DSS specially aim to support, not to replace, the decision maker.

11. End users are able to develop and modify simple system by then selves. Large system can be build with assistance from information system specialist.

12. Models are generally utilized to analyze decision making situation. The modeling capability enables experimentation with different strategies under different configurations. In fact, the model makes a DSS different form most MIS.

13. Access is provide to a variety of data source, formats and types, including GIS, multimedia and objected oriented database.

14. The DSS can be employed as a stand alone tool used by an individual decision maker in one location or distributed throughout an organization and in several organization along the supply chain.

The key DSS characteristics and capabilities allows decision makers to make better, more consistent decision in timely manner and they are provided by the major DSS component.

3.1.2 Decision Support System Classification

DSS Application can be classified in several different way (Shim et al., 2002). The design process as well as the operation and implementation of DSS, Depends in many cases on the type of DSS involved. However it is to be remembered not every DSS fits nearly to one category. Most fit in to the classification provided by the association for Information Systems Special Interest Group on Decision Support Systems (AIS SIGDSS). We discuss the classification first followed by other classification that more or less fit into the former's classification set.

The AIS SIGDSS Classification for DSS

The AIS SIGDSS has adopted a concise classification scheme for DSS that was proposed by Power (2000). It includes the following categories.

- Communications-driven and group DSS (GSS)
- Data Driven DSS
- Document Driven DSS
- Knowledge Driven DSS, Data Mining and Management ES Application
- Model-Driven DSS

There may also hybrids that combine two or more categories. These are called compound DSS.

Communications-Driven and group DSS Communications-driven and group DSS (GSS) includes DSS that user computer, collaboration and communication technologies to support groups in tasks that may or may not includes decision making. Essentially, all DSS that support any kind of group work fall into this category. They include those that support meetings, design collaboration, and even supply chain management. Knowledge management systems (KMS) that are developed around communities that practice collaborative work also also fall into this category (Power, 2008).

Accoridng to Power (2002) Communications-driven DSSs use concept of network and communications technologies to make to make easier decision-relevant collaboration and communication. So in these systems, communication technologies are the dominant architectural element. Usually Tools used include groupware, video conferencing and computer-based bulletin boards etc.

Paper of Engelbart (2001) 'Augmenting human intellect: A conceptual framework' is the pioneer for much of the later work related to communications-driven DSSs. For the first time He demonstrated the hypermedia/groupware system NLS (oNLine System) at the Fall Joint Computer Conference in San Francisco. Beside this Engelbart manage to invent both the computer mouse and groupware.

The article of Joyner and Tunstall (1970) reporting testing of their conference coordinator computer software is the first empirical study in this DSS research area. Also Murray Turoff's (1970) article on the concept of computerized conferencing. Murray developed and implemented the first computer mediated communications system (EMISARI) tailored to facilitate group communications.

But at the beginning of 1980, scientific researchers developed a new type of software to support group decision making called group Decision Support Systems, named as GDSSs (Gray et al. (1981), Huber 1982, Turoff and Hiltz 1982). Also Mindsight by Execucom Systems, GroupSystems intrduced at the University of Arizona, and the SAMM system built by University of Minnesota researchers were few examples of early group DSSs.

Day by day GroupSystems matured into a demanded commercial product. Jay Nunamaker Jr. and his colleagues cited in 1992 that the underlying principle for GroupSystems had its beginning in 1965 with the construction of Problem Statement Language/ Problem Statement Analyzer at the Case Institute of Technology. In 1984, the forerunner to Group DSS Systems called PLEXSYS was completed and a computerised group meeting facility was constructed at the University of Arizona. The first Arizona facility, called the PlexCenter, housed a large U-shaped conference table with central processing workstations (Power, 2002).

According to SAMM, Dickson et al. (1992), Brent Gallupe, a Ph.D. student at the University of Minnesota, decided in 1984 to program his own small GDSS system in BASIC and run it on his university's desktop. In theri article DeSanctis and Gallup (1987) defined two types of GDSSs. Basic or level one GDSSs are systems with tools to reduce communication barriers, such as large screens for display of ideas, voting mechanisms, and anonymous input of ideas and preferences. These are communications-driven DSSs. More Advanced or level two GDSSs provide problem-structuring techniques, such as planning and modeling tools. These are model-driven group DSSs. Since the mid-1980s, many research studies have examined the impacts and consequences of both types of group DSSs. Also, companies have commercialized model-driven group DSS and groupware.

Kersten (1985) developed NEGO, a computerized group tool to support negotiations.

Bui and Jarke (1986) reported developing Co-op, a system for cooperative multiple criteria group decision support. Kraemer and King (1988) introduced the concept of collaborative Decision Support Systems (CDSSs). They defined them as interactive computer-based systems to facilitate the solution of ill-structured problems by a set of decision makers working together as a team.

In 1989, Lotus introduced a groupware product called Notes and broadened the focus of GDSSs to include enhancing communication, collaboration and coordination among groups of people. Notes had its roots in a product called PLATO Notes, written at the Computer-based Education Research Laboratory (CERL) at the University of Illinois in 1973 by David R. Woolley.

In general, groupware, bulletin boards, audio and videoconferencing are the primary technologies for communications-driven decision support. In the past few years, voice and video delivered using the Internet protocol have greatly expanded the possibilities for synchronous communications-driven DSS.

Data Driven DSS: Data driven DSS are primarily involved with data and processing them into information and presenting the information to a decision maker. Many DSS developed in OLAP and Data Mining software system fall into this category. Data-driven DSS is a type of DSS that emphasizes access to and manipulation of a time-series of internal company data and sometimes external data. Simple file systems accessed by query and retrieval tools provide the most elementary level of functionality. Data warehouse systems that allow the manipulation of data by computerized tools tailored to a specific task and setting or by more general tools and operators provide additional functionality. Data-driven DSS with On-line Analytical Processing (OLAP) provides the highest level of functionality and decision support that is linked to analysis of large collections of historical data. Executive Information Systems (EIS) and Geographic Information Systems (GIS) are special purpose Data-Driven DSS.

One of the first data-driven DSSs was built using an APL-based software package called AAIMS, an analytical information management system. It was developednfrom 19701974 by Richard Klaas and Charles Weiss at American Airlines (Alter 1980).

As noted previously, in 1979 John Rockarts research stimulated the development of executive information systems (EIS) and executive support systems (ESS). These systems evolved from single-user model-driven Decision Support Systems and from the development of relational database products. The first EIS used pre-defined information screens maintained by analysts for senior executives. For example, in the fall of 1978 development of an EIS called Management Information and Decision Support (MIDS) began at Lockheed-Georgia (Houdeshel and Watson, 1987).

The first EIS were developed in the late 1970s by Northwest Industries and Lockheed "who risked being on the 'bleeding edge' of technology" Few even knew about the existence of EIS until John Rockart and Michael Treacys article, "The CEO Goes On-line," appeared in the January-February 1982 issue of the Harvard Business Review" (Watson et al. 1997, p.6). Watson et. al (1997) further note. A major contributor to the growth of EIS was the appearance of vendor supplied EIS software in the mid 1980s. Pilot Software's Command Center and Comshare's Commander EIS made it much easier for firms to develop an EIS by providing capabilities for (relatively) easy screen design, data importation, user friendly front ends, and access to news services" (p.6). In a related development in 1984, Teradata's parallel processing relational database management system shipped to customers Wells Fargo and AT&T.

In about 1990, data warehousing and on-line analytical processing (OLAP) began broadening the realm of EIS and defined a broader category of data-driven DSSs (Dhar and Stein 1997). Nigel Pendse (1997), author of the OLAP report, claims both multi-dimensional analysis and OLAP had origins in the APL programming language and in systems like Express and Comshares System W. Nylund (1999) traces the developments associated with business intelligence (BI) to Procter and Gambles efforts in 1985 to build a DSS that linked sales information and retail scanner data. Metaphor Computer Systems, founded by researchers like Ralph Kimball from Xeroxs Palo Alto Research Center (PARC), built the early P&G data-driven DSS. Staff from Metaphor later founded many of the business intelligence vendors: The term BI is a popularized, umbrella term coined and promoted by Howard Dresner of the Gartner Group in 1989. It describes a set of concepts and methods to improve business decision making by using fact-based support systems. BI is sometimes used interchangeably with briefing books, report and query tools, and executive information systems. In general, business intelligence systems are data driven DSSs.

Bill Inmon and Ralph Kimball actively promoted Decision Support Systems built using relational database technologies. For many information systems practitioners, DSSs built using Oracle or DB2 were the first Decision Support Systems they read about in the popular computing literature. Ralph Kimball was the “doctor of DSS” and Bill Inmon was the “father of the data warehouse” By 1995, Wal-Marts data-driven DSS had more than 5 terabytes of on-line storage from Teradata that expanded to more than 24 terabytes in 1997. In more recent years, vendors added tools to create web-based dashboards and scorecards.

Document Driven DSS Document Driven DSS is a relatively new field in Decision Support. Document-Driven DSS is focused on the retrieval and management of unstructured documents. Documents can take many forms, but can be broken down into three categories: Oral, written, and video. Examples of oral documents are conversations that are transcribed; video can be news clips, or television commercials; written documents can be written reports, catalogs, letters from customers, memos, and even e-mail.

Fedorowicz (1988) estimated that American businesses store almost 1.3 trillion documents which can use up to 50 percent of their floor space. Yet only 5 to 10 percent of these documents are available to managers for use in decision making. Fedorowicz defined document as a “chunk” of information. Unfortunately documents are not standardized in a uniform pattern or structure. Managers and IT/IS staff need a way to transform these documents into usable formats that can be compared and processed to support decision making. New information technology and software is making this concept into a reality. (DSS Resource)

Text and document management emerged in the 1970s and 1980s as an important, widely used computerized means for representing and processing pieces of text (Holsapple and Whinston 1996). The first scholarly article for this category of DSS was written by Swanson and Culnan (1978). They reviewed document-based systems for management planning and control. Until the mid-1990s little progress was made in helping managers find documents to support their decision making. Fedorowicz (1993, 1996) helped define the need for such systems. She estimated in her 1996 article that only 5 to 10 percent of stored business documents are available to managers for use in decision making. The world-wide web technologies significantly increased the availability of documents and facilitated the development of document-driven DSSs.

By 1995, the world-wide web (Berners-Lee, 1996) was recognized by a number of software developers and academics as a serious platform for implementing all types of

Decision Support Systems.

Model Driven DSS Model-Driven DSS emphasize access to and manipulation of a model, for example, statistical, financial, optimization and/or simulation models. Simple statistical and analytical tools provide the most elementary level of functionality. Some OLAP systems that allow complex analysis of data may be classified as hybrid DSS systems providing both modeling and data retrieval and data summarization functionality. In general, model-driven DSS use complex financial, simulation, optimization or multi-criteria models to provide decision support. Model-driven DSS use data and parameters provided by decision makers to aid decision makers in analyzing a situation, but they are not usually data intensive, that is very large data bases are usually not need for model-driven DSS. Early versions of Model-Driven DSS were called Computationally Oriented DSS by Bonczek, Holsapple and Whinston (1981). Such systems have also been called model-oriented or model-based Decision Support Systems.

Gorry and Scott Morton (1971) production planning management decision system was the first widely discussed model-driven DSS, but Ferguson and Jones (1969) production scheduling application was also a model driven DSS.

Model driven DSSs emphasize access to and manipulation of financial, optimization, and/or simulation models. Simple quantitative models provide the most elementary level of functionality. Model-driven DSSs use limited data and parameters provided by decision makers to aid decision makers in analyzing a situation, but in general large data bases are not needed for model-driven DSSs (Power, 2002). Early versions of model-driven DSSs were called model-oriented DSSs by Alter (1980), computationally-oriented DSSs by Bonczek et al. (1981) and later spreadsheet-oriented and solver-oriented DSSs by Holsapple and Whinston (1996).

The first commercial tool for building model-driven DSSs using financial and quantitative models was called IFPS, an acronym for interactive financial planning system. It was developed in the late 1970s by Gerald R. Wagner and his students at the University of Texas. Wagners company, EXECUCOM Systems, marketed IFPS until the mid 1990s. Grays Guide to IFPS (1983) promoted the use of the system in business schools. Another DSS generator for building specific systems based upon the analytic hierarchy process (Saaty 1982), called Expert Choice, was released in 1983. Expert Choice supports personal or group decision making. Ernest Forman worked closely with Thomas Saaty to design Expert Choice.

In 1978, Dan Bricklin and Bob Frankston co-invented the software program VisiCalc (visible calculator). VisiCalc provided managers the opportunity for hands-on computer-based analysis and decision support at a reasonably low cost. VisiCalc was the first killer application for personal computers and made possible the development of many model-oriented, personal DSSs for use by managers. The history of microcomputer spreadsheets is described in Power (2000). In 1987, Frontline Systems founded by Dan Fylstra marketed the first optimization solver add-in for Microsoft Excel.

In a 1988 paper, Sharda et al. reviewed the first 15 years of model-driven DSS research. They concluded that research related to using models and financial planning systems for decision support was encouraging but certainly not uniformly positive. As computerized models became more numerous, research focused on model management and on enhancing more diverse types of models for use in DSSs such as multicriteria, optimization, and simulation models.

The idea of model-driven spatial Decision Support System (SDSSs) evolved in the late 1980s (Armstrong et al., 1986) and by 1995 the SDSS concept had become firmly established in the literature (Crossland et al. 1995). Data-driven spatial DSSs are also common.

Knowledge Driven DSS, Data Mining and Management ES Application These DSS involve the application of knowledge technologies to address specific decision support needs. Essentially, all artificial intelligence based DSS fall into ANN and ES are included here. Because the benefits of these intelligent DSS or knowledge based DSS can be large organization have invested in them. These DSS are utilized in the creation of automated decision making system. The basic idea is that rules are used to automate the decision-making prcess. These rule sare basically either an ES or structured like one. This is important when decisions must be made quickly, as in many e-commerce situations.

Knowledge-driven DSSs can suggest or recommend actions to managers. These DSSs are person-computer systems with specialized problem-solving expertise. The expertise consists of knowledge about a particular domain, understanding of problems within that domain, and skill at solving some of these problems (Power 2002). These systems have been called suggestion DSSs (Alter 1980) and knowledge based DSSs (Klein and Methlie 1995). Goul et al. (1992) examined artificial intelligence (AI) contributions to DSS.

In 1965, a Stanford University research team led by Edward Feigenbaum created the DENDRAL expert system. DENDRAL led to the development of other rule-based reasoning programs including MYCIN, which helped physicians diagnose blood diseases based on sets of clinical symptoms. The MYCIN project resulted in development of the first expert-system shell (Buchanan and Shortliffe 1984).

Bonczek et al.s (1981) book created interest in using these technologies for DSSs. In 1983, Dustin Huntington established EXSYS. That company and product made it practical to use PC based tools to develop expert systems. By 1992, some 11 shell programs were available for the MacIntosh platform, 29 for IBM-DOS platforms, four for Unix platforms, and 12 for dedicated mainframe applications (National Research Council 1999). Artificial Intelligence systems have been developed to detect fraud and expedite financial transactions, many additional medical diagnostic systems have been based on AI, and expert systems have been used for scheduling in manufacturing operation and web-based advisory systems. In recent years, connecting expert systems technologies to relational databases with web-based front ends has broadened the deployment and use of knowledge-driven DSS.

Trends suggest that data-driven DSS will use faster, real-time access to larger, better integrated databases. Model-driven DSS will be more complex, yet understandable, and systems built using simulations and their accompanying visual displays will be increasingly realistic. Communications-driven DSS will provide more real-time video communications support. Document-driven DSS will access larger repositories of unstructured data and the systems will present appropriate documents in more useable formats. Finally, knowledge-driven DSS will likely be more sophisticated and more comprehensive.

Holsapple and Whinston's Classification

Holsapple and Whinstone (2000) classified DSS into the following six frameworks: text-oriented DSS, database-oriented DSS, spreadsheet-oriented DSS, solver-oriented DSS, rule oriented DSS and compound DSS. Essentially, these frameworks readily map into the AIS SIGDSS categories:

- The *text-oriented DSS* are the same as the document driven DSS.
- the *database-oriented DSS* are the data- Database-oriented DSS.

- The *Spreadsheet-oriented DSS* are another form of model-driven DSS in which the functions are add-in programs of the spreadsheet are used to create and manage the models because packages such as Excel can include a rudimentary DBMS or can readily interface with one, they can handle some properties of a database-oriented DSS, especialy the manipulation of descriptive knowledge.

- The *Solver-oriented DSS* map directly in to model-driven DSS.

- The *Rule-oriented DSS* include most knowledge-driven DSS, Data Mining and management ES applications.

- The *Compound DSS* The compound DSS integrates two or more of those cited above and is defined the same as by the SIGDSS. (DSS BIS Pg: 101)

Alter's Output Classification

Alter's (1980) classification is based on the "degree of action implicaiton of system output" or extent to which system outputs can directly support (or determine) the decision. According to the classification there are seven categories of DSS.

- *File drawer systems* This type of DSS primarily provides access to data stores or data related items.

- *Data analysis systems* This type of DSS supports the manipulation of data through the use of specific or generic computerized settings or tools.

- Analysis information systems. This type of DSS provides access to sets of decision oriented databases and simple small models.

- *Accounting and financial models* This type of DSS can perform 'what if analysis' and calculate the outcomes of different decision paths.

- *Representational models* This type of DSS can also perform 'what if analysis' and calculate the outcomes of different decision paths, based on simulated models.

- *Optimization models* This kind of DSS provides solutions through the use of optimization models which have mathematical solutions.

- *Suggestion models* This kind of DSS works when the decision to be taken is based on well-structured tasks.

3.1.3 Components of Decision Support System

According to Turban et al. (2011), A DSS can be composed of a data management subsystem, a model management subsystem, a user interface subsystem and a knowledge-based management subsystem. Figure 3.3 shows the full overview of the components of DSS.

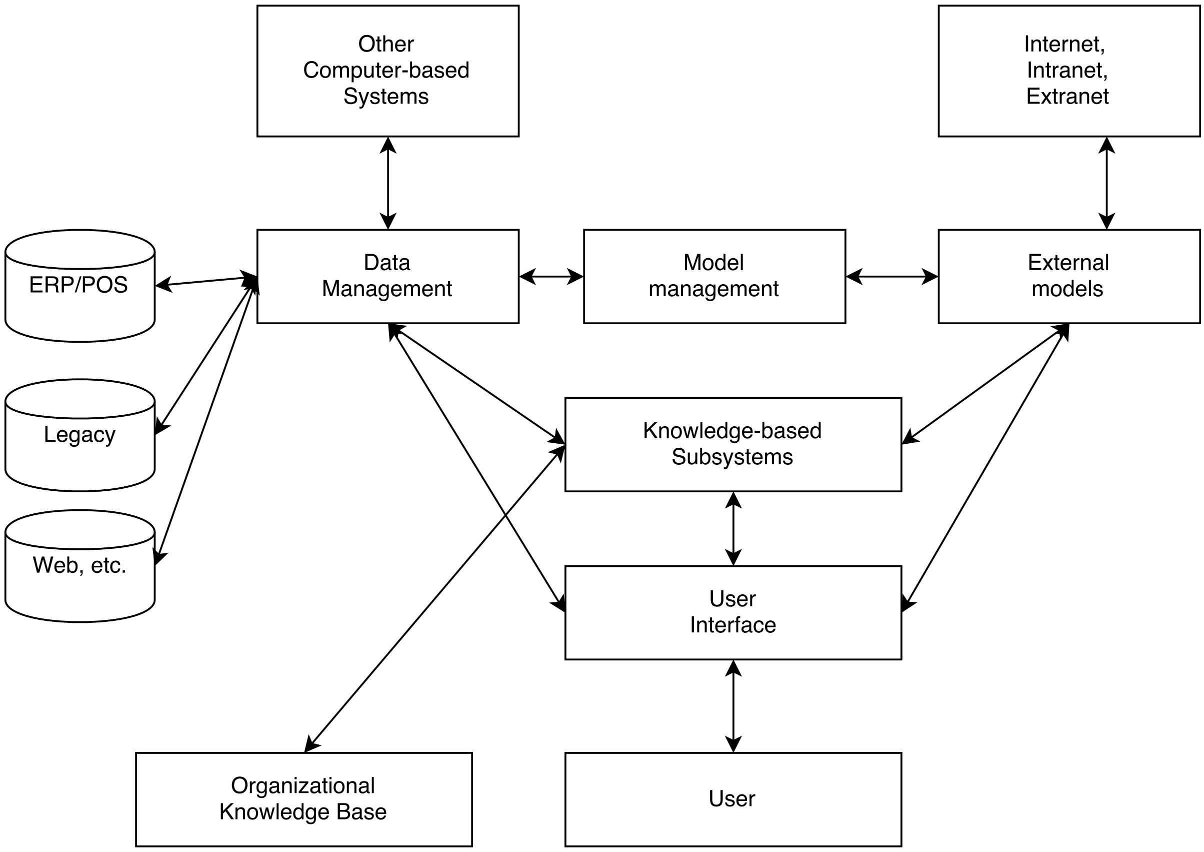


Figure 3.3: Different Key Characteristics and Capabilities of DSS

The Data Management Subsystem

Data management subsystem contains a database that comprise relevant data generated by a system or collected manually and is managed by software called the database management system (DBMS). The data management subsystem can be brought together with the corporate data warehouse, a repository for corporate decision-making data. Usually, the data are stored or accessed via a database Web server with proper authorization. The data and information used in Decision Support System comes from one or more of three sources:

Organizational information; *We may want to use virtually any information available in the organization for Decision Support System proposed by management. What we use, of course, depends on what our need and whether it is available. It is possible to design your Decision Support System which can have access to this information directly from companys database and data warehouse. However, specific information is often copied to other repository dedicated for Decision Support System to save time in searching through the organizations database and data warehouses.*

External information; *some decisions require input from external resources of information. Various branches of public data sources, Dow Jones, Compustat data, and the Internet, other open source data platform to mention just a few, can provide additional information for the use with a Decision Support System.*

Personal information; *we can incorporate our own insights and experience our personal information into Decision Support System. We can design our Decision Support System so that you enter this personal information only as needed, or you can keep the information in a personal database that is accessible by the Decision Support System.*

The Model Management Subsystem

; Model management subsystem is a system that consist of financial, statistical, management science or other process models related to enterprises including public and priveate both. Available Modeling languages for building customized system and process models are also included in model management system. This software is often called a model base management system (MBMS) this component can be connected to corporate or external storage of models. Model solution techniques are management systems are implemented in web development system (such as java, python, c# and .net, ruby, scala etc) to run on application servers. Models can take various following forms:

Businesses Use Models; *This kind of model analyse variables and their relationships. Mostly statistical are used for this kind of model. For example, we can use a statistical model called analysis of variance (ANOVA) to determine whether newspaper, TV, and billboard advertising are equally effective in increasing sales or one is better than the other.*

Decision Support Systems; *In various decision-making situations by utilizing decision making tools that allow us to analyze information in several different ways. The* *models we use in a Decision Support System depend on the goal we want to achieve, consequently, the kind of analysis we require. For example, if we would use what-if analysis to see what is the effect of changing of one or more variables will have on other variables, or optimization to find the most profitable solution given operating restrictions and limited resources as a part of operational research. Even Spreadsheet software such as excel can be used as a Decision Support System tool for basic what-if analysis though with the blessing of modern tool and programming platform other advance tolls is more preferable.*

The model management system*; The model management subsystem stores and maintains the Decision Support System's models. Its functionality is similar to the characteristics of database management system. The model management component can not select the best model to use for a particular problem and it requires expertise and experience but it can help us to create and manipulate models quickly and easily.*

The User Interface Subsystem

The user communicates and execute commands in the DSS through the user interface subsystem (GUI). So the user can be considered as a part of the system. Researchers also assert that some of the unique contributions of DSS are derived form the agile interaction between the computer and the decision maker. The modern web browser equiped with HTML5, CSS3 and javascript can provide familiar, consistent graphical user interface (GUI) structure for most DSS. For locally used DSS a spreadsheet also provides a familiar user interface but it is not interactive as web browser.

The user interface is the built in part of the system. The results and outcome goes visible though it. This is the only component of the system with which user communicate directly. If we have a Decision Support System with a poorly designed graphical user interface and if it is too rigid or too difficult to use then we simply will not use it no matter how much capable it is. The best user interface uses build up on required terminology and methods and is flexible, consistent, simple, and adaptable.

The Knowledge-Based Management Subsystem

The knowledge-based management subsystem can work as a subsystems or can act independently. This system provides knowledge and intelligence to augment the decision maker's own. It can be connected directly with the organization's knowledge base or data base, which is sometime s called the knowledge base or knowledge component of organization. Knowledge may be provided via development framework such as java, python, .net and are easy to integrate into the other decision support component.

For example, when analyzing the production of parts in a factory environment, a Decision Support System should signal if the quality of production exceeds the quality margin that the projected staff can service. Such signaling requires the Decision Support System to incorporate some algorithm for quality control and exact threshold or rules of thumb . Such rules of thumb, also considered as heuristics, make up the knowledge base.

By nature a Decision Support System must include the three major component: DBMS, MBMS, and User Interface. The knowledge-based management subsystem can be optional, but it is also able to provide many benefits by providing intelligence and knowledge to the three major components as in any other MIS, the user may be considered a component of DSS.

We can consider an example to realise the components of a Decision Support System, lets consider the Decision Support System that Lands End has tens of millions of names in its customer database. It sells a wide range of clothing for all different gender, age and various household wares as well. To match the right customer with the catalog, lands end has identified twenty different target markets. Customers in these target markets receive catalogs of merchandise that they are likely to buy. End the expense of sending catalogs of all products to all customers in the list. To predict customer demand end needs to continuously monitor buying trends. And to meet that demand, lands' end must accurately forecast sales levels over time to time. To accomplish theses goals, it uses a Decision Support System which performs three tasks as follows:

Data Management: *The Database Management System (DBMS) stores customer and product information and could be relationnal or not relatoinal DBMS. In addition to this organizational information, Lands' End also needs external information from open source or Private databases, such as demographic information and industry and style trend information.*

Model Management: *The Decision Support System has to processes to analyze the information follow thorug some descriptive methods and predictive modeling as well. The models create new results or knowledge component that decision makers need to plan product lines and inventory levels. For example, Lands End uses a statistical model* *called multiple regression analysis to determine trends in customer buying patterns and forecasting models to predict sales levels.*

User Interface Management: *A user interface enables Lands End decision makers to access output processed high Level information and to specify the models they want to use to create the knowledge they need.*

3.1.4 Difference Between MIS and DSS

Management Information Systems (MIS) and Management Information Systems (DSS) are two abbreviations that are often heard in the field of Business Management. They differ in a few aspects.It is interesting to note that MIS is a type of link that assists in the communication between managers of various disciplines in a business firm or an organization. On the whole it plays a very important role in building up communication among the corporate people.

DSS on the other hand is an improvement of the concept of MIS. It is true that both of them differ in terms of their focus. DSS focuses more on leadership. It is all about senior management in a firm providing innovative vision. MIS focuses more on the information gathered and the information that has poured from different quarters. Experts on managerial behavior say that DSS focuses more on decision making. MIS on the other hand focuses more on planning the report of various topics concerned with the organization that would assist the managers to take vital decisions pertaining to the functioning of the organization.

One of the finest differences between MIS and DSS is that MIS focuses on operational efficiency whereas DSS focuses more on making effective decision or in other words helping the company to do the right thing. Flow of information is from both sides, up and down in the case of MIS. On the other flow of information is only upward in the case of DSS.

In the case of DSS the report can be flexible whereas in the case of MIS the report is usually not flexible. MIS is characterized by an input of large volume of data, an output of summary reports and process characterized by a simple model. On the other hand DSS is featured by an input of low volume of data, an output of decision analysis and a process characterized by interactive model.

DSS are used to provide support to analysts and decision makers within an organization that are relevant to a specific problem or situation and make an evaluation of various different outcomes. They are used by managers and other decision-makers in both unstructured and semi-structured situations and largely in ad-hoc situations although it may involve repetitive decisions. A DSS is used in situations where individual managerial judgment is required.

An MIS provides routine information typically on an ongoing basis in a standardized format. Reports may be run routinely or be provided on demand. While an MIS can be used to solve standardized and routine problems using specified criteria and the data from an MIS forms the base pool of information for a DSS, MIS are also used for routine functions such as production control and monitoring, forecasting, human resources management, financial analysis and research. Because of the broad nature of the information presented, an MIS might be used by lower, middle or upper level management.

While an MIS simply gathers data, a DSS manipulates that data and helps to develop tools that aid in the decision-making process.

3.2 Data Mining (DM) in Decision Support System

Data Mining (DM) assists in finding patterns within business and scientific or research data. So It does not act as a standalone process since it is data driven and human interpreted. It requires a human to realise the knowledge obtained from DM technologies and apply them in problem solving. DM technologies can be scaled to handle significantly large amount of data and to assist in the automation of the knowledge discovery process. Statistics, signal or image processing, artificial intelligence, machine learning, database query tools such as SQL, econometrics and economics, management science, domain-knowledge-based numerical and analytical methodologies, and nonlinear dynamical and stochastic systems such as Markov Model, Bayesian Statistics are examples of the fields that have contributed to the current range of DM technologies. DM tasks can be both descriptive such as clustering with hierarchical or non-hierarchical, correlations, pattern recognition, dimension reduction, and frequent item sets and predictive such as classification, linear and non - linear regression, survival analysis (Burstein and Holsapple, 2008).

The ability to anticipate, react, and adapt to changes in the market and the capability to implement appropriate business strategies are two of the most indispensable characteristics of any successful company. Until recently, uncertainties inherent to forecasting models drove many companies to rely on decision making based on the gut feel of senior management rather than on verifiable data. As business conditions become more dynamic and active responses to subtle changes prove more valuable, continuous visualization and monitoring of the state of business through tools and technologies such as Data Mining (DM), Decision Support Systems (DSS) and business intelligence (BI) are becoming increasingly important. Correspondingly, the development of predictive modeling capabilities is becoming more prevalent for the support of decision making through the verifiable data presented in DM technologies. To date, DM and DSS have generally been studied in relative isolation. Frameworks that bridge the gap between DM analysis and predictions to actions and decisions in decision support are required to ensure better integration of the two methodologies. In this section, wediscuss the use of DM technologies in the creation of DSS and allied systems for human use (Burstein and Holsapple, 2008).

Gorry and Morton (1971) suggested the concept of DSS by integrating decision types, i. e., unstructured, semistructured, and structured, given by Simon (1960) and management activities, i. e., strategic planning, management control, and operational control, given by Anthony (1965). A new paradigm for DSS was suggested by Courtney (2001). In this paradigm, every step of the process influences the centralized knowledge model, which consists of recognizing the problem, creating perspectives to understand the nature of the problem, finding possible solutions, and updating the model continually. DSS is a simple combination of technologies that embed DM techniques and facilitate analysis of what-if scenarios. Every day, billions of bytes of new, unique data are recorded in enterprise-scale data warehouses and databases. Recent advances in remote or other sensor technologies have added to the magnitude and frequency of such data collection. Instead of trying to obtain more information or to design better data management systems, today's challenge lies in how to make the most of such data (Burstein and Holsapple, 2008).

In order to make a right decision, the managers need knowledge. In case of processing large amount of data, issues may occur because of data analysis and necessary knowledge extract. Generally data is analyzed through an automated process with the help of programming languages or enterprise software solutions, known as Knowledge Discovery in Data Mining techniques. Data Mining can be defined as a method of exploring and analysis for large amounts of data with a particular target on discovering significantly important patterns. Data Mining helps finding required knowledge from raw, unprocessed data which could be in structured, semi-structured or unstructured. Using Data Mining techniques allows extracting knowledge from the data mart, data warehouse and, in particular cases, even from operational databases (Burstein and Holsapple, 2008).

In this context, Data Mining plays an necessary and important role in helping organizations to understand their customers and their behavior, keeping clients, stocks anticipation, sale policies optimization as well as other benefits which bring a considerable competitive advantage to the organization. The main purpose of these techniques is to find patterns and hidden (but relevant) relations that might lead to revenue increase and think new way of organising enterprise rules and action for better services. According to Burstein and Holsapple (2008), The essential difference between Data Mining techniques and the conventional database operation techniques is that, for the second ones, the database becomes passive and is only being used for large amounts of data population, therefore helping in future finding of that specific data. Alternatively, the database is not passive anymore, being able to serve useful information regarding the business plans put in discussion.

3.2.1 Data Mining Processes

In order to systematically carry out Data Mining projects, a common process is usually followed. Based on best practice and experience Data Mining researchers, practitioners, business analyst have proposed several processes to maximize the chances of successes in conducting Data Mining projects (Turban et al., 2011). These efforts have led to several standardized processes, some of which are described in this section.

From the standardized processes, arguably the most popular and mostly used one Cross-Industry standard Process for Data Mining **CRISP-DM** was proposed in the mid 1990s by a European consortium of companies to serve as a non proprietary standard methodology for Data Mining (CRISP-DM, 2009). Figure 3.4 illustrate this proposed process, which is a sequence of six steps but can be applied through parallel precessing that starts with a good understanding of the business and the need for the Data Mining project and finishes with the installation of the solution that satisfied the specific business need. Even though these steps are sequential and depends on previous section in nature

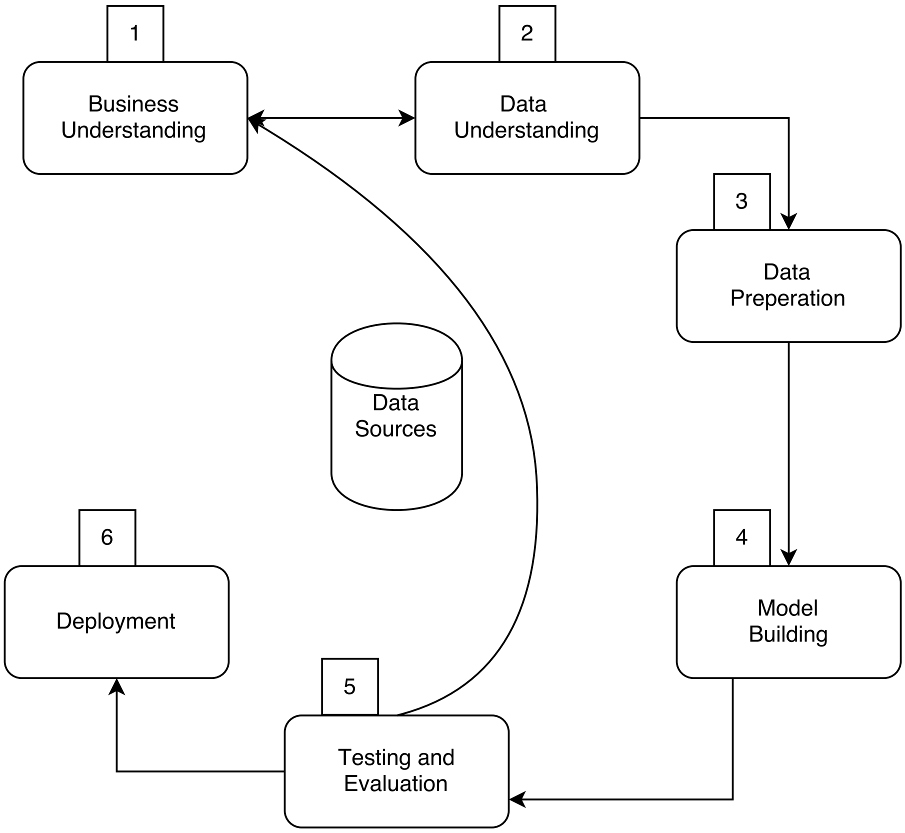


Figure 3.4: Steps of CRISP-DM Mining Processe

there is usually a great deal of backtracking. Because the Data Mining is driven by experience and experimentation, depending on the problem situation and the knowledge of the analyst, the whole process can be very iterative and time consuming if the data does not directs to some feasible solution at the end. Because one steps are build on the outcome of the former ones, one should pay more attention to the beginning steps in order not to put whole study on an misleading path from the onset.

3.2.2 Data Mining Methods

Regarding Data Mining studies, two major types of method exists. One of them is represented by the statistical inference or hypothesis testing, which assumes exposing a theory regarding the relation between actions and their results. The second type of study is represented by the knowledge discovery from existing information. For both type of analysis, relations between data warehouse existing data are tracked time to time. This can be done by using third party data viewing tools or by using fundamental statistical analysis, such as correlation and regression analysis. Data Mining techniques reside from classic statistical and mathematical calculation, from database administration and from advanced artificial intelligence. They are not a substitute for traditional statistical techniques, but an extension of graphical and statistical techniques and built on top of traditional technique (Turban et al., 2011).

Data Mining uses a large variety of statistical and mathematical algorithm, formula shape recognition, classification, fuzzy matching logic, machine learning, genetic algorithms, neural networks, data viewing etc., from which we can mention regression algorithms, decision algorithms, neural networks, clustering analysis. Figure 3.5 shows a complete hierarchy of data mining task (Turban et al., 2011).

Summarization. Summarizing or aggregating data is the best abstraction or generalization of data. A set of task - relevant data is summarized and abstract, resulting a smaller set which gives a general overview of the data and usually with aggregation information. For example, the long distance calls of customer if stored in second unit the it can be summarized in the total minute, total spending in money, total call per day etc. such high level, summary information instead of detail calls to present the scales managers for customer behaviour analysis. The summarization can go up to different abstraction levels and can be viewed from different aspects. For example, the calling minutes and splendid can be totaled along the calling period in a weeks, months, quarters or years and for large data more than yearly but also we do not want to loose a lot of information when summarizing the data (Fu, 1997).

Regression algorithms. Regression is one of the old and basic statistical method for finding relation and dependency between variables. In the case of Data Mining, it is also an important analytical method, used in classification applications through logical regressions as well as forecasted reports measured using the least square or other methods. Non-linear data can be transformed through normalization or other standard method into useful linear data and analyzed using linear regressions. The universal test for Data Mining classification is the coincidence index matrix. It is primarily focused on data classification abilities of the model. For continuous regressions, class inflection points must be identified which is controlled by minimizing mean square error when estimation the model. The applications of the methods into solving business problems are multiple so some other advanced regression such as multivariate regression, non liner regression and time series is introduced (Fu, 1997).

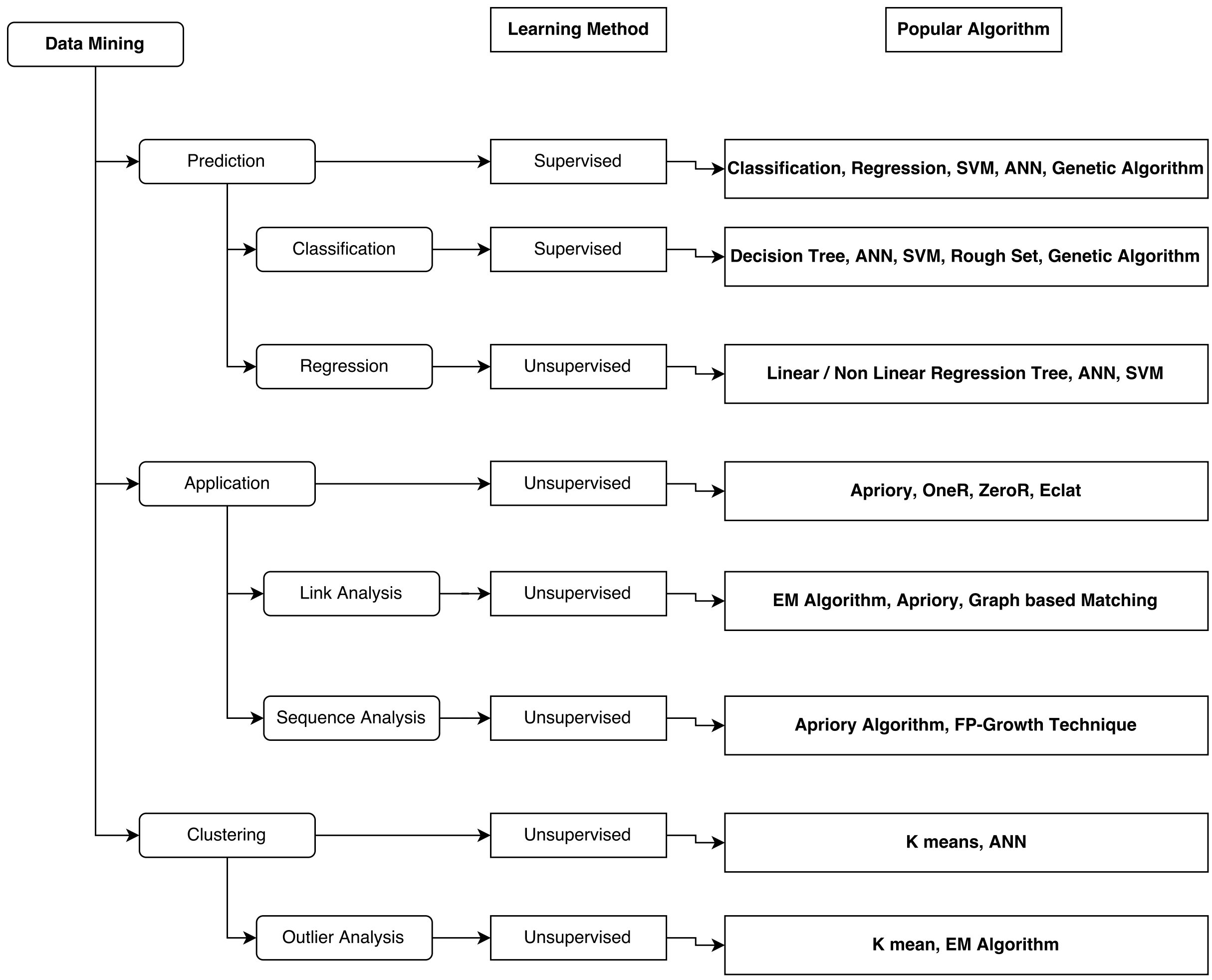


Figure 3.5: Hierarchy of Data Mining Tasks

Decision trees. In Data Mining technology, decision trees refers rules tree-view arrangement, also known as joining rules. The trees' construction mechanism of the trees consists in collecting all the variables the analyst assumes might help the decision making and analyzing them considering their influence into result estimation. The algorithm automatically determines which of the variables or attribute are the most relevant, based on the ease of data sorting. The decision tree algorithms are applied in Business Data Mining in areas like: loan request classification, applicants ranking for various positions (Fu, 1997).

Neural networks. This is one of the most commonly used Data Mining method in advanced level. It consists of taking sets of observations and placing them in a relational system through arc-connected nodes and layers. This idea derives from the way neurons act inside of the human brain when passing signal with out loosing any information. Neural networks are usually structured in at least three layers, having a constant structure allowing reflection of complex non-linear relations. Each entry data has a node in the first layer, while the last layer represents the output summarised data the result. In order to classify the neural network model, the last layer (containing the output) has a corresponding node for each section or category. In most of the cases, this type of networks also have a mid node layer (hidden) which adds complexity to the model. The obtained results are compared to the targeted ones, and the difference is re-entered in the system for node's cost adjustments. The process keeps looping until the network correctly classifies the input data (Fu, 1997).

Clustering analysis. One of the most general and over used un supervised learning forms, of this type of analysis allows the algorithm to determine the number of subsets based on euclidean distance. Partitioning is mainly used for defining new variable categories, which divide raw data in a precise number of regions or neighbourhood (k-means clustering) in a two dimensional space. Considering a random number of centers (k), data is associated to the center which is the closest to it. The basic principle of this analysis is to identify the average characteristic for different indicators in sets of data. Thus, new observations can be measured by reporting the deviation from the average. This analysis is often the base technique applied in a Data Mining study, being used in client segmentation and, implicitly,taking a segment-oriented action (James et al., 2013).

Business planning, forecasting, and decision support applications frequently need to analyze data from diverse source such as past data warehouses and data marts, syndicated data vendors, legacy systems, and the Internet and other public-domain sources. In addition, data entry from external or internal collaborators, professional consultants, and decision makers and executives must be able to be read in real time and/or in increments. Data from different sources are usually mapped to a predefined data model and implemented through extraction, transformation, and loading (ETL) tools.

The use of automated and sophisticated DM techniques to support pre-established DSS tasks is one biggest implementation of emerging technologies. The planning process results are normally published or exported in a predefined placeholder (e. g., a relational database table) that is accessible to operate systems and other planning applications based on schedule. The current emphasis in the DM and DSS communities is on the development of algorithms, practices, and systems that apply new methodologies and scale to large data repositories. The implementation of domain knowledge is more critical and complex however, in scientific applications. DM technologies have the potential to modernize scientific discovery when they are combined with domain-specific knowledge about the physics of data sources, and allied uncertainties, verification, and prediction aspects. This potential is demonstrated in emerging applications related to remote sensing (James et al., 2013).

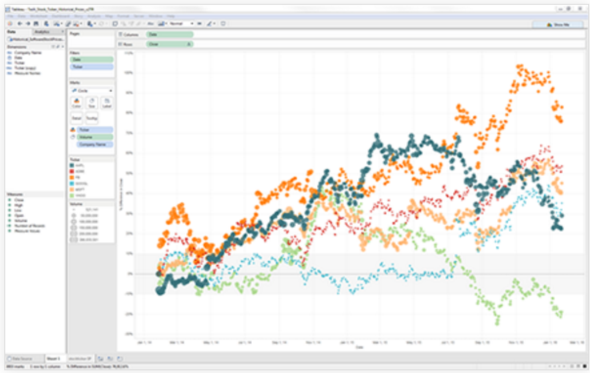
3.3 Information Visualization for Decision support

Information visualization is a process of contracting a visual presentation of abstract quantitative data in a aggrigate form. The main characteristics of visual preparation enable humans to recognize patterns, trends and anomalies inherent in the data with little effort in a visual display. Visualization are there for widely used in contemporary decision support at large enterprises and final steps of building a decision support system.

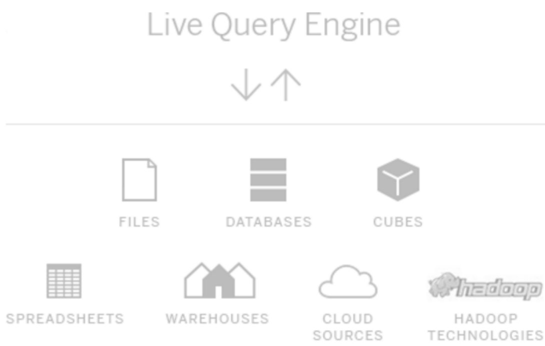
The visual interface are called dashboard now a days, are tools for reporting the status of a company and its business environment facilitate management activities and decision support for customer level. By the grace of modern technology now Decision Support Systems are not considered as a piece of software any more. Rather that building software, enterprises are now more interested to create web application which can be run on cloud and can be accessed any time from any where.

Such kind of decision support application are often created using different technology such as combination of HTML5, CSS3 and Javascript or Combination of $C\#$ and .net or Java. But building decision support application based on those frame works, the process will be time consuming. Also we need significant amount experts to build those application which could be expensive toll. So now companies are interested on subscription based third party software such as Tablau, Microsoft Power BI, DOMO, Good Data for building Decision Support Application and reporting. Those application are user friendly, cost effective and also saves a large amount of money in enterprises. Rather that making Decision Support System as a piece of software companies can build several Decision Support System Application using one software which is the main benefit using those software and also subject of interest of that paper. Through our study we have used Tableau for visualization which offers several different functionality as follows:

1. **VizQL:** Visual query language that translates the performed drag-and-drop ac-



(a) Visual Query Interface



(b) Live Query Engine Connection Supports

Figure 3.6: Tableau Features and Functionality

tions over interface into data queries and then expresses the result data visually. Visual Query supply dramatic gains in person's ability to follow and understand data by abstracting the underlying complexities of query itself and analysis. The outcome is an intuitive user experience that lets people answer questions, make decision analyse system or process as fast as they can think of them. Visual Query represents a foundational advancement in the area of data analysis, data modeling and visualization for building quick analytics from raw data.

2. **Live Query Engine:** Tableau includes Live Query Engine which is the first technology to let people natively query databases, cubes, warehouses, Cloud sources, even Hadoop and amazon data system without any knowledge of SQL or other relevant programming or advance development. It lets people query several different and diverse data sources with a point-and-click interface anyone can use but the query runs in the back end. People connect to database csv file system or plain

text of any size with a few easy clicks in Tableau interface. Layer in additional data sources on the fly. Connecting to and combining data by joining are easy enough for database novices to achieve like relational database system.

3. **In-Memory Data Engine:** For significant large database and tables When data sources are slow or just not right for quick operation, in that case Tableau support own build Data Engine or data extract file (TDE). To overcome the limitations of existing databases and data silos, Capable of being run on ordinary computers the in momery engine is introduced, it leverages the complete memory hierarchy from disk to cache with its possible optimization power but there is possibilities of overflow is computer memory is too small. It drastically shifts the relation between big data and fast analysis and also make it easy to work with. Figure 3.6 shows, the Visual Query interface of Tableau Desktop and data connection capability of Live Query Engine.

3.4 Implementation of DSS and DM in Enterprise

Initial implementations of Data Mining methods and technologies in interpreting business,scientific open data could be data driven (descriptive analysis) or model-driven (predictive) DSS. Which involves the development of predictive models from areas such as management science, operations research, optimization, Monte Carlo simulation, Baysian Statistics and dynamic programming, pattern recognition. Machine Learning, Cognitive science, AI and expert systems, and traditional algorithm based software engineering also contribute to the design of DSS. Generally DSS can cater to multiple audiences as producers, suppliers, and customers share the same results and information through collaborative planning, forecasting, and replenishment processes. Building Full collaboration, development, and sharing of data can lead to one-number forecasts and be used to extend supply chains (Burstein and Holsapple, 2008).

For example, Key performance indicators (KPIs) could be used to trace critical factors on an schedule basis. On-line analytical processing (OLAP) tools allow for easy, quick, consistent, and interactive analyses of multidimensional aggregated enterprise data, and provide decision maker with the ability to gain insights into knowledge hidden in large databases. Some of the current features of OLAP are currently used is Drill-down, slicedice, reach-through, and rotation (Burstein and Holsapple, 2008).

Some functions and application of Data management technologies, DSS, OLAP, geographical information system (GIS), and others are contained within the broad spectrum of analytical information technologies (AIT) as shown in Figure 3.7.

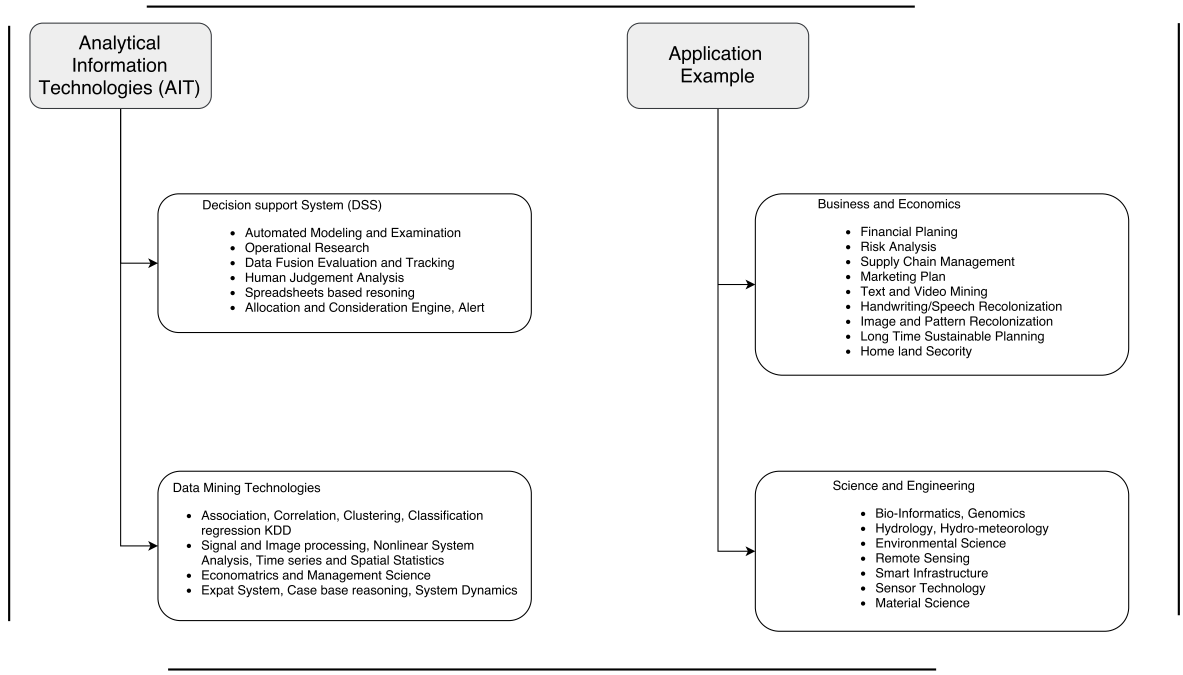


Figure 3.7: Analytical information technologies and their applications

3.5 The process of developing a DSS using Data Mining techniques

Building Decision Support Systems from scratch involves time, high-costs and human resources efforts and the success of the system can be affected by many risks like: system design, data quality, and technology obsolescence. Main objective of Decision Support Systems is to assist the managers and executives business analyst to make decision regarding the benefit of investment, budgeting cash flows and financial planning, especially in the case of public funds.

Presently, many institutions invest in building organizational data warehouses and data marts in order to increase the performance and the efficiency of the analytical reporting activity. Also, there are several expensive tools and software that can be used to analyze the trends and to predict some future characteristics and evolution of the business. Some of these tools analyze data from the statistic perspective or by using neural networks. In our opinion, in order to build an efficient Decision Support System there must be combined several techniques and methods that can improve the performance and the accuracy of the analysis from two major perspectives: historical data and forecasts. This requirement can be obtain by combining data warehousing, OLAP, Data Mining and business intelligence tools for analyzing and reporting into a flexible architecture that must contains: A data model's level where an ETL process must be apply to clean and load data into a data warehouse or data marts; An application level with analytical models where multidimensional reporting like OLAP and Data Mining techniques can be combined to for historical and forecast analysis; An interface level where dashboards and reports can be build with business intelligence tools (Bâra and Lungu, 2012).

These stages can be adapted and applied in Decision Support Systems, but during the development cycle it is mandatory that differences between general system modeling and Decision Support Systems modeling must be treated separately, in order to obtain a successful business requirements of implementing the specifications explained by Bâra and Lungu (2012) as followes:.

Stage 1. The feasibility study consists of identifying the requirements and business opportunities and proposing solutions of improving the decision making process. Each of the proposed solutions must be justified by the implied costs and benefits.

Stage 2. Project planning consists of evaluating project sustainability possibilities, identifying existent infrastructure components and future needs. The result of these activities concludes with the project plan. After its validation and approval, the effective start of the project can begin.

Stage 3. Business requirements analysis. This stage focuses on detailing and analyzing on priority the initial requirements of the organizational management team. Usually, the requirements are identified based on interviews conducted by managers and the project staff. These requirements might suffer slight changes during the project, but the development team must make the managers aware of the capabilities and limitations of a DSS, therefore reducing the risk of un-feasible business requirements to occur.

Data analysis the biggest challenge of a Decision Support System development project consists of identifying necessary data, analyzing its content and the way it relates to other data. Data analysis is focused on business analysis rather than system analysis performed in traditional methodologies. It is preceded by a data cleaning activity.

Data cleaning implies transforming and filtering data sources in order to be used in building the destination module the analysis module. This process is done by: identifying necessary data from the functional modules; analyzing the content of the selected data sources; selecting the appropriate data for the project; implementation of data filtering related specifications; selecting the tools to be used in the filtering/cleaning process. During the source selection process, a few key aspects must be taken into consideration: data integrity, precision, accuracy and data format. These facets are critical in regards to the success of the new ETL process.

An important step in this stage is choosing the technologies used in the prototype development and, later on, in the final system. Based on a comparative analysis over advantages and disadvantages brought by each of the technologies on the project, different approaches might be taken into consideration: usage of data warehouses, including OLAP (Online Analytical Processing) functionalities, usage of knowledge extract algorithms, data source integration tools or, on a final phase and assuming a parallel approach on building the system has been taken, usage of applications integration tools.

Stage 4. System design. Database / data warehouse design. According to the system's requirements, the necessary data will be stored both on a detailed level as well as on aggregate level, therefore relational, object-oriented or multi-dimensional data storage approaches might be taken. During this sub-phase, the logical data model is refined and detailed and the physical model of the new system is developed in order to satisfy the reporting and analysis requirements of the managers. While on "Data analysis", the process has been oriented to data sources (data-in or dataentry) coming from operational modules, in this phase the targets or data destinations (dataout) are set aiming on reports, analysis and queries. Therefore, a list of best practices must be taken into consideration:

Due to the above mentioned aspects, we recommend that the storage, management and data processing solution to consist of a centralized data warehouse on an organizational level. Following logical and physical criteria, the data warehouse can be divided into data marts on departmental level, thus being easier to maintain and developed by separate teams, following the same set of specifications.

The ETL (extract / transform / load) process design this phase is the most complex one in the projects lifecycle and is directly dependant on the data sources quality.

The integration of all the destination databases in a single environment and building the ETL process on it, avoiding a separation of each destination module, thus mitigating the risk of distinct data marts is recommended. The strategy of building data marts in the same environment is also viable, but only on the condition that these are already integrated. The important fact here is that the ETL process must remain the same for all levels (Bâra and Lungu, 2012).

The design of the ETL process needs a series of pre-requisite stages: preliminary processing of data sources, in order to have a standardized format, data reconciliation and redundancy and inconsistency elimination of data.

The steps to be taken in creating an ETL process are the following:

1. Creation of transformation specifications (mapping) of the sources in regards to the specific destinations. This may be done as a matrix or as transformation diagrams.

2. Choosing and testing the ETL tools to be used. At the moment, a series of ETL process modeling and implementation tools exist on market, but choosing one of them would depend on the features they provide and on the support of data source integration inside the same transformation process.

3. The ETL process design several extract and transform operators are used, depending on the data model (sorting, aggregation, joining, dividing operators, etc.). The process can be split into sub-processes that would run separately in order to minimize the execution time. The execution flow of the process will be modeled using flow diagrams.

4. ETL programs design. Depending on the program in which the data is loaded, three phases of data loading are applied:

    - Initial load - the initial load of destinations with current operational data
    - Historical load - the initial load of destinations with archived historical data
    - Incremental load - regular loading of destinations with current data coming from

5. Choosing the environment for running the ETL process represents the decision over using a dedicated server / machine or the process would be divided and run

decentralized. The decision depends on the available resources and on the processing time, as well as on the timeliness that the process is scheduled to run.

The results of these activities is materialized in the data mapping documentation, the flow diagram / diagrams of the ETL process, the transformation programs documentation and the process execution specifications.

Stage 5. Building the system. The technologies that are used for Decision Support Systems development are part of the business intelligence technologies category and consist of: technologies for data warehouse data organization, Data Mining algorithms, extract, transform and load (ETL) tools, CASE (Computer-Aided Software Engineering) modeling tools and web technologies.

Stage 6. System implementation. Represents the stage when the system is being delivered, training sessions are held for implied managers / business owners, the necessary technical support is provided, data loading procedures are run, the application is installed and the performance is being tracked. The stage ends with the release of the system into production (commercial go-live) and with the delivery of the utilities and final project documentation, the user guides and presentation manuals for the application.

A process mode for DMDSS

One of the key issues in the design of DMDSS was to determine the Data Mining process model. The process model for DMDSS is based on the CRISP-DM (CRoss Industry Standard Process for Data Mining) process model. CRISP-DM process model breaks down the Data Mining activities into the following six phases which all include a variety of tasks : business understanding, data understanding, data preparation, modelling, evaluation and deployment. CRISP-DM process model was adapted to the needs of DMDSS as a two stage model: the preparation stage and the production stage. The division into two stages is based on the following two demands. First, DMDSS should enable repeated creation of Data Mining models based on an up-to-date data set for every area of analysis. Second, business users should only use it within the deployment phase with only the basic level of understanding of Data Mining concepts. Area of analysis is a business domain on which business users perform analysis and make decisions (Bâra and Lungu, 2012).

According to Bâra and Lungu (2012) The preparation stage represents the process model for the use of DMDSS for the purposes of preparation of the area of analysis for the production use (Fig 3.8). During the preparation stage, the CRISP-DM phases are performed in multiple iterations with the emphasize on the first five phases starting from business understanding and ending with evaluation. The aim of executing multiple iterations of all CRISP-DM phases for every area of analysis is to achieve step-by-step improvements in any of the phases. In the business understanding phase, slight redefinitions of the objectives can be made, if necessary, according to the results of other phases, especially the results of the evaluation phase. In the data preparation phase the improvements in the procedures which execute recreation of the data set can be achieved. The data set must be recreated automatically every night based on the current state of data warehouse and transactional databases. The problems detected in the data preparation phase can also demand changes in the data understanding phase (Bâra and Lungu, 2012).

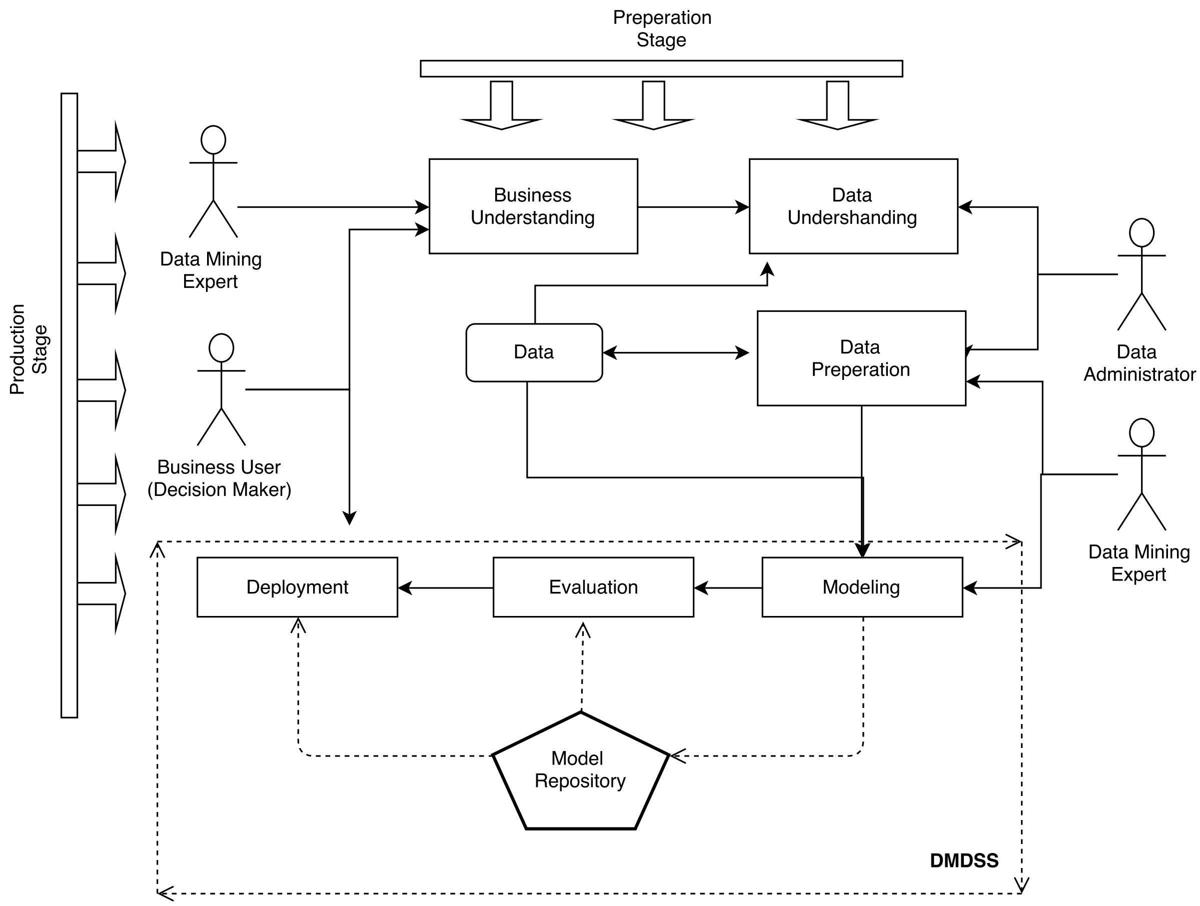


Figure 3.8: Architecture of DMDSS Process in Enterprise Framework

In the modeling and evaluation phase, the model is created and evaluated for several times to allow fine tuning of Data Mining algorithms through finding proper values of the algorithm parameters. It is essential to do enough iterations in order to monitor the level of changes in data sets and Data Mining models acquired and reach the stability of the data preparation phase and parameter values for Data Mining algorithms. The mission of the preparation stage is to confirm the fulfilling of the objectives of the area of analysis for decision support and to assure the stability of data preparation (Bâra and Lungu, 2012).

Production step represents the production use of DMDSS for the particular area of analysis. In that stage the focus is on the phases of modeling, evaluation and deployment, which does not mean that other phases are not encompassed in the production stage. Data preparation, for an example, is executed automatically based on procedures developed in the initial stage. Advanced Modeling and evaluation or validation are performed by a Data Mining expert or data scientist while the deployment phase is performed by a business analyst or system engineers.

4 Application

4.1 Problem Description

A company ABC Fashion is a Australian Clothing retailer company. The company has only been running operations in New South Wales (NSW). However, Board of Directors is considering an expansion of the company in to one of the three following states:

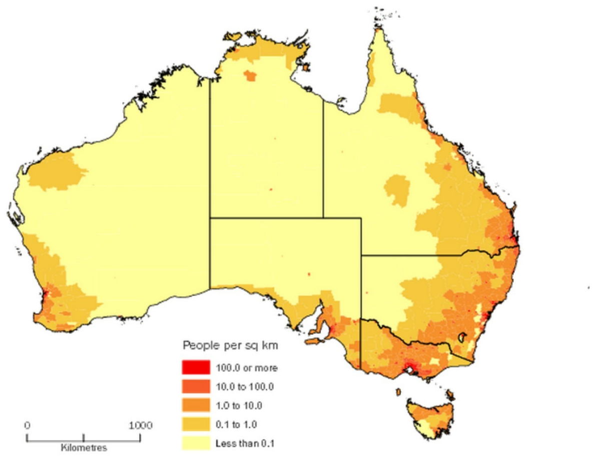
- Queensland (QLD)
- Victoria (VIC)
- Western Australia (WA)

Based on the provide historical data our main goal is to create a data-driven DSS which will provide more insight of the current situation of clothing retail situation and trend in these regions so that decision makers can take much better decision.

4.1.1 Study Area

In geographically Australia (officially known as the Commonwealth of Australia) is a federation of six states, together with ten federal territories. mainland of the country consists of five of the six federated states and three of the federal territories (the "internal" territories). Apart form other state the state of Tasmania is an island about 200 kilometers from the mainland. The remaining seven territories are classified for some purposes as "external" territories. Our main focus is on analysis the Clothing Retail Industry only on Queensland (QLD), Victoria (VIC), Western Australia (WA). From the Figure 4.1 we can imagine the reason of selecting this states as they have the higher population density that the others. Which is very important to know for such kind of industrial expansion.

(a) States and territories of Australia



(b) Population Density Map of Australia by States

Figure 4.1: Map decomposition of Australia by Region and Population

4.1.2 Business Understanding

From the problem Statement section we has already been informed that, the goal of this project is to build a decision support application to analyze Australia Clothing Retail Industry for Company expansion. But the analysis is controlled on three states. So we need some economic indicator from which we can create some statistical summary report to see how other companies competing in those three regions and which is the more secure region for expansion. Population has also great influence on Clothing Retail Industry. So it is important to take account population statistics in the study. Finally as the company will be ruing for longer period after expansion it is necessary to check the historical trend of Clothing Retail Industry for those three region.

4.1.3 Available Data

The primary data contains net profit margin in percentage for different competitors in different regions in the year of 2016. It shows in the year of 2016 how much what is the amount made by several companies in four different states. File contains information on 30 leading companies with respect to states. The following table shows a small glimpse of the data. The data file is available as csv format.

Table 4.1: Net Profit Margin Data of Competitors by State

	Company	QLD	WA	VIC	NSW
1	Competitor 1	4.57%	4.18%	7.90%	8.54%
2	Competitor 2	2.60%	1.88%	8.99%	9.23%
3	Competitor 3	2.26%	7.48%	6.71%	7.84%
4	Competitor 4	6.47%	6.70%	7.94%	9.89%
5	Competitor 5	6.82%	1.17%	8.54%	5.67%
6	Competitor 6	4.68%	6.44%	6.52%	3.03%
7	...	...	...	...	...

From the requirement of the project the provided data was not enough for in-depth analysis. So we decided to use some reliable open data source for more analysis. The most reliable source has found is Australian Bureau of Statistics where Historical demographic data and historical data on Clothing Retailing industries is collected and used in our whole study. d Including the following data we use more two data source through our study. The Retail Turnover data contain monthly turnover value by states for several industries. The unit is measurement is in million dollar. As we are only interested clothing retail industries we can clean the information about other industries for further analysis to get rid of redundant field in the database. Table 4.2 shows the structure of retail truonver data set.

Australian demographic data contains quarterly population of Australia from June 1982 to March 2015. The total population also is divided in to states and also gender.The unit of measurement is in persons. Though we are only interested on only four state

Table 4.2: Retail Turnover Data by Industries and States

	X	NSW.Supermarket	NSW.Liquor	NSW.specialised.food	...
1	Apr-1982	303.10	41.70	63.90	...
2	May-1982	297.80	43.10	64.00	...
3	Jun-1982	298.00	40.30	62.70	...
4	Jul-1982	307.90	40.90	65.60	...
5	Aug-1982	299.20	42.10	62.60	...
6	Sep-1982	305.40	42.00	64.40	...
7	...	...	...	...	...

stated above but for further its better to store the all information. In each column consist estimated population of a particular state for a particular gender. But there also exist a column which refers estimated population in total which is our main interest at the initial level. Table 4.3 shows the structure of the retail turnover data set.

Table 4.3: Quarterly estimated population of Australia by Gender and Total

	X	Male.NSW	Male.VIC	Male.QLD	...
1	Jun-1981	2608351	1958717	1178447	...
2	Sep-1981	2616060	1964139	1189946	...
3	Dec-1981	2624579	1969349	1200504	...
4	Mar-1982	2634534	1975617	1210128	...
5	Jun-1982	2643527	1981619	1219369	...
6	Sep-1982	2649615	1986589	1228791	...
7	...	...	...	...	...

4.1.4 Data Understanding

The Net Profit Margin data contains total 5 variables. The company variables is a object Id to refers the different companies who already has branch on those regions. The Other four variables contains net profit margin in percentage for four different regions.

The Demographic data file contains historical population of different states of Australia and grouped by total, male, female. Though our main interest on the population of only those four region but it is good to store the information on other states also for further use.

The Retail Turnover Data set contains the Turnover by state and by industries in million dollar. It contains data form April 1982 to August 2015. As we only need information on Clothing Retail Industries we only keep the data for retail industries and clean the other variables which is the main focus of our next section.

Turnover is an accounting term that calculates how quickly a business collects cash from accounts receivable or how fast the company sells its inventory. In the investment industry, turnover represents the percentage of a portfolio that is sold in a particular month or year. A quick turnover rate generates more commissions for trades placed by a broker. Two of the largest assets owned by a business are accounts receivable and inventory. Both of these accounts require a large cash investment, and it is important to measure how quickly a business collects cash. Turnover ratios calculate how quickly a business collects cash from its accounts receivable and inventory investments.

4.2 Construction of DSS Application

On the basis of separating the literal description and the data description finally we are able to build our Decision Support System application form the scratch. This section is dedicated to practical approach of data preparation and ETL, descriptive analysis of data, building predictive model with linear regression and time series modeling and Deployment of DSS application to the server.

4.2.1 Data Preparation and ETL

Data cleansing or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

In computing, Extract, Transform, Load (ETL) refers to a process in database usage and especially in data warehousing. The ETL process became a popular concept in the 1970s.[1] Data extraction is where data is extracted from homogeneous or heterogeneous data sources; data transformation where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis; data loading where the data is loaded into the final target database, more specifically, an operational data store, data mart, or data warehouse. So we can the first two processes in ETL export and transform is included in data preparation task.

All the data sources we have explained in the previous section contains some extra or redundant field/information which is removed when preperating data for ETL. For our data preparation a new tool is used called Alteryx Designer. With Alteryx, analysts can easily perform the data preparation work needed for analytics, independently accessing, profiling, cleansing and joining data from multiple sources using drag-and-drop tools and without coding. For many analysts in business groups such as marketing, sales, finance, or customer insight, the process to prep, blend and analyze data is slow and painful. It requires different tools and people to gather, cleanse and join data from different sources, more tools to build and publish analytic models, and even more to get it into the hands of business decision makers. Alteryx Designer solves this by delivering a repeatable workflow for self-service data analytics that leads to deeper insights in hours, not the weeks typical of traditional approaches! Alteryx Designer empowers data analysts by combining data preparation, data blending, and analytics predictive, statistical and spatial using the same intuitive user interface. In addition, Alteryx includes third party packaged data to enrich existing data and improve data quality. It is able to connect with several different type of data source such as csv, text, xlsx, sql database, oracle teradata etc. It also offers scheduled, and incremental data by the help of Alteryx server. Figure 4.2 shows the user interface of the software.

The Net Profit Margin data and Demographic Statistics data has already has structured information as a tabular way. So the main preference is connect the data file through Alteryx and load it in to the main data repository which is Microsoft SQL server in our case. But for other two data sources data cleaning is needed as there are lot of extra information is available what we do not need. To load the data in to repository, data must be transformed as structured way which is a prerequisite for structural database. The retail turnover data contains historical turnover of all retail market such as hardware, household goods, clothing retailing footwear and so on followed by state. As we only need information on clothing retailing market so other attribute is cleaned from the data and

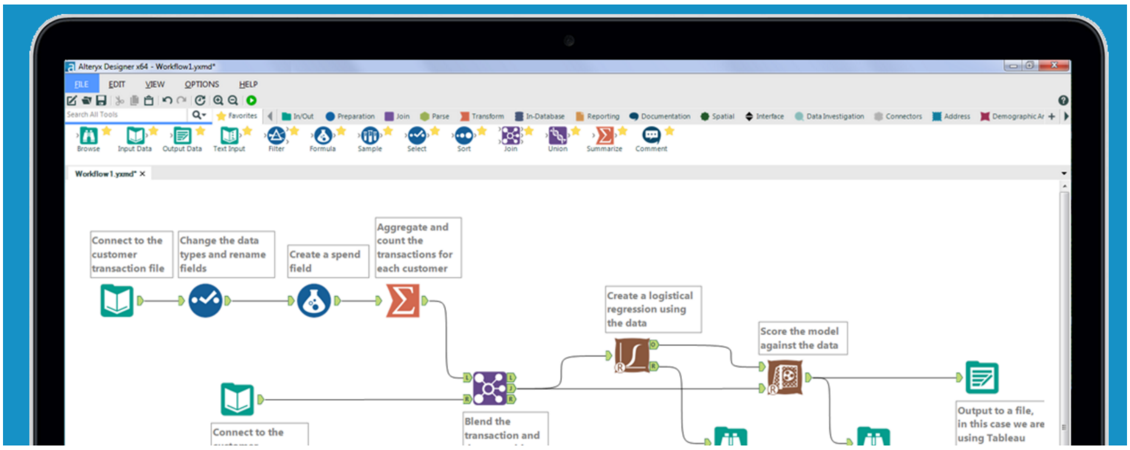


Figure 4.2: Alteryx Designer User Interface

and only clothing retailing data is selected for all states.

Through our process in Alteryx field header name fixed to make it more meaning full in the database. the all data are already structured form except except Net Profit Margin data. In Net Profit data the in the field the value contain % which is cleaned to read the data as numeric format. Then all the data are stored in three different table in data repository for further use.

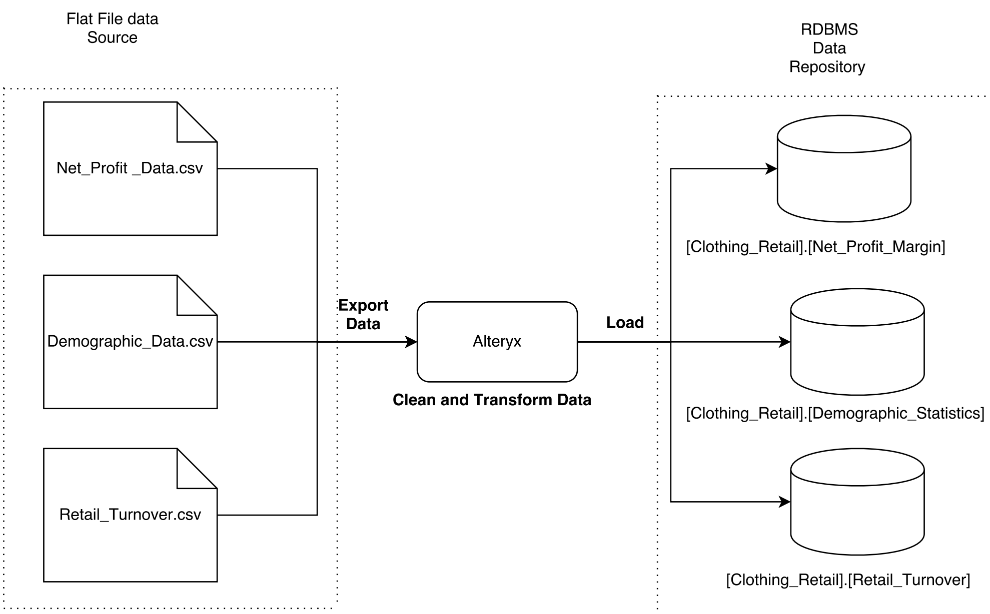


Figure 4.3: Data Preperation and ETL Process Model

Figure 4.4 show the work-flow for data preparation, blending and ELT from Alteryx designer. Each process in Alteryx is call a work-flow which can be created trough user interface and saved for further use. It it possible to read several file in one process and create an ETL in Alteryx. We read three different data sources in one work-flow, process the data and store the data in the database for further analysis.

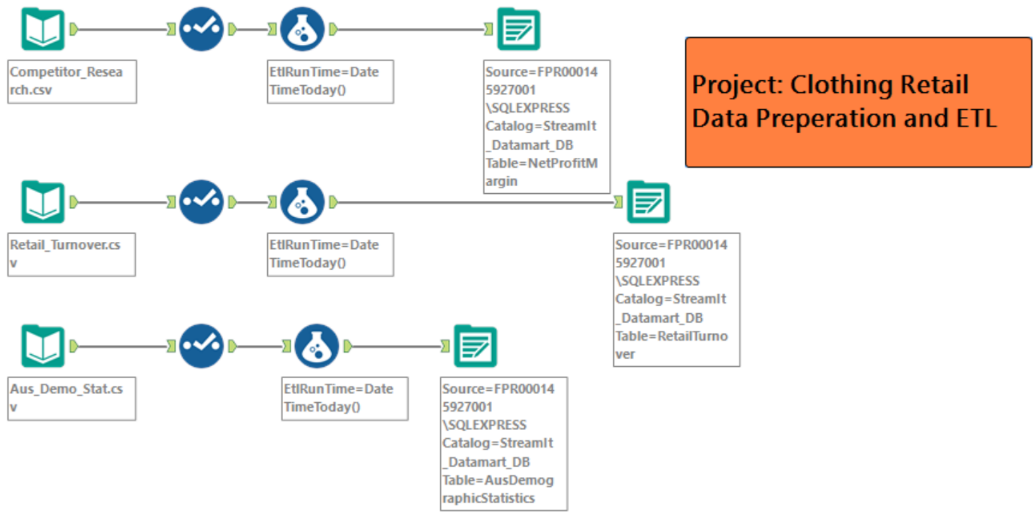


Figure 4.4: Alteryx Designer work-flow for Data Preparation and ETL

We are also able to see the result form the running work-flow when it will finish running. The result pane consist for different tabs error, warning, messages. From the Figure 4.5 we can see, the how many record is written to data base after ETL. Alteryx consider every row as a record.

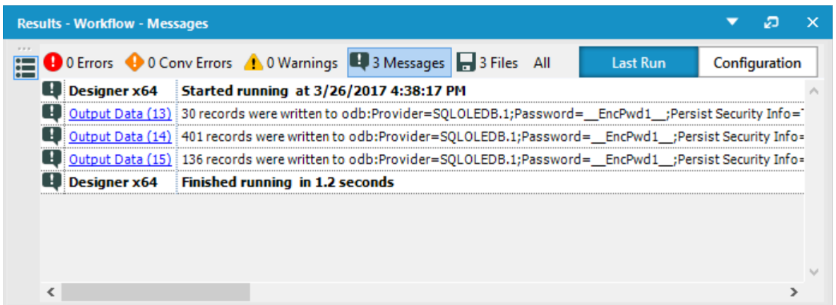


Figure 4.5: Resul view of Alteryx Work-flow

4.2.2 Descriptive Analysis

In data analysis Descriptive Analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods, Which give us information about a statistical model can be used or not through the further analysis, Which is also sometime refers as Exploratory Data Analysis. But primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians and business analyst to explore the data to see if there any usual pattern in the data.

Most Exploratory Data Analysis (EDA) techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of Exploratory Data Analysis (EDA) is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

- Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.
- Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
- Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

Based on our data at a first glance to compare the performance of different competitors on those region histogram plot, density plot and finally box plot. A histogram is a plot that show, the underlying frequency distribution (shape) of a set of continuous data can explain the variation, range of the data and how the frequency is distributed over range. Density plot is a simple estimation of histogram which is helpful to identify the skewness of the data and if it follows some particular probability distribution. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers. It is also very useful for comparison of one distribution to another in terms of mean value, variation and skewness. In our study box plot will be very use full to compare the variation of Turnover between QLD, WA, VIC, NSW.

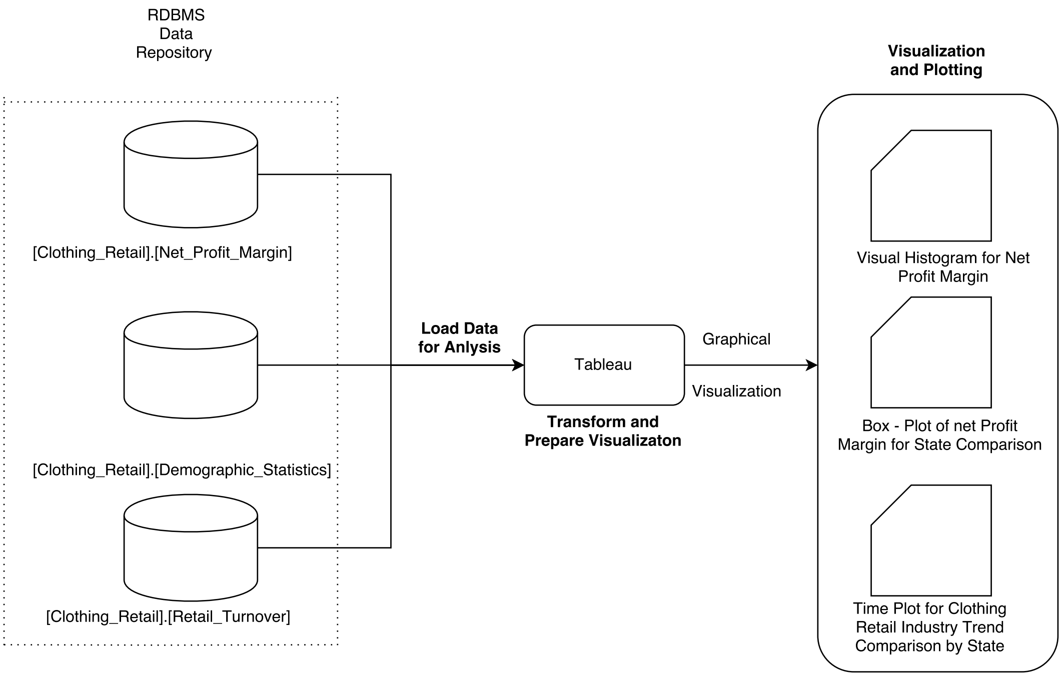


Figure 4.6: Process Model for Data Load and Descriptive analysis in Tableau

The enterprise edition of Tableau Desktop is use to create all these visualization. As we mentioned its just a drag and drop query builder which is based on Visual Query Language. So it is very to build visualization within a very short time through Tableau. We can see the process model for descriptive analysis in Figure 4.6.

As our other data sets are historical data over time so it is necessary to have time series analysis. plot of Turnover over time is the best possible way to start the analysis to see the behaviour such as time trend or seasonality exist in the data or not which is preliminary exploratory analysis for uni-variate time series.

4.2.3 Predictive Modeling

Predictive model should include features which capture all the important qualitative properties of the data: patterns of variation in level and trend, effects of inflation and seasonality, correlations among variables. In our study the predictive modeling is dedicated to time series modeling and forecasting for yearly cost per capita. There is a lot of approach could be used to model the data. Our main interest is to build a general moving average models, test the models and finally use the best models for forecast.

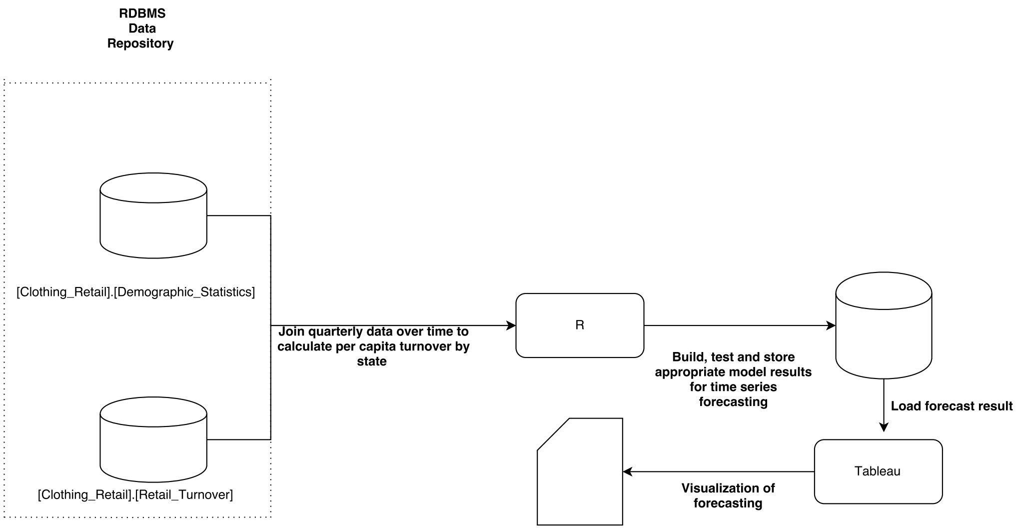


Figure 4.7: Process Model for Data Mining and Predictive analytics in Tableau and R

Exponential Smoothing Model for Time Series

Mathematically a time series is defined by the values of y_i and variable Y(temperature, closing price of a share, per capita turnover, etc) at time $t_1, t_2, t_3...$ Thus Y is a function of T, symbolized by

$$y = f(t) \tag{4.1}$$

According to Shumway and Stoffer (2010), In exponential smoothing procedures the weights assigned to observations are t-potentially decreased, as the observations get older. There are a variety of exponential smoothing methods and they all have a common property that recent values are given-relatively more weight in forecasting than the older observations. Holt (1957) extended single exponential smoothing to linear exponential smoothing to allow forecasting of data with trends. There are two variations to this method that differ in the nature of the seasonal component. The additive method is preferred when the seasonal variations are roughly constant through the series, while the multiplicative method is preferred when the seasonal variations are changing proportional to the level of the series. With the additive method, the seasonal component is expressed in absolute terms in the scale of the observed series, and in the level equation the series is seasonally adjusted by subtracting the seasonal component. Within each year the seasonal component will add up to approximately zero. With the multiplicative method, the seasonal component is expressed in relative terms (percentages) and the series is seasonally adjusted by dividing through by the seasonal component. Through our study additive model has been used to fit the model. The forecast for Holt's linear exponential smoothing is found using three smoothing constants, α and β and γ with values between 0 and 1, and three equations:

$$Level : L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \tag{4.2}$$

$$Trend : b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \tag{4.3}$$

$$Seasonal : S_t = \gamma(Y_t - L_t) + (1 - \gamma)S_{t-s} \tag{4.4}$$

$$Forecast : F_{t+m} = L_t + b_t m + S_{t-s+m} \tag{4.5}$$

Where, L_t is the estimate of the level of the series at time t. b_t is the estimate of the slope of the series at time t. Y_t is the observation at time t.α and β and γ are constant between 0 and 1. F_{t+m} is forecasted value at period $t+m$ and m is the number of periods ahead to be forecast.

Here L_t denotes an estimate of the level of the series at time t and b_t denotes an estimate of the slope of the series at time t. Equation 4.10 adjusts L_t directly for the trend of the previous period, b_{t-1}, by adding it to the last smoothed value, L_{t-1}. This helps to eliminate the lag and brings 14 to the approximate level of the current data value. Equation 4.11 then updates the trend, which is expressed as the difference between the last two smoothed values. This is appropriate because if there is a trend in the data, new values should be higher or lower than the previous ones. Since there may be some randomness remaining, the trend is modified by smoothing with P the trend is the last period $(L_t - L_{t-1})$, and adding that to previous estimate of the trend multiplied by (1-β). Finally, equation 4.12 is used to forecast ahead. The trend, b_t, is multiplied by the number of periods ahead to be forecast, m, and added to the base value, L_t (Shumway and Stoffer, 2010).

4.2.4 Evaluation

The evaluation of a DMDSS process is more about the evaluation of the methods has been using in Data Mining and evaluation of the overall system that it full fill the requirement of the project. The statistic we are using for descriptive statistics could be evaluated by looking over if all different statistics lead us to the same conclusion. For predictive modeling stage the evaluation depends on the methodology of the model. But in general it has been done by statistical significant test or hypothesis testing by validating the assumption of statistical model. Though the teasing tolls depends on data and model structure. The approach is totally different for cross-sectional data, time series data or multivariate data. Because the predictive models are different for different data structure. In case of Decision Support System evaluation the aim to access weather the user need are properly med, the system is suitable for task and user perform better with the system.

As we are using mostly Histogram, Distribution Plot and Box plot for descriptive statistics so the result and interpretation form those all tools lead us to same conclusion. For time series analysis there is a lot of statistical technique to evaluate the model. The ACF plot of error term to see the trend exist or not, Mean square error to see the amount of error the model have made. The structure we have followed through the building our system is agile. So user was always in touch with us to see if they are comfortable with the visuals and user interface. So it was a real time evaluation for project.

4.2.5 Deployment

As we mentioned earlier by the advancement of web technology and cloud computing DSS is not been build as piece of software installed in the local machine anymore. Following the new technology trend we have build our system with the integration of SQL Server, Alteryx, R and finally tableau to process the graphics and user interface. The result of model parameter from predictive modeling has been stored as table in SQL server which is the main data repository. Later it has been connected directly from tableau to build the DSS application locally.

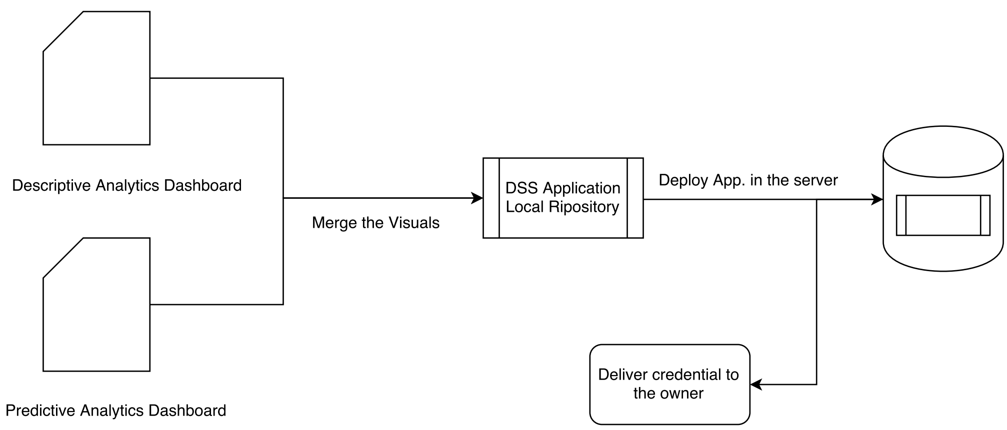


Figure 4.8: Process Model for Deployment of Time series

After evaluation the full DSS application is deployed on tableau server with live connection with the data. There are two method for live connection in tableau server. One is real time connection another is through tableau data extract. The faster possible way is use of tableau data extract connection. I this method table take the required data and store it as extract in the server so that it do not need to query every time to database which can make the whole process slower. The extract could be update daily or weekly basis or manually to extract the new data form database.

5 Results

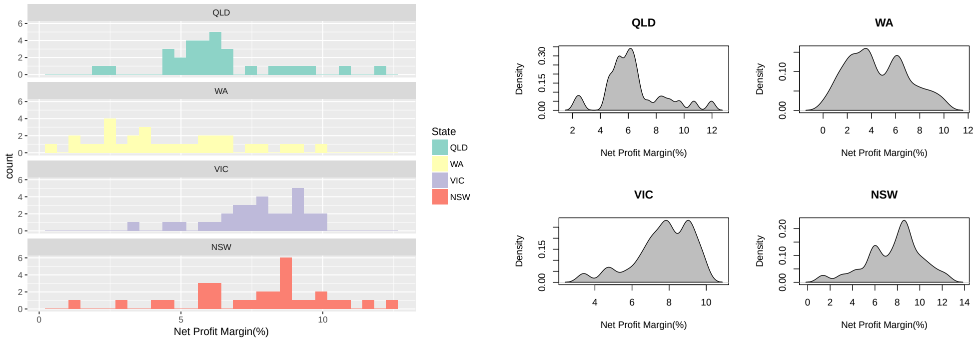
The market research of retail clothing industry is divided in to two main section. In descriptive analysis some basic statistical method is used to compare the performance of companies based on net profit margin of the companies who are already executing operation in four different region. The trend of of retail clotting in four different states is also studied by quarterly turnover and per-capita turnover to see the general trend and trend incorporating the population. In predictive modeling and forecasting section general exponential smoothing model has been introduced for further forecasting of per-capita turnover to see if possible turnover can be made by the proposed expanded branch in the future.

5.1 Descriptive Analysis

Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures or we may measure a large number of people on any measure. Descriptive statistics help us to simplify large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. For our marked research some basic analysis using histogram, density plot and box-plot is introduced to show the significant difference between the companies operating in four difference state. Also general time plot is used with basic linear regression model to see the trend of turnover in four different states.

5.1.1 Competitor Analysis

Net profit margin is the ratio of net profits to revenues for a company or business segment . Typically expressed as a percentage, net profit margins show how much of each dollar collected by a company as revenue translates into profit. The equation to calculate net



(a) Histogram Plot (b) Density Plot

Figure 5.1: R output of distribution and density plot of Net Profit Margin by State

profit margin is:

$$net\ profit\ margin = net\ profit/revenue \times 100 \tag{5.1}$$

The basic tools for descriptive analysis of a cross sectional continuous data distribution plot, density plot and box-plot to see the pattern, variation and how the the data distributed over sample group. Figure 5.1a and Figure 5.1b shows the distribution of net profit margin of different companies who are already operating from these Victoria, Queensland, New South wales and Western Australia. From statistical point of view we can see, Variation of net profit margin is much higher in Queensland and Western Australia than Victoria refers that, clothing retail market is more consistent in Victoria then Western Australia and Queensland in terms of net profit margin. The distribution is right skewed for Queensland and Western Australia and left skewed for Victoria refers that chance of having profit . Also form density plot we can see that possibility of having better profit margin is higher in Victoria than Western Australia and Queensland.

For more detail view we can look over the box-plot generated form Tableau in Figure 5.2. The median net profit margin for New South Wales, Queensland, Victoria and Western Australia are 9%, 6%, 8% and 4% respectively. So companies are performing better in Victoria in terms of median profit margin and in Western Australia has companies has the least performance that the others. The outliers of net profit in Queensland and Victoria refers few companies are under and over perming in those two states. The large box for Western Australia refers that market is more uncertain there than Queensland and Victoria.

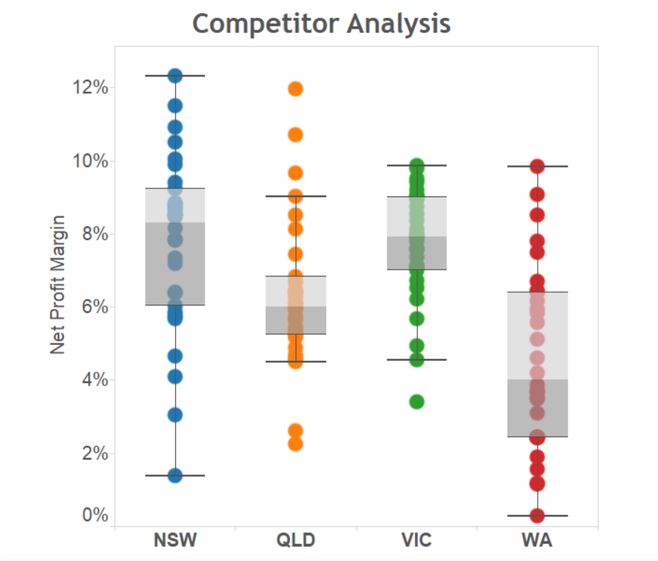


Figure 5.2: Tableau output of Box-Plot of Net profit Margin by States for Compaines

Table 5.1 shows, R output on basic summary statistics. Where we can see that the average net profit margin margin is higher for Victoria Western Australia and Queensland. Also minimum net profit margin is much higher for Victoria that any other region. So from the competitor analysis we can assume that, Victoria could be the best market for expansion in current situation.

Table 5.1: R output of Summary of Net profit margin by State

	Company	QLD	WA	VIC	NSW
1	Length:30	Min. : 2.260	Min. :0.290	Min. :3.400	Min. : 1.370
2	Class :character	1st Qu.: 5.237	1st Qu.:2.433	1st Qu.:7.018	1st Qu.: 6.110
3	Mode :character	Median : 6.000	Median :4.020	Median :7.920	Median : 8.305
4		Mean : 6.324	Mean :4.555	Mean :7.696	Mean : 7.728
5		3rd Qu.: 6.782	3rd Qu.:6.335	3rd Qu.:8.980	3rd Qu.: 9.135
6		Max. :11.960	Max. :9.850	Max. :9.870	Max. :12.310

5.1.2 Industry Analysis

Retail Turnover measures the turnover of local retail trade. The accounts receivable turnover formula tells how quickly we are collecting payments, as compared to your credit sales. If credit sales for the month total $300,000 and the account receivable balance is

$50,000, for example, the turnover rate is six. The goal is to maximize sales, minimize the receivable balance and generate a large turnover rate.

Figure 5.3 shows, exploratory time series plot of retail turnover. The plot shows, retail turnover value in million from 2000 to 2015 august for four different states. Ignoring the correlation between two time point a simple linear regression has been fit as trend line to see the trend of retail turn over for the states. The results shows, for clotting retailing all the states has upward turnover though there is little up and down from 2005 to 2008 for Queensland and Western Australia. But if we see the value of turn over in four regions they are significantly different than each other. For example for example for New South Wales turnover goes more the $600 million but in Western Australia it is even less than $150 million. One possibility is population should have a good good impact on the clothing retailing which we did take in to account so far.

From the preliminary analysis we can tell that, as all three states has upward trend company can be expanded to any of the states but Victoria has more consistent trend than the Queensland and Western Australia.

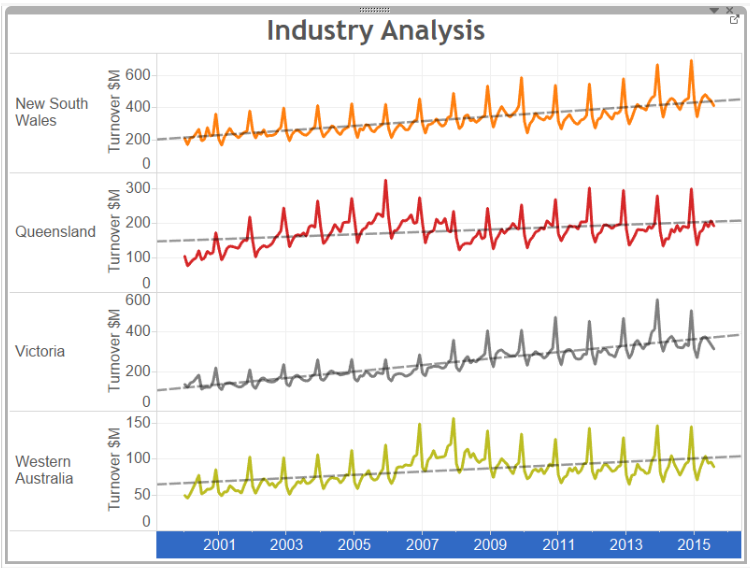


Figure 5.3: Time-Trend plot of Monthly Turonver in Million $ by States

The Tableau output for regression trend line is given in Table 5.2. Though the R-square value is vary low and month of the date parameter is insignificant for some states as P-value is greater 0.5. The reason of insignificance could be ignoring the auto-correlation between time point. But things to be notice here the parameter moth of the date is positive so for all the states the trend line has upward trend, which is also visible from the Figure 5.3.

Table 5.2: R Summary Output of Regression Trend Line Over time by states

State	Parameter	Value of Coefficient	P-value	R-Square
NSW	Month of Date	0.04	< 0.0001	0.52
	intercept	-1219.11	< 0.0001	
QLD	Month of Date	0.01	0.95	0.34
	intercept	-201	0.01	
VIC	Month of Date	0.05	< 0.0001	0.63
	intercept	-1522.7	< 0.0001	
WA	Month of Date	0.01	0.05	0.15
	intercept	-162.2	0.4	

5.1.3 Sales per capita Trend

In the previous section we saw that in all states the trend of turnover is is upward. But the range of turnover value is so different to compare between states. Which also refers that turnover has a great influence on the characteristics of the states. Rater that using the turnover directly so we decided to use a index sales per capita and use in our analysis as follows:

$$sales\ per\ capita\ monthly = monthly\ turnover/quarterlypopulation \tag{5.2}$$

Turnover data for states is available for every month but population statistics is available for every quarter of the year. So we decide to join the data first based on year and quarter then calculate the index. The Figure 5.4 shows, the exploratory time series plot with trend line for the sales per capita index. Which refers some different insights than Figure 5.3 we assumed.

From Figure 5.4 we can see, in terms of sales per capita not all region has upward trend. For all states the per capita sales ranges between \$0 to \$100 million which is very helpful for comparison. In terms of sales per capita The New South Wales and Victoria has upward trend and Queensland and Western Australia has much flatter trend in per capita sales. Performance of clothing retail market is much better in Victoria. Also

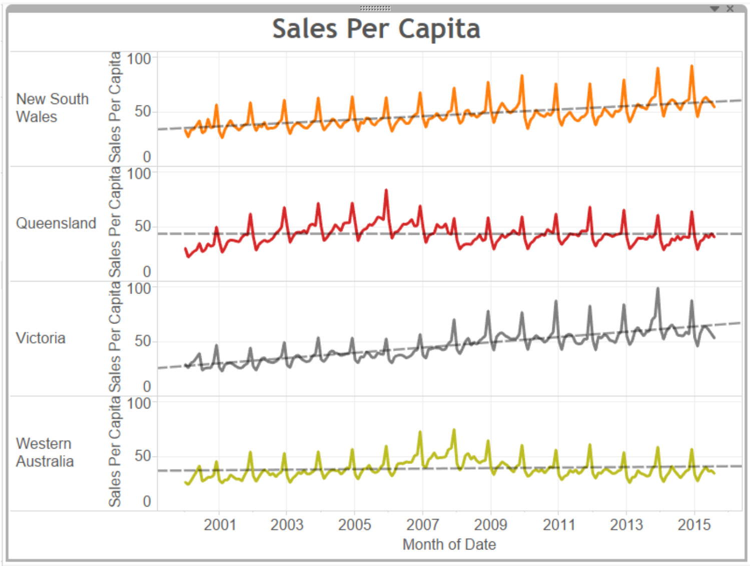


Figure 5.4: Time-Trend plot of Monthly Sales Per Capita States

market is consistent in Victoria that other two region as there is no random up and down in the curve which exist in case of Western Australia and Queensland.

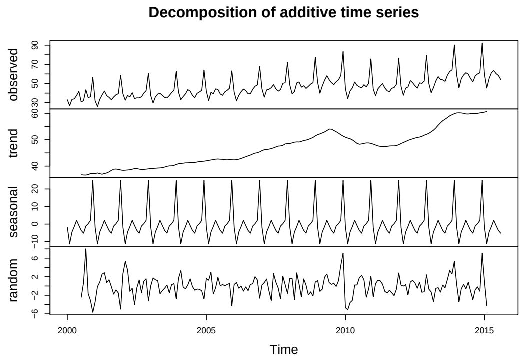
Table 5.3: Result of Sales per Capita Regression Trend Line

State	Parameter	Value of Coefficient	P-value	R-Square
NSW	Month of Date	0.0042	< 0.0001	0.39
	intercept	-119.554	< 0.0001	
QLD	Month of Date	- 0.00026	< 0.95	0.27
	intercept	44.1	0.001	
VIC	Month of Date	0.007	< 0.0001	0.63
	intercept	-213.7	< 0.0001	
WA	Month of Date	0.0007	0.05	0.15
	intercept	12.53	0.4	

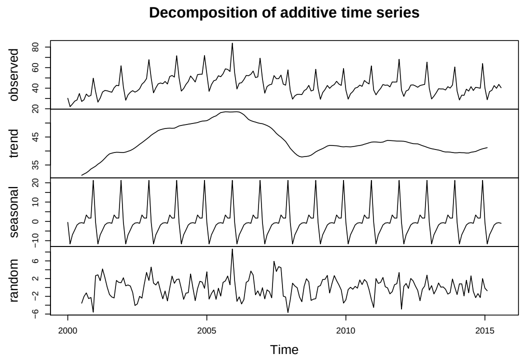
Form table 5.3 it is visible that, Victoria has the higher upper trend that any other states as the slope parameter month of the day is higher that any other states and positive. Also we can see that, slope for Queensland is negative which refers that Queensland has downward trend in terms of sales per capita though there is no statistical significance of it as the P-value is greater that 0.05. So from our analysis we can assume that, Victoria is the best state for company expansion.

5.2 Trend Analysis and Forecasting

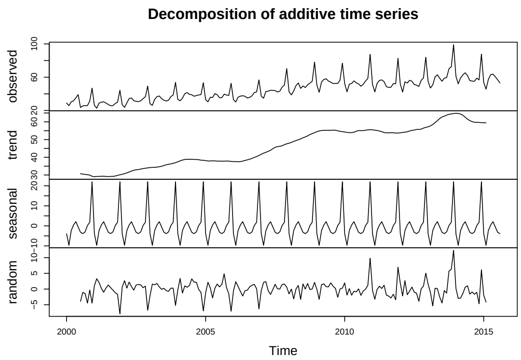
Time series analysis refers to a particular collection of specialised regression methods that use integrated moving averages and other smoothing techniques to illustrate trends in the data. It involves a complex process that incorporates information from past observations and past errors in those observations into the estimation of predicted values. Moving averages provide a useful way of presenting time series data, highlighting any long-term trends whilst smoothing out any short-term fluctuations. They are also commonly used to analyse trends in financial analysis. The calculation of moving averages is described in more detail here (Coghlan, 2015).



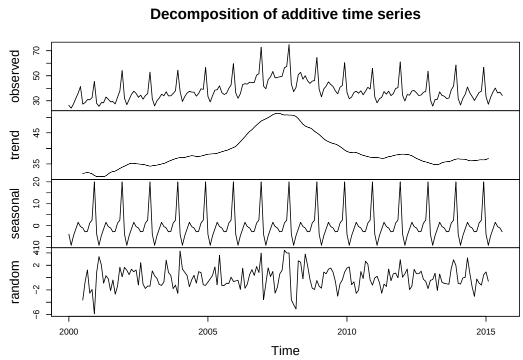
(a) NSW



(b) QLD



(c) VIC



(d) WA

Figure 5.5: R output of decomposition of Per Capita Sales for Four Different Region

Decomposition of Time Series

According to Coghlan (2015) If the time series can be described using an additive model with increasing or decreasing trend and seasonality, Holt Winter's exponential smoothing can be used to make short-term forecasts. But first its is necessary to decompose the several component of time series to see the the seasonality and trend exist in the time series. From the Figure 5.5 we can see, time trend and seasonality exist in all four time series for four different regions and random component refers constant variance and stationary over time period.

Fitting Holt Winter's Model for Seasonal Time Series

In order to select a Holt's linear method for forecasting, we take the first value of the original series as an estimate of the level of the series, i.e, $L_1 = Y_1$, and the difference between the first and second observation of the original series as an estimate of the slope of the series, i.e. $b_1 = Y_2 - Y_1$. Figure 5.6 Show the fitted and actual plot if time series for four different states. We consider different values for the smoothing parameters α, β and γ ranging from 0 to 1 and calculate the forecasted series for the set of values. Then we get the error series by subtracting the forecasted series from the original series. Now through computes programming, we obtain the set of values of a and 0 as smoothing parameters which gives the minimum Mean square error (MSE). Following this procedure we obtain the value of α, β and γ is given in Table 5.4.

Table 5.4: Estimated Value of the Parameter for Four Different Time Series by States

	α	β	γ	State
1	0.42	0.00	0.78	NSW
2	0.39	0.04	0.72	QLD
3	0.39	0.01	0.87	VIC
4	0.43	0.00	0.58	WA

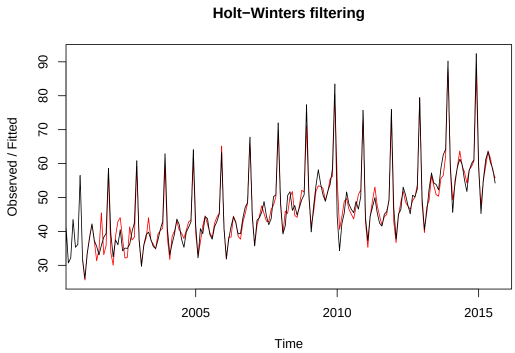
Following this procedure we obtain the value of $\alpha = 0.42$, $\beta = 0.00$ and $\gamma = 0.78$ for the time series of New South Wales which minimize the MSE. Thus our Holt-Winters linear model is:

$$\begin{aligned} L_t &= 0.42Y_t + (1 - 0.42)(L_{t-1} + b_{t-1}) \\ b_t &= 0.00(L_t - L_{t-1)} + (1 - 0.00)b_{t-1} \\ S_t &= 0.78(Y_t - L_t) + .(1 - 0.78)S_{t-s} \end{aligned} \tag{5.3}$$

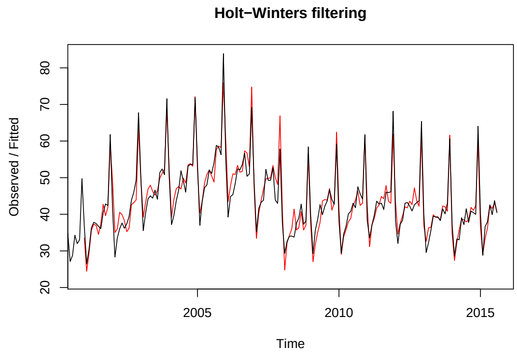
Using m=1 and s=12 we get,

$$F_{t+1} = L_t + b_t + S_{t-12+1} \tag{5.4}$$

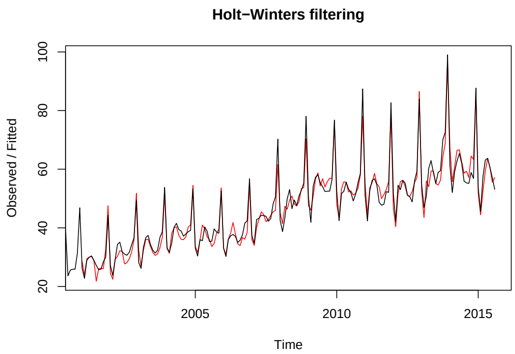
Same model can be constructed for Queensland, Victoria and Western Australia. These models are used to forecast the values.



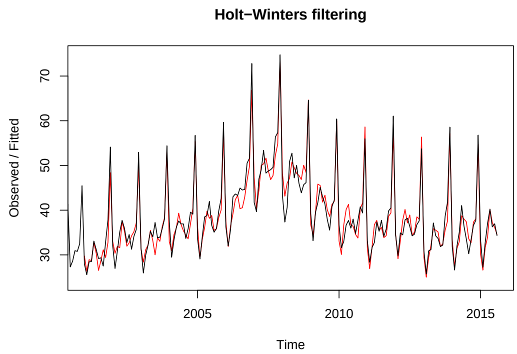
(a) NSW



(b) QLD



(c) VIC



(d) WA

Figure 5.6: Fitted and actual plot of Sales Per Capita by States

If we see the Figure 5.6 we notice the miss match of series at the initial level but as more time point start to contribute in the model the predictive models start to show better results. Table 5.4 show the parameter value of fitted Holt Winter's exponential smoothing model for four different time series for four different states.

For New South Wales The estimated values of α, β and γ are 0.42, 0.00 and 0.78, respectively. The value of α (0.42) is relatively low, indicating that the estimate of the level at the current time point is based upon both recent observations and some observations in the more distant past. The value of β is 0.00, indicating that the estimate of the slope b of the trend component is not updated over the time series, and instead is set equal to its initial value. This makes good intuitive sense, as the level changes quite a bit over the time series, but the slope b of the trend component remains roughly the same. In contrast, the value of γ (0.78) is high, indicating that the estimate of the seasonal component at the current time point is just based upon very recent observations.

For Queensland The estimated values of α, β and γ 0.39, 0.04 and 0.72, respectively. The same inference can be drawn from the value of α and γ what we have considered for New South Wales. But for *beta* the value is non zero refers that the estimate of the slope b of the trend component is updated over the time series but slowly.

For Victoria The estimated values of α, β and γ 0.39, 0.01, and 0.87, respectively. The same inference can be drawn from the value of α and γ what we have considered for New South Wales. But for *beta* the value is almost near to zero refers that the estimate of the slope b of the trend component is updated but slowly that Queensland.

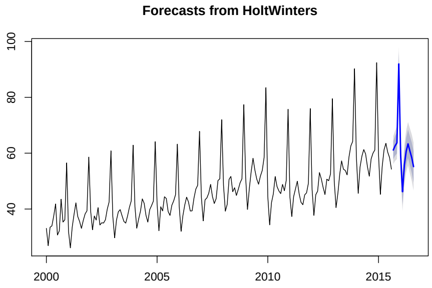
For Victoria The estimated values of α, β and γ 0.43, 0.00, and 0.58, respectively. The same inference is valid for the value of α and β what we have considered for New South Wales. But for γ the value is more lower than any other states. indicating that the estimate of the seasonal component at the current time point is based upon more lag values than compared to other states.

Forecasting

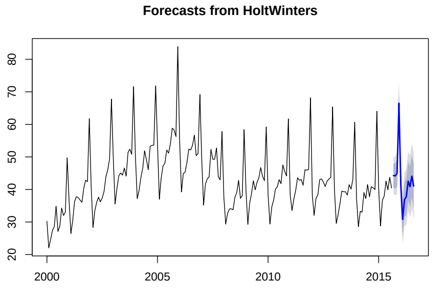
To make forecasts for future times not included in the original time series, we use the "forecast.HoltWinters" function in the forecast package of R. For example, the original data for the souvenir sales is from January 2000 to August 2015. If we wanted to make forecasts for next 12 month and plot the forecasts we will be able to construct the graph which is presented in Figure 5.7.

Table 5.5 at the end to the section represent the forecasted values of sales per capitra for next five month group by states. Lo.80 refers lower limit of estimated value of time series for 80% level of statistical significance and Hi.80 refers the higher limit on 80% level of significance. The same comment is valid for 95% interval estimate.

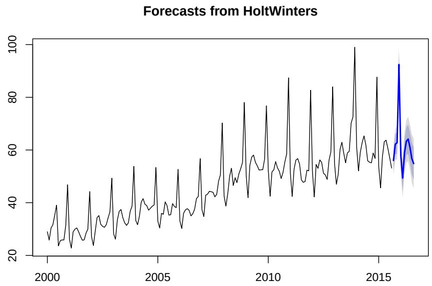
The visualization we have created form R is static and not interactive which means it not possible to see each each and every value when we are observing the plots. In that case Tableau has nice feature for interactive graping. So the result we have got form our analysis we put all in to Tableau desktop and visualize graph including foretasting.



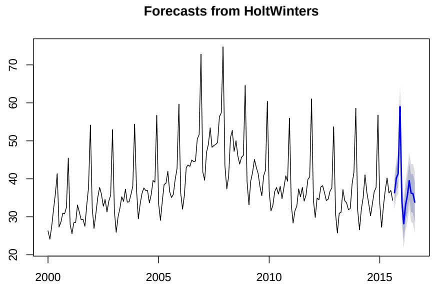
(a) NSW



(b) QLD



(c) VIC



(d) WA

Figure 5.7: R output of Sales Per Capita Forecasting by States

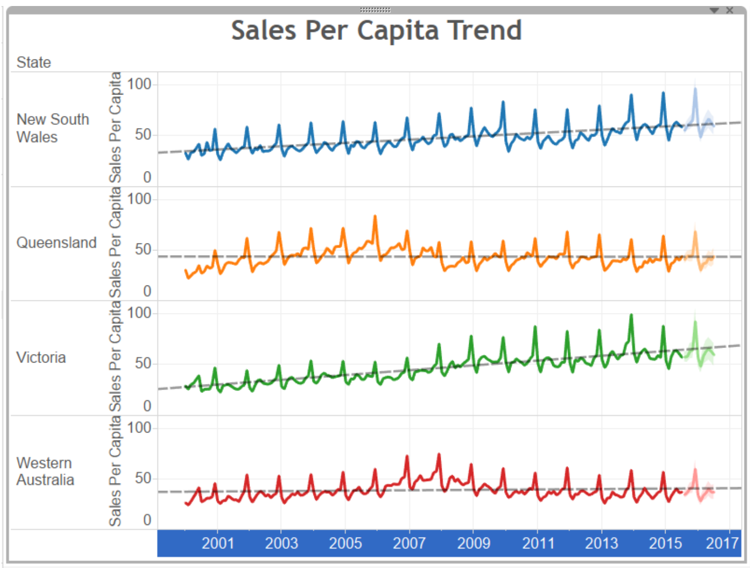
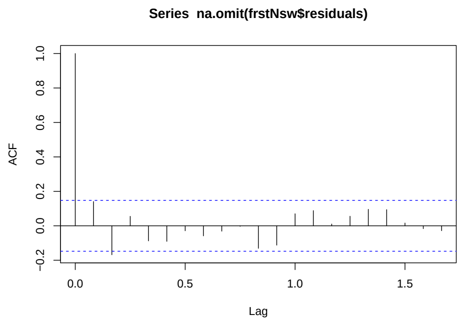


Figure 5.8: Interactive forecasting Plot of Sales Per Capita by States

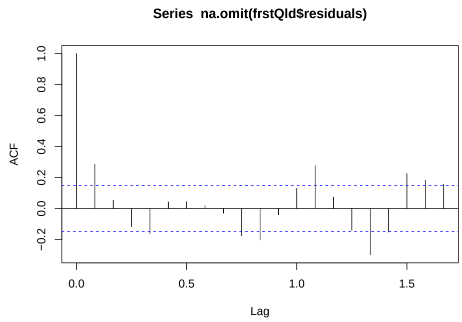
It is also possible to downland the forecasted result as report version for more in depth analysis. Figure 5.8 show the visualization of forecasting for four different time series in a single pan with interactive feature.

Model Diagnosis

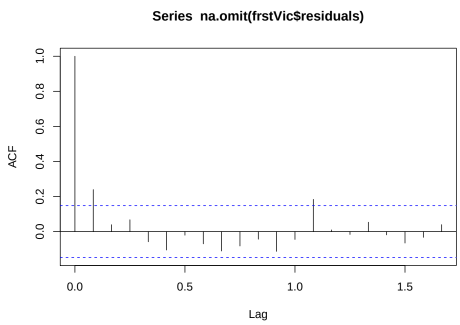
We can investigate whether the predictive model can be improved upon by checking whether the in-sample forecast errors show non-zero auto correlations. The correlogram shows in Figure 5.9 that the auto correlations for the in-sample forecast errors exceed the significance bounds for lag 2 and some other lags which show error terms are correlated at least for lag 1. But as the prediction is just extra feature and for time management we decide to keep the present model and improve the model in later version. Also from residual plot Figure 5.10 we can notice, the variance is constant for all the states but mean is not constant over time which refers error terms could be correlated over time.



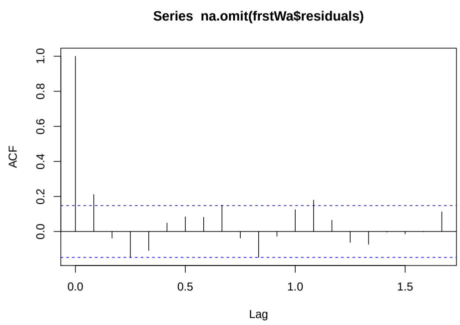
(a) NSW



(b) QLD



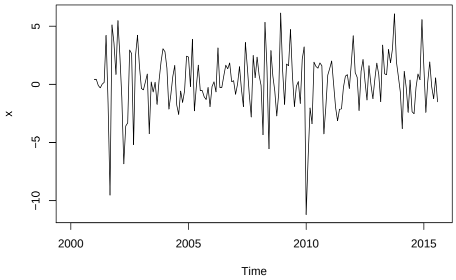
(c) VIC



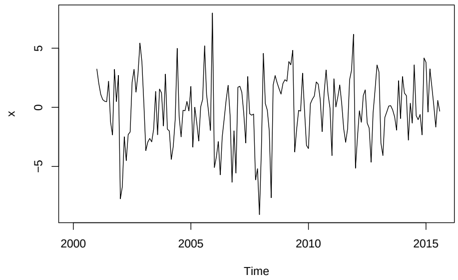
(d) WA

Figure 5.9: R output for ACF Plot of residuals from Fitted Model by State

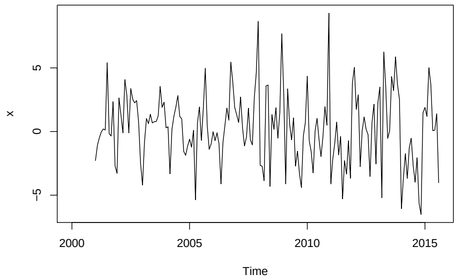
Also from residual plot Figure 5.10 we can notice, the variance is constant for all the states but mean is not constant over time which refers error terms could be correlated



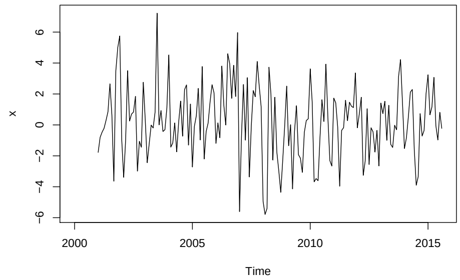
(a) NSW



(b) QLD



(c) VIC



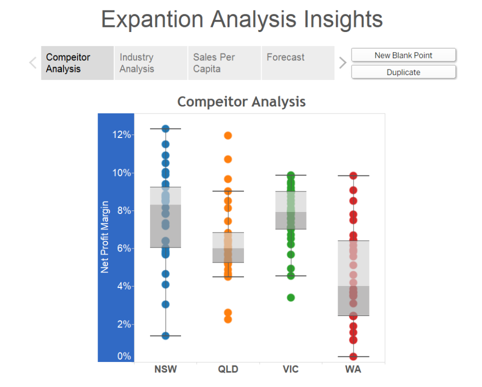
(d) WA

Figure 5.10: R output of residuals of Fitted Model by State

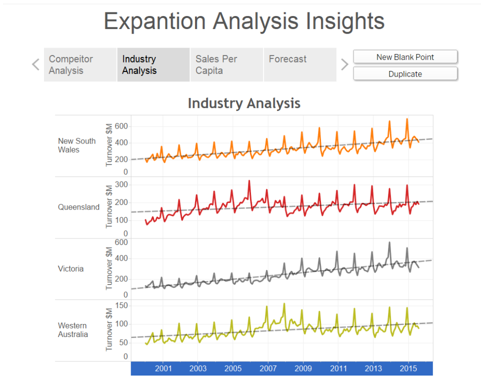
over time. So it is needed to transform the data to take some difference to make it stationary over time.

5.3 Combine the results and Deployment of Dashboard

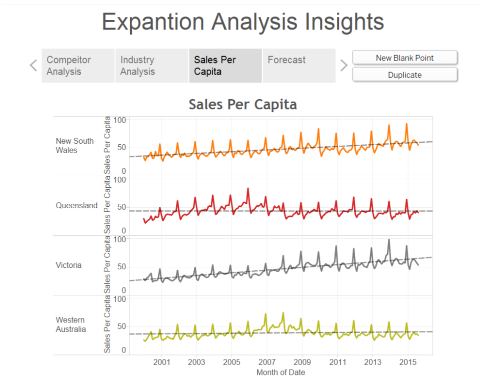
As we saw making graphics though R is not so much interactive. But it can be done through Shiny R application but then we also need a private R server to host the application. So we decide to use Tableau for all visualization. We select some informative form our analysis what we think informative for the management. In reality management are interest to know how we build the model why we did some descriptive analysis and why we test our model so remove all these graphics from the final presentation and we just keep the box-plot of net profit margin to compare the states by performance of the companies which already exit in those states, time trend for plot to net profit margin to see the historical trend of monthly turnover by states, time trend of per capita income



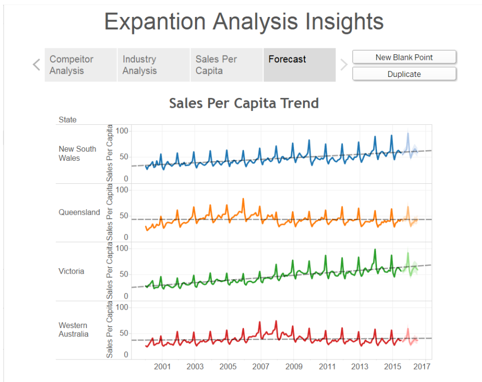
(a) Box-Plot of Net profit Margin in %



(b) Time Series Plot of Monthly Turn Over in Million $



(c) Time Series Plot of Sales Per Capita



(d) Time Series Forecasting of Sales per Capita

Figure 5.11: Deployed DSS Application in the Cloud Server

to see how clothing retailing industries is performing with respect to population of the states and finally the forecasting of per capita turn for next five month. All the visualization is first has been made in Tableau desktop and later on it uploaded to Tableau server with the existing data. As the data are confidential so we enable authentication for the application so only authorized person can see the application. We also schedule the refreshment of the data by weekly basis and for that we use Tableau data extract method.

Table 5.5: Forcasted Results of Next Five time Point by State at Different Level of Significant

Month Year	Point.Forecast	Lo.80	Hi.80	Lo.95	Hi.95	State
Sep 2015	61.11	57.85	64.38	56.12	66.11	NSW
Oct 2015	62.64	59.10	66.19	57.22	68.06	NSW
Nov 2015	63.70	59.90	67.51	57.89	69.52	NSW
Dec 2015	91.96	87.91	96.01	85.77	98.15	NSW
Jan 2016	59.10	54.82	63.37	52.55	65.64	NSW
Sep 2015	44.25	40.62	47.89	38.69	49.81	QLD
Oct 2015	44.18	40.25	48.10	38.18	50.17	QLD
Nov 2015	44.99	40.79	49.20	38.56	51.42	QLD
Dec 2015	66.48	61.99	70.97	59.61	73.35	QLD
Jan 2016	41.47	36.70	46.25	34.17	48.78	QLD
Sep 2015	56.00	52.40	59.61	50.49	61.52	VIC
Oct 2015	62.19	58.31	66.06	56.26	68.11	VIC
Nov 2015	62.77	58.64	66.89	56.45	69.08	VIC
Dec 2015	92.46	88.08	96.84	85.77	99.15	VIC
Jan 2016	57.76	53.14	62.37	50.70	64.82	VIC
Sep 2015	36.34	33.33	39.36	31.73	40.95	WA
Oct 2015	40.19	36.91	43.47	35.17	45.21	WA
Nov 2015	41.37	37.84	44.90	35.97	46.76	WA
Dec 2015	59.03	55.27	62.79	53.28	64.78	WA
Jan 2016	34.09	30.10	38.07	28.00	40.18	WA

5.4 Recommendation

From the deceptive analysis (Table 5.1) we can have noticed that, Victorai (VIC) has the higher average ($7.7% million) and ($7.92% million) median Net Provit Margin also box-plot and distribution plot refers that distribution for VIC left skewed refers that more companies are doing good. So if we randomly through any company in VIC it has higher chance of having good net profit mergint at the end. Also form predictive analysis we can see, VIC is in much better position than the other.

According to sales per capita time series analysis (Figure 5.4) we can see that only Victoria (VIC) and New South Wales (NSW) has the upward trend of per capita over time. Also the time series show more consistent over time than Queensland (QLD) and Western Australia. So after considering all insightful analysis we strongly recommend to choose **Victoria** for next company expansion. From from the Table 5.5 we can see 65.00 is the rough estimate of sales per capita at the first.

6 Conclusion

The use of Data Mining and decision support methods, including novel visualization methods, can lead to better performance in decision making, can improve the effectiveness of developed solutions and enables tackling of new types of problems that have not been addressed before.Though main focus of this paper is to apply all these technique for business environment but the approach can be applied for any research area such as heath care, banking, public and private financial Sectors and other research areas where system generates high volume of data and management wish to see more insightful information form the data within a very short time frame.

According to the objectives, our main purpose of this paper is to create a Model for Data Mining (DM) and Decision Support System (DSS) framework (DMDSS) which can be divided in to three phases: Create a process for Data Mining, create a process for Decision Support, Integrate the process for Data Mining and Decision Support to build a DMDSS process for decision making in a enterprise environment. Finally we apply our model in to real life problem to build a DMDSS application. Through the building of DMDSS application we experience that, all this three process are inter connected. Data Mining task executed first for analyse the data and validate the model then DSS process introduce to consume the results gained from the Data Mining and later on integrating both process interactive visualization has been generated which shows in Figure 5.11. After validation the final DSS application is uploaded in the private server with enabled authentication mode so that only authorised people can have assess the confidential information.

Form our whole study we notice that, Tableau is fully capable of building DSS application but of course some other tool such Alteryx and R is need for data preparation and predictive modeling such as R is needed in the process in which Tableau is not capable. But Tableau is very good in quick interactive visualization based on visual query language. In the other had R is good predictive modeling but lacking in interactive visualization. R is also capable of data preparation but time consuming with respect to time frame so data preperation tool like Alteryx can be plugged in the process to intense the time.

From our decision support application for analysis of clothing retailing market of Australia for a particular company expansion we can successfully choose the best states. Considering the net profit margin, monthly retail turnover and sales per capita Victorai is approprite sate for exapntion among Queensland, Victoria and Western Australia. We were able to add some additional feature of forecasting of time series by Holt Winter's exponential smoothing model. As we see in model diagnosis the residuals are correlated which refer possible transformation of data is needed and also more advance modeling like Auto-regressive and Moving Average with difference (ARIMA) can be used to solve that problem in further analysis.

Bibliography

Marc P Armstrong, Paul J Densham, and Gerard Rushton. Architecture for a microcomputer based spatial decision support system. In *Second International Symposium on Spatial Data Handling*, volume 120, page 131, 1986.

Adela Bâra and Ion Lungu. *Improving Decision Support Systems with Data Mining Techniques.* INTECH Open Access Publisher, 2012.

Tim Berners-Lee. Www: Past, present, and future. *Computer*, 29(10):69–77, 1996.

Frada Burstein and Clyde Holsapple. *Handbook on decision support systems 2: variations.* Springer Science & Business Media, 2008.

Avril Coghlan. A little book of r for time series, 2015.

Douglas C Engelbart. Augmenting human intellect: a conceptual framework (1962). *PACKER, Randall and JORDAN, Ken. Multimedia. From Wagner to Virtual Reality. New York: WW Norton & Company*, pages 64–90, 2001.

Jane Fedorowicz. *Document-based Decision Support Systems: A Technology Assessment.* PRIISM, Pacific Research Institute for Information Systems and Management, 1988.

Yongjian Fu. Data mining. *IEEE Potentials*, 16(4):18–20, 1997.

George Anthony Gorry and Michael S Scott Morton. A framework for management information systems. 1971.

Paul Gray, Nancy W Berry, JS Aronofsky, O Helmer, Gerald R Kane, and Thomas E Perkins. The smu decision room project. In *Transactions of the First International Conference on Decision Support Systems*, volume 1, pages 122–129, 1981.

George Houdeshel and Hugh J Watson. The management information and decision support (mids) system at lockheed-georgia. *MIS quarterly*, pages 127–140, 1987.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

Robert Joyner and Kenneth Tunstall. Computer augmented organizational problem solving. *Management Science*, 17(4):B–212, 1970.

Gregory E Kersten. Negogroup decision support system. *Information & Management*, 8 (5):237–246, 1985.

Nada Lavrač, Marko Bohanec, Aleksander Pur, Bojan Cestnik, Marko Debeljak, and Andrej Kobler. Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics*, 40(4):438–447, 2007.

Dunja Mladenic. *Data mining and decision support: Integration and collaboration.* Springer Science & Business Media, 2003.

Mihaela Muntean. Initiation in olap technology: theory and practice, 2004.

Daniel J Power. *Decision support systems: concepts and resources for managers.* Greenwood Publishing Group, 2002.

Daniel J Power. Decision support systems: a historical overview. In *Handbook on Decision Support Systems 1*, pages 121–140. Springer, 2008.

Rok Rupnik and Matjaž Kukar. *Data mining based decision support system to support association rules.* Pearson Education, 2007.

Jung P Shim, Merrill Warkentin, James F Courtney, Daniel J Power, Ramesh Sharda, and Christer Carlsson. Past, present, and future of decision support technology. *Decision support systems*, 33(2):111–126, 2002.

Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples.* Springer Science & Business Media, 2010.

Reid G Smith and Adam Farquhar. The road ahead for knowledge management: an ai perspective. *AI magazine*, 21(4):17, 2000.

Efraim Turban, Ramesh Sharda, and Dursun Delen. *Decision support and business intelligence systems*. Pearson Education India, 2011.

Hugh J Watson, R Kelly Rainer Jr, and Chang E Koh. Executive information systems: a framework for development and a survey of current practices. *MIS quarterly*, pages 13–30, 1991.