Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Bachelor Thesis

The importance of Big Data technologies for modern businesses

Victoria Polyakova

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Victoria Polyakova

Business Administration

Thesis title

The importance of Big Data technologies for modern businesses

Objectives of thesis

The aim of the bachelor thesis is to describe the use of Big Data technologies in business, and to compare the usage of machine learning methods in different industries. The bachelor thesis will also focus on factors affecting the ability of new data scientists to break into the field of Big Data analysis by analysing selected quaestionnaire results.

Methodology

Theoretical part of the bachelor thesis will be based on selected literature and other scientific sources. In the practical part, basic descriptive statistics, and methods for analysing categorical data will be used.

The proposed extent of the thesis

30-40 pages

Keywords

Big Data, Data Science, Machine Learning, questionnaire survey, technology

Recommended information sources

ABBOTT, D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.

FIELD, Andy P. *Discovering statistics using IBM SPSS statistics.* Thousand Oaks: SAGE Publications, 2013. ISBN 978-1-4462-4917-8.

SIEGEL, E. Predictive Analytics. Hoboken: John Wiley & Sons, 2013. ISBN 978-1-118-35685-2.

TUFFÉRY, S. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8

Expected date of thesis defence

2022/23 SS – FEM

The Bachelor Thesis Supervisor

Ing. Zuzana Pacáková, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 20. 06. 2022	Electronic approval: 27. 10. 2022
prof. Ing. Libuše Svatošová, CSc.	**doc. Ing. Tomáš Šubrt, Ph.D.**
Head of department	Dean

Prague on 23. 05. 2024

Declaration

I declare that I have worked on my bachelor thesis titled "The Importance of Big Data Technologies for modern business" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on date of submission ________15. 03. 2024___________________

Acknowledgement

I would like to thank Ing. Zuzana Pacáková, Ph. D. for her advice and support during my work on this thesis.

The importance of Big Data technologies for modern businesses

Abstract

This bachelor thesis explores the intersection of big data analytics and predictive modeling in the context of airline passenger satisfaction. The theoretical framework delves into the definition and significance of big data analytics, elucidating various types of analytics and the tools utilized in the field. It places predictive analytics techniques such as business intelligence, statistics, and data mining, highlighting their pivotal roles in uncovering insights from vast datasets. Additionally, the study examines the benefits and challenges associated with big data analytics, underlining its potential across diverse industry sectors. Practical application involves leveraging IBM SPSS Statistics to conduct logistic regression and clustering analysis on a dataset concerning airline passenger satisfaction.

This thesis utilizes logistic regression to identify key predictors of passenger satisfaction, while clustering analysis uncovers distinct customer groups. Through theoretical insights and practical application, the study enhances understanding of big data analytics' role in the aviation industry. It emphasizes the significance of advanced analytics tools in extracting actionable insights for strategic decision-making, ultimately fostering customer-centric approaches.

Keywords: Big Data, Machine Learning, IBM SPSS Statistics, technology, analytics, satisfaction, tools, regression, clustering, correlation.

Table of Contents

Introduction

In the contemporary landscape, the resounding declarations such as "We live in an era of big data" by influential figures like Humberty (2015), the World Economic Forum (2014), and Boyd & Crawford (2012) highlight the profound impact of big data technologies on our lives and societal structures. Beneath the surface of these discussions lies a nuanced complexity – the term "big data" often carries diverse interpretations and is intricately linked with broader technological advancements and the ongoing digitalization of businesses and society (Rüping, 2015).

In the contemporary landscape of the aviation industry, characterized by fierce competition, rapid technological advancements, and evolving consumer preferences, the quest to understand and enhance passenger satisfaction has become paramount. With the advent of big data analytics, airlines have been presented with a powerful toolset to decipher the intricate dynamics underlying passenger perceptions and preferences. This bachelor thesis embarks on a comprehensive exploration into the realm of airline passenger satisfaction, leveraging the principles of big data analytics to unravel the underlying factors shaping customer experiences.

Theoretical underpinnings form the cornerstone of this study, providing a robust framework for understanding the essence and significance of big data analytics in contemporary business contexts. Delving into the multifaceted landscape of analytics, the theoretical discourse delineates various types of analytics – from descriptive and diagnostic to predictive and prescriptive – and elucidates their respective roles in extracting actionable insights from vast and disparate datasets. By elucidating the intricacies of analytics methodologies and their applications, this theoretical foundation sets the stage for the empirical investigation into airline passenger satisfaction.

The empirical component of this study is anchored in the practical application of advanced analytics tools, with a particular focus on IBM SPSS Statistics. Leveraging a carefully curated dataset specifically tailored to explore passenger satisfaction within the airline industry, this research endeavors to employ a multifaceted analytical approach. Through the utilization of logistic regression and clustering analysis, this study seeks to uncover the latent patterns and underlying drivers influencing passenger perceptions and satisfaction levels.

By integrating theoretical constructs with empirical exploration, this thesis aspires to contribute to a deeper understanding of the pivotal role played by big data analytics in informing strategic decision-making processes within the aviation sector. Against the backdrop of an increasingly competitive industry landscape and heightened customer expectations, the relevance of this study becomes manifest. By dissecting the complexities of passenger satisfaction and delineating actionable insights, this research endeavors to empower airline companies with the knowledge and tools necessary to cultivate customer-centric approaches. Ultimately, the overarching goal is to enhance to overall passenger experience, foster customer loyalty, and drive organizational success in an increasingly data-driven and customer-centric aviation landscape.

1. Objectives and Methodology

1.1 Objectives

The significance of Big Data technologies in contemporary businesses is undeniably transformative, reshaping operational landscapes and decision-making processes. In acknowledgement of this paradigm shift, the objectives of this thesis encompass a multifaceted exploration, delving into the applications and implications of big data technologies in the business realm.

1. Description of Big Data Technology Implementation: The primary objective is to provide a comprehensive description of the utilization of Big Data technologies in diverse business contexts. This involves an in-depth examination of how modern businesses integrate and leverage Big Data tools, platforms, and analytics to derive meaningful insights, enhance operational efficiency, and gain a competitive edge in today's dynamic markets.
2. Comparison of Machine Learning methods Across Industries: Another key goal is to conduct a comparative analysis of the application of machine learning methods in different industries. By exploring and contrasting how various sector harness machine learning within the framework of Big Data, the thesis aims to identify industry-specific trends, challenges, and successful strategies, offering valuable insights for both practitioners and researchers.
3. Implementation of IBM SPSS Statistics on Selected Big Data Database: this objective involves the practical application of IBM SPSS Statistics, a robust tool in the realm of big data analytics, on a chosen database characterized by substantial volumes and complexities of data. The implementation will encompass the integration of IBM SPSS Statistics into the selected database infrastructure, exploring its functionalities and capabilities in handling and analyzing large-scale and diverse datasets.
4. Demonstration of Big Data Analytics with IBM SPSS Statistics: Building upon the implementation, the thesis seeks to demonstrate the efficacy of IBM SPSS in performing big data analytics within the selected database. This involves showcasing how IBM SPSS can unearth meaningful patterns, trends, and insights from the intricate web of data, emphasizing its role in transforming raw information into actionable intelligence. The demonstration aims to provide a practical understanding of the capabilities of IBM SPSS in addressing the unique challenges posed by big data.
   - To identify significant predictors of passenger satisfaction through binary logistic regression analysis, thereby elucidating demographic and flight-related factors influencing customer perceptions. To develop a predictive model using logistic regression to anticipate passenger satisfaction levels, facilitating targeted strategies for enhancing customer experience and fostering loyalty in the airline industry.
   - To uncover hidden patterns and groupings within the airline passenger dataset through TwoStep Cluster Analysis, enabling the identification of distinct passenger segments based on their satisfaction profiles.

By incorporating these objectives, the thesis aims to not only elucidate the theoretical aspects of Big Data but also provide a hands-on exploration of the practical implementation of IBM SPSS in a real-world big data setting. This holistic approach ensures a comprehensive understanding of the symbolic relationship between cutting-edge technologies, industry applications, and the practical tools that facilitate the extraction of value from vast and complex datasets.

1.2 Methodology

The methodology adopted for this bachelor thesis on big data integrates theoretical insights from selected literature and scientific sources with practical analytical approaches, combining both theoretical and empirical dimensions to provide a comprehensive understanding of the subject.

1. Literature Review and Theoretical Framework:
    - Source Selection: The theoretical foundation of the thesis will be built upon a thorough review of selected literature and relevant scientific sources. Emphasis will be placed on authoritative works, scholarly articles, and industry publications that contribute to the conceptual understanding of big data technologies, machine learning, and their applications in business.
    - Conceptual Framework: The literature review will establish a conceptual framework, elucidating key theories, principles, and perspectives related to big data. This conceptual foundation will serve as the theoretical backstone for subsequent analyses and discussions in both the theoretical and practical components of the thesis.
2. Practical Data Analysis:
    - Descriptive Statistics: Basic descriptive statistics will be employed to summarize and present key characteristics of the datasets. This may involve measures such as mean, median, mode, standard deviation, and other statistical indicators, providing a clear profile of the data under consideration.
    - Categorical Data Analysis: In addition, methods specifically tailored for analyzing categorical data will be applied. This could include techniques such as frequency distribution analysis, cross-tabulations, and chi-square tests to derive meaningful insights from categorical variables present in the datasets.
    - Binary Logistic Regression Analysis: This method was used to identify significant predictors of passenger satisfaction and develop a predictive model to anticipate satisfaction levels. Through this analysis, insights were gained into demographic and flight-related factors influencing customer perceptions.
    - TwoStep Cluster Analysis: This technique was employed to uncover hidden patterns and groupings within the dataset, facilitating targeted strategies for enhancing customer experience. By automatically determining the optimal number of clusters and using hierarchical and K-Means clustering algorithms, this analysis provided valuable insights for decision-making and targeted interventions.
    -

2. Literature Review

2.1 Big Data Definition

Understanding what is regarded as Big Data and its diverse meanings can be a complicated attempt caused by its gradually developing nature and extensive applications. The concept of big data has encountered notable transformations as time goes, and context and perspective can lead to the variance of its definitions.

In historical perspective, even that the idea of operating with huge and complex amounts of data has been going around for several decades, the term “Big Data” originated only in the beginning of 21st century and gained its popularity. In the era before Big Data, data analysis was concentrated on structured data and traditional databases. Data Warehouse existed as a traditional form factor for handling big data analytics efficiently for many decades. At addition to this integrated system, predictive analytics made a difference after introducing Big Data and its new tools, platforms, and methodologies to cope with large, various, and constantly changing data that modern organizations deal with. In the early years of 21st century the adoption of new tools and platforms brought more openness in various types of data to be handled. Through centuries, we went from “Knowledge is power” to “Knowledge is profit”, and big data breathed a new life into possibilities to gain a knowledge.

First thing to realize about big data is that the concept is to the lesser degree about being big. The volume is definitely crucial, despite that it is only one of the 3V’s. The original definition has been explained as far back as 2001 by Doug Laney, an analyst of Gartner (Chen et al., 2014a). It was written in his research report where he brought the 3Vs model: Volume, Velocity, and Variety. This definition was coined a long time ago, but it is still pretty valid. Some huge companies, such as Gartner, IBM, and Microsoft, had used that model for representation of big data during 10 years (Chen et al., 2014a). When we talk about big data, as the matter of fact, we talk about all these things.

The volume is a strong emphasis in McKinsey’s definition of big data: “Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” (Manyika et al., 2011) In this term data volume is a predominant characteristic, and it is common for some individuals to interpret big data in terabytes or even in petabytes.

According to Doug Laney, velocity means it is crucial to act quickly when collecting and analyzing data. To fully capitalize on business opportunities, the speed is important. Examples of such cases (Hamami & Dahlan, 2020):

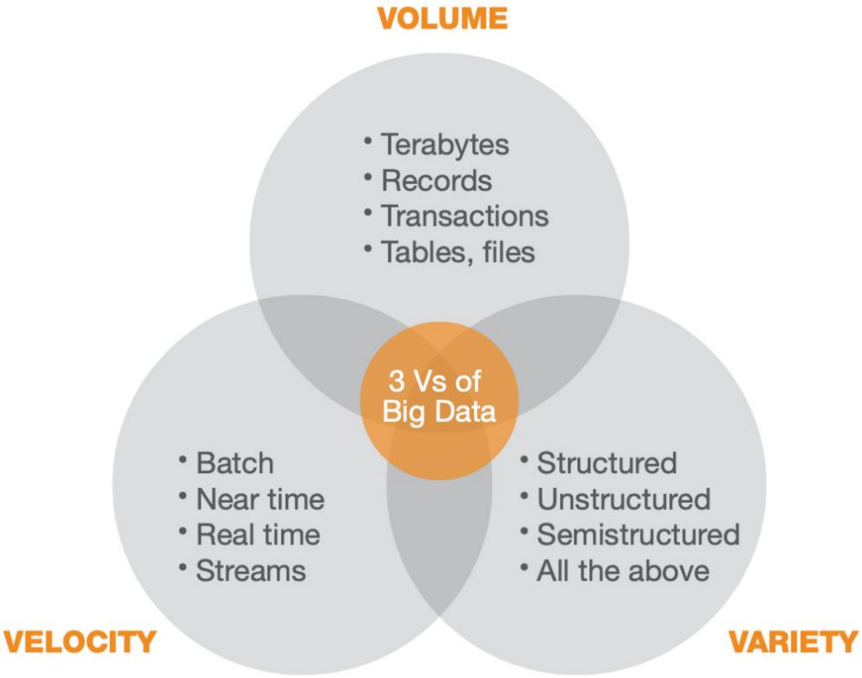
- In e-commerce, take into consideration the continuous flow of information coming from the clickstream of each user who interacts with the website’s products.
- In the healthcare, imagine the swift and seamless transmission of patients’ medical records via smart devices that are linked to health-monitoring sensors.
- In social media, platforms like Twitter, Facebook, and various social media networks constantly produce massive amounts of text, images, and video content worldwide. The continuous flow of social media data creates management and control issues. It is produced instantaneously, creating a dynamic data stream. With its vital insights into the current state of affairs and user behavior,

this data stream is immensely valuable. It empowers quick responses to ongoing events and gives businesses access to real-time data for crucial decisions based on public opinion.

- Data from the financial markets is a prime example of data velocity since it is created and transferred at incredibly fast speeds on a constant basis. To respond quickly to changes in the market and make decisions on time, analysis must be done right way (Ferguson, 2012).
- Other examples include similar requirements for quick processing that can be observed in situations like sensor and machine-generated data, as well as cameras that demand real-time analysis of both video and image data.

What truly amplifies the vastness of big data is its origin from an increasingly diverse array of sources. In this context, variety refers to a broad range of data types, including text, audio, video, webpages, and semi-structured and unstructured data as well as more traditional structured data. Variety holds an equally significant role as volume. Moreover, they frequently act as mutual drivers, enhancing one another. Online searches, history viewing, social media pictures and video sharing, business transactions, and data acquisition devices like Radio Frequency Identification (FRI) and the Internet of Things (IoT) are a few examples of these data sources. All of this data is captured using IT-based innovations, which means that most of it is unstructured (Marr, 2017).

Figure 1. The three Vs of big data



There are four more attributes that complement the three main components of big data, which are volume, velocity, and variety.

- **Value:** The quest for value in large data sets has become a central concern for researchers and practitioners alike. It entails determining the possible advantages and disadvantages of utilizing big data to obtain understanding of the present and future state of an organization. As an instance, Mishra et al. (2016) use data on temperature and weather patterns to demonstrate how Tesco improved its operating margins and highlight the importance of big data analytics in the context of supply chain management. Research suggests that businesses that implemented data-driven decision-making saw notable increases in productivity (Watson, 2014).
- **Variability:** Variability includes all possible problems that could arise from big data and require the creation of complex programs to interpret the data. With Jukić et al. (2015) clarify that the dynamic and evolving nature of data is highlighted by the variability of big data, which allows for the production of multiple interpretations from the same dataset at different times.
- **Veracity:** the existence of biases, noise, and anomalies in the data is related to this big data feature. Demchenko et al. further break down the veracity dimension into two components: data trustworthiness, which indicates the dependability of the data's source, and data certainty, which relates to computer and storage reliability.
- **Visualization:** when analyzing patterns and trends in data, visualization tools like graphs, charts, and pictures are essential. With the use of these tools, data becomes easier to access, comprehend, and remember for later use. Sivarajah et al. (2017) state that eBay creates interactive visuals from complicated datasets using a big data visualization tool called Tableau.

As for whether big data should have every one of these features, Kitchin (2014) points out that different data sets may show different combinations of features, and some data sets may only have one big data feature.

The real reason that the term big data has exploded in the last decades is *technology*. All new terms in this industry come originally from technology and it was coined by the people who were using Hadoop and other no SQL databases. They were looking for a way to talk about the technology that differentiated from the older relational databases. If we go to a big data conference right now, we will find out what they are actually talking about are those data sources. Sharda (2013) draws attention to how the technology landscape is changing and emphasizes that big data research is about combining business ideas, technologies, and statistical techniques that are necessary to build highly scalable systems that can collect, process, store, and analyze large amounts of data, including both structured and unstructured data. Additionally, Tarvis et al. (2013) mentions big data as a term coined by computer scientists to describe the rapid changes in technology.

Every vendor jumped to their definition of big data, and everybody wanted to elevate the debate. They came up with more commercial ways of looking at it. The following definition by Shaun Connolly (Hortonworks) is suggested as transactions, interactions, and observations. Transactions means the data we have been collecting for the long time; interactions are where people go to website or social media; observations is everything it is powered by sensors (machine to machine type data). Similar definition, according to the data types, was found that comes from Barry Devlin (invented the term data warehouse) and was process mediated data, human sourced information, and machine generated data. Another term for big data used by Steve Lucas (served in SAP, Adobe, and now is working for Boomi). He called big data "the signals of the future". The classic example is sentiment: your sales will go down in the future if your sentiment is plummeting now. That means big data helps as predict and intervene in the future before something goes wrong. The final broad definition we are going to talk about is big data as opportunity. It is called dark data from 451 research by Matt Aslett. In fact, it is analyzing data that was ignored before due to technology limitations.

As a result, since the 90s, there has been a steady increase in interest in business intelligence and analytics, a trend that is furthered by the emergence of the big data movement. Modern standardized software tools and well-established organizational and management strategies have been integrated into traditional business intelligence practices over time. The fierce of competition in the global digital economy is driving the next stage of analytics evolution, which is know as big data.

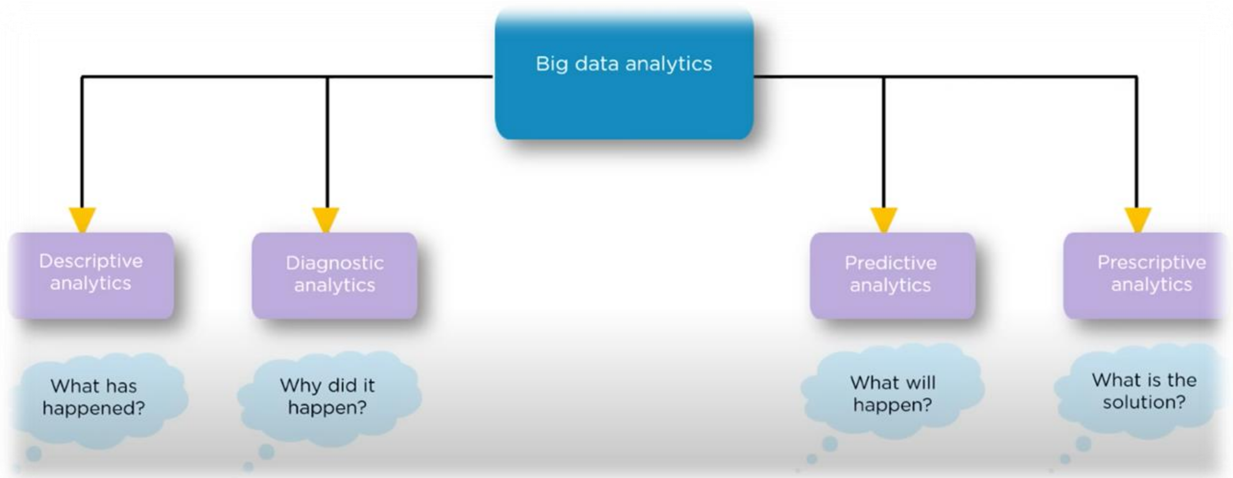
2.2 Big Data Analytics

Using computational techniques, analytics seeks to find and convey important patterns in data. Gaining insights and frequently influencing decisions are the goals of analytics. Analytics, by definition, examines historical data because data is a reflection of past information. The concepts behind analytics are not new; they have been expressed over the years under a variety of names, including cybernetics, data analysis, neutral networks, pattern recognition, statistics, knowledge discovery, data mining, and now data science. The term “analytics” became well-known around 2005, most notably with the launch of Google Analytics (Abbott, 2014). The rise in popularity of analytics recently is useful. It makes sense for organizations to use the data they collect and summarize in order to improve estimates, predictions, decisions, and, eventually, operational efficiency. Big data analytics is a study of huge amounts of stored data in order to extract behavior patterns. It helps companies or public administrations to understand the users better, find previously unnoticeable opportunities, provide a better service, and mitigate fraud. The fields where big data analytics is used are risk management, product development and innovations, decision making, and improving customer experience. A lifecycle begins with a business case which defines the reason and goals behind the analysis. The next step is to identify a broad variety of data source and move to filtering the identified data from previous stage to remove the corrupt data. The data that is not compatible with the tool is extracted and then transformed to a form that is compatible. In the data aggregation stage, data with the same fields across different dataset are integrated. The next step is the process of evaluating data using analytical and statistical tools to discover useful information. Visualization of data is done by graphically communicating the analysis results (tools are used at this stage). Final result is made available to business stakeholders for decision making.

2.2.1. Types of big data analytics

We have four different types of big data analytics. We have the descriptive analytics which asks the question: “What has happened?”. Diagnostic analytics asks: “Why did it happen?”. Predictive analytics asks: “What will happen?”. The last but not least, prescriptive analytics inquire: “What is the solution?”.

Figure 2. Types of big data analytics



In the context of big data, descriptive analytics entails compiling historical data into a human-readable format. This kind of analytics explores historical data and makes a variety of inferences. By using descriptive analytics, for example, a company can evaluate its past performance by looking at variables like revenue over time to identify trends and make financial decisions. Businesses can quickly understand the data by using graphical representations, which helps with decision-making. For businesses, descriptive analytics plays a crucial role in producing detailed reports and tallying different metrics on social media sites like Facebook. When used wisely in organizational strategies, historical data has the potential to do more for cost saving than just provide insights.

In the world of big data, diagnostic analytics aims to provide an explanation for why a particular issue arose. Its main goal is to determine the underlying cause of a problem and understand why it arises. Numerous methods, including data mining, data discovery, and drill-down analyses, are used in diagnostic analysis. This kind of analytics is valuable to businesses because it allows them to investigate the underlying causes of issues. Organizations can more effectively solve problems by proactively preventing the recurrence of similar issues in the future by understanding the root cause. Think about e-commerce websites where users add products to their carts but abandon them before completing the transaction. Analytical diagnostics helps identify the causes of this behavior. An e-commerce platform, for example, might investigate the reasons behind its poor online sales performance in spite of a strong marketing plan. Using diagnostic analytics, factors like expensive shipping costs, broken pages, or few payment options are examined. By going through this process, the business learns more about the particular difficulties, which enables it to make well-informed decisions to improve user experience and deal with underlying problems. Finding the causes guarantees that future occurrences of the same issues will be lessened.

Predictive analytics is the process of utilizing past and present data to make well-informed predictions about the future. The fundamental question of this strategy is “what will happen next?”. Predictive analytics, which makes use of artificial intelligence, machine learning, and data mining, works with probabilities in an effort to foresee consumer trends, changes in the market, and behavioral patterns based on historical and current data. Think of a business that manages millions of accounts, PayPal, and works to reduce online fraud. The concern over ongoing fraudulent incidents is a big problem that affects consumers’ trust in the brand. PayPal makes strategic use of predictive analytics to anticipate and reduce these types of fraudulent activities. PayPal uses customer behavior trends and payment history to build algorithms that forecast likely future transactions. The system can distinguish between legitimate and possibly fraudulent transactions thanks to the utilization of big data and complex algorithms. By being proactive, PayPal is able to protect its customers from possible fraudulent transactions and preserve the integrity of its services by implementing preventative measures.

As the name suggests, prescriptive analytics provides answers to particular issues, especially those that arise instantly. Prescriptive analytics is concerned with providing guidance towards the best possible solution for a given problem. It is related to both predictive and descriptive analytics, where descriptive analytics examines what has occurred and predictive analytics uses artificial intelligence and machine learning to predict what might happen next. Prescriptive analytics utilizes a range of algorithms, computational modeling techniques, and business rules to aid in decision-making. Let’s look at how prescriptive analytics is used for profit optimization in the airline industry. The cost of a plane ticket is determined by a wide range of internal and external variables, such as demand from customers, taxes, seat preference, and oil prices. Demand, availability, and seasonal variations like holidays or school breaks all affect price fluctuations. The destination is also very important because it determines how much the flight will cost based on the volume of air traffic there. By examining these various elements, prescriptive analytics creates an algorithm that automatically modifies airline tickets. By adjusting fares proactively to reflect changes in demand and the state of the market, the airline is able to maximize its profit.

2.2.2. The difference between Data Analysis and BDA

Traditional data analysis is scope-limited and frequently uses cross-sectional surveys, and it is based on structured data that is kept in relational databases. On the other hand, unstructured data that is kept in non-relational databases is included in big data analytics along with other types of data structures. Its scope is wider, as it can analyze and visualize copious amounts of data from various sources in real-time, surpassing the constraints of traditional analytics. The quality of data perceived differently, with traditional analysis benefiting from well-known sources, while big data analytics grapples with data collected from various origins. The primary focus of traditional analytics is on improving key performance indicators, whereas bog data analytics prioritizes following the flow of information and analyzing massive volumes of data in real-time.

	Traditional Data Analysis	***Big Data Analysis***
Structure	*Most data is stored in commercial relational database management systems, collected via legacy systems, and organized. The RDBMS was the main application for processing queries. (Chen, Chiang and Storey, 2012)*	*In addition to RDBMS, new Hadoop is now a competitive choice for BDA. (Chen, Chuang and Storey, 2012)*
Storage	*Relational databases store data in rows and columns that can be accessed with SQL queries.*	*Instead of depending just on SQL, BDA makes use of non-relational databases that can store data in any format. It examines unstructured or semi-structured data in a unique way.*
Content	*Due to limited number of respondents in focus groups and cross-sectional questionnaire surveys, traditional fata*	*By leveraging online reviews as data for analysis, BDA can get around limitations that come with more*

	analytics, which focuses on examining customer satisfaction, has limitations. (Guo, Barnes and Jia, 2017)	*conventional approaches. (Guo, Barnes and Jia, 2017)*
Scope	*Due to narrow scope of traditional data sources, there is less information, fewer facts, and uncertainty in the decisions made.*	*Combining data from various sources allows BDA to analyze and visualize a wider range of information, giving decision-makers more facts to work with.*
Quality	*Because data processing used only few, well-known sources, data quality was once thought to be good. (Hashem et al., 2015)*	*Big Data is often associated with low-quality data because it gathers information from many sources. (Hashem et al., 2015)*
Aim	*In order to gain deeper insights, traditional analytics mainly focus on enhancing key performance indicators.*	*The use of BDAs seeks to track information flow and perform real-time analysis on massive amounts of data.*

Table 1: The Difference between Data Analysis and BDA

2.2.3. BDA Tools and Vendor products

Based on the framework proposed by Chen and Zhang (2014), analytics tools in the realm of Big Data Analytics fall into three distinct classes:

1. Batch Process Tools: These tools are based on batch processing and leverage platforms like Apache Hadoop. Batch processing involves the analysis of data in fixed-size blocks or batches. Hadoop, with its distributed storage and processing components, enables the storage and parallel processing of large datasets across clusters of computers.
2. Stream Processing Tools: This class includes tools designed for real-time processing of streaming data. Examples mentioned by Chen and Zhang include Apache Storm, S4, and Tibco StreamBase. These tools enable the analysis of data as it is generated, allowing for immediate insight and actions.
3. Interactive Analysis Tools: This class encompasses tools that facilitate interactive and ad-hoc analysis of data. Chen and Zhang mention tools like Tableau, which provide users with the ability to explore, visualize, and analyze data interactively. These tools often present data in graphical formats, tables, charts, and maps.

Chen and Zhang’s classification highlights the diversity of tools available for big data analytics, each catering to specific use cases and processing requirements (Chen and Zhang, 2014). It recognizes the importance of considering the nature of the data, the analytical tasks at hand, and the desired outcomes when selecting the appropriate tools and platforms for a given big data analytics project.
Here is an overview of the vendor products and their use in the context of big data analytics:

- Cloudera:
  - Product: Cloudera Enterprise
  - Description: Cloudera specializes in distributing open-source software based on Apache Hadoop. Cloudera Enterprise is designed to enhance the manageability of Hadoop deployments, making it suitable for serious enterprise users. It includes technical support, upgrades, administrative tools, professional services, training, and certification.

    - Use for Big Data Analytics: Cloudera's product makes Hadoop more accessible and manageable for enterprises, ensuring reliability in production. It extends Hadoop's capabilities while retaining its open-source nature, scalability, and flexibility, catering to a wide range of data types.
- IBM:
    - Products:
        - Netezza
        - IBM InfoSphere Biglnsigths
        - IBM InfoSphere Streams
        - IBM SPSS
    - Description: IBM boasts one of the largest product portfolios, with a focus on analytics. Netezza is a data warehouse appliance, and InfoSphere Biglnsights is a Hadoop-based offering. InfoSphere Streams is a platform for real-time analytics. Additionally, IBM SPSS is a comprehensive statistical analysis software widely used for predictive analytics and data mining.
    - Use for Big Data Analytics: IBM SPSS is instrumental in predictive analytics and data mining. It allows users to analyze large datasets, identify trends, make predictions, and derive valuable insights. When combined with other IBM products, it enhances the overall analytics capabilities of the organization, offering a holistic approach to big data analysis.
- SAP:
    - Product: SAP HANA
    - Description: SAO has deepened its commitment to analytics through internal development and acquisitions. SAP HANA is an enterprise software architecture that enables real-time analytic queries without the need for transforming data into optimized models.
    - Use for Big Data Analytics: SAP HANA allows for real-time analytics without predefining queries, providing scalability for big data and speed for discovery analytics.
- SAS:
    - Products:
        - SAS High-Performance Computing
        - SAS On Demand
        - SAS Data Integration Studio
    - Description: SAS is known for predictive analytics. SAS High -Performance Computing supports big data initiatives with in-memory, in-database, and grid computing support. SAS On Demand supports private and public clouds. SAS Data Integration Studio provides support for Hadoop.
    - Use for Big Data Analytics: SAS products are designed to handle big data with capabilities for predictive analytics, data management, and integration.
- Tableau Software:
    - Product: Tableau
    - Description: Tableau is renowned for its strong visualization features, enabling exploratory or discovery analytics. It is used as a BI platform, offering a visual approach that allows uses to interact directly with visualization for queries, reports, and analyses.

- Use for Big Data Analytics: Tableau’s visual approach facilitates high ease of use, enabling self-service analytics. Analysts can explore and discover insights without waiting for IT assistance. With simple drag-and-drop methods, users can access and analyze big data in a self-service manner.

These vendor products cater to different aspects of big data analytics, offering solutions for data storage, processing, real-time analytics, visualization, and self-service analytics, among others. Each product brings unique features and capabilities to the table, contributing to the diverse landscape of tools available for big data analysis.

2.3 Predictive Analytics

With insights from a wide range of fields, including pattern recognition, statistics, machine learning, artificial intelligence, and data mining – some of which have a foundation that dates back more than a century – predictive analytics explores the complex patterns that exist within data. What distinguishes predictive analytics from its competitors? To begin with, it is a model of data-driven methodology, in which algorithms extract key model attributes straight from the data, without depending on analyst-crafted assumptions. By identifying variables, model parameters, weights, coefficients, or even by navigating the complexities of the model, these data-driven algorithms essentially create models (Abbott, 2014). Furthermore, the intricate process of identifying patterns in the data is automated by the algorithms that support predictive analytics. These complex algorithms define the basic structure of the models in addition to locating coefficients or weights for them. A variety of software programs and algorithms feature automation characteristics that skillfully modify input variables so that they can be seamlessly integrated for the best possible use in predictive models.

2.3.1. Business Intelligence

Predictive analytics and business intelligence work hand in hand in the field of data-driven decision-making. As a broad field, business intelligence includes all the instruments, procedures, and methods used to gather, examine, and display business data in order to support well-informed decision-making. It provides the essential framework and insights from both historical and current data, acting as the cornerstone for predictive analytics. Business intelligence is the forerunner in predictive analytics, providing the necessary data foundation. Patterns, trends, and important indicators can be found through data exploration and analysis carried out using business intelligence platforms. Predictive models then use these insights as their building blocks. By utilizing the foundation established by business intelligence, predictive analytics makes predictions about future patterns, actions, and results. Thus, in the context of a larger data analytics landscape, business intelligence offers the context and data exploration tools that enable predictive analytics to draw meaningful conclusions.

On the other hand, predictive analytics and business intelligence are essentially different in how they handle data exploration, modeling, and the analysis’s main motivation. Variables in business intelligence are usually found using requirements and queries driven by the user. The questions are defined by the users, and the analysis centers on answering those particular queries. The emphasis is on producing reports and visualizations that help with business performance understanding, as well as insights into current and historical data. Users can interactively explore data and extract information based on predefined questions with the help of business intelligence. Predictive analytics, on the other hand, adopts a more proactive and data-driven strategy. Predictive modeling algorithms sort through enormous datasets, taking onto account a variety of patterns, and identifying which ones are most predictive of the target measure – which frequently involves the likelihood of an event or outcome – instead of depending on user-defined queries. The data itself finds patterns, which makes it possible to identify correlations and trends that conventional business intelligence queries might miss. This distinction is reflected in the word "which" used in questions related to predictive analytics (Abbott, 2014). In predictive analytics, "which" questions frequently start with "which" because the goal of the analysis is to identify the variables that have the greatest influence on predicting a specific result.

2.3.2. Statistics

Predictive analytics heavily relies on statistics to glean insightful information and generate well-informed forecasts. When it comes to making predictive models more accurate, dependable, and understandable in the context of big, intricate datasets, statistics is essential. The creation of predictive models is one of the main uses of statistics in big data predictive analysis. The large and diverse datasets are analyzed using statistical methods like regressions analysis, time series analysis, and machine learning algorithms to find patterns, relationships, and trends. These models use statistical concepts to find patterns in the data and forecast future outcomes based on past performance.

Statistical techniques help identify the most influential variables in the context of feature selection, an essential step in developing predictive models. The statistical significance of each variable is ascertained using techniques like chi-squared tests and analysis of variance (ANOVA), which facilitate the selection of the variables that are most pertinent to the predictive task.

Table 2 highlights key distinctions between statistics and predictive analytics, emphasizing their differing approaches to modeling and data:

Statistics	Predictive analytics
Models based on theory: There is an optimum.	Models often based on non-parametric algorithms; no guaranteed optimum.
Models typically linear.	Models typically nonlinear.
Data typically smaller; algorithms often geared toward accuracy with small data.	Scales to big data; algorithms not as efficient or stable for small data.
The model is king.	Data is king.

Table 2: Statistics vs. Predictive Analytics (Abbott, 2014)

These variations highlight the divergent approaches in each discipline: predictive analytics emphasizes adaptability, nonlinearity, and the importance of large datasets, while statistics tends toward theoretical models and optimization.

2.3.3. Data Mining

With similar techniques and algorithms, data mining is the forerunner of predictive analytics. Through a variety of computational techniques, it entails the process of obtaining patterns, information, and insights from large datasets. Applications of data mining have a long history in a variety of industries, including marketing, customer relationship management, biotechnology, engineering, finance, and manufacturing. Finding hidden patterns and relationships in data is the main goal of data mining, which frequently has no explicit prediction objective. Finding important knowledge and information that may not be immediately obvious is the goal in order to provide insights into behaviors. Anomaly detection, association rule mining, and clustering are examples of traditional data mining techniques.

Comparison with predictive analytics:

- Algorithms and Approaches: Data mining and predictive analytics frequently employ comparable techniques and algorithms. To analyze and interpret data, both domains make use of computational tools, machine learning strategies, and statistical models.
- Predictive Objective: Predictive analytics primarily attempts to forecast future outcomes, whereas data mining investigates patterns without a predetermined predictive goal. Finding patterns is just one aspect of predictive analytics; its main goal is to use patterns to forecast future events.
- Application Focus: Historically, data mining has been used to provide insights into historical data trough retrospective analysis and knowledge discovery. On the other hand, predictive analytics looks ahead, attempting to predict patterns in the past to identify future trends and behaviors.
- Education and Branding: In graduate programs as well as the software industry, the terms “predictive analytics” and “data mining” are frequently used interchangeably. The two fields share common principles and techniques, which are reflected in their general overlap in use.

2.4. Unveiling the Benefits of Big Data Analytics

In recent years, Big Data Analytics has become a game-changer, reshaping the way organizations handle, interpret, and leverage vast and intricate datasets. This section of the thesis delves into the theoretical underpinnings of the myriad advantages offered by Big Data, focusing on its pivotal role in decision-making, operational efficiency, customer experience, innovation, risk management, and resource optimization.

1. Enhanced decision-Making.
   Big Data Analytics equips decision-makers with a powerful toolset for making informed choices. It is a general observation about the benefits of big data analytics. Real-time analysis of extensive datasets provides deeper insights into market trends, customer behaviors, and operational nuances. Data-driven decision-making enables organizations to fine-tune strategies, navigate risks, and seize emerging opportunities.
2. Improved Operational Efficiency.
   The adoption of Big Data tools streamlines operational processes, leading to heightened efficiency. By scrutinizing historical and real-time data, organizations can identify operational

bottlenecks, optimize workflows, and allocate resources more effectively. Predictive analytics contributes to proactive problem-solving, curbing downtime and reducing operational costs.

3. Personalization and Customer Experience.
   Big Data empowers organizations to grasp individual customer preferences and behaviors. Through the analysis of customer interactions, purchase histories, and feedback, businesses can tailor products and services to specific needs. This personalized approach not only elevates customer satisfaction but also fosters loyalty and retention.
4. Innovation and Product Development.
   The abundance of data accessible through Big Data Analytics fuels innovation and propels product development. Organizations can uncover fresh insights into market demands, identify gaps in existing offerings, and predict future trends. This proactive approach ensures that products align with evolving customer expectations, providing a competitive edge.
5. Effective Risk Management.
   Big Data Analytics plays a crucial role in risk assessment and management. By analyzing historical data and identifying patterns, organizations can anticipate potential risks and implement preventive measures. This not only minimizes the impact of unforeseen events but also enhances organizational resilience in dynamic business environments.
6. Economies of Scale and resource Optimization.
   The scalability of Big Data solutions enables organizations to efficiently manage extensive data without proportional increases in costs. Cloud-based Big Data platforms, particularly, offer flexible and cost-effective solutions, allowing businesses to scale their analytics capabilities as needed, leading to economies of scale and optimal resource utilization.

This theoretical exploration highlights the transformative potential of Big Data Analytics across diverse organizational dimensions. By leveraging the power of data, organizations can make informed decisions, enhance operational efficiency, drive innovation, manage risks effectively, and optimize resources – positioning themselves for sustained success in an increasingly data-centric landscape.

2.5. Challenges in Handling Big Data

The ethical landscape surrounding Big Data applications is rife with challenges, with a predominant concern being privacy, as highlighted by Hull (2015) and Boyd and Crawford (2012). Hull emphasizes the inadequacies of privacy legislation based on the notion of privacy self-management, contending that users lack sufficient information and understanding about the potential uses of their data when providing consent. This gap in awareness raises critical questions about the ethicality of data utilization.

Boyd and Crawford (2012) contribute to the discourse by shedding light on the unforeseen uses of social media data. Their case study involving researchers accessing Facebook profiles without user knowledge exemplifies the ethical dilemma of utilizing data in research without explicit consent. The availability of data, they argue, should not justify unrestricted utilization, urging a reconsideration of ethical norms in the realm of Big Data analytics.

The complexity is further amplified when considering third-party data handling, as elucidated by Hull (2015). The sale of data to entities like Facebook introduces the risk of divergent third-party policies, leading to potential unforeseen consequences. Facebook’s use of user likes to predict personal characteristics without explicit consent serves as an example of the ethical challenges posed by third-party data handling, prompting inquiries into accountability and transparency.

A common thread in both Hull's and Boyd and Crawford's arguments is the paramount importance of user awareness and informed consent. The distinction between being in public and being public becomes a pivotal consideration. Users must be cognizant of how their data will be utilized and the potential consequences, underscoring the need for transparency and user autonomy in the Big Data ecosystem.

Addressing the ethical challenges in applying Big Data requires a reevaluation of existing legislation, an enhancement of user awareness, and the establishment of transparent practices. The ethical application of Big Data necessitates a paradigm shift, ensuring that users are well-informed, empowered to make decisions about their data, and safeguarded against unintended consequences arising from its utilization.

The integration of Big Data into research and decision-making processes has become pervasive, yet the critical reflections of Boyd and Crawford (2012) and Humberty (2015) draw attention to inherent challenges that merit careful consideration. Boyd and Crawford caution against the tendency to treat Big data as an undifferentiated whole, pointing to the prevalent issue of sample bias. Using the example of Twitter data, they illuminate the intricate nature of Twitter users, emphasizing factors like the existence of multiple accounts, shared usage, and the selective availability of public Tweets through APIs. Their assertion that the size of a dataset loses significance when the sample is inherently skewed underscores a fundamental concern in the application of Big Data.

Building on this concern, Humberty (2015) challenges four prevailing false assumptions about the possibilities afforded by Big Data. First, the notion that Big Data renders traditional sampling unnecessary and enables the study of entire populations is debunked. Second, the assumption that one can definitively identify the data sources necessary to comprehend entire populations is deemed overly optimistic. Third, the belief that online behavior is an accurate reflection of offline behavior is questioned, particularly considering the segment of the population that does not engage with online services. Lastly, Humberly contests the idea that historical data can reliably predict future outcomes, emphasizing the dynamic nature of the digital landscape and the ever-changing user preferences.

Moreover, Humberly sheds light on the homogeneity of online user bases, noting that they often represent specific subsets of the larger population. This challenges the assumption that behaviors observed within these online subsets are broadly representative of the entire population. Furthermore, he underscores the dynamic nature of online behavior, pointing out that individuals evolve, as does their interaction with online platforms. The continual evolution of language and user preferences in the digital realm raises questions about the applicability of historical data in predicting future trends.

Effectively navigation the methodological challenges presented above necessitates a nuanced comprehension of sample biases, a meticulous scrutiny of assumptions, acknowledgement of the homogeneity inherent in online user bases, and recognition of the dynamic nature characterizing online behavior. Researchers and practitioners must exercise prudence to guarantee the reliability and applicability of their findings as they harness Big Data for research and decision-making.

The utilization of Big Data presents distinctive challenges in management, as outlined by McAfee and Brynjolfsson (2012). This theoretical exploration identifies five key management challenges imposed by Big Data, starting with the imperative need for skilled leaders capable of recognizing opportunities, articulating clear visions, and effectively translating these visions into reality. Talent management emerges as a pivotal challenge due to the rarity of individuals possessing both technical proficiency in handling extensive datasets and the ability to align business needs with the problem-solving potential of Big Data.

Additionally, Big Data technologies pose unfamiliar terrain for IT departments, constituting a challenge in adapting organizational units to efficiently engage with these novel technologies. The imperative to bring together diverse skill sets – those proficient in understanding business challenges, well-versed in data intricacies, and competent in handling emerging technologies – underscores the challenge of organizational collaboration.

However, perhaps the most profound challenge lies in cultural transformation. Shifting an organizational culture from decision-making based on intuition to one grounded in data-driven insights demands a fundamental shift in mindset. McAfee and Brynjolfsson (2012) contend that data-driven organizations alter their approach, transitioning from asking, "What do we think?" to the more analytical query, "What do we know?" This cultural transformation is intricate, necessitating a comprehensive reevaluation of decision-making processes and organizational values.

Successfully applying Big Data requires strategic management responses to navigate challenges such as effective leadership, talent management, IT department adaptation, organizational collaboration, and cultural transformation. This theoretical framework emphasizes the pivotal role of management in steering organizations through the complexities inherent in the application of Big Data, stressing the need for visionary leaders and adaptive organizational structures to unlock the full potential of Big Data analytics.

2.6. Big Data analysis with IBM SPSS

In the contemporary era of extensive data, the application of advanced analytics tools is paramount for meaningful insights (Field, 2017). This thesis focuses on the practical implementation of IBM SPSS Statistics, drawing from the comprehensive insights and methodologies outlined in the referenced work.

The statement suggests that IBM SPSS Statistics can indeed be implemented for practical application in the context of advanced analytics. IBM SPSS Statistics is a powerful software tool specifically designed for statistical analysis and data mining tasks. It offers a wide range of analytical techniques and methodologies that can be used to derive meaningful insights from data.

By referencing the work of Field (2017) and focusing on the practical implementation of IBM SPSS, the thesis aims to leverage the comprehensive insights and methodologies outlined in Field's work within the framework of IBM SPSS Statistics. This implies that the researcher intends to utilize the capabilities of IBM SPSS Statistics to apply advanced analytics techniques and derive meaningful insights from data, aligning with the overarching objective of the thesis.

Exploratory Data Analysis:

Exploratory Data Analysis holds a foundational role, providing crucial insights into intricate patterns within vast datasets (Field, 2017). Utilizing IBM SPSS, researchers can employ graphical representations and summary statistics, enhancing their understanding of the expansive data landscape.

Descriptive and Inferential Statistics:

Transitioning beyond EDA, there is a spectrum of statistical techniques applicable to big data analytics. From basic descriptive statistics that summarize and describe data, to inferential statistics that draw conclusions and make predictions. IBM SPSS, as the software companion, facilitates the execution of these techniques on large and complex datasets.

Regression Analysis and Predictive Modeling:
Regression analysis, fundamental in predictive modeling, is examined in detail. IBM SPSS can be employed to perform regression analyses, allowing for the prediction of outcomes based on various predictor variables. This capability is particularly relevant in big data scenarios where predictive insights are crucial.

Multivariate Analyses and Machine Learning Integration:
Multivariate analyses, including factor analysis and cluster analysis, find practical applications in big data analytics (Field, 2017). Furthermore, the integration of machine learning concepts with IBM SPSS is explored, showcasing the software's capabilities in classification, clustering, and predictive modeling within extensive datasets.

The insights and methodologies in conjunction with IBM SPSS offer a robust foundation for navigating the challenges of big data analytics. The combination of Field's insights with IBM SPSS equips users with the knowledge and tools essential for extracting meaningful insights from vast and intricate datasets, thus contributing significantly to the evolving landscape of big data analytics.

2.7. Prospects in Various Industry Fields

2.7.1 Manufacturing

The integration of big data is integral to the business models of "born digital" companies such as Google and Facebook, constituting a core element of their operations. In contrast, the manufacturing sector is often perceived as traditional, grounded in well-established practices. A survey conducted by Oxford Economics, polling 363 business executives on their perspectives regarding digitalization, suggests that manufacturing may not undergo a radical transformation through information through information technology (Oxford Economics, 2011).
However, the actual landscape reveals that big data introduces diverse and extensive opportunities for enhancing manufacturing processes. Li et al. (2015) emphasizes the vital role of optimizing manufacturing processes in the current globalized economy, achievable through the analysis of various data sources. Technologies like RFID and wireless sensor networks emerge as valuable tools for acquiring data from almost any manufacturing phase or domain.
Dutta and Bose (2015) provide examples of successful big data initiatives in manufacturing companies. Merck utilizes big data for faster vaccine development, Volvo forecasts component failures, and Xerox employs telemetric data for improved service quality and reduced costs. Li et al. (2015) proposes the use of scientific workflow technology to optimize shop floor workflow and scheduling.

2.7.2. Energy production

The prevailing concept in energy production today is the implementation of smart grids. A smart grid refers to the evolution of the traditional power grid, integrating computation, communications, and control systems to optimize the generation, supply, and consumption of electric energy (Chen et al., 2014a). Given the extensive coverage of power networks across vast geographical areas, they present unique opportunities for data collection related to electricity production and consumption. In the era before smart electric meters, the measurement of electricity consumption was a labor-intensive process, lacking real-time data retention. The convergence of detailed and timely records of power consumption with additional data sources, such as electricity prices or weather data, opens avenues for predicting and optimizing power supply and demand, grid maintenance, and electricity prices (Chen et al., 2014a).

2.7.3. Retail

The retail landscape is currently undergoing a significant transformation, propelled by the wave of digitalization. This shift has unlocked vast opportunities to amass extensive data for optimizing operational and marketing strategies. Major global e-commerce players like eBay and Amazon are at the forefront, leveraging data intensively and challenging the traditional brick-and-mortar retail model. According to Pousttchi and Huffenbach (2014), the utilization of customer data empowers retailers to proactively understand customer needs, tailor personalized offers, and provide guidance on purchasing decisions. They argue that employing data-driven personalized marketing could lead to a substantial increase in response rates and a boost in sales by 20-30%. Tools within the realm of data-driven personalized marketing encompass the use of ratings, analysis of frequently co-purchased items, and location-based marketing. These tools support cross- and up-selling initiatives, ensuring that customers receive the right offers at the right time. Customer-generated product ratings, service ratings, and reviews offer valuable feedback to enhance customer service and satisfaction. Additionally, data analytics can provide insights to optimize pricing strategies for maximum profitability (Pousttchi & Huffenbach, 2014). In addition to marketing enhancements, predictive analytics and machine learning play a crucial role in optimizing physical operations and the supply chain. Amazon, for instance, holds a patent for a system that predicts customer orders before they are placed, potentially enabling the company to reduce delivery times and optimize stock levels (Bensinger, 2014). Accurate and swift trend prediction aids retailers in anticipating demand for various products, facilitating agile adjustments in supply. This is vital for apparel retailers heavily influenced by ever-changing trends. In essence, big data opens up a myriad of opportunities for retailers to enhance customer engagement, optimize operations, and stay ahead of market trends.

2.7.4. ICT

The emergence of big data as a phenomenon unfolds extensive opportunities for the Information and Communication Technology (ICT) sector, serving as the primary conduit for big data applications (Lu et al., 2013). The integration of the Internet of Things (IoT) and big data amplifies the need for data-collecting sensors and robust wireless networks to efficiently transmit and manage the data. Big data assumes a central role in optimizing the capacity of ICT infrastructure, ensuring its reliability and security in delivering services to users. Furthermore, ICT companies can augment their value proposition by offering customer-centric services rooted in data and analytics. In essence, big data serves as a catalyst for transformative advancements within the ICT sector, fostering innovation and enhancing the range of services provided to users.

2.7.5. Banking

The harnessing of big data analytics capabilities opens significant opportunities for banks to extract valuable insights and enhance various aspects of their business. Srivastava and Gopalkrishnan (2015) outline three key domains of possibilities:

1. Customer-Centric Approaches:
   Big data enables customer-centric tools related to marketing and sales. By analyzing customer feedback, including ratings, banks can enhance the overall customer experience. Additionally, it aids in more effective lead and referral management. The analysis of data allows for the assessment of lead quality. Data can b leveraged to micro-segment offerings and identify next-best-offers for customers.
2. Risk Management:
   Big data analytics allows for the analysis of transactions to identify abnormal patterns indicative of potential criminal activities, such as money laundering or unauthorized credit card usage.
3. Transaction Insights:
   Transactions offer valuable insights into trade dynamics, future capital flows, and market sentiment. This information can be effectively utilized to enhance the management of investments and loans, as well as support comprehensive market analysis.

2.7.6. Logistics

Big data is gaining significant attention in supply chain management. Frehe et al. (2014) conducted a literature review highlighting 14 cases of big data initiatives on logistics. These initiatives optimize routes, driving times, and fuel consumption to cut costs and reduce CO2 emissions. Big data is also used for warehouse optimization, generating reports like CO2 emission reports, providing data about transportation and warehousing, offering analysis tools to customers, and supporting pricing negotiations.

3. Practical Part

In today’s airline industry, understanding passenger satisfaction is crucial for airlines striving to stay competitive and build customer locality. To gain insights into passenger experiences and preferences, airlines often conduct extensive surveys. The dataset we are examining contains the results of one such survey, offering a wealth of information on airline passenger satisfaction.

In the practical part of the thesis, several methods were applied to analyze the dataset on airline passenger satisfaction obtained from Kaggle, a prominent platform for datasets and data science competitions. This dataset provides valuable insights into passenger experiences and preferences, which are crucial for airlines aiming to enhance customer satisfaction, loyalty, and competitiveness (TJ Klein, 2020).

The methods employed in the analysis included Binary Logistic Regression Analysis, TwoStep Cluster Analysis, Descriptive Statistics, and Frequency Tables. Through these analyses, the aim was to gain a comprehensive understanding of passenger satisfaction and identify actionable insights to improve customer experience and build customer loyalty in the airline industry.

Comprising various variables, including demographic details and satisfaction ratings across different aspects of air travel, this dataset provides a comprehensive view of factors influencing passenger perceptions. By analyzing this dataset, researchers and industry professionals can uncover valuable insights into what drives passenger satisfaction and identify areas for improvement in airline operations.

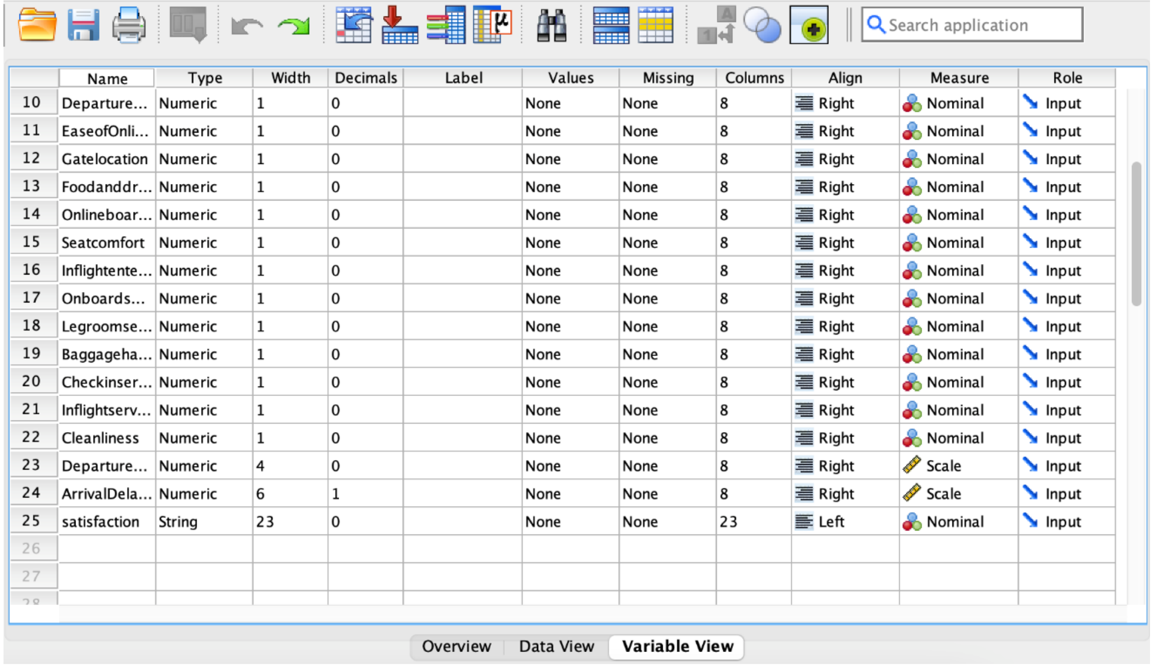
In this prescriptive analysis, our aim is to thoroughly explore the dataset. We will examine the variables, distribution of key metrics, and initial observations on potential trends or patterns. Through this exploration, we lay the groundwork for further analyses to understand the complex relationship between passenger demographics, travel-related factors, and overall satisfaction levels in the airline industry.

The dataset on airline passenger satisfaction contains information collected from a survey conducted among airline passengers. Here is a description of the dataset variables:

1. Gender: Indicates the gender of the passengers (Female or Male).
2. Customer Type: Specifies the type of customer (Loyal customer or Disloyal customer).
3. Age: Represents the actual age of the passengers.
4. Type of Travel: Identifies the purpose of the flight for the passengers (Personal Travel or Business Travel).
5. Class: Specifies the travel class in the plane for the passengers (Business, Eco, or Eco Plus).
6. Flight Distance: Indicates the flight distance of the journey.
7. Inflight Wi-Fi Service: Represents the satisfaction level with the inflight Wi-Fi service (on scale from 0: Not Applicable to 5: Very Satisfied).
8. Departure/ Arrival Time Convenient: Indicates the satisfaction level with departure/ arrival time convenience.
9. Ease of Online Booking: Indicates the satisfaction level with online booking.
10. Gate Location: Indicates the satisfaction level with the gate location.
11. Food and Drink: Represents the satisfaction level with food and drink services.
12. Online Boarding: Indicates the satisfaction level with online boarding.
13. Seat Comfort: Represents the satisfaction level with seat comfort.
14. Inflight Entertainment: Indicates the satisfaction level with inflight entertainment.
15. On-board Service: Represents the satisfaction level with onboard services.
16. Leg Room Service: Indicates the satisfaction level with leg room service.
17. Baggage Handling: Represents the satisfaction level with baggage handling.
18. Check-in Service: Indicates the satisfaction level with check-in service.
19. Inflight Service: Represents the satisfaction level with inflight service.
20. Cleanliness: Indicates the satisfaction level with cleanliness.

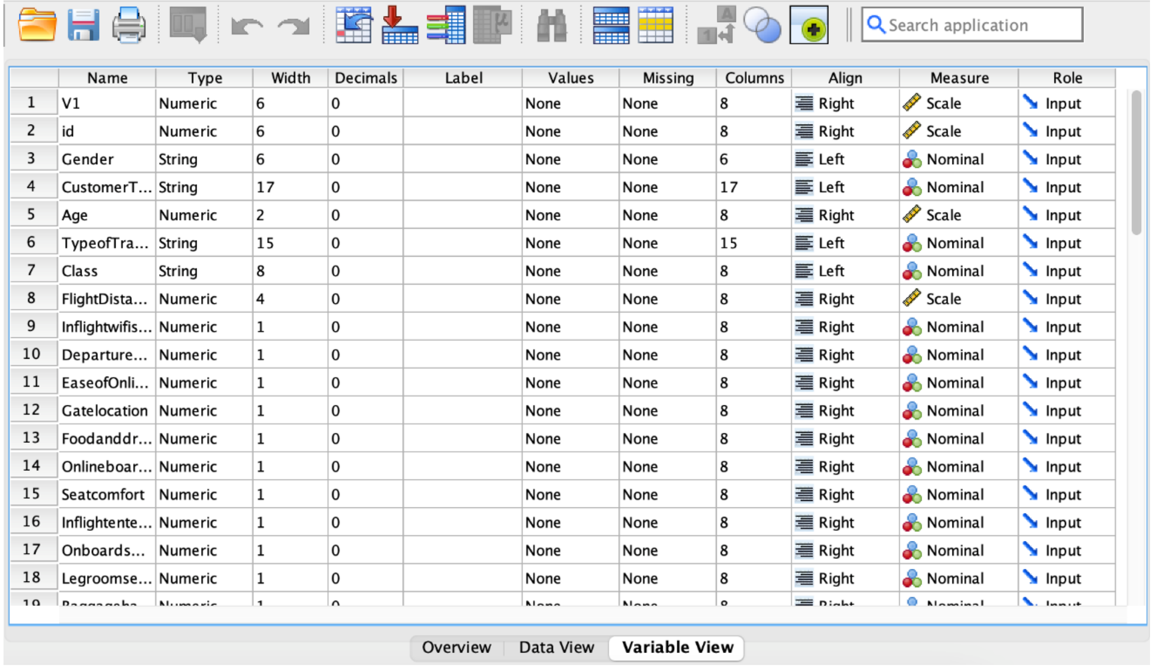
21. Departure Delay in Minutes: Represents the number of minutes delayed during arrival.
22. Arrival Delay in Minutes: Represents the number of minutes delayed during arrival.
23. Satisfaction: Indicates the airline satisfaction level of passengers (Satisfaction, Neutral, or Dissatisfaction).

Figure 3. Airline passengers' satisfaction variable view (1)



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
10	Departure...	Numeric	1	0		None	None	8	Right	Nominal	Input
11	EaseofOnli...	Numeric	1	0		None	None	8	Right	Nominal	Input
12	Gatelocation	Numeric	1	0		None	None	8	Right	Nominal	Input
13	Foodanddr...	Numeric	1	0		None	None	8	Right	Nominal	Input
14	Onlineboar...	Numeric	1	0		None	None	8	Right	Nominal	Input
15	Seatcomfort	Numeric	1	0		None	None	8	Right	Nominal	Input
16	Inflightente...	Numeric	1	0		None	None	8	Right	Nominal	Input
17	Onboards...	Numeric	1	0		None	None	8	Right	Nominal	Input
18	Legroomse...	Numeric	1	0		None	None	8	Right	Nominal	Input
19	Baggageha...	Numeric	1	0		None	None	8	Right	Nominal	Input
20	Checkinser...	Numeric	1	0		None	None	8	Right	Nominal	Input
21	Inflightserv...	Numeric	1	0		None	None	8	Right	Nominal	Input
22	Cleanliness	Numeric	1	0		None	None	8	Right	Nominal	Input
23	Departure...	Numeric	4	0		None	None	8	Right	Scale	Input
24	ArrivalDela...	Numeric	6	1		None	None	8	Right	Scale	Input
25	satisfaction	String	23	0		None	None	23	Left	Nominal	Input
26											
27											

Figure 4. Airline passengers' satisfaction variable view (2)



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	V1	Numeric	6	0		None	None	8	Right	Scale	Input
2	id	Numeric	6	0		None	None	8	Right	Scale	Input
3	Gender	String	6	0		None	None	6	Left	Nominal	Input
4	CustomerT...	String	17	0		None	None	17	Left	Nominal	Input
5	Age	Numeric	2	0		None	None	8	Right	Scale	Input
6	TypeofTra...	String	15	0		None	None	15	Left	Nominal	Input
7	Class	String	8	0		None	None	8	Left	Nominal	Input
8	FlightDista...	Numeric	4	0		None	None	8	Right	Scale	Input
9	Inflightwifis...	Numeric	1	0		None	None	8	Right	Nominal	Input
10	Departure...	Numeric	1	0		None	None	8	Right	Nominal	Input
11	EaseofOnli...	Numeric	1	0		None	None	8	Right	Nominal	Input
12	Gatelocation	Numeric	1	0		None	None	8	Right	Nominal	Input
13	Foodanddr...	Numeric	1	0		None	None	8	Right	Nominal	Input
14	Onlineboar...	Numeric	1	0		None	None	8	Right	Nominal	Input
15	Seatcomfort	Numeric	1	0		None	None	8	Right	Nominal	Input
16	Inflightente...	Numeric	1	0		None	None	8	Right	Nominal	Input
17	Onboards...	Numeric	1	0		None	None	8	Right	Nominal	Input
18	Legroomse...	Numeric	1	0		None	None	8	Right	Nominal	Input
19	Baggageha...	Numeric	1	0		None	None	8	Right	Nominal	Input

This dataset provides a comprehensive overview of various factors potentially influencing passenger satisfaction, including demographic information, travel-related details, and satisfaction ratings for different aspects of the airline experience. It can be used for applying classification and clustering methods.

Before applying classification and clustering techniques to analyze airline passenger satisfaction, it is imperative to define the specific aims of the analysis. The primary objectives include:

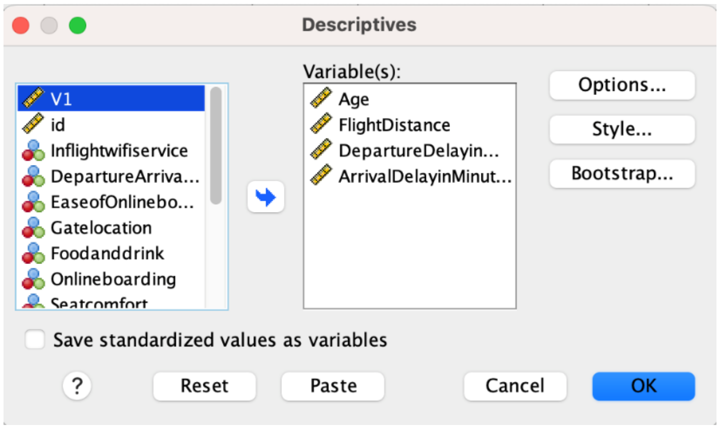
- Identifying factors influencing passenger satisfaction: Through classification analysis, the aim is to detect patterns and relationships within the dataset that contribute to passenger satisfaction or dissatisfaction. This involves determining the key factors that influence satisfaction levels, such as demographic characteristics, travel preferences, and inflight experiences.
- Segmenting passengers based on satisfaction levels or personal characteristics: Utilizing clustering analysis, the goal is to identify distinct customer segments or profiles based on their satisfaction levels, travel behaviors, and demographic attributes. This segmentation can provide valuable insights for targeted marketing strategies and personalized customer experiences.

In addition to clarifying the aims of the analysis, deep data understanding is paramount. This involves thoroughly describing the dataset using appropriate descriptive characteristics and visualizations.

3.1. Descriptive Characteristics

Descriptive statistics provide a summary of the central tendency, dispersion, and shape of the data distribution, allowing researchers to explore key features of the dataset.
In this analysis, we will compute measures such as the mean, median, standard deviation, minimum, and maximum for numerical variables like Age, Flight Distance, Departure Delay in Minutes, and Arrival Delay in Minutes. These statistics will help us understand the typical values and variability of these variables, providing valuable insights into the airline passenger satisfaction dataset.
By following the steps outlined below, we can efficiently generate descriptive statistics for the numerical variables and gain a deeper understanding of the dataset's characteristics.

Figure 5. Descriptives



Descriptives

V1
id
Inflightwifiservice
DepartureArriva...
EaseofOnlinebo...
Gatelocation
Foodanddrink
Onlineboarding
Seatcomfort

Variable(s):
Age
FlightDistance
DepartureDelayin...
ArrivalDelayinMinut...

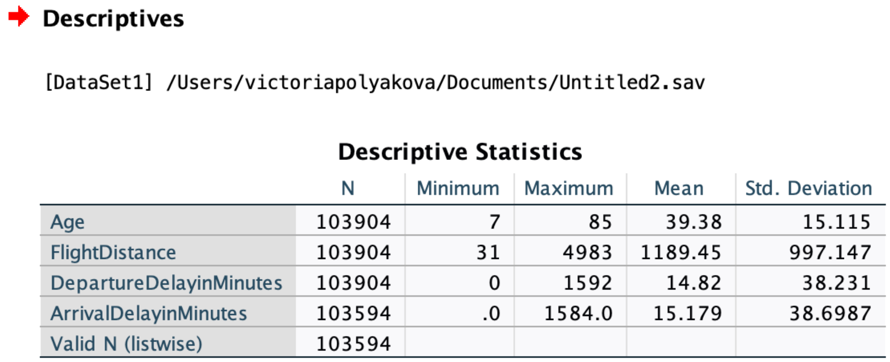
□ Save standardized values as variables

Interpret the results:

- The output will include statistics such as mean, median, standard deviation, minimum, and maximum for each selected variable.

- We will interpret these statistics to understand the central tendency, variability, and range of values for each numerical variable.

Figure 6. Descriptive statistics results



Descriptives

```
[DataSet1] /Users/victoriapolyakova/Documents/Untitled2.sav
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	103904	7	85	39.38	15.115
FlightDistance	103904	31	4983	1189.45	997.147
DepartureDelayinMinutes	103904	0	1592	14.82	38.231
ArrivalDelayinMinutes	103594	.0	1584.0	15.179	38.6987
Valid N (listwise)	103594				

The average age of passengers in the dataset is 39.38 years. This gives us a central value around which ages tend to cluster. With a standard deviation of 15.115, we observe that ages vary from the mean by approximately 15.115 years on average. This indicates a moderate level of dispersion in age distribution. The youngest passenger is 7 years old, while the oldest is 85 years old. This range provides insight into the diversity of ages among passengers. Since we have an even number of observations (103904), the median age will be the average of the 51952nd and 51953rd observation when the data is sorted in ascending order. The median age provides a measure of central tendency that is robust to extreme values. It represents the age at which half of the passengers are older and half are younger. This value gives us insight into the typical age profile of passengers in the dataset.

Passengers in the dataset travel an average distance of 1189.45 miles per flight. This gives us a sense of the typical distance covered by passengers. With a high standard deviation of 997.147, we see significant variability in flight distances. Some passengers may take short flights, while others embark on much longer journeys. The shortest flight distance recorded is 31 miles, while the longest is 4983 miles. This wide range highlights the diversity in travel distances among passengers. The median flight distance would be approximately the value at the 51952nd observation when the data is sorted in ascending order. It indicates the distance traveled by the typical passenger in the dataset. Understanding this value helps us gauge the average travel distance for passengers.

Passengers experience an average departure delay of 14.82 minutes. This indicates that, on average, flights depart slightly behind schedule. The relatively high standard deviation of 38.231 suggests considerable variability in departure delays. While the mean delay is around 14.82 minutes, individual delays can vary significantly. Departure delays range from 0 minutes to 1592 minutes (approx. 26.5 hours). This broad range underscores the potential for significant delays in departure times. The median departure of delay will be the average of the 51952nd and 51953rd observations when the data is sorted in ascending order. It gives us a sense of the central tendency of departure delays.

Passengers experience an average arrival delay of 15.179 minutes. This suggests that, on average, flights arrive slightly later than scheduled. Similar to departure delays, the standard deviation of 38.6987 indicates considerable variability in arrival delays. Some flights may arrive earlier than scheduled, while others experience more significant delays. Arrival delays range from 0 minutes to 1584 minute (approx. 26.4 hours). This wide range highlights the potential for substantial variations in arrival times among flights. The median arrival delay will be the average of the 51797th and 51798th observations when the data is sorted in ascending order. Similarly, the median arrival delay signifies the middle value of arrival delays in the dataset. It offers insight into the typical delay experienced by passengers upon arrival at their destination. Like the median departure delay, this measure is less sensitive to extreme values and provides a representative value for the dataset.

After calculating the medians for each numerical variable, here are the results:

Median Age = 39 years
Median Flight Distance = 862 miles
Median Departure Delay = 0 minutes
Median Arrival Delay = 0 minutes

Frequency tables are essential tools for summarizing and understanding categorical data. They provide a clear overview of the distribution of categories within a dataset, showing the count and percentage of observations in each category. By generating frequency tables, researchers can gain insights into the composition of their data and identify any patterns or trends present.

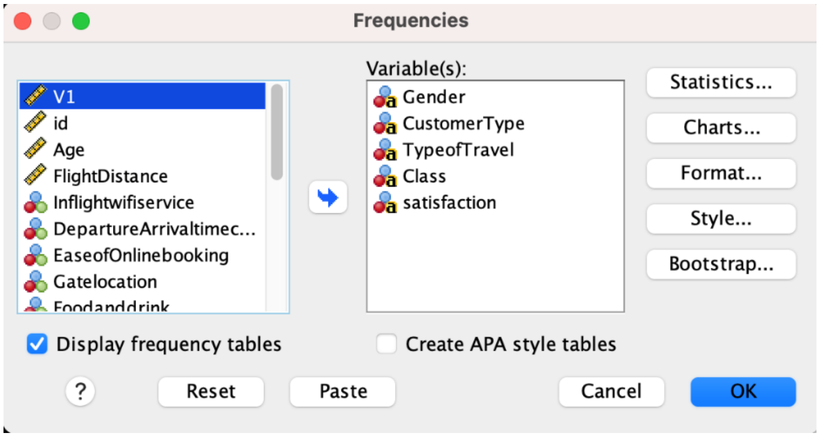
To create frequency tables in IBM SPSS Statistics, we will utilize the Frequencies procedure. This procedure allows us to specify the categorical variables we want to analyze and generates comprehensive tables summarizing the frequency of each category within those variables.

By following the steps outlined below, we can effectively create frequency tables for categorical variables such as Gender, Customer Type, Type of Travel, Class, and Satisfaction. These tables will serve as valuable tools for exploratory data analysis and provide a solid foundation for further statistical analysis and interpretation.

We have chosen gender, customer type, type of travel, class, and satisfaction for making descriptive statistics using frequencies because these variables are categorical and provide essential insights into passenger demographics, preferences, and satisfaction levels. Analyzing these variables allows us to understand the composition of the passenger dataset and identify any patterns or trends related to customer characteristics and satisfaction. By examining the frequencies of different categories within each variable, we can gain valuable insights into the distribution of passengers across various demographic and service-related factors, enabling us to make informed decisions and recommendations to enhance customer experience and satisfaction in the airline industry.

1. Import dataset:
    - We have chosen the data from kaggle.com website called “Airline Passenger Satisfaction”. Import the dataset by going to File > Open > Data and selecting the dataset.
2. Define the analysis:
    - Go to Analyze > Descriptive Statistics > Frequencies.
3. Select the categorical variables:
    - Gender, Customer Type, Type of Travel, Class, and Satisfaction.
4. Run the analysis:
    - Click the “OK” button to run the frequencies.

Figure 7. Frequencies



V1
id
Age
FlightDistance
Inflightwifiservice
DepartureArrivaltimec...
EaseofOnlinebooking
Gatelocation
Foodanddrink

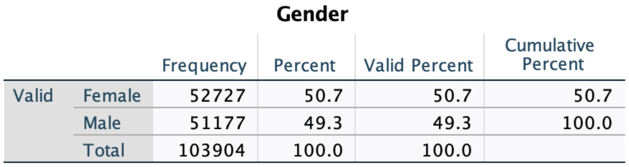
Variable(s):
Gender
CustomerType
TypeofTravel
Class
satisfaction

☑ Display frequency tables

☐ Create APA style tables

5. Interpret the results:
   - We will identify any notable patterns or discrepancies in the distribution of categories. For example, we will observe if certain categories are overrepresented or underrepresented compared to others.
   - Assessing the relative frequencies of categories within each variable will help in understanding the proportion of observations belonging to different groups.

Figure 8. Frequencies results (1)

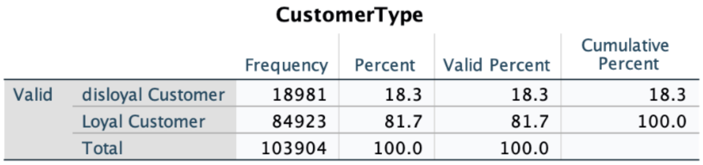


Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	52727	50.7	50.7	50.7
	Male	51177	49.3	49.3	100.0
	Total	103904	100.0	100.0	

There are 52727 occurrences of the “Female” gender category in the dataset. The percentage of female passengers in the dataset is 50.7%. This indicates that roughly half of the passengers are female. The valid percent is the same as the percent value, indicating that all observations are accounted for in the analysis. The cumulative percentage indicates the cumulative proportion of observations up to the current category. In this case, the cumulative percent of 50.7% means that up to this point, 50.7% of the observations are female. That means that the percentage of male is 49.3%.

Figure 9. Frequencies results (2)

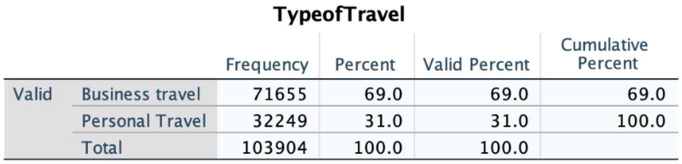


CustomerType

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	disloyal Customer	18981	18.3	18.3	18.3
	Loyal Customer	84923	81.7	81.7	100.0
	Total	103904	100.0	100.0	

There are 18981 instances of “Disloyal Customer” in the dataset. The percentage of disloyal customers in the dataset is 18.3%. This suggests that roughly 18.3% of customers are categorized as disloyal. As for the loyal customer, the percent is equal to 81.7%.

Figure 10. Frequencies results (3)

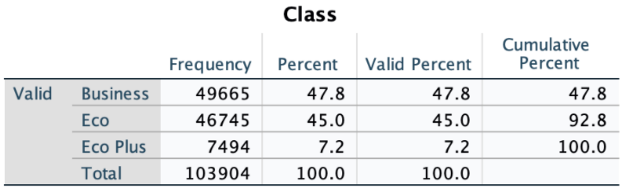


TypeofTravel

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Business travel	71655	69.0	69.0	69.0
	Personal Travel	32249	31.0	31.0	100.0
	Total	103904	100.0	100.0	

There are 71655 instances of "Business Travel" in the dataset and 32249 instances of "Personal Travel". The percentage of business travel occurrences in the dataset is 69%. This indicates that approximately 69% of the passengers are categorized as business travelers. The percentage of personal travel occurrences in the dataset is 31%.

Figure 11. Frequencies results (4)

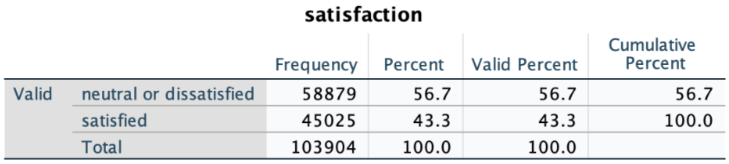


Class

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Business	49665	47.8	47.8	47.8
	Eco	46745	45.0	45.0	92.8
	Eco Plus	7494	7.2	7.2	100.0
	Total	103904	100.0	100.0	

There are 49665 instances of passengers categorized as business class, 46745 instances of economy class, and 7494 of economy premium. The percentage of passengers travelling in business class is 47.8%, in economy class is 45%, and in economy premium is 7.2%.

Figure 12. Frequencies results (5)



satisfaction

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	neutral or dissatisfied	58879	56.7	56.7	56.7
	satisfied	45025	43.3	43.3	100.0
	Total	103904	100.0	100.0	

There are 58879 instances of passengers categorized as neutral or dissatisfied and 45025 instances as satisfied. The percentage of passengers categorized as neutral or dissatisfied is 56.7% and as satisfied is 43.3%.

When selecting variables for visualization out of the provided list of airline service attributes, it's essential to prioritize those that are most likely to influence overall passenger satisfaction and reflect key aspects of the travel experience.

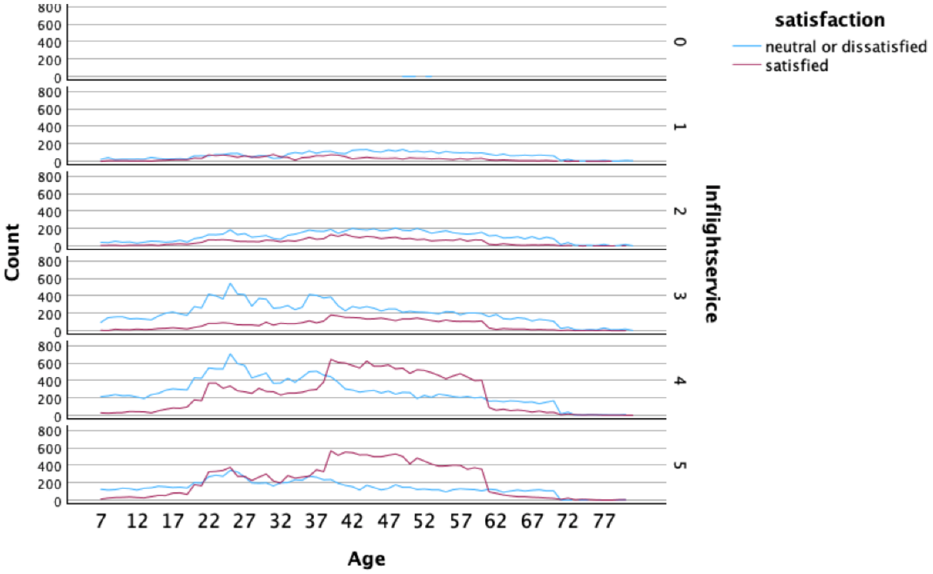
Inflight Service: This variable encompasses various aspects of the onboard experience, including the quality of service provided by flight attendants, responsiveness to passenger needs, and overall hospitality. Inflight service can significantly impact passenger perceptions of the airline and their overall satisfaction with the flight.

Departure and Arrival Convenience: Factors such as the efficiency of the boarding process, punctuality of departure and arrival times, and ease of navigating airport facilities contribute to passenger convenience and satisfaction. Monitoring satisfaction levels related to departure and arrival convenience can help identify areas for improvement in operational efficiency and customer service.

Seat Comfort: Comfortable seating arrangements, including seat size, legroom, cushioning, and reclining options, are crucial factors in passenger comfort and overall satisfaction. Passengers appreciate spacious and ergonomic seating that allows them to relax and enjoy their flight experience.

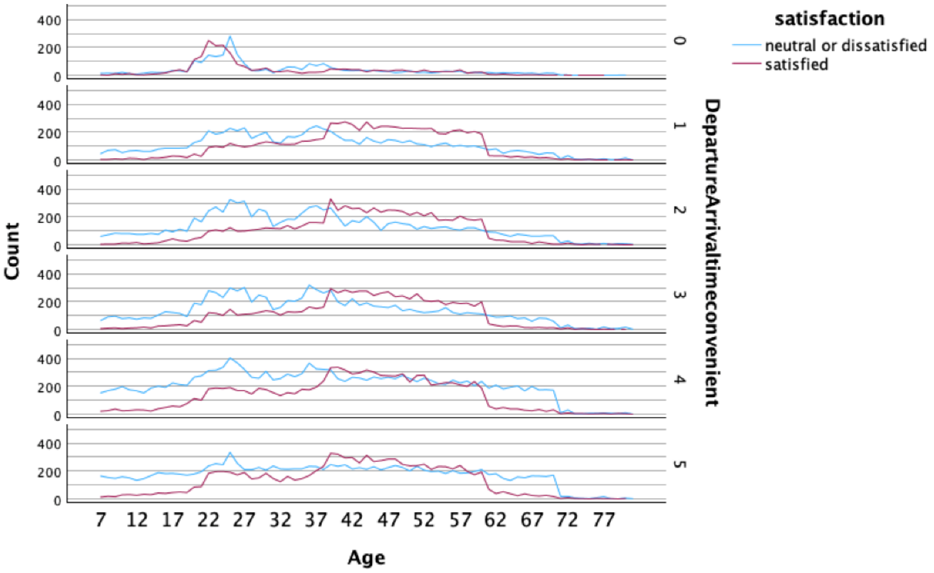
Cleanliness: The cleanliness and hygiene of aircraft cabins, lavatories, and shared spaces are essential for ensuring passenger well-being and satisfaction. Passengers expect clean and well-maintained facilities throughout the flight, which can contribute to a positive overall impression of the airline.

Figure 13. Inflight service satisfaction based on age



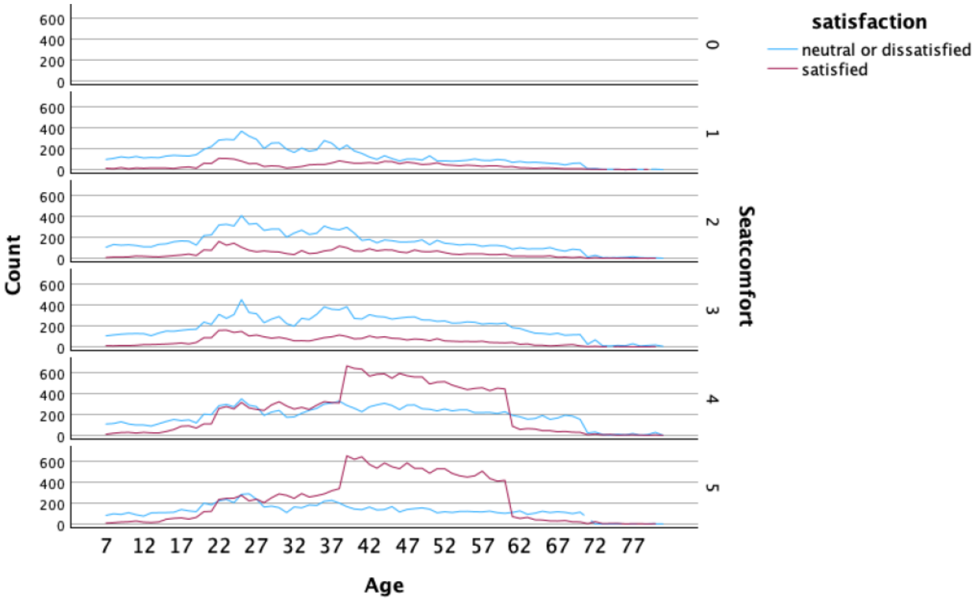
The line chart illustrates the satisfaction level with inflight service across different age groups, revealing a notable trend: elder passengers tend to express higher satisfaction levels compared to younger age groups. This insight underscores the importance of tailoring inflight service offerings to meet the preferences and expectations of diverse age demographics, ultimately enhancing overall passenger satisfaction and loyalty.

Figure 14. Departure arrival time convenient satisfaction based on age



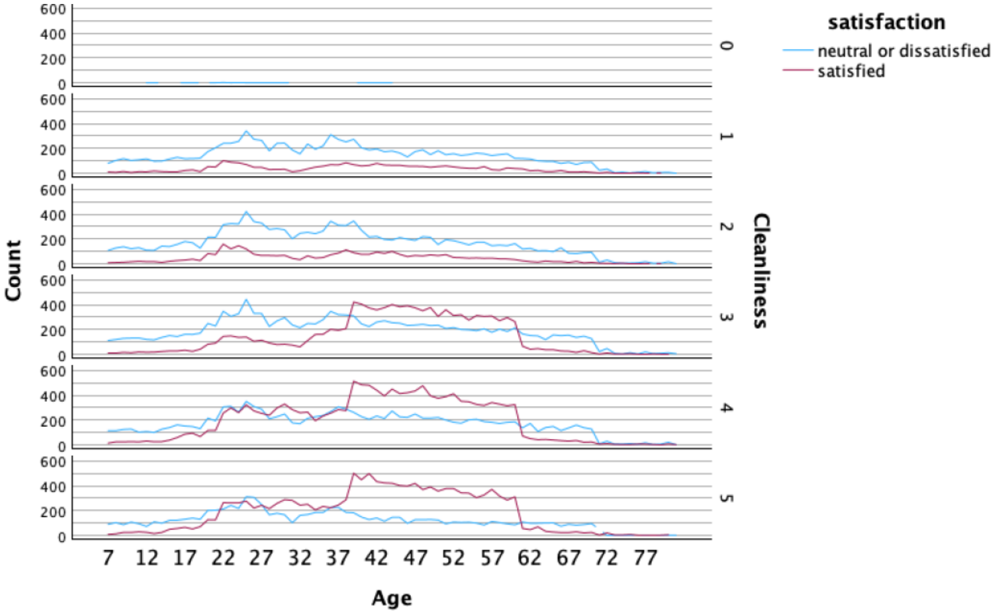
The line chart depicting satisfaction with departure and arrival time convenience across various age groups reveals a consistent pattern: satisfaction levels remain relatively stable across different age demographics. This suggests that age may not significantly influence perceptions of departure and arrival time convenience, highlighting the need for airlines to focus on other factors to enhance overall passenger satisfaction in this aspect.

Figure 15. Seat comfort satisfaction based on age



The line chart illustrating satisfaction with seat comfort across different age groups indicates a notable trend: older passengers tend to report higher levels of satisfaction with seat comfort compared to their younger counterparts. This finding underscores the importance of considering age-related preferences and comfort needs when designing seating arrangements and amenities to improve overall passenger satisfaction, particularly for older travelers.

Figure 16. Cleanliness satisfaction based on age

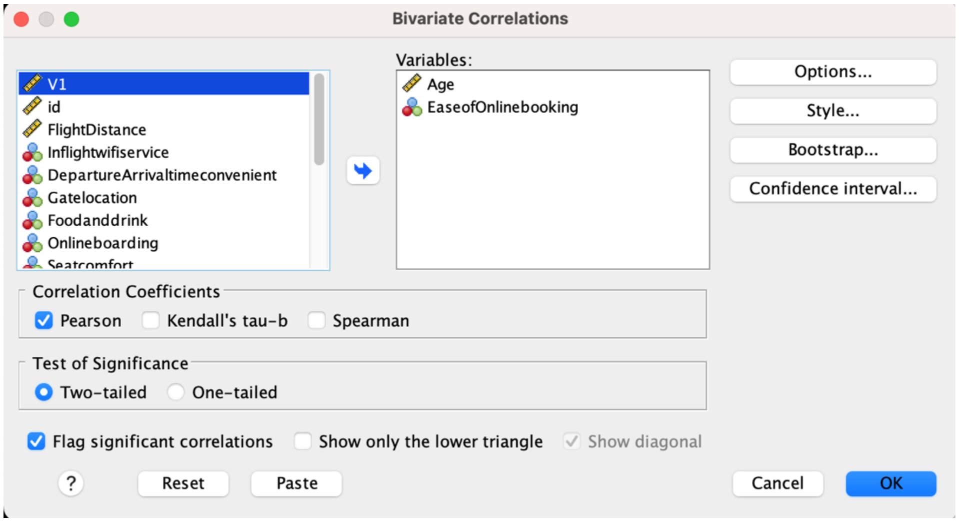


The line chart depicting satisfaction with cleanliness across various age groups reveals a discernible pattern: older individuals express higher levels of satisfaction with cleanliness compared to younger age groups. This insight suggests that maintaining cleanliness standards in airline cabins may contribute significantly to enhancing the travel experience, particularly for older passengers who may prioritize cleanliness and hygiene.

Multicollinearity in regression analysis occurs when independent variables are highly correlated, leading to unreliable coefficient estimates, challenging interpretation, and inflated standard errors. This condition can mislead significance tests, inefficiently utilize data, and render models sensitive to outliers. Ultimately, multicollinearity undermines the reliability and predictive accuracy of regression analyses. To assess multicollinearity in IBM SPSS Statistics, there are following steps:

1. Import the Dataset.
    - We have chosen the data from kaggle.com website called "Airline Passenger Satisfaction". Import the dataset by going to File > Open > Data and selecting the dataset.
2. Define the Analysis:
    - Go to Analyze > Correlate > Bivariate
3. Choose Variables:
    - In the "Bivariate Correlations" dialog box, select independent variables such as Age and Ease of Online Booking or Age and Flight Distance.
    - Move these variables into the "Variables" box on the right by clicking the arrow button.
4. Run Analysis:
    - Click the "OK" button to run the correlation analysis.

Figure 17. Correlations



Variables:
Age
EaseofOnlinebooking

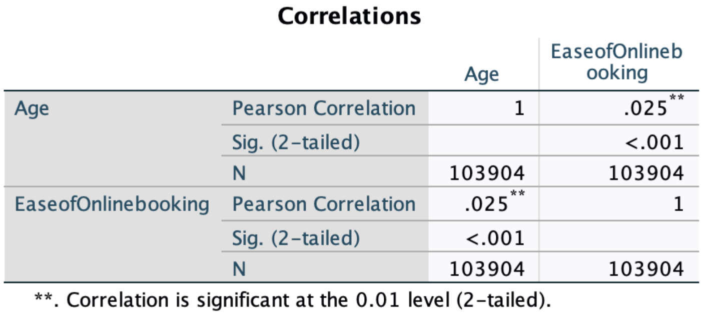
Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations ☐ Show only the lower triangle ☑ Show diagonal

5. Review the Output:
    - Examine the correlation coefficients between each pair of independent variables. High correlations (close to +/- 1) indicate potential multicollinearity issues.

Figure 18. Correlations results (1)



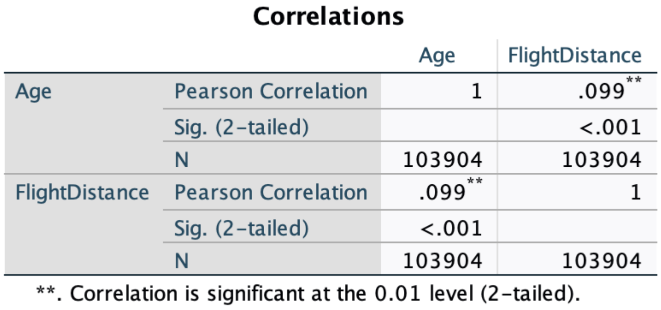
Correlations

		Age	EaseofOnlinebooking
Age	Pearson Correlation	1	.025**
	Sig. (2-tailed)		<.001
	N	103904	103904
EaseofOnlinebooking	Pearson Correlation	.025**	1
	Sig. (2-tailed)	<.001	
	N	103904	103904

**. Correlation is significant at the 0.01 level (2-tailed).

A Pearson correlation coefficient of 0.025 between age and ease of online booking indicates a very weak positive linear relationship between these two variables. The correlation coefficient is close to zero, suggesting that there is almost no linear relationship between age and ease of online booking. The positive sign indicates that as age increases, there is a slight tendency for ease of online booking to also increase, but the relationship is very weak. Therefore, changes in one variable are not likely to have a substantial impact on the other variable.

Figure 19. Correlations results (2)



Correlations

		Age	FlightDistance
Age	Pearson Correlation	1	.099**
	Sig. (2-tailed)		<.001
	N	103904	103904
FlightDistance	Pearson Correlation	.099**	1
	Sig. (2-tailed)	<.001	
	N	103904	103904

**. Correlation is significant at the 0.01 level (2-tailed).

A Pearson correlation coefficient of 0.099 between age and flight distance suggests a very weak positive linear relationship between these two variables such as age and ease of online booking correlation. In practical terms, this correlation suggests a minimal association between age and flight distance. If correlation coefficients are generally low, as in these cases, multicollinearity is less likely to be an issue.

3.2. Classification

The dataset on airline passenger satisfaction contains both categorical and numerical variables, making in suitable for various classification methods. Depending on the specific research question and characteristics of the dataset, we can choose the most appropriate classification method and use IBM SPSS Statistics to perform the analysis.
In the contemporary airline industry, understanding passenger satisfaction is paramount for airlines striving to enhance customer experience and foster loyalty. Through the utilization of binary logistic regression, the research aims to identify significant predictors of passenger satisfaction and develop a predictive model to anticipate satisfaction levels. Through this research, insights gained can inform strategic decision-making processes within the airline industry and contribute to the enhancement of customer satisfaction initiatives.

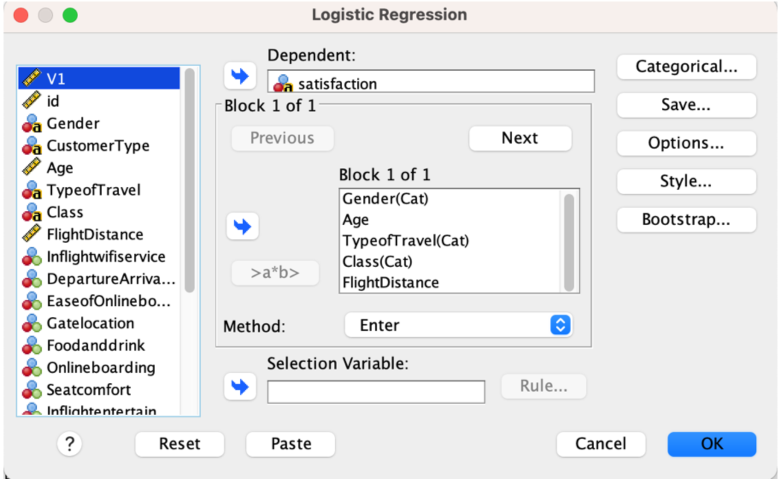
Gender, age, type of travel, class, and flight distance were selected for binary logistic regression analysis due to their theoretical relevance and practical importance in the airline industry. These variables are commonly considered as key determinants of passenger satisfaction, reflecting both demographic characteristics and flight-related factors that could influence overall experience.

Gender is often associated with distinct preferences and behaviors, while age reflects varying expectations and needs across different age groups. Type of travel (business or personal) and class (economy, business, or premium) are directly related to the quality of service and amenities provided during the flight, thereby impacting passenger satisfaction levels. Flight distance is also significant, as longer flights may entail different experiences and challenges compared to shorter ones.

To perform logistic regression analysis for the dataset on airline passenger satisfaction using IBM SPSS Statistics, there are several steps to follow:

1. Import the Dataset:
    - We have chosen the data from kaggle.com website called “Airline Passenger Satisfaction”. Import the dataset by going to File > Open > Data and selecting the dataset.
2. Define the Analysis:
    - Go to Analyze > Regression > Binary Logistic.
3. Set the Dependent Variable:
    - In the Binary Logistic dialog box, select the dependent variable “Satisfaction”.
4. Select Independent Variables:
    - This may include variables like Gender, Age, Type of Travel, Class, Flight Distance, and satisfaction ratings for various services.
5. Run the analysis:
    - Click the “OK” button to run the logistic regression analysis.

Figure 20. Binary Logistic Regression



Logistic Regression

Dependent: satisfaction

Block 1 of 1

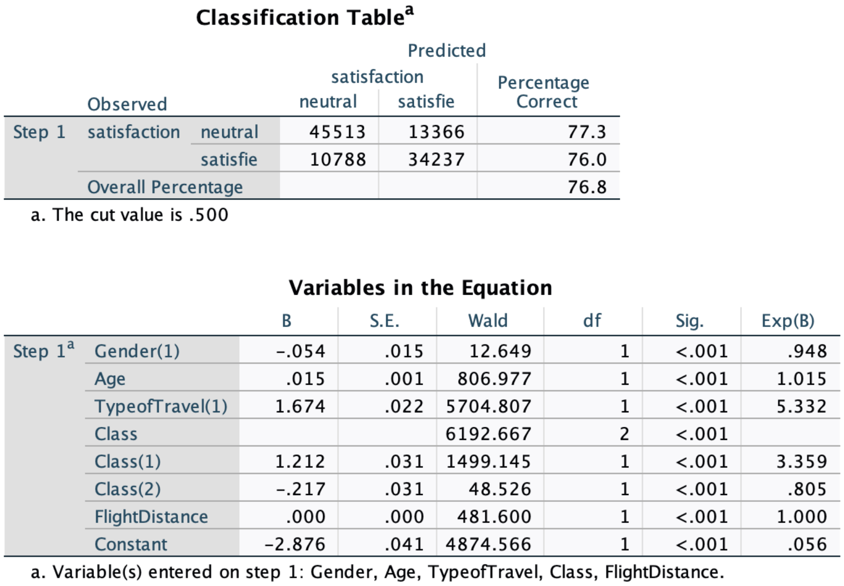
Block 1 of 1
Gender(Cat)
Age
TypeofTravel(Cat)
Class(Cat)
FlightDistance

Method: Enter

Selection Variable:

6. Interpret the results:
    - After the analysis is completed, review the output in the SPSS Viewer window.
    - Interpret the results, paying attention to coefficients, odds ratios, significance levels, and model fit statistics.
    - Determine which independent variables are significantly associated with passenger satisfaction and the direction of their effects.

Figure 21. Binary Logistic Regression Results (1)



Classification Table[a]

Observed			Predicted		
			satisfaction		Percentage Correct
			neutral	satisfie	
Step 1	satisfaction	neutral	45513	13366	77.3
		satisfie	10788	34237	76.0
	Overall Percentage				76.8

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gender(1)	-.054	.015	12.649	1	<.001	.948
	Age	.015	.001	806.977	1	<.001	1.015
	TypeofTravel(1)	1.674	.022	5704.807	1	<.001	5.332
	Class			6192.667	2	<.001	
	Class(1)	1.212	.031	1499.145	1	<.001	3.359
	Class(2)	-.217	.031	48.526	1	<.001	.805
	FlightDistance	.000	.000	481.600	1	<.001	1.000
	Constant	-2.876	.041	4874.566	1	<.001	.056

a. Variable(s) entered on step 1: Gender, Age, TypeofTravel, Class, FlightDistance.

In IBM SPSS Statistics, the “Variables in the Equation” table typically provides information about the coefficients associated with each independent variable in the logistic regression model. Look for the column labeled “B”. The value represents the coefficient of that variable in the logistic regression model. Positive coefficients indicate that an increase in the independent variable is associated with higher odds of the outcome (e.g., passenger satisfaction), while negative coefficients indicate the opposite.

Gender coefficient is equal to -0.054 which means that being male is associated with lower odds of passenger satisfaction. In the other words, male passengers are less likely to be satisfied compared to female passengers. The standard error of 0.015 indicates relatively low variability in the estimate of the coefficient for gender.

Age coefficient is 0.015. Since the coefficient is positive, it suggests that older passengers are more likely to be satisfied than younger passengers. A standard error of 0.001 suggests very low variability in the estimate of the coefficient for age, indicating a highly precise estimate.

Type of travel coefficient is 1.674. Positive coefficient suggests that passengers traveling for business are more likely to be satisfied than those traveling for personal reasons. A standard error of 0.022 indicates moderate variability in the estimate, suggesting relatively lower precision compared to the coefficients for gender and age.

Class coefficients:

- Class (Business): not provided
- Class 1 (Eco): 1.212
- Class 2 (Eco Plus): -0.217

Since the coefficient for Class 1 is positive and for Class 2 is negative, it suggests that passengers in Eco are more likely to be satisfied than those in Business, and in Eco Plus less than in Business. Standard errors of 0.031 suggest moderate variability in the estimates, indicating relatively lower precision compared to the coefficients for gender and age.

Flight distance coefficient is 0.000. Since the coefficient is close to zero, it suggests that flight distance does influence passenger satisfaction in this model. A standard error of 0.000 suggests very low variability in the estimate of the coefficient for flight distance, indicating a highly precise estimate.

Alongside the coefficients and standard errors, the table also provides information significance levels. A significance level (p-value) of <0.001 means that the coefficient for each independent variable in the logistic regression model is statistically significant at a very high level of confidence. There is less than a

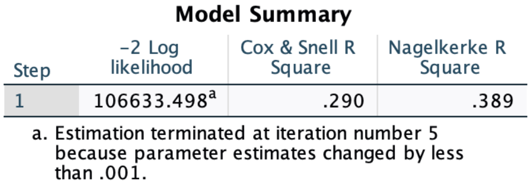
0.1% chance of results to occur by random variation, suggesting a strong relationship between the independent variables and passenger satisfaction.

The Wald chi-square statistics associated with each independent variable in logic regression analysis provide information about the significance of the variable in predicting the outcome (satisfaction):

- Gender Wald Chi-Square: 12.649. The Wald chi-square statistic tests the null hypothesis that the coefficient for gender is equal to zero (i.e., gender has no effect on passenger satisfaction). A higher chi-square value suggests that the variable has a significant effect on the outcome. In this case, the Wald chi-square statistic of 12.649 indicates that gender is statistically significant in predicting passenger satisfaction.
- Age Wald Chi-Square: 806.977. Similarly, with a high value of 806.977, age is highly significant in predicting passenger satisfaction.
- Type of Travel Wald Chi-Square: 5704.807. A very high chi-square value of 5704.807 indicates that type of travel is highly significant in predicting passenger satisfaction.
- Class Wald Chi-Square: 6192.667. It is highly significant in prediction satisfaction the same as type of travel.
- Class 1 Wald Chi-Square is 1499.145 and class 2 is equal to 48.526, which indicating that they are significant predictors of passenger satisfaction.
- Flight Distance Wald Chi-Square: 481.600. With a relatively high value, flight distance is statistically significant in predicting passenger satisfaction.

The classification table provides insights into the accuracy of the logic regression model in predicting passenger satisfaction levels. The results based on the observed and predicted satisfaction categories. Out of all passengers who reported being neutral in observed satisfaction, the model correctly predicted 77.3% of them as neutral. Additionally, 22.7% of passengers who were neutral in satisfaction were incorrectly predicted as satisfied. Among passengers who reported being satisfied, the model correctly predicted 76.0% of them as satisfied. However, 24.0% of passengers who were satisfied were incorrectly predicted as neutral (Classification Table).

Figure 22. Binary Logistic Regression Results (2)



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	106633.498[a]	.290	.389

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The “Model Summary” table provides important information about the overall fit and explanatory power of the logistic regression model.

1. -2 Log likelihood value: 106633.498[a]
   The -2 Log likelihood is a measure of how well the logistic regression model fits the data. A lower value indicates a better fit. In this case, it suggests that the model provides a reasonable fit to the data.
2. Nagelkerke R Square value: 0.389
   A value closer to 1 indicates a better fit. In this case, the value suggests that the model explains approximately 39% of the variance in passenger satisfaction.

3.3. Clustering

For the airline passenger satisfaction dataset, several clustering methods can be applied using IBM SPSS Statistics. The aim of analyzing clusters is to gain valuable insights into passenger preferences and behavior, which can inform strategic decision-making and drive improvements across various aspects of airline operations and customer experience.

When the number of clusters is unknown or when dealing with large datasets, it is useful to use TwoStep Cluster Analysis. This cluster analysis is a scalable clustering method that can handle large datasets efficiently. It automatically determines the optimal number of clusters and uses a combination of hierarchical and K-Means clustering algorithms. By applying these clustering methods in IBM SPSS Statistics, we can uncover hidden patterns and groupings in the airline passenger satisfaction dataset, which can provide valuable insights for decision-making and targeted interventions.

This analysis demonstrates the segmentation of airline passengers into distinct clusters based on their ratings of various airline services. By grouping passengers with similar satisfaction levels and preferences, airlines can better understand the diverse needs of their customer base. The analysis reveals which services are most positively or negatively perceived by different passenger segments, allowing airlines to prioritize improvements and tailor their offerings accordingly. Ultimately, this segmentation facilitates targeted strategies to enhance customer satisfaction, loyalty, and overall airline performance.

Overall, two-step cluster analysis is a useful exploratory technique for uncovering hidden structures or patterns in datasets and for identifying meaningful groupings of cases based on their similarities. It is widely used in various fields such as market research, customer segmentation, and social sciences.

In the clustering analysis, several variables were utilized to segment airline passengers into distinct groups based on their satisfaction levels and preferences. These variables include:

- Baggage Handling: These variable measures passengers' satisfaction with the airline's baggage handling procedures, including check-in, luggage drop-off, and retrieval.
- Check-in Service: This variable assesses passengers' satisfaction with the efficiency and effectiveness of the airline's check-in process, both online and at the airport.
- Cleanliness: This variable reflects passengers' perceptions of the cleanliness and hygiene standards maintained throughout their journey, including aircraft cabins, restrooms, and airport facilities.
- Departure and Arrival Time Convenience: This variable evaluates passengers' satisfaction with the timeliness and convenience of departure and arrival schedules, including punctuality and adherence to announced schedules.
- Ease of Online Booking: These variable measures passengers' satisfaction with the airline's online booking system, including website usability, reservation process, and clarity of booking information.
- Food and Drink Quality: This variable assesses passengers' satisfaction with the quality, variety, and presentation of food and beverages offered during the flight.
- Gate Location: This variable reflects passengers' satisfaction with the accessibility and convenience of gate locations within airport terminals, including proximity to amenities and ease of navigation.
- Inflight Entertainment: This variable evaluates passengers' satisfaction with the range and quality of entertainment options available during the flight, including movies, TV shows, music, and games.
- Inflight Service: These variable measures passengers' satisfaction with the overall quality of service provided by flight attendants, including responsiveness, attentiveness, and professionalism.
- Inflight Wi-Fi Service: This variable assesses passengers' satisfaction with the availability and reliability of in-flight Wi-Fi connectivity, including internet speed and ease of access.

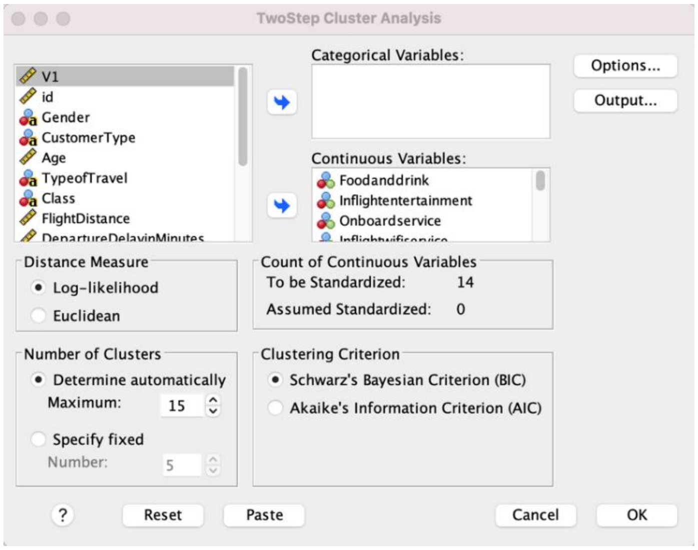
- Leg Room Service: This variable reflects passengers' satisfaction with the comfort and spaciousness of seating arrangements, including legroom and seat recline options.
- Onboard Service: This variable evaluates passengers' satisfaction with onboard amenities and services, including comfort items, complimentary offerings, and assistance from cabin crew.
- Online Boarding Process: These variable measures passengers' satisfaction with the efficiency and user-friendliness of the online boarding process, including mobile check-in, electronic boarding passes, and self-service kiosks.
- Seat Comfort: This variable assesses passengers' satisfaction with the comfort and ergonomics of seating arrangements, including seat size, cushioning, and adjustability.

These variables collectively provide a comprehensive understanding of passengers' experiences and preferences throughout their journey, enabling airlines to identify areas of improvement and tailor their services to better meet customer expectations.

There are following steps for clustering implementation:

1. Import the Dataset.
    - Import the dataset by going to File > Open > Data and selecting the dataset. We have chosen the data from kaggle.com website called “Airline Passenger Satisfaction”.
2. Define the Analysis:
    - Go to Analyze > Classify > TwoStep Cluster
3. Choose Variables:
4. In the TwoStep Cluster Analysis dialog box, select the variables relevant to clustering passengers based on their satisfaction levels.
5. Run Analysis:
    - Click the “OK” button to run the TwoStep Cluster Analysis.

Figure 23. Cluster analysis



Categorical Variables:

Continuous Variables:
Foodanddrink
Inflightentertainment
Onboardservice
Inflightwifiservice

Distance Measure
● Log-likelihood
○ Euclidean

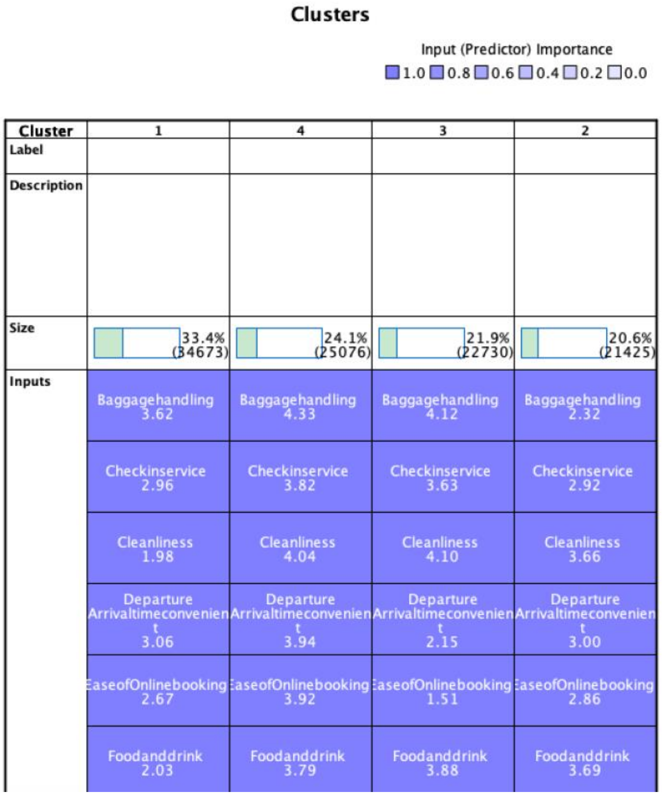
Count of Continuous Variables
To be Standardized: 14
Assumed Standardized: 0

Number of Clusters
● Determine automatically
Maximum: 15
○ Specify fixed
Number: 5

Clustering Criterion
● Schwarz's Bayesian Criterion (BIC)
○ Akaike's Information Criterion (AIC)

6. Interpret Results:
    - Interpret the results to understand how passengers are grouped into clusters based on their satisfaction levels and other relevant variables.

Figure 24. Clusters results

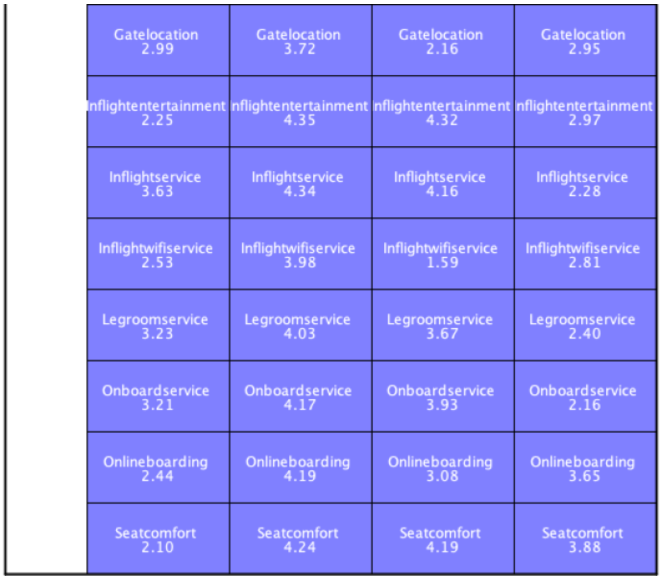


Clusters

Input (Predictor) Importance
■1.0 ■0.8 ■0.6 ■0.4 □0.2 □0.0

Cluster	1	4	3	2
Label				
Description				
Size	33.4% (34673)	24.1% (25076)	21.9% (22730)	20.6% (21425)
Inputs	Baggagehandling 3.62	Baggagehandling 4.33	Baggagehandling 4.12	Baggagehandling 2.32
	Checkinservice 2.96	Checkinservice 3.82	Checkinservice 3.63	Checkinservice 2.92
	Cleanliness 1.98	Cleanliness 4.04	Cleanliness 4.10	Cleanliness 3.66
	Departure Arrivaltimeconvenien t 3.06	Departure Arrivaltimeconvenien t 3.94	Departure Arrivaltimeconvenien t 2.15	Departure Arrivaltimeconvenien t 3.00
	EaseofOnlinebooking 2.67	EaseofOnlinebooking 3.92	EaseofOnlinebooking 1.51	EaseofOnlinebooking 2.86
	Foodanddrink 2.03	Foodanddrink 3.79	Foodanddrink 3.88	Foodanddrink 3.69

Figure 25. Clusters results (continuing)



	Gatelocation 2.99	Gatelocation 3.72	Gatelocation 2.16	Gatelocation 2.95
	Inflightentertainment 2.25	Inflightentertainment 4.35	Inflightentertainment 4.32	Inflightentertainment 2.97
	Inflightservice 3.63	Inflightservice 4.34	Inflightservice 4.16	Inflightservice 2.28
	Inflightwifiservice 2.53	Inflightwifiservice 3.98	Inflightwifiservice 1.59	Inflightwifiservice 2.81
	Legroomservice 3.23	Legroomservice 4.03	Legroomservice 3.67	Legroomservice 2.40
	Onboardservice 3.21	Onboardservice 4.17	Onboardservice 3.93	Onboardservice 2.16
	Onlineboarding 2.44	Onlineboarding 4.19	Onlineboarding 3.08	Onlineboarding 3.65
	Seatcomfort 2.10	Seatcomfort 4.24	Seatcomfort 4.19	Seatcomfort 3.88

The clusters derived from the analysis represent unique segments of airline passengers characterized by their ratings of different airline services. Cluster 1, comprising 33.4% of passengers, exhibits lower satisfaction across multiple service aspects, including baggage handling, cleanliness, and departure/arrival time convenience. Conversely, Cluster 4, representing 24.1% of passengers, stands out as the most satisfied group, demonstrating high ratings across all evaluated services, from check-in to inflight amenities.

Cluster 1 comprises passengers who express dissatisfaction with various service elements, indicating potential areas for improvement in airline services. These passengers may benefit from enhancements in baggage handling procedures, cleanliness maintenance, and optimization of departure and arrival experiences. Moreover, the cluster analysis may reveal specific demographics within Cluster 1, such as age groups, predominant travel class, and average flight distances, providing further insights into the characteristics of these dissatisfied passengers.

Conversely, Cluster 4 represents passengers who consistently rate airline services highly across various dimensions. These satisfied passengers may exhibit distinct characteristics, such as age distribution, preferred travel classes, and typical flight distances. Understanding the demographic and behavioral attributes of Cluster 4 passengers can help airlines identify and replicate successful service delivery strategies to maintain high levels of customer satisfaction.

By delving deeper into the characteristics of each cluster, airlines can tailor their service offerings more effectively to meet the specific needs and preferences of different passenger segments. This targeted approach enables airlines to enhance overall customer satisfaction and loyalty, ultimately driving business growth and competitiveness in the airline industry.

In Cluster 1, passengers tend to exhibit a diverse range of demographic characteristics and travel preferences. Age distribution within this cluster may vary, although it tends to include a mix of younger and older passengers. Predominant travel classes among Cluster 1 passengers may include economy and economy plus, suggesting a preference for more budget-friendly options. Flight distance for passengers in this cluster may vary widely, with some opting for shorter domestic flights while others may undertake longer international journeys. Overall, Cluster 1 represents a heterogeneous group of passengers with differing needs and expectations, highlighting the importance of tailored service offerings to address their diverse requirements.

In Cluster 4, passengers typically exhibit characteristics indicative of high satisfaction levels. The age distribution within this cluster may skew towards older demographics, suggesting a preference for more comfortable and premium travel experiences. Predominant travel classes among Cluster 4 passengers may include business and first class, indicating a preference for luxury and premium services. Flight distance for passengers in this cluster may vary, although they may be more inclined towards longer-haul flights, including international travel. Overall, Cluster 4 represents a segment of passengers who prioritize comfort, convenience, and premium amenities, underscoring the importance of delivering high-quality services to maintain their satisfaction and loyalty.

In Cluster 2, passengers may display characteristics indicating moderate satisfaction levels. The age distribution within this cluster may be relatively evenly distributed across different age groups, reflecting a diverse passenger profile. Predominant travel classes among Cluster 2 passengers may include a mix of economy and economy plus, suggesting a preference for affordability balanced with some added amenities. Flight distance for passengers in this cluster may vary, with a mix of short-haul and medium-haul flights. Overall, Cluster 2 represents a segment of passengers with moderate satisfaction levels, indicating opportunities for airlines to improve certain aspects of service delivery to enhance overall customer satisfaction.

In Cluster 3, passengers may exhibit characteristics indicative of slightly lower satisfaction levels compared to other clusters. The age distribution within this cluster may include a mix of age groups, although it may skew slightly towards younger demographics. Predominant travel classes among Cluster 3 passengers may include economy and economy plus, reflecting a preference for cost-effective travel options. Flight distance for passengers in this cluster may vary, with a mix of short-haul and medium-haul flights. Overall, Cluster 3 represents a segment of passengers with some dissatisfaction with certain aspects of the travel experience, highlighting areas where airlines can focus their efforts to address passenger concerns and improve overall satisfaction levels.

4. Results and Discussion

Big data, a term referring to vast and varied data sets, poses a significant opportunity for organizations to extract valuable insight. Big data analytics delves into this wealth of information, aiming to uncover hidden patterns, correlations, and trends that traditional data analytics methods might overlook. This analytical process encompasses several types, including descriptive, diagnostic, predictive, and prescriptive analytics.

Descriptive analytics provides a retrospective view of data, summarizing past events, while diagnostic analytics delves deeper into understanding the causes behind these events. Predictive analytics forecasts future outcomes based on historical data and statistical models, while prescriptive analytics offers recommendations to optimize future decisions.

Unlike traditional data analytics, BDA deals with large volumes of structured and unstructured data, requiring specialized tools and techniques. Numerous BDA tools and vendor products are available, from open-source platforms like Apache Hadoop to commercial solutions such as IBM SPSS and SAS.

Predictive analytics, a subset of BDA, relies on techniques like business intelligence, statistics, and data mining to forecast future outcomes. Business intelligence transforms raw data into actionable insights, while statistics provides the mathematical foundation for predictive models. Data mining extracts patterns and knowledge from large datasets, contributing to predictive analytics' effectiveness.

The benefits of BDA are manifold, ranging from improved decision-making and operational efficiency to better customer insights and innovation. However, challenges like data privacy, integration, scalability, and skill shortages must be addressed to fully realize these benefits.

IBM SPSS stands as a prominent tool for big data analytics, offering advanced statistical analysis, data mining, and predictive modeling capabilities. Its efficiency in handling large and complex datasets enables organizations to derive actionable insights effectively.

In various industries like manufacturing, energy production, retail, ICT, banking, and logistics, BDA presents promising prospects. It empowers organizations to optimize processes, enhance customer experiences, and drive innovation across diverse sectors.

The analysis conducted using IBM SPSS Statistics on the dataset regarding airline passenger satisfaction yielded valuable insights into factors influencing passenger satisfaction and clustering patterns among passengers. Here, we discuss the findings and implications of both the logistic regression and two-step cluster analysis.

The logistic regression analysis aimed to identify significant predictors of passenger satisfaction using variables such as gender, age, type of travel, class, and flight distance. The results revealed several significant predictors:

1. Gender: Male passengers were associated with lower odds of satisfaction compared to female passengers.
2. Age: Older passengers were more likely to be satisfied than younger passengers.
3. Type of Travel: Business travelers were more likely to be satisfied than those traveling for personal reasons.
4. Class: Passengers in Economy class were more likely to be satisfied than those in Economy Plus.
5. Flight Distance: Flight distance did not significantly influence passenger satisfaction.

Furthermore, all independent variables showed statistically significant associations with passenger satisfaction, as indicated by low p-values (<0.001). The logistic regression model exhibited a reasonable fit to the data, explaining approximately 39% of the variance in passenger satisfaction.

The two-step cluster analysis aimed to group passengers based on satisfaction levels and other relevant variables. The analysis automatically generated two distinct clusters with moderate cohesion and separation. The cluster sizes were slightly imbalanced, with one cluster being larger than the other. Overall, the findings suggest that passenger satisfaction is influenced by various factors, including demographic characteristics, travel preferences, and service quality. Understanding these factors can inform strategic decision-making processes within the airline industry, allowing airlines to tailor services and improve customer experiences effectively.

1. Service Customization: Airlines can tailor their services based on passenger demographics and preferences identified through logistic regression analysis. For example, they can offer personalized amenities and seating options to different passenger segments.
2. Operational Enhancements: Insights from clustering analysis can help airlines streamline operations and allocate resources more efficiently. By understanding distinct passenger groups, airlines can optimize service delivery and enhance overall satisfaction.
3. Marketing Strategies: Airlines can develop targeted marketing campaigns aimed at specific passenger segments identified through clustering analysis. This approach can improve marketing ROI and customer engagement.
4. Continuous Improvement: By regularly analyzing passenger satisfaction data and monitoring trends, airlines can identify areas for improvement and implement corrective measures promptly. This proactive approach fosters continuous improvement in service quality and customer satisfaction.

In conclusion, the combination of logistic regression and two-step cluster analysis offers a comprehensive understanding of passenger satisfaction and segmentation within the airline industry. By leveraging these analytical techniques, airlines can enhance customer experiences, foster loyalty, and maintain a competitive edge in the market.

5. Conclusion

In this thesis, we have explored the intricate dynamics of airline passenger satisfaction through a blend of theoretical discussions and practical analysis using IBM SPSS Statistics. By delving into the realms of big data analytics, predictive modeling, and clustering techniques, we aimed to uncover meaningful insights for the airline industry.

Theoretical insights into big data analytics provided a foundational understanding of the vast potential and challenges associated with analyzing large and diverse datasets. We discussed the various types of analytics, including descriptive, diagnostic, predictive, and prescriptive analytics, highlighting their significance in extracting actionable insights from complex data sources. Moreover, we elucidated the benefits and challenges of big data analytics, emphasizing its pivotal role in enhancing decision-making processes across industries.

Moving from theory to practice, our empirical analysis focused on the dataset concerning airline passenger satisfaction. Leveraging IBM SPSS Statistics, we employed logistic regression and two-step cluster analysis to dissect the intricate web of factors shaping passenger satisfaction and clustering patterns among passengers. Through logistic regression, we identified significant predictors of passenger satisfaction, ranging from demographic characteristics to travel preferences and service quality indicators. Subsequent clustering analysis unveiled distinct passenger segments based on satisfaction levels and other relevant variables.

Combining theoretical knowledge with empirical findings, our research contributes to a deeper understanding of passenger satisfaction dynamics within the airline industry. By identifying critical factors influencing satisfaction levels and segmenting passenger populations, airlines can tailor their services, optimize operations, and refine marketing strategies to better meet customer needs and preferences. The insights gleaned from our analysis pave the way for strategic decision-making processes aimed at enhancing customer experiences, fostering loyalty, and driving sustainable growth in the highly competitive airline market.

In conclusion, the thesis underscores the transformative potential of big data analytics in deciphering complex phenomena such as passenger satisfaction. By embracing advanced analytical techniques and leveraging robust tools like IBM SPSS Statistics, airlines can embark on a journey towards greater customer-centricity, operational efficiency, and strategic agility in an ever-evolving landscape. As we navigate the intricacies of the digital age, harnessing the power of big data analytics emerges as a cornerstone for unlocking new opportunities and driving meaningful change in the airline industry and beyond.

6. References

ABBOTT, D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.

BENSIGER, G. (2014, January 17). Amazon wants to ship your package before you buy it.

BOYD, D & CRAWFORD, K. (2012). Critical questions for big data – Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15(5),* 662–679.

CHEN, H., CHIANG, R. H. L., STOREY, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36(4),* 1165–1188.

CHEN, M., MAO, S., & LIU, Y. (2014a). Big data: a survey. *Mobile Networks and Applications, 19(2),* 171–209

CHEN, C. L. P., ZHANG, C-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences, 275(2014),* 314–347).

DEMCHENKO, Y., GROSSO, P., & MEMBREV, P. (2013). Addressing Big Data Issues in Scientific Data Infrastructure. *Collaboration Technologies and Systems (CTS), 2013 International Conference on (Pp. 48-55). IEEE.*, 48–55.

DUTTA, D. & BOSE, I. (2015). Managing a big data project: The case of Ramaco Cements Limited. *International Journal of Production Economics, 165, 293*–306.

FERGUSON, M. (2012). Architecting A Big Data Platform for Analytics.

FIELD, Andy P. *Discovering statistics using IBM SPSS statistics.* Thousand Oaks: SAGE Publications, 2013. ISBN 978-1-4462-4917-8.

FREHE, V., KLEINSCHMIDT, T., TEUTEBERG, F. (2014). Big Data in Logistics – Identifying Potentials through Literature, Case Study and Expert Interview Analyses.

GUO, Y., BARNES, S. J., & JIA, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467–483.

HAMAMI, F., DAHLAN, I A. (2020). The Implementation of Stream Architecture for Handling Big Data Velocity in Social Media.

HASHEM, I. A. T., YAQOOB, I., ANUAR, N. B., MOKHTAR, S., GANI, A., & ULLAH KHAN, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115.

HULL, G. (2015). Successful failure: what Foucault can teach us about privacy self- management in a world of Facebook and big data. *Ethics of Information Technology, 17(2),* 89–101.

HUMBERTY, M. (2015). Awaiting the second big data revolution: From digital noise to value creation. *Journal of Industry, Competition and Trade, 15(1),* 34–47.

JUKIĆ, N., SHARMA, A., NESTOROV, S., & JUKIĆ, B. (2015). Augmenting Data Warehouses with Big Data. *Information Systems Management, 0530*(July), 150527102705002.

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data?select=train.csv

KITCHIN, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848.

LI, X., SONG, J., & HUANG, B. (2015). A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics. *International Journal of Advanced Manufacturing Technolgy,* 1–13.

LU, T., GUO, X., XU, B., ZHAO, L., PENG, Y., & YANG, H. (2013) Next big thing in big data: the security of the ICT supply chain. Proceedings from ASE/IEEE: *International Conference on Social Computing.* Washington D.C.: IEEE.

McAFEE, A. & BRYNJOLFSSON, E. (2012, October) Big data: The management revolution. *Harvard Business Review*.

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C., & HUNG BYERS, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

MARR, B. (2017). Beyond the big data buzz: how data is disrupting business in every industry in the world.

MISHRA, D., GUNASEKARAN, A., PAPADOPOULOS, T., CHILDE, S. J., & OPER RES, A. (2016). Big Data and supply chain management: a review and bibliometric analysis. *Annals of Operations Research*.

OXFORD ECONOMICS. (2011). The new digital economy.

POUSTTCHI, K. & HUFFENBACH, Y. (2014). Engineering the value network of the customer interface and marketing in the data-rich retail environment. *International Journal of Electronic Commerce, 18(4),* 17–41.

RÜPING, S. (2015). Big data in medicine and healthcare. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 58(11),* 794–798.

SHARDA, R., DELEN, D., TURBAN, E. (2013). Business Intelligence: A Managerial Perspective on Analytics.

SIVARAJAH, U., KAMAL, M. M., IRANI, Z., & WEERAKKODY, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263-286.

SRIVASTAVA, U., GOPALKRISHNAN, S. (2015). Impact of Big Data Analytics on Baking Sector: Learning for Indian Banks.

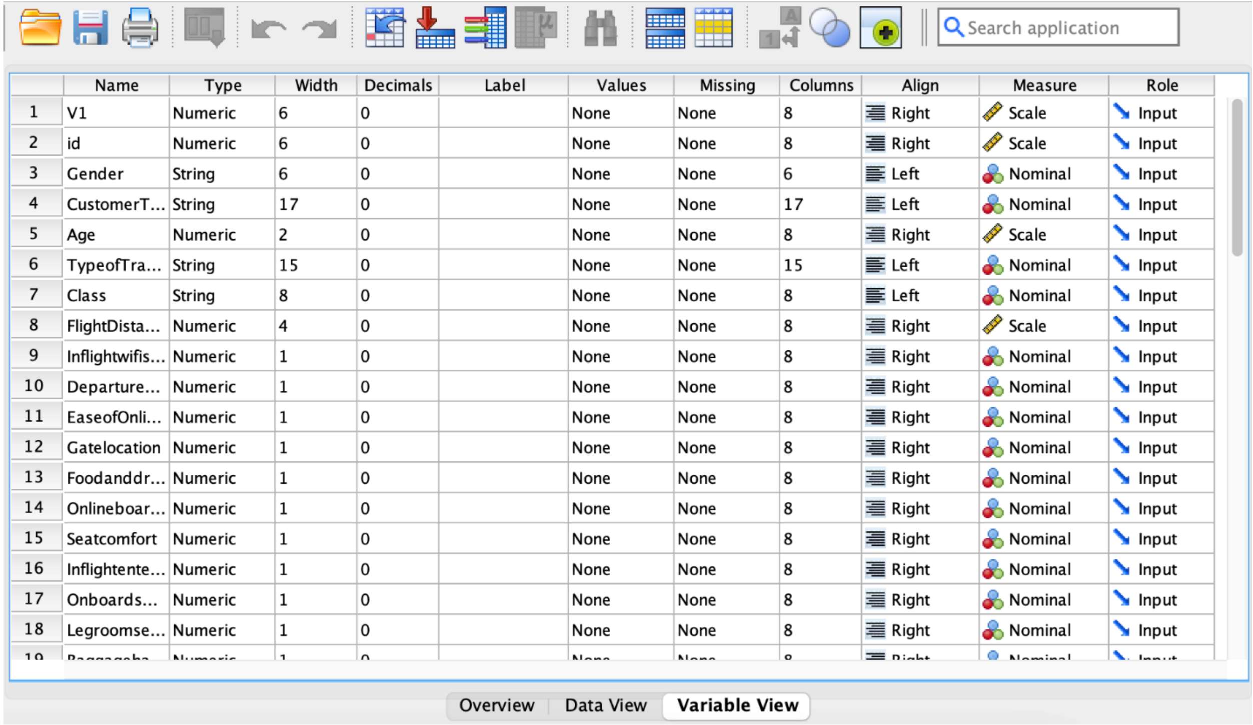
WATSON, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, *34*(1), 65.

WORLD ECONOMIC FORUM. (2014). The global Information technology report 2014 – Rewards and risks of big data.

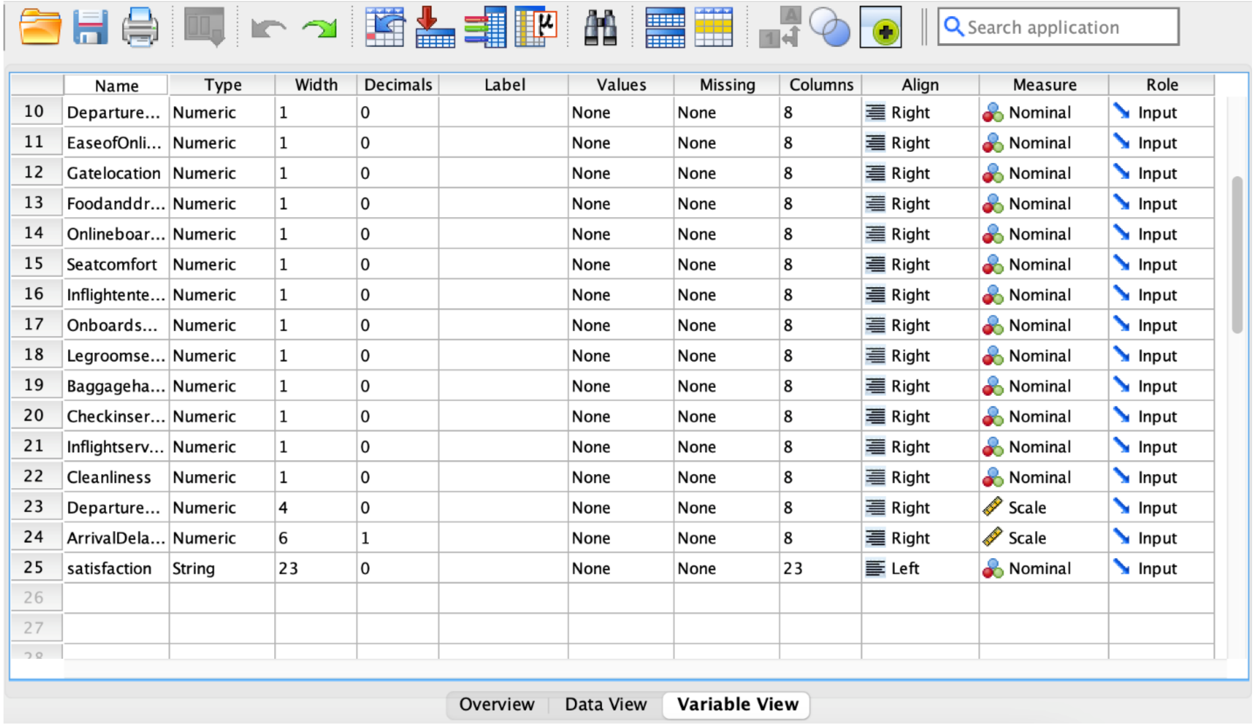
7. List of pictures, tables, and graphs

Appendix

Raw Data



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	V1	Numeric	6	0		None	None	8	Right	Scale	Input
2	id	Numeric	6	0		None	None	8	Right	Scale	Input
3	Gender	String	6	0		None	None	6	Left	Nominal	Input
4	CustomerT...	String	17	0		None	None	17	Left	Nominal	Input
5	Age	Numeric	2	0		None	None	8	Right	Scale	Input
6	TypeofTra...	String	15	0		None	None	15	Left	Nominal	Input
7	Class	String	8	0		None	None	8	Left	Nominal	Input
8	FlightDista...	Numeric	4	0		None	None	8	Right	Scale	Input
9	Inflightwifis...	Numeric	1	0		None	None	8	Right	Nominal	Input
10	Departure...	Numeric	1	0		None	None	8	Right	Nominal	Input
11	EaseofOnli...	Numeric	1	0		None	None	8	Right	Nominal	Input
12	Gatelocation	Numeric	1	0		None	None	8	Right	Nominal	Input
13	Foodanddr...	Numeric	1	0		None	None	8	Right	Nominal	Input
14	Onlineboar...	Numeric	1	0		None	None	8	Right	Nominal	Input
15	Seatcomfort	Numeric	1	0		None	None	8	Right	Nominal	Input
16	Inflightente...	Numeric	1	0		None	None	8	Right	Nominal	Input
17	Onboards...	Numeric	1	0		None	None	8	Right	Nominal	Input
18	Legroomse...	Numeric	1	0		None	None	8	Right	Nominal	Input
19	Baggageha...	Numeric	1	0		None	None	8	Right	Nominal	Input



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
10	Departure...	Numeric	1	0		None	None	8	Right	Nominal	Input
11	EaseofOnli...	Numeric	1	0		None	None	8	Right	Nominal	Input
12	Gatelocation	Numeric	1	0		None	None	8	Right	Nominal	Input
13	Foodanddr...	Numeric	1	0		None	None	8	Right	Nominal	Input
14	Onlineboar...	Numeric	1	0		None	None	8	Right	Nominal	Input
15	Seatcomfort	Numeric	1	0		None	None	8	Right	Nominal	Input
16	Inflightente...	Numeric	1	0		None	None	8	Right	Nominal	Input
17	Onboards...	Numeric	1	0		None	None	8	Right	Nominal	Input
18	Legroomse...	Numeric	1	0		None	None	8	Right	Nominal	Input
19	Baggageha...	Numeric	1	0		None	None	8	Right	Nominal	Input
20	Checkinser...	Numeric	1	0		None	None	8	Right	Nominal	Input
21	Inflightserv...	Numeric	1	0		None	None	8	Right	Nominal	Input
22	Cleanliness	Numeric	1	0		None	None	8	Right	Nominal	Input
23	Departure...	Numeric	4	0		None	None	8	Right	Scale	Input
24	ArrivalDela...	Numeric	6	1		None	None	8	Right	Scale	Input
25	satisfaction	String	23	0		None	None	23	Left	Nominal	Input
26											
27											
28											