Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informatiky a kvantitativních metod

Vybrané metody analýzy dat v managementu s SPSS

Korelační a regresní analýza

Bakalářská práce

Autor: David Pirkl
Studijní obor: aplikovaná informatika

Vedoucí práce: doc. RNDr. Ph.D. Pavel Pražák

Hradec Králové Srpen 2022

Vybrané metody analýzy dat v managementu s SPSS

Prohlášení:

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně a s použitím uvedené literatury.

V Hradci Králové dne 29.8.2022 David Pirkl

Poděkování:

Děkuji doc. RNDr. Pavlu Pražákovi, Ph.D. za pomoc při vedení bakalářské práce, cenné rady a vstřícnost v průběhu její kompletace, bez které by výsledná práce nemohla vzniknou

Anotace

Bakalářská práce se zabývá vybranými metodami analýzy dat v managementu, jmenovitě korelační a lineární regresní analýzou a jejich aplikační možnosti napříč oblastí managementu. Jednotlivá témata korelační a lineární regresní analýzy jsou vysvětleny za využití doložené odborné literatury a s ukázkou jejich užití na konkrétních případech. Je představen statistický software SPSS a jeho komponenty pro provedení analýzy dat při užití vybraných metod, která je následně provedena nad reálným souborem dat zaměřeném na službu vypůjčení kol ve městě Soul. Důraz je kladen na aplikační možnosti vybraných metod analýzy dat pro využití v managementu a generalizace výsledků, díky které je možné činit informovanější rozhodnutí.

Annotation

Selected methods of data analysis in management

The bachelor's thesis deals with selected methods of data analysis in management, namely correlation and linear regression analysis and their application possibilities in the field of management. Individual topics of correlation and linear regression analysis are explained using professional literature and with examples of their use. The statistical software SPSS and its components are introduced for performing data analysis using selected methods, which is subsequently performed on a real data, which focuses on the bicycle rental service in the city of Seoul. Emphasis is placed on the application possibilities of selected methods of data analysis for use in management and the generalization of results, thanks to which it is possible to make more informed decisions.

Obsah

Seznam obrázků

1 Úvod

Téma korelační a regresní analýzy v managementu v SPSS bylo vybráno, jelikož jsem se v rámci pracovního poměru setkal s potřebou jejich aplikace, ve které jsem cítil značnou míru subjektivity a nejistoty. Tím, že je práce s daty zároveň mým osobním zájmem, bylo zvoleno toto téma pro mou bakalářskou práci.

Hlavním cílem je tak nelézt využití vybraných metod datové analýzy v managementu a pro podnikové řízení. Zpracováním tohoto tématu chci ověřit, že vybrané metody datové analýzy mají v managementu využití, která mohou pozitivně ovlivnit chod společnosti, ať už v lepším poznáním vnitropodnikových procesů, pro prediktivní účely, či dalšími cíli vycházející z managementu.
Práce je členěna do 2 částí, které lze popsat jako teoretickou a praktickou část. V první, teoretické části jsou shrnuty poznatky z odborné literatury zabývající se korelací a lineární regresní analýzou proložené několika příklady, které se snaží aplikovat aktuálně řešenou problematiku pro využití v managementu. V druhé, praktické části je představen statistický software SPSS od společnosti IBM a jeho komponenty potřebné pro realizaci korelační a lineární regresní analýzy. Kromě samotných modulů zabývajících se korelací a lineární regresí je předvedeno využití popisné statistiky a možnosti grafů s nastaveními vhodnými pro rozsáhlé datové soubory. Při představení možností a nastavení v SPSS je využit datový soubor obsahující data o složkách a kvalitě betonu a je zaměřen především na představení práce v SPSS. Následně je provedena analýza dat za pomoci vybraných metod nad datovým souborem obsahujícím data o službě výpůjčky kol ve městě Soul v průběhu 1 roku, která je zaměřena především na samotnou analýzu a interpretaci výsledků.

Využitý datový soubor služby výpůjčky kol pro praktickou ukázku datové analýzy byl vybrán především pro svůj obsah, který je podnikového charakteru, má 14 proměnných různého typu a přes 6 000 záznamů, což je dostatečné množství pro účely této práce, tedy nalezení a potvrzení aplikačního využití vybraných metod analýzy dat v managementu.

Oblasti podnikání mohou být různého charakteru, kde výrobní společnosti mohou mít odlišné požadavky a cíle kladené na analýzu dat, než společnosti zabývající se například nákupem a prodejem, avšak i odlišné oblasti podnikání mohou mít společné znaky a využívat stejné výkonnostní ukazatele. V rámci uvedených příkladů a také na praktické ukázce je tak snaha zobecnit nalezené výsledky i mimo oblast aktuálně zkoumaného datové souboru.

2 Cíl práce

Cílem práce je použití vybraných metod analýzy dat s SPSS ve vybraných problémech managementu. Problémy managementu mohou být různorodé a pro každou společnost trochu odlišné, cílem tak je nalézt a potvrdit využití vybraných metod analýzy dat pro typově odlišné příklady v oblasti managementu a zobecnit nalezené poznatky pro jejich využití.

3 Metodika zpracování

Práce je členěna do 2 hlavních celků.

První část je tvořena shrnutím znalostí o korelační a lineární regresní analýze z doložené literatury. Jsou tak popsány jednotlivá témata potřebná pro provedení korelační a lineární regresní analýzy od základnějších po komplexnější. Tato část je proložena několika smyšlenými příklady z oblasti managementu tak, aby reprezentovali možnost jejich využití. I když jsou příklady smyšlené, typově jsou jejich zadání sestavena tak, aby reflektovali možné reálné požadavky plynoucí z oblasti managementu.

Druhá část je tvořena aplikací teoretických znalostí probíraných v části první. Pro účely této praktické ukázky jsou využity 2 reálné datové soubory. Jedním je datový soubor zabývající se kvalitou betonu sloužící především pro ukázku práce se statistickým softwarem SPSS, jeho nastavením a realizací analýzy dat za pomoci vybraných metod. Text pro analýzu dat tohoto datového souboru je doprovázen obrázky především pro nastavení jednotlivých částí SPSS a její výsledky jsou interpretovány velmi stručně. Druhým je datový soubor o službě výpůjčky kol ve městě Soul a zaměřuje se převážně na samotnou analýzu dat za pomoci vybraných metod, její vyhodnocení a využití v oblasti managementu.

V rámci jednotlivých příkladů je snaha generalizovat nalezené výsledky tak, aby byli využitelné i mimo analyzovaný datový soubor, a tak dodávali managementu informace, které mohou využít pro efektivnější řízení.

4 Využitá terminologie

Vybrané metody analýzy dat využívají odbornou terminologii, kterou je vhodné před otevřením jednotlivých témat představit. Problematika definice pojmů je shrnuta například v [1] (2015, str. 17-23), kde pro některé v této práci využité termíny neexistuje jednoznačný všeobecně uznávaný popis. Takovým příkladem je pojem *proměnná*, která je nejen v této práci velmi často se vyskytujícím pojmem. Dle [1] (2015, str. 17) ve významovém anglickém slovníku Websters's college dictionary (1995, str. 1476) pro ni existuje dvanáct významů. Dle těchto definic se dá proměnná pro účely statistické analýzy dat v této práci definovat jako vlastnost sledované statistické jednotky. Její význam a následné členění dále popisuje [1] (2015, str.17) následujícím způsobem:

„Pro statistickou analýzu dat je typické odlišení vysvětlovaných (závisle) proměnných v postavení důsledku, od vysvětlujících (nezávisle) proměnných v postavení příčin. Vysvětlované proměnné jsou (v předchozím uvedeném smyslu) mimo přímou kontrolu a jejich proměnné jsou v ideálním případě nenáhodné, tedy pod kontrolou řešitele úlohy, ale (při určitých omezeních a dodatečných podmínkách) mohou rovněž být náhodnými veličinami. Nenáhodné veličiny bývají též označovány za pevné nebo fixní, ale žádné z uvedených označení není úplně výstižné už jen z toho důvodu, že náhodným měřicím chybám se těžko vyhneme."

V předchozí citaci je též využit pojem *veličina*, která dle [1] (2015, str. 17) je v české statistické literatuře považována za téměř rovnocennou k pojmu *proměnná* s tím, že je používána ke kvantitativnímu popisu jevů a nejrůznějších vlastností. Text předchozí citace je výstižný i pro vybrané metody analýzy dat v rámci managementu. Závislou proměnnou mohou představovat požadované výsledky či ukazatele, jako jsou podíly na trhu, návratnost investic, náklady a mnoho dalších. V roli nezávislých proměnných jsou jevy jak ovlivnitelné (nenáhodné), jako jsou alokace financí či transformace pracovního prostředí, tak jevy neovlivnitelné (náhodné), jako je měsíční příjem zákazníků či konkurenční aktivita. K předchozímu popisu by se dalo namítnout, že ze statistického pohledu příjem zákazníků, stejně tak jako konkurenční aktivita, nejsou náhodnými jevy, lidé mají příjem zasloužený například dle jejich odbornosti a zkušeností, konkurenční aktivity jsou řízeny plánováním a strategií jejich vedením, avšak z pohledu zadavatele či vykonavatele analýzy dat jsou tyto jevy neovlivnitelné, případně je jeho vliv na ně zcela zanedbatelný, a tak budou v této práci považovány za náhodné.

Odlišné metody analýzy dat mohou vyžadovat různé formy proměnných pro jejich využití. [1] (2015, str. 18-20) popisuje 3 důležitá kritéria proměnných, které následně určuje jejich možné využití. Tento výčet sice není kompletní, ale jak je ve zdroji uvedeno, tak dostačující pro všechny zde využité metody analýzy dat. První je obor možných hodnot, který proměnné dělí na diskrétní a spojité. Diskrétní hodnoty, tedy spočitatelné, v oblasti managementu tak mohou být počet zaměstnanců, poboček nebo prodaných kusů výrobků, kdežto spojité mohou nabýt jakékoliv hodnoty z definovaného intervalu, jako jsou finanční výnosy, změna v podílech nebo výkonnost, kde se stanovuje požadovaná přesnost na hodnotu. Dalším kritériem je postavení proměnné podle zadání úlohy, která následně dělí proměnné na vysvětlované (závislé) a vysvětlující (nezávislé). Tedy při požadavku na zjištění nákladovosti pro vývoj nového výrobku je hledaná hodnota nákladů vysvětlovanou proměnnou a proměnné které ji ovlivňují, např. zaměstnanecké náklady, spotřeba energie či materiál na výrobu, jsou vysvětlující proměnné. Třetím kritériem je přesnost měření, ty dělí proměnné do kategorií, které je dle [1] (2015, str. 20) v krátkosti možné popsat následovně:

1) Nominální – jedná se o kvalitativní proměnné ze kterých lze vyčíst pouze to, že jsou odlišné. Nemá smysl je nijak řadit či poměřovat. Například textový kód produktu
2) Ordinální – jsou také kvalitativního charakteru, avšak má smysl je řadit vzestupně nebo sestupně dle jejich obsahu. Například výše dosaženého vzdělání
3) Kardinální – mají nejvyšší úroveň měření, jejich hodnoty dovolují je řadit i poměřovat. Například již zmíněné výnosy a náklady.

V předchozí citaci byl využit pojem statistická analýza dat, kdy právě oblasti analýzy dat a statistiky mají v managementu obdobná využití, kterými mohou být lepší vhled do oblasti podnikání či získání více informací pro důležitá rozhodování. I v tomto případě neexistuje jednoznačný popis a je vhodné je definovat v rámci kontextu ve kterém jsou využity, tedy pro aplikaci v oblasti managementu. [2] (2013, str. 2) popisuje analýzu dat následujícím způsobem:

„Účelem analýzy dat je studovat charakteristiky výběrového souboru dat a jejich zobecnění na soubor populační. Vyvození závěru o populaci na základě výběrového souboru by byl platný pouze v případě, že obsah výběrového souboru vhodně reprezentuje danou populaci. To lze zajistit použitím správné vzorkovací techniky. Velký výběrový soubor také nemusí nutně znamenat lepší výsledky. Není důležité množství, ale kvalita výběrového souboru“

Statistika je obsáhlým oborem, kdy jejímu vysvětlení je mnohdy věnována rozsáhlá úvodní část odborné literatury, která se jí zabývá. Pro její stručný popis však lze citovat [3] (2016, str. 2):

> *„Statistika je věda o tvorbě studií nebo experimentů, shromažďování dat a modelování/analýze dat za účelem rozhodování a vědeckých objevů, pokud jsou dostupné informace omezené, či proměnlivé. To znamená, že statistika je věda o učení se z dat“*

Ze zmíněných citací je tak vidět jistá míra průniku jejich obsahu. Avšak dle užití těchto pojmů z doložené literatury a pro účely managementu je možné je rozdělit následovně. Analýza dat je zaměřená na již existující soubor dat, jeho čištění, transformaci, popisu a rozboru dat a tvorbě modelů. Statistika se zabývá i samotným způsobem sběru potřebných dat, popisu jejich vlastností a s jistou mírou pravděpodobnosti rozšiřovat získané znalosti z výběrového souboru na soubor populační, kde výběrový soubor jsou všechna dostupná data, která jsou podmnožinou všech dat existujících pro danou proměnnou, tedy populačním souborem. Populační soubor však nemusí být vždy k dispozici. Nutno podotknout, že se jedná o velmi zjednodušenou definici snažící se ukázat hlavní znaky obou těchto pojmů pro účely této práce.

5 Korelace

Korelace vyjadřuje vzájemnou souvislost dvou proměnných a bývá základem pro mnohé další statistické metody. Její stručný popis je vystižen v [1] (2015, str.238):

„Předmětem zájmu jsou změny hodnot jedněch veličin, které s větší či menší pravděpodobností vyvolávají nebo doprovázejí změny hodnot jiných veličin. Místo kauzality, kterou se myslí (v této souvislosti tedy vzácné) pevné příčinné spojení jevů, událostí nebo proměnných, se výzkumník musí spokojit s volným příčinným spojením, které se podle úlohy a typu proměnných (nejednoznačně) označuje za asociaci nebo za korelaci“

Zjištění existence a následné míry lineární souvislosti mezi proměnnými může být v managementu cennou informací. Korelace dokáže vysvětlit různé vlivy, které mají dopad jak na nákladové, tak výnosové položky podniku, a tím pomoci k efektivnějšímu řízení a lepšímu pochopení oblasti podnikání. Příkladem tak může být výrobní společnost vlastnící velké haly, kde může korelace ušetřit značné množství peněz, pokud se díky ní zjistí, že kvalitnější pracovní prostředí (investice do klimatizace, lepší cirkulace vzduchu a další) má na efektivitu značně pozitivnější vliv než navýšení finančního bonusu. Pro investiční společnost jsou důležité souvztažnosti mezi jejich finančními aktivity pro úspěšnou strategii diverzifikace (rozložení investic do více, ideálně na sobě co nejméně závislých oblastí), kdy vyšší diverzifikace znamená menší závislost mezi finančními aktivity, a tak propad jednoho aktiva tak vůbec, nebo pouze minimálně, ovlivní hodnotu aktiv ostatních. Tím společnost minimalizuje riziko náhlé ztráty značného kapitálu jednou neočekávanou událostí.

Pro management, který se zabývá řízením společností a jejich částí, jsou tak informace jedním z nejcennějších zdrojů a zjištění souvislosti sledovaných jevů, jak ukazují předchozí příklady, jsou jednou z nich. Avšak současně pro korelaci platí i některá omezení, kdy jednou z nejzásadnějších je v literatuře často zmiňovaný fakt, že korelace neimplikuje kauzalitu, tedy že vysoká míra korelace nutně neznamená příčinnou souvislost mezi sledovanými jevy. Další podmínkou je linearita mezi sledovanými proměnnými. Korelace je tak schopna popsat pouze proměnné, které mají mezi sebou určitou míru lineární souvislosti a o jiných potencionálních souvislostech je nevypovídající. Při potvrzení nulové korelace tak nelze říct, že mezi proměnnými neexistuje žádná souvislost, pouze je možné tvrdit, že jsou lineárně nezávislé a jiné souvislosti tak nelze vyloučit.

5.1 Korelační koeficient

Korelační koeficient je hodnota, kterou se vyjadřuje míra korelace mezi proměnnými. Může nabývat jak kladné, tak i záporné hodnoty. Pokud je hodnota koeficientu záporná, znamená to, že při zvyšující se hodnotě jedné proměnné hodnota druhé proměnné klesá. Při kladné hodnotě korelačního koeficientu se hodnoty proměnných zvyšují nebo snižují souběžně. Typů, a tedy i výpočtů, korelačních koeficientů je více, avšak nejpoužívanějším je Pearsonův korelační koeficient, který i díky četnosti svého využití může být implicitně myšlen při zmínce o korelačním koeficientu, jak dokládá následný text od [4] (2017, str.634):

„korelační koeficient (r) je nejpoužívanější statistikou, která představuje vzájemnou souvztažnost mezi dvěma proměnnými (intervalové nebo poměrové), řekněme X a Y. Používá se k určení, zda mezi X a Y existuje lineární vztah. Udává míru, do jaké míry variace jedné proměnné, X, souvisí s variací druhé proměnné, Y. Protože byl původně navržen Karlem Pearsonem, je také známý jako Pearsonův korelační koeficient a také označován jako jednoduchá korelace, bivariační korelace nebo pouze korelační koeficient.“

Z tohoto důvodu je tak v rámci této práce využit pouze Pearsonův korelační koeficient. Popis jiných korelačních koeficientů, jejich výpočtu a oblasti využití je možné nalézt například v textu [6] (2015, str. 271-289). Před popisem výpočtu Pearsonova korelačního koeficientu je vhodné představit tři pojmy, ze kterých je jeho výpočet složen.

Rozptyl udává, jak moc jsou pro danou proměnnou jednotlivá pozorování vzdálena od střední hodnoty, v tomto případě aritmetického průměru. Výpočet vychází ze součtu umocněných odchylek, tedy součtu jednotlivých hodnot, od kterých je odečtena střední hodnota proměnné a v případě výběrového rozptylu vydělen rozsahem výběrového souboru. Umocnění zajišťuje kladné znaménko a vyjadřuje tak vzdálenost jednotlivých pozorování od střední hodnoty na druhou (součet těchto hodnot tak díky umocnění na druhou bývá nazýván suma čtverců, v angličtině *Sum of Squares*)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

X_i – i-tá hodnota proměnné X
$\overline{X}$ – průměrná hodnota proměnné X

Kovariance je mírou smíšené variability dvou proměnných. Vychází z hodnot zkoumaných proměnných a může nabývat libovolné kladné, nulové nebo záporné hodnoty a ukazuje směr lineární závislosti mezi danými proměnnými.

$$COV_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

X_i / Y_i – i-tá hodnota proměnných X a Y
$\overline{X}$ / $\overline{Y}$ – průměrná hodnota proměnné X a Y
n – počet pozorování

Směrodatná odchylka je spočtena odmocněním rozptylu, čímž se výsledné hodnoty dostanou do původních jednotek. Je vhodné zmínit konvenci značení směrodatné odchylky, pokud je prováděna nad populačním souborem, značí se písmenem σ (sigma), kdežto při výpočtu nad výběrovým souborem je značena písmenem s.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Se znalostí těchto vzorců je následně možné provést transformaci dat, které se říká **normování**. Cílem normování je transformovat náhodnou proměnnou X do sjednocené podoby (normy) tím, že její střední hodnota bude odpovídat nule E(X) = 0 a rozptyl bude odpovídat $s^2(X) = 1$. Normalizaci je tak možné provést za pomocí následujícího vzorce pro každý záznam proměnné X.

$$Z_i = \frac{X_i - E(X)}{s(X)}$$

E(X) – střední hodnota proměnné X
s(X) – rozptyl proměnné X

Výsledné hodnoty normovaných proměnných tak jsou v intervalu <-1;1>, kdy krajní hranice představují silnou, či úplnou souvztažnost proměnných a nula souvztažnost žádnou. Normování je užitečnou formou transformace díky univerzální reprezentativní hodnotě normovaných hodnot. V praxi se ve studovaných jevech a potažmo proměnných, ze kterých jsou složeny, vyskytuje mnoho různých jednotek a rozdíly v jejich hodnotách by vedly ke značné komplikovanosti jak během výpočtu, tak i jejich následnému vyhodnocení. Díky normování tak není potřeba se držet kontextu každé proměnné (tedy jejich rozptylu a umístění střední hodnoty) a je možné napříč všemi proměnnými pracovat unifikovaně, což se je obzvláště výhodné při práce s větším počtem proměnných.

Příklad č. 1

Ve smyšleném příkladu Je k dispozici datový soubor na obrázku č. 1, který obsahuje počet servisovaných motorových vozidel (X) a odpracovaných hodin zaměstnanců (Y) v rámci jedné automobilové prodejny za loňský rok, agregované po měsících. Je požadován popis dat, konkrétně zjištění rozptylu, kovariance a směrodatné odchylky těchto proměnných

měsíc	Počet servisovaných vozidel (X)	Odpracovaných hodin (Y)
1	198	429
2	214	487
3	228	482
4	210	449
5	247	507
6	204	462
7	195	430
8	201	452
9	241	492
10	220	412
11	217	420
12	233	477

Obrázek 1 – Datový soubor servisovaných vozidel, vlastní zpracování

Průměrná hodnota $X_1 = 217$ a $X_2 = 458$

Rozptyl jednotlivých proměnných je následující:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \frac{(198-217)^2 + (214-217)^2 + \dots (233-21}{12-1}$$

$$= \frac{361 + 9 + \cdots + 256}{11} = \frac{2608}{11} = 290$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$= \frac{(429-458)^2 + (487-458)^2 + \dots (477-458)^2}{12-1}$$

$$= \frac{841 + 841 + \cdots + 361}{11} = 968$$

Odmocněním nalezených rozptylů se získá směrodatná odchylka pro obě proměnné, tedy $s_x = \sqrt{290} \doteq 17$ a $s_y = \sqrt{968} \doteq 31$. Tyto hodnoty určují vzdálenost záznamů od střední hodnoty neboli rozptýlenost, a je díky nim možné lépe pochopit, jak moc data kolísají od střední hodnoty. Průměrný počet servisovaných vozidel tak je 217 a směrodatná odchylka představuje 17 vozidel. To, zda je výsledek dobrý či špatný nelze bez širšího kontextu hodnotit, nicméně je možné díky této informaci si lépe představit počty přijatých vozidel v průběhu roku, kdy by k rozptylu od střední hodnoty o více než zmiňovaných 17 vozidel nemělo docházet příliš často, tedy méně než 200 či více než 234 vozidel. Obdobně lze hodnotit i výsledky pro počet odpracovaných hodin, kde je měsíčně průměrně odpracováno 458 hodin se směrodatnou odchylkou 31 hodin.

Kovariance se získá součtem součinů všech hodnot X_i - $\bar{X}$ a $Y_i - \bar{Y}$, který je vydělen rozsahem výběrového souboru:

$$COV_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$= \frac{(198-217)(429-458) + (214-217)(487-458) + \cdots + (233-217)(477-458)}{11}$$

$$= 355$$

Kovariance mezi počtem servisovaných vozidel a počtem odpracovaných hodin je 355. Tato hodnota jejich smíšené variability je sama o sobě pro účely managementu příliš nevypovídající, a to především díky tomu, že výsledek je reprezentován v jejich původních hodnotách a není nijak normalizován. Avšak tato operace je nedílnou součástí Pearsonova korelačního koeficientu, který tuto hodnotu normalizuje pomocí směrodatných odchylek, jak je uvedeno v následujícím oddíle.

5.1.1 Pearsonův korelační koeficient (r)

Pearsonův korelační koeficient je nejčastěji využívaným korelačním koeficientem, jak uvádí [1] (2015, str.239) nebo [4] (2017, str.634). Je definován jako kovariance sledovaných proměnných vydělena jejich směrodatnými odchylkami:

$$r = \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}} = \frac{COV_{xy}}{s_x\, s_y}$$

Hodnota kovariance sledovaných proměnných v čitateli je díky vydělení jejich směrodatnými odchylkami normalizována do intervalu <-1;1>. Tím, že je výpočet Pearsonova koeficientu založen na rozptylech proměnných, je poměrně citlivý na odlehlé hodnoty. Je tedy vhodné před samotným výpočtem koeficientu data o takové hodnoty očistit, případně jinak zpracovat, aby nebyla výsledná hodnota koeficientu zkreslena.

V managementu má tak Pearsonův korelační koeficient mnohá využití, která jsou limitována především dostupností využitelných dat. Příkladem může být expanze kamenných prodejen knihkupectví, kdy úspěšnost prodejny může být hodnocena mnoha způsoby, podle cíle expanze. Ať je hlavním cílem zvýšení ziskovosti nebo rozšíření podílů, postavení nové prodejny může být finančně náročnou a dlouhodobou investicí a množství faktorů ovlivňujících počet návštěvníků a velikost průměrného nákupu je těžko odhadnutelná. Avšak při využití Pearsonova korelačního koeficientu je možné zjistit které faktory, například vzdělání, počet obyvatel či měsíční příjem, ovlivňují prodeje jednotlivých prodejen. Tyto informace pomohou managementu s rozhodnutím, ve kterém městě a v jaké lokaci bude prodejna pravděpodobněji úspěšnější. Avšak nalezené Pearsonovy koeficienty mohou podléhat jisté míře zkreslení, které se dá minimalizovat dostatečně rozsáhlým datovým souborem, jak vysvětluje [1] (2015, str. 239):

„nadhodnocení odhadu dochází v případech, kdy vysvětlující proměnné jsou nenáhodné. Avšak ani pro náhodné vysvětlující proměnné není koeficient nezkresleným odhadem s tím, že s růstem rozsahu náhodného výběru se zkreslení, i když pomalu, zmenšuje“

5.1.2 Testování korelačního koeficientu

Po nalezení korelačního koeficientu je vhodné se ujistit o jeho dostatečné statistické významnosti testem nulového hypotézy. Ta je provedena stanovením nulové hypotézy, která tvrdí, že mezi sledovanými proměnnými neexistuje korelace, tedy r=0. Problematika testování hypotéz je obsáhlé téma a v této práci je popsána pouze velmi stručně, pro hlubší popis celé problematiky je možné odkázat na text [2] (2013, str. 167-215), který prokládá tématiku četnými příklady.

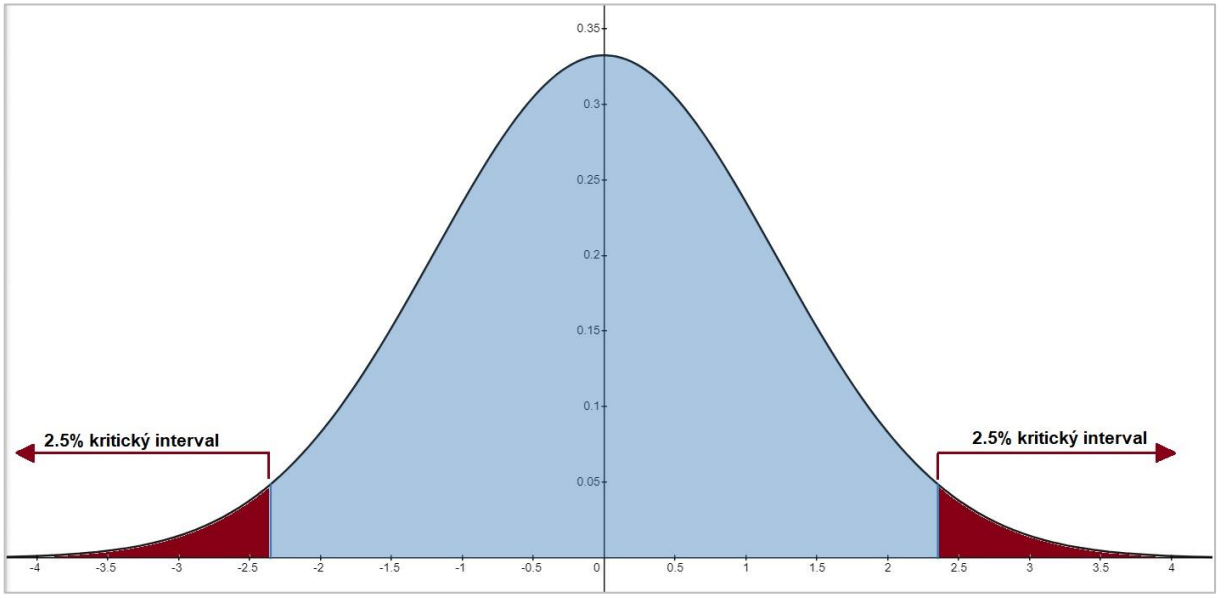
Pro testování Pearsonova korelačního koeficientu tak [4] (2017, str. 637) či [1] (2015, str.240) používají jako vhodnou statistiku:

$$t = r\left(\frac{n-2}{1-r^2}\right)^{\frac{1}{2}}$$

které má studentovo t rozdělení pravděpodobnosti o n–2 stupni volnosti.
Jsou stanoveny následující hypotézy:

- H_0 (nulová hypotéza) – neexistuje statisticky významná korelace mezi sledovanými proměnnými
- H_1 (alternativní hypotéza) – mezi sledovanými proměnnými existuje statisticky významná korelace

V případě testování korelačního koeficientu tak nulová hypotéza tvrdí, že zmíněné proměnné nejsou korelované a alternativní hypotéza tvrdí opak. Hladina významnosti je označována písmenem α a představuje maximální přípustnou pravděpodobnost, do které nulovou hypotézu nezamítáme. Po jejím překročení se hodnota nachází v takzvaném kritickém intervalu, kde se již nulová hypotéza zamítá a přijímá se hypotéza alternativní. Jinak řečeno, při překročení hranice významnosti je možné s dostatečnou důvěrou tvrdit, že je nulová hypotéza zamítnuta a tím se vypočtený korelační koeficient považuje za důvěryhodný. Hodnoty pro tuto hranici významnosti nejsou pevně stanoveny, avšak napříč literaturou, např. [4] (2017, str. 567), jsou nejčastěji využívané hodnoty 0.05 a 0.01 (neboli 5 % a 1 %). Testy mohou být jednostranné nebo oboustranné, kdy v případě oboustranného testu se α dělí 2, kritický interval je tak při využití 5 % následně tvořen 2.5 % na každé straně, jak je vidět na obrázku č. 2.



Obrázek 2 – Vizualizace oboustranného intervalu spolehlivosti pro α=0.05, střední hodnotou 0, a rozptylem 1.2. vlastní zpracování za pomoci online grafového editoru www.desmos.com

Po dosazení vypočteného koeficientu *r* a počtu záznamů *n* do vzorce testovací statistiky tak vyjde hodnota, která pokud se nachází v kritickém intervalu, dovolí zamítnout nulovou hypotézu o nekorelovanosti, a podpoří tak významnost korelačního koeficientu. Výpočty hladiny významnosti již obvykle provádí software, avšak je vhodné předvést alespoň jeden výpočet, kdy pro t statistiku s požadovanými stupni volnosti je nutné v tabulkách nalézt konkrétní hodnotu odpovídající přechodu do kritického intervalu.

Příklad č. 2

Ve smyšleném příkladu je zjišťována hypotetická souvislost mezi spokojeností zákazníka v restauračním zařízením a finančním ohodnocením zaměstnanců. Vypočítaný Pearsonův korelační koeficient je 0.67 a byl spočítán z 32 dotazníků vyplněných zákazníky po zaplacení účtu, který nabízel celkovou spokojenost na škále od 1 (zcela nespokojen) do 10 (zcela spokojen). Nyní je nutné potvrdit, zda je nalezený koeficient významný na zvolené hladině 0.05 a může tak být využit pro další analýzu. Při využití stupně volnosti n-2 pro t studentovo rozdělení s oboustrannou alternativní hypotézou, tedy pro hodnotu 30, je hranice 2.042, kterou je možné nalézt ve statistických tabulkách [5] v části Studentova rozdělení t.

H_0: mezi sledovanými proměnnými neexistuje korelace (r = 0)
H_1: mezi sledovanými proměnnými existuje korelace (r ≠ 0)

$$t = r\left(\frac{n-2}{1-r^2}\right)^{\frac{1}{2}} = 0.67\sqrt{\frac{30}{0.5511}} = 0.67 \cdot 7.38 = \mathbf{4.94}$$

$$t = 4.94 > 2.042$$

Nulová hypotéza o neexistenci korelace mezi spokojenosti zákazníků a finančním ohodnocením zaměstnanců se tak zamítá. Když je nyní dostatečně potvrzena informace, že finanční ohodnocení zaměstnanců má vliv na spokojenost zákazníků, která je jednou z podstatných kritérií pro úspěšné a dlouhodobé vedení restauračního zařízení, je vhodné tuto možnost a její návratnost zvážit při dalším finančním plánování. Toto zjištění, že finanční ohodnocení zaměstnanců má dopad na spokojenost zákazníků, kromě i dalších potencionálních benefitů, je možné využít například pro rozhodnutí přenechání 100 % spropitného zaměstnancům.

5.2 Parciální korelace

Vysoká korelace mezi dvěma proměnnými X a Y nemusí vždy nutně znamenat jejich přímou souvislost. Na sledovaný jev má mnohdy vliv mnoho různých proměnných, které nemusí být vždy k dispozici, ať už z důvodu nedostupnosti dat či složitosti měření požadované proměnné, a každá z nich se může různou mírou podílet na výsledku sledovaného jevu. Zároveň však může docházet k ovlivňování nejen sledované proměnné, ale k vzájemnému ovlivňování mezi ostatními proměnnými, které má pak negativní dopad na nalezenou korelaci a je nutné tento problém řešit. Například na počet prodaných telefonů tak bude mít vliv jejich cena, avšak výsledné prodeje vůči cenně telefonu může ovlivňovat například i průměrná měsíční mzda obyvatel nebo vynaložené finanční náklady na reklamu ve sledovaném období, zavedení nového modelu do prodeje a mnoho dalších vlivů. Pokud se tedy v rámci managementu připravují důležitá rozhodnutí a je nutné znát „čistý“ vliv stanovených proměnných vůči sledované proměnné, parciální korelace je dle [6] (2015, str. 299) způsob, jak nejlépe zjistit souvislost mezi dvěma proměnnými při odstranění vlivu jiných proměnných.

Hledá se tak potencionální vliv třetí (či více) proměnné Z, která může mít s X i Y významnou lineární souvislost, jejíž změna tak ovlivňuje obě proměnné. Při neobjevení takovéto souvislosti je možné špatně interpretovat výsledky analýzy a případné modely, které by z nich vycházely, by ztratili na přesnosti. Dle [6] (2015, str. 299-318) existují 4 možné vlivy třetí proměnné, které mohou mít dopad na souvislost původních dvou interpretovaných proměnných:

1) Po započtení vlivu třetí proměnné původní souvislost mizí, nebo je výrazně redukován. Souvislost mezi X a Y je tak označována za zdánlivou, či klamnou
2) Souvislost X a Y je souvislostí nepřímou a třetí proměnná je tak chápána jako intervenující, tedy je ovlivňována proměnnou X a následně ovlivňuje proměnnou Y.
3) Mezi původními proměnnými nedojde k žádným změnám. Třetí proměnná souvislost nijak neovlivňuje a ten je tak označován za ryzí.
4) Souvislost mezi X a Y je odlišná pro různé kategorie třetí proměnné, takový stav je označován za interakční efekt.

Parciální korelace odstraňuje vliv vybraných náhodných proměnných, a tak umožňuje zjistit souvislost mezi dvěma proměnnými bez jejich vlivu, výsledkem je tak korelační koeficient lépe popisující „čistou“ souvislost mezi nimi. Toho je docíleno sledováním korelace požadovaných proměnných při pevných hodnotách kontrolované proměnné. Jak vysvětluje [4] (2017, str. 639) množství takto

vyloučených vlivů ostatních proměnných se značí tzv. řádem. Parciální koeficient pro X a Y při kontrole Z je tak nazván parciální koeficient prvního řádu (korelace mezi 2 proměnnými je tak vlastně parciální korelace nultého řádu, bez kontroly další proměnné) a ve vzorci je definován v dolním indexu koeficientu, kde kontrolované proměnné jsou odděleny za tečkou, v tomto případě tedy $r_{XY \cdot Z}$. Pro výpočet je dostačující znalost jednotlivých korelačních koeficientů:

$$r_{XY \cdot Z} = \left(\frac{r_{XY} - (r_{XZ} r_{YZ})}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}} \right)$$

$r_{XY \cdot Z}$ – korelační koeficient proměnných X a Y při kontrole proměnné Z
r_{XY} – korelační koeficient proměnných X a Y
r_{XZ} – korelační koeficient proměnných X a Z
r_{YZ} – korelační koeficient proměnných Y a Z

Příklad č. 3

Ve smyšleném příkladu má na množství prodaných knih (proměnná *Y*) vliv dosažené vzdělání obyvatel (proměnná X_1) a výše jejich příjmu (proměnná X_2). Majitelé sítě knihkupectví hledají ideální lokace pro otevření nových prodejen a mají k dispozici zmíněná data agregovaná na úrovni obcí a měst. Cílem je tak zjistit, který z těchto vlivů je podstatnější pro množství prodaných knih a z nalezeného výsledku se poté bude vycházet pro rozhodování o lokalitách k expanzi.

Jednotlivé Pearsonovy koeficienty vypočítané pro tyto souvislosti jsou:
$r_{X_1Y} = 0.47$ (prodané knihy a dosažené vzdělání)
$r_{X_2Y} = 0.41$ (prodané knihy a výše příjmu)
$r_{X_1X_2} = 0.67$ (dosažené vzdělání a výše příjmu)

$$r_{XY\cdot Z} = \left(\frac{r_{XY} - (r_{XZ}r_{YZ})}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}\right)$$

$$r_{X1Y\cdot X2} = \left(\frac{r_{X1Y} - (r_{X1X2}r_{YX2})}{\sqrt{(1 - r_{X1X2}^2)(1 - r_{YX2}^2)}}\right) = \left(\frac{0.47 - (0.67 * 0.41)}{\sqrt{(1 - 0.67^2)(1 - 0.41^2)}}\right)$$
$$= \left(\frac{0.47 - 0.2747}{\sqrt{0.5511 \cdot 0.8319}}\right) = \frac{0.1953}{0.6771} = \mathbf{0.29}$$

$$r_{X2Y\cdot X1} = \left(\frac{r_{X2Y} - (r_{X1X2}r_{YX1})}{\sqrt{(1 - r_{X1X2}^2)(1 - r_{YX1}^2)}}\right) = \left(\frac{0.41 - (0.67 * 0.47)}{\sqrt{(1 - 0.67^2)(1 - 0.47^2)}}\right)$$
$$= \left(\frac{0.41 - 0.3149}{\sqrt{0.5511 \cdot 0.7791}}\right) = \frac{0.0951}{0.6552} = \mathbf{0.15}$$

Výsledná hodnota pro koeficient výše vzdělání při kontrole měsíčního příjmu je vyšší než pro výši příjmu při kontrole dosaženého vzdělání. Výše vzdělání je tak relevantnější proměnnou pro výsledný prodej knih a je tak významnější proměnnou než měsíční příjem. Hodnoty korelačního koeficientu tak ani v jednom případě nejsou příliš vysoké, avšak pokud by nebyl jiný zdroj informací, tak výsledky naznačují že města s vyšším počtem lidí, kteří dosáhli vyššího vzdělání jsou preferovanou volbou.

6 Lineární regresní analýza

Korelace je schopna poskytnout informace o souvztažnosti dvou proměnných, odpoví tak na otázky typu, jak moc je prodej aut spjat s cenou benzínu, či zda při prodloužení pracovních hodin v restauraci stoupá, klesá, či zůstává stabilní počet zákazníků. Je tak možné lépe poznat vzájemný vliv mezi proměnnými a díky tomu dělat informovanější rozhodnutí. Pro mnohá rozhodování však nemusí být nalezená informace o korelaci dostačují. Pro oblast managementu a řízení znamená přechod z popisu souvztažností jevů k tvorbě prediktivních modelů, které jsou schopny s jistou mírou pravděpodobnosti kvantifikovat výsledky sledované proměnné i do blízké budoucnosti, respektive při různých hodnotách vysvětlujících proměnných, rozšíření znalostí, a tak možnosti dělat informovanější rozhodnutí. Pokud tak nedostačuje odpověď, že při zvýšení cen benzínu klesá prodej aut, ale je potřeba vědět o kolik procent klesne prodej aut při zvýšení cen benzínu o 5,-Kč, lineární regresní analýza může být vhodnou metodou analýzy dat.

Kromě prediktivních účelů je regrese vhodná i pro explanační účely, kdy pomocí získaných dat je možné zjistit míru vlivu jednotlivých vysvětlujících proměnných a lépe tak popsat sledovaný jev, jako jsou například hlavní příčiny poklesu tržeb nebo jaké nezávislé proměnné a do jaké míry mají vliv na výkonnost skladu. [6] (2015, str. 319-320) jako cíl explanace uvádí vysvětlení míry vlivu a velikost zastoupení jednotlivých nezávislých proměnných na vysvětlované proměnné, a tak co nejlépe popsat jejich vliv na výslednou hodnotu. Prediktivní modelování je zaměřeno především k hledání konkrétní hodnoty závislé proměnné na základě hodnot nezávislých proměnných. Toto je však značné zjednodušení a tato problematika je, jak zmiňuje [7] (2013, str. 39), značně rozsáhlejší:

„Téma explanačního modelování oproti prediktivnímu modelování může vyvolat rozsáhlou debatu, která značně přesahuje naše zaměření. Co je však důležité si uvědomit je, že použité techniky se značně překrývají, ale ne všechna poučení získaná z explanačního modelování platí pro prediktivní modelování. Takže čtenář se zkušenostmi v regresní analýze se může setkat s novými, a dokonce zdánlivě protichůdnými lekcemi“

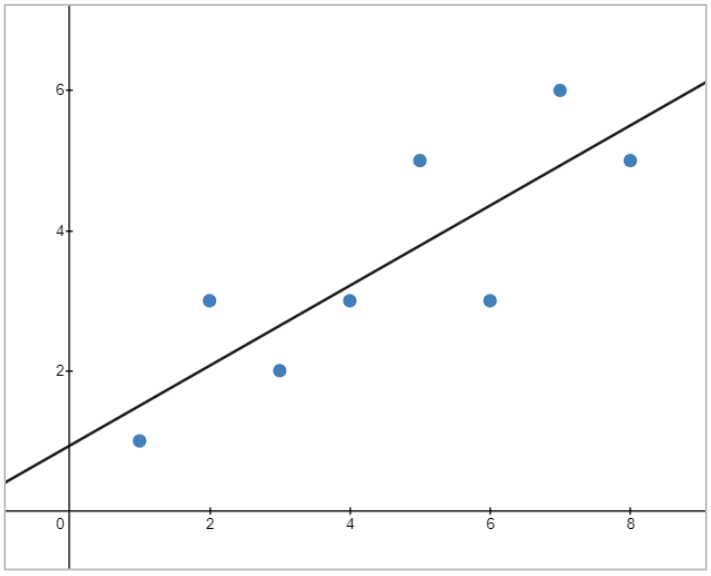
Oba přístupy jsou pro oblast managementu přínosné, kde explanace popisuje vliv a míru jednotlivých nezávislých proměnných na proměnnou závislou, a tak vysvětluje proč se daný jev děje, z jakých nezávislých proměnných je tvořen, a predikce za pomoci modelu dodává hodnotu závislé proměnné a předpovídá jaký bude stav jevu na základě hodnot vysvětlujících proměnných.

6.1 Jednoduchá lineární regrese

Lineární regrese je základní a zároveň často využívanou formou regrese. Její základní model, jednoduchá lineární regrese, zkoumá souvislost mezi dvěma proměnnými. Podstatou je nalézt takovou přímku, která co nejlépe prochází daty a následně se stane aproximační funkcí pro sledovaný jev, tedy matematickým vyjádřením, které se při co nejmenší chybě snaží co nejlépe popsat souvislost pomocí přímky. Avšak možností, jak proložit přímku daty je nespočetně mnoho, co je tedy myšleno pojmem „co nejlépe"? Metod výpočtu posazení přímky je více, avšak mnohdy využívanou je metoda nejmenších čtverců. Čtvercem je myšlena umocněná vzdálenost bodu (svisle) od přímky na druhou a slovem nejmenší je vyjádřen požadavek na položení přímky takovým způsobem, aby součet umocněných vzdáleností všech bodů od přímky byl co možná nejmenší. Samotný výpočet je náročnou operací, kterou dnes provádí software a její popis je možné nalézt např. v [1] (2015, str. 247-248). Předpis jednoduché lineární regrese odpovídá směrnicovému tvaru rovnice přímky:

$$y = \beta_0 + \beta_1 x$$

y – závislá proměnná
x – nezávislá proměnná.
β_1 – koeficient ovlivňující sklon přímky
β_0 – konstanta označující protnutí přímky s osou Y (tedy hodnotu závislé proměnné y při nulovém vstupu nezávislé proměnné x).



Obrázek 3 – Graf lineární regrese, vlastní zpracování

Lineární regresní analýza vyžaduje splnění několika podmínek, které jsou kladeny na využitá data, aby mohla být úspěšně aplikována. Nedodržení podmínek kladených na lineární regresi vede k nevhodně sestavenému modelu, který ztrácí na přesnosti. Podmínky pro jednoduchou lineární regresi jsou shrnuty [6] (2015, str. 320) následovně: „

1) souvislost mezi analyzovanými proměnnými musí být lineární
2) Závisle proměnná Y je měřena na intervalové úrovni a nezávisle proměnná X je buď intervalová, nebo dichotomická
3) Obě proměnné by měly být přibližně normálně rozloženy. Pokud ovšem máme dostatečně velký soubor (cca N>100), nemusíme se tímto předpokladem příliš trápit, neboť díky centrální limitní větě platí, že v takové situaci „nenormální“ rozložení nemá na výsledky velký vliv.“

Vzhledem k četným využitím lineární regrese a díky moderním výpočetním možnostem je prakticky okamžitě přístupná, a tak dochází k jejímu nevhodnému použití i v případech, kde by jiné statistické metody byly vhodnější. To může být využití nad daty, která nesplňují některý z dříve uvedených požadavků, například pokud by se dala souvislost lépe aproximovat jinou než lineární funkcí, takový lineární regresní model by mohl dodávat zkreslené výsledky, které by byly přesnější při využití adekvátní regresní funkce, či jiné statistické metody.

Při správném využití má však lineární regrese v managementu a řízení svá využití. Je možné díky ní lépe poznat vnitropodnikové procesy, pokud tak management potřebuje zvýšit měsíční produkci, pomocí lineární regrese je možné zjistit jaké proměnné mají na výslednou produkci vliv a výsledky modelu zajistí lepší představu o dopadu potencionálních rozhodnutí. Pro plánování prodejních a dalších cílů je možné využít historická data, které společnosti obvykle uchovávají a nastavit je dle sezónnosti či jiných vlivných faktorů, jenž by byly bez matematického vyjádření stěží odhadnutelné. Koeficient lineární regrese (β_1) může management využít pro počítání vnitropodnikových norem a pomoci tak například rovnocennému hodnocení zaměstnanců. Například, pokud má společnost velké sklady s celodenním aktivním pohybem, nemusí být vhodné zaměstnance odměňovat pouze dle odpracované doby, ale spíše za reálně odvedenou práci. Výkonnost každého zaměstnance tak může být odlišná, kdy například čas strávený přesunem po skladu, počtem zpracovaných zakázek a typem skladovací operace za celý pracovní den mohou mít 2 zaměstnanci značně odlišnou. Pomocí lineární regrese je tak možné pro každou takovouto operaci stanovit normu, která odhalí rozdíly napříč zaměstnanci a může pomoci nastavit optimální hodnotící podmínky.

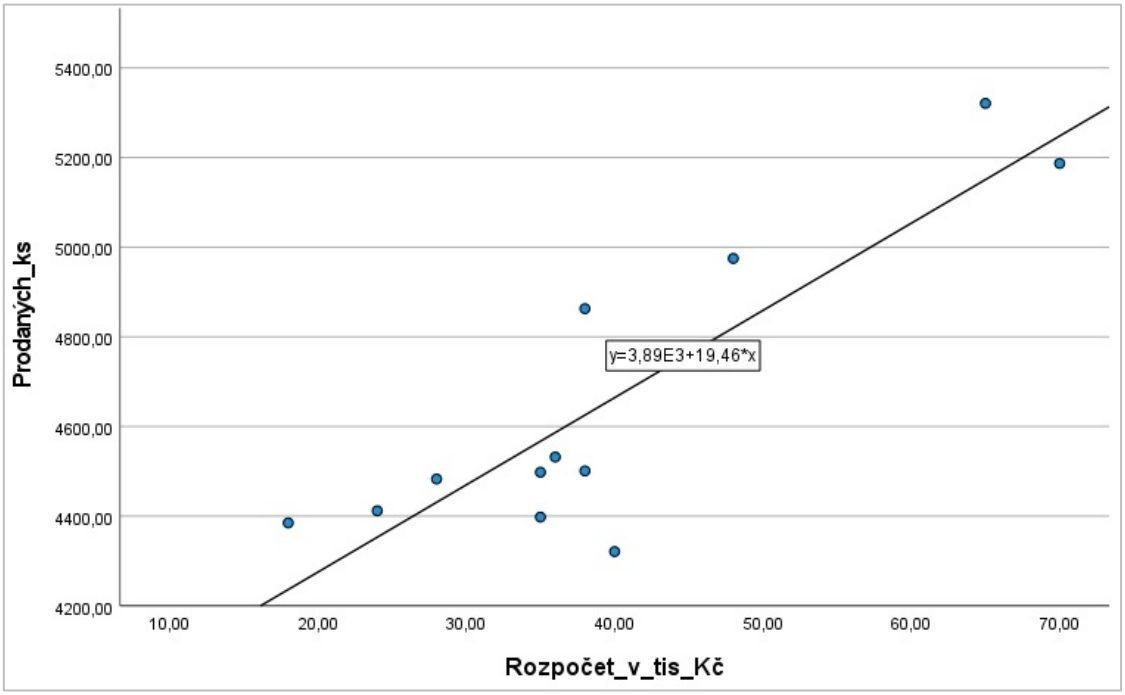
Příklad č. 4

Ve smyšleném příkladu se společnost zabývající se prodejem knih snaží sestavit rozpočet pro příští rok. Téměř ve všech bodech se vedení jednoznačně shodlo, avšak když došla řeč na přidělení peněz marketingovému oddělení, došlo k patové situaci. Zástupci obchodního oddělení tvrdí, že přijetím nového obchodního zástupce, jehož roční náklady jsou 500 000,-Kč zvýší prodej knih o 8 000ks, kdežto marketingové dopady jsou za stejné peníze údajně menší. S tím však polovina zúčastněných osob nesouhlasí. Ke zjištění, čí názor je pravdivý se využijí data z loňského roku, reprezentované v tabulce na obrázku č. 4, obsahující náklady marketingového oddělení a celkový počet prodaných knih po měsících. Pro jednoduchost, nechť platí předpoklad, že data jsou očištěna od všech ostatních vlivů.

měsíc	Rozpočet (v tis. Kč)	prodaných knih
Leden	18	4385
Únor	28	5247
Březen	38	6709
Duben	24	3985
Květen	65	9432
Červen	48	6521
Červenec	36	5217
Srpen	35	4679
Září	40	6019
Říjen	35	3895
Listopad	70	9876
Prosinec	38	5483

Obrázek 4 – Tabulka měsíců, rozpočtu a prodaných knih, vlastní zpracování

Na první pohled lze z dat vyčíst vliv investovaných peněz marketingovým oddělením. Avšak konkrétní tvrzení obchodního oddělení nelze ze samotného pohledu na pozitivní dopad vyvrátit ani potvrdit. Pro tento účel se sestaví lineární regresní model na obrázku č. 5

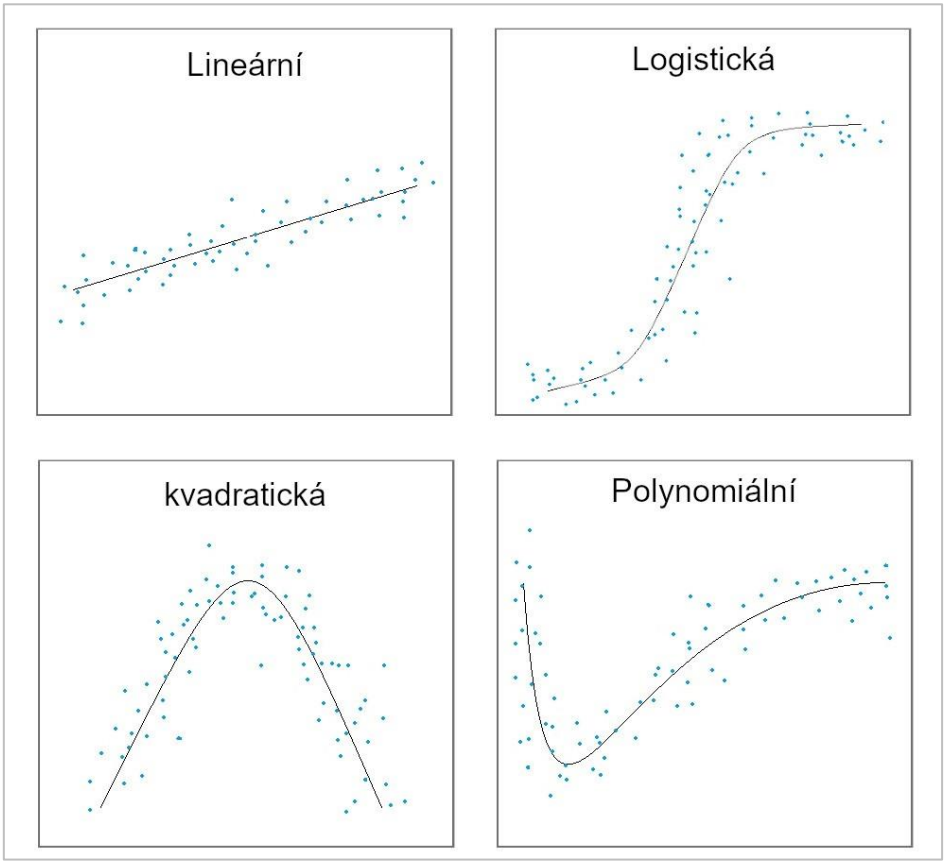


Obrázek 5 – Bodový graf rozpočtu marketingového oddělení a prodaných knih s vypočtenou lineární regresí ve statistickém softwaru SPSS, vlastní zpracování

Výsledný model má podobu: $y = 3\,886 + 19.46x$

Konstanta, tedy hodnota 3 886, představuje počet prodaných knih, pokud by oddělení marketingu nedostalo žádné peníze. Směrnice, tedy hodnota 19.46, představuje o kolik se změní počet prodaných knih při jednotkové změně množství využitelných peněz, v tomto případě každý jeden tisíc korun společnosti přináší navýšení o necelých 20 knih. Po dosazení hodnoty 500 000, které si obchodní oddělení nárokuje, tak vyjde celkem 13 631 prodaných knih. Potvrzuje se tak opoziční názor vůči obchodnímu oddělení, že diskutovaných 500 000,-kč dokáže marketing využít efektivněji a za tuto částku navýší prodej knih o 9 745ks (od spočtené hodnoty dodané modelem je nutné odečíst konstantu – tedy hodnotu, na kterou nemá vliv množství využitých peněz marketingem). Tato data tak vedení využije k tomu, aby mohlo udělat finální rozhodnutí.

Jak ukazuje předchozí příklad, regresní analýzu je vhodné začít grafickým znázorněním, které dokáže leccos prozradit o dostupných datech. Jedním z prvních úkolu tak je potvrzení lineárního souvislosti zobrazených proměnných. Pokud je na první pohled vidět, že se v bodovém grafu tvoří jiná než lineární souvislost, je nutné vyhodnotit, zda se nejedná o závažné porušení linearity. Na obrázku č. 6 je zobrazena matice čtyř datových sad k nimž jsou přiřazeny aproximované funkce, jenž nejlépe odpovídají tvaru dat.



Obrázek 6 – Grafické znázornění aproximací odlišných funkcí mezi 2 proměnnými, vlastní zpracování

Je nutné zdůraznit, že vizualizovaná data a následně pro ně aproximované funkce na obrázku č. 6 jsou účelně sestaveny pro reprezentativnost. Ve skutečnosti se s takto jednoznačně přiřaditelnými funkcemi lze setkat jen zřídka. Grafické zobrazení dat je tak vhodnou pomůckou, jak si získat lepší představu o rozložení datového souboru, ale je vždy vhodné potvrdit či vyvrátit lineárnost výpočtem. Dalším vhodným přínosem vizualizace dat je zjištění odlehlých pozorování. Odlehlé hodnoty je vhodné odstranit (případně jinak zpracovat), jinak mohou mít negativní vliv na vypočtený model. Jelikož lineární regrese pracuje s průměry jednotlivých proměnných, je tak citlivá na tyto extrémní hodnoty. Ze všech měr centrální tendence, tedy mediánu, modální hodnoty a průměru, je právě průměr díky způsobu výpočtu na odlehlé hodnoty citlivý a tento efekt je ještě silnější,

pokud se vychází z malého souboru dat. Avšak tato vlastnost platí i obráceně, pokud je datový soubor dostatečně obsáhlý, malé množství odlehlých hodnot má na sestavený model zanedbatelný vliv (za předpokladu, že odlehlé hodnoty nejsou opravdu extrémní).

S vizualizací zkoumaných dat se v kontextu velikosti datového souboru pojí ještě jeden potenciální problém, a tím je takové množství záznamů, že prakticky vyplní grafové plátno či jeho oblasti barvou zvolenou pro reprezentaci bodů do takového míry, že není poznat rozmístění záznamů. Velikost datového souboru může být opravdu rozmanitá, od jednotek či nízkých desítek, např. z kvalitativního dotazníkového průzkumu, až po soubory se statisíci či miliony záznamy, např. počítačově sbíraná data o nákupních zvycích lidí na e-shopu. Velká data je tak třeba pro vhodnou vizualizaci upravit. Na to existují 2 efektivní způsoby. V obou případech je v rámci jednoho zobrazeného bodu do něj agregováno určité množství dat, které tento bod reprezentuje. U grafu je následně legenda se škálou, která prozrazuje kolik je, v jakém bodu agregováno dat, a to dle zvoleného způsobu:

1) množství dat je reprezentováno velikostí bodu – čím více dat je v jednom bodu agregováno, tím větší je jeho průměr vůči ostatním
2) množství dat je reprezentováno barvou bodu – obvykle sytostí, čím je sytější barva, tím více dat je v daném bodu agregováno.

Jakmile jsou splněny všechny podmínky, je možné sestavit model. Avšak kromě samotné regresní rovnice je vhodné si spočítat i další ukazatele kvality sestaveného modelu, jako jsou ANOVA, *R* a R^2.

6.1.1 ANOVA

ANOVA je zkratkou anglického názvu analysis of variance, tedy analýzy rozptylu, kterou definuje [4] (2017, str. 604) jako statistickou techniku pro zjištění rozdílů mezi průměry dvou a více populací. Využívá se tak pro testování průměrů, kdy nulová hypotéza obvykle tvrdí, že průměry jsou stejné. V tomto kontextu jsou nezávislé proměnné nazývány faktory, které následně dělí ANOVA do 2 skupin, a to jednofaktorovou ANOVA a n-faktorovou ANOVA, která může nabývat 2 a více faktorů. Analýza rozptylu je často využívaná v lineární regresi, protože lze díky ní určit hodnotu R^2, která slouží pro hodnocení modelu. Test analýzy rozptylu se dle [6] (2015, str. 229) zakládá na dělení meziskupinového rozptylu vnitroskupinovým rozptylem. Její výstup má formu tabulky, viz obrázek č. 7, ve které lze nalézt F statistiku, která je dle [6] (2015, str. 233) definována jako podíl variability mezi skupinami a variabilitou uvnitř skupin a její hladinou významnosti.

Při platnosti nulové hypotézy, tedy $R^2=0$, by se tyto hodnoty sobě rovnaly a výslednou hodnotou F by byla 1. Čím je hodnota vzdálenější od 1, tím větší je rozptyl mezi skupinami. V případě lineární regrese se tak porovnává rozptyl regresního modelu a reziduí. Reziduum definuje [1] (2015, str. 247) jako chybu odhadu zvolené proměnné v regresním modelu. Jak zmiňuje [6] (2015, stra. 326), pokud je hranice významnosti nižší než 0,05 nulová, je F statisticky signifikantní. Dostatečně velké F a splnění hranice významnosti je tak vhodným testem užitečnosti regresního modelu.

Variabilita	Součet čtverců	Stupně volnosti	Průměrný čtverec	F statistika	Významnost
Meziskupinová	24 101	1	24 101	46, 78	0,00
Vnitroskupinová	9 273	18	515		
Celkem	33 374	19			

Obrázek 7 – tabulka ANOVA, vlastní zpracování

Pro příklad jednofaktorové ANOVA lze uvést zjišťování managementu autobazaru, zda existuje preference ohledně výkonu motoru vozidla udávanou v hp pro 3 skupiny řidičů – občasný řidič s nájezdem do 2 000 km za rok, aktivní řidič s nájezdem do 10 000 km za rok a řidič z povolání, u kterého se předpokládá roční nájezd přes 10 000 km. Nulová hypotéza by tak tvrdila, že mezi těmito skupinami řidičů neexistuje rozdílná preference na výkon motoru, kterou využívají. Pro zmíněný autobazar by mohlo zjištění rozdílných preferencí těchto skupin řidičů znamenat lepší zacílení na nákup vozidel s požadovaným motorovým výkonem, aby jejich nabídka lépe odpovídala jejich obvyklé klientele, což by mohlo mít pozitivní dopady na výsledný prodej. Nezávislé proměnné v ANOVA musí být kategorického typu.

6.1.2 R a R^2

Jak uvádí [6] (2015, str. 325), hodnota **R** je v jednoduché lineární regresi hodnotou Pearsonova korelačního koeficientu, zde však nabývá pouze kladných hodnot a nemůže tak sloužit jako vyjádření korelačního koeficientu. R díky normalizaci nabývá hodnot od 0 (nekorelovanosti) až do 1 (úplná korelovanost) a reprezentuje tak lineární míru souvztažnosti mezi proměnnými.

R^2 vyjadřuje, jak přesné budou vypočítané hodnoty lineární regresní funkce tím, že jeho hodnota v intervalu od <0;1> popisuje jakou část celkového rozptylu model vysvětluje. [6] (2015, str. 325) tuto hodnotu po vynásobení 100, a tedy převedením na procentuální vyjádření, nazývá koeficientem determinace. Obě verze jsou tak odlišným vyjádřením jednoho a toho samého výsledku, tedy jak velkou část rozptylu závislé proměnné je možné vysvětlit pomocí nezávislé proměnné. Pokud by tak v extrémním případě existovala úplná korelace, znamenalo by to, že všechny body leží přesně na vypočtené přímce, a tudíž by bylo R^2 rovno jedné, a tak by byl vysvětlen celý, tedy nulový, rozptyl. To je však v praxi zcela vzácný jev a většinou je nutné počítat s chybou, která představuje nevysvětlenou část rozptylu. Tyto chyby mohou mít různý původ, jako je chyba měření nezávislé proměnné nebo souvztažnost jiných proměnných. Hodnota R^2 se tak dá popsat dle [4] (2017, str. 637) následující rovnicí:

$$R^2 = \frac{vysvětlovaný\ rozptyl}{celkový\ rozptyl} = \frac{SS_x}{SS_y}$$
$$= \frac{celkový\ rozptyl - variační\ chyba}{celkový\ rozptyl} = \frac{SS_x - SS_{error}}{SS_y}$$

SS = rozptyl (Sum of Squares)

Příklad č. 5

Ve smyšleném příkladu hledá společnost zabývající se prodejem pleťových krémů způsob, jak rozšířit podíly na trhu. Management po celodenní diskusi a rozhodování vybral dvě potencionální nové řady, které by měli podílům pomoci, avšak mají finanční zdroje pro uvedení pouze jedné řady. Buď bude na trh uvedena nový řada s kyselinou hyaluronovou, nebo s koenzymem Q10. Obě řady by byli časově i nákladově obdobně náročné pro uvedení, management tak požaduje zjistit, která z uvedených látek by měla lepší prodejní potenciál. K dispozici jsou zakoupená prodejní data všech konkurenčních krémů, které obsahují tyto látky jako svou hlavní složku a je potřeba zjistit, zda množství těchto látek (nezávislá proměnná) má vliv na jejich celkový prodej (závislá proměnná). Pro každou z nich se tak sestaví lineární regresní rovnice a zjistí se, která látka je vhodnější. V obou případech se jedná o 50ml krémy s 5% zastoupením aktivní látky.

Koenzym Q10	**Kyselina hyaluronová**
Z vypočteného modelu bylo získáno R = 0.784, tedy R^2 = 0.61. To jsou poměrně vysoké hodnoty. Výsledek tak říká, že 61 % rozptylu celkového prodeje zkoumaných krémů je vysvětleno tím, že obsahuje 2,5ml koenzymu Q10. Zbylých 39 % jsou všechny ostatní faktory, jako je vliv reklamy, cenových akcí a další, případně chyba měření. Pro koenzym Q 10 je tak možné prohlásit, že je tato látka u lidí populární a její samotný obsah v krému zajistí dostatečně velké prodeje, aniž by bylo potřeba investovat větší sumy peněz do reklamy či jiných metod podpory prodeje.	Z vypočteného modelu bylo získáno R = 0.457, tedy R^2 = 0.21. To nejsou příliš přesvědčivé hodnoty. Výsledek tak říká, že pouhých 21 % prodeje těchto krémů je vysvětleno zastoupením 2,5ml této aktivní látky, je zde tak stále nevysvětleno 79 % celkového prodeje. Pro kyselinu hyaluronovou je tak možné učinit závěr, že krémy s touto látkou potřebují další formy podpory prodeje a tato látka samotná není příliš velkým lákadlem pro zákazníky.

Managementu je tak možné z dostupných dat doporučit novou řadu s koenzymem Q10, protože krémy s jejím obsahem mají značnou část prodejnosti zajištěnou právě díky jejímu výskytu v krému, oproti kyselině hyaluronové, která prodejnosti vděčí spíše jiným, zde nezjištěným vlivům.

6.2 Mnohonásobná lineární regrese

V mnoha případech je k analýze sledovaného jevu jedna nezávislá proměnná nedostačující a je nutné do regrese zahrnout další vstupy, o kterých je odůvodněné se domnívat, že mají na závislou proměnnou znatelný vliv. Mnohonásobnou lineární regresi je tak možné získat přidáním dodatečných nezávislých proměnných do jednoduché lineární regrese. Díky tomu je možné do lineární regrese zahrnout co nejvíce vlivů, pro které jsou dostupná data. Je tak možné vyčíslit nejen o kolik se změní prodej aut při konkrétní změně ceny benzínu, ale i o kolik se změní výsledný prodej při změně ceny benzínu, průměrné mzdy, počtu obyvatel a dalších nezávislých proměnných, které mohou mít na závislou proměnnou dopad. V rámci managementu a řízení je jen zřídka možné dostatečně vysvětlit sledovaný jev pomocí jediné proměnné, a tak mnohonásobná lineární regrese rozšiřuje tuto možnost o vložení libovolného množství nezávislých proměnných. Výsledná funkce mnohonásobné lineární regrese je tak podobná její jednoduché verzi, stále obsahuje konstantu, pouze se pro každou novou proměnnou vypočítá její korelační koeficient. Každá proměnná tak skrze svůj koeficient, také nazvaný jako parciální, má svůj podíl na výsledku regrese:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

y – závislá proměnná
n – počet nezávislých proměnných
x_1 až x_n – hodnota nezávislých proměnných,
β_0 – konstanta určující průsečík přímky s osou y
β_1 až β_n – koeficienty nezávislých proměnných

Při vhodném využití více proměnných je možné vysvětlit větší část rozptylu závislé proměnné, než by tomu bylo při využití jednoduché lineární regrese. Avšak zavedení více proměnných do lineární regrese s sebou přináší novou problematiku a podmínky, které je nutné splnit pro sestavení vhodného modelu pro účely managementu. Výčet těchto podmínek se dá shrnout dle textu [6] (2015, str. 354) následovně:

1. Závisle proměnná Y musí být proměnná metrická
2. Nezávisle proměnné jsou měřeny rovněž na intervalové úrovni. Pokud tuto podmínku nesplňují, je možné využít umělých proměnných (více v kapitole 5.2.2 umělé proměnné)
3. Nezávislé proměnné by neměly být mezi sebou příliš korelovány. Tím by došlo k porušení podmínky na absenci multikolinearity, která má za

následek nespolehlivé výsledky regrese a může shledat jinak vhodnou nezávislou proměnnou jako statisticky nevýznamnou

4. V datech se nevyskytují odlehlé hodnoty, na které je lineární regrese citlivá
5. Proměnné jsou v lineární vztahu. Jelikož je vícenásobná lineární regrese založena na Pearsonově korelačním koeficientu, nelineární vztahy zůstanou neodhaleny
6. Vztahy mezi proměnnými vykazují homoskedasticitu, tedy homogenitu rozptylu. Což znamená, že rozptyl v datech jedné proměnné bude víceméně shodný pro všechny hodnoty druhé proměnné. Například pokud bude rozptyl v příjmech shodný pro všechny věkové skupiny, pak mezi věkem a příjmem bude existovat homoskedasticita. Opakem homoskedasticity je heteroskedasticita

Část těchto podmínek je dostatečně sebevysvětlujících nebo byly vysvětleny již v předchozím textu, než aby jim byla věnována další pozornost, avšak na některé je vhodné se podívat blíže, a to na problematiku multikolinearity a umělých proměnných.

Při využití mnohonásobné lineární regrese je také kromě testu užitečnosti regresního modelu probíraného v části ANOVA vhodné při hodnocení modelu provést testy hodnot jednotlivých koeficientů využitých v modelu. Jak popisuje [4] (2017, str. 656), na rozdíl od testování užitečnosti modelu, kde se využívá F statistika, se pro testování jednotlivých koeficientů využívá t testu.

Pro jednotlivé koeficienty se u mnohonásobné lineární regrese vypočítává standardizovaná verze koeficientů, také označována jako *beta*. Jak udává [4] (2017, str. 653), ty jsou vypočteny standardizací všech proměnných tak, aby jejich průměr odpovídal 0 a rozptyl 1 ještě před sestavením regresního modelu. Výsledná hodnota tak díky standardizaci do jednotného intervalu <-1;1> umožňuje srovnat míru souvstažnosti jednotlivých nezávislých proměnných vůči proměnné závislé. Vyšší hodnota znamená větší vliv na závislou proměnnou.

6.2.1 Multikolinearita

Multikolinearita označuje stav vysoké korelace mezi jednotlivými nezávislými proměnnými. Popis problematiky multikolinearity a její příklady jsou vysvětleny v [4] (2015, str. 661-672), ze kterého vychází následující seznam potencionálních problému, které může její výskyt způsobit:

1) Parciální regresní koeficienty nebudou správně odvozeny. Je velká pravděpodobnost vysoké chyby odhadu.
2) Velikost, stejně tak jako znaménko parciálních regresních koeficientů se mohou měnit mezi jednotlivými vzorky
3) Je náročnější odvodit relativní vliv nezávislé proměnné na vysvětlení rozptylu závislé proměnné
4) Nezávislé proměnné mohou být nesprávně zahrnuty nebo odstraněny při metodě vkládání dat *stepwise* (stepwise je metoda postupného vkládání proměnných do modelu řešeného v dalším oddíle)

Mezi jednotlivými nezávislými proměnnými bude téměř vždy existovat jistá míra korelace, která však bývá z pohledu vlivu na sestavení regresního modelu zanedbatelná. Otázkou tak zůstává, kdy už je vliv multikolinearity dostatečně velký na to, aby ovlivnil kvalitu modelu a je nutné se jí zabývat a odstranit její negativní dopady. Bohužel neexistuje jedna konkrétní metoda, která by jednoznačně určila, zda se jedná či nejedná o multikolinearitu, je tak zapotřebí využít nápomocných ukazatelů a aplikovat je na konkrétní datovou sadu.

Existuje více způsobů, jak nalézt multikolinearitu v datech, jako je grafické znázornění dat, výpočet jednoduché párové korelace, nebo ukazatel VIF (z anglického *Variance Inflation Factor*, neboli faktor zvětšení rozptylu). I tyto jednotlivé metody mají svá úskalí, například jaká hodnota VIF, kterou obvykle vypočítává software, by již mohla znamenat vliv multikolinearity na model se výzkumníci dle [4] (2015, str. 369) jednoznačně neshodnou, kdy někteří uvádějí jako hodnotu větší než 10 za problematickou, jiní hodnotu nad 5. Pro zjištění multikolinearity za pomoci grafického znázornění či párových korelací se využívá matice, aby bylo možné zobrazit každý pár nezávislých proměnných zvlášť. Oba tyto způsoby jsou zobrazeny v následujícím příkladu č. 6, ve kterém je využit statistický software SPSS od IBM. Pro účely tohoto příkladu jsou tak využity procedury pro výpočet korelační matice, korelačního koeficientu a zobrazení grafů. Detailnější popis využití softwaru SPSS a těchto procedur je uveden v praktické části práce.

Příklad č. 6

Ve smyšleném příkladu vedení obce Hraběholy požaduje sestavení lineárního regresního modelu pro zjištění spokojenosti rezidentů v obci. Na obrázku č. 8 jsou k dispozici údaje o věku, IQ, měsíčním příjmu a subjektivní spokojenosti měřené na škále 1-10 (1 zcela nespokojený, 10 zcela spokojený) zjištěné od rezidentů, kteří v obci bydlí alespoň 5 let. Kromě samotného modelu, který je schopen s jistou mírou pravděpodobnosti předpovědět očekávanou spokojenost na základě těchto nezávislých proměnných, je požadováno zjistit, jaký vliv na spokojenost mají jednotlivé nezávislé proměnné

Věk	IQ	Měsíční příjem (v tis. Kč)	Spokojenost
22	98	22	8
22	107	24	6
24	104	21	5
25	94	26	7
25	124	25	5
27	129	25	3
28	101	32	5
29	102	28	6
30	89	24	9
30	111	31	4
30	104	34	5
32	124	30	3
33	104	27	5
33	102	34	6
37	92	45	8
37	124	38	4
38	117	35	6
38	99	40	8
40	104	45	7

Obrázek 8 – Tabulka spokojenosti rezidentů obce Hraběholy, vlastní zpracování

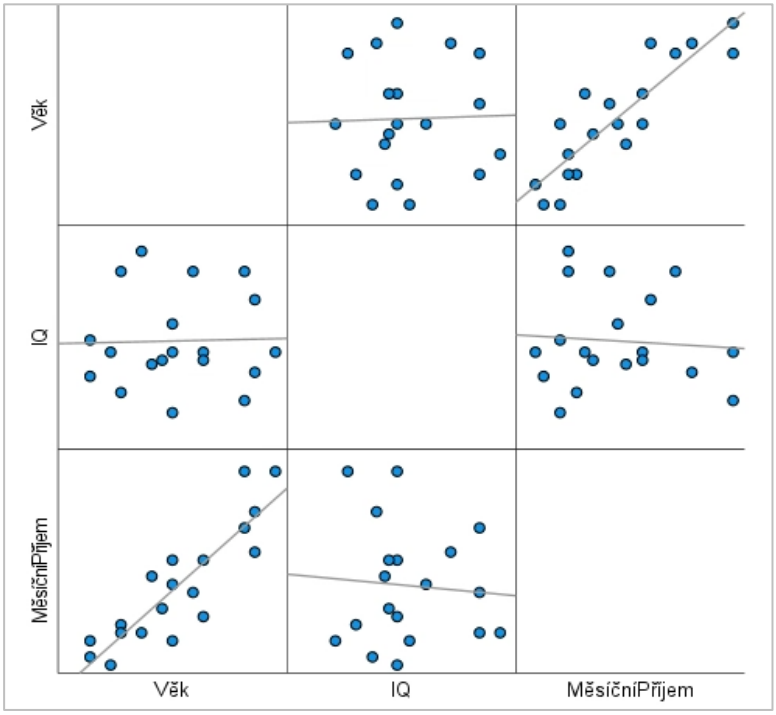
Korelace

		Věk	IQ	Měsíční příjem
Věk	Pearsonův korelační koef.	1	,028	,878**
	Významnost (oboustranná)		,909	,000
	Počet záznamů	19	19	19
IQ	Pearsonův korelační koef.	,028	1	-,077
	Významnost (oboustranná)	,909		,755
	Počet záznamů	19	19	19
Měsíční příjem	Pearsonův korelační koef.	,878**	-,077	1
	Významnost (oboustranná)	,000	,755	
	Počet záznamů	19	19	19

**. Korelace je významná na hranici 0.01 (oboustranná).

Obrázek 9 – Korelační matice pro datovou sadu obce Hraběholy v SPSS, vlastní zpracování

Na obrázku č. 9 s korelačními koeficienty je vidět značná korelace mezi věkem a měsíčním příjmem 0,877, stejně tak je vidět tato souvislost v grafu na obrázku č. 10. S vyšším věkem tak nejspíše přibývá zkušeností a zaměstnavatelé jsou ochotni tuto zkušenost lépe zaplatit. 0.877 je dostatečně vysoká hodnota na to, aby bylo možné předpokládat vliv multikolinearity. Co z toho však plyne pro vyhodnocení a následně sestavený model? Protože jsou měsíční příjem a věk silně korelované, je složité pro ně vyhodnotit korelační koeficienty – obě proměnné tak přináší do modelu téměř totožnou informaci, avšak obě zvyšují směrodatnou odchylku, která se projeví sníženou důvěryhodností modelu, kdy se modelu dostatečně nenavýší hodnota R^2 na to, aby se vykompenzoval touto proměnnou navýšený rozptyl. Vhodným řešením by tak mohlo být vypuštění jedné z korelovaných nezávislých proměnných, v tomto případě tak například proměnné věk a ponechá se měsíční příjem. Změny je možné pozorovat na následujících obrázcích č. 11 kde jsou vypočítány koeficienty pro všechny nezávislé proměnné a obrázku č. 12 kde jsou vypočteny koeficienty bez nezávislé proměnné věk.



Obrázek 10 *– Párový bodový graf pro datovou sadu obce Hraběholy v SPSS, vlastní zpracování*

Koeficienty[a]

Model		Nestandardizované koeficienty B	Nestandardizované koeficienty Standardní chyba	Standardizované koeficienty Beta	t	Významnost
1	(Konstanta)	17,493	2,607		6,710	,000
	IQ	-,122	,022	-,830	-5,660	,000
	Věk	,061	,093	,198	,648	,527
	Měsíční příjem	-,016	,072	-,068	-,221	,828

a. Závislá proměnná: Spokojenost

Obrázek 11 – Vypočtené koeficienty pro datovou sadu města Hraběholy v SPSS, vlastní zpracování

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	17,766	2,526		7,033	<,001
	IQ	-,119	,021	-,811	-5,749	<,001
	MěsíčníPříjem	,025	,033	,108	,763	,456

a. Dependent Variable: Spokojenost

Obrázek 12 – Vypočtené koeficienty pro datovou sadu města Hraběholy v SPSS, vlastní zpracování

Proměnná IQ je touto změnou téměř nedotčena, avšak změny v měsíčním příjmu už jsou znatelnější. Nejen že došlo k menší číselné změně koeficientu, ale také se téměř o polovinu zmenšila hodnota významnosti, která dává vyšší důvěru v model při splnění stanoveného intervalu spolehlivosti. Ve většině případů se za dostatečně významnou hladinu považuje 0.05, pokud je tak významnost <0.05, je možné výsledný koeficient považovat za důvěryhodný. Samozřejmě tak hodnota 0.456 je stále velmi vysoká a nedá se považovat za spolehlivou. Nalezené koeficienty by nebyli v praxi příliš přínosné, nicméně samotné takto výrazné snížení významnosti je dostačujícím důkazem vlivu multikolinearity nezávislých proměnných v modelu.

Jak tedy zvolit vhodnou kombinaci proměnných do modelu? Pokud je model sestavován z velkého množství proměnných, kde se dá očekává jistá míra multikolinearity, dalo by se nadneseně říct, že se z každé vysoce korelované skupiny proměnných hledá ta, která nejlépe vystihuje jejich vliv na sledovanou proměnnou, viz. předchozí příklad vlivu proměnných věk a měsíční příjem na spokojenost, kdy je spíše pravděpodobné, že je člověk spokojený díky vyššímu příjmu, než že je starší.

Po kontrole multikolinearity a ujištění se, že data splňují všechny předpoklady nutné pro vícenásobnou lineární regresi je možné sestavit model. Prvním krokem je zvolení způsobu vložení nezávislých proměnných do modelu, který může mít na sestavený model vliv. [6] (2015, str. 358) tak uvádí 3 možné postupy vložení proměnných: „

1) **Metoda standardní** (tzv. metoda *enter)* – Všechny proměnné, které analytik určí, jsou do výpočtu vloženy najednou
2) **Metoda postupného vkládání** (*stepwise*) – Proměnné jsou vkládány do výpočtu regrese postupně podle předem zadaných matematických kritérií. V této metodě výzkumník nekontroluje pořadí proměnných, jak postupně vstupují do analýzy, o pořadí nerozhoduje SPSS či jiný software – to je algoritmus výpočtu a kritéria vkládání. Je to metoda, které se s trochou nadsázky říká metoda pro nalezení „nejlepšího" modelu.
3) **Metoda hierarchická** (*blocks*) – Pořadí, v němž proměnné vstupují do výpočtu, řídí výzkumník, proměnné se postupně přidávají po skupinách (blocích) a odvíjí se od jeho kauzálního modelu, který testuje."

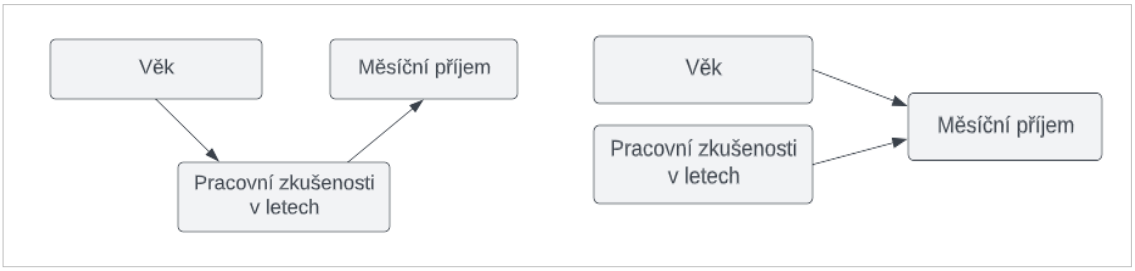
Dále vysvětluje jejich vhodné využití, kdy standardní metoda je díky vložení všech proměnných najednou vhodná k vysvětlení vlivů jednotlivých nezávislých proměnných na vysvětlované proměnné, tedy jak velkou část rozptylu každá tato proměnná vysvětlí.

Metoda postupného vkládání, v literatuře občas popisovaná jako metoda k nalezení „nejlepšího" modelu. Jejím principem je sestavení tolika modelů, kolik je proměnných. Nejdříve se vytvoří model s jednou proměnnou, vyhodnotí, přidá se nová proměnná, vyhodnotí a postupuje se až do vyčerpání nezávislých proměnných. Výsledkem je tak tabulka s výčtem modelů, ve kterých jsou dobře vidět změny hodnot modelu při zavedení nové proměnné.

Metoda hierarchická dovoluje výzkumníkovi zvolit pořadí vkládaných proměnných. To umožňuje zkoumat, zda nově přidaná proměnná měla vliv na výsledné R^2 a tím, zda mezi nově přidanou a vysvětlovanou proměnou existuje nějaká souvislost, či je zprostředkován tzv. intervenující proměnou, která byla řešená v části 4.1.3 parciální korelace. Díky postupnému vkládání je tak možné odhalit problematické proměnné, jak je ukázáno na příkladu č. 7

Příklad č. 7

Ve smyšleném příkladu využívá personální agentura převzatý lineární regresní model pro určení mzdového rozpětí klientů. Z modelu však dostává neuspokojivé výsledky vůči realitě a potřebuje model upravit. Jeden z analytiků naznačil, že může být problém v rámci nezávislých proměnných věk a pracovních zkušenostech v letech, které jsou nejspíše korelovány. Je tak zapotřebí otestovat, zda tomu tak skutečně je a doporučit následný postup. K otestování je možné nechat sestavit tentýž model znovu, avšak za pomoci hierarchické metody a nezávislou proměnnou věk přidat jako poslední a určit závěry dle změn v modelu.



Obrázek 13 – Diagramy intervenující proměnné pracovní zkušenosti v letech (vlevo) a neintervenující proměnné pracovní zkušenosti v letech (vpravo), vlastní zpracování

Scénář 1 **Neexistence intervenující proměnné**	**Scénář 2** **Existence intervenující proměnné**
Pokud by se modelu po přidání nezávislé proměnné věk zvýšilo R^2, znamenalo by to, že informace o věku vysvětlila část rozptylu, která byla v modelu před jejím vložením nevysvětlena. Zvětšení R^2 tak úměrně zmenšuje zbylou část rozptylu, která zůstává nevysvětlena a odpovídá 1 – R^2, to znamená že je v modelu využita oprávněně a je potřeba hledat zdroj problému jinde.	Pokud by se modelu po přidání nezávislé proměnné věk nezvýšilo R^2, nebo se navýšilo pouze zanedbatelně, a zároveň by byla potvrzena souvislost mezi věkem a pracovními zkušenostmi v letech, proměnná pracovní zkušenosti v letech by byla označena za intervenující a následně odstraněna z modelu. Po této úpravě by se znovu spočítali problematické hodnoty a pokud by došlo k požadovanému zpřesnění, bylo by možné problém uzavřít.

6.2.2 Umělé proměnné

Výše příjmu, počet sestavených výrobků za hodinu, náklady na reklamu, spotřeba nafty u vozidla a mnoho dalších proměnných jsou standardně měřitelné hodnoty. Co když je ale do modelu potřeba zahrnout proměnné, které nejsou normálními způsoby měřitelné a mají spíše kvalitativní charakter, např. výše dosaženého vzdělání nebo míra spokojenosti zákazníků. Jedná se tak o proměnné, které by bylo vhodné přidat do modelu, ale nemusí být z datového souboru přímo k dispozici. Ty často nabývají 2 podob:

1) kategoriální proměnné – s omezeným počtem přípustných hodnot. příkladem může být kategorizace věku do několika skupin (nezletilí, mladí, dospělí, starší) s definovanými hranicemi pro začlenění do předem stanovených skupin, míra dosaženého vzdělání (základní, středoškolské, univerzitní) a jiné.
2) dichotomické proměnné – nabývající jen a pouze dvou hodnot, jako je například pohlaví (muž a žena) nebo příslušnost ke konkrétní skupině (sportovec či nesportovec).

Transformace dat do umělých proměnných je ze své podstaty ztrátová operace (za předpokladu, že se transformací nemyslí pouze změna názvu či obdobné přemapování proměnných do alternativních názvů), kdy v průběhu transformace dochází k jejich částečnému znehodnocení.

Umělé proměnné jsou velmi flexibilním nástrojem, jak obohatit model o nová data. Vždy je ale potřeba pečlivě zvážit jejich přínos a rizika. Díky této flexibilitě se však dá jen stěží popsat vhodný postup, jak tyto proměnné vytvářet a používat. V oblasti managementu je tak možné je využít pro modely predikující prodejnost výrobků v rámci předem stanovených věkových skupin, jejichž rozsah se může lišit dle prodejních požadavků. Při přípravě ročních prodejních plánů je možné rozšířit datovou sadu o sezónnost specifickou pro oblast podnikání. Pokud tak dochází k výraznému navýšení prodejů na začátku prázdnin a měsíc před Vánoci, může být vhodné stanovit datová rozmezí a označit prodejní dny jako sezónní či nesezónní. Reprezentace výsledků regresní analýzy v rámci sezónních a mimo sezónních dní budou jednodušší a uchopitelnější, než je popisovat na úrovni jednotlivých dní, nebo naopak tuto informaci vynechat a přijít tak o možnost zkvalitnění výsledků. Umělé proměnné je tak možné v lineární regresní analýze využít kdykoliv je to pro účely managementu vhodné, jen je třeba se ujistit, že jsou použity správně a výsledný model zkvalitňují.

7 Realizace analýzy v SPSS

V této praktické části je realizována analýza nad dvěma datovými soubory v softwaru SPSS. Nejdříve je představen samostatný software SPSS a jeho komponenty potřebné k provedení korelační a regresní analýzy za účelem jejich využití v managementu. Tato část je převážně vlastní prací a veškeré interpretace a závěry provedené jak v průběhu analýzy, tak i v závěrech této kapitoly jsou tak snahou o co nejlepší využití znalostí získaných z nastudované literatury, shrnuté v teoretické části této práce. V oddílech 7.2 je využit datový soubor [11] kvality betonu v příloze č. 1. Tato část se zabývá především ukázkou provedení datové analýzy pomocí softwaru SPSS a je doplněna o obrázky ukazující, jak v SPSS nastavit požadované procedury, v rámci prostoru tak nejsou jednotlivé výsledky podrobněji zkoumány. V oddíle 7.3 využívající datový soubor [12] zabývajícího se množstvím vypůjčených kol ve městě Soul je pak věnována pozornost především datové analýze, dílčím vyhodnocením a možným inferencím z nalezených výsledků, které najdou využití v managementu a datový soubor je blíže představen na začátku tohoto oddílu a popsán v příloze č. 2.

7.1 O SPSS

Následující text vychází z oficiálních webových stránek společnosti IBM, která je aktuálním vlastníkem softwaru SPSS [8]. Jedná se o velmi rozšířený univerzální statistický softwarový nástroj. Na webu spss.com.hk v sekci history [9] je možné zjistit, že vznik sahá až do roku 1968, kdy tři studenti Norman H. Nie, C. Hadlai (Tex) Hull and Dale H. Bent vyvinuli software, který měl za využití statistiky přeměnit hrubá data do vhodných informativních zdrojů, na jejichž základě je možné dělat důležitá rozhodnutí. To, co původně sloužilo jako nástroj pro Stanfordskou univerzitu se však brzy ukázalo jako žádány produkt napříč akademickou sférou a potažmo i v komerčním prostoru. Postupem času a s příchodem digitální doby tak SPSS rostlo dále, až v roce 2009 IBM oznámilo jeho akvizici, pod kterým je dnes licencováno.

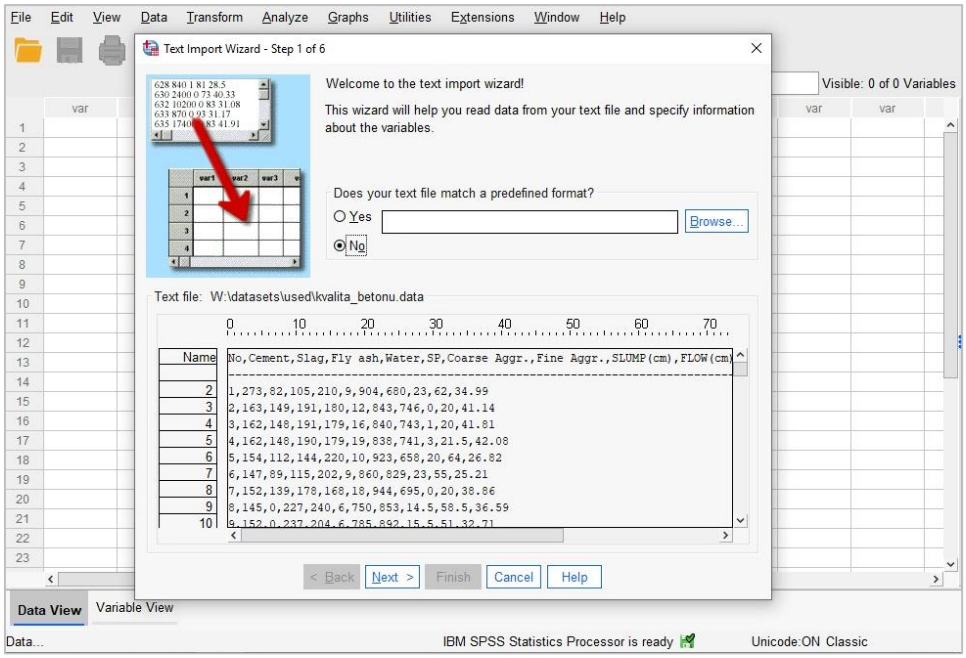
Název je zkratkou pro Statistical Package for the Social Science a i přesto, že se dodnes používá, jedná se spíše i zažité jméno, jelikož jeho dnešní možnosti přesahují hranice referované v oněm názvu. Vyniká tak v oblastech statistiky, těžení dat a data science. Velké oblibě bezesporu přispívá fakt, že dává možnost využití těchto poměrně náročných disciplín i lidem, kteří nemají programátorské zkušenosti či jsou v této oblasti začátečníky. S tím souvisí i jednoduché GUI, či různé pomocné tutoriály na oficiálním webu, které celý proces naučení se využití SPSS velmi ulehčí.

7.2 Ukázka způsobu provedení datové analýzy s SPSS

V následující ukázce je provedena analýza dat za využití korelační a lineární regresní analýzy nad souborem dat obsahujícím komponenty a cílové vlastnosti betonu [11]. Zaměření je především na způsob provedení a nastavení softwaru SPSS pro provedení analýzy než na podrobnější hodnocení jejích výsledků, které je cílem příkladu v oddíle 7.3 Realizace vybraných metod analýzy dat pro účely managementu. Datový soubor obsahuje 7 vstupních proměnných a 3 cílové proměnné, konkrétně sledované vlastnosti betonu pevnost v tlaku, průtok a sednutí. Popis souvztažností jednotlivých komponent betonu a tvorba modelu předpovídající požadované vlastnosti má pro management četná využití, jako jsou časové i finanční úspory díky sestavenému lineárnímu regresnímu modelu, který na základě několika testů je schopen dopočítat potřebné poměry komponent pro cílové vlastnosti. Více informací je možné zjistit v příloze č. 1. Pro všechny následné ukázky a výpočty Je využita verze SPSS 28.0.1.0(142) při využití licence FIM UHK platné do 01.01.2032

7.2.1 Import a úprava dat

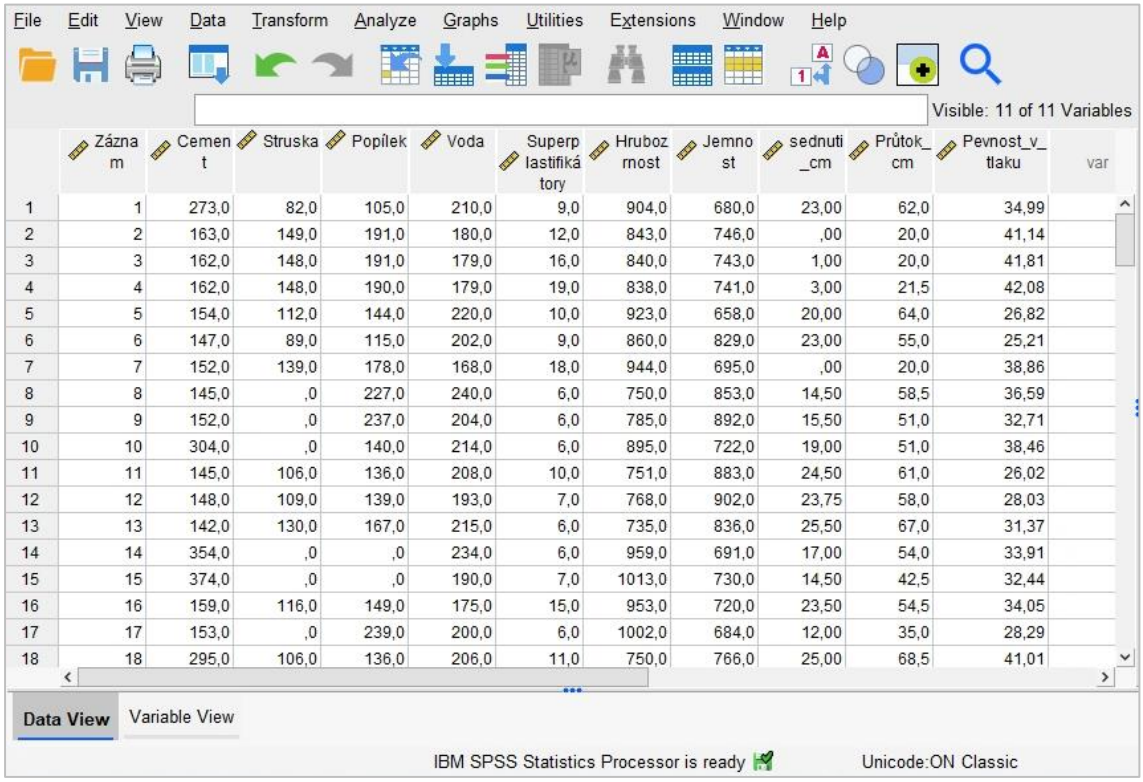
Prvním krokem pro využití softwaru SPSS a jeho statistických procedur je načtení dat. SPSS import dat usnadňuje svým průvodcem importu, který je možné otevřít skrze menu *File* -> *Open* -> *Data…* a celý proces tak činí jednoduchým i bez předchozích zkušeností práce s SPSS. Na obrázku č. 13 je vidět 6-kroková verze pro data typu .csv nebo .data. V jednotlivých krocích se doplní a nastaví vlastnosti importovaných dat, jako je způsob oddělení sloupců, typ znaku oddělující desetinná místa či zda první (či více) řádek obsahuje názvy atributů. Průvodce zároveň zobrazuje náhled, aby bylo možné špatná nastavení odhalit ještě před zpracováním a celý proces tak nebylo nutné opakovat. Do SPSS je možné načíst data v různých formátech, ať už to je to často využívaný formát .csv, tabulkové typy jako je .xlsx nebo jiné, méně používané formáty. Na obrázku č. 14 je tak vidět první krok průvodce s náhledem na importovaná data.



Obrázek 14 – průvodce importem dat v SPSS pro formát .data, vlastní zpracování

SPSS je nápomocné s rozčleněním a správným nastavením dat, které se snaží automaticky nastavit dle jejich obsahu. Avšak je možné, že jsou v datech obsažené neodpovídající hodnoty, udělala se v průběhu průvodce importem dat chyba nebo je struktura dat taková, že ho SPSS bez manuální pomoci nedokáže rozpoznat. Příkladem je v různých proměnných využití odlišných oddělovačů desetinných hodnot, pokud jedna proměnná využívá čárku jako oddělovač a druhá tečku, je nutné takovéto rozdíly manuálně sjednotit.

Po importu datové sady se zobrazí tabulkový list s daty. SPSS nabízí 2 hlavní pohledy, které je možné přepínat v levém dolním rohu tabulátorového editoru. Pohled na data (Data View, obrázek č. 15), ve kterém každý řádek reprezentuje jeden záznam z načteného datového souboru. V tomto pohledu je možné se podívat a upravovat záznamy v datové sadě, řadit záznamy dle obsahu konkrétních sloupců (pravý klik na požadovaný sloupec a možnosti *Sort Ascending/Descending*) a operace obdobného charakteru.



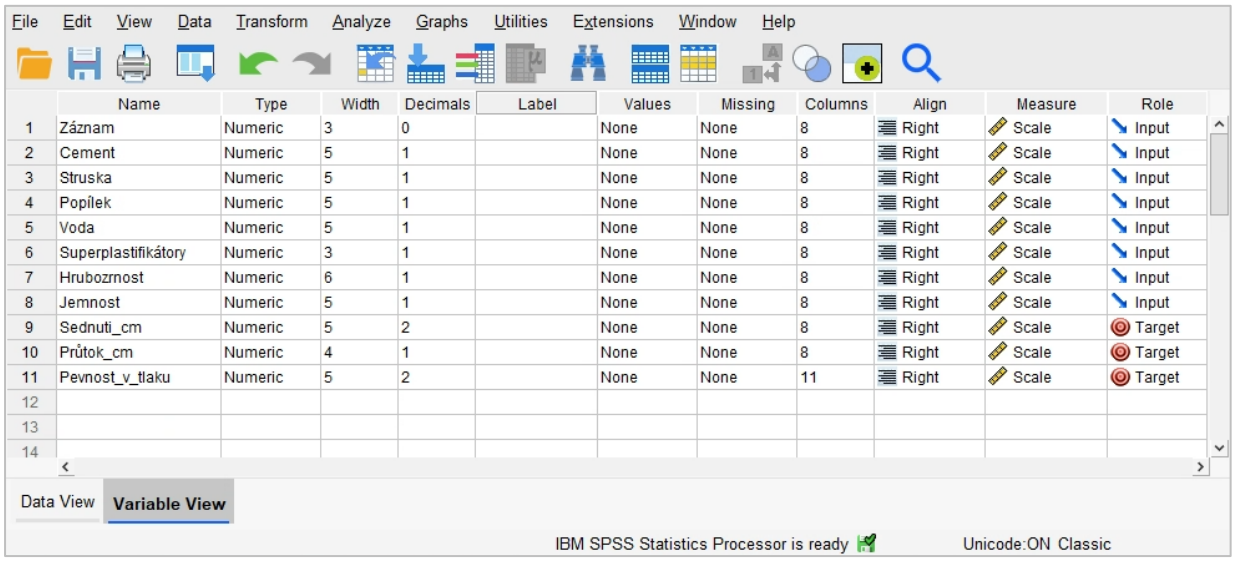
	Zázna m	Cemen t	Struska	Popílek	Voda	Superp lastifiká tory	Hruboz rnost	Jemno st	sednuti _cm	Průtok_ cm	Pevnost_v_ tlaku	var
1	1	273,0	82,0	105,0	210,0	9,0	904,0	680,0	23,00	62,0	34,99	
2	2	163,0	149,0	191,0	180,0	12,0	843,0	746,0	,00	20,0	41,14	
3	3	162,0	148,0	191,0	179,0	16,0	840,0	743,0	1,00	20,0	41,81	
4	4	162,0	148,0	190,0	179,0	19,0	838,0	741,0	3,00	21,5	42,08	
5	5	154,0	112,0	144,0	220,0	10,0	923,0	658,0	20,00	64,0	26,82	
6	6	147,0	89,0	115,0	202,0	9,0	860,0	829,0	23,00	55,0	25,21	
7	7	152,0	139,0	178,0	168,0	18,0	944,0	695,0	,00	20,0	38,86	
8	8	145,0	,0	227,0	240,0	6,0	750,0	853,0	14,50	58,5	36,59	
9	9	152,0	,0	237,0	204,0	6,0	785,0	892,0	15,50	51,0	32,71	
10	10	304,0	,0	140,0	214,0	6,0	895,0	722,0	19,00	51,0	38,46	
11	11	145,0	106,0	136,0	208,0	10,0	751,0	883,0	24,50	61,0	26,02	
12	12	148,0	109,0	139,0	193,0	7,0	768,0	902,0	23,75	58,0	28,03	
13	13	142,0	130,0	167,0	215,0	6,0	735,0	836,0	25,50	67,0	31,37	
14	14	354,0	,0	,0	234,0	6,0	959,0	691,0	17,00	54,0	33,91	
15	15	374,0	,0	,0	190,0	7,0	1013,0	730,0	14,50	42,5	32,44	
16	16	159,0	116,0	149,0	175,0	15,0	953,0	720,0	23,50	54,5	34,05	
17	17	153,0	,0	239,0	200,0	6,0	1002,0	684,0	12,00	35,0	28,29	
18	18	295,0	106,0	136,0	206,0	11,0	750,0	766,0	25,00	68,5	41,01	

Obrázek 15 – Pohled na data v SPSS, vlastní zpracování

Druhý pohled na jednotlivé proměnné (Variable View, obrázek č. 16), kde každý řádek představuje jednu proměnnou a ve sloupcích jsou jednotlivé atributy, které jsou dle oficiální dokumentace [10] následující:

- Name – název proměnné
- Type – datový typ (textový, číselný, datový a další)
- Width – šířka (u číselné hodnoty počet číslic, u textu počet znaků)
- Decimals – počet desetinných míst
- Label – alternativní název pro proměnnou dovolující používat mezery a další znaky, které by jinak porušovali podmínky pro validní název proměnné
- Values – možnost mapování alternativních textů pro jednotlivé hodnoty v proměnné
- Missing – možnost definice hodnot (nebo rozpětí hodnot), s kterými bude následně zacházeno, jako by v datové sadě nebyli. Hodnoty nejsou odstraněny, pouze se s nimi pracuje jako s chybějícími
- Columns – definice šířky sloupce v Data View
- Align – možnost centrování textu v Data View doleva (Left), na střed (Center) nebo doprava (Right)
- Measure – typ proměnné, tedy nominální (Nominal), ordinální (Ordinal) a intervalová (Scale)

- Role – role proměnné, zde je typů více ale nás zajímají především dva, a to vstupní, které představují nezávislé proměnné (Input) a výstupní, které představují závislou proměnnou (Target)

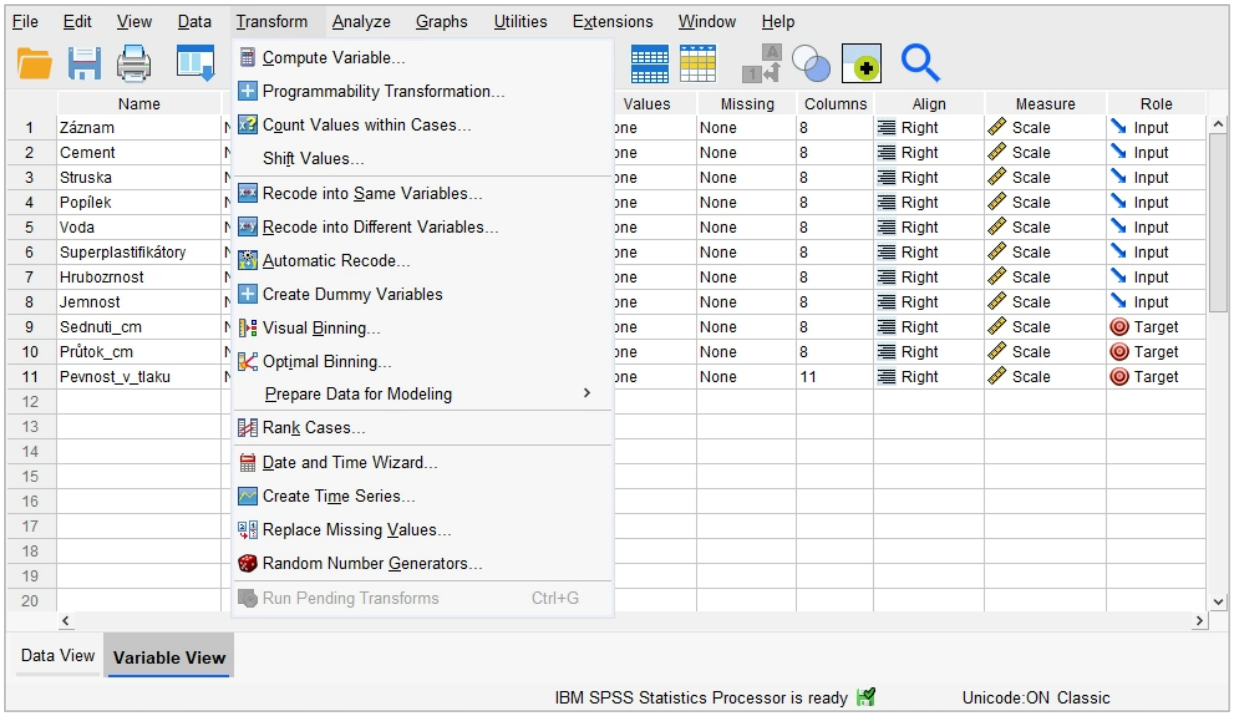


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Záznam	Numeric	3	0		None	None	8	Right	Scale	Input
2	Cement	Numeric	5	1		None	None	8	Right	Scale	Input
3	Struska	Numeric	5	1		None	None	8	Right	Scale	Input
4	Popílek	Numeric	5	1		None	None	8	Right	Scale	Input
5	Voda	Numeric	5	1		None	None	8	Right	Scale	Input
6	Superplastifikátory	Numeric	3	1		None	None	8	Right	Scale	Input
7	Hrubozrnost	Numeric	6	1		None	None	8	Right	Scale	Input
8	Jemnost	Numeric	5	1		None	None	8	Right	Scale	Input
9	Sednuti_cm	Numeric	5	2		None	None	8	Right	Scale	Target
10	Průtok_cm	Numeric	4	1		None	None	8	Right	Scale	Target
11	Pevnost_v_tlaku	Numeric	5	2		None	None	11	Right	Scale	Target
12											
13											
14											

Obrázek 16 – Pohled na proměnné v SPSS, vlastní zpracování

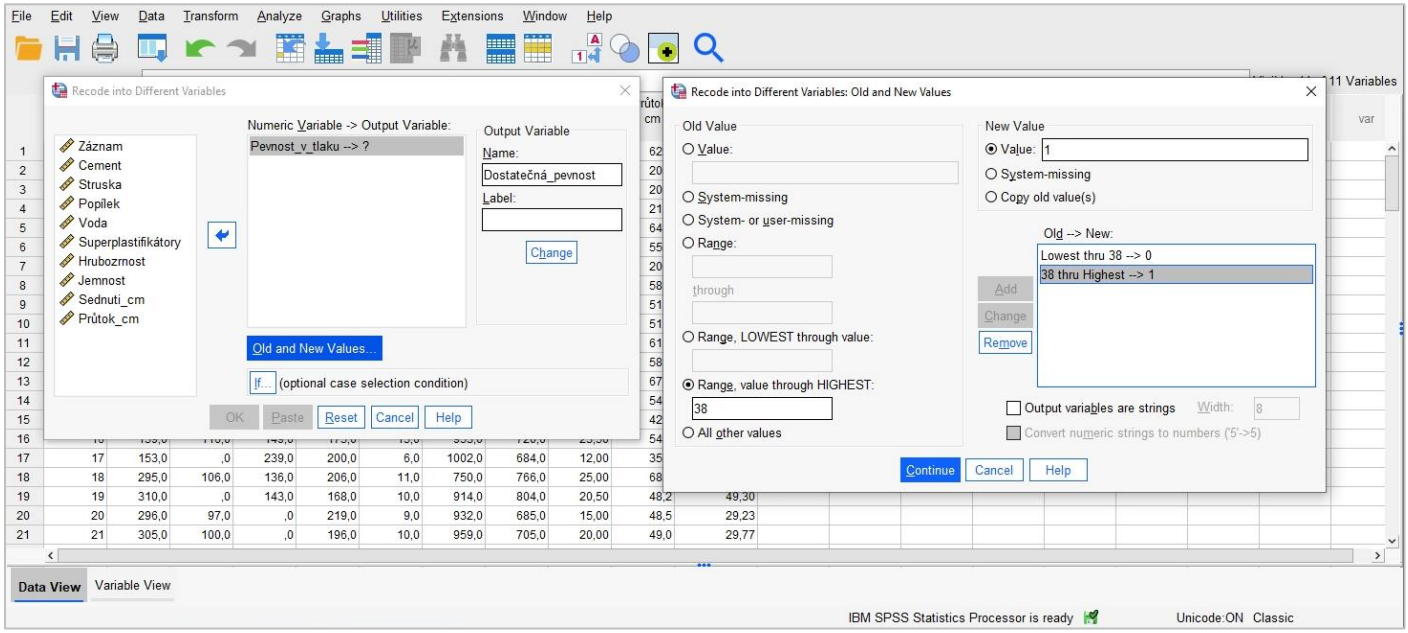
V pohledu proměnných je možné nastavit metadata jednotlivých proměnných, aby odpovídala jejich obsahu a cílům prováděné analýzy. Úprava metadat v tomto pohledu však nemění načtená data. Pokud by bylo v rámci přehlednosti lepší odstranit desetinná místa například u množství vody a cementu, je možné nastavit atribut *Decimals* na 0, to může usnadnit přehled v datovém pohledu a na provedené výpočty to nemá vliv, jelikož data jsou stále uchována v původní přesnosti.

Někdy je však nutné data skutečně upravit. Pro úpravy na úrovni záznamů je možné manuálně přepsat jednotlivé záznamy prostým kliknutím na požadovanou buňku a přepsáním požadované hodnoty, kdy je však nutné dodržet typ proměnné (při jejím nedodržení SPSS automaticky zahodí vloženou hodnotu a zanechá původní). Pro úpravu všech nebo části záznamů v rámci celé proměnné je možné využít transformačních procedur, jak je zobrazeno na obrázku č. 17.



Obrázek 17 – Transformační procedury v SPSS, vlastní zpracování

Pro příklad lze uvést procedury SPSS *Recode into Same Variables…* a *Recode into Different Variables…*, které transformují načtená data, v prvním případě v rámci stejné proměnné a v druhém případě se tvoří nová proměnná. Pokud by tak bylo potřeba vytvořit například kategoriální proměnnou *Dostatečná_pevnost* pro vyhodnocení typu splňuje/neslňuje, je možné vytvořit novou proměnnou z hodnot proměnné *pevnost_v_tlaku*. Využije se tak funkce *Recode into Different Variables…*, která otevře nové okno kde se zadá zdrojový a nový sloupec, jednotlivé proměnné je možné namapovat v novém okně po kliknutí na tlačítko *Old and New Values…* jak je vidět na obrázku č. 18, jsou namapovány hodnoty od minima do hodnoty 38 jako 0 (nesplňuje) a hodnoty nad 38 jako 1 (splňuje)

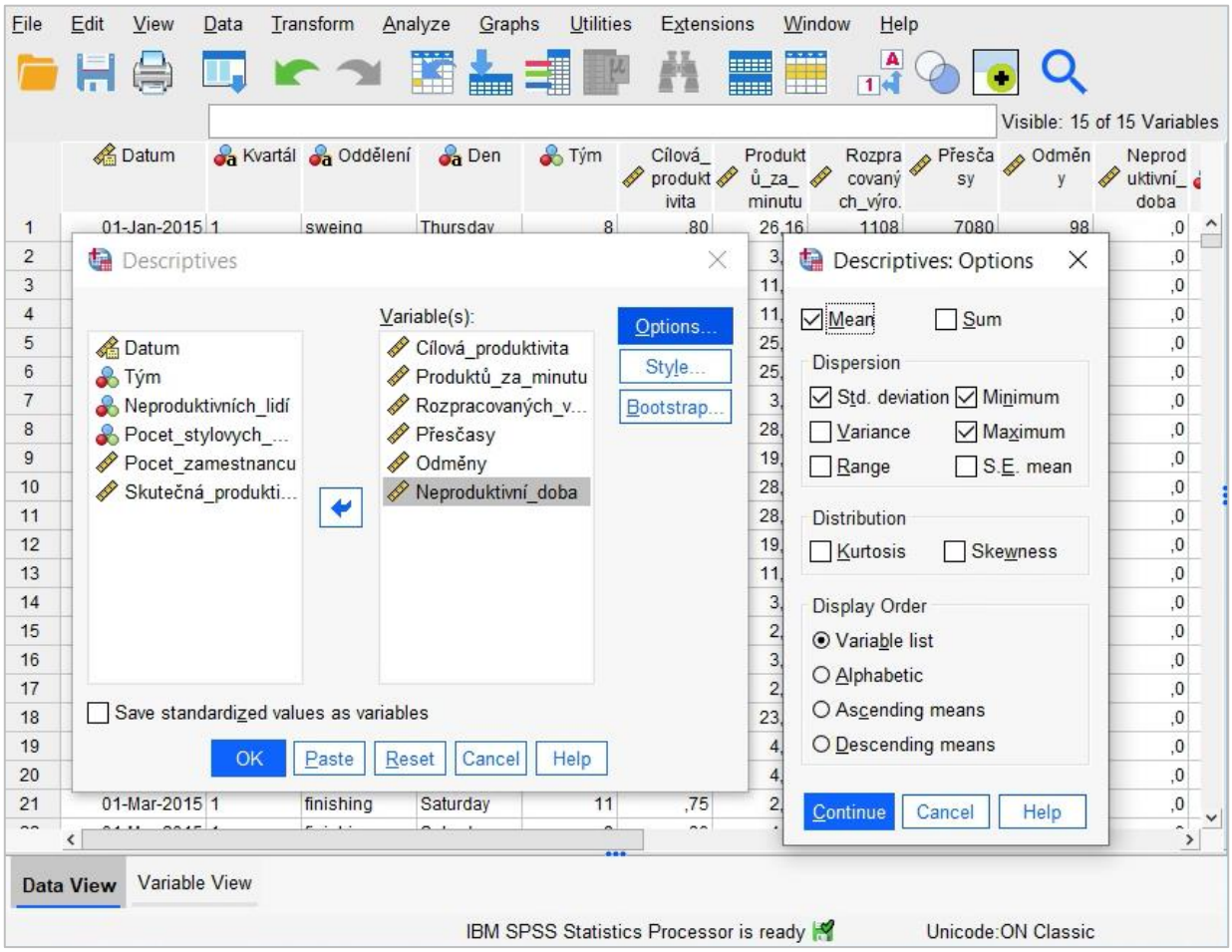


Obrázek 18 – Transformace do nové proměnné v SPSS, vlastní zpracování

Je tak rozdíl mezi přemapováním názvů, v SPSS značeno jako *Label,* kdy v datovém pohledu lze přepínat mezi originálními názvy a přemapovanými názvy pomocí tlačítka *Value labels*, které nemá vliv na výpočty prováděné v SPSS a slouží především pro lepší přehlednost v datové sadě a transformacemi, jako je redukce počtu kategorií proměnné, či úpravě vedoucí ke změně typu proměnné, které dopady na výpočty či využitelnost proměnné pro konkrétní statistické metody mít mohou.

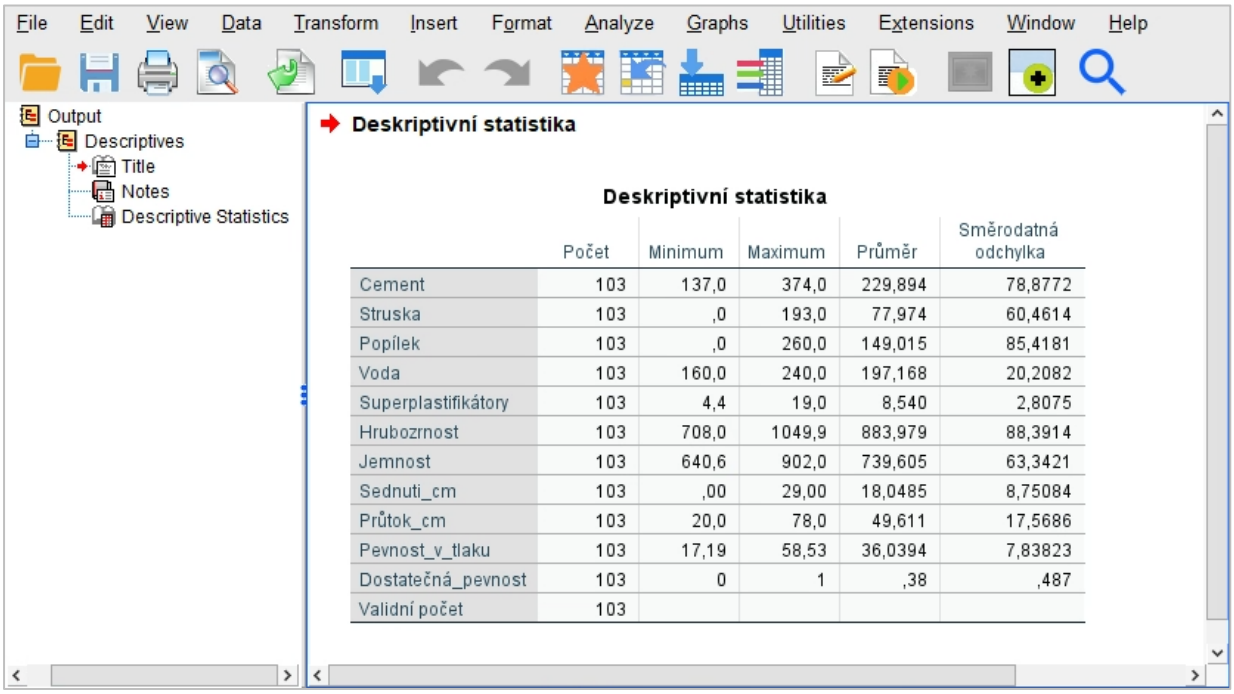
7.2.2 Popisná statistika

Jakmile jsou data transformována a označena do požadované podoby, je možné využít popisné statistiky SPSS. Tu je možné najít v SPSS pod možností *Analyze -> Descriptive Statistics – Descriptives…*, následně se zobrazí okno viz obrázek č. 19, ve které je možné zvolit pro které proměnné bude deskriptivní statistika vypočtena a tlačítka pro další možnosti. Tlačítko *Options…* otevře podokno s výběrem jednotlivých popisných statistik, které je možné spočítat, viz obrázek č. 18, tlačítko *Style…* pro kondiční formátování a tlačítko *Bootstrap…*, což je dle [10] technika pro odhad přesnosti výběrových statistik, která není v rámci této práce využita.



Obrázek 19 *– Okno deskriptivní statistiky s podoknem Options v SPSS, vlastní zpracování*

Po kliknutí na tlačítko *OK* se otevře nové okno v SPSS s požadovanými statistikami, viz obrázek č. 20. V levé části nového podokna je vidět seznam zobrazených úloh, v tomto případě jediný output s deskriptivní statistikou a na hlavním plátně zobrazené statistiky. Při nové akci, například zobrazení deskriptivní statistiky pro jiné proměnné, se v levém sloupci přidá nová úloha pod tu předchozí a obdobným způsobem se na hlavním plátně zobrazí požadovaná statistika pod předchozí úlohu. Je zde také možné změnit jednotlivé názvy, upravit vzhled tabulek, dodatečných vizuálů a provést mnoho dalších úprav, které zlepšit přehlednost, případně připravit vizuály pro prezentaci managementu. Stačí tak například dvojklik na požadovanou tabulku, ta se následně otevře v novém podokně s možnostmi dovolujícími ji upravit do požadované podoby. Obdobným způsobem je možné upravit i grafy a ostatní výstupní objekty v SPSS.



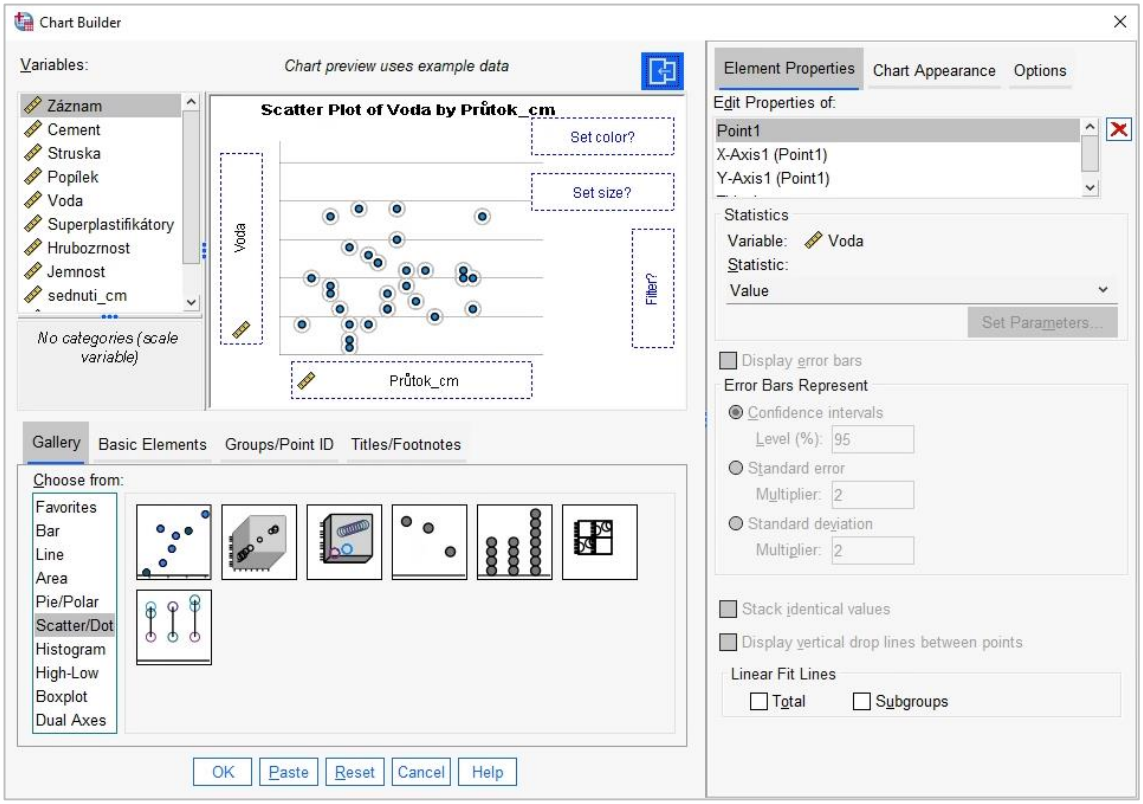
Deskriptivní statistika

	Počet	Minimum	Maximum	Průměr	Směrodatná odchylka
Cement	103	137,0	374,0	229,894	78,8772
Struska	103	,0	193,0	77,974	60,4614
Popílek	103	,0	260,0	149,015	85,4181
Voda	103	160,0	240,0	197,168	20,2082
Superplastifikátory	103	4,4	19,0	8,540	2,8075
Hrubozrnost	103	708,0	1049,9	883,979	88,3914
Jemnost	103	640,6	902,0	739,605	63,3421
Sednuti_cm	103	,00	29,00	18,0485	8,75084
Průtok_cm	103	20,0	78,0	49,611	17,5686
Pevnost_v_tlaku	103	17,19	58,53	36,0394	7,83823
Dostatečná_pevnost	103	0	1	,38	,487
Validní počet	103				

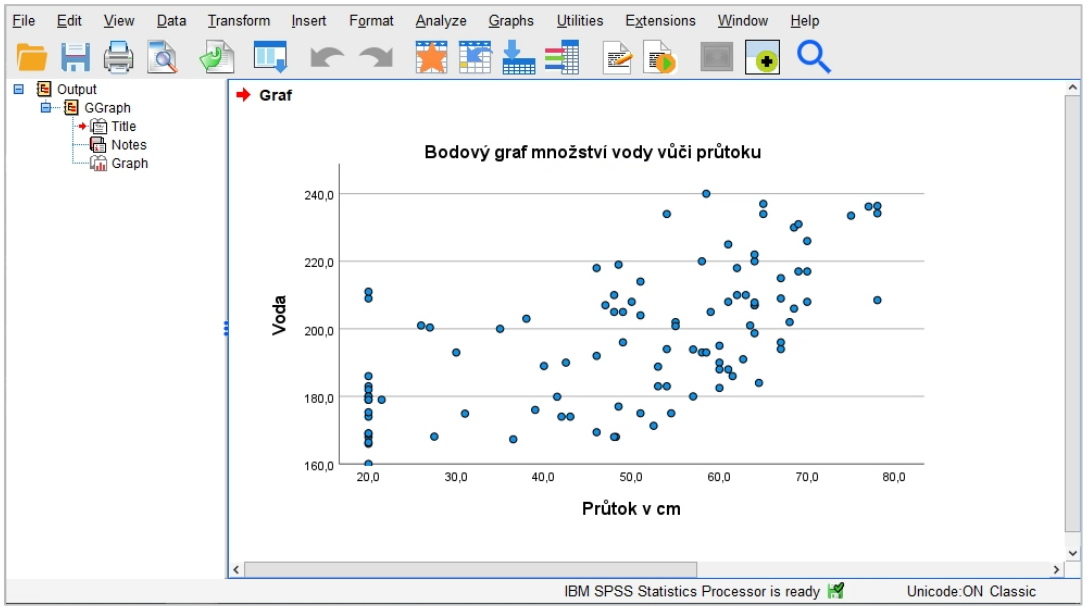
Obrázek 20 – Deskriptivní statistika v SPSS, vlastní zpracování

7.2.3 Grafy v SPSS

Grafy se dají v SPSS zobrazit různými způsoby, avšak skrze menu *Graphs -> Chart Builder…* je možné otevřít okno se všemi dostupnými grafy, které je možné využít i s jejich náhledem (obr. č.21). Ten se dá rozdělit do tří částí, v levé dolní části jsou jednotlivé grafy dle jejich typů. Při kliknutí na konkrétní graf se zbytek okna upraví tak, aby bylo možné nastavit parametry a vlastnosti požadovaného grafu. V levé horní části je sloupec s proměnnými z datové sady, kde se zobrazí pouze ty proměnné, které splňují podmínky pro daný graf a na plátně je vidět náhled na volený typ grafu. Přetažením jednotlivých proměnných do čerchovaně ohraničených obdélníků je možné umístit proměnné na požadované místo. V pravé části okna je možné nastavit vlastnosti a vzhled grafu. Ty je případně možné upravit i po vykreslení, pokud by zde zvolené nastavení bylo nevyhovující. Pro případné filtrování je možné do pole *Filter* vložit požadovanou proměnnou a v pravé části *Chart Builder* v *Element Properties* se doplní nová možnost *Filter Panel,* u které je následně možné odstranit nežádoucí hodnoty jejich přesunutím do okna *Excluded*.

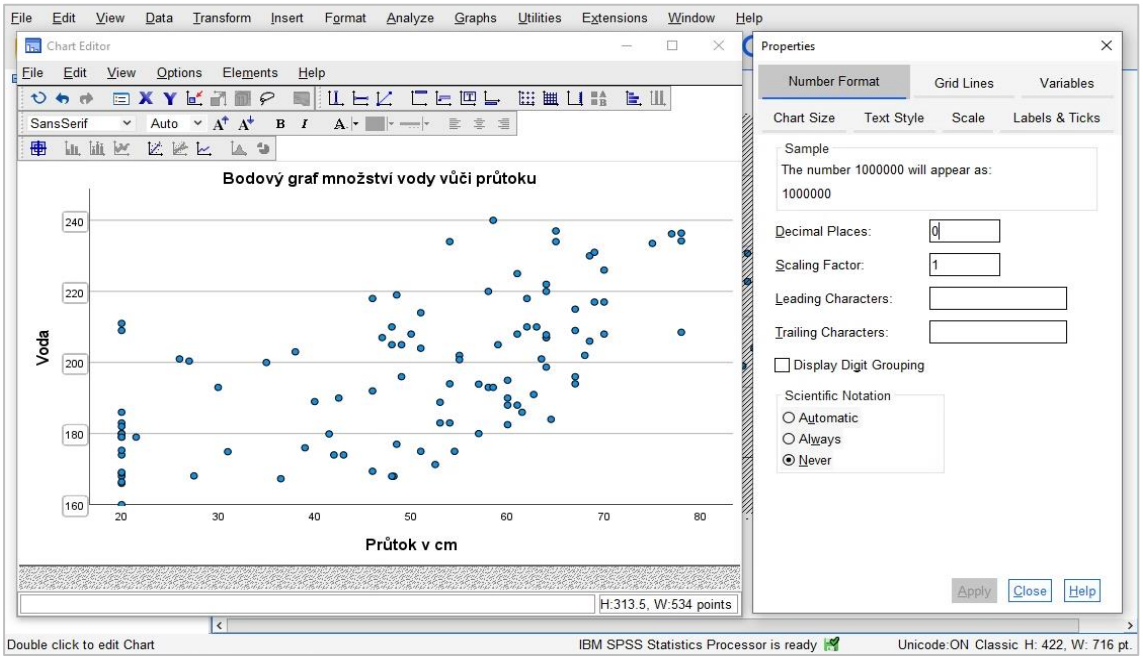


Obrázek 21 – Okno pro sestavování grafů v SPSS, vlastní zpracování

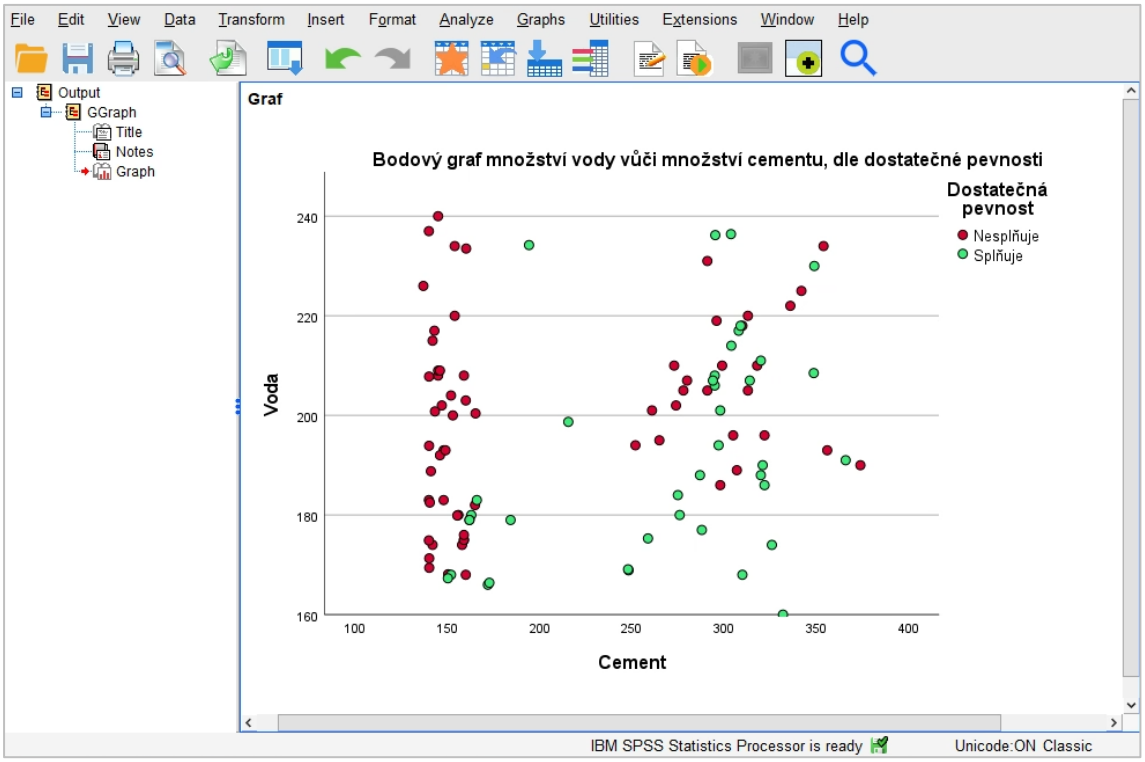


Obrázek 22 – Bodovy graf množství vody vůči průtoku, vlastní zpracování

Po stisknutí tlačítka *OK* se tak na plátno vykreslí bodový graf s požadovanými proměnnými (obrázek č. 22), kde je možné provést další úpravy. V tomto případě je vhodné zmenšit počet desetinných míst pro obě proměnné. Dvojklikem na graf se zobrazí nové podokno *Chart Editor (obrázek č. 23),* ve kterém je možné jednotlivé části grafu upravovat. Dalším dvojklikem na libovolnou hodnotu na ose x se otevře nové podokno, kde je na kartě *Number Format* možné upravit počet desetinných míst pro hodnoty osy x pomocí *Decimal Places*. I zde se jedná pouze o vizuální úpravu v grafu, záznamy mají původní desetinnou přesnost. Obdobným způsobem lze upravovat jednotlivé části grafu



Obrázek 23 – podokna *Chart Editor* a *Properties* pro bodový graf množství vody vůči průtoku, vlastní zpracování



Obrázek 24 – Bodový graf množství vody vůči množství cementu, dle dostatečné pevnosti, vlastní zpracování

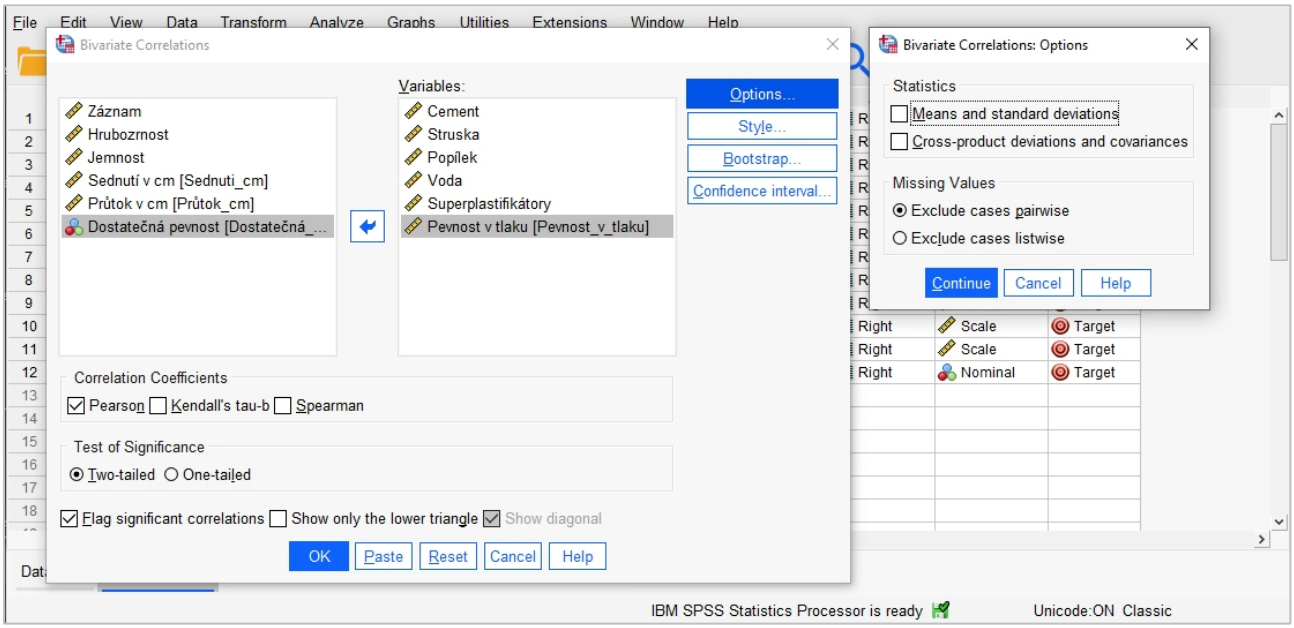
Na obrázku č. 24 je vidět bodový graf množství vody vůči množství cementu při využití barevného rozlišení pro uměle vytvořenou proměnnou *Dostatečná pevnost* v *Chart Builder*. To způsobí barevné rozlišení jednotlivých záznamů podle toho, zda splňují nebo nesplňují podmínku pro uměle vytvořenou proměnnou *Dostatečná pevnost*. Je tak lépe vidět, jaká kombinace množství těchto dvou složek má na požadovanou pevnost vliv. Z grafu lze usuzovat, že malé množství cementu má negativní dopad na požadovanou pevnost a větší množství cementu s menším množstvím vody má pozitivní dopad na požadovanou pevnost, ačkoliv v dané oblasti není příliš záznamů.

To je však zobrazení pouze pro tyto dvě proměnné. Pro zobrazení více párových bodových grafů lze využít možnosti maticového bodového grafu, do kterého lze vložit více proměnných najednou, které jsou následně zobrazené v maticovém rozložení. V SPSS jsou další možné úpravy, jako je například tlačítko *Bin elements* v *Chart editor*, který umožní seskupit velké množství záznamů do jednoho bodu, což je obzvláště vhodné při práci s rozsáhlým datovým souborem.

7.2.4 Korelační analýza v SPSS

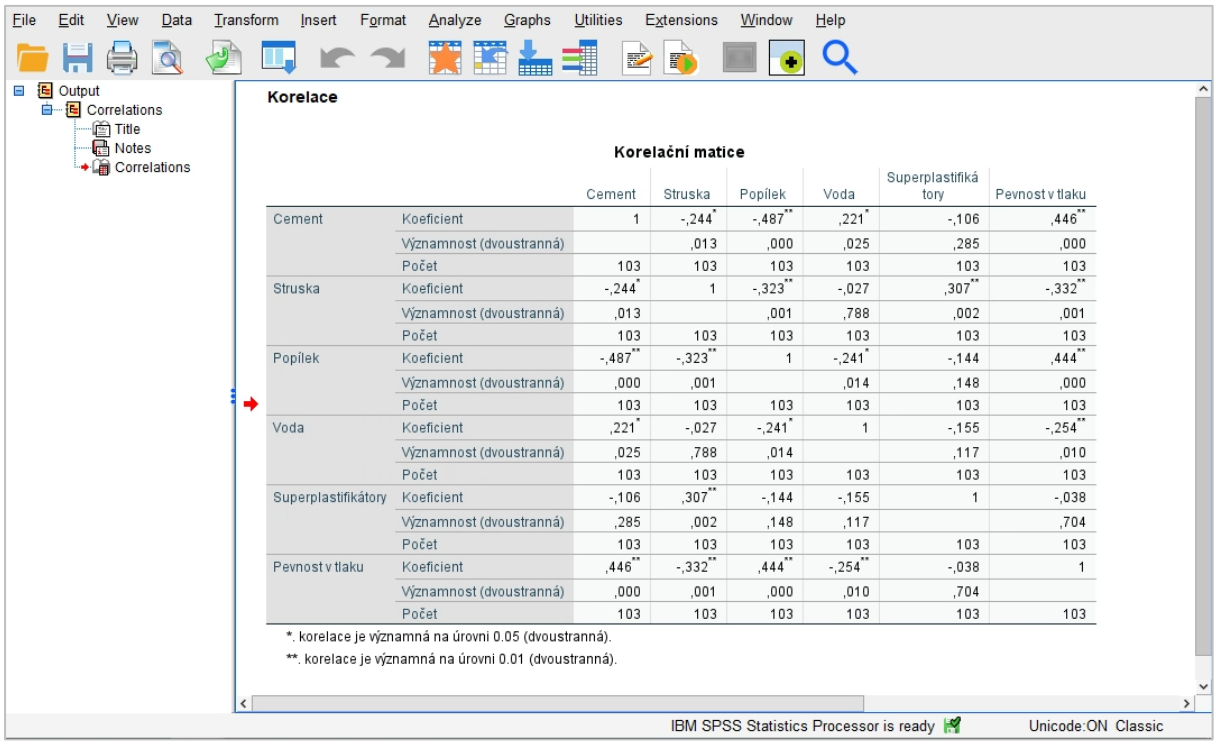
Korelační analýzu je možné v SPSS nalézt v menu *Analyze -> Correlate*. Ta obsahuje podmenu s dalšími možnostmi, jako je dvourozměrná korelace či parciální korelace. Při zvolení dvourozměrné korelační analýzy se otevře okno *Bivariate Correlations*, ve kterém je možné zvolit požadované proměnné pro realizaci korelační analýzy. Dále je možné zvolit korelační koeficient, kde je předem nastaven Pearsonův korelační koeficient, tlačítko *Options...*, viz obrázek č. 25, s dalšími možnostmi, obsahujícími například výpočet průměrů a standardní odchylky, *Style..., Bootstrap...* a *Confidence interval...*, které v obsahu této práce nejsou využity

Pro ukázku jsou využity proměnné *Cement, Struska, Popílek, Voda, Superplastifikátory* a *Pevnost_v_tlaku*, avšak pouze korelace pevnosti v tlaku vůči ostatním proměnným má smysl vyhodnocovat, jelikož ostatní proměnné jsou nenáhodné. Cílem je tak zjistit, jak každá tato proměnná ovlivňuje výslednou pevnost.



Obrázek 25 – Nastavení dvourozměrné korelace s otevřeným podoknem Options v SPSS, vlastní zpracování

Po vložení požadovaných proměnných a stisknutí tlačítka OK se zobrazí korelační matice pro zvolené proměnné, jak je vidět na obrázku č. 25. Pro účely ukázky matice vykreslené SPSS bez dodatečného nastavení jsou ponechány všechny záznamy, avšak smysl má vyhodnocovat pouze pevnost v tlaku vůči ostatním proměnným, tedy poslední sloupec nebo poslední 3 řádky zobrazené matice.



Obrázek 26 – Párová korelační matice v SPSS, vlastní zpracování

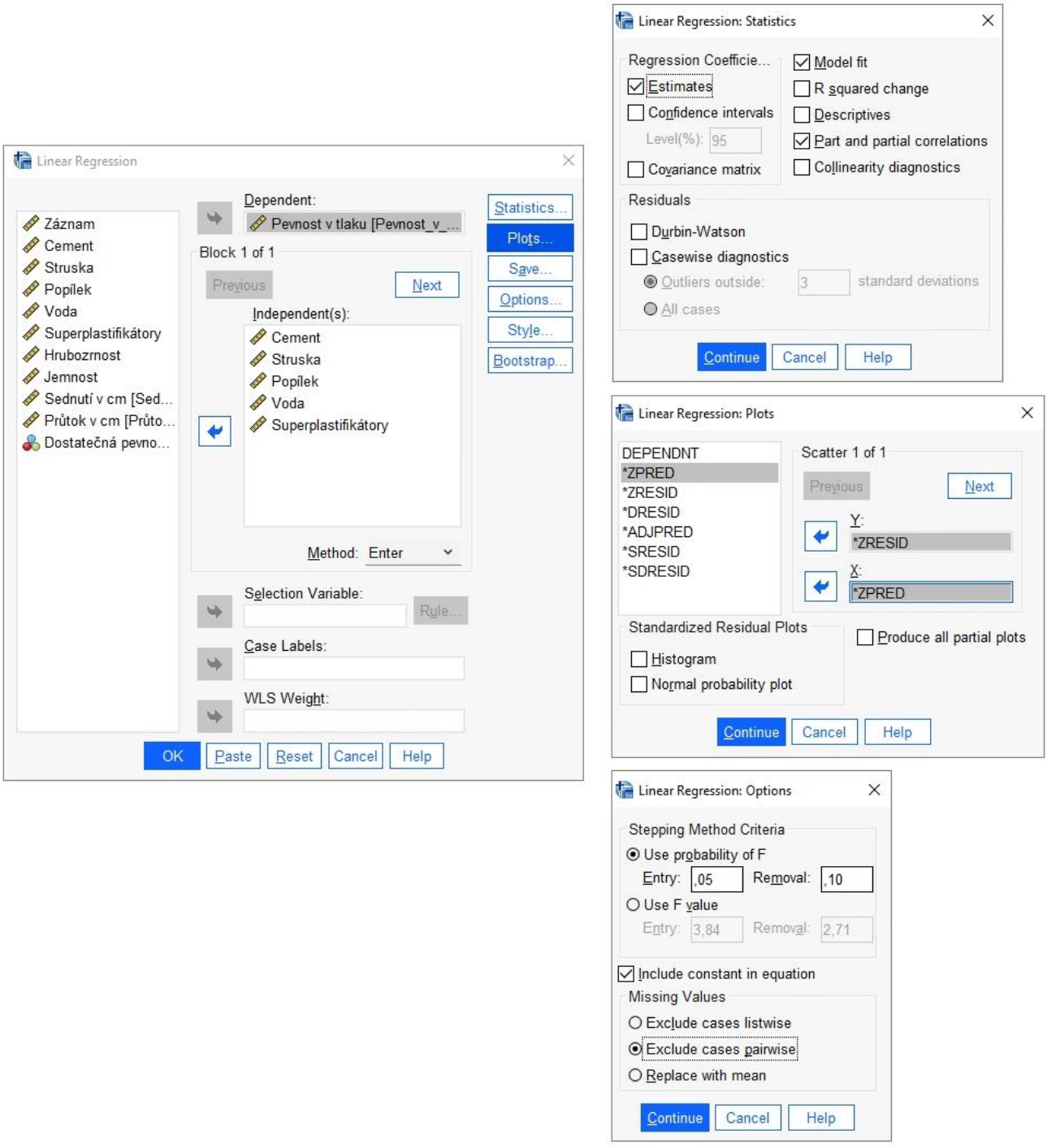
V matici jsou vidět jednotlivé výsledky pro zvolené proměnné. Pro každý záznam lze vidět vypočtený Pearsonův korelační koeficient, hranice významnosti pro oboustranný interval spolehlivosti a počet záznamů. Při větším počtu proměnných je pro lepší přehlednost vhodné ponechat pouze řádek koeficientu s tím, že v nastavení korelační matice se označí nepotřebné řádky, pravým tlačítkem myši se klikne na jeden z označených řádku a zvolí se možnost *Delete*. Informace o počtu záznamů je zbytečná, protože deskriptivní statistika již ukázala, že datová sada obsahuje 103 zcela vyplněných záznamů a splnění hranice významnosti je vidět u jednotlivých koeficientů pomocí počtu hvězdiček, které jsou vysvětleny v legendě pod grafem. Z tabulky na obrázku č. 26 tak lze zjistit, že na cílovou vlastnost *pevnost v tlaku* má kladný vliv především větší množství cementu a popílku, naopak více vody nebo strusky vede ke snížení výsledné pevnosti. Pro superplastifikátory je výsledná hodnota zanedbatelná, a především nesplňuje stanovené hranice významnosti, není tak možné usuzovat její vliv na požadovanou pevnost.

7.2.5 Lineární regresní analýza v SPSS

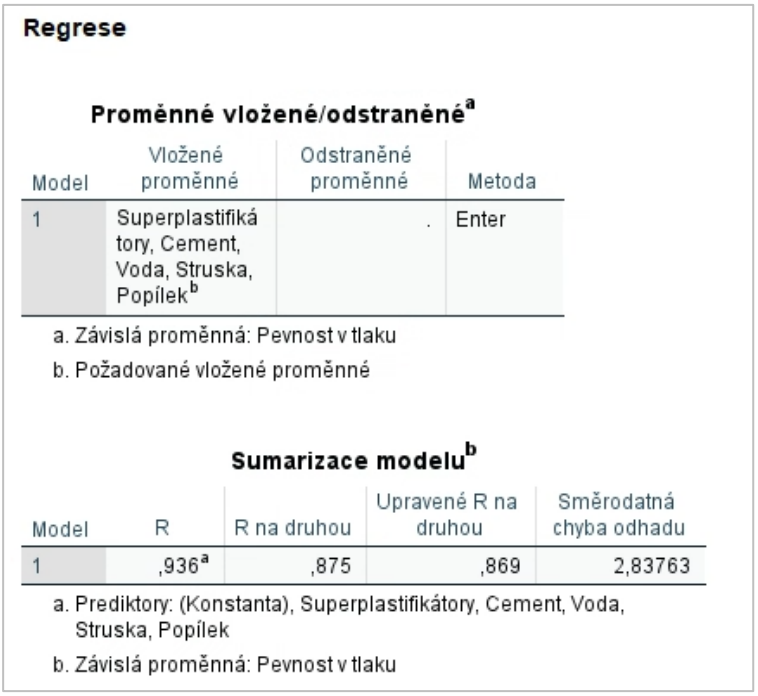
Oddíl lineární regresní analýzy v SPSS čerpá především z [6] (2015, str. 363-367), kde je v této práci využito obdobného nastavení a popsány výsledky lineární regresní analýzy. V SPSS se nachází procedura lineární regrese v menu *Analyze -> Regression -> Linear…*, které otevře okno s volbami pro závislou proměnou a seznam proměnných nezávislých. Po pravé straně jsou tlačítka, která otevírají nová podokna s dalšími možnostmi, jako:

- *Statistics* – pro volbu požadovaných statistik
- *Plots* – pro vykreslení grafů
- *Options* – pro další možnosti, jako je způsob zpracování záznamů s chybějícími daty a několika dalších, které v této práci nejsou využity. ,

Na obrázku č. 27 jsou zmíněná okna s využitým nastavením, které jsou následně prezentovány v nadcházejících obrázcích



Obrázek 27 – Okno Lineární regrese s podokny Statistics, Plots a Options v SPSS, vlastní zpracování



Regrese

Proměnné vložené/odstraněné[a]

Model	Vložené proměnné	Odstraněné proměnné	Metoda
1	Superplastifikátory, Cement, Voda, Struska, Popílek[b]	.	Enter

a. Závislá proměnná: Pevnost v tlaku

b. Požadované vložené proměnné

Sumarizace modelu[b]

Model	R	R na druhou	Upravené R na druhou	Směrodatná chyba odhadu
1	,936[a]	,875	,869	2,83763

a. Prediktory: (Konstanta), Superplastifikátory, Cement, Voda, Struska, Popílek

b. Závislá proměnná: Pevnost v tlaku

Obrázek 28 – Tabulka vložených a odstraněných proměnných a sumarizace modelu v SPSS, vlastní zpracování

Na obrázku č. 28 je vidět tabulka vložených a odstraněných proměnných sestaveného lineárního regresního modelu. V tomto případě byla zvolena metoda enter, tedy vložení všech proměnných najednou. V sumarizaci modelu je podstatná hodnota *R na druhou*, která představuje množství vysvětleného rozptylu sestaveným modelem. Hodnota 0,875 znamená, že sestavený model vysvětluje 87,5 % celkového rozptylu závislé proměnné. Jinak řečeno, stále existuje 12,5 % nevysvětleného rozptylu neznámými nebo nevyužitými proměnnými. [6] (2015, str. 366) odkazuje na učebnice, ve kterých se doporučuje využívat hodnoty *upravené R na druhou*. Tento údaj tak výslednou hodnotu upravuje vzhledem k počtu vložených proměnných, kdy s větším počtem nezávislých proměnných může uměle růst i *R na druhou*.

ANOVA[a]

Model		Suma čtverců	Stupně volnosti	Odhad rozptylu	F	Významnost
1	Regrese	5485,608	5	1097,122	136,253	,000[b]
	Reziduály	781,055	97	8,052		
	Celkem	6266,664	102			

a. Závislá proměnná: Pevnost v tlaku

b. Prediktory: (Constant), Superplastifikátory, Cement, Voda, Struska, Popílek

Obrázek 30 – Analýza rozptylu v SPSS, vlastní zpracování

Koeficienty[a]

Model		Nestandardizované koeficienty		Standardizované koeficienty			Korelace			Statistika multikolinerity	
		B	Standardní chyba	Beta	t	Významnost	Nultého řádu	Parciální	Část	Tolerance	VIF
1	(Konstanta)	11,940	3,752		3,182	,002					
	Cement	,102	,005	1,023	21,692	,000	,446	,911	,778	,578	1,730
	Struska	,024	,006	,187	4,221	,000	-,332	,394	,151	,655	1,526
	Popílek	,089	,005	,965	19,665	,000	,444	,894	,705	,533	1,874
	Voda	-,087	,015	-,224	-5,927	,000	-,254	-,516	-,212	,898	1,114
	Superplastifikátory	,327	,107	,117	3,056	,003	-,038	,296	,110	,872	1,147

a. Závislá proměnná: Pevnost v tlaku

Obrázek 29 – Tabulka s koeficienty lineární regrese v SPSS, vlastní zpracování

V tabulce analýzy rozptylu na obrázku č. 29 lze zjistit, zda je platná nulová hypotéza $R^2 = 0$. Významnost je při zaokrouhlení na 3 desetinná místa stále menší než 0,000 a tedy splňuje obvykle využívanou hranici menší než 0,05.

V tabulce s koeficienty na obrázku č. 30 jsou všechny v modelu využité proměnné, včetně závislé proměnné pojmenované jako *(Konstanta)*. Podle [6] (2015, str. 367) je to hodnota závislé proměnné, v uvedené tabulce tedy pevnost v tlaku, při nulovém vstupu ostatních nezávislých proměnných. Tedy s nulovou hodnotou nezávislých proměnných by měla být pevnost v tlaku 11,94 N/mm2. Avšak s nulovou hodnotou vložených nezávislých proměnných nelze cement vytvořit, v tomto případě tak koeficient závislé proměnné nemá smysl interpretovat. Nestandardizovaný koeficient představuje změnu v závislé proměnné při jednotkové změně nezávislé proměnné. V tomto případě by tak s každým jednotkovým navýšením, například množství cementu, zvýší pevnost v tlaku o 0,102 N/mm2. Naopak navýšením množství vody se snižuje výsledná pevnost v tlaku. Standardizované koeficienty jsou hodnoty upravené do intervalu <-1;1> a je díky nim možné porovnávat vliv jednotlivých nezávislých proměnných na výsledný model i přes jejich odlišné hodnoty, kdy střední hodnota a rozptyl mezi těmito proměnnými mohou být zásadně odlišné. Čím více se standardizovaný koeficient blíží nule, tím menší vliv má na závislou proměnnou, naopak hodnoty blížící se -1 a 1 znamenají výrazné ovlivnění závislé proměnné. Na první pohled je zde vidět

porušení zmiňovaného pravidla u proměnné *Cement*, která přesahuje hodnotu 1. Dle [13] se tento jev objevuje v případě pozitivní či negativní korelace mezi prediktory. V takovém případě může být vhodné sestavit model znovu, avšak za využití metody vkládání proměnných *stepwise*, tedy postupného vkládání proměnných, jak je vidět na obrázku č. 31.

Koeficienty[a]

Model		Nestandardizované koeficienty		Standardizované koeficienty	t	Významnost	Korelace			Statistika multikolinearity	
		B	Standardní chyba	Beta			Nultého řádu	Parciální	Část	Tolerance	VIF
1	(Konstanta)	25,857	2,150		12,025	,000					
	Cement	,044	,009	,446	5,004	,000	,446	,446	,446	1,000	1,000
2	(Konstanta)	4,381	1,781		2,461	,016					
	Cement	,086	,005	,867	15,851	,000	,446	,846	,758	,763	1,310
	Popílek	,079	,005	,866	15,835	,000	,444	,846	,757	,763	1,310
3	(Konstanta)	23,731	3,585		6,619	,000					
	Cement	,090	,005	,902	18,986	,000	,446	,886	,782	,752	1,330
	Popílek	,075	,004	,822	17,210	,000	,444	,866	,709	,744	1,343
	Voda	-,099	,017	-,255	-5,972	,000	-,254	-,515	-,246	,928	1,078
4	(Konstanta)	16,427	3,597		4,567	,000					
	Cement	,101	,005	1,013	20,669	,000	,446	,902	,772	,581	1,722
	Popílek	,087	,005	,947	18,662	,000	,444	,883	,697	,541	1,848
	Voda	-,095	,015	-,244	-6,276	,000	-,254	-,535	-,234	,924	1,082
	Struska	,028	,006	,214	4,741	,000	-,332	,432	,177	,683	1,464
5	(Konstanta)	11,940	3,752		3,182	,002					
	Cement	,102	,005	1,023	21,692	,000	,446	,911	,778	,578	1,730
	Popílek	,089	,005	,965	19,665	,000	,444	,894	,705	,533	1,874
	Voda	-,087	,015	-,224	-5,927	,000	-,254	-,516	-,212	,898	1,114
	Struska	,024	,006	,187	4,221	,000	-,332	,394	,151	,655	1,526
	Superplastifikátory	,327	,107	,117	3,056	,003	-,038	,296	,110	,872	1,147

a. Závislá proměnná: Pevnost v tlaku

Obrázek 31 – Tabulka s koeficienty lineární regrese v SPSS, vlastní zpracování

Obrázek č. 31 prozrazuje, že k překročení hranice 1 pro standardizovaný koeficient cementu dochází po přidání proměnné Struska, což naznačuje možnou korelaci s jednou, či více již dosazených nezávislých proměnných. Při sestavení nového modelu bez proměnné *Struska* (obrázek č. 32) je výsledná hodnota *R na druhou* 0,852, tedy došlo k redukci vysvětleného rozptylu o 2,3 % a standardizovaný koeficient cementu je 0,933.

Sumarizace modelu[b]

Model	R	R na druhou	Upravené R na druhou	Směrodatná chyba odhadu
1	,923[a]	,852	,846	3,07150

a. Prediktory: (Konstanta), Superplastifikátory, Cement, Voda, Popílek

b. Závislá proměnná: Pevnost v tlaku

Koeficienty[a]

Model		Nestandardizované koeficienty		Standardizované koeficienty	t	Významnost
		B	Standardní chyba	Beta		
1	(Konstanta)	16,806	3,865		4,348	,000
	Cement	,093	,005	,933	20,490	,000
	Voda	-,089	,016	-,229	-5,582	,000
	Popílek	,079	,004	,865	18,609	,000
	Superplastifikátory	,419	,114	,150	3,684	,000

a. Závislá proměnná: Pevnost v tlaku

Obrázek 32 – sumarizace modelu a koeficienty lineární regrese bez proměnné Struska, vlastní zpracování

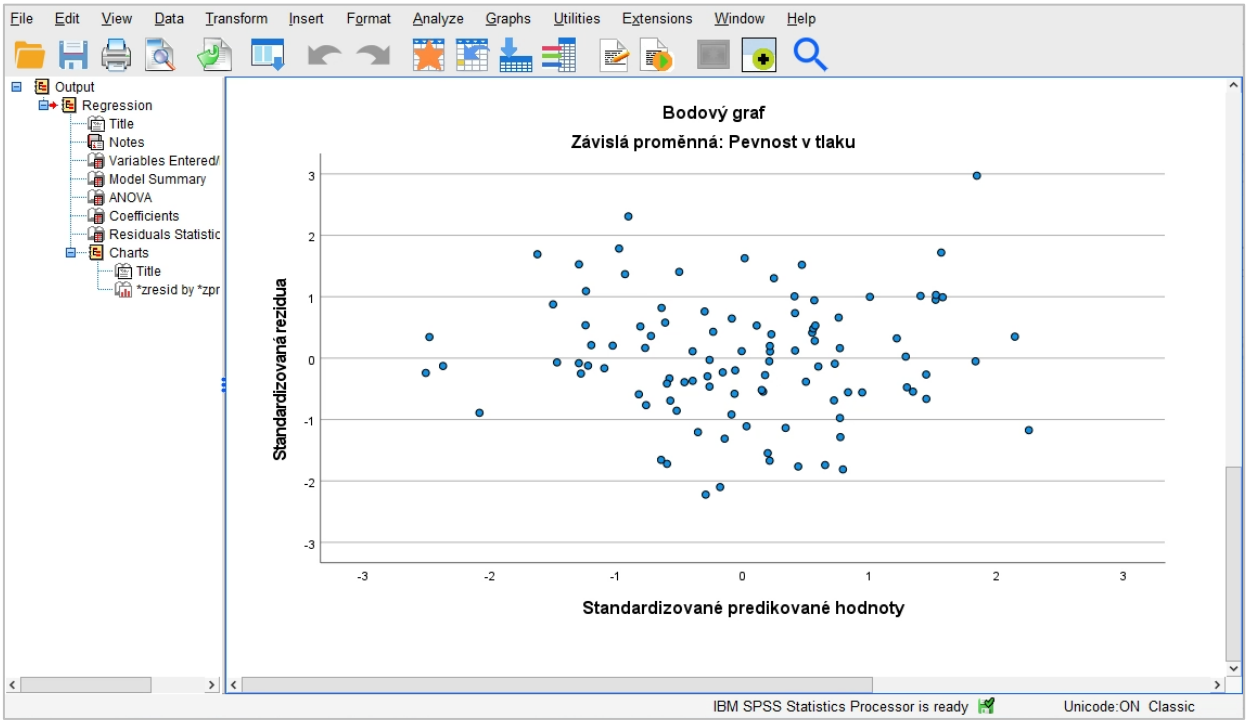
Avšak požadovaná pevnost je pouze jedna ze sledovaných proměnných a pevnost samotná nemusí být nic platná, pokud nebude mít beton při aplikaci požadované vlastnosti průtoku a sednutí, tak může být v praxi nepoužitelný.

Bodový graf na obrázku č. 33 představuje standardizované predikované hodnoty proti standardizovaným reziduím. Standardizace jednotlivých hodnot je provedena následujícím vzorcem:

$$Z = \frac{X - \mu}{\sigma}$$

μ – průměrná hodnota
σ – směrodatná odchylka

Tento graf by tak podle [6] (2015, str. 370) neměl vykazovat žádný vzorec v uspořádání proměnných. V případě že tomu tak není, je to znak možného nenaplnění předpokladu o linearitě a homoskedasticitě, což se nezdá být tento případ. Samotný test homoskadisticity zde není prováděn, nicméně je v rámci SPSS možné využít oficiálních stránek IBM, kde byl tento problém řešen, viz [14].



Obrázek 33 – Bodový graf standardizované regresní predikované hodnoty a standardizovaných regresních reziduí, vlastní zpracování

7.3 Realizace vybraných metod analýzy dat pro účely managementu

Pro realizaci datové analýzy byl vybrán datový soubor [13], příloha 2, obsahující informace o službě vypůjčení jízdních kol ve městě Soul. Datový soubor obsahuje celkem 14 proměnných:

- Datum – Ve formátu rok-měsíc-den
- Počet_vypůjčených_kol – počet vypůjčených kol
- Hodina – Hodina daného dne
- Teplota – Teplota ve stupních Celsia
- Vlhkost – Vlhkost v procentech
- Rychlost_větru_ms – Rychlost větru udaná v metrech za sekundu
- Vizibilita – viditelnost na vzdálenost 10 metrů
- Teplota_rosného_bodu – Teplota rosného bodu ve stupních Celsia
- Solární_radiace – solární radiace udaná v MJ/m2
- Dešťové_srážky_mm – Dešťové srážky udané v milimetrech (1 litr vody na 1 metr čtvereční)
- Sněhové_srážky_cm – Sněhové srážky udané v centimetrech
- Roční_období – výčet textových hodnot: Spring, Léto, Podzim, Zima
- Svátek – výčet textových hodnot: Holiday/No holiday
- Provozni_den – výčet textových hodnot: Yes (standardní pracovní den), No (den, kdy půjčovna nebyla v provozu)

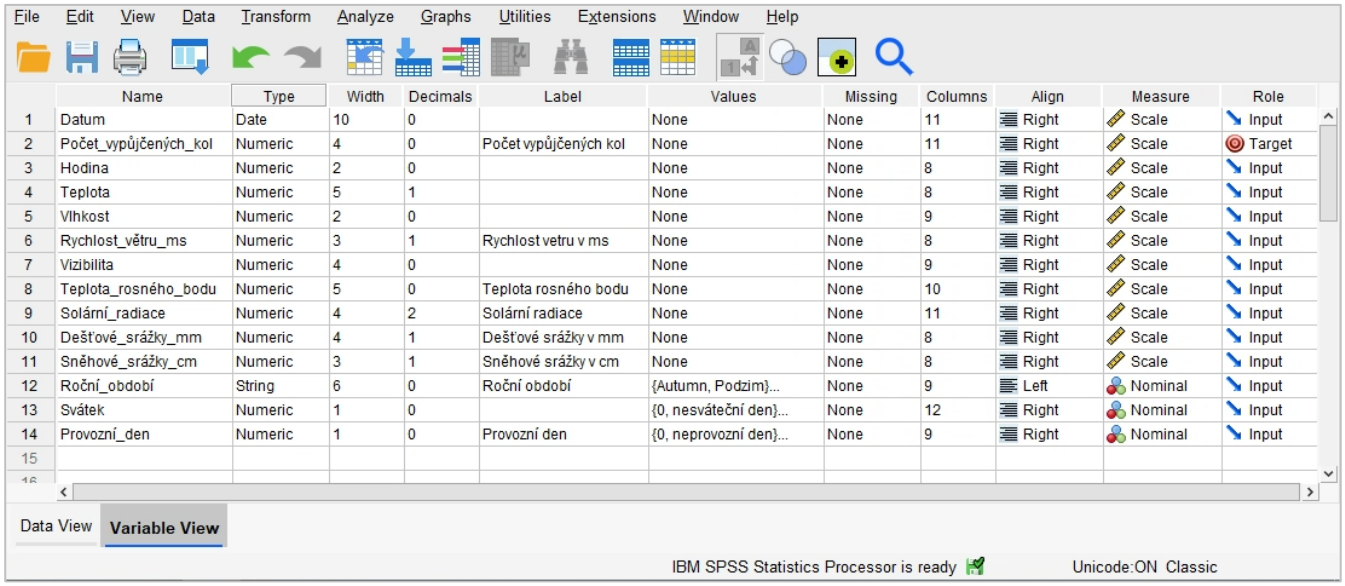
Datový soubor obsahuje 8 760 záznamů s rozsahem od 1. prosince 2017 do 30. listopadu 2018. Počet záznamů tak odpovídá 1 záznamu za hodinu po dobu 1 celého roku. Soubor obsahuje nominální, ordinální i intervalové proměnné a množství dat je dostatečné k provedení datové analýzy s vybranými metodami.

Cílem vybraných metod analýzy dat v této úloze je zjistit, které proměnné mají zásadní vliv na množství vypůjčených kol, sumarizovat nalezené výsledky do formy, kterou je možné předat managementu k následným rozhodnutím a dodat jakákoliv doporučení k dosažení efektivnějším výsledkům společnosti.

7.3.1 Rozbor datového souboru

Po načtení datového souboru do SPSS je vhodné jako první upravit všechny proměnné do požadovaného stavu, aby se s nimi dalo lépe pracovat. V tomto případě se jedná o následující úkony, které vedou k výslednému nastavení proměnných na obrázku č. 34:

1) Datum obsahuje anglické názvy měsíců, a tak je upraven atribut *Type* na *Date* ve formátu dd.mm.yyyy, který upraví zobrazení datumu na číselnou reprezentaci i v rámci měsíce, na podobu den-měsíc-rok.
2) Přemapování ročního období na české ekvivalenty. Využije se tak procedury *Value Labels,* kterou lze otevřít klikem na atribut Values v pohledu na proměnné, která překryje anglické názvy, které v datovém souboru zůstanou nezměněny.
3) Pro proměnnou Svátek se využije procedura transformace, a to *Recode into Same Variables*. Transformace je provedena z textového typu na celočíselný, kde bude využita konvence 0 reprezentující nepravdu a 1 pravdu. Nesváteční dny tak budou označeny hodnotou 0. Celočíselné typy jsou pro výpočetní účely lépe zpracovatelné než textové, a proto je v tomto případě využita transformace. Aby se zabránilo případným nejasnostem, jsou tyto hodnoty označeny pomocí procedury *Value Labels* a jsou namapovány jako sváteční a nesváteční den.
4) Pro proměnnou provozní den bude provedena ta samá změna jako v kroku číslo 3 z důvodu konzistence s tím rozdílem, že namapování bude na hodnoty provozní a neprovozní den. *Label* také povolují jinak nepoužitelné znaky v záznamech, jako jsou mezery, které je činí ve výsledcích lépe čitelné.
5) Nastaví se závislá proměnná, která je v tomto případě počet vypůjčených kol pomocí změny atributu *Role* na *Target*
6) Pro proměnné svátek a provozní den se po změnách upraví *Type* na *Numeric* a nastaví *Width* na 1, tedy celočíselnou hodnotu se šířkou 1 číslice.
7) Pro všechny víceslovné proměnné se nastaví hodnota *Label* na verze s mezerami pro lepší čitelnost



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Datum	Date	10	0		None	None	11	Right	Scale	Input
2	Počet_vypůjčených_kol	Numeric	4	0	Počet vypůjčených kol	None	None	11	Right	Scale	Target
3	Hodina	Numeric	2	0		None	None	8	Right	Scale	Input
4	Teplota	Numeric	5	1		None	None	8	Right	Scale	Input
5	Vlhkost	Numeric	2	0		None	None	9	Right	Scale	Input
6	Rychlost_větru_ms	Numeric	3	1	Rychlost vetru v ms	None	None	8	Right	Scale	Input
7	Vizibilita	Numeric	4	0		None	None	9	Right	Scale	Input
8	Teplota_rosného_bodu	Numeric	5	0	Teplota rosného bodu	None	None	10	Right	Scale	Input
9	Solární_radiace	Numeric	4	2	Solární radiace	None	None	11	Right	Scale	Input
10	Dešťové_srážky_mm	Numeric	4	1	Dešťové srážky v mm	None	None	8	Right	Scale	Input
11	Sněhové_srážky_cm	Numeric	3	1	Sněhové srážky v cm	None	None	8	Right	Scale	Input
12	Roční_období	String	6	0	Roční období	{Autumn, Podzim}...	None	9	Left	Nominal	Input
13	Svátek	Numeric	1	0		{0, nesváteční den}...	None	12	Right	Nominal	Input
14	Provozní_den	Numeric	1	0	Provozní den	{0, neprovozní den}...	None	9	Right	Nominal	Input
15											

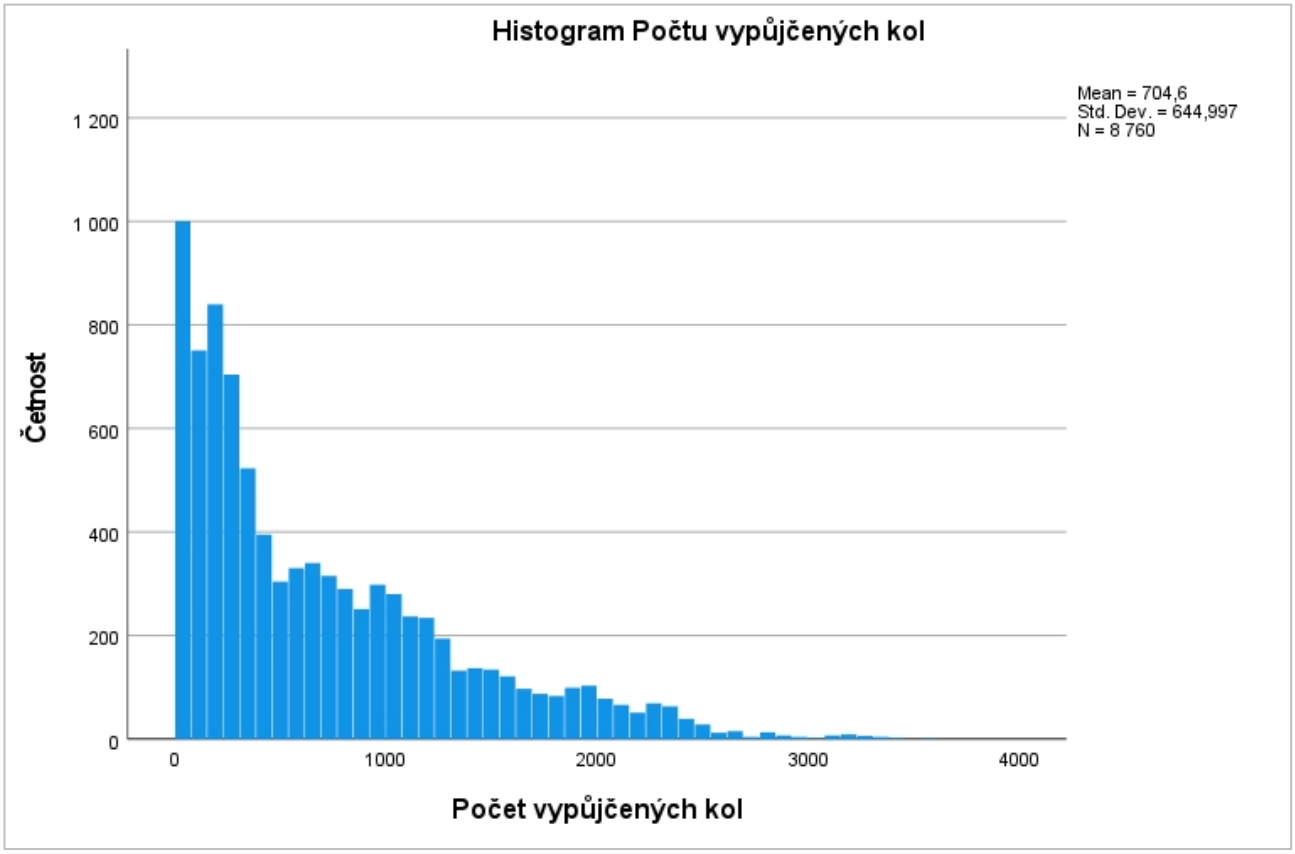
Obrázek 34 *– Pohled na proměnné datového souboru služby výpůjčky kol ve městě Soul v SPSS, vlastní zpracování*

Po načtení a nastavení všech proměnných požadované podoby je vhodné zobrazit popisnou statistiku pro všechny proměnné v datovém souboru, která je vidět na obrázku č. 35. Dobrou zprávou je u všech proměnných hodnota 8760 pro sloupec N, který dle [10] ukazuje množství vyplněných záznamů, tedy že data neobsahují prázdné záznamy, které se obvykle značí jako *null*. Minimum a Maximum jsou vhodné pro zjištění odlehlých hodnot. Například teplota se pohybuje v rozmezí od -17,8 °C do 39,4 °C, které se zdají být co se odlehlých hodnot týče v pořádku, což lze prohlásit i o ostatních proměnných. Zároveň lze potvrdit, že se v datovém souboru na první pohled nevyskytují případné nesmyslné údaje, kterými by mohli být například záporné hodnoty počtu vypůjčených kol a jiné. Zajímavostí je hodnota průměru u přemapovaných proměnných provozní den a svátek na binární hodnoty 0 a 1, kdy průměr ukazuje přechod z 0 na 1 a tedy zastoupení provozních a neprovozních dní, potažmo svátečních a nesvátečních dní. I když jsou tyto proměnné v tomto případě nominálního charakteru, díky tomu že jsou zastoupeny číselným typem pro ně SPSS spočítalo průměr, a tak lze z tabulky vyčíst, že zhruba 97 % dní je pracovních a zhruba 5 % dní v roce bylo svátečních. U závislé proměnné počtu vypůjčených kol je maximální hodnota 3 556 vypůjčených kol, avšak průměrně je vypůjčeno 704.6 kola, což znamená, že většina záznamů se bude vyskytovat v nižší polovině celkového rozptylu počtu vypůjčených kol. Avšak pro tuto závislou proměnou je vhodné znát její rozložení co nejlépe, a tak je lepší pro ni nechat sestavit histogram, viz. obrázek č. 36. V rámci deskriptivní statistiky je kromě procedury *Descriptives* také možné využít proceduru *Frequencies*, ve které je možné zjišťovat také kvartily, případně jiné volitelné percentily.

Deskriptivní statistika

	N	Minimum	Maximum	Průměr	Směrodatná odchylka
Datum	8760	01.12.2017	30.11.2018	01.06.2018	105 08:55:44,...
Počet vypůjčených kol	8760	0	3556	704,60	644,997
Hodina	8760	0	23	11,50	6,923
Teplota	8760	-17,8	39,4	12,883	11,9448
Vlhkost	8760	0	98	58,23	20,362
Rychlost vetru v ms	8760	,0	7,4	1,725	1,0363
Vizibilita	8760	27	2000	1436,83	608,299
Teplota rosného bodu	8760	-306	272	36,81	124,049
Solární radiace	8760	,00	3,52	,5691	,86875
Dešťové srážky v mm	8760	,0	35,0	,149	1,1282
Sněhové srážky v cm	8760	,0	8,8	,075	,4367
Svátek	8760	0	1	,05	,217
Provozní den	8760	0	1	,97	,180
Platných záznamů N (listwise)	8760				

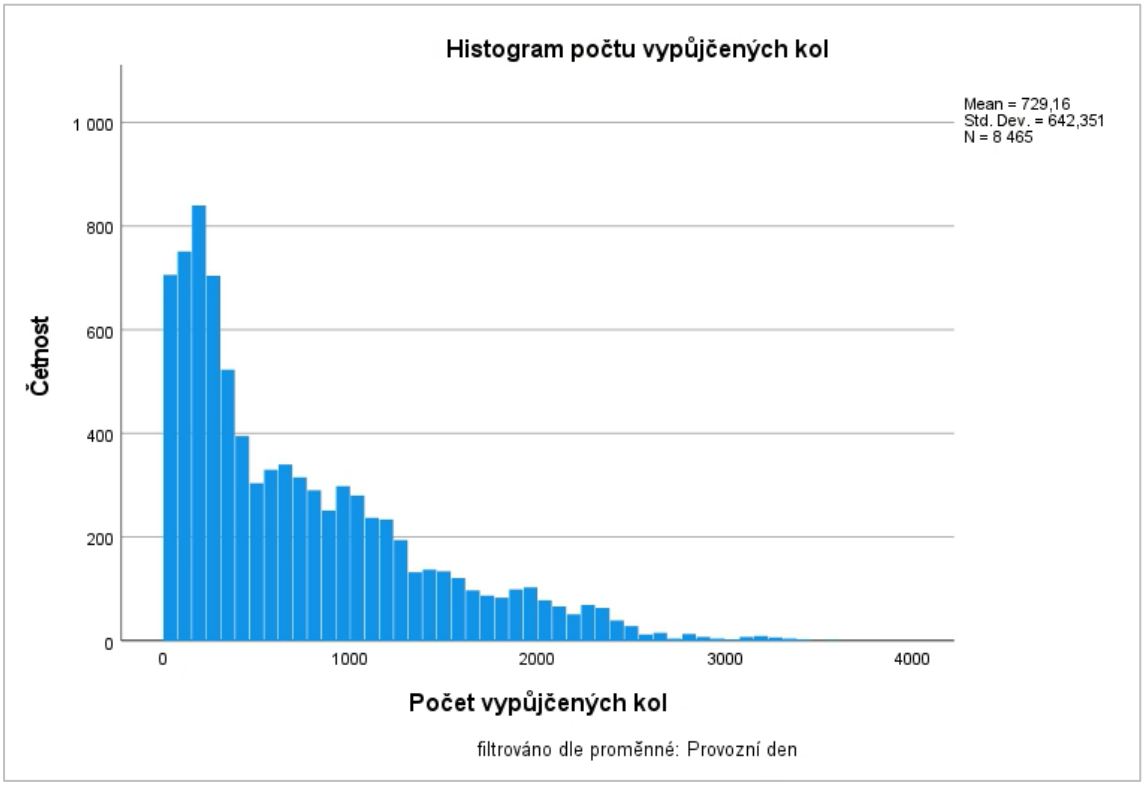
Obrázek 35 *– Deskriptivní statistika pro datový soubor služby vypůjčení kol ve městě Soul, vlastní zpracování*



Obrázek 36 *– Histogram počtu vypůjčených kol v SPSS, vlastní zpracování*

Histogram ukazuje postupně se snižující četnost záznamů pro zvyšující se hodnotu vypůjčených kol. Nejčastějším záznamem je 0 vypůjčených kol s hodnotou kolem 1 000 hodin, což je v přepočtu zhruba 42 dní v roce, kdy nebylo vypůjčeno jediné kolo. To je poměrně vysoká hodnota, kdy jsou kola zcela nevyužita. Zde se dá očekávat vliv neprovozních dní, avšak i jiné nezávislé proměnné mohou toto způsobit. Je tak vhodné v daném histogramu vyfiltrovat neprovozní dny, které jsou na obrázku č. 37.

Touto filtrací jsou z histogramu odstraněny záznamy, ve které služba nebyla v provozu, a tedy nebyla vypůjčena žádná kola. I když je pro management vhodné počítat se zastoupením neprovozních dní v průběhu roku, pro účely zjištění vlivu jiných nezávislých proměnných je dobré tyto záznamy vypustit, aby nezkreslovaly výsledky pro běžné provozní dny.



Obrázek 37 *– Histogram počtu vypůjčených kol s filtrem na nesváteční a provozní dny v SPSS, vlastní zpracování*

7.3.2 Jednoduchá lineární regrese

Po seznámení se s daty a provedením potřebných úprav je možné začít s lineární regresní analýzou. Jako první se sestaví jednoduchá lineární regrese. Pro výběr ideální nezávislé proměnné z datového souboru je vhodné sestavit párovou korelační matici, ze které lze vyčíst vhodnost jednotlivých proměnných. Vybrané proměnné jsou na obrázku č. 38.

Koeficienty

		Počet vypůjčených kol	Teplota	Vlhkost	Rychlost vetru v ms	Vizibilita	Dešťové srážky v mm	Sněhové srážky v cm
Počet vypůjčených kol	korelační koef.	1	,539**	-,200**	,121**	,199**	-,123**	-,142**
Teplota	korelační koef.	,539**	1	,159**	-,036**	,035**	,050**	-,218**
Vlhkost	korelační koef.	-,200**	,159**	1	-,337**	-,543**	,236**	,108**
Rychlost vetru v ms	korelační koef.	,121**	-,036**	-,337**	1	,172**	-,020	-,004
Vizibilita	korelační koef.	,199**	,035**	-,543**	,172**	1	-,168**	-,122**
Dešťové srážky v mm	korelační koef.	-,123**	,050**	,236**	-,020	-,168**	1	,008
Sněhové srážky v cm	korelační koef.	-,142**	-,218**	,108**	-,004	-,122**	,008	1

**. Korelace je významná na hranici 0.01 (oboustranná).

Obrázek 38 – korelační matice vybraných proměnných, vlastní zpracování

Z korelační matice vyčnívá vliv teploty na počet vypůjčených kol více než u jiných proměnných. Ta je tak vhodnou nezávislou proměnnou pro sestavení jednoduchého lineárního regresního modelu. Zároveň lze vidět určitou souvztažnost mezi vlhkostí a viditelností s negativním znaménkem, kdy s větší vlhkostí klesá viditelnost.

Na obrázku č. 39 je vidět sestavený model jednoduché lineární regrese počtu vypůjčených kol při využití teploty jako nezávislé proměnné. Při sestavení byl využit filtr pouze pro záznamy pracovních dní. Toho bylo docíleno v okně *Linear Regression* pomocí pole *Selection Variable*, kam byla vložena proměnná *Provozní_den* a nastaveno pravidlo, kde je hodnota této proměnné rovna 1 (díky využití *Label* je na obrázcích vidět textové namapování na „provozní den“)

Sumarizace modelu

Model	R Provozní den = provozní den (Zvolené)	R na druhou	Upravené R na druhou	Směrodatná chyba odhadu
1	,563[a]	,317	,317	531,021

a. Prediktory: (Konstanta), Teplota

ANOVA[a,b]

Model		Suma čtverců	Stupně volnosti	Odhad rozptylu	F	Významnost
1	Regression	1105952631,1	1	1105952631,1	3922,056	,000[c]
	Residual	2386420903,2	8463	281982,855		
	Total	3492373534,3	8464			

a. Závislá proměnná: Počet vypůjčených kol

b. Zvoleny záznamy, které splňují podmínku: Provozní den = provozní den

c. Prediktory: (Konstanta), Teplota

Koeficienty[a,b]

Model		Nestandardizované koeficienty		Standardizované koeficienty		
		B	Standardní chyba	Beta	t	Významnost
1	(Konstanta)	347,771	8,390		41,449	,000
	Teplota	29,863	,477	,563	62,626	,000

a. Závislá proměnná: Počet vypůjčených kol

b. Zvoleny pouze záznamy, které splňují podmínku: Provozní den = provozní den

Obrázek 39 – Sumarizace modelu, ANOVA a vypočtené koeficienty pro jednoduchou lineární regresi, vlastní zpracování

V sumarizaci modelu lze zjistit hodnotu R^2 = 0,317, což není dostatečná hodnota pro tvorbu závěrů ze sestaveného modelu. Nicméně vzhledem k množství proměnných v datovém souboru a faktu, že se jedná o reálná data, na které budou mít vliv i další proměnné, je vysvětlení zhruba 32 % procent rozptylu počtu vypůjčených kol jedinou proměnnou poměrně dobrý výsledek. Z tabulky *Koeficienty* na obrázku č. 39 lze z nestandardizovaného koeficientu zjistit, že každé navýšení teploty o 1 °C navýší počet vypůjčených kol zhruba o 30 kusů s tím, že při nulovém vstupu teploty se vypůjčí zhruba 348 kol.

Zajímavý je také pohled na rozdíl jednoduché lineární regrese pro stejnou nezávislou proměnnou, tedy teplotu, avšak pro odlišné roční období. Na následujících obrázcích je vidět jednoduchý lineární regresní model pro počet vypůjčených kol dle teploty v letních měsících (obr. č. 40) a v zimních měsících (obr. č. 41).

Sumarizace modelu

Model	R Roční období = Léto (Zvolené)	R na druhou	Upravené R na druhou	Směrodatná chyba odhadu
1	,163[a]	,027	,026	681,115

a. Prediktory: (Konstanta), Teplota

Koeficienty[a,b]

Model		Nestandardizované koeficienty		Standardizované koeficienty		
		B	Standardní chyba	Beta	t	Významnost
1	(Konstanta)	390,751	83,926		4,656	,000
	Teplota	24,201	3,110	,163	7,782	,000

a. Závislá proměnná: Počet vypůjčených kol

b. Zvoleny pouze záznamy, které splňují podmínku: Roční období = Léto

Obrázek 40 – jednoduchá lineární regrese počtu vypůjčených kol dle teploty v letních měsících, vlastní zpracování v SPSS

Sumarizace modelu

Model	R Roční období = Zima (Zvolené)	R na druhou	Upravené R na druhou	Směrodatná chyba odhadu
1	,383[a]	,147	,146	138,948

a. Prediktory: (Konstanta), Teplota

Koeficienty[a,b]

Model		Nestandardizované koeficienty		Standardizované koeficienty		
		B	Standardní chyba	Beta	t	Významnost
1	(Constant)	252,280	3,297		76,529	,000
	Teplota	10,525	,547	,383	19,252	,000

a. Závislá proměnná: Počet vypůjčených kol

b. Zvoleny pouze záznamy, které splňují podmínku: Roční období = Zima

Obrázek 41 – jednoduchá lineární regrese počtu vypůjčených kol dle teploty v zimních měsících, vlastní zpracování v SPSS

Pro letní měsíce je hodnota R a potažmo R^2 značně menší než v zimních měsících, což naznačuje, že v zimním období je teplota vlivnějším faktorem než v letním období. A tak i díky hodnotě nestandardizovaného koeficientu závislé proměnné (konstanta), která je v létě 390 oproti zimním 252, tedy v létě se bez započtení vlivu teploty vypůjčí více kol než v zimě, se dá předpokládat, že hodnoty teploty v letním období jsou pro zákazníky dostatečně vhodné na to, aby rozdíl v teplotě byl zanedbatelným faktorem (hodnota R^2 = 0,027 je v lineární regresi nevyužitelná) a tak v létě mají na počet vypůjčených kol vliv jiné, zde nevyužité proměnné. V zimě je teplota o něco podstatnějším faktorem, i když zhruba 15 % vysvětleného rozptylu není příliš velká hodnota. Toto jsou užitečná zjištění, které je vhodné vzít v potaz při tvorbě doporučení a vyhodnocení datové analýzy.

7.3.3 Mnohonásobná lineární regrese

Jedna proměnná však nebývá dostatečná k tomu, aby výsledný model dodával dostačující výsledky a je vhodné využít všech dostupných dat. Předchozí jednoduchou lineární regresi využívající nezávislou proměnnou teplota je tak možné rozšířit o další nezávislé proměnné z datového souboru. Do lineární regrese při využití metody postupného vkládání proměnných vložíme další nezávislé proměnné, díky čemu je možné sledovat vliv nově přidané nezávislé proměnné na výsledný model. Tento postup usnadní nalezení optimální kombinace nezávislých proměnných při využití dostupného datového souboru. Na obrázku č. 42 je tak vidět sumarizace jednotlivých modelů a na obrázku č. 43 matice koeficientů pro jednotlivé modely, která je na první pohled poměrně rozsáhlá, avšak obsahuje potřebné informace.

Sumarizace modelu

Model	R Provozní den = provozní den (Zvolené)	R na druhou	Upravené R na druhou	Směrodatná chyba odhadu
1	,563[a]	,317	,317	531,021
2	,667[b]	,445	,445	478,389
3	,701[c]	,491	,491	458,383
4	,710[d]	,504	,503	452,650
5	,715[e]	,512	,511	449,075
6	,716[f]	,512	,512	448,810
7	,716[g]	,512	,512	448,722

a. Prediktory: (Konstanta), Teplota

b. Prediktory: (Konstanta), Teplota, Hodina

c. Prediktory: (Konstanta), Teplota, Hodina, Vlhkost

d. Prediktory: (Konstanta), Teplota, Hodina, Vlhkost, Dešťové srážky v mm

e. Prediktory: (Konstanta), Teplota, Hodina, Vlhkost, Dešťové srážky v mm, Solární radiace

f. Prediktory: (Konstanta), Teplota, Hodina, Vlhkost, Dešťové srážky v mm, Solární radiace, Vizibilita

g. Prediktory: (Konstanta), Teplota, Hodina, Vlhkost, Dešťové srážky v mm, Solární radiace, Vizibilita, Teplota rosného bodu

Obrázek 42 – Sumarizace modelu při metodě postupného vkládání proměnných, vlastní zpracování

Koeficienty[a,b]

Model		Nestandardizované koeficienty		Standardizované koeficienty			Statistika multikolinearity	
		B	Směrodatná chyba	Beta	t	Významnost	Tolerance	VIF
1	(Konstanta)	347,771	8,390		41,449	,000		
	Teplota	29,863	,477	,563	62,626	,000	1,000	1,000
2	(Konstanta)	-8,372	11,030		-,759	,448		
	Teplota	27,508	,433	,518	63,549	,000	,985	1,015
	Hodina	33,564	,757	,362	44,335	,000	,985	1,015
3	(Konstanta)	432,360	19,202		22,516	,000		
	Teplota	29,867	,424	,563	70,517	,000	,945	1,059
	Hodina	28,151	,752	,303	37,453	,000	,917	1,090
	Vlhkost	-7,026	,256	-,224	-27,491	,000	,906	1,104
4	(Konstanta)	379,011	19,305		19,633	,000		
	Teplota	29,871	,418	,563	71,420	,000	,945	1,059
	Hodina	28,953	,744	,312	38,903	,000	,912	1,096
	Vlhkost	-6,098	,260	-,194	-23,442	,000	,852	1,173
	Dešťové srážky v mm	-66,418	4,512	-,116	-14,721	,000	,939	1,065
5	(Konstanta)	511,887	22,281		22,975	,000		
	Teplota	32,608	,477	,614	68,414	,000	,716	1,397
	Hodina	28,520	,739	,307	38,578	,000	,910	1,099
	Vlhkost	-8,076	,309	-,258	-26,158	,000	,596	1,679
	Dešťové srážky v mm	-64,246	4,480	-,113	-14,340	,000	,937	1,067
	Solární radiace	-84,889	7,273	-,115	-11,672	,000	,598	1,674
6	(Konstanta)	423,651	34,725		12,200	,000		
	Teplota	32,200	,492	,607	65,444	,000	,671	1,490
	Hodina	28,686	,741	,309	38,736	,000	,906	1,104
	Vlhkost	-7,376	,374	-,235	-19,728	,000	,406	2,465
	Dešťové srážky v mm	-63,660	4,481	-,112	-14,207	,000	,936	1,069
	Solární radiace	-79,013	7,482	-,107	-10,560	,000	,564	1,773
	Vizibilita	,033	,010	,031	3,311	,001	,642	1,557
7	(Konstanta)	475,757	42,819		11,111	,000		
	Teplota	30,032	1,153	,566	26,042	,000	,122	8,191
	Hodina	28,721	,741	,309	38,781	,000	,906	1,104
	Vlhkost	-7,953	,465	-,254	-17,085	,000	,262	3,822
	Dešťové srážky v mm	-63,166	4,486	-,111	-14,079	,000	,933	1,072
	Solární radiace	-77,948	7,498	-,105	-10,396	,000	,561	1,782
	Vizibilita	,032	,010	,031	3,240	,001	,642	1,559
	Teplota rosného bodu	,254	,122	,050	2,079	,038	,101	9,909

a. Závislá proměnná: Počet vypůjčených kol

b. Zobrazeny pouze záznamy, které splňují podmínku: Provozní den = provozní den

Obrázek 43 – tabulka koeficientů pro jednotlivé modely sestavené metodou postupného vkládání proměnných, vlastní zpracování

Ze sumarizace modelu lze vidět nejvyšší dosažená hodnota R^2 = 0,512. Ideální by byla hodnota vyšší než 0,7, nicméně i hodnota přes 0,5, tedy vysvětlení více než 50 % rozptylu počtu vypůjčených kol je dostatečná na to, aby bylo možné z datové analýzy vyvodit závěry a případná doporučení pro management. Mezi modely 5, 6 a 7 jsou již minimální rozdíly v atributech sumarizace modelu, což jsou přidané proměnné *vizibilita* a *teplota rosného bodu*. Z korelační matice již byla vidět potencionální multikolinearita mezi vizibilitou a vlhkostí, která se tak nejspíše projevuje i v modelu tím, že již nenavyšuje hodnotu R^2. Pro nezávislou proměnnou teplota rosného bodu lze díky statistice multikolinearity očekávat obdobný problém díky VIF hodnotě blížící se 10. To lze potvrdit z korelační matice na obrázku č. 44.

Korelační matice

		Teplota	Teplota rosného bodu
Teplota	Koeficient	1	,870**
Teplota rosného bodu	Koeficient	,870**	1

**. Korelace je významná na hranici 0.01 (dvoustranná).

Obrázek 44 – párová korelační matice teploty a teploty rosného bodu, vlastní zpracování

Ze sestavených modelů metodou postupného vkládání tak bude využit model číslo 5, tedy při využití proměnných teploty, hodiny, vlhkosti, dešťových srážek a solární radiace. Ten má hodnotu R^2 = 0,512 a výsledné koeficienty jsou zobrazeny na obrázku č. 45. Výsledný funkční vztah se tak dá zapsat následovně:

$$\boldsymbol{počet\ vypůjčených\ kol} = 511{,}9 + 32{,}6 \cdot teplota + 28{,}5 \cdot hodina - 8{,}1 \cdot vlhkost - 64{,}3 \cdot dešťové\ srážky - 84{,}9 \cdot solární\ radiace$$

Koeficienty[a,b]

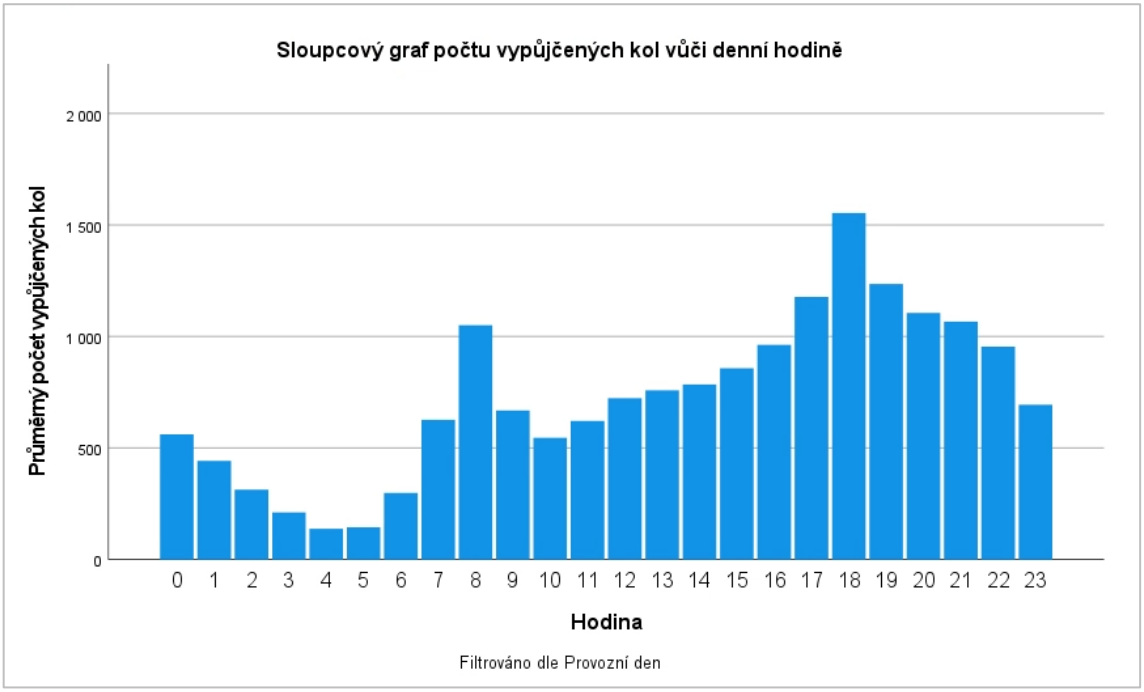
Model		Nestandardizované koeficienty		Standardizované koeficienty			Korelace			Statistika multikolinearity	
		B	Standardní chyba	Beta	t	Významnost	Nultého řádu	Parciální	Část	Tolerance	VIF
1	(Konstanta)	511,887	22,281		22,975	,000					
	Teplota	32,608	,477	,614	68,414	,000	,563	,597	,520	,716	1,397
	Hodina	28,520	,739	,307	38,578	,000	,425	,387	,293	,910	1,099
	Vlhkost	-8,076	,309	-,258	-26,158	,000	-,202	-,274	-,199	,596	1,679
	Dešťové srážky v mm	-64,246	4,480	-,113	-14,340	,000	-,129	-,154	-,109	,937	1,067
	Solární radiace	-84,889	7,273	-,115	-11,672	,000	,274	-,126	-,089	,598	1,674

a. Závislá proměnná: Počet vypůjčených kol

b. Zvoleny pouze záznamy, které splňují podmínku: Provozní den = provozní den

Obrázek 45 – koeficienty zvolených proměnných pro výsledný lineární regresní model, vlastní zpracování

Z tabulky koeficientů na obrázku č. 45 je vidět, že nezávislé proměnné teplota a hodina jsou důležitými faktory pro počet vypůjčených kol. Ačkoliv denní hodina přispívá k vysvětlení rozptylu, interpretace že vyšší hodina, ordinální proměnná, má pozitivní vliv na počet vypůjčených kol je nesmyslná. Když už jsou známy proměnné využité v modelu, je vhodné některé z nich ještě blíže analyzovat, aby bylo možné nalézt co nejvíce vhodných informací využitelných v managementu. Například zjištění rozložení počtu vypůjčených kol v průběhu dne může být přínosnou informací. Na obrázku č. 46 je tak vidět sloupcový graf počtu vypůjčených kol vůči denní hodině, filtrováno pouze pro provozní dny.

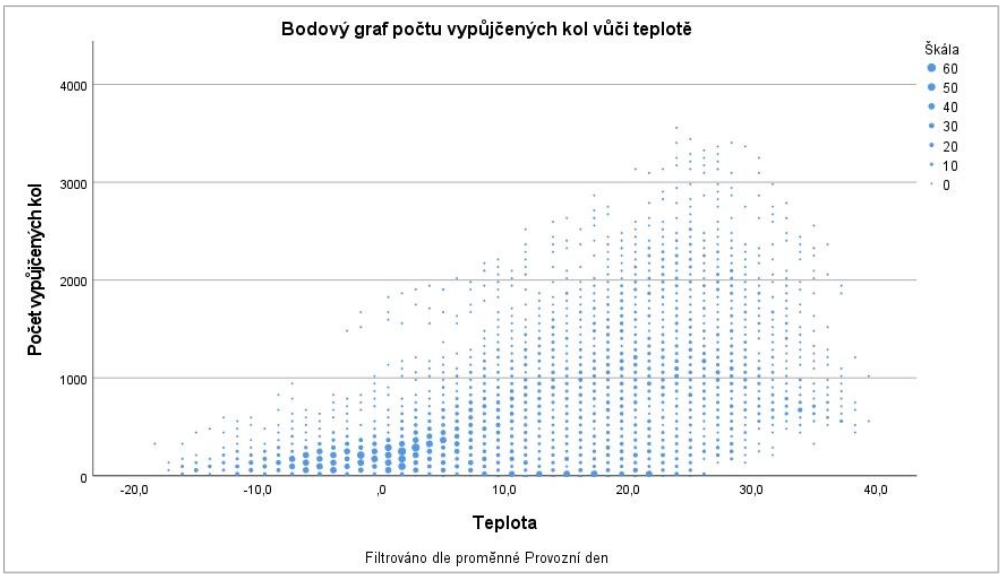


Obrázek 46 – Sloupcový graf počtu vypůjčených kol vůči denní hodině pro provozní dny, vlastní zpracování

Ve sloupcovém grafu je vidět pokles k brzkým ranním hodinám a naopak kolem 8 hodiny ranní a 18 hodiny večerní jsou vidět značná navýšení. Jedno z možných vysvětlení by mohlo být takové, že je služba využívána jako dopravní prostředek do práce a zpět domů. To však z dostupných dat nelze jednoznačně potvrdit ani vyvrátit.

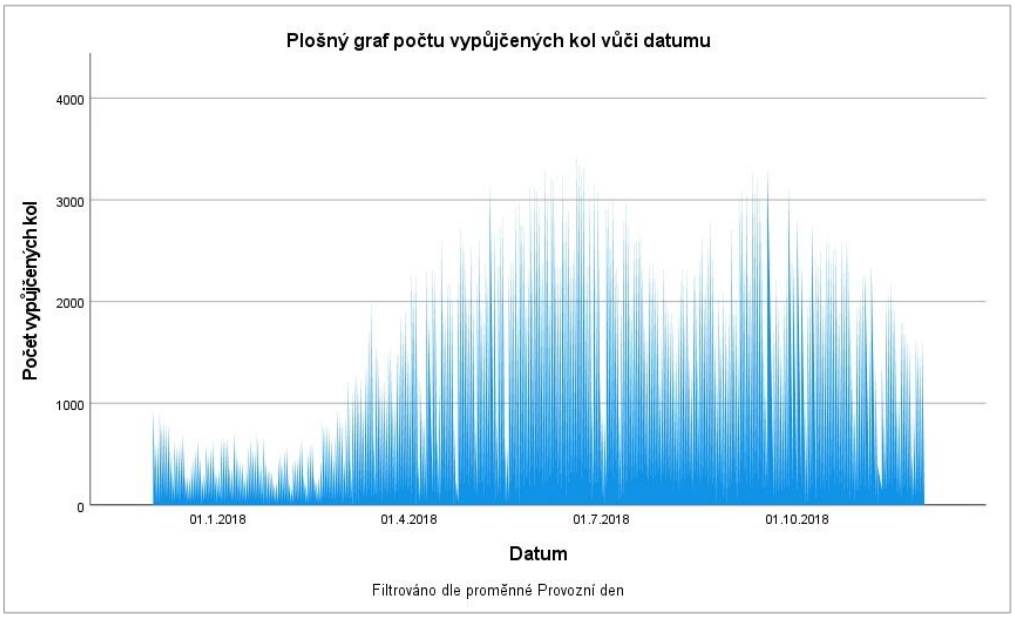
Dále je vhodné zobrazit bodový graf počtu vypůjčených kol vůči teplotě a podívat se na toto rozložení. Vzhledem k počtu záznamů v datovém souboru bude využita procedura *Bin elements*, která sumarizuje nejbližší body do jednoho. Počet takto sumarizovaných záznamů v jednom bodě je v případě obrázku č. 47 reprezentován velikostí bodu a škála je v pravém horním rohu. Toto nastavení je možné provést v okně *Chart Editor* a následně v podokně *Properties*, kde na kartě *Binning* se nastaví indikátor počtu záznamů na *Marker Size* a na kartě *Marker* je následně upravena požadovaná barva. Z grafu lze usuzovat, že při překročení jisté teplotní

hranice začíná počet vypůjčených kol klesat, což je v grafu vidět zhruba od 30 °C a výše, nicméně mimo tuto část je zde rostoucí tendence počtu vypůjčených kol s vyšší teplotou



Obrázek 47 – Bodový graf počtu vypůjčených kol vůči teplotě při využití procedury Bin elements, vlastní zpracování

Dobré by bylo znát tako rozložení množství vypůjčených kol v průběhu celého roku, jak je vidět na plošném grafu na obrázku č. 48 (pro prezentaci může být preferovaná volba sloupcového grafu agregovaného například do týdnů či měsíců, avšak pro přehled využitý spojnicový graf dostatečně dobře ukazuje celkové rozložení a zároveň i případné výkyvy, které by v agregovaném grafu nebyly vidět). Z tohoto grafu lze tak vidět znatelný pokles v zimních měsících, který bude pravděpodobně souviset s teplotou. Z grafu je též vidět pokles počtu vypůjčených kol ke konci letních prázdnin, tedy na přelomu srpna a září.



Obrázek 48 – plošný graf počtu vypůjčených kol vůči datumu filtrováno dle provozního dne, vlastní zpracování

Bodový graf na obrázku č. 47 ukázal vliv teploty na počet vypůjčených kol, avšak pro konkrétní hodnoty ideální teploty pro maximální počet vypůjčených kol je nutné provést ještě dodatečný výpočet. Nejdříve se vytvoří nová umělá proměnná, která bude agregovat teplotu do kategorií po 5 °C (např. hodnota 2,5 bude představovat rozpětí 0–5 °C). Následně je potřeba sečíst počet vypůjčených kol pro každou kategorii. To je možné provést na kartě *Data – Aggregate…* kde se zvolí funkce *Sum* pro proměnnou počtu vypůjčených kol, funkce *Number of cases* opět pro proměnnou vypůjčených kol (slouží jako ukazatel počtu záznamů) a do pole *Break Variable* se vloží proměnná agregované teploty. Výsledná tabulka se ještě rozšíří o vypočtenou hodnotu vypůjčených kol na záznam, tedy pomocí karty *Transform – Compute Variable*, kde se vydělí suma vypůjčených kol počtem záznamů. Výsledkem je tabulka na obrázku č. 49.

	Agregovaná_teplota	Počet_záznamů	Počet_vypůjčených_kol_suma	Vypůjčených_kol_na_záznam
1	-17,5	46	4,814	104,65
2	-12,5	182	27,435	150,74
3	-7,5	470	83,782	178,26
4	-2,5	756	174,073	230,26
5	2,5	1064	351,438	330,30
6	7,5	1079	558,699	517,79
7	12,5	1042	706,286	677,82
8	17,5	1222	1,025,887	839,51
9	22,5	1397	1,464,797	1048,53
10	27,5	990	1,208,626	1220,83
11	32,5	423	486,421	1149,93
12	37,5	89	80,056	899,51

Obrázek 49 – tabulka agregované teploty, počtu záznamů a sumě vypůjčených kol, vlastní zpracování

Z tabulky je tak důležitý údaj počtu vypůjčených kol na záznam, což je tedy průměrný počet vypůjčených kol pro danou teplotní agregaci. Zde je vidět ideální rozsah mezi 20–35 °C, kdy 27,5 °C je nejideálnější teplotou.

Závěry a doporučení pro management

Datový soubor obsahuje dostatečné množství dat na to, aby bylo možné vyhodnotit i sezónnost služby vypůjčení kol a je tak možné konstatovat její poměrně velkou sezónnost, kdy ve městě Soul v zimním období množství vypůjčených kol nepřesáhne 1 000 kusů a blíží se spíše hranici 500 kusů, vůči létu, kde se vypůjčuje i přes 3 000 kol. Zimní měsíce tak snižují výnosnost tím, že je značné množství kol nevyužitých, a navíc se dá předpokládat navýšení nákladů jejich uskladněním. V ideálním případě by bylo vhodné kola využít pro jiné účely, pokud tato možnost neexistuje, tak dle dosavadních dat přes zimní období je dostačující množství 1 000 kusů připravených kol a zbytek kol je vhodné uskladnit. Naopak v letním období se dosáhlo maxima 3556 vypůjčených kol, avšak záznamů s takto vysokými hodnotami je opravdu málo. Dle obrázku č. 50 má 99 % percentil hodnotu 2 530, což je o více než 1 000 kol méně, než je maximum, a přitom se jedná o pouhé 1 % všech záznamů. Jelikož nejsou známy náklady a výnosy na 1 kolo, není možné dát konkrétní doporučení, ale určitě stojí za zhodnocení potřebné množství připravených kol pro zákazníky. Za předpokladu, že maximum počtu vypůjčených kol je zároveň množství vlastněných kol, by tak při snížení počtu vlastněných kol na hodnotu 2 530 nebylo možné zcela uspokojit poptávku v 1 % všech případů, avšak by se snížil počet vlastněných kol o 29 %, které by jistě značně snížilo náklady na údržbu a nákup kol. Do finálního rozhodnutí však mohou vstupovat další proměnné, jako je spokojenost zákazníků, a tedy nárok vždy obsloužit všechny zákazníky a další případné požadavky.

Percentily

Počet vypůjčených kol

N	Validní	8760
	Chybějící	0
Percentily	5	22,00
	10	64,00
	50	504,50
	95	2043,00
	99	2530,34

Obrázek 50 – Tabulka percentilů pro počet vypůjčených kol v SPSS, vlastní zpracování

Se sezónností tak souvisí i teplota, která má dopad na počet vypůjčených kol. Výsledky analýzy ukázaly, že větší teplota nutně neznamená větší počet vypůjčených kol. Z tabulky č. 49 je tak vidět ideální rozsah mezi 20–35 °C, kdy teplota 27,5 °C je tou nejideálnější pro výslednou hodnotu vypůjčených kol

V datech se dá předpokládat rozdělení účelu vypůjčených kol jako dopravního prostředku a také jako volnočasová sportovní aktivita. Bohužel z datového souboru takovéto rozdělení nelze určit, natož potvrdit či vyvrátit. Nicméně za předpokladu využití vypůjčení kol jako volnočasové sportovní aktivity je tyto údaje možné generalizovat všeobecněji na venkovní sportovní aktivity, kdy jsou pro tyto účely nalezené vhodné podmínky pro teplotu a sezónnost platné. To může být podstatnou informací pro potencionální rozšiřování služeb, kdy je dobré dopředu počítat s problematikou zimního poklesu zájmu o tyto služby, případně rozšíření do oblastí s chladnějším podnebím.

Určitě je žádoucí vždy před hlavními sezónami pro vypůjčení kol, které dle plošného grafu na obrázku č. 48 jsou před zahájením letních prázdnin a v září, zkontrolovat stav všech kol a provést opravy, případně navýšit kapacity pro potřeby okamžitých oprav v těchto obdobích, protože se očekává nejvyšší vytíženost a při nefunkčním stavu jen malého množství kol by hrozila nedostupnost služby pro část zákazníků.

Pro případnou expanzi do dalších měst či jiných zemí jsou na základě využitého datového souboru ideální volbou lokace, které mají v průběhu roku co nejvyšší podíl dní s teplotami okolo 27,5 °C. Z regresního modelu však bylo zjištěno, že dostupné proměnné vysvětlují zhruba půlku celkového rozptylu vypůjčených kol, stále tak existují proměnné, které ovlivňují počet vypůjčených kol, o kterých nejsou dostupné informace. Důležitým faktory mohou být množství obyvatel ve městě, kde je služba zavedena, infrastruktura, sociálně-demografické faktory a mnoho dalších. Pro konkrétnější doporučení k expanzi by tak bylo vhodné model doplnit o další relevantní data, pokud by je bylo možné zajistit.

8 Shrnutí výsledků

Využitá literatura pro vybrané metody analýzy dat potvrdila značnou rozsáhlost této problematiky a jelikož byla cílem aplikace vybraných metod analýzy dat pro účely managementu a jejich následné provedení, ukázalo se náročné metody na jednotlivé případy správně aplikovat. Korelace jako statistická metoda pro aplikaci v managementu byla vybrána pro svou schopnost vyjádřit souvztažnost mezi dvěma sledovanými proměnnými. To se v rámci příkladů provedené v této práci potvrdilo jako přínosné, protože lze díky jejím výsledkům lépe poznat nastalé jevy, jako je míra souvztažnosti mezi počasím a počtem vypůjčených kol, které má nezanedbatelný vliv. Matematické vyjádření takovýchto jevů tak společnostem pomůže cíleněji zaměřit jejich aktivity pro splnění podnikových cílů. Lineární regrese byla zvolena pro její prediktivní vlastnosti a kvantifikaci nalezených výsledků. V dnešní digitální době vede značná část společností záznamy o prodejích a jiných vlastních aktivitách. Cílem tak bylo využití takovýchto dat pro sestavení lineárního modelu, který je schopen popsat a kvantifikovat jednotlivé jevy mající vliv na výsledky v historii, dnes i v budoucnu. Tento cíl se v rámci jednotlivých příkladů podařilo splnit, například v datové sadě kvality betonu to umožňuje společnosti nalézt vhodné složení jednotlivých komponent tak, aby splnilo požadavky na jeho vlastnosti zadavatelem. To tak pomůže společnosti flexibilněji reagovat na různá výběrová řízení

Postupně tak byly vybrány ty oblasti, které jsou buď pro konkrétní metodu analýzy dat nepostradatelné, nebo jsou využitelné v oblasti managementu. Na dílčích menších příkladech, které jsou vzhledem k obtížně získatelným reálným datům vymyšlené, jsou předvedena využití jednotlivých komponent analýzy dat při využití korelační a lineární regresní analýzy, které potvrzují jejich pozitivní informační přínos. Tyto příklady však byly z části inspirovány skutečnými problémy, se kterými se autor práce setkal a reprezentují tak skutečně řešené problémy v managementu, ve kterých vybrané metody analýzy dat najdou své využití. Konkrétní matematické vyjádření výsledků, například že při navýšení teploty v jarních měsících o 2 °C dojde k navýšení počtu vypůjčených kol o 65 kusů, je při správné aplikaci lineární regresní analýzy lepším podkladem pro rozhodování, než subjektivní znalost a zkušenost jednotlivých osob v managementu.

Výpočet konkrétních ukazatelů pro management, jako je odhadovaný prodej při určitém stavu nezávislých proměnných, výpočet norem a jim podobné metriky jsou díky možnostem dostupného statistického softwaru relativně jednoduchým úkonem, avšak jak ukázala praktická část nad reálnými daty, některé výsledky je obtížné interpretovat nebo z nich generalizovat nalezené poznatky. To platí

obzvláště pro nejednoznačné grafy, kdy například doporučení z odborné literatury, že by sestavený graf neměl vykazovat známky vzorce uspořádání proměnných může být náročné. Obdobně to platí i pro korelační koeficienty, u kterých sice lze v odborné literatuře nalézt doporučení pro konkrétní hodnoty a hranice, avšak ve chvíli, kdy se z nich mají sestavit výsledná doporučení pro management, která mohou značně ovlivnit chod celé společnosti se již nejeví tak jednoznačně.

Jako neobtížnější úkol provedení analýzy dat při využití korelace a lineární regrese pro účely managementu se tak jeví převedení teoretických poznatků na reálnou datovou sadu a tvorbu konkrétních výsledků a doporučení pro oblast zájmu managementu. Právě různorodost obsahu takovýchto datových sad a odlišných požadavků na cíle analýzy dat vede k nemožnosti vytvoření univerzálního postupu a vyhodnocení analýzy dat při využití korelace a lineární regrese. A tak co nejlepší znalost oblasti podnikání, respektive datové sady, je bezpochyby velkým přínosem pro její úspěšné provedení.

9 Závěry a doporučení

V průběhu práce se potvrdil přínos metod korelační a lineární regresní analýzy dat v managementu. Na otázky, které management může klást a na které potřebuje odpovědi pro úspěšné vedení podniku, tak tyto metody mohou dodat konkrétní, případně přibližné hodnoty. Tyto metody tak sice ne vždy jsou schopny dodat jednoznačné výsledky, ale při správném využití jsou schopny dodat nové poznatky, které managementu pomohou dělat informovanější rozhodnutí. Takováto rozhodnutí tak jsou následně podepřena i datově, což je důležité pro dlouhodobé konzistentní vedení společnosti. Dal by se namítnout přínos dílčích příkladů v teoretické části, které jsou vymyšlené, že byly sestaveny účelně pro potřebu dané ukázky. To nelze jednoznačně vyvrátit, avšak vzhledem k jejich obsahu je vidět jejich reprezentativní hodnota pro aplikaci vybraných metod analýzy dat v managementu jako celku a nejedná se o typově jedinečné úlohy, které daná metoda je schopna řešit.

Výsledky lineární regresní analýzy tak mohou pomoci s vnitropodnikovými procesy, jako je tvorba norem pro zaměstnance, nastavením realizovatelných cílů a úpravou jednotlivých částí procesů pro jejich efektivnějších plnění. Dále zjištěním vlivu vnějších jevů na aktivity podniku, ať jsou prodejního nebo jiného charakteru. To může zahrnovat chování trhu, zákazníků, aktivity konkurence a další, ke kterým je obvykle možné požadovaná data zakoupit, a tedy využít možností korelační a regresní analýzy pro získání lepšího vhledu do oblasti podnikání a jejího aktuálního stavu. Korelační a lineární regresní analýza jsou tak schopny pomoci společnosti jeho vnitřní procesy usměrňovat tak, aby lépe směřovaly k plnění cílů a reflektovaly aktuální podnikatelské prostředí, vytvářet krizové scénáře a nalézt způsoby, jak na ně efektivně reagovat. Obdobným způsobem je možné tvořit modely chování trhu pro oblast zájmu, což může společnosti pomoci lépe poznat jejich klíčového zákazníka a zmiňované procesy adekvátně upravit. Je tak více oblastí v managementu, kde je možné využít korelační a lineární regresní analýzu dat, které lze vzhledem k různorodosti tohoto prostředí jen stěží jednoznačně definovat, avšak jednotlivé příklady ukázali šíři využitelnosti těchto metod.

Za účelem zkvalitňování aplikace korelační a lineární regresní analýzy dat, tak lze doporučit kritickou debatu nad nalezenými výsledky s kvalifikovanými osobami, ať už s jinými výzkumníky či osobami znalými oblasti zájmu podnikání. Samotné provádění analýzy dat za pomoci těchto metod je jedním z nejlepších způsobů, jak se naučit lépe zpracovávat, analyzovat a vyhodnocovat nové datové soubory. Čím více takovýchto analýz dat výzkumník provede, tím jistější si může být v nalezených výsledcích a snad i nacházet možné vzorce napříč spektrem analyzovaných dat, které můžou mít značný přínos pro management.

Aplikace těchto metod analýzy dat tak díky softwaru SPSS je, co se do samotného provedení týče, poměrně jednoduchá i rychlá, avšak je potřeba si uvědomit, že na jejich základě se budou činit rozhodnutí, která mohou ovlivnit chod celé společnosti a jejich výsledků. Je tak potřeba s takovouto vážností k analýze dat přistupovat a zcela porozumět každé využité proceduře a grafu v SPSS, ujistit se, že je datový soubor relevantní a vhodný k zadanému úkolu a před samotnou prezentací výsledku pro management případně nálezy prodiskutovat s kvalifikovanou osobou, pokud je to možné.

10 Seznam použité literatury

[1] Hebák, Petr. Statistické myšlení a nástroje analýzy dat. 2. vydání. Praha: Informatorium, 2015. ISBN 978-80-7333-118-4.

[2] Verma, J., 2012. Data Analysis in Management with SPSS Software. Springer Science & Business Media.

[3] Ott, R. L. and Longnecker, M. (2016). *An Introduction to Statistical Methods and Data Analysis*, 7th Edition, Cengage Learning.

[4] Malhotra, Naresh K. a David F. Birks. *Marketing research: an applied approach*. 2nd European ed. Harlow: Financial Times /Prentice Hall, c2003. ISBN 0273657445.

[5] Homepage | statistika.vse.cz [online]. Copyright © [cit. 10.08.2022]. Dostupné z: https://statistika.vse.cz/download/materialy/tabulky.pdf

[6] Mareš, Petr, Ladislav Rabušic a Petr Souku. Analýza sociálněvědních dat (nejen) v SPSS. Brno: Masarykova univerzita, 2015. ISBN 978-80-210-6362-4.

[7] Provost, Foster a Tom Fawcett. *Data science for business: what you need to know about data mining and data-analytic thinking*. Sebastopol: O´Reilly, c2013. Data Science/Business. ISBN 978-1-4493-6132-7.

[8] SPSS Statistics | IBM. [online]. Copyright © Copyright IBM Corporation 2022 [cit. 12.06.2022]. Dostupné z: https://www.ibm.com/products/spss-statistics

[9] SPSS - About SPSS Inc.. SPSS, Data Mining, Statistical Analysis Software, Predictive Analysis, Predictive Analytics, Decision Support Systems [cit. 12.06.2022]. Dostupné z: http://www.spss.com.hk/corpinfo/history.htm

[10] IBM Docs. [online]. Copyright © Copyright IBM Corporation 2022 [cit. 12.06.2022]. Dostupné z: https://www.ibm.com/docs/en

[11] Yeh, I-Cheng, "Modeling slump flow of concrete using second-order regressions and artificial neural networks," Cement and Concrete Composites, Vol.29, No. 6, 474-480, 2007.

Dostupné z: https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test
Datum stažení: 8.7.2022

[12] **a)** Sathishkumar V E, Jangwoo Park, and Yongyun Cho. 'Using data mining techniques for bike sharing demand prediction in metropolitan city.' Computer Communications, Vol.153, pp.353-366, March, 2020

b) Sathishkumar V E and Yongyun Cho. 'A rule-based model for Seoul Bike sharing demand prediction using weather data' European Journal of Remote Sensing, pp. 1-18, Feb, 2020

Dostupné z:
https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand
Datum stažení: 17.1.2022

[13] Standardized regression coefficients outside (-1,1). [online]. Copyright © Copyright IBM Corp. 2022 [cit. 10.08.2022]. Dostupné z: https://www.ibm.com/support/pages/standardized-regression-coefficients-outside-11

[14] A statistical test for the presence of heteroscedasticity. [online]. Copyright © Copyright IBM Corp. 2022 [cit. 15.08.2022]. Dostupné z: https://www.ibm.com/support/pages/statistical-test-presence-heteroscedasticity

11 Přílohy

Příloha 1 – datový soubor kvality betonu

O souboru dat

Beton je velmi složitý materiál. Požadované vlastnosti betonu nejsou určeny pouze obsahem vody, ale jsou ovlivněny i dalšími přísadami. Datový souboru obsahuje celkem 103 záznamů, 7 vstupních proměnných a 3 cílové proměnné

Atributy datového souboru

Vstupní proměnné

- vstupní proměnné složky jsou uvedeny v kilogramech na m^3 betonu

1) Cement
2) Struska – černý hrubozrnný materiál s ostrými hranami a skelným leskem
3) Popílek
4) Voda
5) Superplastifikátor – vysoce vodoredukující přísady snižují obsah vody
6) Hrubost – agregovaná
7) Jemnost – agregovaná

Cílové proměnné

1) Sednutí – zkouška konzistence betonu udávaná v cm
2) Průtok – zkouška průtoku betonu udávaná v cm
3) Pevnost v tlaku – udává velikost napětí, kterým je působeno na vzorek betonu při jeho porušení udávaného v Mpa

Dostupné z: https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test
Datum stažení: 8.7.2022

Zdroj dat udávaný na zdrojovém webu:
Dárce: I-Cheng Yeh
Email: icyeh '@' chu.edu.tw
Instituce: Department of Information Management, Chung-Hua University (Republic of China)
Další kontaktní informace: Department of Information Management, Chung-Hua University, Hsin Chu, Taiwan 30067, R.O.C.

Příloha 2 – datový soubor o službě vypůjčení kol ve městě Soul

O souboru dat

V současné době jsou v mnoha městech zavedeny půjčovny kol pro zvýšení pohodlí mobility. Je důležité, aby půjčovna kol byla k dispozici a přístupná veřejnosti ve správný čas, protože to zkracuje dobu čekání. Poskytnout městu stabilní nabídku půjčovny kol se nakonec stává hlavním problémem. Stěžejní částí je predikce počtu kol potřebných v každou hodinu pro stabilní zásobování půjčovnami kol.

Datový soubor obsahuje informace o počasí (teplota, vlhkost, rychlost větru, viditelnost, rosný bod, sluneční záření, sněžení, déšť), počet kol pronajatých za hodinu a informace o datu.

Atributy datového souboru

1. Datum: Ve formátu rok-měsíc-den
2. Počet_vypůjčených_kol – počet vypůjčených kol
3. Hodina – Hodina daného dne
4. Teplota – Teplota ve stupních Celsia
5. Vlhkost – Vlhkost v procentech
6. Rychlost_větru_ms – Rychlost větru udaná v metrech za sekundu
7. Vizibilita – Viditelnost na vzdálenost 10 metrů
8. Teplota_rosného_bodu – Teplota rosného bodu ve stupních Celsia
9. Solární_radiace – solární radiace udaná v MJ/m2
10. Dešťové_srážky_mm – Dešťové srážky udané v milimetrech (1 litr vody na 1 metr čtvereční)
11. Sněhové_srážky_cm – Sněhové srážky udané v centimetrech
12. Roční_období – výčet textových hodnot: Spring, Léto, Podzim, Zima
13. Svátek – výčet textových hodnot: Holiday/No holiday
14. Provozni_den – výčet textových hodnot: Yes (standardní pracovní den), No (den, kdy půjčovna nebyla v provozu)

Dostupné z:
https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand
Datum stažení: 17.1.2022

UNIVERZITA HRADEC KRÁLOVÉ
Fakulta informatiky a managementu
Akademický rok: 2020/2021

Studijní program: Aplikovaná informatika
Forma studia: Kombinovaná
Obor/kombinace: Aplikovaná informatika (ai3-k)

Podklad pro zadání BAKALÁŘSKÉ práce studenta

Jméno a příjmení: **David Pirkl**
Osobní číslo: **I1900129**
Adresa: Okružní 139, Srnojedy, 53002 Pardubice 2, Česká republika

Téma práce: Vybrané metody analýzy dat v managementu s SPSS
Téma práce anglicky: Selected methods of Data Analysis in Management with SPSS Software

Vedoucí práce: doc. RNDr. Pavel Pražák, Ph.D.
Katedra informatiky a kvantitativních metod

Zásady pro vypracování:

Cílem práce je použití vybraných metod analýzy dat s SPSS ve vybraných problémech managementu

1. úvod
2. Korelační analýza
3. Regresní analýza
4. Použití ve statistickém programu SPSS

Seznam doporučené literatury:

J.P. Verma: Data Analysis in Management with SPSS Software, Springer 2013.
N.K. Malhotra, D.F. Birks: Marketing Research, An Applied Approach, Prentice Hall, Harlow, 2006.

Podpis studenta: Datum:

Podpis vedoucího práce: Datum: