Univerzita Pardubice

Fakulta ekonomicko-správní

Ústav systémového inženýrství a informatiky

Regresní úlohy pro Data Mining

Josef Rubáš

Bakalářská práce

2013

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2012/2013

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Josef Rubáš**
Osobní číslo:	**E090117**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Regionální a informační management**
Název tématu:	**Regresní úlohy pro Data Mining**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Práce bude zaměřena na vypracování vzorových příkladů na použití lineární a logistické regrese pro potřebu předmětu Data Mining I.
Práce bude obsahovat:

- vysvětlení základních pojmů z oblasti DM a metod regrese;
- popis způsobu zpracování dat v data miningu;
- lineární a logická regrese dle metody CRISP;
- aplikace získaných poznatků na vybraná data.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

ANDĚL, Jiří. Matematická statistika. první. Praha: SNTL- Nakladatelství technické literatury, 1978, 352 s. 3918.
NG, Wee Keong. Advances in knowledge discovery and data mining: 10th Asia-Pacific Conference, PAKDD 2006, Singapore, April 9-12, 2006 : proceedings. New York: Springer, c2006, xxiv, 879 p. Lecture notes in computer science, 3918. ISBN 978-354-0332-060.
PYLE, Dorian. Data preparation for data mining: 10th Asia-Pacific Conference, PAKDD 2006, Singapore, April 9-12, 2006 : proceedings. San Francisco: Morgan Kaufmann, 1999, xix, 540 s. Lecture notes in computer science, 3918. ISBN 15-586-0529-0.

Vedoucí bakalářské práce: **doc. Ing. Pavel Petr, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **1. října 2012**
Termín odevzdání bakalářské práce: **30. dubna 2013**

doc. Ing. Renáta Myšková, Ph.D.
děkanka

L.S.

prof. Ing. Jan Čapek, CSc.
vedoucí ústavu

V Pardubicích dne 1. října 2012

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako Školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Univerzity Pardubice.

V Pardubicích dne 30. 4. 2013

Josef Rubáš

PODĚKOVÁNÍ:

Rád bych touto cestou poděkoval vedoucímu mé bakalářské práce doc. Ing. Pavlu Petrovi, Ph.D. za cenné rady, připomínky, ochotu, trpělivost a pomoc při zpracování této práce.

ANOTACE

Práce je věnována regresním úlohám v data miningu. Teoretická část seznamuje s historií a definicí data minigu. Podrobněji je pojednáno o CRISP-DM metodologii i o regresní analýze. Stěžejní částí bakalářské práce je aplikace regresních úloh (logistické a lineární) na třech praktických případech. Tištěná práce je doplněna o CD s multimediálními ukázkami řešení příkladů.

KLÍČOVÁ SLOVA

Data mining, historie data miningu, CRISP-DM, lineární regrese, logistická regrese.

TITLE

Regression problems in data mining.

ANNOTATION

The labour is devoted to regression problems in data mining. The theoretical part of this bachelor's dissertation briefly deals about history of data mining and definition of data mining. More attention is devoted to the CRISP-DM methodology and regression analysis. The fundamental point of the bachelor's dissertation is the application of regression problems (logistic and linear) on free practical cases.

KEYWORDS

Data mining, history of data mining, CRISP-DM, linear regression, logistic regression

OBSAH

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

SEZNAM ZKRATEK

CRISP-DM	Cross Industry Standart Proces for Data Mining
CSV	Comma separated value
ČR	Česká republika
ČSÚ	Český statistický úřad
DM	Data mining
EIS	Executive information System
KDD	Knowledge Discovery from Databases
OLAP	OnLine Analytical Processing
OLS	Ordinary least squares
SAS	Statistical analysis system
SEMMA	Sample, explore, modify, model, assess
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language

ÚVOD

Dobývání znalostí z databází se do povědomí odborné veřejnosti dostalo až počátkem 90. let. Databáze představují uspořádanou množinu dat a v průběhu vývoje se staly spolehlivým prostředkem k uchovávání rozsáhlých dat a poskytují také možnost vyhledávat v nich informace dle kritérií. Jedním z prostředků manipulace s daty je statistické šetření, jež je osvědčený prostředek, jak modelovat a analyzovat závislosti v datech. Po dlouhou dobu se tyto disciplíny vyvíjely nezávisle, až do chvíle, kdy rozsah automaticky nasbíraných dat překročil určitou mez a klasické analýzy již nestačily. Data mining se stal spojením databází a jejich analýzy pro potřeby rozhodovacích procesů. Jedním z typických příkladů je analýza nákupního koše zákazníka, na jejímž základě je upravován strategický plán prodeje.

Bakalářská práce je logicky rozčleněna na teoretickou a praktickou část. Teoretická část nejprve seznamuje se samotnou definicí data miningu, jeho základními zdroji a stručnou historií. Dále už se práce věnuje samotné metodice CRISP-DM a vymezuje mimo jiné fáze procesu data miningového zpracování. V poslední části teoretické přípravy jsou popsány regresní úlohy.

Stěžejní částí bakalářské práce je navazující praktická část, v níž budou nabyté poznatky aplikovány na třech vybraných příkladech z různých prostředí, a to ekonomického, ekonomicky-sociálního a ekonomicky-ekologického. Pro poslední dva uvedené příklady budou využity údaje Českého statistického úřadu. Experimentální část je v souladu s fázemi metodiky CRISP-DM, a to v krocích porozumění problému, porozumění datům, příprava dat, modelování dat, hodnocení výsledků a na závěr je provedena implementace. V každé databázi budou nalezeny vztahy, na základě kterých budou vytvořeny tři jednorozměrné a jedna vícerozměrná lineární regrese. Následně bude na stejných datech, se shodnými parametry, provedena aplikace logistické regrese.

Cílem bakalářské práce je najít skryté závislosti mezi prezentovanými daty, provést statistickou analýzu dat a zhodnotit získané výsledky.

1 DATA MINING A REGRESNÍ METODY

V rámci data miningu je možné aplikovat široké spektrum metod pro modelování dat s využitím různých analytických procedur. Jednou z nejpoužívanějších jsou regresní metody.

1.1 Data mining

Definice data miningu

Pojem data mining pochází z angličtiny a překládá se jako dolování dat nebo vytěžování dat. Data mining (DM) v sobě skýtá velké množství technik, postupů a algoritmů používaných k objevování smysluplných korelací, předloh a trendů při procházení velkého množství dat uložených nejčastěji v datových skladech. Jako obecně přijatá definice dolování dat se považuje:

„DM je netriviální dobývání skrytých předem neznámých a potenciálně užitečných informací z dat. Při jejich objevování se využívají expertní systémy a grafické a statistické techniky a prezentují se způsobem srozumitelným lidem“. (Everitt, str. 108)

DM se využívá v mnoha odvětvích, například k vyhodnocení návratnosti kampaní, analýze zákaznické databáze, hledání souvislostí a možných korelací, ve snaze vypátrat z dat všechno to, co pomůže efektivně řídit firmu, projekt či zaměstnance. [1]

Systémem DM je možné data i jednoduše zpracovat prostřednictvím rozpoznávacích technologií i grafických a statistických technik. Umožní tedy získaná data prezentovat srozumitelnou formou pro veřejnost či zadavatele. [1], [8]

Nicméně DM lze aplikovat i na malé databáze dat. Důležitou vlastností DM je, že se jedná o analýzy odvozované z obsahu dat, nikoliv předem určené uživatelem. Jedná se především o odvozování prediktivních informací, nikoliv pouze deskriptivních.

Přínosnou odlišností DM, oproti jiným statistickým nástrojům, je právě zaměření na uživatele. Statistické úlohy DM jsou realizovány automaticky podle určených algoritmů a uživatel tedy nemusí mít speciální znalosti statistiky. U jiných programů je uživatelem specialista zhotovující zprávy pro koncového zadavatele. [1], [8], [9]

Základní zdroje DM

V těchto systémech je možné definovat tři základní zdroje, a to databáze, statistiku a strojové učení.

1. Databáze

Do této skupiny řadíme veškeré systémy, v nichž jsou data shromažďována a které umožňují vyvolávání specifické skupiny dat. Patří sem tedy: relační databáze, EIS (Executive information system), OLAP (On-Line Analytical Processing), datový sklad a dotazovací jazyk SQL. [1]

2. Statistika

Představuje nástroje pro matematické zpracování dat, tj. kontingenční tabulky, regresní analýza, distribuční a korelační analýza. [1]

3. Strojové učení

Strojové učení je moderní, rychle neustále se rozvíjející technologie pro získávání znalostí a dat. Příkladem jsou teorie řízení a umělá inteligence. [1]

Stručná historie dobývání znalostí z databází

V podstatě první impulzy, jež předcházely vzniku DM, je možné datovat až do 30. a 40. let 20. století, kdy se společnosti SAS a SPSS zabývaly matematickými analýzami, jako jsou např. směrodatná odchylka, rozptyl, shluková analýza a intervaly spolehlivosti. První metody zaměřené k prozkoumávání databází a hledání skrytých spojitostí mezi daty začaly vznikat asi před 50 lety, tedy v 60. letech 20. století, kdy docházelo k masivnějšímu rozvoji výpočetní techniky, a vyvstala potřeba získávat užitečné údaje z uložených dat. Ale většinou šlo jen o ojedinělý výzkum na univerzitní půdě. [1], [3], [8]

V 70. a 80. letech 20. století byl rozvoj podpořen zvýšením výkonu a rychlosti počítačů. Díky tomu došlo i k rozmachu statistických metod, databázových aplikací a umělé inteligence. V této době se však o DM hovoří spíše jako o vytěžování dat (tedy DM), odvozování znalostí (data extraction), odkrývání informací, získávání znalostí a tyto pojmy mají spíše hanlivý význam, protože nebylo ještě možné zajistit spolehlivost jejich výsledků. Navíc šlo spíše o vyhledávání korelací ve velkých datových souborech. [1], [3], [8]

O skutečném dobývání znalostí z databází (z angličtiny Knowledge Discovery from Databases, tj. KDD) se začalo mluvit až počátkem 90. let minulého století. Poprvé byl představen ve Spojených státech Amerických, kde se roku 1989 na Mezinárodní konferenci o umělé inteligenci IJCAI´89 tento pojem ustanovil. KDD je definován jako „*proces netriviálního objevování implicitních, dopředu neznámých a potenciálně použitelných znalostí v datech.“* (Berka, str. 15)

V tomtéž roce na prvním workshopu o KDD byla zdůrazněna potřeba získávat z databází využitelné znalosti. Tedy až od 90. let, po objevu nových statistických metod, můžeme jednoznačně hovořit o DM jako o samostatném a rovnocenném oboru aplikované vědy. Z počátku se pro tuto oblast razily nejrůznější názvy: information harvesting, data archeology, data destilery, data dredging. Nakonec ovšem zvítězila hornická metafora – dobývání a dolování znalostí z dat (DM). [1], [3], [8]

1.2 Metodiky data miningu

„S postupem doby začaly vznikat metodiky, jejichž cílem je poskytnout uživatelům jednotný rámec pro řešení různých úloh z oblasti dobývání znalostí. Tyto metodiky umožňují sdílet a přenášet zkušenosti z úspěšných projektů.„ (Berka, str. 22)

Prostřednictvím DM lze řešit tyto typy úloh:

- Popis dat – vizualizace, sumarizace.
- Hledání „nugget“ – dominantní struktury, asociační pravidla, segmentace, shluková analýza, popis rozdělení dat.
- Predikce – klasifikace (predikce kategoriální proměnné), regrese (predikce spojité proměnné), časové řady (predikce proměnné závislé na čase).

Mezi metody DM patří například metodika 5A firmy SPSS nebo metodika SEMMA firmy SAS. Pro potřebu bakalářské práce se zaměřím na metodiku CRISP-DM, jejímž cílem je umožnit řešení rozsáhlých úkolů v dobývání znalostí z databází.

Metodika CRISP-DM

Metodika CRISP-DM (CRoss Industry Standard Processfor Data Mining) vznikla v rámci Evropského výzkumného projektu. Cílem bylo navrhnout univerzální postup, který bude možné aplikovat v nejrůznějších oblastech (př. komerční aplikace). Hlavními přednostmi CRISP-DM je rychlost, efektivita, spolehlivost a nižší náklady na „dobývání dat“. Kromě návrhu standardního postupu má CRISP-DM nabízet také „průvodce'“ potenciálními problémy a řešeními, které se mohou vyskytnout v reálných aplikacích.

Hierarchické uspořádá CRISP-DM

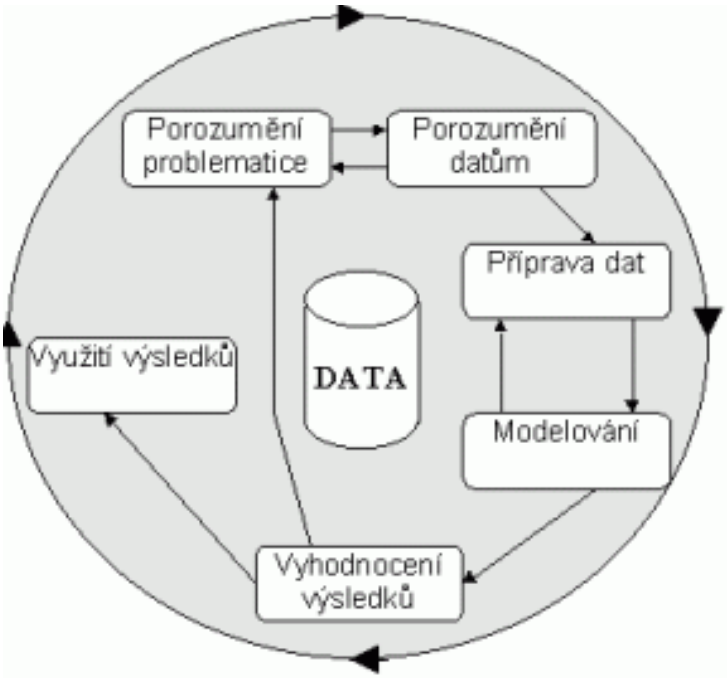
V rámci hierarchického uspořádání metodiky CRISP-DM rozeznáváme čtyři úrovně abstrakce: fáze projektu, obecné úkoly, specializované úkoly, aplikace v procesu. [7], [8]

Úrovně jsou uspořádány od obecných ke specifickým. V prvním kroku jsou definovány jednotlivé řady fází a na každou tuto fázi navazuje v druhé úrovni několik úkolů. Nicméně

druhá fáze musí být pořád velmi obecná, aby podchytila všechny možné DM situace. Úkoly na této úrovni jsou ale již úplné a stabilní. Pod pojmem úplné se v takovémto případě rozumí, že pokrývá celý postup DM a všechny možné DM aplikace. A za stabilní se považuje, když je model validní i pro zatím nepředvídatelný vývoj. V třetí fázi, specializované úlohy, jsou již obecné úlohy z druhé fáze převedeny na konkrétní akce podle řešeného problému. Čtvrtá úroveň představuje technickou realizaci specializovaných úloh. [1], [8]

Fáze CRISP-DM

Metodologie CRISP-DM rozděluje celý proces DM do šesti základních fází, v rámci nichž ještě rozlišujeme další kroky. Jednoznačně časově nejnáročnější fází je příprava dat, v procentuálním vyjádření se jedná přibližně o 55 % z celkového času. Pro fázi porozumění problému by mělo stačit 5 % času, další fáze projektu, porozumění datům, zabere přibližně 10 % času projektu. 15 % času zabere modelování a dalších 15 % připadne na sepsání závěrečné zprávy. [1], [8], [10]



Obrázek 1: Grafické vyjádření následnosti kroků v DM

Zdroj: [1]

- *Definování cílů* (Business understanding)

Tato úvodní fáze je zaměřena na pochopení cílů úlohy a požadavků, jež směřují k řešení formulovanému z manažerského hlediska. Manažerská formulace musí být následně převedena do zadání úlohy pro DM. V této fázi se rovněž provádí inventura zdrojů (datových,

výpočetních i lidských). Hodnotí se možná rizika, náklady a přínos použití metod DM. V neposlední řadě se stanovuje předběžný plán prací. [1], [8], [10]

- *Porozumění datům* (Data understanding)

V této etapě dochází k převzetí dat a seznámení se s nimi. Jedná se zejména o posouzení kvality dat a o vytipování možných podmnožin záznamů v databázi. V rámci porozumění datům zjišťujeme různé deskriptivní charakteristiky dat, jako jsou četnosti hodnot, průměr, minimum, maximum, atd. Tato analýza je zaměřena na DM otázky, které mohou být zodpovězeny použitím dotazů, vizualizací a reportů. Toto zahrnuje rozdělení klíčových atributů, vazby mezi páry, jednoduché statistické analýzy, četnosti hodnot různých atributů, průměrné hodnoty, minima a maxima apod. Tyto analýzy se mohou zaměřit přímo na cíl DM projektu a sloužit tak pro formulaci hypotéz, nebo pouze přispívat k popisu dat. Pokud je to vhodné, tak mohou být součástí diagramy a grafy, které poukazují na datové charakteristiky nebo které mohou být určitým vodítkem k zajímavým podskupinám v datech. V této fázi také dochází k vytváření prvních hypotéz, které se v průběhu celého procesu snažíme potvrdit. Někdy však můžeme hypotézu vyvrátit nebo naopak najít jiné řešení. [1], [8], [10]

- *Příprava dat* (Data preparation)

Příprava dat zahrnuje činnosti, které vedou k vytvoření datového souboru, který bude zpracován jednotlivými analytickými metodami. Dochází k integraci více datových zdrojů, čištění, úpravě a formátování dat do podoby vyžadované analytickými nástroji a metodami, které později budou na data aplikovány. Součástí přípravy dat je často standardizace dat. Tato fáze je obvykle nejpracnější částí řešení celé úlohy. Proces nelze správně provést bez znalosti dat. Špatná integrace dat by mohla vést ke znehodnocení zdrojů dat a ovlivnění celkové kvality řešení. [1], [8], [10]

- *Modelování* (Modeling)

Modelování představuje fázi aplikace analytických metod. Obvykle existuje řada různých metod pro řešení dané úlohy. Je třeba vybrat ty nejoptimálnější a vhodně nastavit jejich parametry pro řešení definovaného problému. Z tohoto kroku vybíráme několik nejlepších získaných řešení, které postupují do dalšího kroku. Ovšem jde tedy o iterační činnost, a použití analytických algoritmů může vést k potřebě modifikovat data, a tedy k návratu k datovým transformacím. Součástí této etapy je také ověřování nalezení znalostí z pohledu metod dobývání znalostí z databází. [1], [8], [10]

- *Hodnocení výsledků* (Evaluation)

V této fázi dochází ke konečnému hodnocení a selekci získaných modelů podle různých vlastností a ověření správnosti získaných řešení za pomoci těchto modelů. Dle získaných výsledků je již možno zvážit případnou implementaci celého procesu. [1], [8], [10]

- *Využití výsledků* (Deployment)

Vytvořením vhodného modelu řešení úlohy obecně nekončí. Dosažené výsledky je třeba vyhodnotit z pohledu zadavatele (př. manažera), zda byly splněny cíle formulované při zadání úlohy. A dále je nutné data zpracovat do podoby pochopitelné pro zadavatele, např. sepsání závěrečné zprávy, nebo zavedení systému pro automatickou klasifikaci nových případů atd. Ve většině případů je to až zadavatel, a nikoliv analytik, kdo provádí kroky k využití výsledků analýzy. [1], [8], [10]

1.3 Regresní úlohy

Definice regresní funkce: „*Nechť X a Y jsou náhodné veličiny. Podmíněnou střední hodnotou E(Y|x), považovanou za funkci proměnné x, budeme nazývat regresní funkcí náhodné veličiny X vzhledem k Y. Regresní funkce vyjadřuje změny podmíněné střední hodnoty jedné náhodné veličiny při změně hodnot druhé náhodné veličiny. Graf regresní funkce nazýváme regresní křivka*.“ (Kubanová, str. 108)

Definice Stochasticky závislé veličiny: „*Nechť X, Y jsou dvě náhodné veličiny. Jestliže změna hodnoty jedné náhodné veličiny vyvolá změnu rozdělení pravděpodobností druhé náhodné veličiny, říkáme, že náhodné veličiny X, Y jsou stochasticky závislé*.“ (Kubanová, str. 108)

Stochastické závislosti se projevují ve změnách střední hodnoty jedné náhodné veličiny souvisejících se změnami hodnot druhé náhodné veličiny, to znamená, že se projevují prostřednictvím podmíněných středních hodnot. [6]

Hlavním úkolem regresní analýzy je zjištění tvaru stochastické závislosti a parametrů regresní funkce. Regresní analýza se zabývá závislostí náhodné veličiny Y na nezávislé proměnné x, která není náhodná a je obecně m-rozměrná. Náhodná veličina Y má pro danou hodnotu $x = (x_1, x_2, ..., x_m)$ a parametry $(\beta_0, \beta_1, \beta_2, ..., \beta_k)$ podmíněnou střední hodnotu $E(Y/x) = g(x, \beta_0, \beta_1, \beta_2, ..., \beta_k)$. Funkce g proměnné x se nazývá regresní funkce a parametry $(\beta_0, \beta_1, \beta_2, ..., \beta_k)$ se nazývají regresní koeficienty. [6]

Modely lineární vzhledem k parametrům mají regresní funkci tvaru

$$g(x, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \dots, \boldsymbol{\beta_k}) = \sum_{i=0}^{k} \boldsymbol{\beta_i} \cdot g_i(x) \tag{1}$$

kde g_i jsou funkce nezávisle proměnných $x = (x_1, x_2, \dots, x_m)$. [6]
$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ jsou parametry

1.3.1 Lineární regrese a metoda nejmenších čtverců

Při budování regresních modelů se běžně užívá metody nejmenších čtverců. Tato metoda poskytuje postačující odhady parametrů jenom při splnění předpokladů o datech a o regresním modelu. Pokud tyto předpoklady:

- regresní parametry β mohou nabývat libovolných hodnot;
- regresní model je lineární v parametrech a platí aditivní model měření;
- matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných X má hodnost rovnou právě m;
- náhodné chyby ε_i mají nulovou střední hodnotu $E(\varepsilon_i)=0$. To musí u korelačních modelů platit vždy. U regresních modelů se může stát, že $E(\varepsilon_i)= K, i = 1, \dots, n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení bude $E(\varepsilon_i')=0$, kde $\varepsilon_i' = y_i - \hat{y}_{P,i} - K$;
- náhodné chyby ε_i mají konstantní a konečný rozptyl $E(\varepsilon_i^2)= \delta^2$. Také podmíněný rozptyl $D(y/x) = \delta^2$ je konstantní a jde o homoskedastický případ;
- náhodné chyby jsou ε_i vzájemně nekorelované a platí $cov(\varepsilon_i\varepsilon_j)= E(\varepsilon_i\varepsilon_j)=0$;
  chyby ε_i mají normální rozdělení $N(0, \delta^2)$;

nejsou splněny, ztrácí výsledky metodou nejmenších čtverců své vlastnosti.

Lineární regresní analýza se používá v těchto případech [4]:

a) Popis empirických dat – hledá se vztah, lineární regresní model, který sumarizuje vazby mezi sloupci v datech.
b) Určení parametrů – běžným cílem regresní analýzy je vyčíslení odhadů neznámých parametrů regresního modelu. Uživatel navrhne regresní model a regresní analýzou se snaží model prokázat. Často tento cíl překrývá i ostatní záměry regresní analýzy.
c) Predikce – cílem regresní analýzy je často predikce, tj. vyčíslení hodnot závisle proměnných pro zadané kombinace vstupních parametrů. Predikce jsou důležité při plánovaní, monitorování a vyhodnocování procesů.

d) Řízení – lze využít také k monitoringu a řízení systémů.
e) Výběru důležitých proměnných – výběr proměnných se provádí s ohledem na nezávisle proměnné, které vysvětlují významný podíl proměnlivosti na závisle proměnné.

Jednoduchým lineárním modelem lineární regrese nazýváme takový lineární model, kdy grafem regresní funkce je přímka. Předpokládejme, že $Y_1, Y_2, ..., Y_n$ je n-tice náhodných veličin s vlastnostmi $EY_i = \alpha + \beta x_i$, $DY_i = \sigma^2, i = 1,2, ..., n$, kde α, β, σ^2 jsou neznámé parametry a $x_1, x_2, ..., x_n$ je n-tice známých hodnot.

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

kde: ε_i jsou nezávislé náhodné veličiny, pro které platí $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$, $i = 1,2, ..., n$.

Náhodná složka zahrnuje působení náhodných vlivů nebo působení veličin, které nejsou zahrnuty v modelu. [6]

Regresní přímka $y = \alpha + \beta x$ se nazývá regresní přímka, β je její směrnice. Úkolem je nyní na základě naměřených dvojic hodnot $(x_1,y_1), (x_2,y_2), ..., (x_n,y_n)$ odhadnout neznámé parametry α, β, σ^2 daného modelu. Tyto odhady budeme značit po řadě a, b, s^2. [6]

Bodové odhady a, b parametrů α, β získáme metodou nejmenších čtverců. Princip této metody spočívá v tom, že hledáme takovou funkci $\hat{y} = a + bx$, aby v jistém smyslu co nejvíce přiléhala k bodům $(x_1, y_1), (x_2,y_2), ..., (x_n y_n)$, kde přiléhání měříme součtem rozdílu hodnot $\hat{y}_i$ a y_i. [4], [6]

Výpočet parametrů a, b:

$$a = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \tag{2}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad \text{[4], [6]} \tag{3}$$

kde: x_i, y_i jsou tabulkové hodnoty nezávislé a závislé proměnné

n značí počet dvojic proměnných, kterých musí být víc než parametrů

1.3.2 Logistická regrese

Je používaná především v případech, kdy vysvětlovaná proměnná není spojitá. Metoda logistické regrese byla původně vyvinuta pro případ, kdy je závislá proměnná binární (dichotomická, alternativní), to znamená, že může nabývat jen dvou hodnot. V takovém případě, již nelze použít k odhadu parametrů „klasickou“ regresní analýzu s odhadem regresních koeficientů prostřednictvím metody nejmenších čtverců. K odhadu těchto parametrů β_0, β_1 používáme metodu maximální věrohodnosti. Obvykle předpokládáme, že je závisle proměnná náhodnou veličinou s normálním rozdělením. Pro odvození modelu pak používáme metodu nejmenších čtverců. [4], [6]

Hledané koeficienty $\beta i, i = 0, 1, ..., p$ odhadneme pomocí metody maximální věrohodnosti, protože použití „klasické“ regresní analýzy s odhadem regresních koeficientů pomocí metody nejmenších čtverců by v tomto případě mohlo způsobit jisté problémy [4], [6]:

- Heteroskedasticita: u binárních proměnných není dodržen předpoklad homogenity rozptylu pro všechny rezidua. Důsledkem by byly špatné výsledky standardních odchylek a testování hypotéz by bylo nekorektní.
- Nenormální rozdělení reziduí: dalším předpokladem metody nejmenších čtverců je normální rozdělení reziduí. Který nemůže být splněn, když rezidua v případě binární proměnné jsou získána pouze ze dvou hodnot, protože by se opět došlo ke špatnému odhadu standardní chyby.
- Nelinearita: odhad pravděpodobnosti úspěchu ($Y = 1$) se pohybuje v intervalu od 0 do 1. V metodě nejmenších čtverců nejsou ale žádná omezení, která by nám jako odhad dala také číslo z tohoto intervalu, a můžeme proto mít za výsledek jakékoliv číslo z intervalu $(-\infty; \infty)$, který by nám říkal, že pravděpodobnost je záporná nebo větší než jedna, což je nesmyslné.

Označme si pravděpodobnosti: $P(Y_i = 1) = \pi_i, P(Y_i = 0) = 1 - \pi_i$; kde Y je binární proměnnou. Je zřejmé, že $Yi \sim Bi(1;\pi_i)$. V případě nezávislosti jednotlivých pozorování můžeme věrohodnost zapsat jako součin pravděpodobností:

$$L(\beta|y) = \prod_{i=1}^{n} \pi(x_i)^{y_1} \, [1 - \pi(x_i)]^{1-y_i} \tag{4}$$

kde: L je věrohodností funkce, $\pi(x_i)$ představuje podmíněnou pravděpodobnost a y_i hodnota pozorování

Kde pravděpodobnost $\pi(x_i) = \frac{e^{g(x_i)}}{1+ e^{g(x_i)}}$, kde $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ a i je vektor neznámých koeficientů.

Pro lepší výpočetní vlastnosti se výraz zlogaritmuje a po úpravě dostáváme:

$$\ln L(\beta|y) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \quad (5)$$

Kde L je věrohodností funkce, $\pi(x_i)$představuje podmíněnou pravděpodobnost a y_i hodnota pozorování

Pro nalezení věrohodností funkce je nutné řešit rovnici $\frac{\partial lnL}{\partial \beta} = 0$. Z této soustavy se poté vypočtou odhady neznámých parametrů $\beta_0, \beta_1, \ldots, \beta_p$. [4], [6]

2 PRAKTICKÁ APLIKACE DM

DM bude proveden metodikou CRISP-DM. Ke zpracování využiji software Microsoft Excel a SPSS Clementine.

2.1 Aplikování lineární regrese na vhodně vybraných datech

Aplikace lineární regrese proběhne na třech datových souborech, přičemž každý má původ v odlišné oblasti výzkumu.

Porozumění problému

Tato fáze metodiky CRISP-DM se zaměřuje na pochopení cílů úlohy a požadavků na řešení formulovaných z manažerského hlediska. Manažerská formulace musí být následně převedena do zadání úlohy pro DM. [8]

Data pro první úlohu budou pocházet z ekonomického prostředí. Datový soubor bude shromažďovat informace o vlastnostech jednotlivých typů automobilů. Cílem práce je najít závislost mezi spotřebou automobilu a vhodně zvolenými proměnnými a na základě výsledků pak rozhodnout, zda lze predikovat spotřebu automobilu.

Druhý soubor dat, kterým se v práci budu zabývat, pochází z více statistických výzkumů a bude se věnovat problematice ekonomicko-sociální. Datový soubor bude obsahovat statistická data, která budou roztříděna dle krajů České republiky. Cílem práce s daty je hledat závislost mezi průměrnou hrubou mzdou v kraji a vhodně zvolenými proměnnými s následným rozhodnutím, zda lze průměrná hrubá mzda predikovat.

Poslední datový soubor se bude skládat rovněž z více statistických výzkumů a bude se zabývat ekologicky-ekonomickou problematikou. Databáze bude obsahovat statistická data, která budou roztříděna dle krajů České republiky. Cílem bude zjistit, zda lze z dat vyvodit závislost mezi investovanými peněžními prostředky na ochranu životního prostředí a zvolenými proměnnými. Z výsledků provedených analýz bude učiněno rozhodnutí, zda tato závislost skutečně existuje a jestli je možné predikovat pro jednotlivé kraje množství investovaných peněz na ochranu životního prostředí.

Porozumění datům

Ve fázi porozumění datům je přiblížen původ dat, deskriptivní charakteristika dat a kvalita dat, tak abychom získali ucelený přehled o surových datech, s kterými bude dále pracováno.

Popis jednotlivých proměnných lze vyčíst v datových slovnících (Příloha 2, Příloha 3, Příloha 4). Bližší charakter dat ukazují Obrázky 2, 5 a 7.

Vstupní datové tabulky jsou na datovém mediu CD-R v Příloze 1 v adresáři \\datové_soubory. Zde lze vidět celý obsah, rozsah a strukturu vstupních datových tabulek „auta.csv“, „prumerna_mzda.csv“ a „investice.csv“.

Porozumění datovému souboru „auta.csv“

Získání těchto dat nebylo jednoduché. Všechny vstupní datové matice byly zpracovány samostatně za pomoci vhodného zdroje a softwaru Microsoft Excel. Data pro soubor „auta.csv“ byla získána z webového portálu www.katalog.auto.cz. Ze zvolených padesáti profilů aut byla data postupně přenášena do prostředí Microsoft Excel, kde byla následně exportována do tabulky ve formátu *.csv, která byla pojmenována „auta.csv“.

	nazev	znacka	karoserie	spotreba	palivo	zdvihovy_objem	vykon	typ_motoru	tocivy_moment	zrychleni	max_rychlost	emise	dvere	mista	hmotnost
1	Škoda Octavia 1.2	Škoda	liftback	5.900	benzin	1197	77	TSI	175	10.800	192	136	5	5	1890
2	Škoda Octavia 2.0	Škoda	liftback	4.800	nafta	1968	103	TDI	320	9.500	211	126	5	5	1995
3	Škoda Octavia RS	Škoda	kombi	7.500	benzin	1984	147	TSI	280	7.300	239	175	5	5	1930
4	Škoda Octavia 1.6	Škoda	liftback	7.100	benzin	1595	75	MPI	148	12.300	190	166	5	5	1880
5	Škoda Roomster 1.2	Škoda	MPV	4.500	nafta	1199	55	TDI	180	15.500	162	119	5	5	1763
6	Škoda Octavia Scout 2.0	Škoda	kombi	5.900	nafta	1968	103	TDI	320	10.100	199	155	5	5	2155
7	Škoda Rapid 1.2	Škoda	sedan	5.400	benzin	1197	77	TSI	175	10.300	195	125	4	5	1635
8	Škoda Rapid 1.6	Škoda	sedan	4.000	nafta	1598	66	TDI	230	11.700	206	114	4	5	1595
9	škoda Yeti 1.4	Škoda	SUV	6.800	benzin	1390	90	TSI	200	10.500	185	159	5	5	1920
10	Škoda Superb 1.8	Škoda	sedan	7.200	benzin	1798	118	TSI	250	8.600	220	169	5	5	2074
11	Škoda Superb 2.0	Škoda	sedan	5.700	nafta	1968	125	TDI	350	8.800	222	149	5	5	2118
12	Škoda Superb GL	Škoda	kombi	4.400	nafta	1598	77	TDI	250	12.600	190	114	5	5	2109
13	Volkswagen Golf Plus ...	Volks...	MPV	5.800	benzin	1197	62	TSI	160	13.400	175	133	5	5	1930
14	Volkswagen Golf Varia...	Volks...	kombi	4.900	nafta	1968	103	TDI	320	9.700	210	128	5	5	2030

Obrázek 2: Náhled obsahu souboru „auta.csv“

Zdroj: [vlastní]

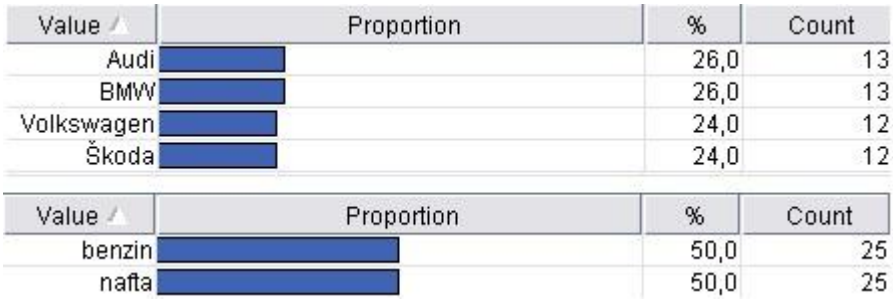
K lepšímu porozumění datům a k možnému nalezení nesrovnalostí v datovém souboru se provede výpočet popisných statistik pomocí uzlu *Data audit*, který je znázorněn na Obrázku 3.

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
karoserie		Set	--	--	--	--	--	7	50
spotreba		Range	3.800	13.900	6.272	1.875	1.869	$null$	50
palivo		Flag	--	--	--	--	--	2	50
zdvihovy_objem		Range	1197	4395	2124.440	756.533	1.272	$null$	50
vykon		Range	55	412	136.960	73.323	2.216	$null$	50

Obrázek 3: Náhled popisných statistik pro soubor „auta.csv“

Zdroj: [vlastní]

Výsledek pozorování popisných statistik neukázal žádnou zásadní chybu. Nebyla zjištěna přítomnost chybných ani odlehlých hodnot. Sloupec *Valid* nabývá pro všechny proměnné stejné hodnoty, takže nebyla zjištěna přítomnost chybějících hodnot. Dále je možné si povšimnout, že tabulka obsahuje dva typy dat a to typy Range a Set. Obrázek 3 slouží spíše pro lepší představu o datech a struktuře dat. Dále je z Obrázku 4 patrné rovnoměrné zastoupení aut čtyř značek a poměr aut spalujících benzín nebo naftu.



Value	Proportion	%	Count
Audi		26,0	13
BMW		26,0	13
Volkswagen		24,0	12
Škoda		24,0	12

Value	Proportion	%	Count
benzin		50,0	25
nafta		50,0	25

Obrázek 4: Procentuální zastoupení proměnné Značka a Palivo

Zdroj: [vlastní]

Porozumění datovému souboru „prumerna_mzda.csv“

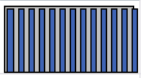
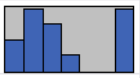
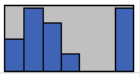
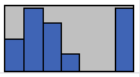
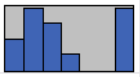
Data pro druhý soubor „prumerna_mzda.csv“ pochází z webového portálu Českého statistického úřadu na adrese www.czso.cz. Datová matice byla vytvořena na základě souhrnných informací ČSÚ o 14 krajích České republiky za rok 2011. I zde ke zpracování dat byl využit software Microsoft Excel a exportováním byl získán výsledný soubor ve formátu *.csv.

	kraj	počet obyvatel	podíl_zen	pocet_obyv_15-64	podíl_zen_15-64	hustota	mest_obyv	pocet_obci	mesta	HDP	prumerna_hruba_mzda	mira_nezamest	registr_subjekty	nemocen_pojis
1	hlavní město Praha	1241664	634173	863517	431699	2502...	100.000	1	1	786057	33546	3.950	529377	993771
2	Středočeský	1273094	644325	880832	434088	116.100	53.000	1145	82	322868	25651	7.700	317598	399430
3	Jihočeský	636138	322720	439059	216625	63.300	64.200	623	54	309006	23199	7.530	158543	230186
4	Plzeňský	571497	288718	394320	193482	75.600	67.400	501	56	325753	24036	7.100	147419	215666
5	Karlovarský	303165	153733	212394	104820	91.500	82.700	132	37	259180	21723	9.830	83396	90733
6	Ústecký	828595	419616	577193	283860	155.200	79.700	354	58	292658	23174	12.940	178718	258251
7	Liberecký	438132	223432	303874	150815	138.600	77.800	215	39	279039	23422	9.460	118766	138344
8	Královehradecký	554050	281774	377554	186329	116.400	67.300	448	48	315316	22837	7.490	134689	196960
9	Pardubický	516260	261378	354670	174094	114.300	62.300	451	38	296796	22978	8.440	114072	186915
10	Vysočina	511972	258028	351776	171393	75.300	57.300	704	34	300309	22918	9.440	105185	172691
11	Jihomoravský	1164633	594770	803165	396916	162.100	62.300	673	49	340093	24651	9.810	291162	475983
12	Olomoucký	638848	326736	440747	218395	121.300	56.600	399	30	279902	22825	11.370	138970	202946
13	Zlínský	589596	301335	406257	200394	148.600	59.800	305	30	309386	22655	9.350	136725	213157
14	Moravskoslezský	1232626	629722	857430	424036	227.000	75.500	300	42	318155	24174	11.180	248824	436517

Obrázek 5: Náhled obsahu souboru „prumerna_mzda.csv“

Zdroj: [vlastní]

K lepšímu porozumění datům a k možnému nalezení nesrovnalostí v datovém souboru se provede výpočet popisných statistik pomocí uzlu *Data audit*, který je znázorněn na Obrázku 6.

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
kraj		Set	--	--	--	--	--	14	14
počet obyvatel		Range	303165	1273094	750019.286	334653.638	0.671	$null$	14
podíl_zen		Range	153733	644325	381461.429	170844.960	0.672	$null$	14
pocet_obyv_15-64		Range	212394	880832	518770.571	232762.088	0.680	$null$	14
podíl_zen_15-64		Range	104820	434088	256210.429	115711.766	0.688	$null$	14

Obrázek 6: Ukázka popisné statistiky souboru „prumerna_mzda.csv“

Zdroj: [vlastní]

Výsledek pozorování popisných statistik ukázal jednu zásadní chybu, která by negativně ovlivnila výsledky. Tato chyba spočívá v odlehlosti dat pro kraj Hlavní město Praha. Tento kraj je specifický tím, že ho tvoří jen hlavní město Praha, takže hodnoty pro něj naměřené se diametrálně liší od průměru v České republice. Ve zbývajících datech nebyla zjištěna přítomnost chybných ani odlehlých hodnot. Sloupec *Valid* nabývá pro všechny proměnné stejné hodnoty, takže nebyla zjištěna přítomnost chybějících hodnot.

Porozumění datovému souboru „investice.csv“

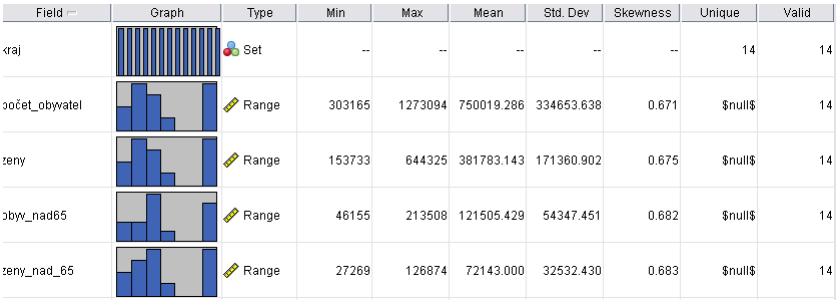
Datový soubor „investice.csv“ a druhý soubor dat „prumerna_mzda.csv“ sdílí společně několik proměnných. Data byla pořízena na webovém portálu ČSÚ na adrese www.czso.cz a rovněž roztříděna dle 14 krajů České republiky. Exportováním ze softwaru Microsoft Excel dostáváme soubor „investice.csv“.

	kraj	počet_obyvatel	zeny	obyv_nad65	zeny_nad_65	index_SO2	index_Nox	index_CO	investice	zemreli	smrtelne_nehody	obyv_na_lekare	obyv_lehatko	prumerny_vek	hustota
1	hlavní město Praha	1241664	638677	213508	126874	3.140	13.940	30.820	4235	9.800	39	133	130	41.900	2503
2	Středočeský	1273094	644325	195120	114610	2.100	3.200	4.900	2016	9.900	97	319	219	40.300	116
3	Jihočeský	636138	322720	103144	60334	1.000	1.200	2.200	1193	10.000	67	244	187	41.200	63
4	Plzeňský	571497	288718	95476	55422	1.000	1.600	2.600	1204	10.300	45	218	161	41.500	76
5	Karlovarský	303165	153733	46155	27269	2.900	3.200	2.400	745	10.100	21	238	200	40.900	91
6	Ústecký	828595	419616	122483	72912	10.900	10.800	4.600	1904	10.700	54	288	162	40.400	155
7	Liberecký	438132	223432	67587	40027	0.800	1.300	3.500	1500	9.700	26	266	168	40.600	139
8	Královehradecký	554050	281774	94861	55952	1.300	1.700	3.500	1465	10.400	57	223	160	41.500	116
9	Pardubický	516260	261378	84711	49998	2.900	3.600	3.400	1150	10.400	48	257	200	41.000	114
10	Vysočina	511972	258028	84830	49970	0.400	1.800	3.100	988	9.800	33	277	192	41.100	75
11	Jihomoravský	1164633	594770	195117	116769	0.500	2.400	3.600	1712	9.800	67	205	156	41.300	162
12	Olomoucký	638848	326736	104919	62773	0.800	2.000	3.100	902	10.300	45	221	194	41.200	121
13	Zlínský	589596	301335	98870	59968	1.200	1.900	2.700	1001	10.500	40	259	193	41.400	149
14	Moravskoslezský	1232626	629722	194295	117124	4.100	5.100	25.400	4793	10.900	70	250	196	40.900	227

Obrázek 7: Náhled obsahu souboru „ investice.csv“

Zdroj: [vlastní]

K lepšímu porozumění datům a k možnému nalezení nesrovnalostí se provede výpočet popisných statistik pomocí uzlu *Data audit*, který je znázorněn na Obrázku 8.



Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
kraj		Set	--	--	--	--	--	14	14
počet_obyvatel		Range	303165	1273094	750019.286	334653.638	0.671	$null$	14
ženy		Range	153733	644325	381783.143	171360.902	0.675	$null$	14
obyv_nad65		Range	46155	213508	121505.429	54347.451	0.682	$null$	14
ženy_nad_65		Range	27269	126874	72143.000	32532.430	0.683	$null$	14

Obrázek 8: Ukázka popisné statistiky souboru „investice.csv“

Zdroj: [vlastní]

Díky podobnosti s datovým souborem „prumerna_mzda.csv“ je i v tomto datovém souboru jediným nedostatkem odlehlost hodnot pro kraj Hlavní město Praha. Ve zbývajících datech nebyla zjištěna přítomnost chybných ani odlehlých hodnot.

Příprava dat

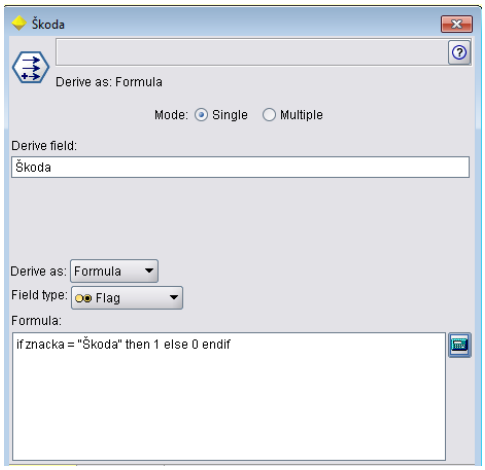
Z hlediska metodiky CRISP-DM by tato fáze měla obsahovat selekci, čištění, transformaci, vytváření, integrování a formátování dat. Ze zmíněných kroků se pozornost zaměří zejména na čistotu dat a požadavky ohledně formátování, které budou spočívat především v transformaci proměnných na dummy proměnné, aby mohla být využita při modelování lineární regrese.

Úprava vstupních dat souboru „auta.csv“

Po provedení všech úprav ve formátování a transformaci proměnných na dummy proměnné, bude výsledná tabulka vhodná pro modelování lineární regrese. Provedené úpravy:

- Načtení souboru „auta.csv“ pomocí uzlu *Var.File* a definováním oddělovacího znaku na středník.
- Dále u zvolených proměnných budou odstraněny proměnné typu Set přetvořením na dummy proměnné. Pomocí uzlu *Derive* vytvoříme nová příznačně pojmenovaná pole. Postup vytvoření rozkladu proměnné typu Set na jednotlivé dummy proměnné typu Flag bude objasněno na jednom příkladu, a to na proměnné „znacka“. Na Obrázku 9 je znázorněno nastavení uzlu *Derive*. Stejný postup aplikujeme i na zbylé vybrané proměnné typu Set.
    - Uzel *Derive*, název: *Audi*, příkaz: *if znacka = "Audi" then 1 else 0 endif*, typ: *flag*
    - Uzel *Derive*, název: *Škoda*, příkaz: *if znacka = "Škoda" then 1 else 0 endif*, typ: *flag*

- Uzel *Derive*, název: *BMW*, příkaz: *if znacka = "BMW" then 1 else 0 endif*, typ: *flag*
- Uzel *Derive*, název: *Volkswagen*, příkaz: *if znacka = "Volkswagen" then 1 else 0 endif*, typ: *flag*



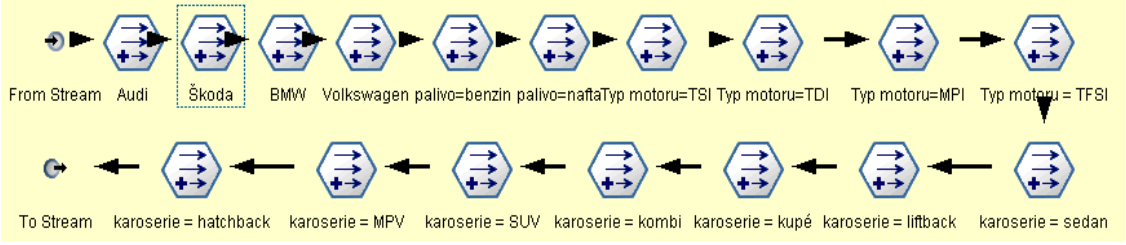
Obrázek 9: Náhled obsahu uzlu *Derive* pro vytvoření dummy proměnných

Zdroj: [vlastní]

- Na Obrázku 10 je znázorněno použití uzlu *SuperNode*, který slouží pro lepší přehlednost a orientaci v prostředí SPSS Clementine. Do tohoto uzlu bylo vloženo 17 dummy proměnných, které můžeme vidět na Obrázku 11. Na Obrázku 12 jsou tyto proměnné prezentovány v tabulkové formě.

Obrázek 10: Náhled na uzel *SuperNode* v SPSS Clementine

zdroj:[vlastní]



Obrázek 11: Náhled do uzlu SuperNode, který obsahuje vytvořené dummy proměnné

Zdroj: [vlastní]

Audi	Škoda	BMW	Volkswagen	palivo=benzin	palivo=nafta	Typ motoru=TSI	Typ motoru=TDI	Typ motoru=MPI	Typ motoru = TFSI	karoserie = sedan
0	1	0	0	1	0	1	0	0	0	0
0	1	0	0	0	1	0	1	0	0	0
0	1	0	0	1	0	1	0	0	0	0
0	1	0	0	1	0	0	0	1	0	0

Obrázek 12: Náhled na tabulku po vytvoření dummy proměnných

Zdroj: [vlastní]

Úprava vstupních dat souboru „prumerna_mzda.csv“

V datech se nevyskytuje žádná proměnná, která by potřebovala převést na dummy proměnnou, takže jedinou úpravou, kterou je na datech třeba provést, je odstranění odlehlého záznamu z tabulky a vytvoření nové proměnné. Provedené úpravy:

- V části porozumění datům bylo pojednáno o odlehlosti záznamu pro kraj Praha. Aby bylo možné s daty dále pracovat a dosáhnout co nejpřesnějších výsledků, bylo rozhodnuto o vyřazení tohoto záznamu. Vyřazení záznamu bylo provedeno uzlem *Select.*
- Načtení souboru „prumerna_mzda.csv“ pomocí uzlu *Var.File* a definováním oddělovacího znaku na středník.
- Pomocí uzlu *Derive* vytvořím novou proměnnou „*% obyv s VS*“, která bude ukazovat procentuální zastoupení vysokoškolsky vzdělaných obyvatel v kraji.
    - Uzel *Derive, název: % obyv s VS,* příkaz: *round(obyv s VS/('počet obyvatel'/100)),* typ: *Range*

Úprava vstupních dat souboru „investice.csv“

V datech se nevyskytuje žádná proměnná, která by potřebovala převést na dummy proměnnou, takže jedinou úpravou, kterou na datech provedeme je odstranění odlehlého záznamu z tabulky. Provedené úpravy:

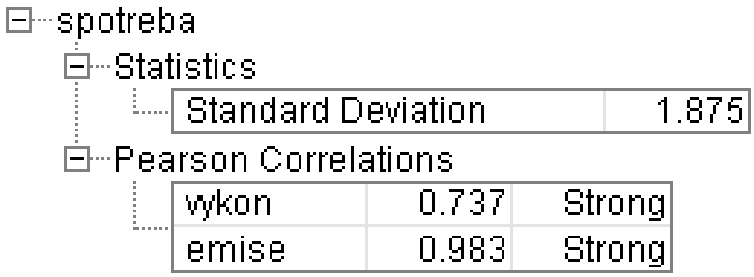
- V části porozumění datům bylo pojednáno o odlehlosti záznamu pro kraj Hlavní město Praha. Aby bylo možné dál s daty pracovat a dosáhnout co nejpřesnějších výsledků, bylo rozhodnuto o vyřazení tohoto záznamu. Vyřazení záznamu bylo provedeno uzlem *Select.*
- Načtení souboru „prumerny_vek.csv“ pomocí uzlu *Var.File* a definováním oddělovacího znaku na středník.
- Při modelaci lineární regrese pro nezávislou proměnnou *Index_CO* byla zjištěna velká odlehlost hodnoty pro Moravskoslezský kraj. Pro zajištění maximální přesnosti modelu pro proměnnou *Index_CO* byl Moravskoslezský kraj vyřazen pomocí uzlu *Select.*

Modelování dat

Tato část metodiky CRISP-DM je rozdělena do tří částí. V každé části budou postupně provedeny tři jednorozměrné a jedna vícerozměrná lineární regrese. Cílem každé části je analyzovat data, tak aby mezi jednotlivými analýzami byly nalezeny souvislosti, které povedou k zisku dat s informační hodnotou. Pro modelaci využiji softwaru SPSS Clementine.

Analýza vstupních dat souboru „auta.csv“

Pro aplikaci lineární regrese byla proměnná „spotřeba“ automobilu zvolena jako závislá a k ní na základě výšky korelačních koeficientů byly zvoleny nezávislé proměnné. Nezávislé proměnné dokumentuje Obrázek 13.



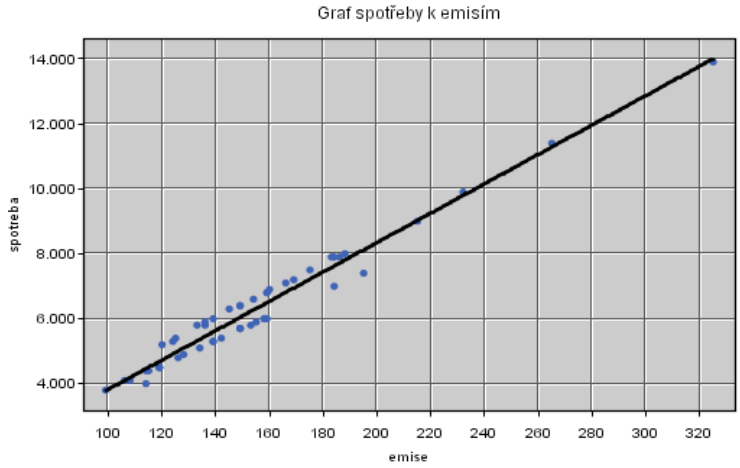
Obrázek 13: Závislé proměnné na spotřebě a jejich korelační koeficienty

Zdroj: [vlastní]

Tabulka 1: Tabulka informací pro modelaci lineární regrese - spotřeba vzhledem k emisím

Vstupní informace			
Závislá proměnná	spotreba	Nezávislá proměnná	emise
Korelační koeficient	0.983		
Doplňující informace			
Vyhodnocení modelu			
Koeficient determinace	0.967	Regresní funkce	y = 0.0452*x - 0.694
Vyhodnocení	Hodnota spolehlivosti je vysoká. Model vyjádřen Obrázkem 14.		

Zdroj: [vlastní]



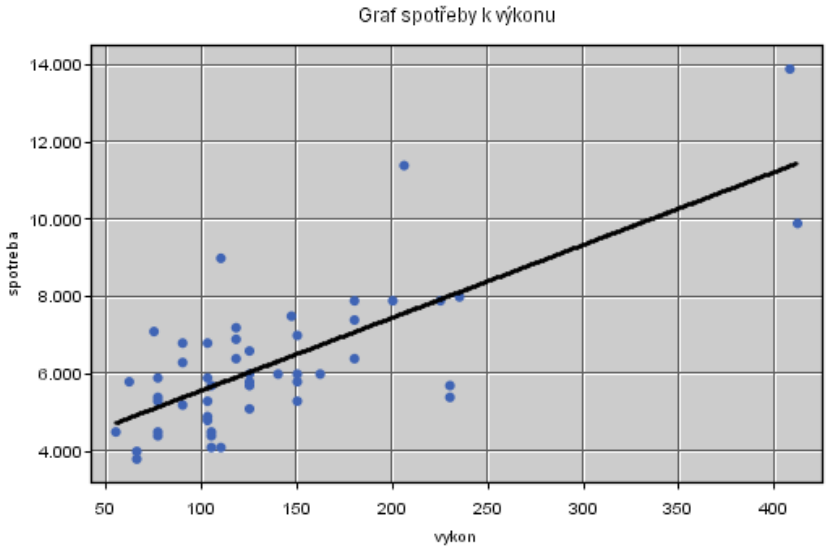
Obrázek 14: Bodový diagram spotřeby a emisí CO_2 proložený regresní funkcí y = 0.0452*x - 0.694

[zdroj vlastní]

Tabulka 2: Tabulka informací pro modelaci lineární regrese - spotřeba vzhledem k výkonu

Vstupní informace			
Závislá proměnná	spotreba	Nezávislá proměnná	vykon
Korelační koeficient	0.737		
Doplňující informace			
Vyhodnocení modelu			
Koeficient determinace	0.543	Regresní funkce	y = 0.01885*x - 3.691
Vyhodnocení	Hodnota spolehlivosti je dostačující. Model je vyjádřen na Obrázku 15.		

Zdroj:[vlastní]



Obrázek 15: Bodový diagram spotřeby a výkonu proložený regresní funkcí y = 0.01885*x + 3.691

Zdroj:[vlastní]

Poslední jednorozměrná lineární regrese je prezentována na proměnné, která byla původně typu Set a v části věnované přípravě dat byla rozpracována do dummy proměnných. Jako závislá proměnná zde byla zvolena spotřeba a nezávislé proměnné jsou Audi, Škoda, BMW a Volkswagen. Hodnota korelační determinace R^2=0.355 shledala vztah mezi proměnnými jako nedostatečný. Na Obrázku 16 je zobrazena predikovaná spotřeba pro jednotlivé typy motorů.

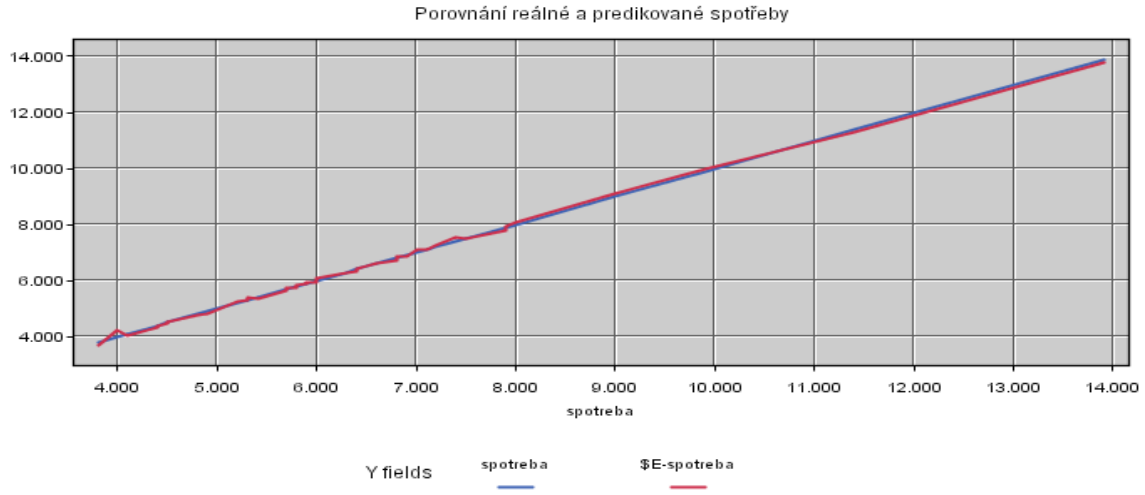
Value		Proportion	%	Count
TDI	5.212		50,0	25
TFSI	6.325		8,0	4
MPI	7.100		2,0	1
TSI	7.545		40,0	20

Obrázek 16: Graf predikovaných spotřeb pro jednotlivé typy motorů po využití lineární regrese

Zdroj:[vlastní]

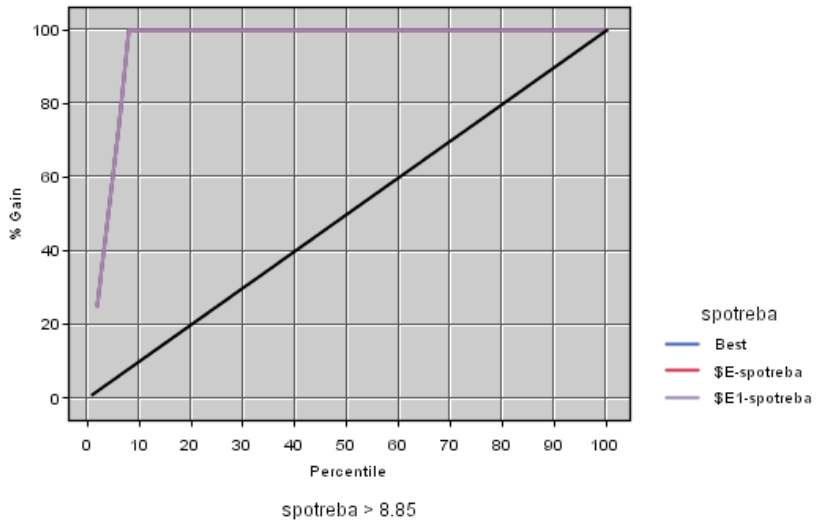
Než se pustíme do vícerozměrné lineární regrese, bylo by vhodné si za pomoci uzlu *Feature Selection* vygenerovat *Filter*, kterým jsou odstraněny nedůležité proměnné. Z původních 25 proměnných máme nyní 12, z kterých se vytvoří model vícerozměrné lineární regrese. Výsledek analýzy vyšel velmi pozitivně. Na základě koeficientu determinace, který nabyl hodnoty 0.999, je možné s minimální odchylkou predikovat spotřebu automobilu. Z regresní funkce je možné vyčíst koeficienty pro jednotlivé proměnné a na Obrázku 17 je porovnána spotřeba garantována výrobcem a predikovaná spotřeba. Na Obrázku 18 nalezneme porovnání výpočtu modelu metodami enter a stepwise. Regresní funkce má tvar:

*y=-0.4047+zrychleni*0.01384+zdvihovy_objem*0.00008406+vykon*0.003113+ tocivy_moment*0.001327+max_rychlost*0.001396+emise*0.04077+hmotnost*0.0002121+BMW*-0.00878+Typ motoru=TSI*0.003126+Typ motoru=TDI*-0.5677+karoserie = SUV*-0.075+karoserie = hatchback*0.09931.*



Obrázek 17: Graf popisující odchylku spotřeby dané výrobcem a predikované spotřeby

Zdroj: [Vlastní]

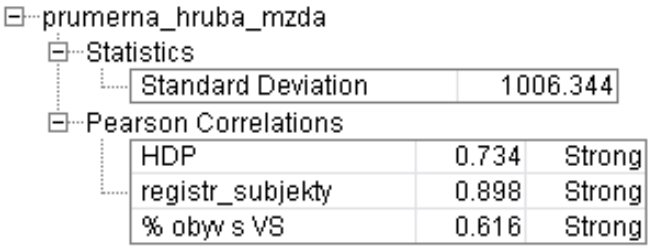


Obrázek 18: Graf typu Evaluation pro srovnání výpočtu metodou enter, stepwise

Zdroj:[vlastní]

Analýza vstupních dat souboru „prumerna_mzda.csv“

Jako vstupní závislá proměnná do modelů lineární regrese byla zvolena průměrná hrubá mzda a k ní zvolené nezávislé proměnné, které jsou společně s jejich korelačními koeficienty uvedeny na Obrázku 19.



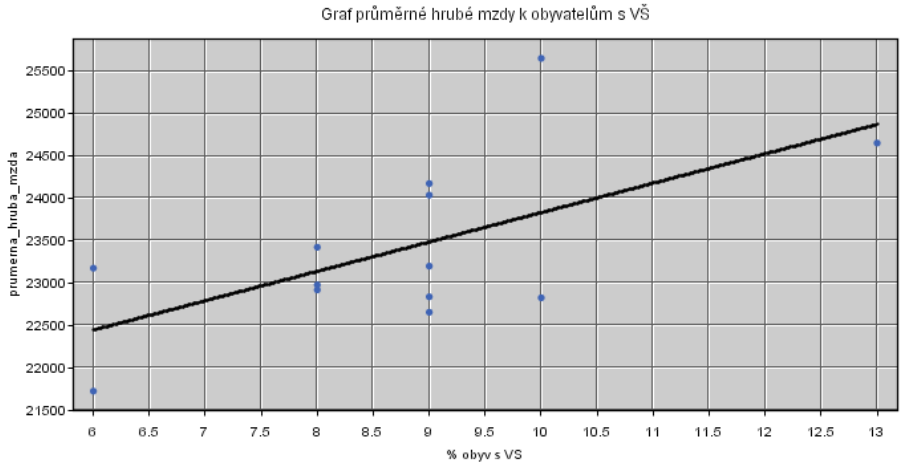
Obrázek 19: Závislé proměnné na mzdě a jejich korelační koeficienty

Zdroj:[vlastní]

Tabulka 3: Tabulka informací pro modelaci lineární regrese - mzda vzhledem k % obyv s VS

Vstupní informace			
Závislá proměnná	prumerna_hruba_mzda	Nezávislá proměnná	% obyv s VS
Korelační koeficient	0.616		
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Koeficient determinace	0.379	Regresní funkce	y = 346.7*x + 20363.1
Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen Obrázkem 20.		

Zdroj: [vlastní]



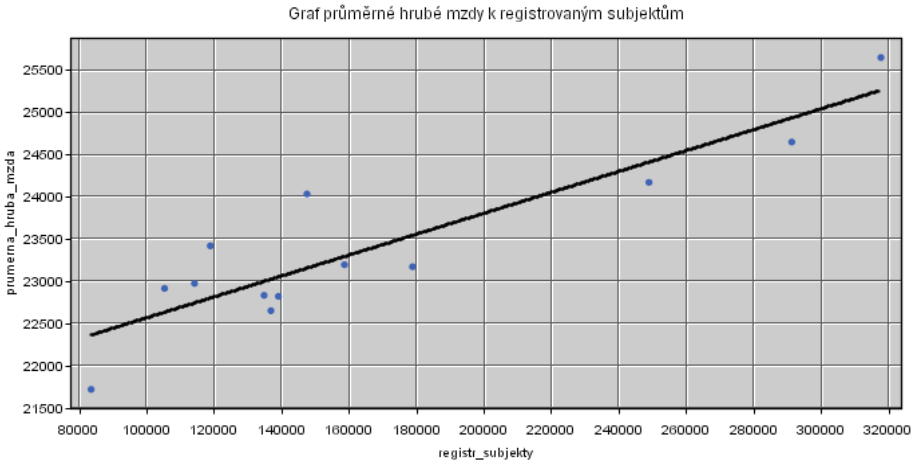
Obrázek 20: Bodový diagram průměrné hrubé mzdy a obyvatel s ukončenou VŠ v %, proložený regresní funkcí y = 346.7*x + 20363.1

Zdroj:[vlastní]

Tabulka 4: Tabulka informací pro modelaci lineární regrese - mzda vzhledem k registr_subjekty

Vstupní informace			
Závislá proměnná	prumerna_hruba_mzda	Nezávislá proměnná	registr_subjekty
Korelační koeficient	0.898		
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Koeficient determinace	0.806	Regresní funkce	y = 0.01237*x + 21335.0
Vyhodnocení	Hodnota spolehlivosti je vysoká. Model vyjádřen Obrázkem 21.		

Zdroj: [vlastní]



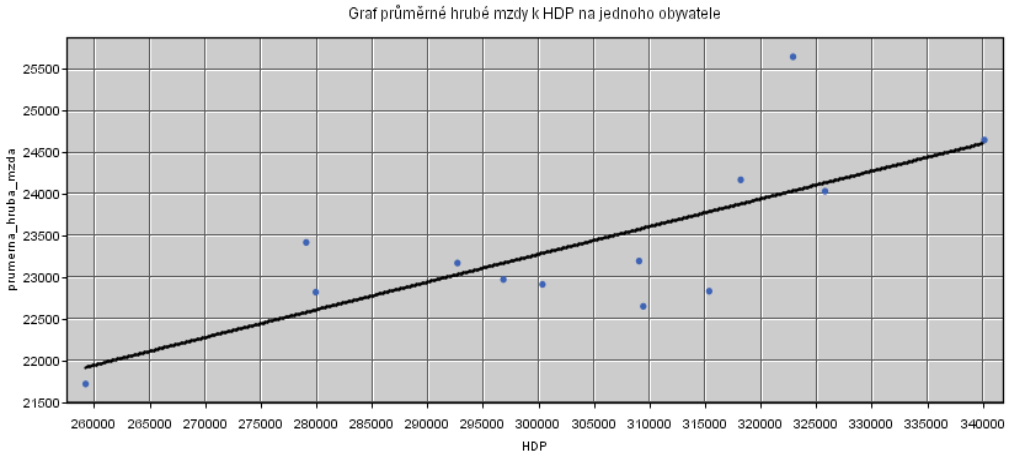
Obrázek 21: Bodový diagram průměrné hrubé mzdy a počtu registrovaných subjektů proložený regresní funkcí y = -0.01237*x + 21335.0

Zdroj: [vlastní]

Tabulka 5: Tabulka informací pro modelaci lineární regrese - mzda vzhledem k HDP

Vstupní informace			
Závislá proměnná	prumerna_hruba_mzda	Nezávislá proměnná	HDP
Korelační koeficient	0.734		
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Koeficient determinace	0.539	Regresní funkce	y = 0.03326*x +13302.1
Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen Obrázkem 22.		

Zdroj: [vlastní]

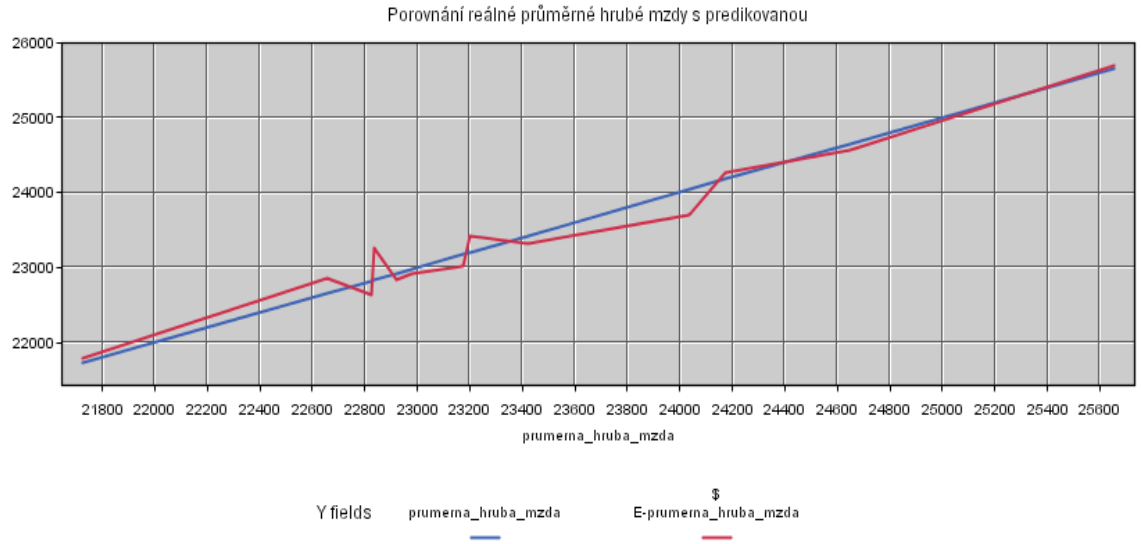


Obrázek 22: Bodový diagram průměrné hrubé mzdy a HDP na 1 obyvatele v Kč proložený regresní funkcí y = 0.03326*x +13302.1

Zdroj:[vlastní]

Pomocí uzlu *Feature Selection* vygenerujeme *Filter*, jehož prostřednictvím se odstraní nedůležité proměnné. Na základě koeficientu determinace, který nabyl hodnoty 0.961, je možné s přijatelnou odchylkou predikovat průměrnou hrubou mzdu v kraji. Z regresní funkce je možné vyčíst koeficienty pro jednotlivé proměnné a na Obrázku 23 je porovnána reálná průměrná hrubá mzda a predikovaná průměrná hrubá mzda. Regresní funkce má tvar:

*y=20263.2+podíl_zen*0.007364+počet_obci*2.284+mesta*150.3+registr.subjekty*0.06344 +nemocen_pojis*-0.01316+obyv_s_VS*-0.0198+zemedel_lesnictvi_rybolov*-0.2243+prumysl_stavebnictvi*0.02639+sluzby*-0.0276+% obyv s VS*-31.67.*

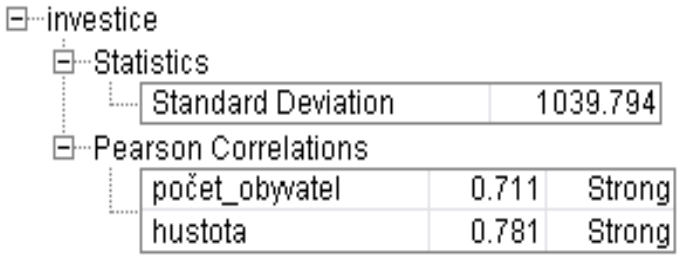


Obrázek 23: Graf popisující odchylku průměrné hrubé mzdy od predikované průměrné hrubé mzdy

Zdroj:[vlastní]

Analýza vstupních dat souboru „investice.csv“

Pro potřebu lineární regrese byla zvolena závislá proměnná *investice* a na základě síly korelace byly zvoleny nezávislé proměnné, které jsou na Obrázku 24. Na obrázku není hodnota pro proměnnou *index_CO*, protože ta se pro zachování objektivnosti výpočtu musela změřit až po odstranění odlehlých hodnot.



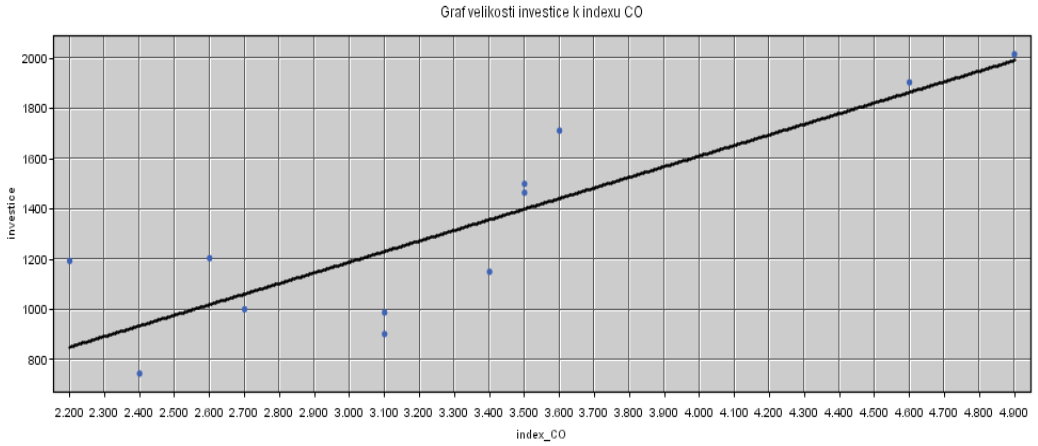
Obrázek 24: Závislé proměnné na investici a jejich korelační koeficienty

Zdroj:[vlastní]

Tabulka 6: Tabulka informací pro modelaci lineární regrese - investice vzhledem k indexu CO

Vstupní informace			
Závislá proměnná	investice	Nezávislá proměnná	index_CO
Korelační koeficient	0.854		
Doplňující informace	Pro zachování objektivity bylo nutné vyřadit odlehlé záznamy pro kraj Hl. město Praha a Moravskoslezský kraj.		
Vyhodnocení modelu			
Koeficient determinace	0.729	Regresní funkce	y = 422.5*x - 79.32
Vyhodnocení	Po odstranění odlehlých hodnot, lze model vyhodnotit jako dostačující. Model je vyjádřen Obrázkem 25.		

Zdroj: [vlastní]



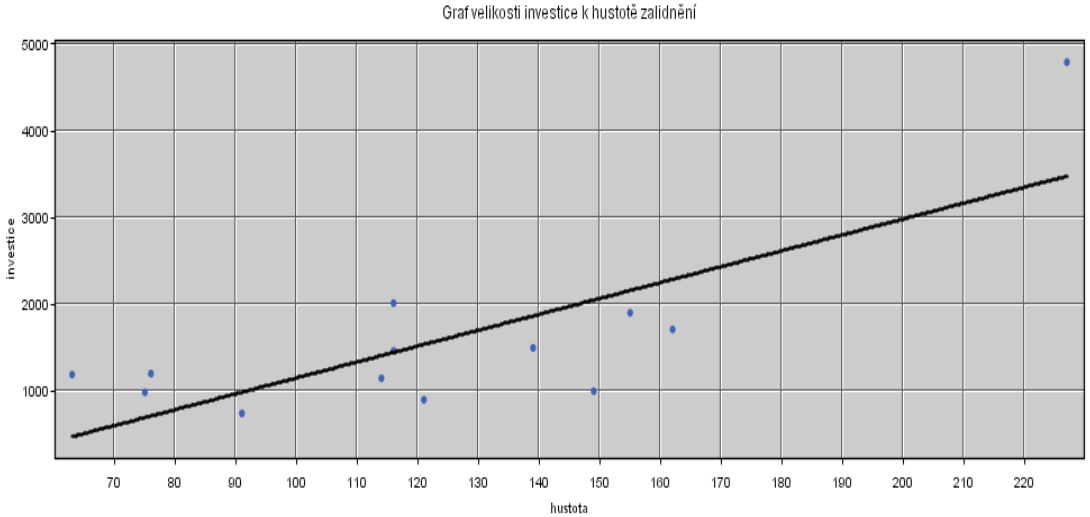
Obrázek 25: Bodový diagram investic na ochranu životního prostředí a indexu CO proložený regresní funkcí y = 422.5*x - 79.32

Zdroj: [vlastní]

Tabulka 7: Tabulka informací pro modelaci lineární regrese - investice vzhledem k hustotě

Vstupní informace			
Závislá proměnná	investice	Nezávislá proměnná	hustota
Korelační koeficient	0.781		
Doplňující informace	Model počítán po odstranění odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Koeficient determinace	0.610	Regresní funkce	y = 18.31*x - 676.7
Vyhodnocení	Hodnota spolehlivosti je dostačující. Model vyjádřen Obrázkem 26.		

Zdroj: [vlastní]



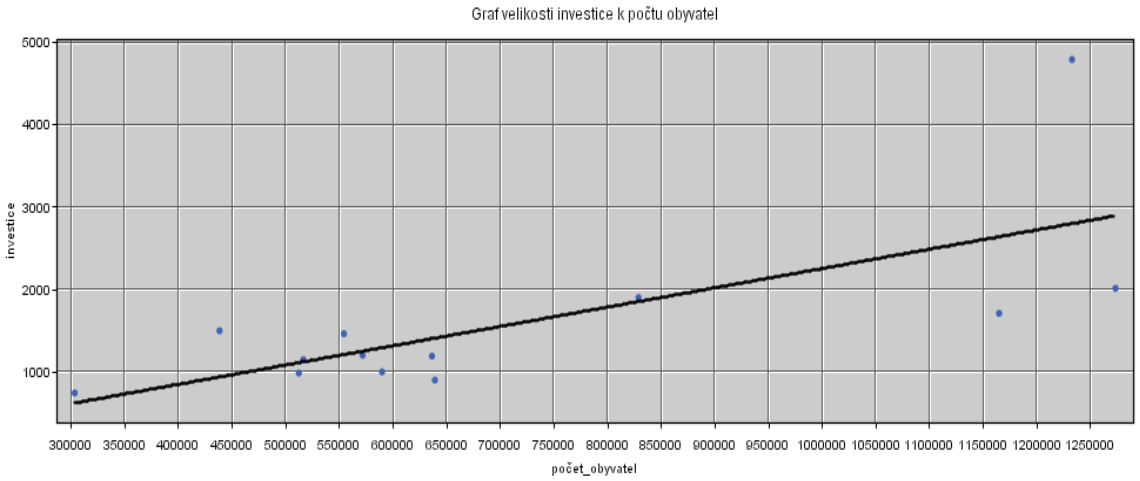
Obrázek 26: Bodový diagram investic na ochranu životního prostředí a hustoty zalidnění proložený regresní funkcí y = 18.31*x - 676.7

Zdroj:[vlastní]

Tabulka 8: Tabulka informací pro modelaci lineární regrese - investice vzhledem k počtu obyvatel

Vstupní informace			
Závislá proměnná	investice	Nezávislá proměnná	počet_obyvatel
Korelační koeficient	0.711		
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Koeficient determinace	0.506	Regresní funkce	y = 0.002343*x - 86.34
Vyhodnocení	Hodnota spolehlivosti je hraniční, ale spíše vyznívá k hodnocení dostačující. Model vyjádřen Obrázkem 27.		

Zdroj:[vlastní]



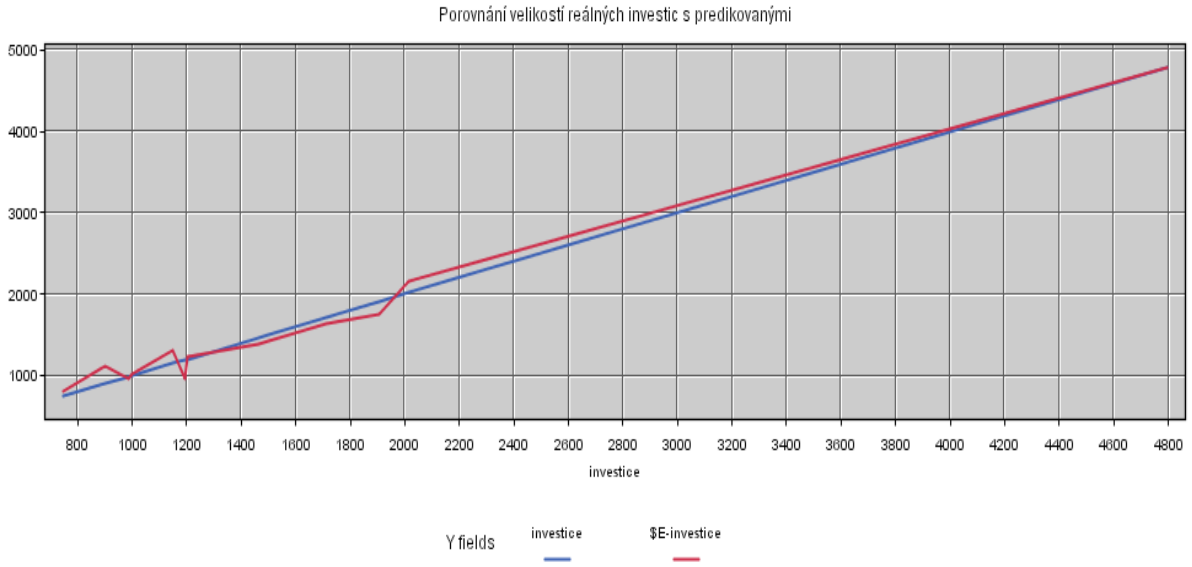
Obrázek 27: Bodový diagram investic na ochranu životního prostředí a počtu obyvatel proložený regresní funkcí y = 0.002343*x - 86.34

Zdroj: [vlastní]

Opět využit uzel *Feature Selection*. Na základě koeficientu determinace, který nabyl hodnoty 0.986, je možné s přijatelnou odchylkou predikovat investované prostředky na ochranu životního prostředí v kraji. Z regresní funkce je možné vyčíst koeficienty pro jednotlivé proměnné. Tyto proměnné pak byly použity pro stejný model, který však byl počítaný metodou backwards. Na Obrázku 29 je potom porovnání zmíněného modelu s modelem vypočítaným metodou enter a na Obrázku 28 jsou porovnány reálné investice s predikovanými. Regresní funkce má tvar:

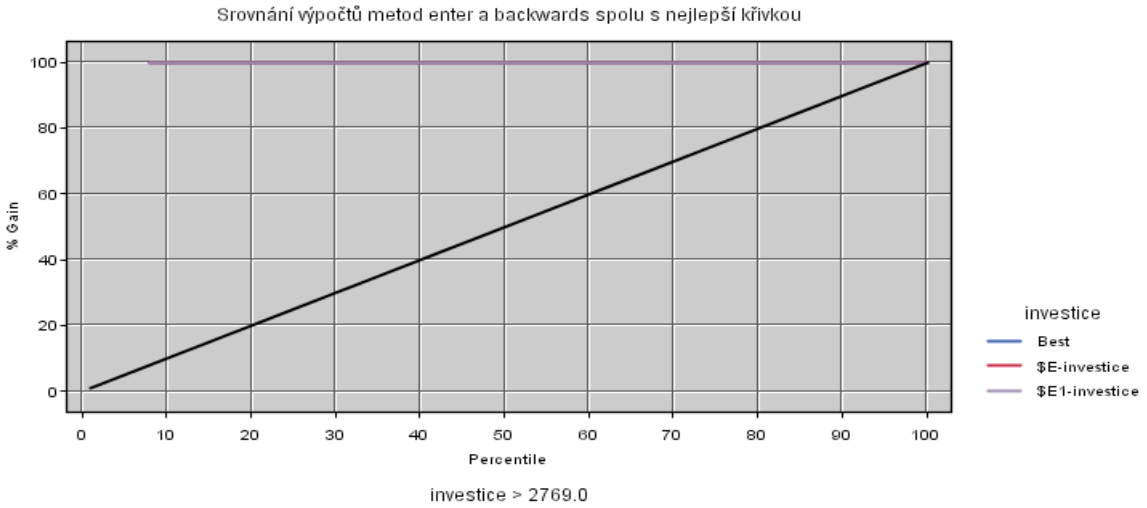
*y = -476.9+počet_obyvatel*0.00004312+obyv_nad65*0.2094+zeny_nad_65*-0.3476+index_CO*132.*

*+hustota*8.74*



Obrázek 28: Porovnání reálných investic na ochranu životního prostředí a predikovaných

Zdroj:[vlastní]



Obrázek 29: Graf Evaluation pro srovnání výpočtů metodou enter, backwards

Zdroj:[vlastní]

Vyhodnocení výsledků

V této fázi se hodnotí přesnost, s jakou model dosahuje očekávaných cílů a snaží se popsat důvody, proč model je hodnocen jako dostatečný či nedostatečný.

Analýza datového souboru „auto.csv“

Definovaným cílem bylo pomocí lineární regrese zjistit, zda se dá predikovat spotřeba automobilu z parametrů obsažených v datovém souboru „auto.csv“. Jako závisle proměnná byla zvolena *spotřeba*, kde na základě síly korelačních koeficientů byla vybrána trojice nezávisle proměnných. Tyto závislé a nezávislé proměnné posloužily jako vstupy do modelů

jednorozměrných lineárních regresí. Z koeficientů determinace se ukázalo, že ze tří zkoumaných proměnných má největší prediktivní úspěšnost proměnná *Emise,* pro kterou vyšel R^2=0.967, středně silná spolehlivost (R^2=0.543) vyšla pro proměnnou *vykon* a pro proměnnou *typ_motoru* bylo naměřeno (R^2=0.355). O výsledku ukazující existenci reálné závislosti, na jejímž základu by se dala predikovat spotřeba, rozhodla vícerozměrná lineární regrese. Korelační determinant pro vícerozměrnou regresi má hodnotu R^2=0.999; to znamená, že model je schopen predikce se spolehlivostí 99,9 %. Při srovnání maxima pro korelační determinant, který dosahuje hodnoty 1, s výsledkem modelu, můžeme konstatovat, že model pro predikci spotřeby vyšel nad očekávání dobře.

Analýza datového souboru „prumerna_mzda.csv“

V druhém modelu bylo cílem definovat pomocí lineární regrese, zda je možné predikovat průměrná hrubá mzda v kraji z parametrů obsažených v datovém souboru „prumerna_mzda.csv“. Na základě síly korelačních koeficientů byla vybrána trojice nezávislých proměnných, které společně se závisle proměnnou *prumerna_hruba_mzda* poslouží jako vstupy do lineární regrese. Hodnoty koeficientů determinace ukázaly, že ze tří zkoumaných proměnných má největší prediktivní úspěšnost proměnná *registr_podniky*, pro kterou vyšel R^2=0.806, hraniční spolehlivost (R^2=0.539) byla naměřena proměnné *HDP* a pro poslední proměnnou *% obyv s VS* bylo naměřeno zanedbatelné R^2=0.379. O výsledku ukazující existenci reálné závislosti, na jejímž základu by se dala predikovat spotřeba, rozhodla vícerozměrná regrese. Korelační determinant pro vícerozměrnou regresi nabyl hodnoty R^2=0.961, model je tedy schopen predikce průměrné hrubé mzdy se spolehlivostí na 96,1 %. Při srovnání maxima pro korelační determinant, který dosahuje hodnoty 1, s výsledkem modelu, můžeme konstatovat, že model pro predikci spotřeby vyšel dobře.

Analýza datového souboru „investice.csv“

Posledním definovaným cílem bylo zjistit, zda se dá predikovat velikost investic na ochranu životního prostředí v kraji z parametrů obsažených v datovém souboru „investice.csv“. K závislé proměnné *investice* byly na základě síly korelačních koeficientů vybrány tři proměnné. Tyto vstupní informace poslouží k modelaci tří jednorozměrných lineárních regresí. Z výsledků koeficientů determinace pro jednotlivé modely lineární regrese (R^2(*index_CO*)=0.729, R^2(*hustota*)=0.610 a R^2(*počet obyvatel*)=0.506) je patrné, že největší spolehlivost pro predikci velikosti investovaných peněžních prostředků dosáhla proměnná *index_CO*. Vícerozměrná lineární regrese vykázala výsledek R^2=0.980, z něhož lze usuzovat,

že je možné s velkou spolehlivostí predikovat objem vynaložených peněžních prostředků na ochranu životního prostředí v krajích České republiky.

Implementace

U všech tří modelů bylo prokázáno, že sledovanou proměnou lze predikovat, a to s poměrně malou odchylkou. Bohužel přínos tohoto zkoumání je pro praxi zanedbatelný, protože v praxi se jednotlivé hodnoty měří samostatně bez využití predikce s cílem dosažení maximální přesnosti.

2.2 Aplikování logistické regrese na vhodně vybraných datech

Aplikace logistické regrese bude provedena na datových souborech, které už byly použity pro modelaci lineární regrese. Budou sledovány stejné závislé a nezávislé proměnné.

Porozumění problému

Vzhledem k tomu, že pro kapitolu 2.2 jsou použity stejné datové soubory a proměnné, jako při aplikaci modelu lineární regrese v kapitole 2.1 a zároveň se neliší ani požadavky na logistickou regresi dle metody CRISP-DM, tak není v této fázi nutné opakovaně provádět úpravy, a proto se obsah této části shoduje s obsahem věnovaným porozumění problému pro lineární regresi.

Porozumění datům

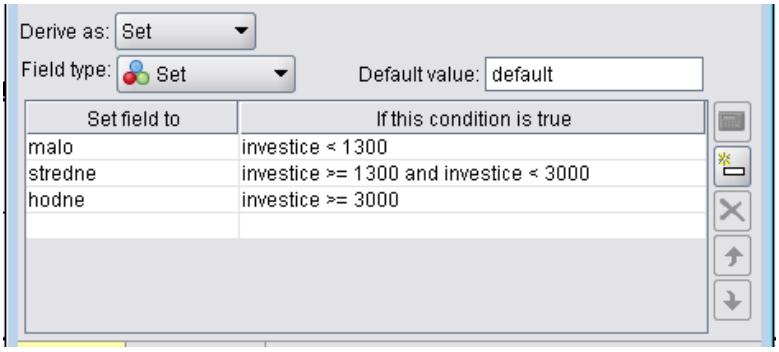
Jak už bylo zmíněno v předchozím odstavci, tak na základě shodnosti dat a požadavků na oba typy regresí nebylo v této fázi nutné rovněž provádět žádné změny, a proto je tato část věnovaná porozumění datům pro logistickou regresi shodná s částí pro lineární regresi.

Příprava dat

Při přípravě dat proběhly všechny kroky, které byly popsány v části věnované přípravě dat pro lineární regresi. Následně byly vytvořeny kategorizované proměnné ze spojitých hodnot proměnné *spotřeba* z datového souboru „auta.csv“, proměnné *prumerna_hruba_mzda* ze souboru „prumerna_hruba_mzda.csv“ a proměnné *investice* ze souboru „investice.csv“. Tyto hodnoty poslouží pro modelaci logistické regrese jako závisle proměnné. Náhled této úpravy pomocí uzlu *Derive*, je možné vidět na Obrázku 30. Ukázka vytvoření kategoriálních proměnných:

- Uzel *Derive,* název: *spotreba_k,* Derive as: *Set,* kategorie: *mala (spotreba<5), stredni (spotreba>=5 and spotreba<8), vysoka (spotreba>=8),* datový soubor: *auta.csv*

- Uzel *Derive*, název: *prumerna_hruba_mzda_k*, Derive as: *Set*, kategorie: *mala (prumerna_hruba_mzda<23000), stredni (prumerna_hruba_mzda>=23000 and prumerna_hruba_mzda<24300), vysoka (prumerna_hruba_mzda>=24300)*, datový soubor: *prumerna_mzda.csv*
- Uzel *Derive*, název: *investice_k*, Derive as: *Set*, kategorie: *malo (investice<1300), stredne (investice>=1300 and investice<3000), hodne (investice>=3000)*, datový soubor: *investice.csv*



Obrázek 30: Náhled uzlu Derive

Zdroj: [vlastní]

Modelování dat

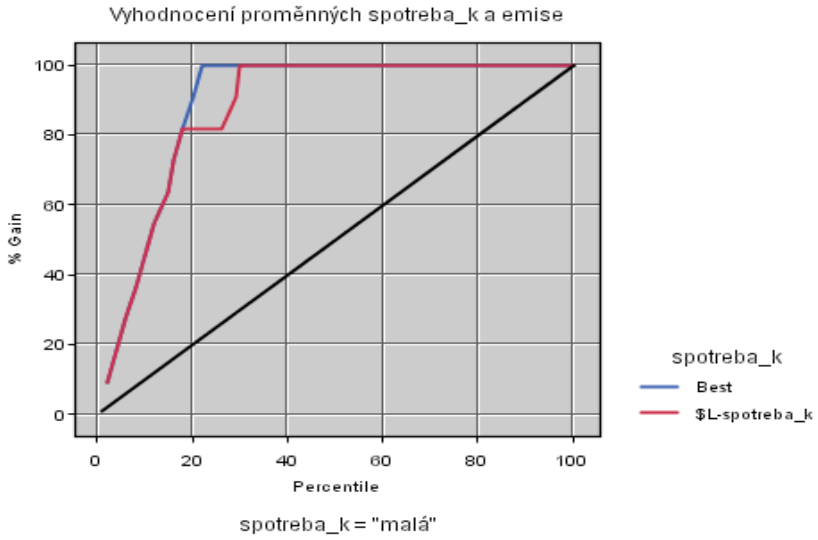
Tato fáze bude obsahovat tři podkapitoly rozlišené dle použitých datových souborů. Jednotlivé části budou obsahovat tři jednorozměrné a jednu vícerozměrnou logistickou regresi. Cílem každé části bude analyzovat data, tak aby mezi jednotlivými analýzami byly nalezeny souvislosti, které povedou k zisku dat s informační hodnotou. Jako vyhodnocující ukazatel byl z pseudo koeficientů determinace zvolen typ Nagelkerke. Tento typ byl zvolen z důvodu, že může reálně dosáhnout hodnoty 1 pro maximální spolehlivost. Pro modelaci je využit softwaru SPSS Clementine.

Analýza vstupních dat souboru „auta.csv“

Tabulka 9: Tabulka informací pro modelaci logistické regrese - spotreba_k vzhledem k emisím

Vstupní informace			
Závislá proměnná	spotreba_k	Nezávislá proměnná	emise
Doplňující informace			
Vyhodnocení modelu			
Pseudo koeficient determinace	0.897	Vyhodnocení	Hodnota spolehlivosti je vysoká. Model vyjádřen na Obrázku 31.

Zdroj: [vlastní]



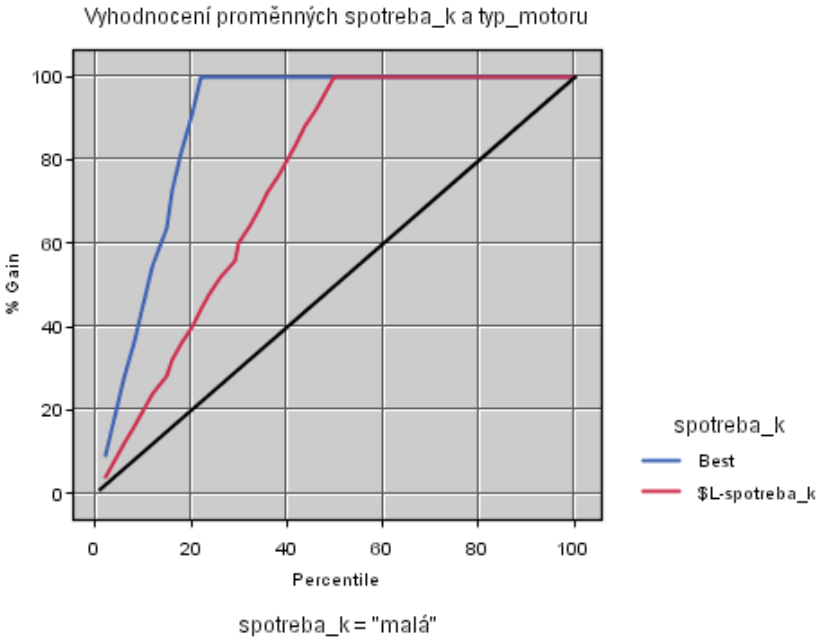
Obrázek 31: Graf typu Evaluation pro nezávislou proměnnou *emise*

Zdroj: [vlastní]

Tabulka 10: Tabulka informací pro modelaci logistické regrese - spotřeba_k vzhledem k typu motoru

Vstupní informace			
Závislá proměnná	spotreba_k	Nezávislá proměnná	typ_motoru
Doplňující informace			
Vyhodnocení modelu			
Pseudo koeficient determinace	0.498	Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen na Obrázku 32.

Zdroj: [vlastní]



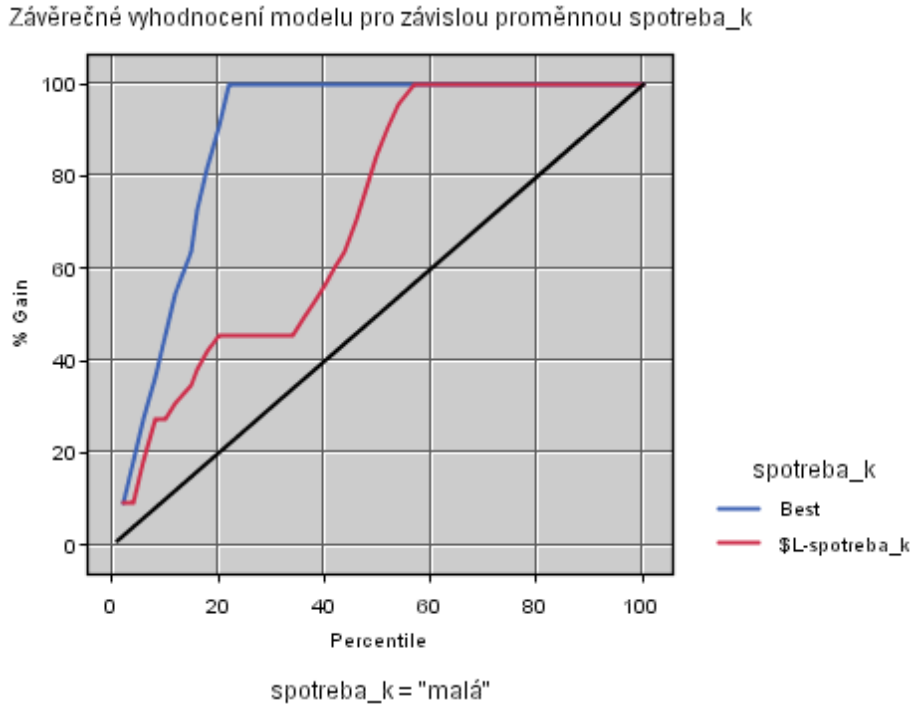
Obrázek 32: Graf typu Evaluation pro nezávislou proměnnou *typ_motoru*

Zdroj: [vlastní]

Tabulka 11: Tabulka informací pro modelaci logistické regrese - spotřeba_k vzhledem k výkonu

Vstupní informace			
Závislá proměnná	spotreba_k	Nezávislá proměnná	vykon
Doplňující informace			
Vyhodnocení modelu			
Pseudo koeficient determinace	0.495	Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen na Obrázku 33.

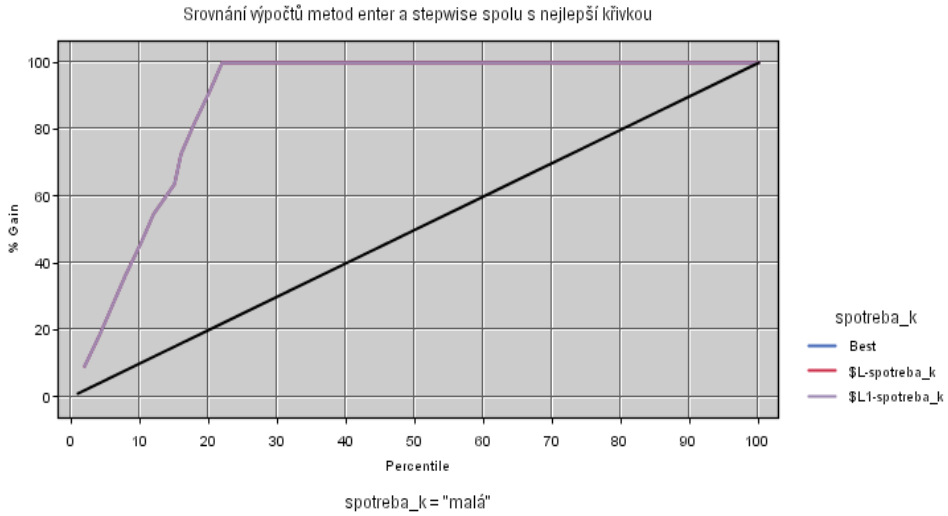
Zdroj: [vlastní]



Obrázek 33: Graf typu Evaluation pro nezávislou proměnnou *vykon*

Zdroj: [vlastní]

Z koeficientu determinace, který nabyl hodnoty 1, je patrné, že model vícerozměrné logistické regrese dokázal klasifikovat data bez jediné chyby. Model je graficky znázorněn na Obrázku 34.



Obrázek 34: Graf typu Evaluation pro srovnání výpočtů metodou enter a stepwise

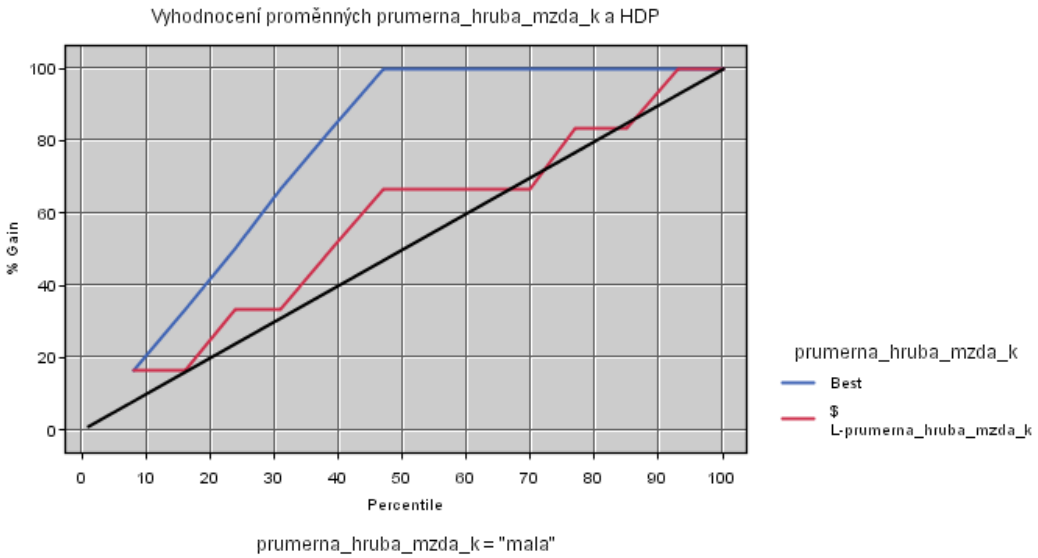
Zdroj: [vlastní]

Analýza vstupních dat souboru „prumerna_mzda.csv“

Tabulka 12: Tabulka informací pro modelaci logistické regrese - mzdy vzhledem k HDP

Vstupní informace			
Závislá proměnná	prumerna_hruba_mzda_k	Nezávislá proměnná	HDP
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Pseudo koeficient determinace	0.533	Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen na Obrázku 35.

Zdroj: [vlastní]



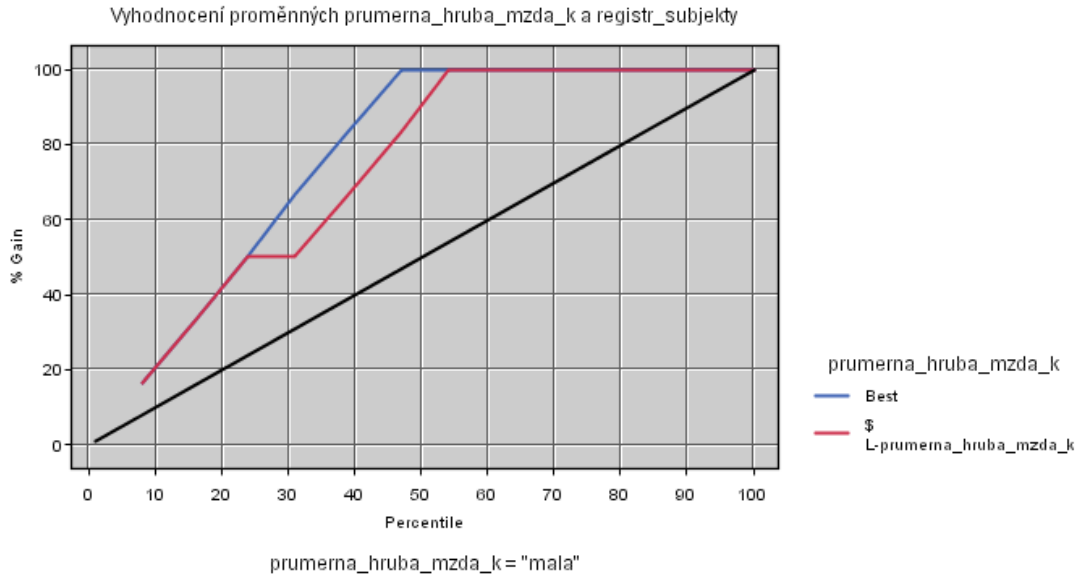
Obrázek 35: Graf typu Evaluation pro nezávislou proměnnou *HDP*

Zdroj: [vlastní]

Tabulka 13: Tabulka informací pro modelaci logistické regrese - mzdy vzhledem k registr_subjekty

Vstupní informace			
Závislá proměnná	prumerna_hruba_mzda_k	Nezávislá proměnná	registr_subjekty
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Pseudo koeficient determinace	0.850	Vyhodnocení	Hodnota spolehlivosti je vysoká. Model vyjádřen na Obrázku 36.

Zdroj: [vlastní]



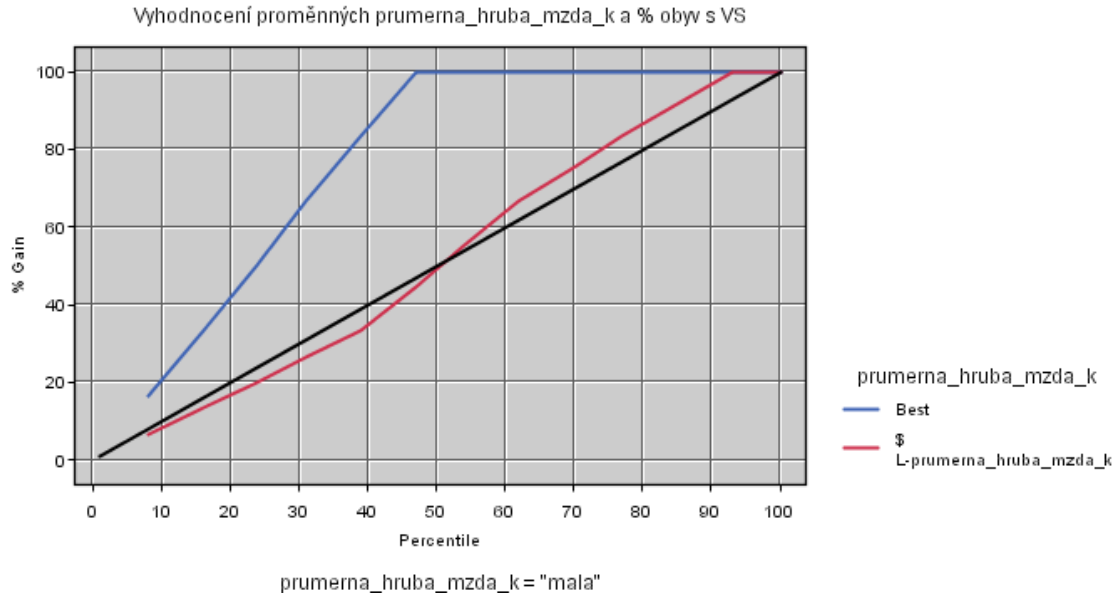
Obrázek 36: Graf typu Evaluation pro nezávislou proměnnou *registr_subjekty*

Zdroj: [vlastní]

Tabulka 14: Tabulka informací pro modelaci logistické regrese - mzdy vzhledem k% obyv s VS

Vstupní informace			
Závislá proměnná	prumerna_hruba_mzda_k	Nezávislá proměnná	% obyv s VS
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Pseudo koeficient determinace	0.549	Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen na Obrázku 37.

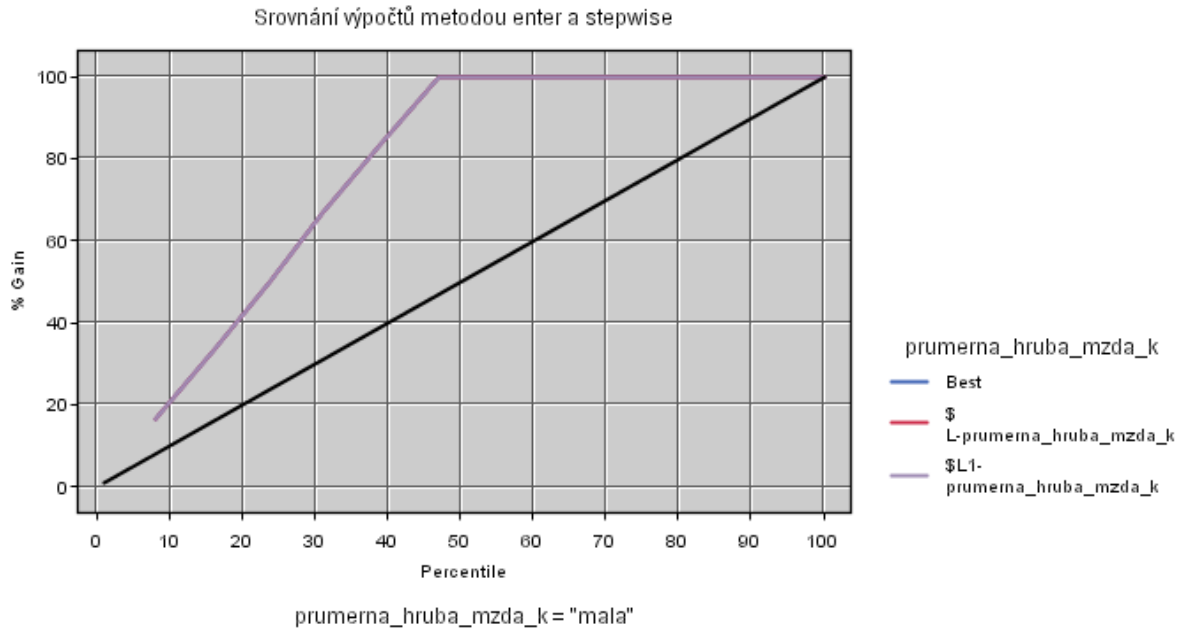
Zdroj:[vlastní]



Obrázek 37: Graf typu Evaluation pro nezávislou proměnnou *% obyv s VS*

Zdroj: [vlastní]

Z koeficientu determinace, který nabyl hodnoty 1 pro metodu enter i stepwise je patrné, že model vícerozměrné logistické regrese dokázal klasifikovat data bez jediné chyby. Model je graficky znázorněn na Obrázku 38. Vstupní nezávislé proměnné jsou *počet_obyvatel*, *podíl_zen*, *pocet_obyv_15-64*, *podíl_zen_15-64*, *pocet_obci*, *mesta*, *registr_subjekty*, *nemocen_pojis*, *zamestnani*, *sluzby*, *% obyv s VS*, *obyv_s_VS*, *zemedel_lesnictvi_rybolov*, *prumysl_stavebnictvi*.



Obrázek 38: Graf typu Evaluation pro srovnání výpočtů metodou enter a stepwise

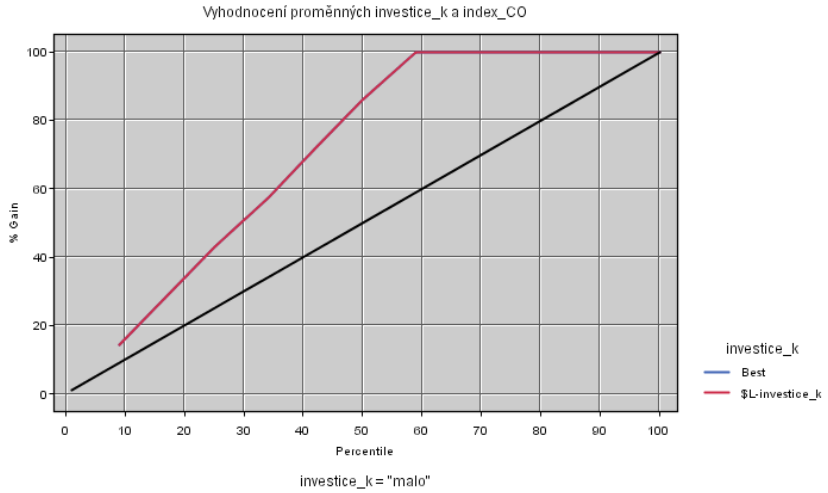
Zdroj: [vlastní]

Analýza vstupních dat souboru „investice.csv“

Tabulka 15: Tabulka informací pro modelaci logistické regrese - investice_k vzhledem k Index_CO

Vstupní informace			
Závislá proměnná	investice_k	Nezávislá proměnná	index_CO
Doplňující informace	Pro zachování objektivity bylo nutné vyřadit odlehlé záznamy pro kraj Hlavní město Praha a Moravskoslezský kraj.		
Vyhodnocení modelu			
Pseudo koeficient determinace	1	Vyhodnocení	Hodnota spolehlivosti je maximální. V grafu je zobrazena přímka nejlepšího možného výsledku, která se překrývá s přímkou modelu. Model je vyjádřen pomocí Obrázku 39.

Zdroj: [vlastní]



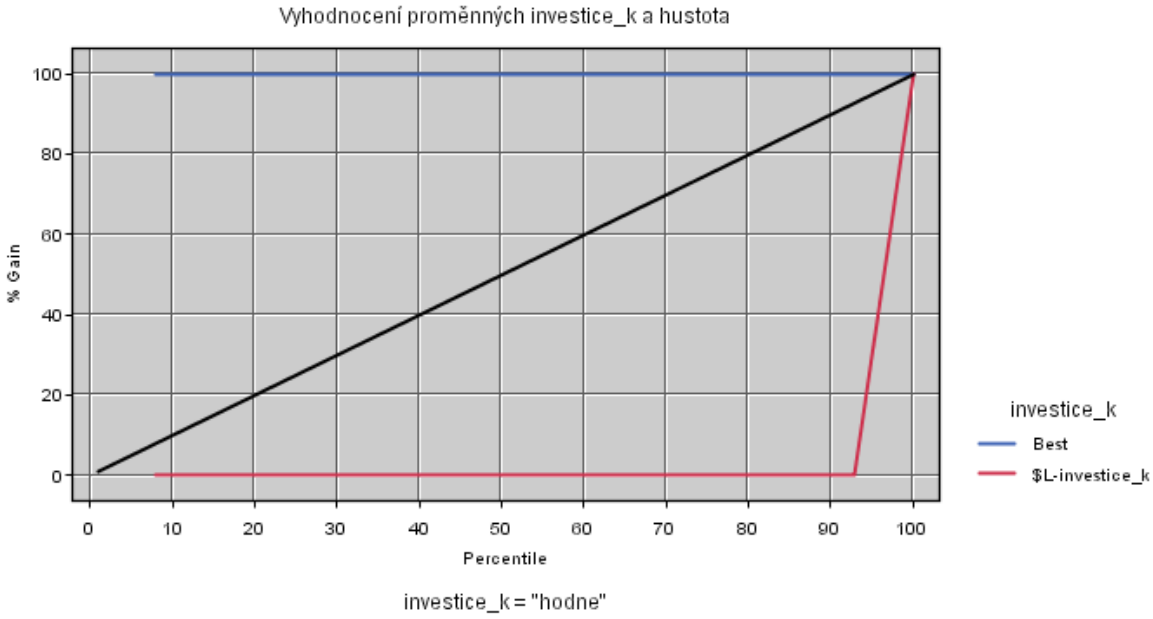
Obrázek 39: Graf typu Evaluation pro nezávislou proměnnou index_CO

Zdroj: [vlastní]

Tabulka 16: Tabulka informací pro modelaci logistické regrese - investice_k vzhledem k hustotě

Vstupní informace			
Závislá proměnná	investice_k	Nezávislá proměnná	hustota
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Pseudo koeficient determinace	0.480	Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen na Obrázku 40.

Zdroj: [vlastní]



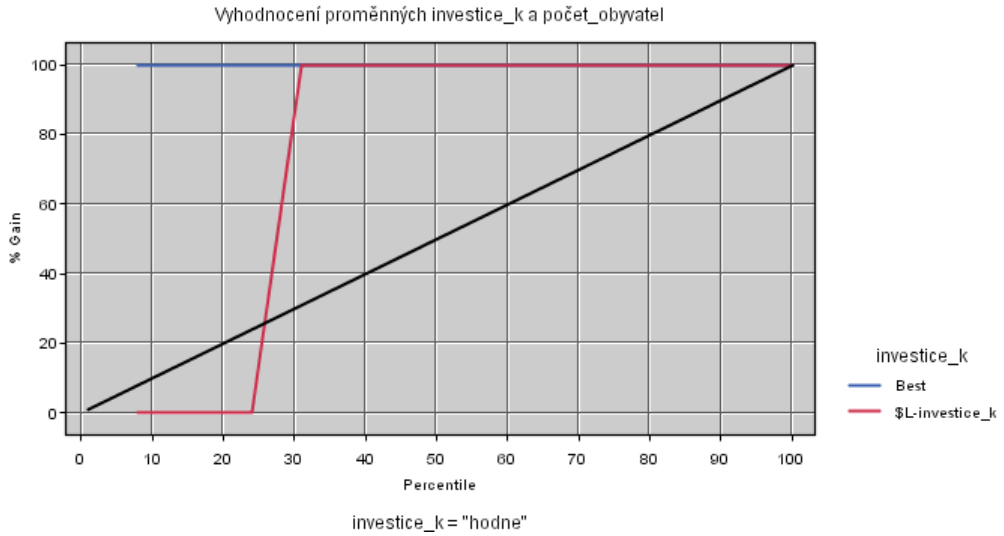
Obrázek 40: Graf typu Evaluation pro nezávislou proměnnou *hustota*

Zdroj:[vlastní]

Tabulka 17: Tabulka informací pro modelaci logistické regrese - investice_k vzhledem k počtu obyvatel

Vstupní informace			
Závislá proměnná	investice_k	Nezávislá proměnná	počet_obyvatel
Doplňující informace	Model počítán po vyřazení odlehlého záznamu proměnné kraj Hlavní město Praha.		
Vyhodnocení modelu			
Pseudo koeficient determinace	0.534	Vyhodnocení	Hodnota spolehlivosti je nedostačující. Model vyjádřen na Obrázku 41.

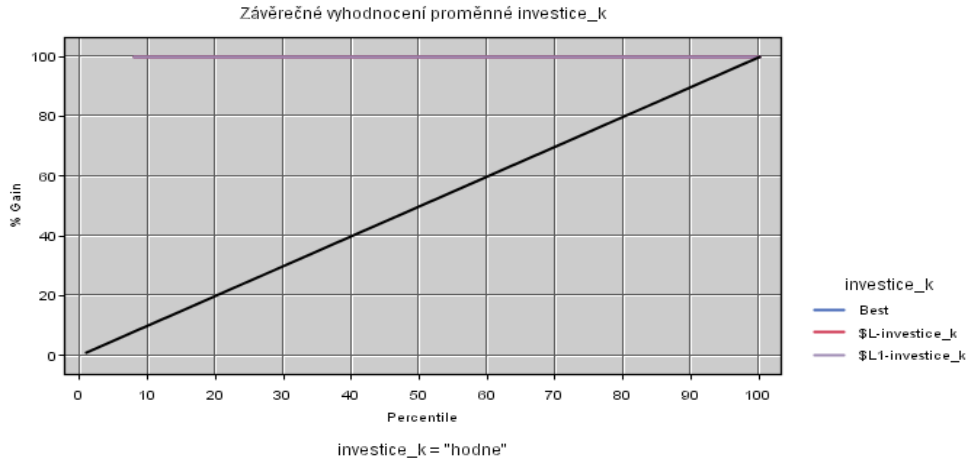
Zdroj: [vlastní]



Obrázek 41: Graf typu Evaluation pro nezávislou proměnnou *počet_obyvatel*

Zdroj: [vlastní]

Z koeficientu determinace, který nabyl hodnoty 1, je patrné, že model vícerozměrné logistické regrese dokázal klasifikovat data bez jediné chyby. V grafu je zobrazena přímka nejlepšího možného výsledku, která se překrývá s přímkou modelu. Model je znázorněn na Obrázku 42. Vstupní nezávislé proměnné jsou *počet_obyvatel*, *zeny*, *obyv_nad65*, *zeny_nad_65*, *index_CO*, *hustota*.



Obrázek 42: Graf typu Evaluation pro srovnání výpočtů metodou enter a backwards stepwise

Zdroj: [vlastní]

Vyhodnocení dat

V této fázi se hodnotí přesnost, s jakou model dosahuje očekávaných cílů, a popisují se důvody, proč je model hodnocen jako dostatečný či nedostatečný. Jako možné vodítko pro odhad výsledků pro logistickou regresi lze použít závěry z kapitoly 2.1.

Analýza datového souboru „auto.csv“

Výsledky logistické regrese se příliš neliší od výsledků lineární regrese. Na základě porovnání pseudo koeficientů determinace pro logistickou regresi, byla jako klíčová proměnná vyhodnocená proměnná *emise*, u které nabyl koeficient determinace hodnoty 0,897. Zbylé dvě proměnné byly shledány jako nedostatečné z důvodu velké nepřesnosti při třídění dat do určených kategorií. Za použití všech proměnných při vstupu do vícerozměrné logistické regrese vyšla splehlivost modelu 100 %. Z tohoto zjištění byl model pro klasifikaci proměnné *spotřeba_k* shledán jako dostatečný se schopností nulové chyby.

Analýza datového souboru „prumerna_mzda.csv“

Lineární regrese opět předpověděla pravděpodobný výsledek logistické regrese. Stejně jako v lineární regresi byla na základě výsledků za klíčovou proměnnou pro logistickou regresi označena *registr_subjekty*, pro kterou pseudo koeficient determinace nabyl hodnoty 0,850. Výsledky jednorozměrných logistických regresí pro proměnné *HDP* a *% obyv s VS*

byly vyhodnoceny jako nedostatečné. A při vstupu všech nezávislých proměnných do vícerozměrné logistické regrese se došlo k závěru, že model je schopný klasifikace se 100% spolehlivostí.

Analýza datového souboru „investice.csv“

Posledním cílem bylo zjistit, zda lze klasifikovat data dle kategorizované proměnné *investice_k*. Z hodnot pseudo koeficientů determinace byla jako klíčová proměnná zjištěna *index_CO*, pro kterou vyšla spolehlivost 100 %. Tento výsledek je diskutabilní vzhledem k možnému zkreslení v důsledku malého počtu záznamů a zvolených kategorií pro proměnnou *investice_k*. Na základě tohoto výsledku by se mohlo zdát zbytečné dělat vícerozměrnou logistickou regresi, ale v důsledku možné zkreslenosti budeme vyvozovat závěry, až z výsledku vícerozměrné logistické regrese. Pro vícerozměrnou logistickou regresi vyšla spolehlivost 100 %, proto lze s jistotou konstatovat, že lze klasifikovat proměnnou *investice_k*.

Implementace

U všech tří modelů bylo pomocí vícerozměrné logistické regrese prokázáno, že sledované proměnné lze bezchybně klasifikovat. Bohužel v praxi tato skutečnost nemá velký přínos. Zkoumané kategorie poskytují vhodný rozsah jen pro odhad. V dnešní době, kdy se dbá na přesnost, mají tyto odhady zanedbatelnou hodnotu.

ZÁVĚR

V části věnované teorii byly zmíněny hlavní body, které posloužily k dostatečnému seznámení a porozumění problematice, jíž se tato práce zabývá. Obsah první kapitoly je tvořen zejména definicí data miningu, seznámením s metodikou CRISP-DM, která zajistila plynulý přechod do části praktické, a v poslední části teoretické přípravy jsou popsány regresní úlohy.

Cílem bakalářské práce bylo najít skryté závislosti mezi prezentovanými daty, provést statistickou analýzu dat a zhodnotit získané výsledky. Data mining byl proveden prostřednictvím metodiky CRISP-DM. Z důvodu vytvoření modelů pro lineární a logistickou regresi byla provedena analýza dat a popisné statistiky. Výsledkem hledání závislostí byly vstupní závislé a nezávislé proměnné do modelů jednorozměrných a vícerozměrných lineárních regresí. Pro každý model byla vypracována informační tabulka a výsledek prezentován grafem. Informační tabulka obsahovala poznámky, parametry a vyhodnocení modelu.

Data mining byl proveden v prostředí softwaru SPSS Clementine, který je velmi efektivní pro získávání znalostí z databází. Manipulace s tímto programem byla podpořena absolvováním předmětu Data mining I. Prostředí tohoto softwaru bych hodnotil jako velmi přehledné, logicky uspořádané a plně vhodné pro vypracování práce dle metody data miningu CRISP-DM. Pozornost byla věnována především přípravě a modelování dat.

Z fáze modelování jsem se musel několikrát vrátit do fáze přípravy dat, když při modelování byly zjištěny nedostatky, které by mohly zkreslit výsledky. Fáze přípravy dat obsahovala zásadní úpravy především na vstupních proměnných a vytvoření nových proměnných. Stěžejní částí bakalářské práce je fáze modelování dat, kde bylo celkem vypracováno 24 modelů pro lineární a logistickou regresi. K lepšímu porozumění, jak fungují jednotlivé metody výpočtu regrese, byly vypracovány modely vícerozměrné regrese s využitím různých metod zpracování. Porovnání metod enter a stepwise přineslo několik zjištění. Metody se zanedbatelně liší v přesnosti výsledku, ale co je hlavní, liší se pořadím, jakým jsou proměnné vkládány do výpočtu. Při použití standardní metody enter, jsou všechny proměnné vloženy do výpočtu najednou, ale při použití metody stepwise, jsou jednotlivé proměnné vkládány do výpočtu postupně, podle předem zadaných matematických kritérií. O tom, v jakém pořadí proměnné vstupují do analýzy, nerozhoduje člověk, ale software.

Celkové zhodnocení výsledků pro logistickou a lineární regresi jsem pak provedl ve fázi vyhodnocení dat.

POUŽITÁ LITERATURA

[1] BERKA, P.: *Dobývání znalostí z databází*. 1. vyd. Praha: Academia, 2003, 368 s. ISBN 80-200-1062-9

[2] *Data mining*. [online]. 2013. Wikipedia [cit. 2013-03-23] Dostupný z: http://cs.wikipedia.org/wiki/Data_mining

[3] EVERITT, B.S a SKRONDAL, A.: *The Cambridge Dictionary of Statistics*.4.vyd, Cambridge University Press, 2010, 478 s. ISBN:978-0521766999

[4] HEBÁK, P.: *Regrese I. Část*. Praha: Vysoká škola ekonomická v Praze, 1998, 138 s. ISBN: 80-7079-909-9.

[5] KREJČÍ, J.: *Automatizované získávání znalostí z dat.* [online]. Praha: Komix s. r. o., 1992. [cit. 2006-04-03]. Dostupný z: http://www.komix.cz/home/komix_cz/podpora/ke_stazeni/prezentace.aspx#MD_1999.

[6] KUBANOVÁ, J.: *Statistické metody pro ekonomickou a technickou praxi*. Bratislava: Static, 2003, 187 s. ISBN: 80-85659-31-X.

[7] KULHAVÝ, L.: *Data Mining, „Dobývání znalostí z databází“*. [online], 2013. [cit 2013-3-25] Dostupný z: http://www.corporateict.cz/12031-pro-strategicke-rozhodovani-na-zakl-dat/data-mining-dobyvani-znalosti-z-databazi.html

[8] PETR, P.: Data mining, díl 1. Pardubice: Univerzita Pardubice, 2006, 144 s. ISBN 80-7194-886-1.

[9] POSPÍŠIL, J. a NEMRAVA, M.: *Dolování dat a jeho aplikace*. [online], 2006 [cit. 2013-03-26] Dostupný z: http://axpsu.fpf.slu.cz/~sos10um/trendy/DM.pdf

[10] PROCHÁZKA, M.: *Data mining: Jiný pohled na problém.* [online], 2010 [cit. 2013-03-26] Dostupný z: http://vtm.e15.cz/aktuality/data-mining-jiny-pohled-na-problem

[11] *Rozhodovací stromy*. [online]. 2013. Wikipedia [cit. 2013-03-28] Dostupný z: http://cs.wikipedia.org/wiki/Rozhodovac%C3%AD_stromy

[12] RUD, O. Parr: Data Mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: ComputerPress, 2001, 329 s. ISBN 80-7226-577-6

[13] *Shluková analýza.* [online]. 2013 Wikipedia [cit. 2013-04-01] Dostupný z: http://cs.wikipedia.org/wiki/Shlukov%C3%A1_anal%C3%BDza

[14] SCHAUER, P.: *Metoda nejmenších čtverců*. [online], VUT Brno, 2013. [cit 2013-3-25] Dostupný z:http://fyzika.fce.vutbr.cz/doc/vyuka_schauer/metoda_nejmensich_ctvercu.pdf

SEZNAM PŘÍLOH

Přílohy

Příloha 1: Elektronická příloha

Vložené CR-R médium obsahuje tyto soubory:

\\vstupní_data

auta.xlsx

prumerna_mzda.xlsx

investice.xlsx

\\datové_soubory

auta.csv

prumerna_mzda.csv

investice.csv

\\streamy

Obsahuje soubory streamů s prací v SPSS Clementine

Analyza1.str

Analyza2.str

Analyza3.str

Příloha 2: Datový slovník pro „auta.csv“

Název atributu	Popis atributu	Typy dat	Hodnoty
nazev	Přesný název automobilu	Set	[A-Z]
znacka	Výrobce automobilu	Set	[Audi, Škoda, BMW, Volkswagen]
karoserie	Karoserie automobilu	Set	[SUV,….., Sedan]
spotreba	Kombinovaná spotřeba auta [l/100 km]	Range	[3.8, 13.9]
palivo	Spalované palivo	flag	[benzin, nafta]
zdvihovy_objem	Zdvihový objem motoru [cm3]	Range	[1197, 4395]
vykon	Výkon motoru v [kW/min-1]	Range	[55, 412]
typ_motoru	Typ motoru	Set	[MPI, TDI, TSI, TFSI]
tocivy_moment	Točivý moment motoru [Nm/min-1]	Range	[148, 680]
zrychleni	Zrychlení z 0-100 km/h v sekundách	Range	[4.4, 15.5]
max_rychlost	Nejvyšší možná rychlost [km/h]	Range	[162, 250]
emise	Naměřené emise [g/km]	Ordered set	[99, 325]
dvere	Počet dveří	Ordered set	[2, 3, 4, 5]
mista	Počet míst k sezení	Range	[4, 5]
hmotnost	Celková hmotnost [kg]	Range	[1570, 2995]

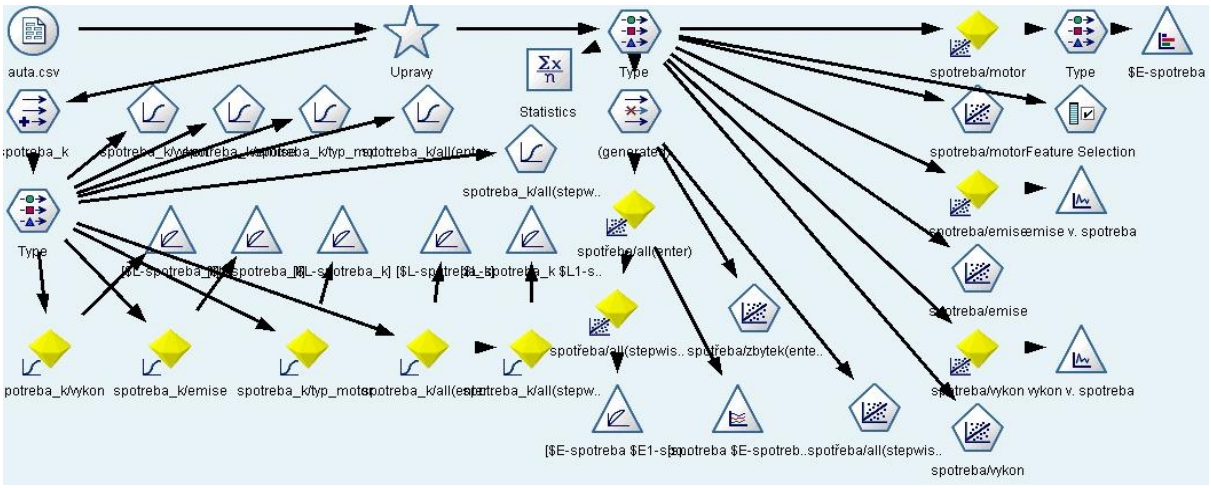
Příloha 3: Datový slovník pro „prumerna_mzda.csv“

Název atributu	Popis atributu	Typ dat	Hodnota
kraj	Název Kraje	Set	[A-Z]
počet obyvatel	Počet obyvatel v kraji	Range	[303165, 1273094]
podíl_zen	Počet žen v kraji	Range	[153733, 644325]
pocet_obev_15-64	Počet obyvatel ve věku mezi 15-64 lety	Range	[212394, 880832]
podíl_zen_15-64	Počet žen ve věku mezi 15-64 lety v kraji	Range	[104820, 434088]
hustota	Hustota zalidnění [obyvatel/km2]	Range	[63.3, 2502.7]
mest_obyv	Podíl městského obyvatelstva	Range	[53.0, 100.0]
pocet_obci	Počet obcí v kraji	Range	[1, 1145]
mesta	Počet měst v kraji	Range	[1, 82]
HDP	HDP na jednoho obyvatele [Kč]	Range	[259180, 786057]
prumerna_hruba_mzda	Průměrná mzda [Kč]	Range	[21723, 33546]
mira_nezamest	Míra nezaměstnanosti v kraji [%]	Range	[3.95, 12.94]
registr_subjekty	Počet registrovaných podnikatelských subjektů	Range	[83396, 529377]
nemocen_pojis	Počet obyvatel s pojištěním proti nemoci	Range	[90733, 993771]
prum_prac_neschop	Průměrná pracovní neschopnost [%]	Range	[2.927, 4.383]
prumerny_vek	Průměrný věk obyvatele	Range	[40.3, 41.9]
obyv_s_VS	Počet obyvatel s ukončenou VŠ	Range	[17700, 263272]
zamestnani	Počet zaměstnaných obyvatel v kraji	Range	[143400, 650300]
zemedel,lesnictvi_rybolov	Počet obyvatel pracujících v zemědělství, lesnictví nebo rybolovu	Range	[1800, 17400]
prumysl_stavebnictvi	Počet obyvatel pracujících v průmyslu a stavebnictví	Range	[56700, 238000]
sluzby	Počet obyvatel pracujících v tržních a netržních službách	Range	[83000, 537100]
ekonom_aktivita	Míra ekonomické aktivity obyvatelstva [%]	Range	[56.0, 61.4]

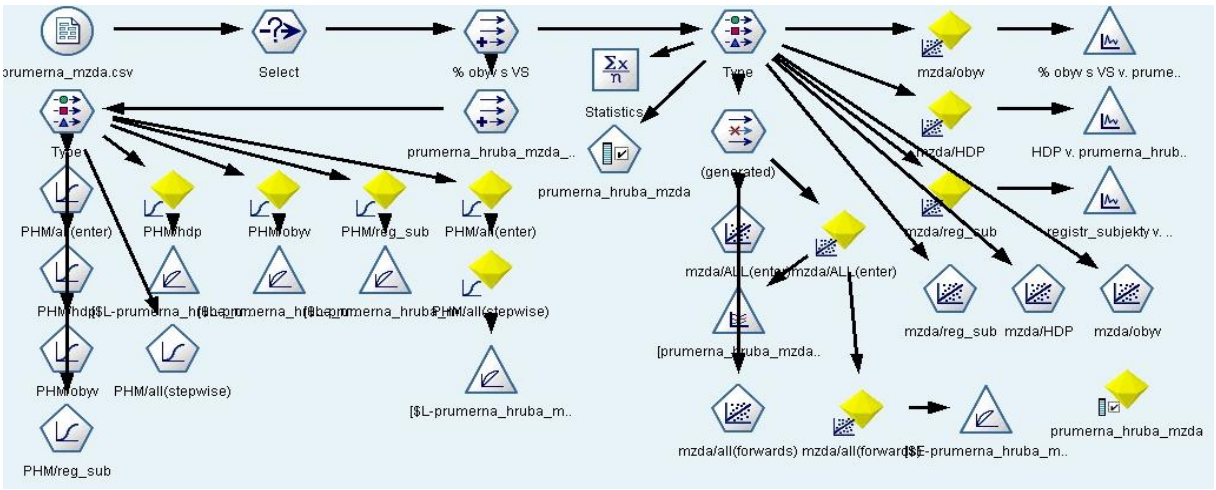
Příloha 4: Datový slovník pro „investice.csv“

Název atributu	Popis atributu	Typ dat	Hodnota
kraj	Název kraje	Set	[A-Z]
počet_obyvatel	Počet obyvatel v kraji	Range	[303165, 1273094]
zeny	Počet žen v kraji	Range	[153733, 644325]
obyv_nad65	Počet obyvatel starších 65 let v kraji	Range	[46155, 213508]
zeny_nad_65	Počet žen starších 65 let v kraji	Range	[27269, 126874]
index_SO2	Index znečištění oxidem siřičitým [t/km^2]	Range	[0.4, 10.9]
index_Nox	Index znečištění oxidy dusíku [t/km^2]	Range	[1.2, 13.94]
index_CO	Index znečištění oxidem uhelnatým [t/km^2]	Range	[2.2, 30.82]
investice	Investice na ochranu životního prostředí [miliony Kč]	Range	[745, 4793]
zemreli	Podíl zemřelých v daném kraji [%]	Range	[9.7, 10.9]
smrtelne_nehody	Počet smrtelných nehod	Range	[21, 97]
obyv_na_lekare	Počet obyvatel na 1 lékaře	Range	[133, 319]
obyv_lehatko	Počet obyvatel na 1 nemocniční lůžko	Range	[130, 219]
prumerny_vek	Průměrný věk obyvatelstva v kraji	Range	[40.3, 41.9]
hustota	Hustota obyvatelstva na 1 km^2	Range	[63, 2503]
prijem	Průměrný čistý peněžní příjem na 1 člena rodiny	Range	[124704, 193993]
chudoba	Podíl osob ohrožených chudobou [%]	Range	[4.5, 16.2]

Příloha 5: Stream pro „auta.csv“



Příloha 6: Stream pro „prumerna_mzda.csv“



Příloha 7: Stream pro „investice.csv“