

České vysoké učení technické v Praze  
Fakulta jaderná a fyzikálně inženýrská

Katedra softwarového inženýrství

Obor: Aplikace informatiky v přírodních vědách



**Analýza příčin vzniku shrinku  
produktů společnosti na základě  
logistických dat**

**Root Cause Analysis of Shrinkage  
Based on Logistics Data**

DIPLOMOVÁ PRÁCE

Vypracoval: Bc. Anna Gruberová

Vedoucí práce: Ing. Martin Plajner, Ph.D.

Rok: 2023



České vysoké učení technické v Praze  
Fakulta jaderná a fyzikálně inženýrská

Katedra softwarového inženýrství

Akademický rok 2022/2023

## ZADÁNÍ DIPLOMOVÉ PRÁCE

|                              |   |
|------------------------------|---|
| <b>Student:</b>              | <b>Bc. Anna Gruberová</b>   |
| <b>Studijní program:</b>     | <b>Aplikace informatiky v přírodních vědách</b>                                       |
| <b>Název práce:</b>          | <b>Analýza příčin vzniku shrinku produktů společnosti na základě logistických dat</b> |
| <b>Název práce anglicky:</b> | <b>Root Cause Analysis of Shrinkage Based on Logistics Data</b>                       |

### **Pokyny pro vypracování:**

1. Seznamte se s procesy v dodavatelském řetězci. Pozornost věnujte štihlé logistice a plýtvání v logistických procesech.
2. Provedte přípravu dat – definujte shrink, rozpoznajte a označte jej v obdržených datech.
3. Na základě předchozího kroku najděte a analyzujte příčiny, které vedou ke vzniku shrinku produktů dané společnosti. Zjištěné výsledky vhodně vizualizujte.
4. Navrhněte automatizaci tohoto postupu, kterou by bylo možné aplikovat i na logistická data dalších společností.

### Doporučená literatura:

- [1] JIRSÁK, P., MERVART M., VINŠ M. *Logistika pro ekonomy - vstupní logistika*. 1. vydání. Praha: Wolters Kluwer ČR, 2012. ISBN 978-80-7357-958-6.
- [2] SUTHERLAND J., BENNET B. *The Seven Deadly Wastes of Logistics: Applying Toyota Production System Principles to Create Logistics Value*. Bethlehem, PA: Lehigh University, 2007.
- [3] HASTIE, T., TIBSHIRANI R., FRIEDMAN J. H. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer, c2009. Springer series in statistics. ISBN 978-0-387-84857-0.

### Jméno a pracoviště vedoucího práce:

**Ing. Martin Plajner, Ph.D.**

Oddělení matematické teorie rozhodování, Ústav teorie informace a automatizace AV ČR, v.v.i.

.....  
vedoucí práce

**Datum zadání diplomové práce:** 12. 10. 2022

**Termín odevzdání diplomové práce:** 3. 5. 2023

Doba platnosti zadání je dva roky od data zadání.

.....  
garant oboru

.....  
vedoucí katedry

.....  
děkan

V Praze dne 12. 10. 2022

## **Prohlášení**

Prohlašuji, že jsem svou bakalářskou práci vypracovala samostatně a použila jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v příloženém seznamu.

V Praze dne .....

.....

Bc. Anna Gruberová

## **Poděkování**

Chtěla bych poděkovat Ing. Václavu Máchovi za vedení mé bakalářské práce, za cenné rady a připomínky k tvorbě této práce a za čas strávený touto pomocí.

Bc. Anna Gruberová

*Název práce:*

**Analýza příčin vzniku shrinku produktů společnosti na základě logistických dat**

*Autor:* Bc. Anna Gruberová

*Studijní program:* Aplikace přírodních věd

*Obor:* Aplikace informatiky v přírodních vědách

*Druh práce:* Diplomová práce

*Vedoucí práce:* Ing. Martin Plajner, Ph.D.  
Oddělení matematické teorie rozhodování, Ústav teorie informace a automatizace AV ČR, v.v.i.

*Konzultant:* –

*Abstrakt:*

*Klíčová slova:* Datová analýza, Logistika

*Title:*

**Root Cause Analysis of Shrinkage Based on Logistics Data**

*Author:* Bc. Anna Gruberová

*Abstract:*

*Key words:* Data Analysis, Logistics





# Obsah

|   |           |
|---|-----------|
| <b>Úvod</b>                                       | <b>11</b> |
| <b>1 Logistika</b>                                | <b>13</b> |
| 1.1 Definice Logistiky . . . . .                  | 13        |
| 1.2 Štíhlá logistika . . . . .                    | 13        |
| 1.2.1 Toyota Production System . . . . .          | 14        |
| 1.2.2 Plýtvání v logistických procesech . . . . . | 18        |
| <b>2 Použité metody</b>                           | <b>19</b> |
| 2.1 Redukce dimenzionality . . . . .              | 19        |
| 2.1.1 Analýza hlavních komponent . . . . .        | 19        |
| 2.1.2 Korespondenční analýza . . . . .            | 21        |
| <b>3 Shrink</b>                                   | <b>23</b> |
| 3.1 Definice . . . . .                            | 23        |
| 3.2 Typy shrinků . . . . .                        | 23        |
| <b>4 Zpracování dat</b>                           | <b>27</b> |
| 4.1 Popis obdržení dat . . . . .                  | 27        |
| 4.2 Předzpracování dat . . . . .                  | 30        |
| 4.3 Statistické zpracování dat . . . . .          | 30        |
| 4.3.1 Období jednoho roku . . . . .               | 30        |
| 4.3.2 Období jednoho měsíce . . . . .             | 33        |
| <b>Závěr</b>                                      | <b>41</b> |
| <b>Literatura</b>                                 | <b>42</b> |
| <b>Přílohy</b>                                    | <b>45</b> |
| <b>A Obsah přiloženého CD</b>                     | <b>45</b> |



# Úvod

Tato diplomová práce se zabývá

První kapitola se věnuje

V následující kapitole

Ve třetí kapitole se nachází

Poslední kapitola



# Kapitola 1

## Logistika

### 1.1 Definice Logistiky

Logistika zahrnuje všechny operace, které se týkají doručení zboží nebo služeb od výrobce k zákazníkovi, s výjimkou samotné výroby zboží nebo provádění služby. Výrobou je naopak rozuměno vše, co mění podobu materiálu. Během výroby se však logistika uplatňuje, například jako přesun materiálu nebo polotovarů mezi jednotlivými výrobními zařízeními. Operace lze rozdělit do tří hlavních toků: materiálový, informační a finanční tok. Materiálový obsahuje všechny pohyby týkající se fyzického materiálu, tedy jeho získávání, přesuny a skladování, a to jak mezi zákazníky, dodavateli či výrobními areály a sklady, tak i vnitřní pohyby mezi produkčními linkami nebo skladovými pozicemi. Informační tok popisuje procesy vznikající během materiálového toku, dále se do něj řadí analýzy již proběhlých toků a plánování a předpovědi budoucích toků. Poslední kategorie, finanční tok mapuje náklady způsobené předešlými dvěma zmíněnými toky.[1]

Pojem logistika je úzce propojen s pojmem Supply Chain Management (SCM)<sup>1</sup>. Zatímco logistika se zabývá toky zboží, služeb či lidí, Supply Chain Management zahrnuje operace logistiky, navíc ale sleduje vztahy mezi procesory, které koordinuje a optimalizuje za účelem naplnění určitých cílů. Tímto cílem bývá často snížení nákladů v rámci částí procesu nebo zvýšení konkurenceschopnosti podniku [3]. Supply Chain Management se tedy prolíná s pojmem logistika a bývají často zaměňovány. Důvodem může být i to, že se jedná o nový pojem, který byl poprvé použitý v roce 1982.[2]

### 1.2 Štíhlá logistika

Štíhlost neboli „lean“ je koncept neustálého vylepšování procesu vytváření produktu nebo služby pomocí odstranění jakéhokoli plýtvání. Plýtváním rozumíme jakoukoli

---

<sup>1</sup>Do češtiny lze Supply Chain Management přeložit jako řízení či správa dodavatelského řetězce. V českém prostředí se používá jak anglická tak česká podoba.

činnost, která v očích zákazníka nezvyšuje hodnotu produktu a tedy není ochotný za tuto činnost zaplatit ve formě vyšší prodejní ceny. Z této definice plýtvání je patrné, že pohled zákazníka hraje důležitou roli při vytváření hodnoty produktu ve štihlých systémech.[14, ?]

Svůj původ nachází štihlá logistika na začátku 20. století, kdy Henry Ford zavedl pohyblivou montážní linku při výrobě automobilu Ford modelu T. Tato linka měla za následek několikanásobné snížení výrobního času a odstartovala sériovou výrobu aut. Díky čemuž se snížila prodejní cena, a automobily tak byly dostupné nejen nejbohatší vrstvě společnosti. Po druhé světové válce navázala automobilová společnost Toyota Motor Company na Fordovu efektivní montážní linku a vytvořila systém nazvaný Toyota Production System (TPS), který je přímým předchůdcem štihlé logistiky.[13]

### 1.2.1 Toyota Production System

Toyota Production System je založen na pěti základních principech. Nejdůležitějším krokem je odstranit plýtvání. Je třeba se soustředit na jednotlivé procesy a na vazby mezi nimi. Pomocí metody genchi genbutsu<sup>2</sup> se nasbírají data a informace o procesech přímo na místě, kde procesy probíhají, aby případné problémy a zdroje plýtvání mohly být přesně určeny. Po této analýze se aplikuje přístup řešení problémů zvaný kaizen<sup>3</sup>, jehož cílem je kontinuální zlepšování procesů. Posledním z principů je dodržování vzájemného respektu mezi všemi oddělení společnosti, jak vedoucími pracovníky, tak zaměstnanci u výrobních linek. [13]

V TPS je plýtvání rozděleno do tří kategorií - Muda (plýtvání), Mura (nevyváženost) a Muri (přetěžování) [16]. V následující části jsou podrobněji popsány jednotlivé typy.

#### Muda

Japonské označení Muda v překladu znamená plýtvání, neúčinnost či marnost. Muda zahrnuje všechny činnosti, které nepřispívají ke zvyšování hodnoty produktu. Mudu lze rozdělit na dva podtypy – 1. typ zahrnuje aktivity, které jsou nezbytné pro koncového zákazníka, např. testování, zda je produkt nebo služba bezpečná. Druhý typ obsahuje ty procesy, které již zákazník nepotřebuje, či dokonce nechce, neboť mohou mít vliv na rychlost výroby produktu (výkonu služby) nebo přímo na jeho kvalitu.

Taiichi Ohno, manažer ve společnosti Toyota, identifikoval sedm typů plýtvání, někdy nazývané *seven deadly wastes*. Klasifikace a popis včetně příkladů je uveden níže [13]:

1. **Nadprodukce** – Pokud je vyrobeno více produktů, než je možné expedovat k zákazníkovi, nebo více materiálu, než kolik je požadováno k další výrobě či okamžité spotřebě.

---

<sup>2</sup>Genchi v překladu znamená skutečná lokace a genbutsu skutečná věc.

<sup>3</sup>Kaizen je japonský překlad slova zlepšení.

2. **Zpoždění/čekání** – Jakákoli prodleva mezi dvěma na sebe navazujícími procesy, např. čekání jedné montážní linky na meziprodukty z jiné linky vlivem rozdílných výrobních časů nebo vlivem nedostatečné výrobní kapacity jednoho ze strojů, dále sem patří také čekání zaměstnanců z důvodu kontroly odvedené práce, pomalého načítání počítačového programu nebo čekání na konkrétní instrukce k výkonu práce [14].
3. **Transport** – Zbytečný přesun produktů, materiálů nebo informací. Tento transport navíc může vést k poškození produktu. Příkladem tohoto typu plýtvání může být situace, kdy materiál, který je nejvíce potřebný pro výrobu produktů je umístěn v největší vzdálenosti, nebo pokud přístup k jedné položce ve skladu je blokován jinými položkami.
4. **Pohyb** – Zbytečný pohyb lidí, vzniklý špatným rozmístěním objektů v prostoru, např. nepřiměřeně dlouhotrvající chůze, natahování se pro předměty, vyhýbání se lidem či předmětům.
5. **Skladování** – Pokud je naskladněno více surovin, rozpracovaných výrobků a hotových produktů, než kolik je požadováno, např. předčasná dovážka položek do skladu, chyba v dodávce, naskladnění položek do zásoby tzv. pro jistotu nebo z důvodu množstevní slevy.
6. **Nadbytečné zpracování** – Při výrobě dochází k použití více energie nebo prostředků než nutné, nebo je vytvořen koncový produkt, který má vyšší hodnotu, než jaký je dohodnutý a požadovaný standard.
7. **Defekty** – Produkty či meziprodukty, které je nutné přepracovat nebo odstranit z výroby z důvodu vady.

Tyto podoby plýtvání aplikované v TPS byly inspirací pro identifikaci sedmi typů plýtvání v logistice [13, 4]:

1. **Nadprodukce** – V případě logistiky je nadprodukce chápána jako doručení produktů dříve nebo ve větším množství než bylo požadováno.
2. **Zpoždění/čekání** – Jakákoli prodleva mezi dvěma na sebe navazujícími procesy, např. čekání na převoz meziproductů mezi dvěma výrobními linkami, příjezd kamionu mimo časové okno, doba mezi příjezdem kamionu a jeho naložením nebo čas mezi přijetím objednávky a zahájením její realizace.
3. **Transport** – Zbytečný přesun produktů, materiálů nebo informací, např. materiál, který je nejvíce potřebný pro výrobu produktů je umístěn v největší vzdálenosti, nebo pokud přístup k jedné položce ve skladu je blokován jinými položkami.
4. **Pohyb** – Zbytečný pohyb lidí, např. vzniklý špatnou organizací předmětů ve skladu, kdy položky, ke kterým se nejčastěji přistupuje, jsou v méně přístupných pozicích skladu, nebo dokonce sklad není strukturovaný vůbec, nebo nutnost změnit trasu při převozu položek ve skladu kvůli nedostatečně širokým uličkám.

5. **Skladování** – Pokud je naskladněno více surovin, rozpracovaných výrobků a hotových produktů, než kolik je požadováno, např. předčasná dovážka položek do skladu, chyba v dodávce, naskladnění položek do zásoby tzv. pro jistotu.
6. **Prostor** – Neoptimální využití dostupného místa, např. nedostatečná výška regálů ve skladech, nevyužitá kapacita regálů, neoptimální naložení kamionu, přetížení dostupných kapacit.
7. **Defekty** – Činnosti, které způsobí nutnost opakovat určitý proces, znehodnocení produktu nebo zvýší náklady, např. špatné zavezení produktu, špatné nebo chybějící označení produktu, chyby v evidenci.

V devadesátých letech, kdy se metody TPS začaly aplikovat ve společnostech, byl mezi sedm typů plýtvání Muda začleněn osmý typ - Dovednosti. V tomto případě dochází k neefektivitě kvůli nevyužití lidského potenciálu a talentů jednotlivých zaměstnanců. K tomu může docházet například striktním rozdělením na manažery a zaměstnance, kde role zaměstnanců je poslouchat nařízení shora a vykonávat práci tak, jak byla navržena vedoucími pracovníky. Avšak právě zaměstnanci pracující přímo v terénu lépe identifikují případné problémy a snadněji naleznou řešení díky svým zkušenostem.[14]

## Mura

Mura lze přeložit jako nestejnomyšlnost, nevyrovnanost a nepravidelnost. Jedná se o plýtvání vznikající špatnou provázaností jednotlivých procesů a to jak interních, tak externích. Následkem nevyváženosti je pak vznik plýtvání Muda. [16, 4]

Plýtvání v podobě Mura se rozlišuje jak v procesech informačního, tak hmotného toku. V případě informačního toku je nejvýznamnějším zdrojem plýtvání situace, kdy je chybně predikována poptávka mezi jednotlivými články logistického řetězce. Ignorování vztahů mezi procesy může vést k chybovosti i v řádu desítek procent. Informace, jejichž opomíjení způsobuje chybovost předpovídání poptávky, mohou být např. v jaké fázi životního cyklu se výrobek nachází, plánování promoakcí nebo výrobní a logistická omezení dodavatelů. Další zdroj Mura v informačním toku je nedostatečná znalost stavu zásob mezi dodavatelem a odběratelem. Následkem čehož dochází k méně častým závázkám avšak s větším objemem, což vede k vyšším pojistným zásobám ve skladech. Většinu zmíněných situací lze eliminovat aplikováním konceptu „Just-in-Time“ do jednotlivých procesů. Plýtvání také vzniká při administrativě, pokud nejsou vhodně standardizované dokumenty používané v logistickém řetězci. Příkladem může být špatná evidence pohybů ve skladu či tvorba objednávek. Nesjednocenost v administrativních procesech vede ke zpomalení navažujících činností nebo dokonce k chybám, které způsobí nemožnost dokončení celého procesu. Pak je nutné vybrané procesy provést znovu a napravit chyby.[4]

Plýtvání v hmotném toku je přímým důsledkem chyb vznikajících v informačním toku. Lze identifikovat i takové zdroje plýtvání, které nesouvisejí přímo s informačním tokem, a to například dodržování různých standardů přepravních prostředků na



straně dodavatele a odběratele. To má pak za následek nadbytečné překládání materiálu do podoby, kterou druhá strana používá a se kterou je schopna následně efektivněji manipulovat.[4]

## **Muri**

Pojem Muri označuje přetěžování. Muri často vzniká při snaze zvýšit produktivitu a odstranit tak předešlé typy plýtvání, v konečném důsledku může ale vést k výrazně větší chybovosti i celkovému selhání. Přetížení mohou být zaměstnanci, ale i stroje. V obou případech vytížení na více než 100 % se může projevit na snížení kvality výstupu. Lidé mohou být méně pozorní a může docházet k nehodám, které mohou v menší či větší míře negativně ovlivnit i větší část logistického řetězce. Stroje mohou produkovat zmetkové výrobky, nebo může dojít k jejich poškození až zničení.[4, 16]

## **Příklad plýtvání Muda, Mura a Mudi**

Všechny tři zmíněné typy plýtvání Muda, Mura a Muri jsou navzájem propojené. Tuto skutečnost je třeba brát v potaz při řešení zefektivňování procesů a eliminaci plýtvání. Pro představu je uvedena následující situace. Společnost potřebuje zákazníkovi přivést šest tun materiálu, uloženého ve stejných jednotunových kontejnerech. Možné způsoby řešení této úlohy jsou znázorněny na obr. 1.1.[16]

Nejjednodušší možností je naložit na jeden kamion veškerý požadovaný materiál. V takovém případě společnost ušetří na počtu vozidel a eliminuje tak plýtvání přepravními prostředky, ušetří čas při nakládce a vykládce, protože není nutné obsluhovat více vozidel, zároveň . Na druhou stranu ale hrozí přetížení kamionu. Následkem přetížení se může zvýšit riziko nehody vozidla, firma může být pokutována nebo vozidlu nemusí být umožněn vjezd na určitá místa.

Opačným extrémem je použít tři kamiony, každý se dvěma tunami materiálu. Potom ale není efektivně využita dostupná kapacita a je patrné, že dochází k mnoha druhům plýtvání typu Muda.

Třetí možností je využití dvou kamionů, kdy první je naložen čtyřmi a druhý dvěma tunami. Toto rozložení nepodléhá žádným pravidlům a patrně proces nakládky není dostatečně spjatý s ostatními procesy nebo neprobíhá správný přenos informací o požadavcích mezi jednotlivými procesy. Nakládka a vykládka prvního velmi naloženého kamionu vyžaduje více času než druhého kamionu. Z toho plyne, že buď není možné v dostupném čase stihnout obsloužit první kamion a dochází k přetížení, anebo v případě druhého kamionu je zbude velké množství času a zaměstnanci zbytečně čekají. Z této volby plyne, že plýtvání typu Mura může způsobit Mudu i Muru.[16]

Optimální řešení je naložit dva kamiony po třech tunách, což je jejich ideální kapacita. V takovém případě společnost minimalizuje za daných podmínek všechny tři typy plýtvání. V reálném světě jsou situace mnohonásobně komplexnější a ne vždy existuje jednoznačné optimální řešení, které je navíc snadno dosažitelné. Důležité ale



Muri = overburdened



Mura = unevenness, fluctuation, variation



Muda = waste



No Muri, Mura, or Muda

Obrázek 1.1: Příklady plýtvání Muda, Mura a Muri při transportu šesti tun materiálu.[16]

je soustředit se na všechny tři typy současně, protože optimalizace pouze jednoho kritéria může způsobit jiný druh plýtvání nebo kolaps části systému.

V roce 2011 bylo realizováno dotazníkové šetření Vysokou školou ekonomickou v Praze, které mapovalo, kolik procent logistických expertů se zabývá odstraněním zmiňovaných tří typů plýtvání. Plýtvání Muda se snaží odstranit z logistických procesů 72 % respondentů, Murou se zabývá 39 % a plýtvání Muri řeší 30 % dotazovaných.[4]

## 1.2.2 Plýtvání v logistických procesech

Tato sekce se zabývá třinácti vybranými logistickými procesy z hlediska plýtvání, jak jsou uvedeny v knize *Logistika pro ekonomy – Vstupní logistika*. Analýza vychází z již zmíněného dotazníkového šetření z roku 2011.

# Kapitola 2

## Použité metody

### 2.1 Redukce dimenzionality

#### 2.1.1 Analýza hlavních komponent

Analýza hlavních komponent (anglicky *Principal component analysis*, dále jako PCA) je statistická metoda využívaná pro extrakci proměnných, redukci vícedimenzionálních dat nebo vizualizaci dat. Lze ji aplikovat pouze na kvantitativní data s numerickými, spojitými hodnotami, neboť metoda využívá lineární algebraické techniky, jako je například kovarianční matice, pro jejíž výpočet se předpokládají spojitě hodnoty.

Jednotlivá pozorování obsažená v datech bývají popsána několika různými proměnnými. Tyto proměnné jsou často vzájemně korelované a obsahují šum. Metoda PCA umožňuje extrahovat pouze důležité informace z proměnných a snížit šum. k tomu je třeba vypočítat nové ortogonální proměnné, nazývané hlavní komponenty, které se získají jako lineární kombinace původních proměnných [6]. Hlavní komponenty reprezentují směry největšího rozptylu původních dat a jsou řazeny podle své významnosti. Jinými slovy, první hlavní komponenta zachycuje co nejvíce variability v datech, druhá hlavní komponenta zachycuje co nejvíce variability, která nebyla zachycena první hlavní komponentou, pro zbylé komponenty analogicky.

#### Princip

Předpokládáme množinu dat  $X = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ , kde  $N$  je počet pozorování a každý vektor  $\mathbf{x}_i$  přísluší jednomu pozorování popsanému  $M$  proměnnými.  $\mathbf{X}$  je potom matice rozměru  $N \times M$  vstupních dat. Dále je definovaný výběrový průměr  $\bar{\mathbf{x}}$  jako

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i, \quad (2.1)$$

a výběrová kovarianční matice  $\mathbf{C}$

$$\mathbf{C} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}). \quad (2.2)$$

První hlavní komponentu, která popisuje největší rozptyl dat označíme  $y_{1i}$  a vypočteme následovně jako lineární kombinaci původních proměnných

$$y_{1i} = \mathbf{a}_1^\top (\mathbf{x}_i - \bar{\mathbf{x}}), \quad \text{pro } i = 1, \dots, N, \quad (2.3)$$

kde  $\mathbf{a}_1 = (a_{11}, \dots, a_{M1})^\top$  je vektor vah.

Optimální vektor  $\mathbf{a}_1$  je takový vektor, který maximalizuje výběrový rozptyl nové proměnné  $y_{1i}$  za podmínky  $\mathbf{a}_1^\top \mathbf{a}_1 = 1$ . Pakliže je výběrový rozptyl  $y_{1i}$  definován jako

$$D(y_{11}, \dots, y_{1N}) = \mathbf{a}_1^\top \mathbf{C} \mathbf{a}_1 \quad (2.4)$$

můžeme maximalizační úlohu vyřešit pomocí metody Lagrangeových multiplikátorů. Lagrangeova funkce s parametrem  $\lambda_1$  má následující tvar

$$\mathcal{L}(\mathbf{a}_1, \lambda_1) = \mathbf{a}_1^\top \mathbf{C} \mathbf{a}_1 - \lambda_1 (\mathbf{a}_1^\top \mathbf{a}_1 - 1). \quad (2.5)$$

Derivaci funkce položíme rovnou nule

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{a}_1} &= 2\mathbf{C}\mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 \stackrel{!}{=} 0 \\ (\mathbf{C} - \lambda_1 \mathbf{I})\mathbf{a}_1 &\stackrel{!}{=} 0, \end{aligned}$$

kde  $\mathbf{I}$  je jednotková matice.

Řešíme soustavu lineárních rovnic pro neznámý parametr  $\mathbf{a}_1$ , která má řešení právě tehdy, když je matice  $\mathbf{C} - \lambda_1 \mathbf{I}$  singulární, tedy platí, že její determinant je roven nule.  $\lambda_1$  je pak největší vlastní číslo matice  $\mathbf{C}$  a  $\mathbf{a}_1$  vlastní vektor příslušný tomuto vlastnímu číslu. Toto tvrzení se matematicky zapíše následovně

$$\mathbf{C}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1. \quad (2.6)$$

Po vynásobení vektorem  $\mathbf{a}_1$  zleva získáme řešení pro maximální rozptyl proměnné  $y_{1i}$

$$D(y_{11}, \dots, y_{1N}) = \mathbf{a}_1^\top \mathbf{C} \mathbf{a}_1 = \lambda_1. \quad (2.7)$$

Druhá hlavní komponenta

$$y_{2i} = \mathbf{a}_2^\top (\mathbf{x}_i - \bar{\mathbf{x}}), \quad \text{pro } i = 1, \dots, N, \quad (2.8)$$

se vypočte obdobným způsobem s přidáním podmínkou ortogonalitě vzhledem k první hlavní komponentě – druhá hlavní komponenta nesmí být korelovaná s předchozí, první hlavní komponentou. Potom popisuje druhý největší možný rozptyl v datech. Znázornění dvou hlavních komponent ve dvoudimenzionálním prostoru je vyobrazeno na obrázku 2.1. Vektor  $\mathbf{a}_2$  se opět získá jako jednotkový vlastní vektor kovarianční matice  $\mathbf{C}$  příslušící druhému největšímu vlastnímu číslu  $\lambda_2$ . [6]

Obrázek 2.1: Znázornění dvou hlavních komponent na dvoudimenzionálních datech.

Získání předpisů pro další hlavní komponenty je analogické. Obecně lze zapsat metodu PCA a převod původních proměnných následujícím maticovým zápisem

$$\mathbf{Y} = \mathbf{X}\mathbf{A}, \quad (2.9)$$

kde  $\mathbf{X}$  je matice vstupních dat,  $\mathbf{A}$  je matice vlastních vektorů kovarianční matice  $\mathbf{C}$ . Pro matici  $\mathbf{A}$  zároveň platí  $\mathbf{C} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top$ , kde  $\mathbf{\Lambda}$  je diagonální matice vlastních čísel  $\mathbf{C}$ . [7]

### 2.1.2 Korespondenční analýza

Vícenásobná korespondenční analýza (anglicky *Multiple correspondence analysis*, dále jako MCA) je metoda, která umožňuje popsat vztahy mezi daty, které jsou popsány kategorickými proměnnými, vytvořením kontingenční tabulky. V případě, že se popisuje vzájemná relace pouze dvou proměnných, se použije základní korespondenční analýza<sup>1</sup>. MCA je alternativou k PCA, pokud jsou analyzovány daty kategorická data. [8]

#### Značení

Nechť  $\mathbf{N}$  je matice dat s rozměry  $I \times J$ , kde  $I$  odpovídá počtu pozorování a  $J$  je počet kategorií. Matice  $\mathbf{N}$  je převedena na korespondenční matici  $\mathbf{P}$  vydělením matice  $\mathbf{N}$  jejím celkovým součtem  $n = \sum_{i=1}^I \sum_{j=1}^J n_{ij} = \mathbf{1}_I^\top \mathbf{N} \mathbf{1}_J$ . To zaručuje, že součet prvků matice  $\mathbf{P}$  je roven jedné. Tyto kroky lze shrnout následujícím matematickým zápisem

$$\mathbf{P} = \frac{1}{n} \mathbf{N}, \quad \mathbf{P} = \{p_{ij}\}, \quad \sum_{i=1}^I \sum_{j=1}^J p_{ij} = 1. \quad (2.10)$$

Součet  $i$ tého řádku, resp. součet  $j$ tého sloupce je značen následovně

$$r_i = \sum_{j=1}^J p_{ij} \quad \text{pro } i = 1, \dots, I,$$

$$c_j = \sum_{i=1}^I p_{ij} \quad \text{pro } j = 1, \dots, J.$$

Vektor  $\mathbf{r} = \mathbf{P} \mathbf{1}_J$  obsahuje všechny řádkové součty matice  $\mathbf{P}$ , analogicky vektor  $\mathbf{c} = \mathbf{P}^\top \mathbf{1}_I$  obsahuje všechny sloupcové součty téže matice.

Pro další výpočty zavedeme značení pro diagonální matice, které mají na diagonále řádkový, resp. sloupcový součet

$$\mathbf{D}_r = \text{diag}(\mathbf{r}), \quad \text{resp.} \quad \mathbf{D}_c = \text{diag}(\mathbf{c}). \quad (2.11)$$

---

<sup>1</sup>anglicky *correspondence analysis* (CA)

## Výpočetní algoritmus základní korespondenční analýzy [10, 9]

Označme  $\mathbf{S} = \{s_{ij}\}$  následující matici

$$\mathbf{S} := \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-\frac{1}{2}}. \quad (2.12)$$

Po té provedme singulární rozklad této matice

$$\mathbf{S} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^\top, \quad (2.13)$$

kde  $\mathbf{\Delta} = \mathbf{\Delta}^2$  je matice vlastních čísel  $\lambda_k$  pro  $k = 1, \dots, K$ , kde  $K = \min\{I-1, J-1\}$ . Potom rozměry matice  $\mathbf{U}$ , resp.  $\mathbf{V}$  jsou  $I \times k$ , resp.  $J \times k$ . Dále platí  $\mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbf{I}$ .

Korespondenční analýza měří míru váženého rozptylu, tzv. inercií pomocí vlastních čísel  $\lambda_k$  matice  $\mathbf{S}$ ,  $\lambda_k$  se pak nazývají hlavní inercie. Celková inercie je rovna

$$I = \sum_{k=1}^K \lambda_k = \sum_{i=1}^I \sum_{j=1}^J s_{ij}^2. \quad (2.14)$$

Hlavní komponenta řádků  $\mathbf{F}$  je rovna

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{\Delta}. \quad (2.15)$$

Hlavní komponenta sloupců  $\mathbf{G}$  je rovna

$$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{\Delta} \quad (2.16)$$

## Výpočetní algoritmus MCA

Předpokládejme, že původní matice kategorických dat má tvar  $N \times Q$ , tj.  $N$  pozorování a  $Q$  proměnných. Matici dat převedeme na indikátorovou matici. Indikátorová matice  $\mathbf{Z}$  je vytvořena tak, že kategorická data jsou rozepsána do pomocných proměnných. Pokud  $q$ tá proměnná je má  $J_q$  typů kategorií, tak příslušná indikátorová matice bude mít  $J = \sum_{q=1}^Q J_q$  sloupců a  $N$ . Tzn. počet proměnných byl tímto rozepsáním rozšířen z počtu původních  $Q$  proměnných na  $J$  proměnných. První způsob MCA aplikuje základní algoritmus korespondenční analýzy na matici  $\mathbf{Z}$ , takto se získají souřadnice pro  $N$  pozorování a  $J$  kategorií.

# Kapitola 3

## Shrink

Cílem této práce je analyzovat shrinky produktů, které byly zaznamenány v datech dané společnosti, a zjistit příčiny jejich vzniku. V následující části je vysvětlen pojem shrink a popsány kategorie, které vybraná společnost rozeznává ve svých datech.

### 3.1 Definice

Slovem shrink se označuje ztráta zisku z neuskutečněného prodeje hotového produktu. Tento produkt je vyroben, či naskladněn, ale z nějakého důvodu nemohl být prodán zákazníkovi. Tímto důvodem může být například poničení produktu, jeho ztráta nebo prošlá doba spotřeby. Za shrink produktu lze označovat i stav, kdy cena produktu je neplánovaně snížena v důsledku zmíněných důvodů. Shrinkem je potom rozdíl plánované prodejní ceny a ceny, za kterou byl produkt skutečně prodán.[17]

### 3.2 Typy shrinků

Vybraná společnost rozlišuje ve svých datech tři kategorie shrinku – inventory, damages, price downs. Dále se budu věnovat popisu jednotlivých typů v rámci kategorií. Každý typ má přiřazeno jednoznačné identifikační číslo, podle kterého je zaznamenáván v databázi. Z důvodů anonymizace dat v práci nejsou uvedené přesné hodnoty těchto ID, namísto toho jsou uvedeny názvy, které definují shrinky.

#### Damages

Do kategorie damages (česky škody) jsou řazeny zbylé důvody k odstranění produktu z prodeje z důvodu degradace produktu. V následující tabulce 3.1 jsou vypsané všechny typy, které moho být evidovány.

Tabulka 3.1: Přehled jednotlivých typů shrinků z kategorie damages.

| Název                                 | Popis  |
|---------------------------------------|--|
| Poškození                             | Odpis zboží, které bylo poškozené. Např. nedopečené, spálené, špatně vyrobené nebo poškozené zaměstnancem nebo zákazníkem (kdy nelze uplatnit reklamaci na zákazníka.)   |
| Prošlé a zkažené zboží                | Odpis zboží, kterému prošla doba spotřeby (v případě výrobků, kde je datum uvedené), zkažené či shnilé zboží (ovoce, zelenina) nebo ztvrdlé pečivo.  |
| Potravinová banka                     | Odpis potravinářského zboží, které bylo darováno potravinovým bankám. Jedná se o produkty, které nebylo možné zařadit znovu do oběhu.  |
| Zvířecí útulky                        | Odpis potravinářského zboží, které bylo darováno do útulků zvířat. Jedná se o produkty, které nebylo možné zařadit znovu do oběhu.   |
| Uznané zákaznické reklamace           | Odpis zboží, které zákazník reklamoval a reklamace byla uznána, ale zároveň nelze toto zboží reklamovat u dodavatele.  |
| Neupl. reklamace distribučního centra | Odpis zboží, které fyzicky nedorazilo z distribučního centra a nebylo možné ho reklamovat z důvodu nesplnění limitu pro vytvoření reklamace na distribučním centru. Také obsahuje odpisy neprodaných EXIT položek po ukončení výprodeje. |
| Kompostéry                            | Odpis zboží, které je prošlé nebo poškozené a které prodejna zlikviduje v kompostéru.  |
| Poškozeno vnějšími vlivy              | Odpis zboží, které bylo poškozeno nebo zničeno vlivem třetí strany (výbuch, vytopení, poškození majetku vloupáním) nebo přírodními živly. Zboží se tedy na prodejně nenachází a nemůže proto být zlikvidováno.                           |
| Zničeno rozmražením                   |  |



## Inventory

Pojem inventory, který lze do češtiny přeložit jako inventář, sdružuje všechny shrinky týkající se změn v inventáři, tj. stavu zásob či inventory. V tabulce 3.2 se nachází přehled všech evidovaných typů.

Tabulka 3.2: Přehled jednotlivých typů shrinků z kategorie inventory.

| Název                                   | Popis   |
|---|---|
| Inventura+                              | Kladné připsání zboží během inventory.  |
| Inventura−                              | Záporné odepsání zboží během inventory.   |
| Inventura                               | Velká inventura skladu.   |
| Oprava inventory                        | Dodatečné opravy, které bylo třeba provést po dokončení velké inventory.  |
| Částečná inventura                      | Odpis, nebo naskladnění zboží při inventuře položek.  |
| Neuznané reklamace distribučního centra | Odpis zboží, které bylo fyzicky dodané distribučním centrem na prodejnu, ale prodejna jej vrátila, ale distribuční centrum vratku neuznalo. |
| Inventura                               | Starší verze ID používaného pro inventuru.  |
| Partes merchandise                      | Odpis prokazatelně ukradeného zboží nebo i ztraceného zboží.  |

## Price down



# Kapitola 4

## Zpracování dat

Úvod ke kapitole....

### 4.1 Popis obdržených dat

Všechna data poskytnutá společností jsou uložena v databázi, ke které je byl zhotoven omezený přístup pro účely získání dat pro analýzy shrinku produktů společnosti. Zároveň s možností přístupu jsem obdržela i tabulku, která stručně komentuje všechny tabulky v databázi a sloupce v jednotlivých tabulkách. Celkem se v databázi nachází přes čtyři sta tabulek, z nichž je potřeba vybrat ty, které obsahují relevantními data pro úlohu shrinků.

Z důvodu ochrany dat nelze uvádět přesné názvy tabulek, nicméně pro lepší orientaci v textu, každé použité tabulce přiřadím název, který odpovídá obsaženým datům v tabulce.

#### Produkty

Základní číselník s údaji o produktech, se nachází v tabulce **produkt** se 27 sloupci. Pro analýzu vzniku shrinků jsem z této tabulky vybrala jako možné významné údaje následující sloupce:

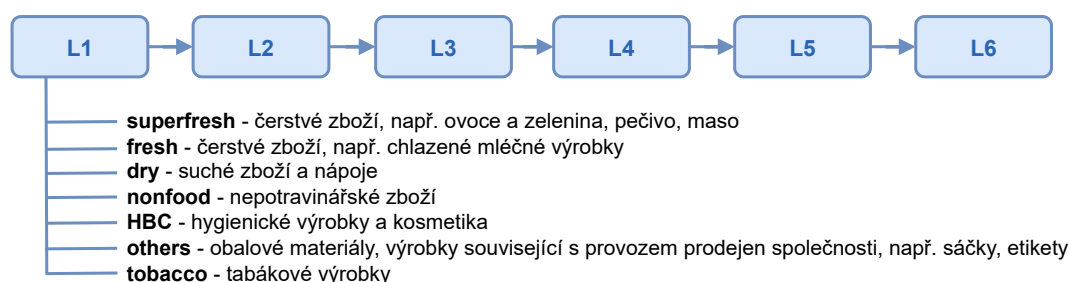
- ID produktu a ID prodejní varianty, spolu tvoří primární klíč.
- Krátký a dlouhý název produktu (tento sloupec slouží především k lepší orientaci v datech)
- Expirace produktu ve dnech (hodnoty 0, 999 a NULL označují neomezenou expiraci)
- Rozměry produktu a jeho hmotnost

Tabulka 4.1: Počet podkategorií v L1 kategorii a zastoupení L1 kategorie v rámci produktového portfolia.

| L1         | Počty kategorií |    |    |     |     | Zastoupení kategorie |
|------------|-----------------|----|----|-----|-----|----------------------|
|            | L2              | L3 | L4 | L5  | L6  |                      |
| nonfood    | 1               | 7  | 27 | 76  | 179 | 76,12%               |
| dry        | 3               | 13 | 33 | 147 | 494 | 7,28%                |
| HBC        | 1               | 4  | 21 | 59  | 193 | 7,07%                |
| fresh      | 5               | 11 | 27 | 111 | 469 | 4,27%                |
| superfresh | 6               | 10 | 31 | 92  | 271 | 4,04%                |
| others     | 4               | 4  | 4  | 5   | 5   | 1,06%                |
| tobacco    | 1               | 1  | 1  | 3   | 8   | 0,17%                |

- ID kategorie produktu v číselné struktuře (pro lepší interpretaci, o jakou kategorii zboží se jedná, je vhodnější použít strukturu 4BOX, kterou lze získat napojením na tabulku `produkt_kategorie`.)

Tabulka `produkt_kategorie` obsahuje převod z číselné struktury, která popisuje kategorii produktu na 4BOX strukturu. 4BOX struktura se skládá ze šesti úrovní. Struktura 4BOX je znázorněná na obrázku 4.1.



Obrázek 4.1: Hierarchie struktury 4BOX používané pro kategorizaci produktů.

V tabulce 4.1 jsou uvedeny počty podkategorií pro každou z kategorií L1 a také procentuální zastoupení kategorie L1 v rámci produktového portfolia vybrané společnosti. Zastoupení je odvozeno podle počtu produktů v kategorii.

Zároveň kategorie L6 je přímo napojená na hodnotu číselné struktury, která je uvedena v číselníku produktů (v tabulce `produkt`).

Tabulka, která obsahuje

## Tabulky transakcí

V tabulce `transakce` se nachází údaje o všech provedených transakcích, a to jak skladové transakce, tak prodeje na prodejnách. V případě prodejen jsou údaje agre-

gované podle prodejny, konkrétního produktu a dne transakce, tzn. v této tabulce nelze rozlišit konkrétní prodeje na jednotlivých pokladnách, ale pouze souhrn za jeden den. Tabulka obsahuje údaje za posledních dvanáct měsíců.

Tabulka transakcí obsahuje 21 sloupců, jako podstatné pro analýzu jsem vybrala následující sloupce:

- ID transakce, jedinečné pro každou transakci
- Typ transakce
- ID produktu, kterého se transakce týká
- ID skladu (včetně prodejen)
- Datum transakce (tzv. business datum, pokud samotná transakce proběhne až po půlnoci uvedeného dne, tak se posílá s datem z předchozího dne, neboť do toho dne businessově náleží.)
- ID promoce
- Motive\_type z hlediska hledání shrinku produktu se jedná o klíčový sloupec, neboť obsahuje označení jednotlivých typů shrinků.
- Množství produktu v dané transakci
- Hodnota transakce v nákladové ceně
- Hodnota transakce v prodejní ceně včetně DPH – v případě prodejů se jedná o skutečnou cenu, u zbylých transakcích je uvedena odpovídající cena podle ceníku.

Tabulku, která obsahuje údaje o jednotlivých prodejkách na prodejnách společnosti, jsem pro účely této práce nazvala **transakce\_prodeje**. Celkem obsahuje třináct sloupců, z nichž jsem vybrala následující:

- ID transakce
- ID produktu, kterého se transakce týká
- ID prodejny
- Datum a čas transakce
- ID promoce
- Množství produktu v dané transakci
- Hodnota transakce v prodejní ceně bez DPH
- Hodnota DPH dané transakce

## 4.2 Předzpracování dat

Množství produktů -

## 4.3 Statistické zpracování dat

### 4.3.1 Období jednoho roku

Sledovala jsem tři různé veličiny, kterými lze hodnotit transakce - počet záznamů (tj. počet řádků) v tabulce **transakce**, celkové množství produktů a celková nákladová cena produktů.

Z databáze jsem vybrala všechny transakce za rok 2022, u kterých byl evidován nějaký typ shrinku. Celkem se jednalo o záznamů. Data jsem agregovala po jednotlivých měsících a pro každý měsíc vypočítala pomocí nástroje pivottables.

typ shrinku v závislosti na expiraci typ shrinku v závislosti hlavní kategorii produktu  
typ shrinku v závislosti na dni v měsíci typ shrinku v závislosti na dni v týdnu

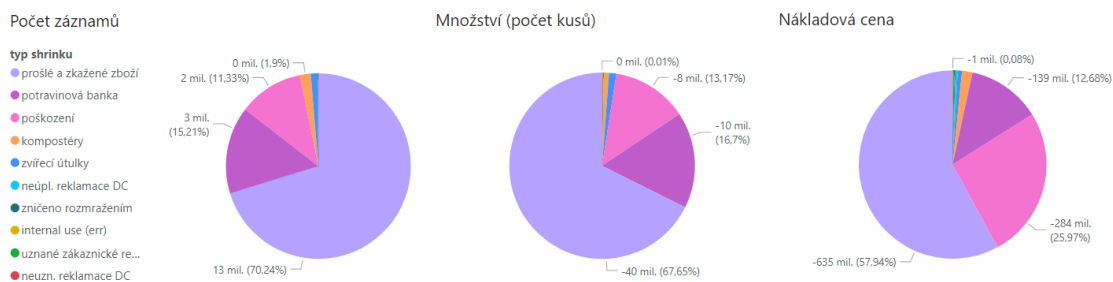
typ shrinku v závislosti na prodejně, neboli typu prodejny nebo v závislosti na centrálním skladu. Závislost mezi prodejnami a typem produktu

### Celkový přehled

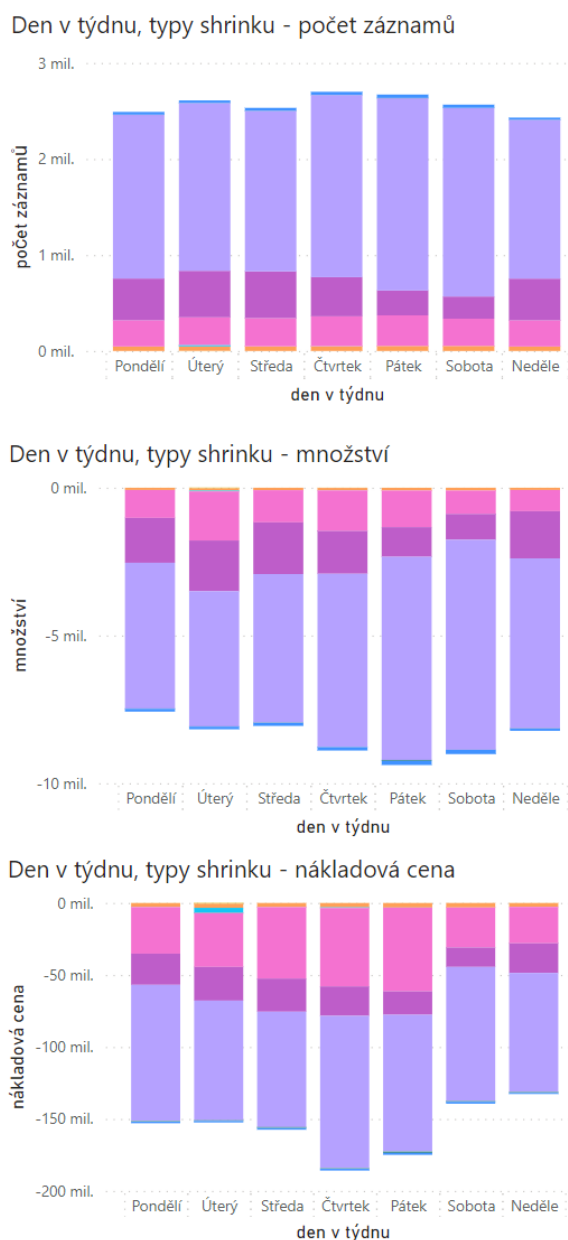
Nejprve jsem určila zastoupení jednotlivých shrinků během celého roku, a to bez uvažování závislosti na jiných faktorech. Poměr rozdělení lze vidět na obrázku 4.2. Největší zastoupení, z pohledu všech tří sledovaných charakteristik, má shrink označující prošlé a zkažené zboží. Z pohledu počtu záznamů činí 70,24 % ze všech shrinků typu damage, což odpovídá 13 milionům řádkům záznamů. V případě množství neprodaných kusů produktů bylo v roce 2022 odepsáno 40 milionů kusů (tj. 65,67 %). Podle nejdůležitějšího ukazatele - nákladové ceny - tento shrink měl za následek ztrátu 635 milionů Kč (tj. 57,94) %. Další damage shrinky, které mají zastoupení větší než 10 % je shrink označující poškozené zboží a shrink s produkty, které byly odevzdány potravinové bance. Zbylé shrinky mají v ročním pohledu malý počet výskytů a jejich výskyt bude analyzován v závislosti na dalších faktorech.

### Závislost typu shrinku na čase

Porovnála jsem množství zaznamenaných shrinků v závislosti na dnech v týdnu, porovnání lze vidět na obr. 4.3. Jednotlivé typy jsou zastoupeny analogicky jako v souhrnném přehledu shrinků za jeden rok, tj. prošlé zboží, zboží zaslané do potravinové banky a poškozené zboží. Počty záznamů pro všechny dny jsou v rozmezí 1,6 až 1,9 milionů záznamů za jeden rok.

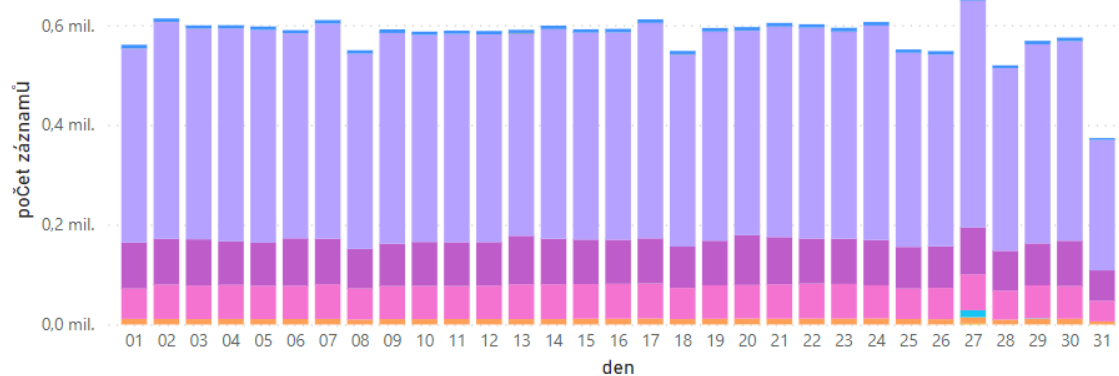


Obrázek 4.2: Zastoupení shrinků typu damage v roce 2022.

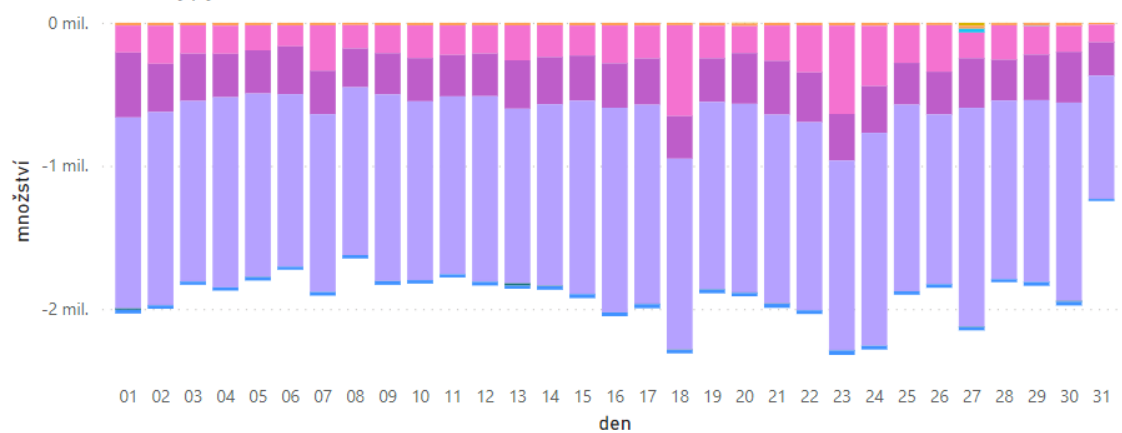


Obrázek 4.3: Zastoupení shrinků typu damage v závislosti na dni v týdnu (údaje pro rok 2022).

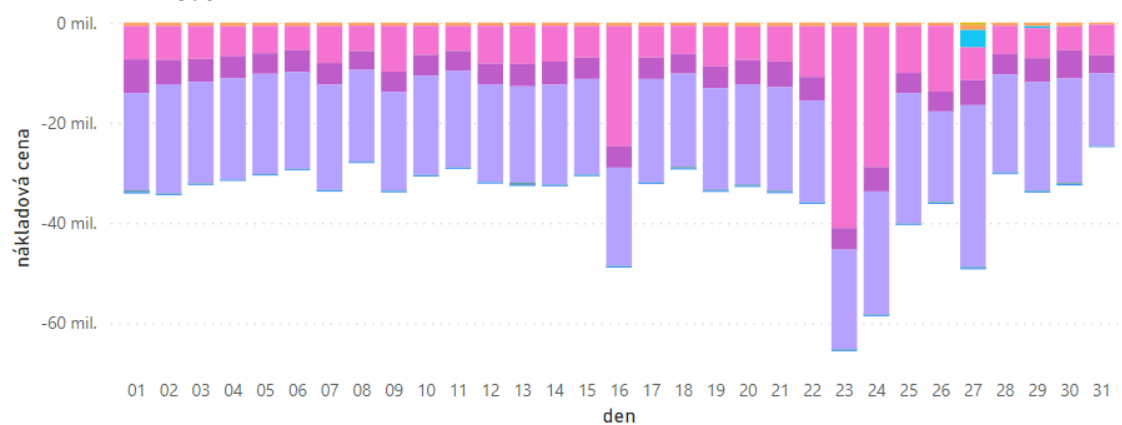
Den v měsíci, typy shrinku - počet záznamů



Den v měsíci, typy shrinku - množství



Den v měsíci, typy shrinku - nákladová cena



Obrázek 4.4: Zastoupení shrinků typu damage v závislosti na dni v měsíci (údaje pro rok 2022).



## Závislost typu shrinku na hlavní kategorii zboží

Přehled zastoupení shrinků v závislosti na kategorii zboží je znázorněn na obr. ?? Kategorie, u které bylo evidováno nejvíce damage shrinků je kategorie produktů *superfresh*. Celkově za rok 2022 ztracené náklady byly v hodnotě 0,6 miliardy korun. V rámci této kategorie je opět nejčastější příčinou odpisu zboží překročená doba expirace, u *superfresh* produktů spíše viditelné zkažení zboží, neboť část *superfresh* produktů nemá uvedenou dobu spotřeby. Druhý nejčastější shrink je shrink označující potravinovou banku. Zbylé shrinky jsou pro tuto kategorii již méně zastoupeny. Kategorie *superfresh* má největší zastoupení většiny typů shrinků nad ostatními kategoriemi - přehled zastoupení jednotlivých shrinků odděleně je na obr. ?? Druhá kategorie, která má nejvíce zaznamenaných shrinků, je kategorie *fresh food*

## Závislost typu shrinku na typu prodejny

### 4.3.2 Období jednoho měsíce

Vzhledem k vysokému počtu dat pro jeden kalendářní rok, jsem se rozhodla provést analýzu na měsíčním výběru dat z tohoto období. Jako zkoumaný měsíc jsem vybrala měsíc říjen, neboť v tomto měsíci bylo v databázi evidováno přes dva a půl milionu pozorování, což je nejvíce ze všech měsíců v roce 2022. Zároveň v porovnání s letními měsíci a Vánocemi se v říjnu nevyskytují významné sezónní výkyvy.

Zkoumaná říjnová data obsahují 2634614 řádků a patnáct sloupců. Každý řádek odpovídá jednomu záznamu v databázi shrinku daného produktu. Sledované údaje ve sloupcích jsou následovné,

- id prodejny, kategorická proměnná,
- id produktu, kategorická proměnná,
- datum transakce, kategorická proměnná,
- typ shrinku, kategorická proměnná,
- L1, kategorická proměnná,
- L2, kategorická proměnná,
- L4, kategorická proměnná,
- L5, kategorická proměnná,
- L6, kategorická proměnná,
- expirace, kategorická proměnná,
- množství, spojitá proměnná,

- ztracená nákladová cena, spojitá proměnná,
- den v týdnu, kategorická proměnná,
- číslo den, kategorická proměnná,
- období v měsíci (rozdělení měsíce na pět částí), kategorická proměnná,

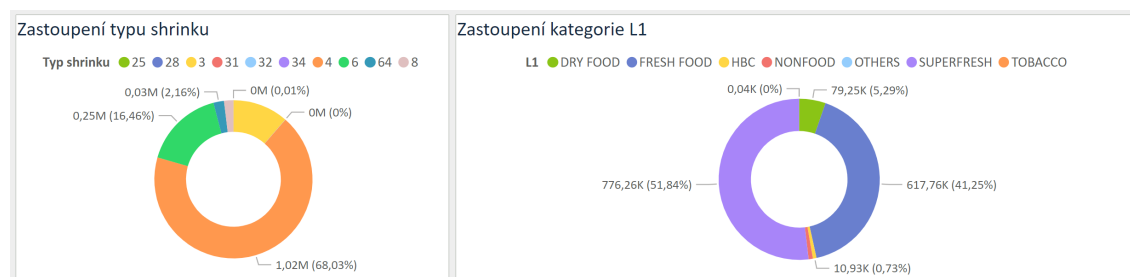
Původní sloupec datum jsem rozdělila na tři jiné proměnné, a to den v týdnu, číslo dne a období v měsíci a sloupec datum jsem vynechala. Z důvodu vysokého počtu záznamů a odlišné povahy dvou typů shrinků jsem data dále rozdělila na shrinky typu damages a shrinky typu inventory.

## Damages shrinky

Následující část text bude věnována rozboru dat pro shrinky typu damages.

### Výběr dat

Nejprve jsem graficky analyzovala zastoupení shrinků v závislosti na vybraných proměnných pomocí nástroje Power BI, viz obr. 4.5. V návaznosti na zjištěné zastoupení shrinků v datech jsem se rozhodla vybrat pouze ty typy shrinků, které tvoří více jak jedno procento z celkových nákladů (tj. náklady činily alespoň jeden milion korun). Vynechala jsem tedy shrinky s označením 25, 28, 31, 32, 34 a naopak shrinky 3,4,6,8,64 byly ponechány. Obdobně jsem přistupovala k záznamům i z hlediska kategorie produktu úrovně L1, jelikož z grafu je patrné, že majoritní zastoupení mají pouze dvě kategorie, a to kategorie superfresh a fresh produktů. Všechny záznamy se zbylými kategoriemi (HBC, others, nonfood, dry food a tobacco) jsem z datasetu odstranila. Těmito kroky jsem zredukovala původní počet řádků datasetu z 1497485 na 1393223 řádků.



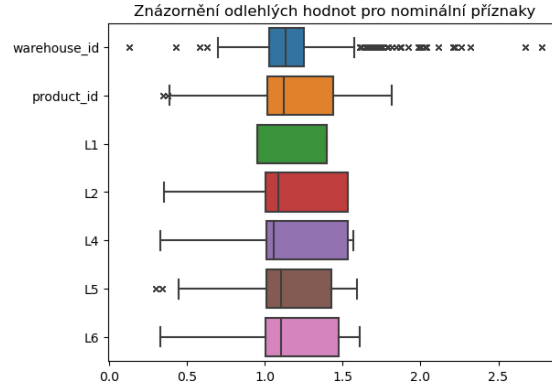
Obrázek 4.5: Zastoupení shrinků typu damage a zastoupení kategorie L1 v datech z října roku 2022.

Jako cílové sloupce (target sloupec) jsem určila sloupec s typem shrinku, množství produktu a nákladovou cenu. Zbylých jedenáct sloupců slouží jako vysvětlující proměnné, dále budou označovány jako příznaky, pro cílový sloupec. Všechny vybrané příznaky

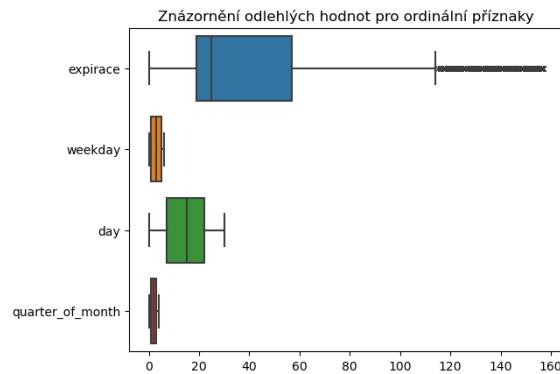
jsou kategorické proměnné, které lze dále rozdělit na nominální a ordinální. Nominální proměnné jsou id prodejny, id produktu, kategorie L1, L2, L4, L5 a L6. Ordinální expirace, den v týdnu, číslo dne a období měsíce.

Target hodnotu pro cílový sloupec typu shrinku jsem přeznačila na hodnoty od nuly do  $n_p$ , kde  $n_p$  je počet kategorií v  $p$ -tém příznaku. Pro další výpočty bylo vhodné přesunout se z nominálních kategorických hodnot na číselné hodnoty. Pro tyto účely jsem zvolila metodu *target encoding*

Dále jsem se zabývala identifikací odlehých hodnot. Nejprve jsem vizualizovala hodnoty pomocí grafu, obrázky 4.6 a 4.7.



Obrázek 4.6: Znázornění odlehých hodnot pro nominální příznaky.

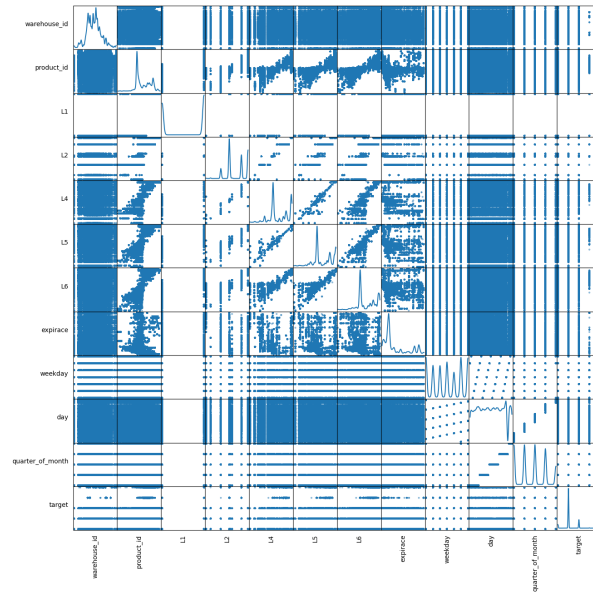


Obrázek 4.7: Znázornění odlehých hodnot pro ordinální příznaky.

Pomocí Tukeyho testu jsem identifikovala přes 150000 outlierů pro příznak id prodejny (*warehouse\_id*), čímž se dataset zredukoval na 1218453 řádků.

V dalším kroku jsem se zaměřila na míru korelace mezi proměnnými. Vizualizovala jsem data pomocí scatter matice pro všechny proměnné, matice je možné vidět na obr. č. 4.8.

Jako první jsem zvolila  $\chi^2$  test. Vzhledem k vysokému počtu dat je matice příliš řídká, a proto nejsou výsledné hodnoty odpovídající a test je tedy pro tuto úlohu nespolehlivý. Jiným měřítkem pro korelaci mezi proměnnými je Pearsonův korelační koeficient.



Obrázek 4.8: Scatter matice příznaků.

Výslednou matici popisující korelační vztahy mezi příznaky jsme vizualizovala teplovní mapou, která je zobrazena na obrázku 4.9. Z výsledků je patrné, že mezi jednotlivými kategoriemi produktů a produkty je silná korelace. Toto zjištění je zcela logické neboť jde jedná o stromovou strukturu kategorií. Zároveň existuje korelace mezi produktovými kategoriemi a expirací produktu.

Dále jsem použila výpočet koeficientů vzájemné informace <sup>1</sup>, která určuje kolik informace sdílí dvě proměnné. Matice vypočítaných koeficientů je na obr.4.10.

Opět pro znázornění závislosti mezi proměnnými jsem použila koeficient Cramers V a Thiels U.

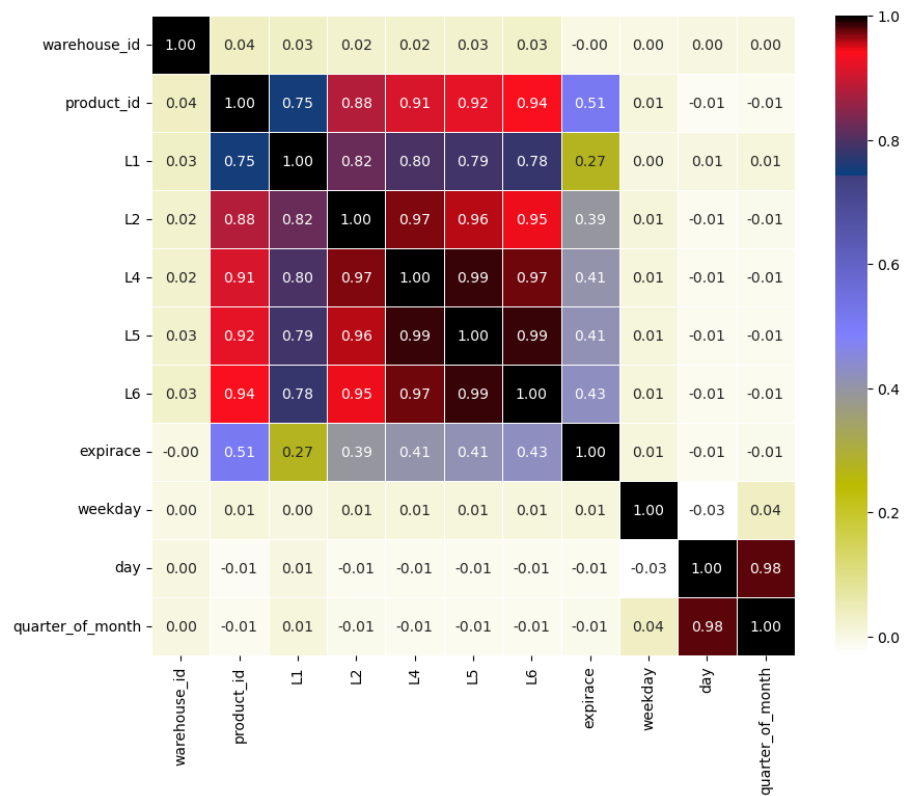
V dalším testu jsem otestovala multikolinearitu dat pomocí rozptylového inflačního faktoru (VIF). Jako hraniční faktor jsem zvolila hodnotu 40 VIF. Čímž došlo k redukci příznaků z jedenácti na pět na kategorii L1, číslo dne, období měsíce, id prodejny a den v týdnu.

Jako další metodu po výběr příznaků jsem vypočítala hodnotu koeficientů vzájemné informace mezi všemi příznaky s target sloupcem. Na obrázku 4.14 lze vidět jak hodnotu jednotlivé proměnné souvisí s cílovým sloupcem.

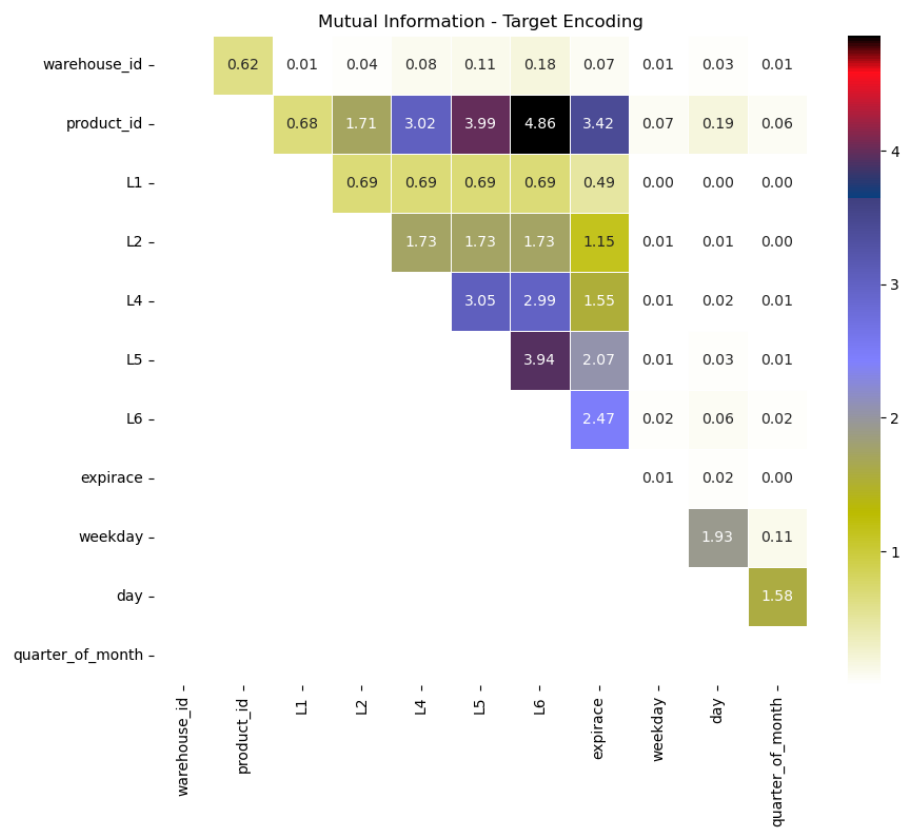
Potom jsem aplikovala na data analýzu hlavních komponent 4.15. Na obrázcích 4.15 a 4.16 jsou znázorněny prvních deset komponent a rozptyl který v datech vysvětlují. Na základě hodnot jsem vybrala prvních pět komponent, poté jsem vypočítala příspěvky příznaků k těmto komponentám a vybrala jsem ty příznaky, které přispívají nejvíce - to jsou příznaky id prodejny, den v týdnu, expirace, den a období v měsíci.

---

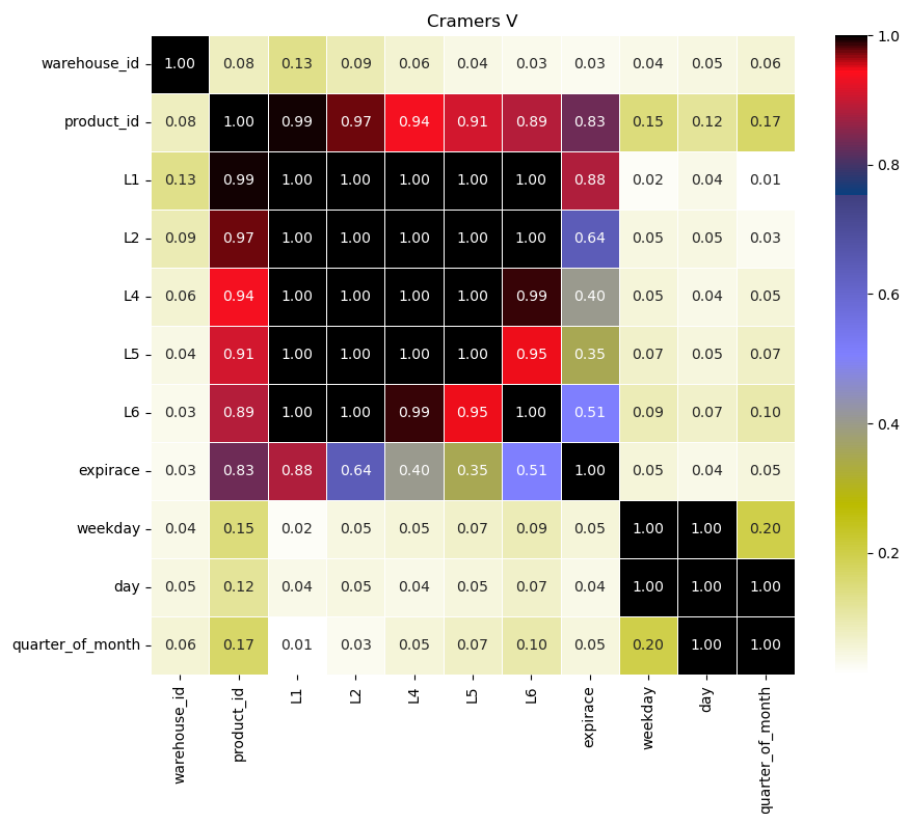
<sup>1</sup> *mutual information*



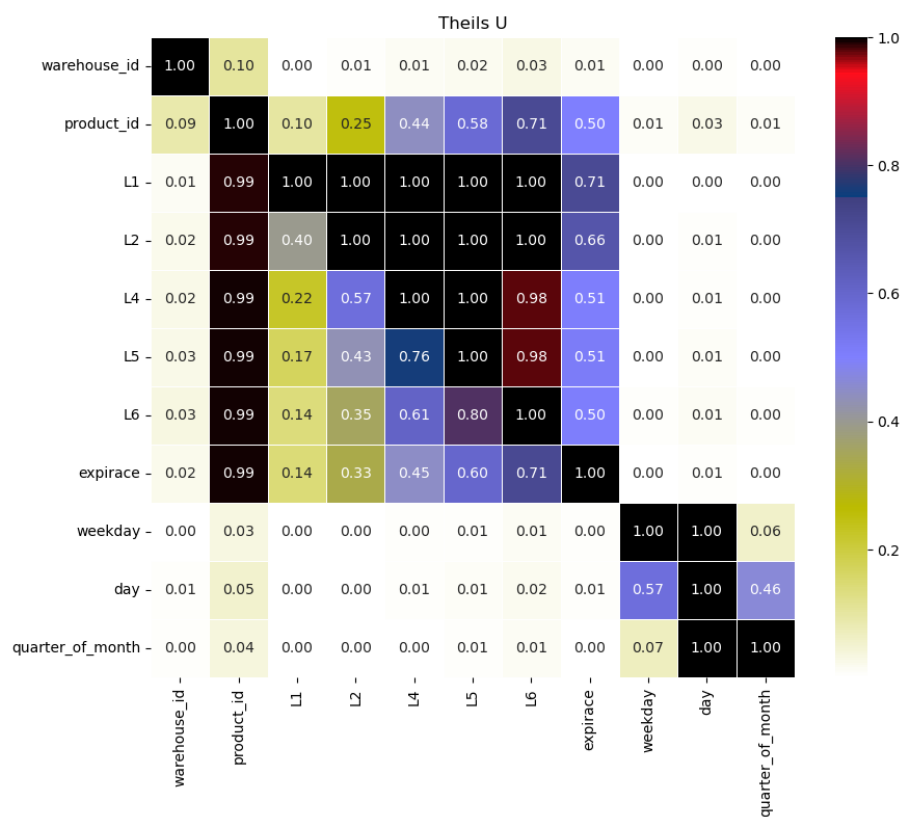
Obrázek 4.9: Matice korelačních koeficientů mezi příznaky.



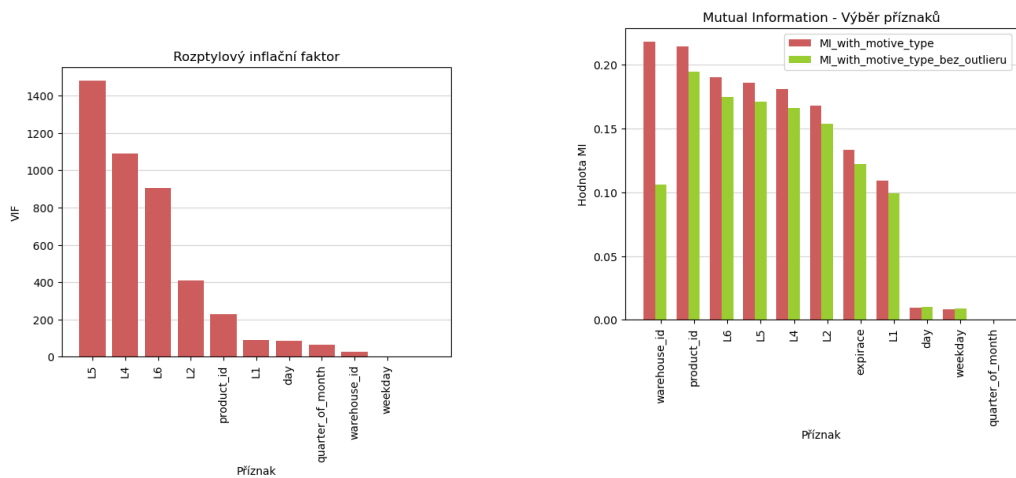
Obrázek 4.10: Matice koeficientů vzájemná informace mezi příznaky.



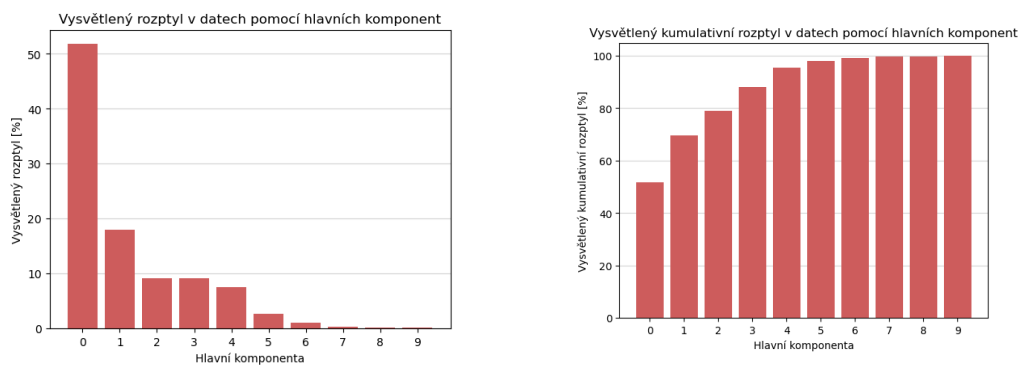
Obrázek 4.11: Matice koeficientů Cramers U mezi příznaky.



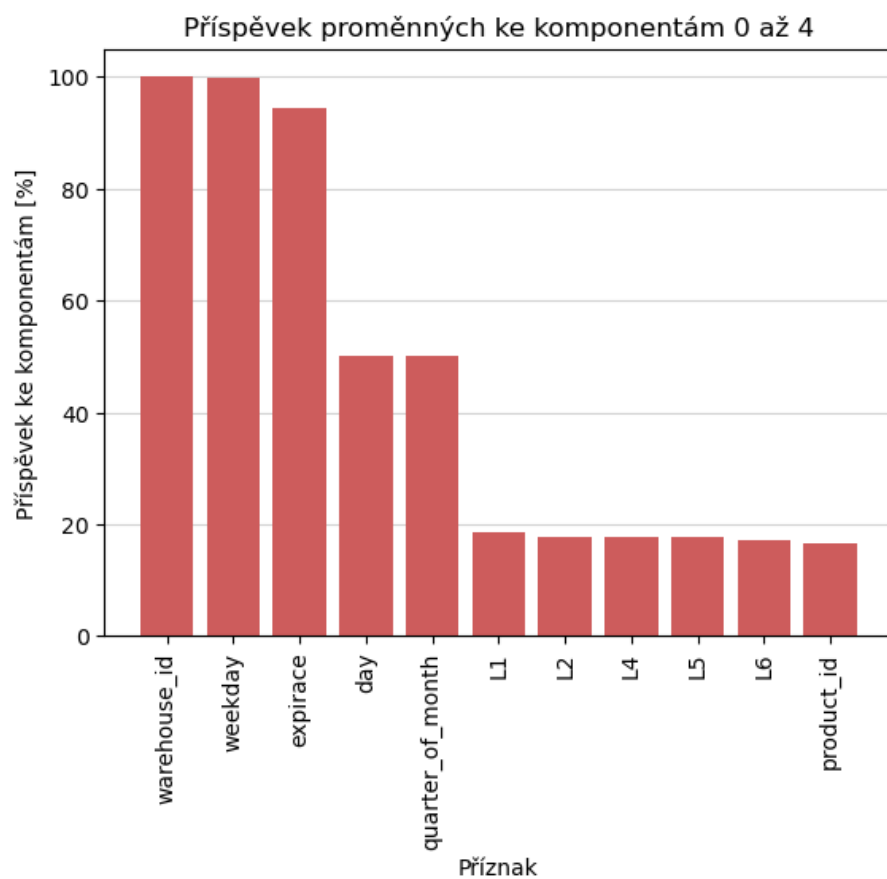
Obrázek 4.12: Matice koeficientů Thiels V mezi příznaky.



Obrázek 4.13: Rozptylový inflační faktor. Obrázek 4.14: Matice koeficientů vzájemná informace mezi příznaky a cílovým sloupcem typ shrinku.



Obrázek 4.15: PCA - vysvětlený rozptyl hlavních komponent. Obrázek 4.16: PCA - kumulativní vysvětlený rozptyl hlavních komponent.



Obrázek 4.17: Příspěvek proměnných ke komponentám 0 až 4.



# Závěr

Cílem práce bylo



# Literatura

- [1] BAUDIN, Michel. *Lean Logistics: The Nuts and Bolts of Delivering Materials and Goods*. New York: Productivity Press, 2005. ISBN 978-1563272967.
- [2] CHRISTOPHER, Martin. *Logistics & Supply Chain Management*. 5th ed. Harlow: Pearson Education Limited, 2016. ISBN 9781292083797.
- [3] What is the difference between Logistics and Supply Chain Management. In: *IIM Udaipur Chronicles* [online]. 11. 10. 2019. [cit. 2022-11-07] Dostupné z: <https://www.iimu.ac.in/blog/what-is-the-difference-between-logistics-and-supply-chain-management/>
- [4] JIRSÁK, Petr, MERVART, Michal, VINŠ, Marek. *Logistika pro ekonomy – vstupní logistika*. 1. vydání. Praha: Wolters Kluwer ČR, 2012.
- [5] JONES, Daniel T., Peter HINES a Nick RICH. Lean logistics. *International Journal of Physical Distribution & Logistics Management*. 1997, **27**(3/4), 153-173. ISSN 0960-0035. Dostupné z: doi:10.1108/09600039710170557
- [6] KURITA, Takio. Principal component analysis (PCA). Computer Vision: A Reference Guide, 2019, 1-4. [cit. 2022-11-07] Dostupné z: [https://link.springer.com/content/pdf/10.1007/978-3-030-03243-2\\_649-1.pdf](https://link.springer.com/content/pdf/10.1007/978-3-030-03243-2_649-1.pdf)
- [7] Analýza hlavních komponent. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2022 [cit. 2023-03-04]. Dostupné z: [https://cs.wikipedia.org/wiki/Analýza\\_hlavních\\_komponent](https://cs.wikipedia.org/wiki/Analýza_hlavních_komponent)
- [8] DI FRANCO, Giovanni. Multiple correspondence analysis: one only or several techniques?. *Quality & Quantity*, 2016, 50.3: 1299-1315. [cit. 2023-03-05]. Dostupné z: doi:10.1007/s11135-015-0206-0
- [9] ABDI, Hervé, VALENTIN, Dominique. Multiple correspondence analysis. *Encyclopedia of measurement and statistics*, 2007, 2.4: 651-657. [cit. 2023-03-05]. Dostupné z: <https://personal.utdallas.edu/~Herve/Abdi-MCA2007-pretty.pdf>
- [10] GREENACRE, Michael. Correspondence analysis in practice. Chapman and Hall/CRC, 2017. [cit. 2023-03-05].
- [11] Correspondence analysis. In *Wikiwand* [online], [cit. 2023-03-06]. Available from: [https://www.wikiwand.com/en/Correspondence\\_analysis](https://www.wikiwand.com/en/Correspondence_analysis)

- [12] WRONKA, Anna. LEAN LOGISTICS. *Journal of Positive Management*. 2017, 7(2), 55-63. ISSN 2392-1412. Dostupné z: doi:10.12775/JPM.2016.012
- [13] SUTHERLAND Joel, BENNETT Bob. *The Seven Deadly Wastes of Logistics: Applying Toyota Production System Principles to Create Logistics Value*. Bethlehem, PA: Lehigh University, 2007. Dostupné z: <https://www.researchgate.net/publication/265356600>
- [14] SKHMOT, Nawras. The Lean Way Blog: The 8 Wastes of Lean. *The Lean Way* [online]. 5. 8. 2017 [cit. 2022-11-17]. Dostupné z: <https://theleanway.net/The-8-Wastes-of-Lean>
- [15] SKHMOT, Nawras. The Lean Way Blog: What is Lean?. *The Lean Way* [online]. 5. 8. 2017 [cit. 2022-11-17]. Dostupné z: <https://theleanway.net/what-is-lean>
- [16] SKHMOT, Nawras. The Lean Way Blog: What is Muda, Mura, and Muri?. *The Lean Way* [online]. 5. 8. 2017 [cit. 2022-11-17]. Dostupné z: <https://theleanway.net/muda-mura-muri>
- [17] Learning the Lingo: 3 definitions related to unsold food inventory. In: *Blog - Spoiler Alert* [online]. 17. 06. 2019. [cit. 2022-02-07] Dostupné z: <https://blog.spoileralert.com/3-definitions-unsold-food-inventory>
- [18] HALFORD, M. Prince., *Prince* [online]. [cit. 2023-03-08]. Dostupné z: <https://github.com/MaxHalford/prince>

# Příloha A

## Obsah přiloženého CD

BP\_Gruberova.pdf – soubor s elektronickou verzí této bakalářské práce.

PreprocessData – složka se zdrojovými kódy naimplementovaného balíku  
PreprocessData.jl.