

**ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE**

**FAKULTA ŽIVOTNÍHO PROSTŘEDÍ**

**BAKALÁŘSKÁ PRÁCE**

2018

Irina Georgievová

**ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE**

**FAKULTA ŽIVOTNÍHO PROSTŘEDÍ**

**KATEDRA VODNÍHO HOSPODÁŘSTVÍ  
A ENVIRONMENTÁLNÍHO MODELOVÁNÍ**

Vizualizace enviromentálních dat

**BAKALÁŘSKÁ PRÁCE**

Vedoucí práce: **doc. Ing. Martin Hanel, Ph.D.**

Bakalant: **Irina Georgievová**

2018

# ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Fakulta životního prostředí

## ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Irina Georgievová

Vodní hospodářství

Název práce

**Vizualizace environmentálních dat**

Název anglicky

**Visualization of environmental data**

---

### Cíle práce

Představení klíčových poznatků týkajících se vizualizace a průzkumové analýzy dat z teoretického hlediska i z hlediska praktické implementace v R. Zhodnoceny budou jak nástroje obsažené v základní distribuci R, tak nástroje dostupné v balících lattice, grid, ggplot2, raster, rasterVis, případně i nástroje pro tvorbu dynamických vizualizací (htmlwidgets, shiny apod.).

Součástí práce bude i vytvoření webové aplikace pro vizualizaci a analýzu hydrologické bilance a předpověď sucha v útvarech povrchových vod ČR.

### Metodika

Teoretická část:

- rešerše základních poznatků o vizualizaci dat
- popis vizualizačních prostředků se zaměřením na využití v hydrologii, porovnání výhod/nevýhod
- popis nejpoužívanějších R balíků, jejich základních funkcí a demonstrace jejich využití

Praktická část:

- tvorba aplikace s využitím Shiny dle průběžné specifikace
- stručný popis aplikace v BP

**Doporučený rozsah práce**

40-60 stran

**Klíčová slova**

vizualizace dat, grammar of graphics, průzkumová analýza dat

---

**Doporučené zdroje informací**

CLEVELAND, W S. *The elements of graphing data*. Murray Hill: AT&T Bell Laboratories, 1994. ISBN 0-9634884-1-4.

TUFTE, E.R. *The Visual Display of Quantitative Information*. Graphics Press, 1983. ISBN 978-0-961-39210-9.

TUKEY, J.W. *Exploratory Data Analysis*. Addison-Wesley, 1977. ISBN 0201076160

WICKHAM, H. *Ggplot2 : elegant graphics for data analysis*. Dordrecht: Springer, 2009. ISBN 978-0-387-98140-6.

WILKINSON, L. *The Grammar of Graphics*. Springer, 2006. ISBN 978-0-387-28695-2.

---

**Předběžný termín obhajoby**

2017/18 LS – FŽP

**Vedoucí práce**

doc. Ing. Martin Hanel, Ph.D.

**Garantující pracoviště**

Katedra vodního hospodářství a environmentálního modelování

---

Elektronicky schváleno dne 8. 3. 2018

**doc. Ing. Martin Hanel, Ph.D.**

Vedoucí katedry

---

Elektronicky schváleno dne 9. 3. 2018

**prof. RNDr. Vladimír Bejček, CSc.**

Děkan

V Praze dne 09. 03. 2018

**Prohlášení:**

Prohlašuji, že jsem bakalářskou práci *Vizualizace enviromentálních dat* zpracovala samostatně. Veškerou literaturu a další podkladové materiály uvádím v seznamu na straně

V Praze dne ..... ....

Irina Georgievová

**Poděkování:**

# Abstrakt

Vložte abstrakt o rozsahu cca 100–200 slov. K problému vícenásobné marginalizace (VM) dochází, pokud články dodavatelského řetězce nastavují cenu svého výstupu způsobem, který by optimalizoval zisk v podmínkách prodávajícího na trhu s koncovým zbožím. V takové situaci dochází k cenové spirále a výsledná cena koncové produkce svou přílišnou výši poškozuje jak spotřebitelský užitek, tak i zisk řetězce jako celku. Kvantifikace dopadu VM byla již předmětem našeho dřívějšího výzkumu. Tento příspěvek se zabývá otázkou, za jakých podmínek k cenové spirále dochází a jakým způsobem probíhá konvergence k výsledným (rovnovážným) cenám. Pro tyto účely byl navržen a implementován model řetězce v podobě multiagentního systému, se kterým je možné analyzovat celý proces pomocí počítačové simulace.

**Klíčová slova:** vizualizace dat, grammar of graphics, průzkumová analýza dat

# Abstract

Double (or multiple) marginalization is often identified as the main source of a decentralized supply chain's (SC's) inefficiency. In its core lies the fact that if the agents constituting the SC choose their output prices according to the golden rule of profit maximization (which normally applies to a single firm that produces independently and sells directly to the end consumer), the prices in the SC tend to spiral up to an inefficient level (equilibrium prices) where both the consumer surplus and the SC's total profit are diminished. The level of equilibrium prices and their impact on the SC's profit and efficiency had been studied in our earlier works. Our focus in this paper was the properties of the process of convergence of the prices inside the SC to equilibrium levels. The analysis was carried out using computer experiments with an agent-based simulation model of a SC with limited information. Only serial chain structure was considered.

**Keywords:** Data visualization, grammar of graphics, exploratory data analysis

# Obsah

<b>Úvod</b>	<b>10</b>
<b>Teoretická část</b>	<b>11</b>
1 Vizualizace dat . . . . .	11
1.1 Historie vizualizace dat . . . . .	11
1.2 Zásady vizualizace dat . . . . .	14
1.2.1 Edward Tufte . . . . .	14
1.2.2 Wiliam S. Cleveland . . . . .	16
1.3 Grammar of graphics . . . . .	19
2 Základní grafy v R . . . . .	20
2.1 Bodový graf . . . . .	20
2.2 Liniový graf . . . . .	20
2.3 Vykreslení rozdělení v R . . . . .	21
2.3.1 Q-Q graf a P-P graf . . . . .	23
2.3.2 Krabicový graf . . . . .	24
2.4 Sloupcový graf . . . . .	25
2.4.1 Histogram . . . . .	25
2.4.2 Koláčový graf . . . . .	27
2.4.3 Číslicový histogram ( <i>stem-and-leaf</i> ) . . . . .	28
3 Průzkumová analýza dat . . . . .	29
3.1 Odlehlá pozorování . . . . .	30
3.1.1 <i>Jackknife</i> . . . . .	30
3.1.2 Mahalanobisovy vzdálenosti . . . . .	31
3.1.3 Leverages . . . . .	32
3.2 Náhrada chybějících pozorování . . . . .	33
3.3 Transformace dat . . . . .	33
3.4 Ověřování normality . . . . .	34
4 Pokročilá vizualizace v R . . . . .	36
4.1 Balíčky pro vizualizaci dat . . . . .	36
4.1.1 <i>grid</i> . . . . .	36
4.1.2 <i>lattice</i> . . . . .	36
4.1.3 <i>ggplot2</i> . . . . .	36
4.2 Balíčky pro interaktivní vizualizaci dat ( <i>htmlwidgets</i> ) . . . . .	36
4.2.1 <i>plotly</i> . . . . .	36
4.2.2 <i>dygraphs</i> . . . . .	37
4.2.3 <i>leaflet</i> . . . . .	37
4.2.4 <i>ggvis</i> . . . . .	37
4.2.5 <i>rgl</i> . . . . .	37
4.3 Balíčky pro prostorová data . . . . .	37
4.3.1 <i>ggmap</i> . . . . .	37
4.4 . . . . .	37
4.4.1 <i>raster</i> . . . . .	37
4.4.2 <i>rasterVis</i> . . . . .	37

4.5 Balíčky pro webové aplikace . . . . .	37
4.5.1 shiny . . . . .	37
4.5.2 flexdashboard . . . . .	37
4.5.3 dashboard . . . . .	37
<b>Praktická část</b>	<b>37</b>
<b>Seznam obrázků</b>	<b>38</b>
<b>Seznam tabulek</b>	<b>38</b>
<b>Literatura</b>	<b>39</b>

# Úvod

Vizualizace dat vždy hrála a neustále hraje významnou roli ve vědě. Je to jeden z nejlepších a jednoduchých způsobů pochopení dat. Vizualizace poskytuje jasnou představu o konfiguraci dat, odhaluje skryté struktury v datech a shrnuje informace. Proces vizualizaci je nedílnou součásti mnoha analýz a téměř všechny vědy využívají grafického zobrazení dat k vizualizaci a komunikaci svých výsledků. Sbírané a analyzované po dobu mnoha let data se v současné době převádějí do grafické formy. Masivní příliv dat a jejich dostupnost vedly k novým metodám a novým přístupům. Kombinace programovacích dovedností, matematických a statistických znalostí a odborných znalostí týkajících se obsahu přijala název "*Data Science*". Objevily se pozice takzvaných "*information designers*", které vyvíjí vlastní softwary pro vizualizaci dat, zakládají poradenské firmy, pořádají globální workshopy nebo vytvářejí blogy s tisíci registrovanými uživateli. [1] Přes všechny výhody vizualizace, jedná se pouze o nástroj datové analýzy, obecně dostupný každému. Nesprávné či nevhodné použití tohoto nástrojů vede k tomu, že existují grafy, které se považují za moc barevné a rušivý, postrádající smysl až zavádějící. Z tohoto důvodu se obracíme na takzvané zásady vizualizace.

Táto bakalářská práce se zabývá shrnutím klíčových poznatků o vizualizaci a průzkumové analýze dat a to jak z teoretického hlediska, tak i z hlediska praktické implementace v programovacím jazyku R. Budou popsány zásady vizualizace, její zařazení do datové analýzy, moderní způsoby vizualizace (současně používané balíčky v R, interaktivní grafy). Součástí práce je také webová aplikace pro vizualizaci a analýzu hydrologické bilance a předpověď sucha v útvarech povrchových vod ČR.

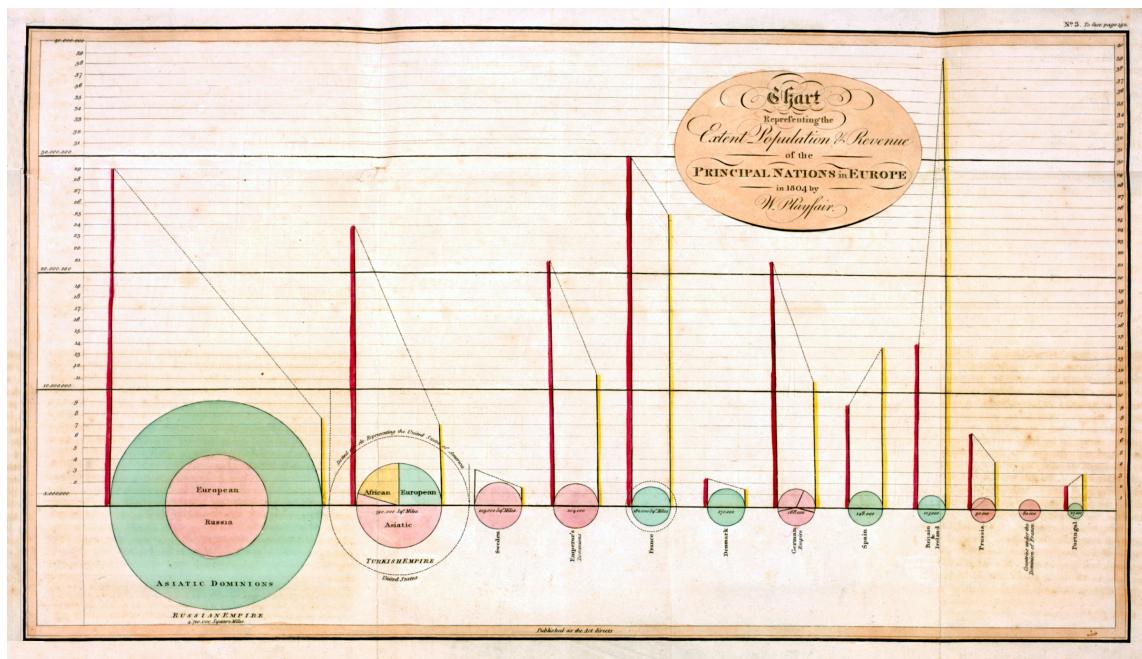
# Teoretická část

## 1 Vizualizace dat

### 1.1 Historie vizualizace dat

Před 17. stoletím jediné co by se dalo klasifikovat jako vizualizaci dat byly mapy pro navigaci a průzkum, ale také diagramy, geometrická schémata a tabulky pozic hvězd a jiných nebeských těles. Postupný vývoj statistické teorie a růst zájmu o data na konci 18. století vedly k inovacím a expanzi nových grafických forem. Kartografové se pokoušeli zaznamenat více, než pouhou geografickou polohu na mapě a objevili se první pokusy o tematické mapování geologických, ekonomických a medicínských dat.

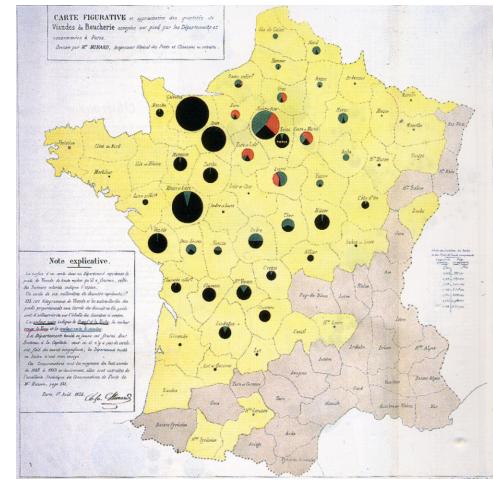
Wilam Playfair (1759-1823) je obecně znám jako průkopník v oblasti vizualizace dat a je považován za vynálezce několika typů grafů. Například liniový a sloupový grafy a grafy časových řád byly popsány v jeho práci z roku 1786 *"Commercial and Political Atlas"*<sup>1</sup>. Později popsal i koláčový graf ve své práci *"Statistical Breviary"* v roce 1801. Obrázek 1 ukazuje příklad jeho kreativní kombinace různých vizualizačních technik (kruhy, koláče, linie), pomocí které se snažil porovnat daňovou zátěž mezi Británií a dalšími zeměmi. Na tomto grafu také ukázal možnost použítí více měřítek pro různé ukazatele (v grafu populace a daně).



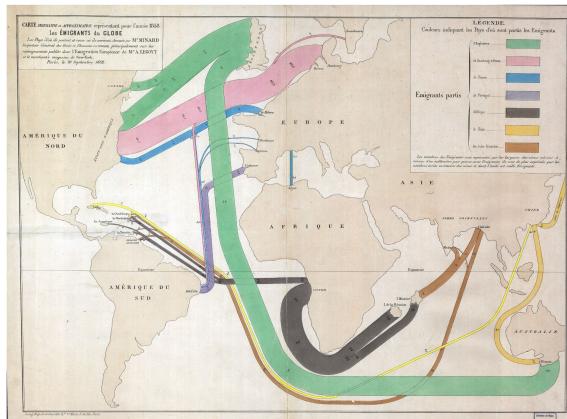
Obrázek 1: Kombinace různých využitelných technik, Playfair 1801

<sup>1</sup> "Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century"

V polovině 19. století byly vytvořeny všechny podmínky pro rychlý růst vizualizace. V důsledku rostoucí významnosti číselných informací pro sociální plánovaní, industrializaci, obchod a dopravu, byli zřízeny oficiální statistické úřady po celé Evropě. Vývoj statistické teorie, iniciovaný Gaussem a Laplaceem, měl odezvu ve společnosti a poskytl prostředky ke zpracování velkého množství dat. Pro vizualizaci se stalo dat období 1850-1900 "Zlatý věkem", s jedinečnou krásou a velkým množstvím inovací. S těmito inovacemi je hlavně spojené jméno Charlese Josepha Minarda (1781-1870). Například, Minardem bylo zavedeno použití koláčových grafů s výsečemi na mapách (obrázek 2), kde velikost koláčového grafu ukazuje sumu za oblast nebo každý grafický region na mapě a výšeče reprezentují dílčí součty za jednotlivé kategorie. Dále se také zabýval znázorněním geografických pohybu a dopravy lidí, zboží, importu a exportu úměrně jejich velikostí. Tento typ vizualizace se nazývá

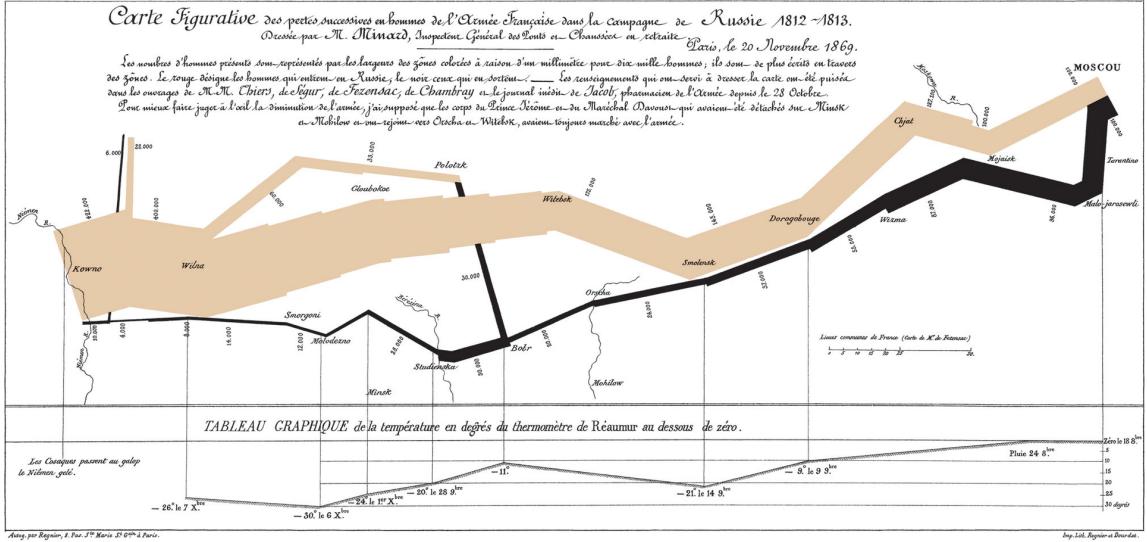


Obrázek 2: Dobytka odeslaná z celé Francie ke spotřebě v Paříži, Minard 1858



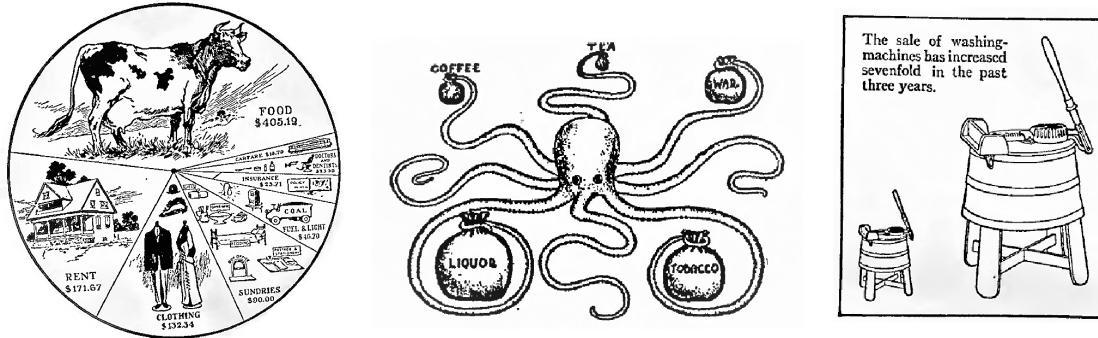
Obrázek 3: Mapa světové migrace, Minard 1858

"flow maps", viz obrázek 3. Jednou z nejslavnějších jeho práci je zobrazení postupných ztrát mužů francouzské armády během Napoleonského tažení na Moskvu v letech 1812-1813 (obrázek 4). Je považovaná za nejlepší informativní vizualizaci. I přestože v tomto grafu je celkem 6 proměnných (množství, lokace ve dvou rozměrech, postup armády, teplota, datum a skupiny), podařilo vše zobrazit tak, aniž by graf byl přeplněný a matoucí.



Obrázek 4: Postup Napoleonských vojsk v letech 1812-13, Minard 1869

Začátek 20. století je občas nazýván "moderním temným věkem" vizualizace. V letech 1900-1950 bylo jen málo grafických inovací. Nadšení pro vizualizaci, které charakterizovalo 19. století bylo nahrazeno formálními (z velké části statistickými) grafy a modely z oblasti sociologie. Hlavní zájem byl o přesná čísla, odhady parametrů, směrodatné odchylky. Vizualizace byly považované za pouhé hezké obrázky bez schopnosti podat přesná data. [2] Ve své práci *"Graphic Methods for Presenting Facts"* z roku 1919 Willord C. Brinton [1880-1957] kritizoval a vysvětloval chyby takovýchto grafů. Například koláčový graf rozdělení rodinných příjmů (od 900\$ do 1000\$) na obrázku 5. Tento graf je příkladem nepovedené vizualizace: oko preferenčně soudí dle velikostí obrázků a ne dle uhlů výsečí. Obrázek uprostřed znázorňuje druhý utracení: je to zábavný způsob vizualizace, avšak nelze přesně určit velikost brašen, ani je porovnat mezi sebou. Další obrázek by měl čtenáři sdělit informaci, že prodej praček za poslední tří roky vzrostl sedmkrát. Z obrázku není patrný poměr sedmi ku jedné ani přesné roky kdy bylo provedeno porovnání údajů. Dále Brinton ve své práci upozorňoval, že neúspěšná prezentace dat může vést k chybám závěrům a také zmiňoval potřebu jakéhosi standardu, souhrnu "gramatických pravidel pro grafický jazyk". [3]



Obrázek 5: Ukázky vizualizaci ze začatku 20. století, Brinton 1919

Ke "znovuzrození" vizualizace došlo v polovině šedesátých let 20. století, po napsání Johnem W. Tukey [1915-2000] článku "*The Future of Data Analysis*", ve kterém vyzývá společnost k uznání analýzy dat jako samostatného oboru statistiky odlišného od matematické statistiky. [4] Brzy poté začal Tukey s vývojem široké řady nových a efektivních grafů pod společným tématem "průzkumové analýzy dat" (popsány v jeho práci "*Explanatory Data Analysis*" z roku 1977, viz o tématu kapitola 3). [5] Mezi těmito novými grafy jsou například číslicový histogram (popsaný v kapitole 2.4.3), boxplot nebo krabicový graf (popsaný v kapitole 2.3.2) a další. Mnoho z nich je aktivně používáno ve statistické praxi a implementováno do většiny softwarů. [2]

## 1.2 Zásady vizualizace dat

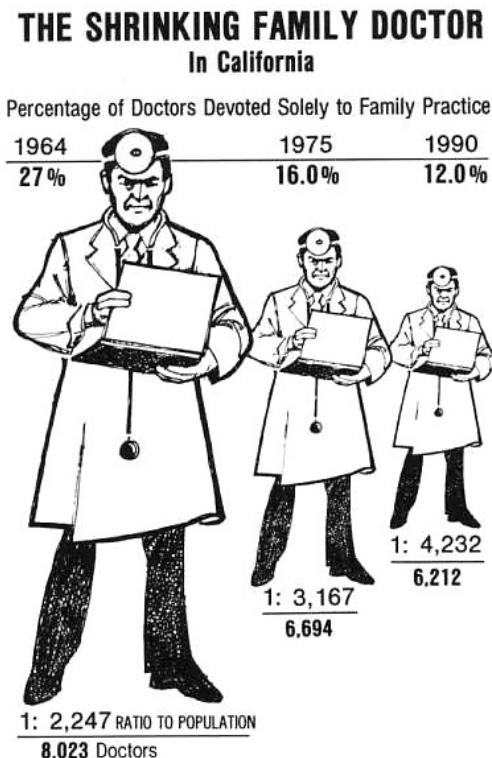
Od roku 1975 se vyvíjí statistické výpočetní systémy a s nimi i nové metody analýzy a vizualizace dat. V tomto období vizualizace začala být vnímána jako vlastní odvětví a to především díky Williamu S. Clevelandu a Edwardu Tufte, kteří položili vědecké základy tohoto odvětví. Tufte vyvinul a popularizoval terminologii a základní principy grafické integrity. Cleveland se zabýval studii grafického vnímání, kognitivních procesů, které lidi používají k pochopení grafů, a rozvíjel teorii o správném provedení vizualizaci. [6] Důsledek jejich práce se promítá i do současné doby kvalitní, interaktivní a dynamickou vizualizaci. [2]

### 1.2.1 Edward Tufte

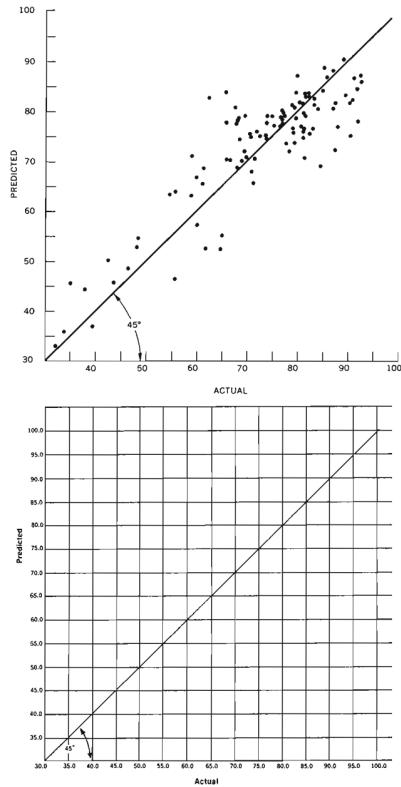
Za revoluční průlom se považuje kniha Edwarda Tufte *The Visual Display of Quantitative Information* z roku 1983, v kombinaci s dvěma následně publikovanými pracemi *Envisioning Information* z roku 1990 a *Visual Explanations* z roku 1997, patří mezi nejznámější publikace na téma vizualizace dat. Právě v těchto publikacích Tufte originálním způsobem definuje "standard" vizualizace. [1] Ideální způsob vizualizace dle Tufte je stručný, elegantní a informativní. Příkladem ideálního grafu je pro Tufte graf postupu Napoleonských vojsk v letech 1812-13, vytvořený Minardem (viz obrázek 4). Tufte říká, že grafická elegance se často nachází v jednoduchosti návrhu

a komplexnosti dat. [7] Tafte formuluje základní principy vizualizace jako grafickou dokonalost a grafickou integritu.

- **Grafická dokonalost** - grafika by měla:
  - být o datech a během jejich předvedení by nemělo dojít ke zkreslení
  - vyvolávat otázky o datech, ne o metodologii a technikách vizualizace
  - ukazovat velké množství dat v malém prostoru
  - předvádět velké datasety souvisle a logicky promyšleně
  - sloužit rozumnému a jasnemu cíli (popisu, průzkumu, ...)
  - být jednotná se statistickým nebo slovním popisem datasetu
- **Grafická integrita** neboli grafická celistvost a jednoznačnost
  - reprezentace čísel, zobrazené v grafu by měli být přímo úměrné číselným veličinám datasetu
  - jasné, detailní a svědomité označení v grafech by mělo potlačit zkreslení, nejasnost a dvojznačnost, popisky jsou důležitá
  - ukazovat variaci dat, nikoliv designu
  - v případě časových řad, představujících peníze, používat obecně známé jednotky
  - počet rozměrů představených v grafu by neměl přesahovat počet proměnných datasetu
  - reprezentace by neměla zahrnovat neúmyslný kontext



Obrázek 6: Zmenšující se rodinný lékař,  
Los Angeles Times, 1979



Obrázek 7: Vztah skutečné míry volební registrace k předpovídáným hodnotám, přetištěně E. Tufte, 1983

Ve spojení s těmito principy byly zavedeny Edwardem Tuftem následující terminy:

- **Lie factor** je definován jako poměr velikosti efektu zobrazeného v grafu oproti velikosti efektu v datech. Pokud se rovná jedničce, považuje se reprezentované hodnoty za přesné. Pokud je faktor větší než 1.05 či menší než 0.95, indikuje se podstatné zkreslení, přesahující míru drobných nepřesnosti vyskytujících se při vykreslování grafů. Tafte ve své práci uvádí jako jeden z příkladů graf na obrázku 6. Tento graf zobrazuje zmenšující se procento lékařů věnujících se výhradně rodinné praxi má *lie factor* odpovídající hodnotě 2.8, tedy skutečný pokles je značně nadhodnocen.
- **Data ink ratio** - poměr, který vyhodnocuje hustotu grafu a obsah informací. Dal by se vyjádřit vzorcem

$$\text{Data ink ratio} = \frac{\text{data-ink}}{\text{celkový inkoust použitý v datech}},$$

kde *data-ink* je nezbytné jádro grafu a smazání jakékoliv jeho části znamená ztrátu informaci. Tento vztah také odpovídá podílu grafického inkoustu požitého k vykreslení nepodstatných informací. Dalo by se to také vyjádřit jako jedna mínus *podíl grafiky, která může být vymazána bez ztráty informací*. Tafte doporučuje tento faktor maximalizovat v rozumných mezích, nejlépe se vyhnout těžkým mřížkovým liniím na pozadí (dokonce i horizontálním referenčním liniím). V příkladu na obrázku 7 jsou zobrazené dvě verze stejného grafu. Horní má hodnotu *data ink ratio* kolem 0.7, dolní graf ale neobsahuje informaci o datech, pouze nápomocné čáry, proto *data ink ratio* se rovná nule.

- **Chartjunk** - se vztahuje ke všem vizuálním elementům, které neslouží ke komunikaci informací zobrazených v grafu nebo odvádějí pozornost od těchto informací. [8]

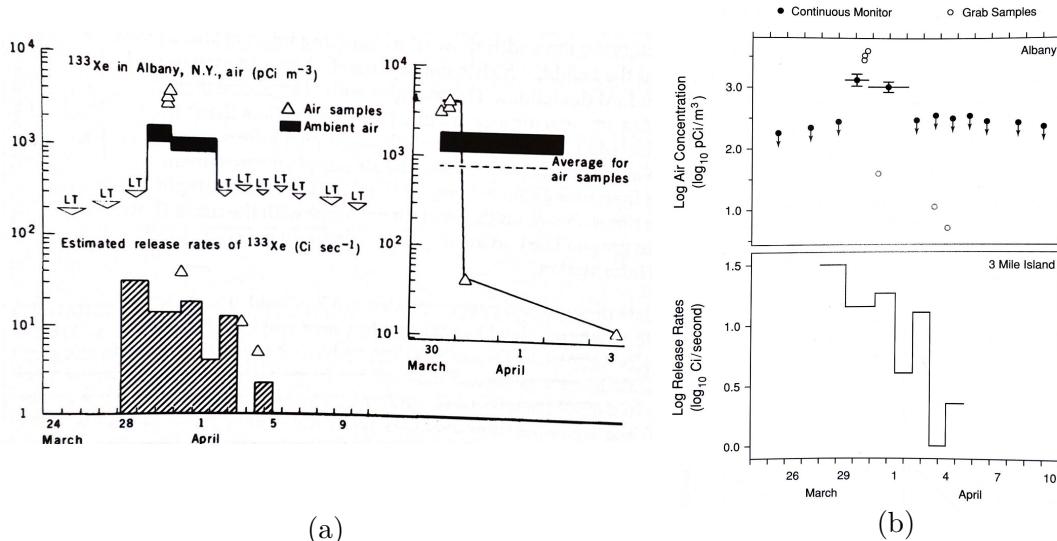
### 1.2.2 Wiliam S. Cleveland

Kromě práci Edwarda Tufta velký vliv měli i publikace Wiliama S. Clevelanda. Cleveland se svým kolegou Robertem McGillem publikovali v roce 1984 článek o grafickém vnímání [9]. Prováděli studie na rozdíl ve vnímání sloupcových grafů (pozice a obecné měřítko), koláčových grafů (úhel), skládaných sloupcových grafů (plocha), barevných a stínovaných map (saturace barev a stínování) a další. [6] Ve svých pracích *Visualizing data* z roku 1993 a *The Elements of Graphing Data* z roku 1994 Cleveland se zabýval principy vizualizace, grafickými metodami a technikami, vykreslením tří a více proměnných (rozměrů). Některé z jeho principů se schodují s principy vymezenými Tuftem, avšak práce Clevelanda v této oblasti přesahovala práci Tufta. Zásady a principy dle Clevelanda by se dali shrnout do čtyř hlavních kategorií: jasný vzhled, jasná srozumitelnost, měřítka, obecné postupy. [10]

- **Jasný vzhled**
  - Data by měla vyčnívat, vykreslení nadbytečných prvků (neboli *chartjunk* dle Tufta) by se mělo vyhnout.

- Ke zobrazení dat by se měli používat výrazné grafické prvky.
- Pro každou proměnnou by měla být použita dvojice os, prostor v takto vytvořeném obdélníku je určen k vykreslení grafu, značky na osách by měli směrovat mimo oblast grafu.
- Prostor grafu by neměl být přeplněný (legenda mimo oblast grafu atd.).
- S počtem značek na osách by se nemělo přehánět.
- Pokud je vhodné, referenční linie mohou být použity, avšak nesmějí zasahovat do dat.
- Popisky by neměli zasahovat do kvantitativních dat a nesmějí znepřehledňovat graf.
- Značky a klíče by měli vyskytovat mimo oblast grafu (případně v legendě), totéž se týká poznámek a nadpisů, které mohou být také umístěny do textu.
- Překrývající se data sety či symboly musí být visuálně snadně rozpoznaatelné.
- Jasnost obrazu musí být zachována při reprodukci i při snížení kvality a zmenšení rozlišení.

Cleveland jako příklad špatně zpracované vizualizace vybral graf 8a, na kterém je zobrazeno množství izotopu xenonu  $^{133}\text{Xe}$  ve vzduchu ( $\text{pCi} \cdot \text{m}^{-3}$ ) po havárii elektrárny Three Mile Island v Albany, New York koncem března a začátkem dubna roku 1979. Vše, včetně popisků os, klíčů a popisků dat bylo umístěno do oblasti grafu, není dodržená žádná ze zásad Clevelanda. Výsledkem je matoucí graf, který je obtížné číst. Stejná data byla vizualizována Clevelandem na grafu 8b s dodržením veškerých zásad: odstranění zbytečných objektů a detailu z oblasti grafu, rozlišné datasetsy se zobrazují ve vlastních panelech, oprava popisků popisujících měření.



Obrázek 8: Radioaktivní oblak při havárii elektrárny Three Mile Island:  $^{133}\text{Xe}$  ve vzduchu ve vzdálenosti 375 km (a) a stejný graf přepracovaný Clevelandem (b), 1994

- **Jasná srozumitelnost**

- Hlavní závěry by měli být obsaženy v grafické formě. Legenda a nadpisy by měli být srozumitelné a vyčerpávající.
- Grafy by měli být zkонтrolovaný.
- Mělo by se usilovat o přehlednost (viz “jasný vzhled”).

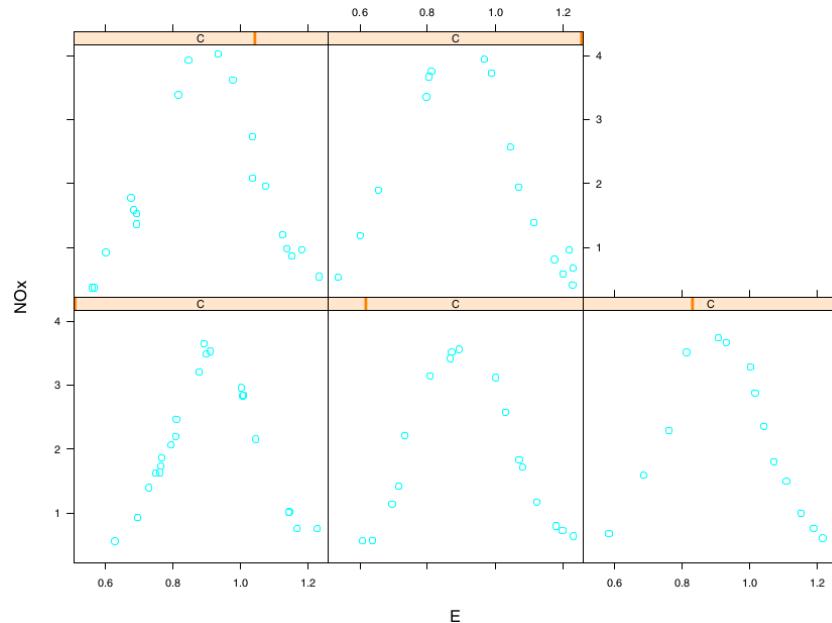
- **Měřítka**

- Volit rozsah os tak, aby obsahoval, případně téměř obsahoval, rozsah dat.
- Volit takové měřítko, aby data vyplňovala co největší prostor.
- Občas je užitečný mít pro proměnnou dvě osy pro rozdílná měřítka.
- Volit vhodné měřítko pokud data jsou porovnávány na více panelech.
- Osy grafu nemusejí vždy nutně zahrnovat nulu pro ukázku rozsahu.
- Použít logaritmická měřítka, když je důležitý pochopit procentní změny nebo multiplikativní faktory.
- Použít přerušené měřítko pouze v případě potřeby. Logaritmování může zbavit této potřeby.

- **Obecné postupy**

- Velké množství kvantitativní informace může být vměštěno do relativně malých oblastí.
- Tvorba grafů by měla být opakující, iterativní, experimentální činností.
- Data by měla být vykreslena tolíkrát kolikrát je třeba.
- Užitečné grafy vyžadují pečlivou a detailní práci.

Cleveland se mimo jiné podílel na tvorbě řady technik pro prohlížení komplexních datových sad s více proměnnými, kterým se říká *Trellis Graphics* nebo *Trellis Plots*. Technika obdržela svůj název *trellis* kvůli obvyklému výsledku řady obdélníkových grafů, připomínajících zahradní mříž. Na obrázku 9 je ukazka *trellis* grafu, zobrazující údaje o emisích motoru. [11]



Obrázek 9: *Trellis* graf, zobrazující údaje o emisích motoru, Becker et al., 1996

### 1.3 Grammar of graphics

*The Grammar of Graphics* publikována Lelandem Wilkinsonem v roce 2005 [12], detailně popisuje prvky, které tvoří základ všech statistických grafů, a odpovídá na základní otázku: co je statistická grafika? [13] Tato publikace měla extrémně velký vliv na myšlení o grafech. Hadley Wickham na základě Wilkinsonovy gramatiky publikoval v roce 2009 článek *A Layered Grammar of Graphics* [14], který se zaměřuje primárně na vrstvy a jejich zapojení do R. Následně také pro něj posloužila jako inspirace pro tvorbu balíčku `ggplot` (viz kapitola 4.2.1).

*The Grammar of Graphics* říká, že statistické grafy jsou mapováním z dat k estetickým atributům (barva, tvar, velikost) nebo geometrickým objektům (body, linie, sloupce). Graf také může obsahovat statistickou transformaci dat a být vykreslen ve specifickém souřadnicovém systému. *Faceting* může být použito k vygenerování stejného grafu pro různé podmnožiny datasetu. Kombinace těchto nezávislých komponent tvoří grafiku. Jednotlivé komponenty tvořící graf, dle Wilkinsonovy syntaxe, lze zapsat následovně:

- Vizualizované **data** a soubor estetických mapování (**mappings**), popisujících, jak proměnné z dat jsou mapovány na vnímané estetické atributy.
- Geometrické objekty (**geoms**) reprezentují to, co je doopravdy na grafu: body, linie, polygony atd.
- Statistické transformace (**stats**) summarizují data mnoha užitečnými způsoby. Jako příklad by se dál použít výpočty intervalů a počtu pozorování při tvorbě histogramu (kapitola 2.4.1) nebo tvorbu lineárního modelu. Statistické transformace patří k nepovinným, ale velmi užitečným komponentům.
- Měřítka (**scales**) reprezentují hodnoty v datovém prostoru převedené na hodnoty v estetickém prostoru, ať už se jedná o barvu, velikost či tvar. Na měřítku závisí legenda a osy, tvořené inverzním mapováním, umožňující číst z grafu původní hodnoty datasetu.
- Souřadnicový systém (**coord**) popisuje, jak jsou souřadnice dat mapovány do roviny grafiky. Rovněž poskytuje osy a mřížky, umožňující čtení grafů. Běžně se používá kartézský souřadnicový systém, ale je k dispozici i řada dalších systémů, včetně polárních souřadnicích a mapových projekcí.
- Specifikace **faceting** popisuje, jaké proměnné by měly být použity k rozdělení dat na podmnožiny a jak tyto podmnožiny by měly být uspořádány. Jedná se o mocný nástroj pro zkoumání toho, zda jsou modely stejné nebo odlišné v různých podmírkách.

Je také důležité zmínit, o čem Wilkinsonova gramatika není. Nenaznačuje, jaký typ grafů by se měl použít k zodpovězení otázek o datech, jak to dělali Cleveland [15] nebo Tukey [5], zaměřuje se konkrétně na jejich tvorbu. Ironií je, že *The Grammar of Graphics* neurčuje, jak by měla vypadat grafika, nespecifikuje velikost písma ani barvu pozadí. Otázkou vzhledu grafů se zabývali Tufte a Cleveland (kapitoly 1.2.1 a 1.2.2). Dále Wilkinsonova gramatika nepopisuje interakce, obsahuje pouze statické grafy. Při tvorbě dynamických či interaktivních grafů je třeba se obrátit na jiný zdroj. [13]

## 2 Základní grafy v R

Pro vytváření základních grafů v R používáme vestavěný balíček `graphics` [16], který obsahuje mnoho užitečných funkcí pro tvorbu grafických prvků. První kapitola se soustředí na tyto funkce tohoto balíčku a v dalších kapitolách jsou popsány funkce balíčků dalších (například `lattice`, `ggplot2`, ...), které zastávají podobné funkce, avšak s různým rozsahem nastavení [17].

V následujících příkladech nejsou grafy doplněny o barvy, popisky os, legendy ani názvy a to především proto, že záměrem této kapitoly je popsat základní grafy a funkce pro jejich tvorbu v prostředí R. Všechny tyto prvky mohou být přidány do grafu, ale tím by příkazy obsahovali irrelevantní parametry vzhledem k zaměření této kapitoly. Základní funkce `plot(x)` jejímž voláním se obdrží pole s grafickou reprezentací proměnné "x", by při doplnění kódu o veškeré parametry vypadala následovně [17]:

```
plot(x, main = "Název grafu", xlab = "popis osy x",
+     ylab = "popis osy y", col = c("red", "black", "green"))
```

Záměrem je tedy používání příkazů s pouze relevantními parametry.

### 2.1 Bodový graf

Bodový graf je rychlým způsobem, jak znázornit vztahy a souvislosti mezi proměnnými datasetu, případně k zjištění jejich neexistence. Data jsou zobrazeny v kartézském souřadném systému a mají pro každou hodnotu proměnné dané místo na vodorovné a svislé ose. V případě existence závislostí mezi proměnnými lze tuto závislost interpolovat přímkou, křivkou či dalším vhodným vyobrazením této závislosti.

Pro vytvoření bodového grafu v základním prostředí R (pomocí `graphics`) použijeme funkci `plot()`, která má tento typ grafu předdefinovaný pro numerické hodnoty. Viz obrázek 10 (a). Nečíselná data vytvoří jiný typ grafu.

```
plot(cars)
```

### 2.2 Liniový graf

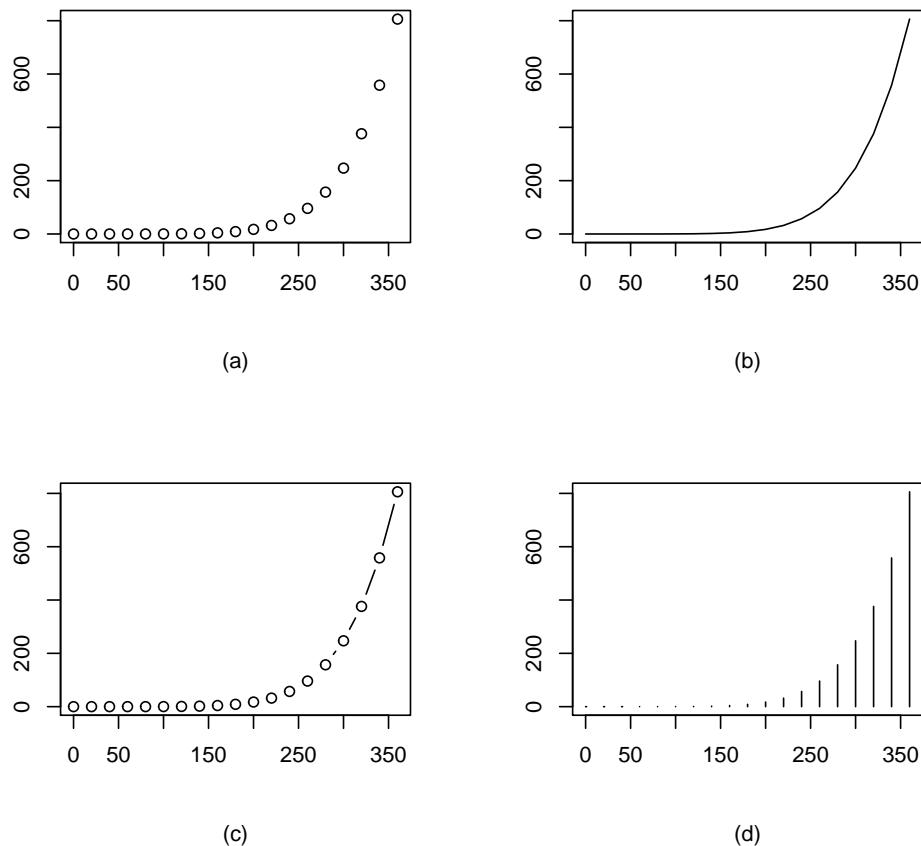
Jediný rozdíl mezi bodovým a liniovým grafem je, že jeden zobrazuje body a druhý je spojuje.[17] (viz. obrázek 10 (a), (b)). Pro vykreslení liniového grafu se používá již několikrát zmíněná funkce `plot()`, kterou doplníme o požadovaný typ vykreslení:

```
plot(x, type="l")
```

V tabulce 1 jsou uvedené některé základní atributy parametru `type`, které mohou být použity [18]:

	Anglický popis	Český popis
p	points	bodový
l	lines	liniový
b	both	složený
h	histogram	histogram
n	no plotting	bez vykreslení

Tabulka 1: Základní atributy parametru ‘type’



Obrázek 10: Porovnání základních typů grafů: (a) - bodový, (b) - liniový, (c) - složený, (d) - histogram

Popis a všechny atributy dalších parametrů funkce `plot()` lze nalézt v návodě zadáním příkazu `?plot()`.

### 2.3 Vykreslení rozdělení v R

Teorie pravděpodobnosti je základem statistiky a R má hodně nástrojů pro práci s pravděpodobností, rozdělením pravděpodobnosti a náhodnými proměnnými. R má zkrácený název pro každé rozdělení pravděpodobnosti. [17] Tyto názvy slouží

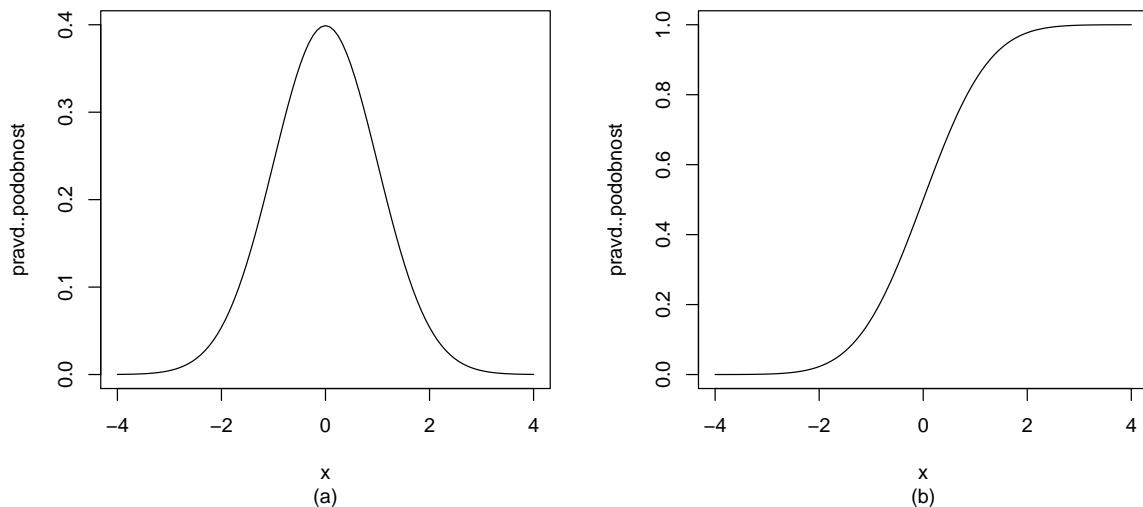
k identifikaci funkcí spojených s rozděleními. Například zkrácený název “norm” pro normální rozdělení, “exp” pro exponenciální rozdělení a další. Funkce pak mají formu:

Funkce	Účel
dxxxx	Hustota pravděpodobnosti
pxxxx	Distribuční funkce
qxxxx	Kvantilová funkce
rxxxx	Generátor náhodných čísel z daného rozdělení

Tabulka 2: Funkce pro práci s rozděleními

Funkce v R lze vykreslovat pomocí funkce `curve()` z balíčku `graphics`. Lze vykreslit jak standardní funkce, tak i funkce definované uživatelem. Například hustotu pravděpodobnosti normálního rozdělení a její distribuční funkci můžeme vykreslit tímto způsobem (Obrázek 11):

```
curve(dnorm(x))
curve(pnorm(x))
```



Obrázek 11: Hustota pravděpodobnosti normálního rozdělení (a) a její distribuční funkce (b)

### 2.3.1 Q-Q graf a P-P graf

Q-Q (*quantile-quantile*) graf a P-P (*probability-probability* nebo *percent-percent*) graf (Obrázek 12) se používají hlavně k testování normality při průzkumové analýze dat 3.4. Další způsob, jak zjistit zda-li data mají normální rozdělení je sestrojení histogramu (viz. sekce 1.4.1), avšak použití Q-Q grafu je přesnější.

Princip Q-Q grafu spočívá v porovnání dvou rozdělení pravděpodobnosti pomocí vykreslení jejich kvantilů proti sobě. Na jedné ose se nacházejí teoretické kvantily normálního rozdělení a na druhé ose kvantily naměřené (pozorované). Pokud data mají přesně normální rozdělení, všechny body grafu leží na přímce  $45^\circ$ . Vztah hustoty rozdělení a Q-Q grafu je znázorněn na obrázku 21. [17] [10]

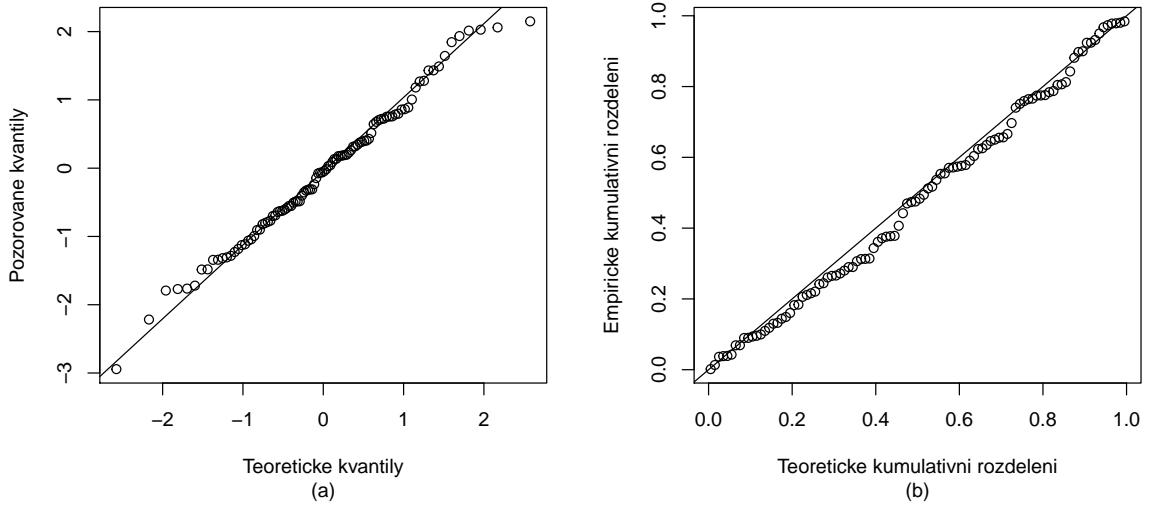
Princip P-P grafu je obdobný jako u Q-Q grafu: vykreslují se dvě distribuční funkcí proti sobě (jedná teoretická a jedná pozorovaná) a pokud všechny body grafu leží přibližně na přímce, jedná se o normální rozdělení. Z velké části se P-P graf používá k vyhodnocení koeficientu šikmosti rozdělení.[19]

V R se Q-Q graf vykreslí takto:

```
qqnorm(x)
qqline(x)
```

P-P graf v R lze vykreslit například následovně:

```
plot(ppoints(length(x)), sort(pnorm(x)))
abline(0,1)
```

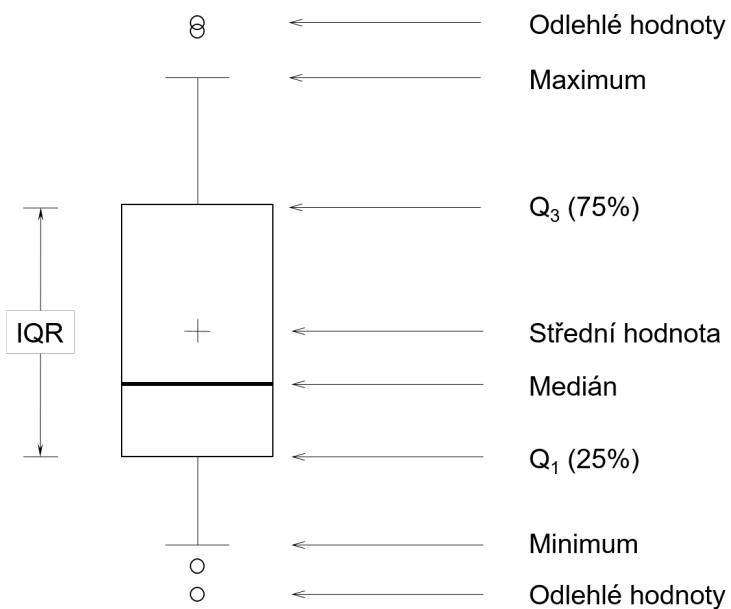


Obrázek 12: Q-Q Graf (a) a P-P Graf (b)

### 2.3.2 Krabicový graf

Krabicový graf poskytuje rychlé a jednoduché vizuální shrnutí datasetu. V základním prostředí R se vykreslí pomocí funkce `boxplot()` z balíčku `graphics`. Obrázek 13 znázorňuje typický krabicový graf, kde silná čára je medián, krabice kolem ní určuje polohu prvního a třetího kvartilu (dolní  $Q_1$  kvantil 25% a horní  $Q_3$  kvantil 75%). "Vousy" (*whiskers*) nad a pod krabici znázorňují rozpětí dat bez odlehlých hodnot. Odlehlé hodnoty jsou definované jako hodnoty ležící ve větší vzdálenosti od krabice než  $1,5 \times \text{IQR}$ , kde IQR je mezikvartilové rozpětí (*interquartile range*) neboli  $Q_3 - Q_1$ .

```
boxplot(x)
```



Obrázek 13: Boxplot

## 2.4 Sloupcový graf

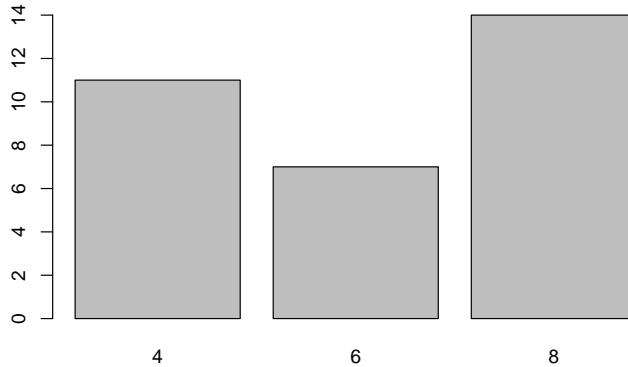
Sloupcový graf je jedním z nejvíce používaných způsobů vizualizace dat. Obvykle se používá pro zobrazení kvantitativních hodnot na ose y a kvalitativních na ose x. Výška sloupců může reprezentovat jak četnosti výskytu hodnot, tak i samotné hodnoty.[20]

V R lze tento typ grafu vykreslit pomocí funkce `barplot()`. V příkladu (Obrázek 14) je použit data set `mtcars`, konkrétně atribut `cyl` - počet válců v motoru.

```
table(mtcars$cyl)
```

```
##  
## 4 6 8  
## 11 7 14
```

```
barplot(table(mtcars$cyl))
```



Obrázek 14: Ukázka jednoduchého sloupcového grafu

### 2.4.1 Histogram

Sloupcový graf s četnostmi na souvislé ose je taky známý jako histogram.[20] Četnosti mohou být absolutní či relativní. Absolutní četnost zobrazuje počet statistických jednotek s hodnotou znaku, který patří do určitého intervalu. Podíl příslušné četnosti a rozsahu datového souboru se nazývá relativní četnost.[21] Šířka sloupce reprezentuje jednotlivé intervaly, které mají stejnou délku. Pro výpočet optimální délky intervalu existují různé metody. Základní histogram se vytváří pomocí funkci `hist()` a její atribut `breaks` udává buď hranice intervalů, jejich preferovaný počet nebo metodu výpočtu intervalu. V R jsou vestavěny 3 metody výpočtu:

1. Sturges [22]

```
hist(x, breaks = "Sturges")
```

$$k = \lceil \log_2(n) \rceil + 1$$

Kde  $k$  je počet intervalů a  $n$  je počet prvků neboli počet pozorování výběru  $x$ . Tato metoda je výchozí pro funkci `hist()`.

2. Scott [22]

```
hist(x, breaks = "Scott")
```

Scotovo pravidlo je následující:

$$h = \frac{3.5\sigma}{n^{\frac{1}{3}}}$$

kde  $\sigma$  je směrodatná odchylka a  $h$  je předpokládaná šířka intervalu.

Počet intervalů může být vypočítán pomocí vztahu:

$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil$$

Případně oba vztahy lze shrnout do jednoho:

$$k = \left\lceil n^{\frac{1}{3}} \frac{\max(x) - \min(x)}{3.5\sigma} \right\rceil$$

3. Freedman–Diaconis [23]

```
hist(x, breaks = "FD")
```

Freedman–Diaconisovo pravidlo pro stanovení předpokládané šířky intervalu je:

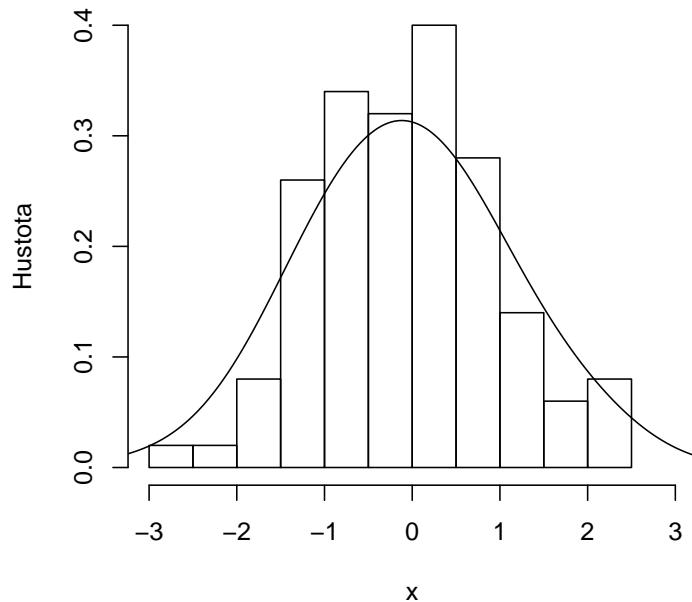
$$h = 2 \frac{IQR(x)}{n^{\frac{1}{3}}}$$

Po dosazení:

$$k = \left\lceil n^{\frac{1}{3}} \frac{\max(x) - \min(x)}{2IQR(x)} \right\rceil$$

kde  $IQR$  je mezikvartilové rozpětí, které definujeme jako rozdíl třetího a prvního kvartilů.

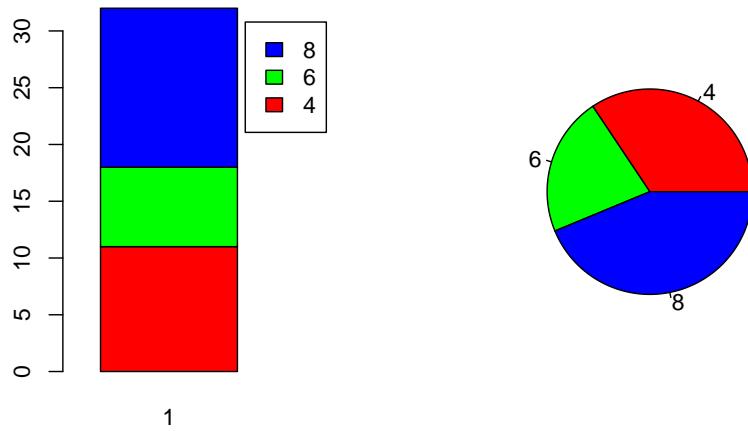
Histogram je jedním ze standardních způsobů, používaných k odhadu tvaru rozdělení, přesto se ale tento způsob považuje za nepřesný, vzhledem k ovlivnění tvaru počtem použitých intervalů. Při normálním rozdělení by měl histogram mít zvoncovitý tvar schodný s Gaussovou křivkou (Obrázek 15).



Obrázek 15: Histogram s odhadem hustoty pravděpodobnosti

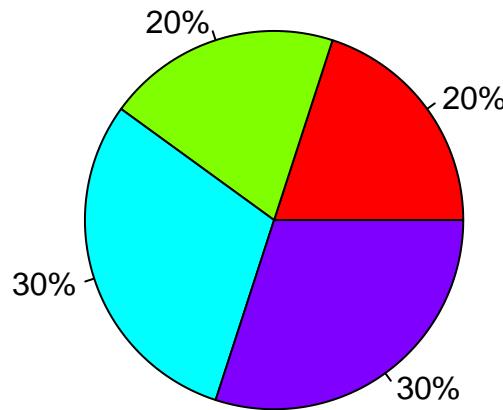
#### 2.4.2 Koláčový graf

Koláčový graf představuje plný kruh ( $360^\circ$ ), který je rozdělen na jednotlivé výseče pro znázornění číselných proporcí mezi proměnnými. Koláčový graf je tvořen transformací skládaného sloupcového grafu do polárního souřadnicového systému (Obrázek 16). [12]



Obrázek 16: Skládaný sloupcový graf transformovaný do polárního souřadnicového systému

Jednoduché koláčové grafy se vykreslují pomocí funkci `pie()` (Obrázek 17).



Obrázek 17: Ukázka jednoduchého koláčového grafu

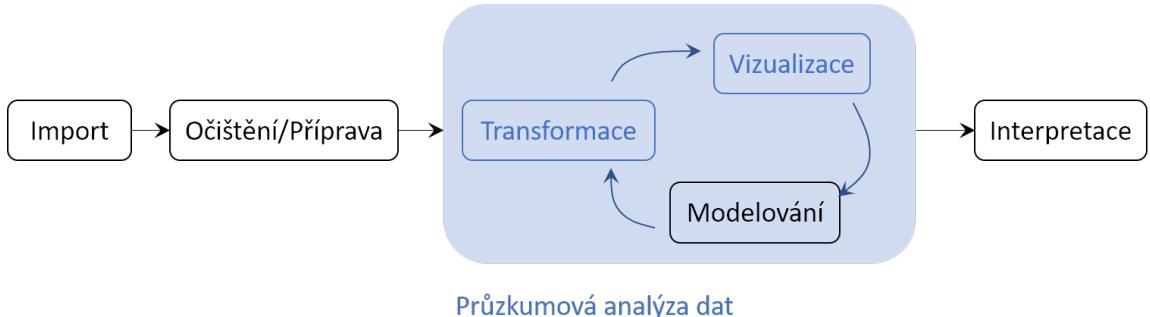
### 2.4.3 Číslicový histogram (*stem-and-leaf*)

Číslicový histogram, jinak známý jako *stem-and-leaf plot*, podobně jako histogram pomáhá vizualizovat tvar rozdělení. Jedná se spíše o historický typ grafu, který byl populární v osmdesátých letech, kvůli obtížnějšímu vykreslování velkých datasetu. Vstupní údaje jsou rozdělené vertikální linií na dva sloupce. Pravý sloupec obsahuje listy (*leaf*) - poslední číslice po desetinné čárce a levý sloupec obsahuje stonky (*stem*) - číslice před desetinnou čárkou. Každý stonk je uveden pouze i pokud neobsahuje žádné listy. Listy se uvádějí od nejmenšího po největší. [5] Proto v příkladu uvedeném níže je v prvním řádku stonkem číslice -2 a listy jsou číslice 9 a 2. Víme tak, že v datasetu se vyskytly čísla -2.9 a -2.2. Tento typ grafu v prostředí R se vykresluje pomocí funkce `stem()`:

```
stem(x)
```

```
##      The decimal point is at the |
## -2 | 92
## -1 | 888755333332211100
## -0 | 9988887766666655554433332111100
##   0 | 0011222223333444456777788888999
##   1 | 0233445689
##   2 | 0012
```

### 3 Průzkumová analýza dat



Obrázek 18: Posloupnost datové analýzy

Úkolem průzkumové analýzy dat (*Explanatory Data Analysis*, zkráceně EDA) je vizualizace a transformace dat systematickým způsobem za účelem maximálního pochopení dat, určení vztahu mezi nimi a posouzení jejich kvality. EDA je důležitou částí datové analýzy a měla by být jedním z jejích prvních kroků.

Zařazení průzkumové analýzy dat do procesu datové analýzy je zobrazeno v diagramu 18. Prvním krokem datové analýzy je **import** dat. Obecně to v tomto případě znamená nahrání obdržených dat ze souboru či databáze do prostředí R. Bez tohoto kroku datová analýza nemůže být vykonána. V momentě když data jsou importována do R je vhodné je **ocístit** neboli **přípravit**. Tím je myšleno ukládání dat v konzistentní a systematické formě, odpovídající sémantice původního datasetu. Zkrátka očištěná data jsou taková data, ve kterých sloupce odpovídají proměnným a řádky odpovídají pozorováním. Takováto příprava dat usnadňuje další práci s nimi.

Jakmile jsou data očištěna, je obvyklým krokem jejich **transformace**. Transformací se rozumí omezení pozorování (například dle zájmového území či povodí), vytváření nových proměnných na základě již existujících, agregace (např. z denního do měsíčního kroku), výpočet souhrnných statistik (středních hodnot, kvantilů atd.), odstranění odlehlcích pozorování a normalizace. Poté, co jsou data očištěná a obsahují veškeré potřebné proměnné, je možné na ně aplikovat dva nejdůležitější nástroje k zjištění informací: vizualizaci a modelování. Jakákoliv analýza tyto nástroje opakovaně využívá, ačkoliv samozřejmě mají své výhody a nevýhody.

**Vizualizace** je schopná odhalit neočekávané chování dat a napovědět další směr analýzy. Vizualizaci lze odhalit nevhodně zvolená či špatně připravená data a nekorektní dotazování. I přesto, že vizualizace je dobrým nástrojem datové analýzy, její aplikace na větší datasety je značně náročná a interpretace výsledků je subjektivní, tudíž závisí na analytikovi.

**Modelování** je v ramci průzkumové analýzy dat doplňkem vizualizace. Jedná se o zásadně matematický a výpočetní nástroj, který se obecně hodí i na větší datasety. Téměř každý model musí splňovat své předpoklady, které by měli být ověřeny před jejich aplikací, na rozdíl od vizualizace, která žádné předpoklady nevyžaduje.[24]

Důležitou součástí analýzy je **interpretace** výsledků a formulace závěrů. Vyhodnocuje, jak dobře zvolený model či vizualizace slouží k pochopení dat a jejich popisu. Je také důležité si uvědomit komu jsou výsledky interpretovány, kdo je cílová skupina. Dobře provedené grafické výstupy podložené jejich správnou interpretaci jsou jedním z nejlepších způsobů prezentace dat.

Průzkumová analýza dat není specifikována jako konkrétní soubor pravidel a postupů, ale jako přístup k analýze dat. Obvykle zahrnuje následující kroky:

- vyhledávání vybočujících (odlehlých) pozorování
- nahraď chybějících hodnot
- transformace dat
- změny typu proměnných
- ověřování normality

### 3.1 Odlehlá pozorování

Odlehlá pozorování (*outliers*) jsou významně odlišná vůči ostatním hodnotám datasetu. Definice toho, jak moc odlišná taková pozorování mají být je dáno analytikem na základě konkretního datasetu a kontextu problematiky. Tato pozorování mohou být indikátorem chybných dat nebo vzácných událostí. Důvody proč se tato pozorování vyskytují by měli být pečlivě zkoumány. Dále je důležité posoudit, jak je jimi výsledek analýzy ovlivněn, případně zdali je předpoklady metody připouštějí.

Hledání odlehlých, vybočujících, pozorování a jiných anomálií pro jednotlivé veličiny lze provést graficky například pomocí boxplotu (viz. sekce 2.3.2), bodových grafů (2.1) nebo číslicových histogramů (2.4.3). Dají se také vypočítat pomocí různých statistik, například metodou *jackknife*, která je popsána v následující kapitole (3.1.1). V momentech, kdy je vizualizace obtížná (velké datasety, větší množství navzájem se ovlivňujících proměnných, atd.), využívají se nástroje vícerozměrné, například Mahalanobisovy vzdálenosti (3.1.2), *leverages* (3.1.3) a další.

#### 3.1.1 *Jackknife*

Metoda byla původně představená Johnem W. Tukey v roce 1958 v „*The Annals of Mathematical Statistic*“ [25] a jedná se o speciální případ metody *bootstrap* (více o metodě B. Efron a R. Tibshirani v „*An Introduction to the Bootstrap*“ [26]).

Postup metody *jackknife* je založen na celkem jednoduché myšlence. Zjišťují se souhrnné statistiky podsouborů (*Jackknife Samples*), které se vytvářejí postupným vypouštěním jednotlivých pozorování z původního datasetu. Jinými slovy existuje  $n$  unikátních Jackknife podsouborů a  $i$ -tý Jackknife podsoubor je definován jako vektor.

Pomocí porovnání souhrnných statistik původního datasetu a vytvořených Jackknife podsouborů se odhadne vliv jednotlivých pozorování na původní dataset. Jedná ze souhrnných statistik, kterou lze použít je střední hodnota  $\bar{x}$ . Pro původní dataset

obsahující  $n$  pozorování lze střední hodnotu odhadnout dle vzorce  $\bar{x} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i$ . Střední hodnota Jackknife podsouborů se vyhodnotí následovně:

$$\bar{x}_i = \frac{1}{n-1} \sum_{j=1, j \neq i}^n x_j, \quad \text{kde } i = 1, \dots, n.$$

Porovnání lze provést dle vzorce  $Var(\bar{x}) = \frac{n-1}{n} \sum_{i=1}^n (\bar{x}_i - \bar{x})^2$ , kde  $Var(\bar{x})$  je odhad rozptylu, který indikuje, jak moc jednotlivá pozorování ovlivňují dataset, tj. přítomnost odlehlych pozorování. Metoda může být také použita k odhadu skutečné, neovlivněné střední hodnoty datasetu. [27]

### 3.1.2 Mahalanobisovy vzdálenosti

K měření vzdálenosti mezi objekty se často používá euklidovská vzdálenost. Euklidovská vzdálenost je jednoduchá na výpočet a interpretaci, ale není schopná brát v úvahu vztahy mezi daty. Proto je v řadě případů vhodné použít mahalanobisovou vzdálenost. Je definovaná matice  $\mathbf{X}(n \times p)$ , obsahující  $n$  objektů  $\mathbf{x}_i$  a  $p$  proměnných. Euklidovská vzdálenost mezi vektorem  $i$ -tého řádku  $\mathbf{x}_i(1 \times p)$  této matice a vektoru středních hodnot  $\bar{\mathbf{x}}(1 \times p)$  se spočítá jako

$$ED_i = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T}, \quad \text{pro } i = 1, \dots, n$$

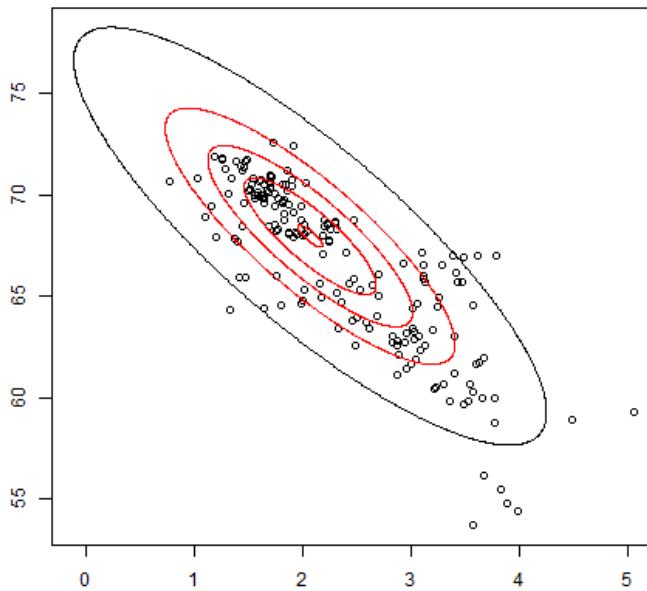
zatímco mahalanobisova vzdálenost se spočítá jako

$$MD_i = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{C}_x^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^T}, \quad \text{pro } i = 1, \dots, n$$

kde  $\mathbf{C}_x$  je kovarianční matice. [28]

Na obrázku 19 jsou znázorněny elipsy mahalanobisových vzdáleností, kde každá elipsa představuje vzdálenost od průměru. Z tohoto je zřejmé, že vzdálenost roste pomaleji ve směru korelace. Pozorování, které je výrazně vzdáleno od středu, ale leží ve směru závislosti, má nižší mahalanobisovou vzdálenost než pozorování, které je stejně vzdáleno od středu, ale neleží ve směru závislosti. Tato vlastnost mahalanobisových vzdáleností umožňuje identifikaci odlehlych pozorování.

Metoda byla představena P.C. Mahalanobisem v roce 1936 ve článku “On the Generalized Distance in Statistics” [29]. Mahalanobisové vzdálenosti se používají nejenom k nalezení odlehlych pozorování, ale i ke zkoumání reprezentativity mezi dvěma data sety, aplikuje se v algoritmu  $k$ -nejbližších sousedů, v diskriminační analýze a má mnoho dalších uplatnění.



Obrázek 19: Mahalanobisovy vzdálenosti

### 3.1.3 Leverages

Leverage (případně též efekt, vliv nebo projekční  $h$  prvek) se používá v regresní analýze k měření velikosti vlivu pozorování na regresní odhad. Princip metody spočívá v kontrole diagonálních prvků projekční matice  $\mathbf{H}$ , která je produktem metody nejmenších čtverců a je definována

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T.$$

Model lineární regrese může být zapsán následovně:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde vektor vysvětlované proměnné je  $\mathbf{y}$ , matice vysvětlujících proměnných je  $\mathbf{X}$ , vektor regresních koeficientů, který je odhadován, je  $\boldsymbol{\beta}$  a vektor náhodné složky je  $\boldsymbol{\varepsilon}$ . Metoda nejmenších čtverců poskytuje řešení regresních rovnic:

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Lze dosadit:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Výsledný vektor má tvar  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ , kde  $\mathbf{H}$  je projekční matice. [30]

### 3.2 Náhrada chybějících pozorování

Problém chybějících pozorování spočívá v neschopnosti jejich zpracovávání některými metodami. Takové hodnoty lze vynechat nebo doplnit (nahradit) jednou z řady metod. Vynechání hodnot vede k nežádoucímu zmenšení datasetu, proto je výhodnější chybějící údaje doplnit. Nejednodušším nástrojem pro náhradu chybějících hodnot je aritmetický průměr příslušné proměnné. Tento způsob může vést ke zkresleným odhadům (neplatí-li předpoklad, že chybějící údaje jsou zcela náhodné) a podhodnocuje variabilitu a kovarianci datasetu, a proto se nedoporučuje v případě vyššího podílu chybějících údajů. Další možnou metodou je náhrada náhodným číslem generovaným z příslušného rozdělení (parametry jsou odhaduty z výběru). V tomto případě se respektuje variabilita datasetu, ale nerespektuje se jeho kovariance. Chybějící údaje lze také odvodit pomocí známých hodnot na základě pomocné jednoduché lineární regresní funkce. Tato metoda respektuje nejenom variabilitu vzorku, ale i jeho korelační strukturu. [31]

### 3.3 Transformace dat

Jedním z cílů transformace dat je dosažení srovnatelnosti proměnných: sjednocení měřítka, variace a typu proměnných. Hlavním využitím je splnění podmínek vyžadovaných metodami, například podmínky normality, kde je snaha převést data na normální rozdělení, snížení vlivu rušivých proměnných (odlehlych hodnot) atd. [32] Rozdělujeme transformaci lineární (centrování, normování) a nelineární (plynoucí z typu a charakteru dat).

Lineární transformace zachovává lineární vztahy mezi proměnnými. Jedním z příkladů takovéto úpravy dat je metoda centrování, která se používá u vícerozměrných analýz. Podstata metody spočívá v zachování měřítka vzorku při změně hodnot: od původních hodnot se odečítá průměr proměnné (od prvků sloupce se odečte jejich sloupcový průměr), průměry získaných nových proměnných se tudíž rovnají nule. Toto lze zapsat následovně:

$$v_{ij} = x_{ij} - \bar{x}_j$$

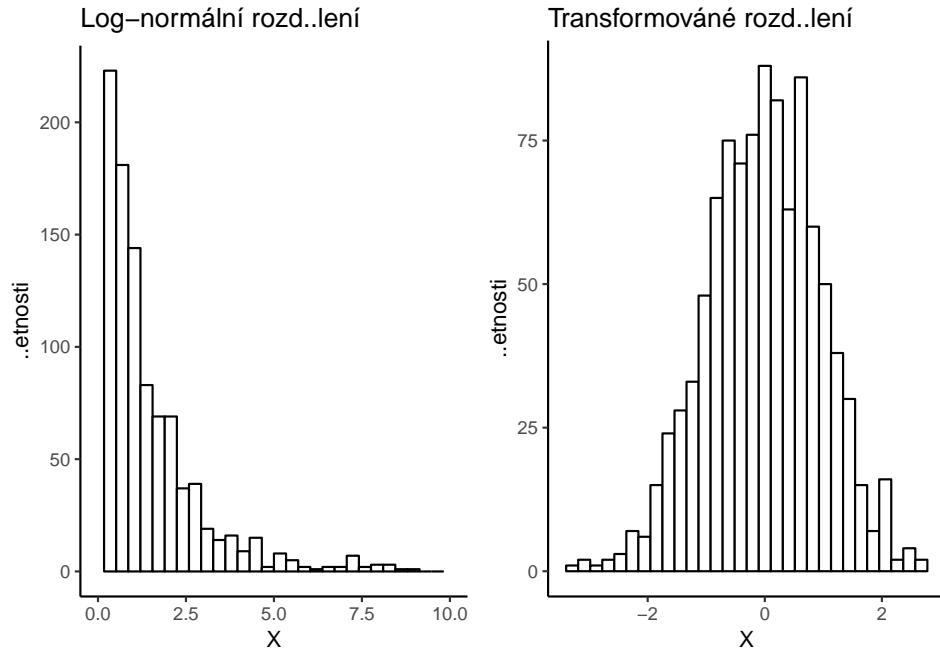
Vektor průměrů  $\bar{v}$  je nulový, kovariance a korelace proměnných zůstává nezměněna. [33] Další často využívanou metodou je metoda normalizace dat. Tato metoda transformuje měřítka vzorků pro možnost jejich porovnání (eliminuje jednotky měření), po úpravě střední hodnota vzorku tedy odpovídá nule a odchylka jedničce (normální rozdělení).

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma(x_j)}$$

$\sigma(x_j)$  je směrodatná odchylka sloupce proměnné, vektor průměrů  $\bar{z}$  je nulový a kovariance vektoru nových proměnných se shoduje s korelací původního vektoru. [34]

Nelineární transformace vyplývá z typu dat a mění (snižuje či zvyšuje) lineární vztahy mezi proměnnými a to znamená, že nezachovává korelací mezi nimi. Pokud data

mají charakter absolutní četnosti, používá se odmocninová transformace  $X' = \sqrt{X}$ , pokud odpovídají log-normálnímu rozdělení, používá se logaritmická transformace  $X' = \log_{10} X$  atd. Logaritmus náhodné veličiny s log-normálním rozdělením má normální rozdělení (viz obrázek 20). Logaritmická transformace může být použita pouze u nezáporných rozdělení. [35] [36]

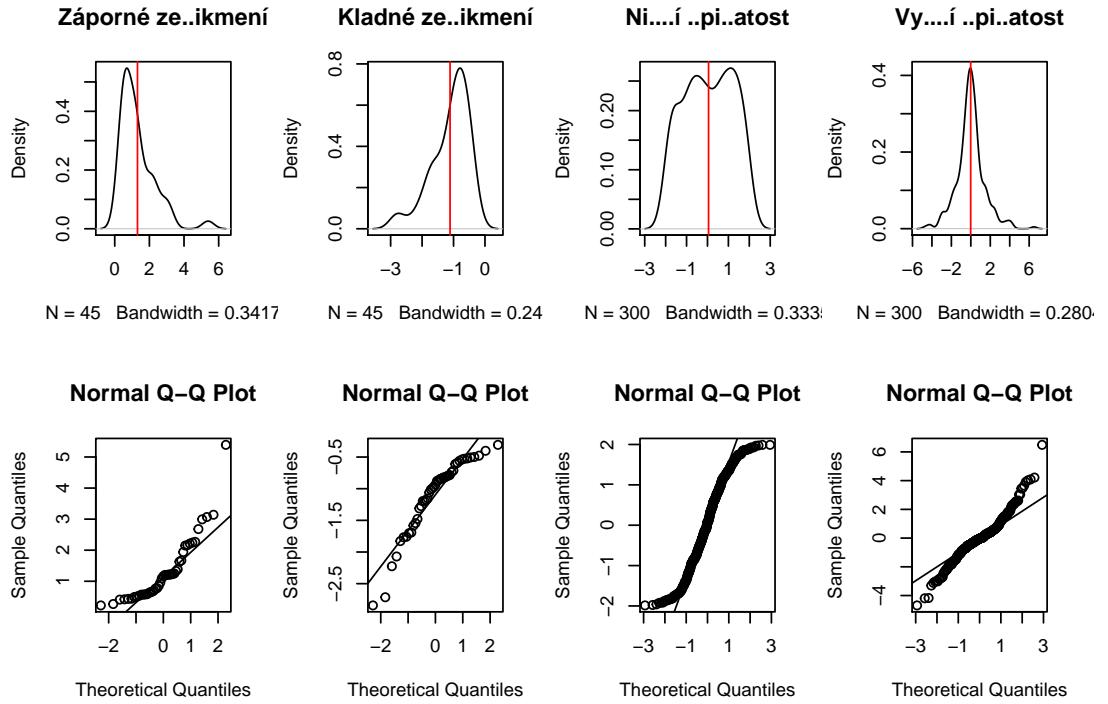


Obrázek 20: Log-normální rozdělení transformováné na normální rozdělení

### 3.4 Ověřování normality

Důležitým aspektem popisu proměnné je tvar jejího rozdělení, který udává četnosti hodnot z různých rozsahů proměnné. Většina statistických testů a metod se zakládá na předpokladu, že proměnná má normální rozdělení. Z tohoto důvodu je vhodné ověřovat normalitu rozdělení analyzovaného vzorku.

Zjistit zda-li vzorek pochází z normálního rozdělení lze grafickým posouzením nebo pomocí testů normality. Mezi nástroje grafického posouzení normality se řadí histogram rozdělení četnosti (kapitola 2.4.1), graf výběrové distribuční funkce (2.3), Q-Q graf a P-P graf (2.3.1). Vztah hustoty rozdělení a Q-Q grafu je znázorněn na obrázku 21. Dále existuje řada testů normality, zde jsou popsány testy Shapiro-Wilk (SW) a Jarqua-Bera (JB).



Obrázek 21: Vztah hustoty rozdělení a Q-Q grafu pro různá narušení normality

Shapiro-Wilk test byl poprvé představen v roce 1965 S. S. Shapirem a M. Wilkem [37]. Metoda dokáže pracovat se vzorky velikosti 12 až 5000 pozorování. Nulová hypotéza tohoto testu předpokládá, že vzorek má normální rozdělení. Pokud  $p$ -hodnota je menší, než zvolená hladina významnosti, zamítá se nulová hypotéza, jinými slovy vzorek nemá normální rozdělení. Statistika testu vypadá následovně:

$$W = \frac{\left( \sum_{i=1}^n a_i x_{(i)} \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

kde  $x_{(i)}$  je  $i$ -tý nejmenší prvek (statistika  $i$ -tého řádu),  $\bar{x}$  je průměr vzorku,  $n$  je počet pozorování.

Jarqua-Bera test závisí na koeficientech šikmosti a špičatosti. Statistika JB testu může být zapsána:

$$T = n \left( \frac{(\sqrt{b_1})^2}{6} + \frac{(b_2 - 3)^2}{24} \right),$$

kde  $n$  je velikost vzorku,  $\sqrt{b_1}$  je koeficient šikmosti vzorku a  $b_2$  je koeficient špičatosti. Nulová a alternativní hypotéza se schoduje s SW testem. Používá se pro větší datasety nad 2000 pozorování. [38]

## 4 Pokročilá vizualizace v R

### 4.1 Balíčky pro vizualizaci dat

Vestavený balíček **base** byl vyvinut Rossem Ihaka na základě zkušeností s implementací grafických ovladačů do S (předchůdce R). Grafy v **base** mají charakter grafů na papíře: stávající obsah nelze modifikovat, ani odstranit. Do grafu lze přidávat potřebné prvky, které jsou vykreslovány na povrch grafu a po vykreslení nemohou být dále měněny. V **base** neexistuje jiná uživateli přístupná reprezentace, než ta, co se objeví na obrazovce. **base** obsahuje nástroje pro kreslení jak primitivních, tak i kompletních výkresů. Funkce tohoto balíčku jsou obecně rychlé, ale mají omezené možnosti. [12] Vykreslení základních grafů v R je popsány v kapitole 2.

Mimo **base** uživatel má možnost využít rozsahlou nabídku dalších balíčků. Následující kapitoly obsahují krátký popis vybraných balíčku. Pro jejich instalaci se používá příkaz `install.packages()`.

#### 4.1.1 **grid**

**grid** je alternativou jednoduché grafiky z **base**, s rozsáhlejšími možnostmi. Vývoj balíčku začal v roce 2000. Autorem je Paul Murrell a balíček vzníkl na základě jeho doktorské práce *Investigations in Graphical Statistics* z roku 1998 [39]. Grafické objekty v **grid** můžou být reprezentovány nezávislé na grafu a později upraveny. Systém *viewportů* (každý obsahuje vlastní souřadnicový systém) usnadňuje rozvržení komplexní grafiky. Balíček umožňuje výkres primitivních výkresů, avšak nemá nástroje pro tvorbu statistických grafů. [40]

#### 4.1.2 **lattice**

**lattice** vyvinutý Deepayanem Sarkarem používá grafiku **grid** k implementaci Clevelandového mřížkového grafického systému, což vedlo k značnému vylepšení grafiky oproti **base**.

#### 4.1.3 **ggplot2**

Balíček **ggplot2** byl vyvinut v roce 2005 Hadley Wickhamem na základě *The Grammar of Graphics* Lelanda Wilkinsona. **ggplot2** přebírá přednosti balíčků **base** a **lattice** a vylepšuje je silným základním modelem, který podporuje tvorbu libovolného statistického grafu, založeného na principech popsaných v kapitole 1.3.

## 4.2 Balíčky pro interaktivní vizualizaci dat (`htmlwidgets`)

### 4.2.1 **plotly**

**4.2.2 dygraphs**

**4.2.3 leaflet**

**4.2.4 ggvis**

**4.2.5 rgl**

**4.3 Balíčky pro prostorová data**

**4.3.1 ggmap**

**4.4 ...**

**4.4.1 raster**

**4.4.2 rasterVis**

**4.5 Balíčky pro webové aplikace**

**4.5.1 shiny**

**4.5.2 flexdashboard**

**4.5.3 dashboard**

**Praktická část**

## Seznam obrázků

1	Kombinace různých vuzuálních technik, Playfair 1801 . . . . .	11
2	Dobytek odeslaný z celé Francie ke spotřebě v Paříži, Minard 1858 . .	12
3	Mapa světové migrace, Minard 1858 . . . . .	12
4	Postup Napoleonských vojsk v letech 1812-13, Minard 1869 . . . . .	13
5	Ukázky vizualizaci ze začatku 20. století, Brinton 1919 . . . . .	14
6	Zmenšující se rodinný lékař, Los Angeles Times, 1979 . . . . .	15
7	Vztah skutečné míry volební registrace k předpovídaným hodnotám, přetištěné E. Tufte, 1983 . . . . .	15
8	Radioaktivní oblak při havárii elektrárny Three Mile Island: $^{133}\text{Xe}$ ve vzduchu ve vzdálenosti 375 km (a) a stejný graf přepracovaný Clevelandem (b), 1994 . . . . .	17
9	<i>Trellis</i> graf, zobrazující údaje o emisích motoru, Becker et al., 1996 .	18
10	Porovnání základních typů grafů: (a) - bodový, (b) - liniový, (c) - složený, (d) - histogram . . . . .	21
11	Hustota pravděpodobnosti normálního rozdělení (a) a její distribuční funkce (b) . . . . .	22
12	Q-Q Graf (a) a P-P Graf (b) . . . . .	23
13	Boxplot . . . . .	24
14	Ukázka jednoduchého sloupcového grafu . . . . .	25
15	Histogram s odhadem hustoty pravděpodobnosti . . . . .	27
16	Skládaný sloupcový graf transformovaný do polárního souřadnicového systému . . . . .	27
17	Ukázka jednoduchého koláčového grafu . . . . .	28
18	Posloupnost datové analýzy . . . . .	29
19	Mahalanobisovy vzdálenosti . . . . .	32
20	Log-normální rozdělení transformované na normální rozdělení . . . . .	34
21	Vztah hustoty rozdělení a Q-Q grafu pro různá narušení normality .	35

## Seznam tabulek

1	Základní atributy parametru ‘type’ . . . . .	21
2	Funkce pro práci s rozděleními . . . . .	22

## Literatura

- [1] RAHLF, T. *Data Visualisation with R: 100 Examples*. B.m.: Springer, 2017. ISBN 978-3-319-49751-8.
- [2] FRIENDLY, M. A Brief History of Data Visualization. In: *Handbook of Computational Statistics: Data Visualization*. B.m.: Springer-Verlag, 2006. ISBN 978-3-540-32825-4.
- [3] BRINTON, W.C. *Graphic Methods for Presenting Facts*. B.m.: The Engineering Magazine Company, New York, 1919. ISBN 978-1155058870.
- [4] TUKEY, J.W. The Future of Data Analysis. *Annals of Mathematical Statistics* [online]. 1962, roč. 33, s. 1–6. Dostupné z: doi:10.1214/aoms/1177704711
- [5] TUKEY, J.W. *Exploratory Data Analysis*. B.m.: Addison-Wesley, 1977. ISBN 0201076160.
- [6] *How William Cleveland Turned Data Visualization Into a Science* [vid. 2.11.2017] [online]. Dostupné z: <https://priceconomics.com/>
- [7] TUFTE, E.R. *Envisioning Information*. B.m.: Graphics Pr, 1990. ISBN 978-0961392116.
- [8] TUFTE, E.R. *The Visual Display of Quantitative Information*. B.m.: Graphics Press, 1983. ISBN 978-0961392109.
- [9] CLEVELAND, W.S. a R. MCGILL. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association* [online]. 1984, roč. 79. Dostupné z: doi:10.1080/01621459.1984.10478080
- [10] CLEVELAND, W.S. *The Elements of Graphing Data*. B.m.: Hobart Press, 1994. ISBN 0963488414.
- [11] BECKER, R.A., W.S. CLEVELAND, M.-J. SHYU, aj. A tour of Trellis graphics. *Murray Hill, NJ: AT & T Bell Laboratories*. 1996, roč. 44.
- [12] WILKINSON, L. *The Grammar of Graphics*. B.m.: Springer, 2005. ISBN 9780387286952.
- [13] WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis (Use R!)*. B.m.: Springer, 2010. ISBN 978-0387981406.
- [14] WICKHAM, H. A layered grammar of graphics. *Journal of Computational and Graphical Statistics* [online]. 2010. Dostupné z: doi:10.1198/jcgs.2009.07098
- [15] CLEVELAND, W.S. *Visualizing Data*. B.m.: Hobart Press, 1993. ISBN 978-0963488404.
- [16] *R: The R Graphics Package* [vid. 22.4.2017] [online]. Dostupné z: <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/graphics-package.html>

- [17] TEETOR, P. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. B.m.: O'Reilly Media, 2011. ISBN 9781449307264.
- [18] R: Generic X-Y Plotting [vid. 11.5.2017] [online]. Dostupné z: <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html>
- [19] P-P plot - Wikipedia [vid. 11.8.2017] [online]. Dostupné z: [https://en.wikipedia.org/wiki/P%E2%80%93P\\_plot](https://en.wikipedia.org/wiki/P%E2%80%93P_plot)
- [20] CHANG, W. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. B.m.: O'Reilly Media, 2012. ISBN 9781449363116.
- [21] NOVOVIČOVÁ, J. *Pravděpodobnost a matematická statistika*. B.m.: Praha: Vydavatelství ČVUT, 2006. ISBN 80-01-01980-2.
- [22] MACIEJEWSKI, R. *Data Representations, Transformations, and Statistics for Visual Reasoning*. B.m.: Morgan & Claypool Publishers, 2011. ISBN 978-1-608-45625-3.
- [23] Histogram - Wikipedia [vid. 6.8.2017] [online]. Dostupné z: <https://en.wikipedia.org/wiki/Histogram>
- [24] WICKHAM, H. a G. GROLEMUND. *R for Data Science*. B.m.: O'Reilly Media, 2017. ISBN 1491910399.
- [25] TUKEY, J.W. Bias and Confidence in Not Quite Large Samples. *Annals of Mathematical Statistics*. 1958, roč. 29.
- [26] EFRON, B. a R. TIBSHIRANI. *An Introduction to the Bootstrap*. B.m.: Taylor & Francis Ltd, 1994. ISBN 0-412-04231-2.
- [27] MCINTOSH, A. The Jackknife Estimation Method. *arXiv*. 2016.
- [28] DE MAESSCHALCK, R., D. JOUAN-RIMBAUD a D.L. MASSART. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*. 2000, roč. 50, č. 1.
- [29] MAHALANOBIS, P.C. On the generalised distance in statistics. *Proceedings National Institute of Science*. 1936, roč. 2, č. 1.
- [30] CARDINALI, C. Observation influence diagnostic of a data assimilation system. *Advanced Data Assimilation for Geosciences: Lecture Notes of the Les Houches School of Physics: Special Issue*. 2014.
- [31] PECÁKOVÁ, I. Problém chybějících dat v dotazníkových šetřeních. *Politická ekonomie*. 2014. ISSN 2336-8225.
- [32] Transformation of Data [vid. 3.2.2018] [online]. Dostupné z: <http://statisticalconcepts.blogspot.cz/2010/02/transformation-of-data-validity-of.html>
- [33] HEBÁK, P., J. HUSTOPECKÝ, E. JAROŠOVÁ, aj. *Vícerozměrné statistické metody (1)*. B.m.: Informatorium, Praha, 2007. ISBN 80-7333-025-3.

- [34] ABDI, H. a L. WILLIAMS. Normalizing data. Encyclopedia of research design. *Thousand Oaks, CA: Sage*. 2010.
- [35] ZUMEL, N. a J. MOUNT. *Practical Data Science with R*. B.m.: Manning, 2014. ISBN 9781617291562.
- [36] KUTNER, M.H., C.J. NACHTSHEIM, J. NETER, aj. *Applied Linear Statistical Models*. B.m.: McGraw-Hill/Irwin, 2004. ISBN 0-07-238688-6.
- [37] SHAPIRO, S.S. a M.B. WILK. An analysis of variance test for normality (complete samples)†. *Biometrika* [online]. 1965, roč. 52, č. 3-4. Dostupné z: doi:10.1093/biomet/52.3-4.591
- [38] ÖZTUNA, D., A.H. ELHAN a E. TÜCCAR. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*. 2006, roč. 36, č. 3.
- [39] MURRELL, P.R. *Investigations in Graphical Statistics*. B.m., 1998. PhD thesis. The University of Auckland.
- [40] MURRELL, P. *grid: The grid graphics package*. červenec 2003