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

Autorka práce:	Irina Georgievová
Studijní program:	Krajinářství
Obor:	Vodní hospodářství
Vedoucí práce:	doc. Ing. Martin Hanel, Ph.D.
Garantující pracoviště:	Katedra vodního hospodářství a environmentálního modelování
Jazyk práce:	Čeština
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.).
Metodika:	- rešerše základních poznatků - popis vizualizačních prostředků se zaměřením na využití v hydrologii, porovnání výhod/nevýhod - popis nejpoužívanějších R balíků, jejich základních funkcí a demonstrace jejich využití
Doporučený rozsah práce:	40-60 stran
Klíčová slova:	vizualizace dat, grammar of graphics, průzkumová analýza dat

Doporučené zdroje informací:

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

Předběžný termín obhajoby: 2017/18 LS - FŽP

Elektronicky zamítnuto: 25. 4. 2017
doc. Ing. Martin Hanel, Ph.D.
Vedoucí katedry

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:

Obsah

Úvod

Teoretická část

1 Grammar of graphics

1.1 Historie vizualizace dat (stručná)

1.1.1 Minard & Playfair

1.2 Zásady vizualizace dat

1.2.1 Tuft

(*The Visual Display of Quantitative Information*; *Tufte's principles*)

1.2.2 Cleveland

(*Elements of Graphing Data*; *Visualizing Data*)

2 Základní grafy v R

Pro vytváření základních grafů v R používáme vestavěný balíček `graphics` [9], 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 [10].

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 irelevantní parametry vzhledem k zaměření této kapitoly. Základní funkce `plot(x)` jejímž voláním se obdrží pole s grafickou reprezentaci proměnné "x", by při doplnění kódu o veškeré parametry vypadala následovaně [10]:

```
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 1 (a). Nečíselná data vytvoří jiný typ grafu.

```
plot(cars)
```

2.2 Liniový graf

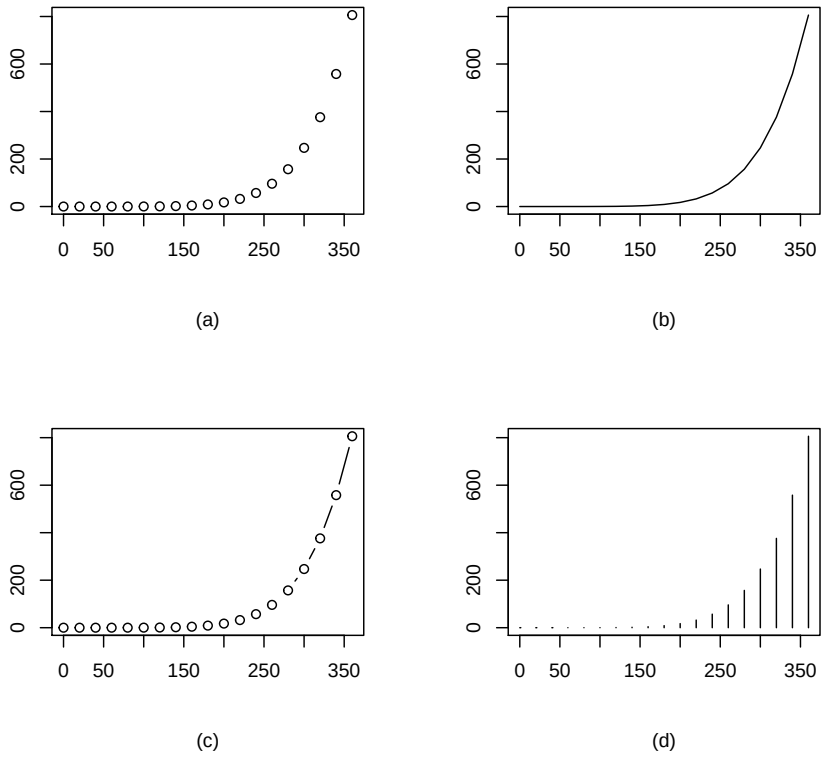
Jediný rozdíl mezi bodovým a liniovým grafem je, že jeden zobrazuje body a druhý je spojuje.[10] (viz. obrázek 1 (a), (b)). Pro vykreslení liniového grafu se používá již několikrát zmíněná funkce `plot()`, kterou doplníme o požadovaný typ vykreslení:

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

V tabulce 1 jsou uvedené některé základní atributy parametru `type`, které mohou být použity [8]:

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

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

2.3 Vykreslení rozdělení v R

Teorie pravděpodobnosti je základem statistiky a R má hodně nástrojů pro práci s pravděpodobnosti, rozdělením pravděpodobnosti a náhodnými proměnnými. R má zkrácený název pro každé rozdělení pravděpodobnosti. [10] Tyto názvy slouží

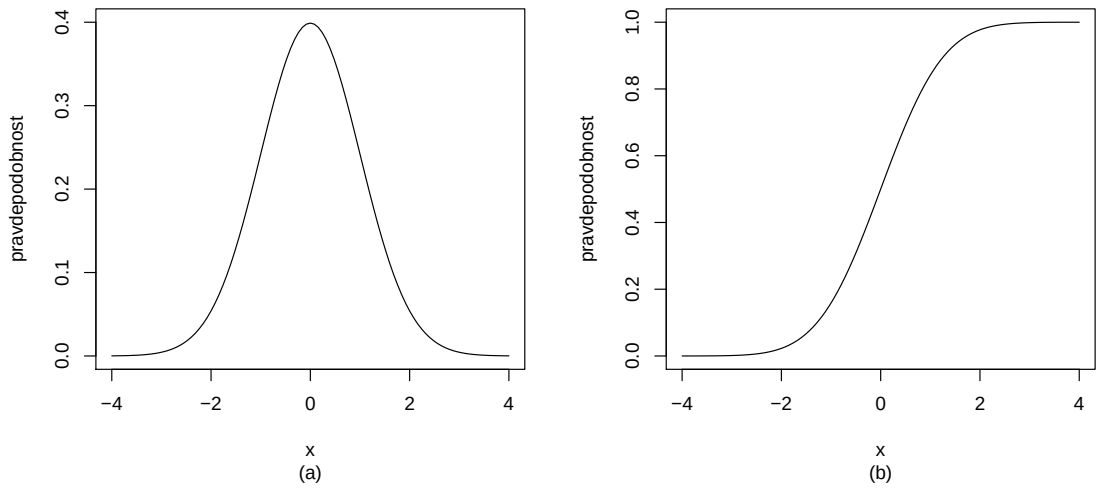
k identifikaci funkcí spojených s rozděleními. Například zkrácený název "norm" pro normální rozdělení, "exp" pro exponenciální rozdělení a další. Funkce pak mají formu:

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 2):

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



Obrázek 2: 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 3) se používají hlavně k testování normality při průzkumové analýze dat. 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°. [10] [2]

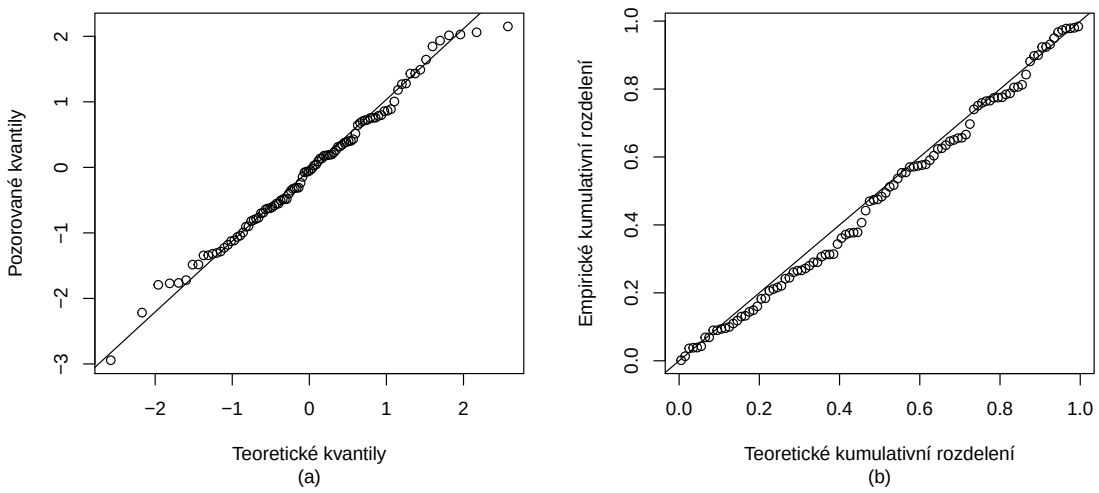
Princip P-P grafu je obdobný jako u Q-Q grafu: vykreslují se dvě distribuční funkcí proti sobě (jedná teoretická a jedná pozorovaná) a pokud všechny body grafu leží přibližně na přímce, jedná se o normální rozdělení. Z velké části se P-P graf používá k vyhodnocení koeficientu šikmosti rozdělení.[7]

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)
```

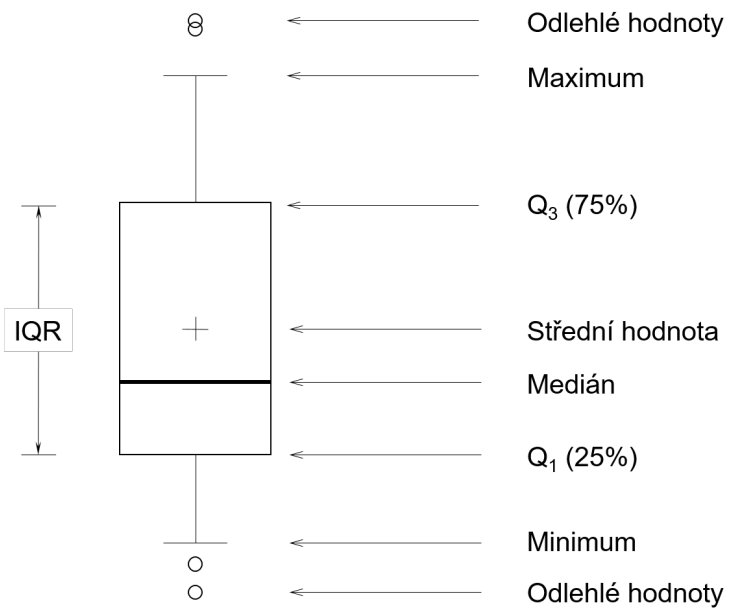


Obrázek 3: Q-Q Graf (a) a P-P Graf (b)

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 4 znázorňuje typický krabicový graf, kde silná čára je medián, krabice kolem ní určuje polohu prvního a třetího kvartilů (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 definovaný jako hodnoty ležící ve větší vzdálenosti od krabice než 1,5 × IQR, kde IQR je mezikvartilové rozpětí (*interquartile range*) neboli $Q_3 - Q_1$.

```
boxplot(x)
```



Obrázek 4: Boxplot

2.4 Sloupcový graf

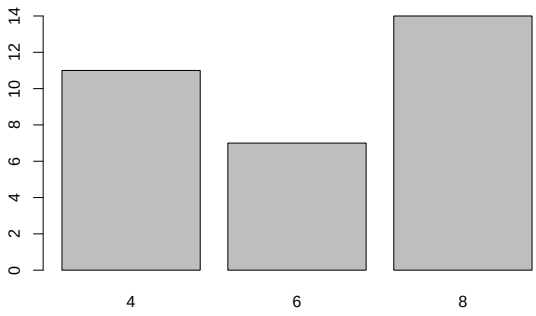
Sloupcový graf je jedním z nejvíce používaných způsobů vizualizace dat. Obvykle se používá pro zobrazení kvantitativních hodnot na ose y a kvalitativních na ose x. Výška sloupců může reprezentovat jak četnosti výskytu hodnot, tak i samotné hodnoty.[1]

V R lze tento typ grafu vykreslit pomocí funkce `barplot()`. V příkladu (Obrázek 5) je použit data set `mtcars`, konkretně atribut `cyl` - počet válců v motoru.

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

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

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



Obrázek 5: Ukázka jednoduchého sloupcového grafu

2.4.1 Histogram

Sloupcový graf s četnostmi na souvislé ose je taky známý jako histogram.[1] 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.[6] 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 [5]

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

$$k = [log_2(n)] + 1$$

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

2. Scott [5]

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

Scotovo pravidlo je následující:

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

kde σ je směrodatná odchylka a h je předpokládaná šířka intervalu.

Počet intervalů může být vypočítán pomocí vztahu:

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

Případně oba vztahy lze shrnout do jednoho:

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

3. Freedman–Diaconis [4]

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

Freedman–Diaconisovo pravidlo pro stanovení předpokládané šířky intervalu je:

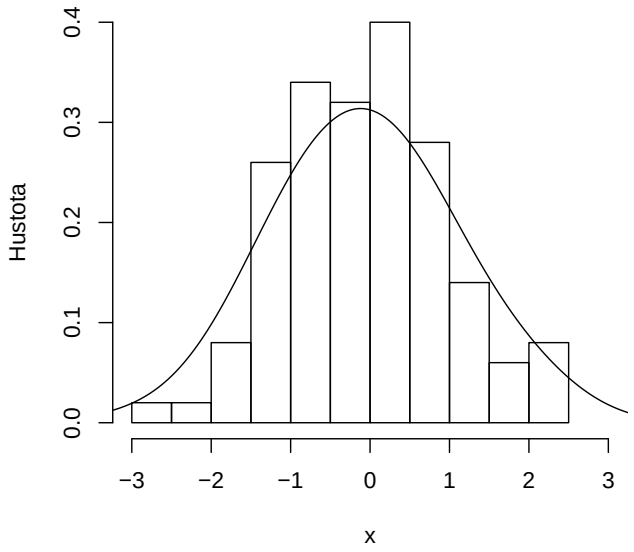
$$h = 2\frac{IQR(x)}{n^{\frac{1}{3}}}$$

Po dosazení:

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

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

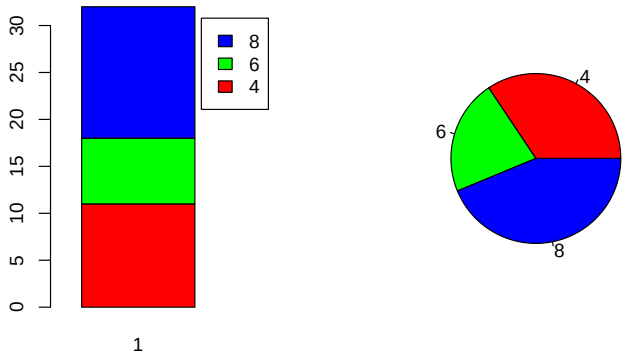
Histogram je jedním ze standardních způsobů, používaných k odhadu tvaru rozdělení, přesto se ale tento způsob považuje za nepřesný, vzhledem k ovlivnění tvaru počtem použitých intervalů. Při normálním rozdělení by měl histogram mít zvoncovitý tvar schodný s Gaussovou křivkou (Obrázek 6).



Obrázek 6: Histogram s odhadem hustoty pravděpodobnosti

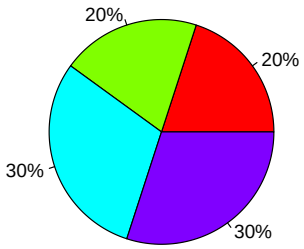
2.4.2 Koláčový graf

Koláčový graf představuje plný kruh (360°), který je rozdělen na jednotlivé výseče pro znázornění číselných proporci mezi proměnnými. Koláčový graf je tvořen transformaci skládaného sloupcového grafu do polárního souřadnicového systému (Obrázek 7). [12]



Obrázek 7: Skládaný sloupcový graf transformovaný do polárního souřadnicového systému

Jednoduché koláčové grafy se vykreslují pomoci funkci `pie()` (Obrázek 8).



Obrázek 8: Ukázka jednoduchého koláčového grafu

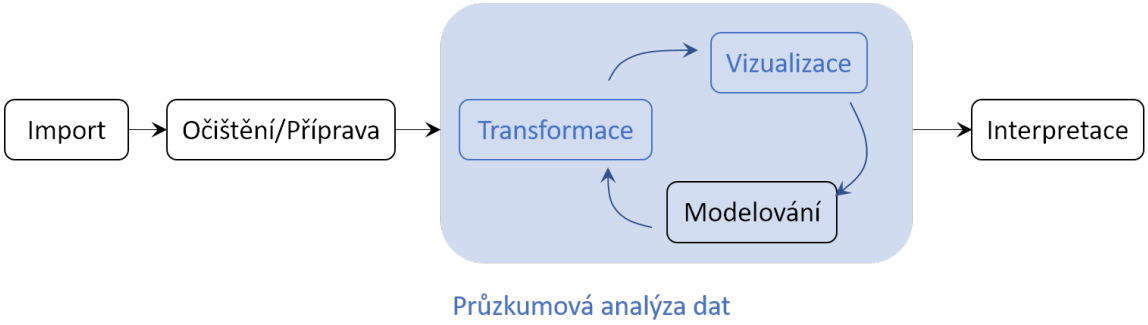
2.4.3 Číslicový histogram (*stem-and-leaf*)

Číslicový histogram je jinak známy jako *stem-and-leaf plot* a podobně histogramu pomáhá vizualizovat tvar rozdělení. Jedná se spíš o historický typ grafu, populární v osmdesátých letech, kdy vykreslování velkých datasetu bylo obtížnější. Vstupní údaje se rozdělí na dva sloupce rozdělené vertikální linii. Pravý sloupec obsahuje listy (*leaf*) - poslední číslici čísla či číslo po desetinné čárce a levý sloupec obsahuje stonek (*stem*) - zbylé číslice (čísla před desetinnou čárkou). Každý stonek se uvádí pouze jednou, aniž by se nějaký vynechal i kdyby to znamenalo, že nebude mít žádné listy. Listy se uvádějí v rostoucím pořadí. Tak v příkladu uvedeným níže, v prvním řádku stonkem je číslice -2, listy jsou číslice 9 a 2 a tak víme že v datasetu se vyskytli čísla -2.9 a -2.2. Tento typ grafu v prostředí R se vykresluje pomoci funkci `stem()`:

```
stem(x)
```

```
##
##   The decimal point is at the |
##
##   -2 | 92
##   -1 | 887755333332211100
##   -0 | 998888776666666555544333332111100
##    0 | 00111222223333444445677778888899
##    1 | 0233445689
##    2 | 0012
```

3 Průzkumová analýza dat



Obrázek 9: 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 jejích kvality. Průzkumová analýza dat není specifikována jako konkretní soubor pravidel a postupu. EDA je důležitou části datové analýzy a měla by být jedním z jejích prvních kroků. [3]

Zařazení průzkumové analýzy dat do procesu datové analýzy je zobrazeno v diagramu 9. Prvním krokem datové analýzy je **import** dat. Obecně v tomto případě to znamená nahrání obdržených dat ze souboru či databaze do prostředí R. Bez tohoto kroku datová nalýza nemůže být vykonána. V momentě když data jsou importována do R je dobře je **očistit** neboli **přípravit**. Tím je myšleno ukládání dat v kontzístentní a systematické formě, odpovídající semantice původního datasetu. Zkratka očištěné data jsou taková data, ve kterých sloupce odpovídají proměnným a řadky odpovídají pozorováním. Taková příprava dat usnadňuje další práci s nimi. Jakmile data jsou očištěná obvýklým krokem je jejich **transformace**. Transformaci se rozumí omezení pozorování (například dle zajmového území), vytvaření nových proměnných, na základě již existujících, agregace a výpočet souhrných statistik. Po tom co jsou data očištěná a obsahují veškeré potřebné proměnné je možné na ně aplikovat dva nejdůležitěší nástroje k zjištění informaci: **vizualizaci** a **modelování**. Tyto nástroje mají svoje výhody a nevýhody a jakákoliv skutečná analýza se na ně opakováně obrácí. [3]

Fisher & Tukey [11]

3.1 Odlehlá pozorování

Odlehlá pozorování (*outliers*) jsou pozorování, která jsou vzdálená vůči ostatním hodnotám datasetu. Definice toho, jak moc vzdálené takováto pozorování mají být je na analytikovi, který to určí na základě pozorování daných konkretním datasetem a kontextu problematiky. Tato pozorování mohou být indikátorem chybných dat

nebo vzácných událostí a mělo by být pečlivě zkoumány proč se vyskytli a jak moc ovlivňují zájmový dataset.

Hledání odlehlých, vybočujících pozorování a jiných anomálií datasetu grafický pro jednotlivé veličiny se dá udělat pomoci boxplotu (viz. sekce 2.3.2), bodových grafů (2.1) nebo číslicových histogramů (2.4.3). Dají se také vypočítat pomocí různých statistik, například metoda *Jacknife*, která bude popsána v následující kapitole.

V momentech, kdy je vizualizace obtížná (velké datasety, větší množství proměnných, atd.), využivají se nástroje vícerozměrné, například Mahalanobisovy vzdalenosti (*Mahalanobis distance*), Leverages, OLS ...

3.1.1 *Jacknife*

3.1.2 Mahalanobisovy vzdalenosti

3.1.3 Leverages

3.1.4 OLS

3.2 Náhrada chybějících pozorování

3.3 Transformace dat

3.4 Změna typu proměnných

3.5 Ověřování normality

Praktická část

4 Praktická vizualizace dat

4.1 Prostředí R

4.1.1 Balíčky

4.1.2 ...

4.2 Balíčky pro vizualizaci dat

4.2.1 ggplot2

4.2.2 lattice

4.2.3 rgl

4.3 Balíčky pro interaktivní vizualizaci dat (htmlwidgets)

4.3.1 plotly

4.3.2 dygraphs

4.3.3 leaflet

4.3.4 ggvis

4.4 Balíčky pro prostorová data

4.4.1 ggmap

4.5 ...

4.5.1 raster

4.5.2 rasterVis

4.6 Balíčky pro webové aplikace

4.6.1 shiny

4.6.2 flexdashboard

4.6.3 dashboard

Literatura

[1] Chang, W. 2012. *R graphics cookbook: Practical recipes for visualizing data*. O'Reilly Media.

[2] Cleveland, W.S. 1994. *The elements of graphing data*. Hobart Press.

[3] Grolemund, G. and Wickham, H. 2016. *R for data science*. Hadley Wickham.

[4] Histogram - wikipedia: [vid. 6.8.2017]. *https://en.wikipedia.org/wiki/Histogram*.

[5] Maciejewski, R. 2011. *Data representations, transformations, and statistics for visual reasoning*. Morgan & Claypool Publishers.

[6] Novovičová, J. 2006. *Pravděpodobnost a matematická statistika*. Praha: Vydavatelství ČVUT,

[7] P–P plot - wikipedia: [vid. 11.8.2017]. *https://en.wikipedia.org/wiki/P%E2%80%93P_plot*.

[8] R: Generic x-y plotting: [vid. 11.5.2017]. *https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html*.

[9] R: The r graphics package: [vid. 22.4.2017]. *https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/graphics-package.html*.

[10] Teetor, P. 2011. *R cookbook: Proven recipes for data analysis, statistics, and graphics*. O'Reilly Media.

[11] Tukey, J.W. 1977. *Exploratory data analysis*. ADDISON WESLEY PUB CO INC.

[12] Wilkinson, L., Wills, D., Rope, D., Norton, A. and Dubbs, R. 2006. *The grammar of graphics*. Springer New York.