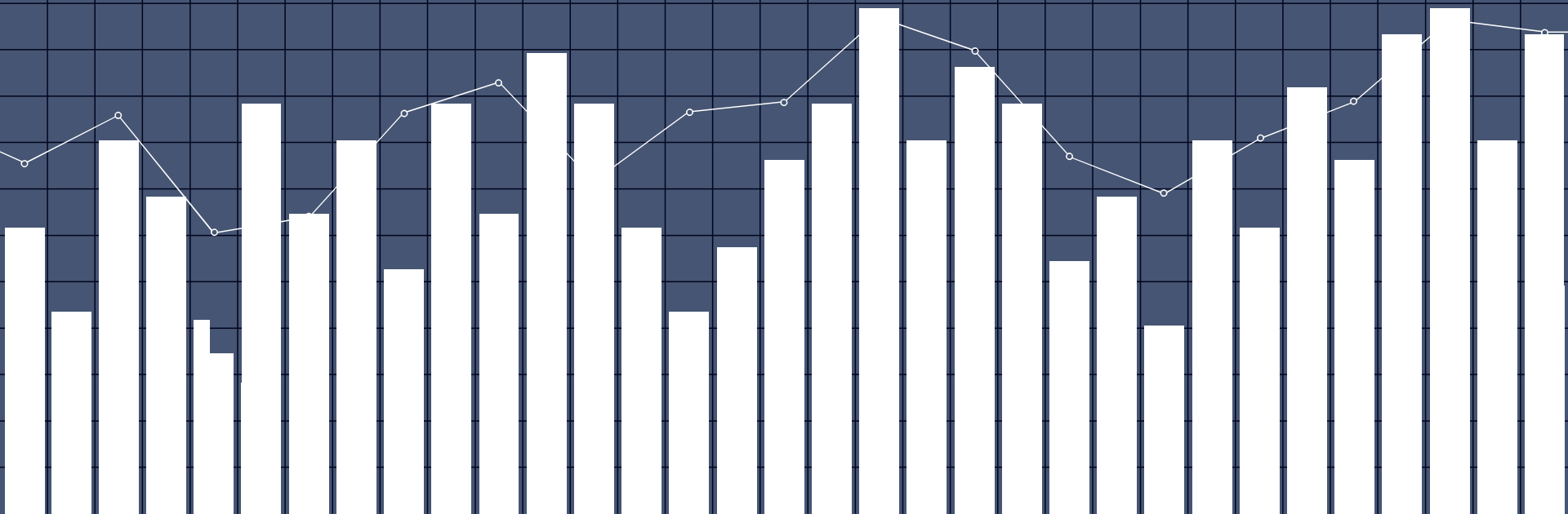


04. Čištění dat



Harmonogram

- 01. Rekapitulace
- 02. Explorace
- 03. Příprava pro analýzu
- 04. Chybějící a odlehlé hodnoty



Čištění dat

Explorace hrubých dat - [base](#)

Matice

```
bmi_1 = read_excel("bmi.xlsx", sheet = 2)
```

Check the class of bmi

```
class(bmi_1)
```

Check the dimensions of bmi

```
dim(bmi_1)
```

View the column names of bmi

```
colnames(bmi_1)
```

Struktura dat

```
str(bmi_1)
```

Glimpse

```
# install.packages("dplyr")
```

```
library(dplyr)
```

```
glimpse(bmi_1)
```

Sumarizace

```
summary(bmi_1)
```

Prvních 10 a posledních 10 řádků

```
head(bmi_1, n = 10)
```

```
tail(bmi_1, n = 10)
```

Čištění dat

Explorace hrubých dat - [psych](#)

```
# Load psych
```

```
install.packages("psych")
```

```
library(psych)
```

```
# Check the structure of bmi, the psych way
```

```
describe(bmi_1)
```

Čištění dat

Explorace hrubých dat - [summarytools](#)

```
# Load summarytools
install.packages("summarytools")
library(summarytools)

# Data
Manpower = read.csv("Manpower.csv")

# Check the structure of bmi, the psych way
view(dfSummary(Manpower))
```

Čištění dat

Příprava dat pro analýzu

```
# Matice
```

```
Infrastructure = read.csv2("Infrastructure.csv")
```

```
# Preview Infrastructure with str()
```

```
str(Infrastructure)
```

```
# Coerce Country to character
```

```
Infrastructure$Country <- as.character(Infrastructure$Country)
```

```
# Coerce Rank to factor
```

```
Infrastructure$Rank <- as.character(Infrastructure$Rank)
```

```
# Look at Infrastructure once more with str()
```

```
str(Infrastructure)
```

Čištění dat

Příprava dat pro analýzu – dílčí manipulace se strings

```
# Load the stringr package
install.packages("stringr")
library("stringr")
```

```
# Trim all leading and trailing whitespace
name = c(" Filip ", "Nick ", " Jonathan")
str_trim(name)
```

```
# Pad these strings with leading zeros
pad = c("23485W", "8823453Q", "994Z")
str_pad(pad, width = 9, side = "left", pad =
"0")
```

```
# Print state abbreviations
Manpower$Country
```

```
# Make states all uppercase and save result
# to states_upper
states_upper =
toupper(Manpower$Country)
states_upper
```

```
# Make states_upper all lowercase again
states_lower = tolower(Manpower$Country)
states_lower
```

Čištění dat

Příprava dat pro analýzu – dílčí manipulace se strings

```
# Look at the head of Infrastructure  
head(Infrastructure)
```

```
# Detect all "Republic" in Country  
str_detect(Infrastructure$Country,  
"Republic")
```

```
# In the Country column, replace "Republic"  
with "R"...
```

```
Infrastructure$Country <-  
str_replace(Infrastructure$Country,  
"Republic", "R")
```


Čištění dat

Příprava dat pro analýzu – chybějící data

- In R, represented as NA

- May appear in other forms

- #N/A (Excel)

- Single dot (SPSS, SAS)

- Empty string

- character: "treatment", "123", "A"

- numeric: 23.44, 120, NaN, Inf

- [integer](#): 4L, 1123L

- factor: factor("Hello"), factor(8)

- logical: TRUE, FALSE, NA

- Inf - "Infinite value" (indicative of outliers?)

- 1/0

- 1/0 + 1/0

- 33333^33333

- NaN - "Not a number" (rethink a variable?)

- 0/0

- 1/0 - 1/0

Čištění dat

Příprava dat pro analýzu – chybějící data

```
name = c("Jerry", "Beth", "Rick", "Morty")  
n_friends = c(NaN, NA, Inf, 2)  
status = c("Listening to human music", "Happy Family", "Garage", "")  
social_df = data.frame(cbind(name, n_friends, status))
```

```
# Call is.na() on the full social_df to spot all NAs  
is.na(social_df)
```

```
# Use the any() function to ask whether there are any NAs  
# in the data  
any(is.na(social_df))
```

```
# View a summary() of the dataset  
summary(social_df)
```

```
# Call table() on the status column  
table(social_df$status)
```

```
# Replace all empty strings in status with NA  
social_df$status[social_df$status == ""] <- NA
```

```
# Print social_df to the console  
social_df
```

```
# Use complete.cases() to see which rows have no missing values  
complete.cases(social_df)
```

```
# Use na.omit() to remove all rows with any missing values  
na.omit(social_df)
```

Čištění dat

Odlehlé hodnoty – explorace grafy

Matice

```
Infrastructure = read.csv2("Infrastructure.csv")
```

Histogram

```
hist(Infrastructure$Ports)
```

Boxplot

```
boxplot(Infrastructure$Airports)
```

Scatterplot

```
plot(Infrastructure$Railway_Coverage, Infrastructure$Roadway_Coverage)
```