Exercise Sheet #3

Fortgeschrittene Statistische Software für NF

Minh Son Tran (12424799)

‘2025-06-05

Contents


```
#install.packages("tidyverse")
#install.packages("palmerpenguins")
#install.packages("dplyr")
#install.packages("knitr")
#install.packages("easystats")
#install.packages("stargazer")
#install.packages("margins")
#install.packages("reticulate")
#install.packages("effects")
#install.packages("carData")
#install.packages("sf")
library(readr)
library(tidyverse)
library(palmerpenguins)
library(dplyr)
library(knitr)
library(ggplot2)
library(report)
library(parameters)
library(see)
library(reticulate)
library(margins)
library(lattice)
library(effects)
library(sf)
```

0.1 Exercise 1

0.1.1 d)

Git’s Strengths and Weaknesses:

2 strengths:

- It allow user to create branches, makes it easy to experiment without affecting the main codebase.
- It is open source and actively maintained.

2 weaknesses:

- User requires time to learn how to use Git properly because of confusing commands.
- Git is not very good for project with binary or very large files

0.2 Exercise 2

0.2.1 a)

GitHub repo: https://github.com/MinhSonTran97/exeRcise-sheet-3.git

0.3 Exercise 3

0.3.1 a)

```
pixar_films <- read_csv("data/pixar_films.csv") %>%
  filter(!is.na(film))
```

```
## Rows: 27 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): film, film_rating
## dbl  (2): number, run_time
## date (1): release_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
pixar_films <- pixar_films %>%
  mutate(rating_factor = factor(film_rating))
```

The possible film_rating values are G, PG, N/A:

- "G": General audiences
- "PG": Parental guidance is advised
- "N/A": Film Rating is not available / no rating provided. This is included since we only removed entries with missing titles

Converting film_rating into rating_factor is appropriate because:

- film_rating represents a set of discrete categories, not numeric values. These categories describe the type of audience the film is suitable for, not a measurable quantity.
- Treating film_rating as a factor helps R recognize it as a categorical variable, which allow us to better summarize, plot, and handle the data in modeling.

0.3.2 b)

```
film_series <- pixar_films %>%
  filter(str_detect(film,"Toy Story|Cars|Incredibles|Finding|Monsters")) %>%
  mutate(series = case_when(
    str_detect(film,"Toy Story") ~ "Toy Story",
    str_detect(film,"Cars") ~ "Cars",
    str_detect(film,"Incredibles") ~ "Incredibles",
    str_detect(film,"Finding") ~ "Finding",
    str_detect(film,"Monsters") ~ "Monsters"
  )) %>%
  group_by(series) %>%
  summarise(films = paste(film, collapse = " - "), number_of_entries = n())
film_series
```

```
## # A tibble: 5 x 3
##   series      films                                          number_of_entries
##   <chr>       <chr>                                                      <int>
## 1 Cars        Cars - Cars 2 - Cars 3                                         3
## 2 Finding     Finding Nemo - Finding Dory                                    2
## 3 Incredibles The Incredibles - Incredibles 2                                2
## 4 Monsters    Monsters, Inc. - Monsters University                           2
## 5 Toy Story   Toy Story - Toy Story 2 - Toy Story 3 - Toy Sto~               4
```