Slovenska Technická Univerzita

Fakulta Informatiky a Informačných Technológii

DSP Zadanie 1

Vývoj počtu hráčov platformy STEAM vzhľadom na vývoj pandémie

Peter Plevko, Marek Adamovič

Cvičenie: Piatok 10:00 – 11:40

doc. Mgr. Michal Kováč, MSc., PhD. 2020/2021

Datasety

Pre našu prácu sme si vybrali 2 datasety, s ktorými ideme pracovať. Jedným je dataset hráčov na platforme STEAM a druhým je covid dataset o denných prírastkoch nakazených a denných prírastkoch úmrtí v jednotlivých krajinách.

STEAM dataset

Dataset obsahujúci priemerný počet hráčov za mesiac pre 100 najhranejších hier na platforme STEAM. Bol získaný pomocou služby https://steamcharts.com/top.
Link na dataset:
https://www.kaggle.com/jackogozaly/steam-player-data

```
> data_steam
# A tibble: 5,271 x 8
   Month_Year     Avg_players      Gain Percent_Gain Peak_Players URL                    Date       Game_Name
   <chr>                <dbl>     <dbl> <chr>              <dbl> <chr>                  <date>     <chr>
 1 September 2021     512351.     269.  +0.05%            942519 steamcharts.com/app/730 2021-09-01 Counter Strike
 2 August 2021        512082.    6015.  +1.19%            802544 steamcharts.com/app/730 2021-08-01 Counter Strike
 3 July 2021          506067.  -43280.  -7.88%            763523 steamcharts.com/app/730 2021-07-01 Counter Strike
 4 June 2021          549347. -110542.  -16.75%           929940 steamcharts.com/app/730 2021-06-01 Counter Strike
 5 May 2021           659889.  -63458.  -8.77%           1087197 steamcharts.com/app/730 2021-05-01 Counter Strike
 6 April 2021         723347.  -17581.  -2.37%           1148077 steamcharts.com/app/730 2021-04-01 Counter Strike
 7 March 2021         740928.     -85.4 -0.01%           1198581 steamcharts.com/app/730 2021-03-01 Counter Strike
 8 February 2021      741013.   -2196.  -0.30%           1123485 steamcharts.com/app/730 2021-02-01 Counter Strike
 9 January 2021       743210.   25406.  +3.54%           1124553 steamcharts.com/app/730 2021-01-01 Counter Strike
10 December 2020      717804.   49049.  +7.33%           1164396 steamcharts.com/app/730 2020-12-01 Counter Strike
# ... with 5,261 more rows
```

Počet záznamov a atribúty

STEAM dataset obsahuje 5 271 záznamov s 8 nasledujúcimi atribútmi:

- Month_Year, typ <char>, udáva dátum, kedy bol záznam nameraný, vo formáte mesiac(názvom), rok
- Avg_players, typ <dbl>, hovorí o priemernom počte hráčov danej hry v daný mesiac
- Gain, typ <dbl>, uvádza rozdiel medzi priemerným počtom hráčov tohto a minulého mesiaca
- Percent_Gain, typ <chr>, vyjadruje stĺpec Gain vo formáte počet percent %
- Peak_Players, typ <dbl>, hovorí o najvyššom počte hráčov v jeden deň počas daného mesiaca
- URL, typ <chr>, uvádza link na stránku grafu danej hry
- Date, typ <date>, udáva dátum, kedy bol záznam nameraný, vo formáte YYYY-MM-DD
- Game_Name, typ <chr>, udáva názov hry

Covid dataset

Tento dataset obsahuje 218 krajín. Každá krajina má záznamy začinajúce 2020-2-15 a končiace 2021-07-31 to znamená máme záznamy o dĺžke 532 dní. Máme aj krajinu v ktorej disponujeme záznamami zo skoršieho dátumu a touto krajinou je Čína, krajina v ktorej coronavirus začal. V týchto záznamoch sa dozvieme informácie o priebehu covidu v každej z krajín.
Link na dataset:

https://www.kaggle.com/josephassaker/covid19-global-dataset

```
# A tibble: 10 x 7
   date       country  cumulative_total_cases daily_new_cases active_cases cumulative_total_deaths daily_new_deaths
   <chr>      <chr>                     <dbl>           <dbl>        <dbl>                  <dbl>             <dbl>
 1 2021-7-21 Zimbabwe                  91120            2705        28684                   2809                62
 2 2021-7-22 Zimbabwe                  93421            2301        28828                   2870                61
 3 2021-7-23 Zimbabwe                  95686            2265        29739                   2961                91
 4 2021-7-24 Zimbabwe                  97277            1591        29599                   3050                89
 5 2021-7-25 Zimbabwe                  97894             617        28887                   3094                44
 6 2021-7-26 Zimbabwe                  99944            2050        28944                   3173                79
 7 2021-7-27 Zimbabwe                 101711            1767        27935                   3280               107
 8 2021-7-28 Zimbabwe                 103567            1856        28844                   3340                60
 9 2021-7-29 Zimbabwe                 105656            2089        28841                   3421                81
10 2021-7-30 Zimbabwe                 107490            1834        29438                   3490                69
```

Počet záznamov a atribúty

Korona dataset obsahuje 117 064 záznamov so 7 stĺpcami. Týmito stĺpcami sú:

- date, typ <chr>, určuje dátum pozorovania vo formáte YYYY-MM-DD
- country, typ <chr>, určuje krajinu v ktorej boli namerané dáta daného stĺpca
- cumulative_total_cases, typ <dbl>, určuje suhrnny počet potvrdených prípadov
- daily_new_cases, typ <dbl>, určuje počet denných potvrdených prípadov
- active_cases, typ <dbl>, určuje počet aktívnych prípadov ochorenia
- cumulative_total_deaths, typ <dbl>, určuje suhrnny počet potvrdených prípadov
- daily_new_cases, typ <dbl>, určuje počet denných potvrdených úmrti

```
#########################DESCRIPTIVE#########################
tail(data_covid)
nrow(data_covid)
ncol(data_covid)
map(data_covid, typeof)
sum(is.na(data_covid))

tail(data_steam)
nrow(data_steam)
ncol(data_steam)
map(data_steam, typeof)
sum(is.na(data_steam))
#############################################################
```

Transformácia dát

Našim prvým krokom bolo načítanie si knižníc.

```
#########################LIBRARIES###########################
library(tidyverse)
library(ggplot2)
library(readxl)
library(lubridate)
#############################################################
```

Následne sme si načítali do premenných naše datasety.

```
#######################READ_FILES#########################
data_covid <- read_csv("covid.csv", col_names = T)
data_steam <- read_csv(file = "players.csv", col_names = T)
##########################################################
```

Následne sme išli pracovať s jednotlivými datasetmi predtým ako sme ich išli spojiť. Pri covid datasete sme mali dátumy meraní pre jednotlivé dni v jednotlivých krajinách a nie pre celé mesiace, ako to bolo pri STEAM datasete. Keďže sme datasety chceli spojiť do jedného, museli sme zjednotiť tento formát. Chceli sme jeden záznam pre jeden mesiac. Postupovali sme následovne, najskôr sme groupli dataset podľa dátumov, čím sme zbavili viacerých meraní (v rôznych krajinách) pre jeden deň (dané hodnoty sme sčítali pomocou funkcie sum(), aby sa nám žiaden pozitívny prípad nestratil). Následne sme každému dátumu priradili deň 1, vďaka čomu sme vedeli záznamy znovu groupnúť podľa dátumu (znovu sme využili funkciu sum()), čím nám vznikla jedinečná hodnota pre jeden mesiac daného roka.

```
#######################COVID_DATASET######################
covid <- data_covid %>%
  group_by(date) %>%
  summarise(daily_new_cases = sum(daily_new_cases, na.rm = TRUE),
            daily_new_deaths = sum(daily_new_deaths, na.rm = TRUE))
covid$date <- as.Date(covid$date)
day(covid$date) <- 1

covid <- covid %>%
  group_by(date) %>%
  summarise(daily_new_cases = sum(daily_new_cases, na.rm = TRUE),
            daily_new_deaths = sum(daily_new_deaths, na.rm = TRUE)) %>%
  arrange(date)
##########################################################
```

Pri STEAM datasete sme mali podobný problém. Pre každý mesiac sme mali za každú jednu hru jeden záznam. Keďže sme chceli globálny záznam hráčov pre celý mesiac, groupli sme dataset podľa dátumu, čím sme získali jednu hodnotu pre každý mesiac (záznamy z jednotlivých hier sme znovu sčítali pomocou funkcie sum()). Ďalším problémom bol fakt, že v datasete STEAMu sme mali formát dátumu typu <chr>, ktorý mal mesiac zadaný ako string. Keďže formát <date> vyžaduje, aby bol zadaný aj deň, tak každému záznamu sme dali deň 1, čo sa aj zhodovalo s formátom pri covid datasete.

```
#######################STEAM_DATASET######################
steam <- data_steam %>%
  select(Month_Year, Avg_players, Peak_Players) %>%
  group_by(Month_Year) %>%
  summarise(Avg_players = sum(Avg_players, na.rm = TRUE),
            Peak_Players = sum(Peak_Players, na.rm = TRUE))
steam$date <- paste(sapply(strsplit(steam$Month_Year, " "),"[[", 2),
  match(sapply(strsplit(steam$Month_Year, " "),"[[", 1), month.name), 1, sep = "/")
steam <- steam %>%
  select(date, Avg_players, Peak_Players)
steam$date <- as.Date(steam$date)
##########################################################
```

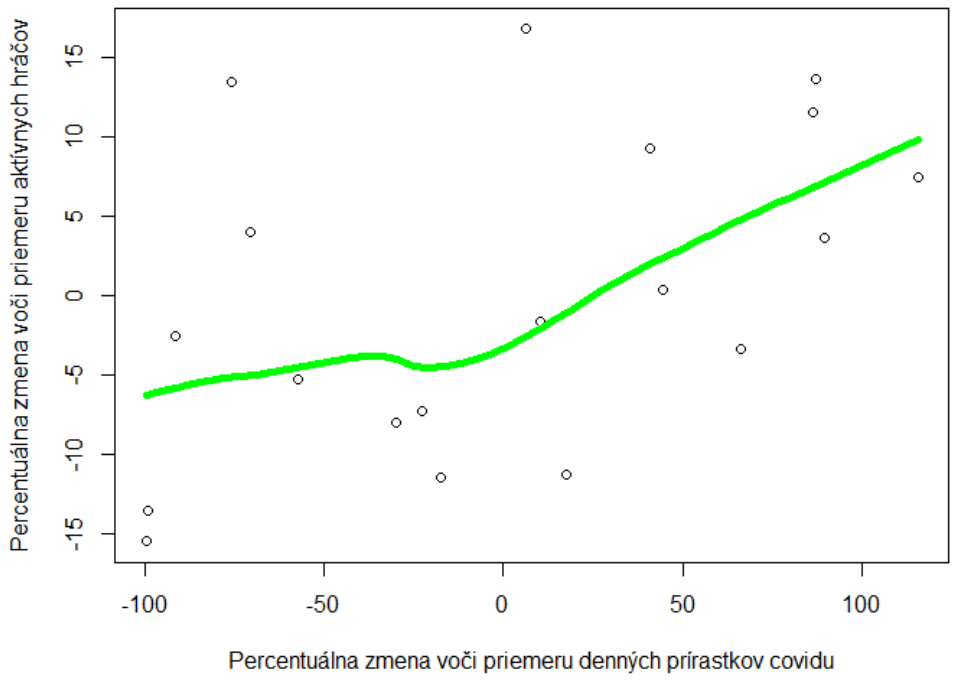
Následne sme spojili tabuľky covid a steam podľa dátumu. Potom sme pomocou výpočtov vytvorili nové stĺpce avrt_chng, avrg_prcntg_chng a daily_new_cases_milions. Pre ukážku ako vyzerajú dáta sme ich následne zobrazili na grafe pomocou knižnice ggplot.

```
#######################RESULTS_GRAPHS#####################
result <- merge(covid, steam, "date")

avrg <- mean(result$Avg_players)
result$avrg_chng <- result$Avg_players - avrg
result$avrg_prcntg_chng <- round((result$Avg_players / avrg - 1) * 100, 2)

result$daily_new_cases_milions = result$daily_new_cases / 1000000

ggplot(data = result, aes(date, avrg_prcntg_chng)) + geom_col() +
  scale_x_continuous(breaks = seq(date("2020-01-01"), by="3 months", length.out=7))
ggplot(data = result, aes(date, daily_new_cases_milions)) + geom_col() +
  scale_x_continuous(breaks = seq(date("2020-01-01"), by="3 months", length.out=7))
##########################################################
```

Na záver sme testovali hypotézu, či nárast počtu covid prípadov spôsobuje nárast počtu hráčov na platforme STEAM.

```
#######################HYPOTHESIS#########################
hypo_data <- result %>%
  select(date, player_avrg_prcntg_chng, covid_avrg_prcntg_chng)

half <- map(1:9, ~ hypo_data[sort(sample(1:dim(hypo_data)[1], size = 0.5*dim(hypo_data)[1])),])

models <- map(half, ~ lm(.x$player_avrg_prcntg_chng ~ .x$covid_avrg_prcntg_chng))

listOfFunctions <- list(coefficients = coef, residuals = residuals)
f <- function(x) {sapply(listOfFunctions, function(g) g(x))}
extractedData <- map(models, ~ f(.x))

sd(map_dbl(models, ~ coef(.x)[1]))
sd(map_dbl(models, ~ coef(.x)[2]))

rss <- map_dbl(models, ~ sum(resid(.x)^2))
rse <- map_dbl(rss, ~ sqrt(.x/(0.5*dim(hypo_data)[1]-2)))
boxplot(rss)
boxplot(rse)

cfs <- map_dbl(models, ~ coef(.x)[2])
t.test(cfs, mu=0)

scatter.smooth(hypo_data$covid_avrg_prcntg_chng, hypo_data$player_avrg_prcntg_chng,
               xlab = "Percentuálna zmena voči priemeru denných prírastkov covidu",
               ylab = "Percentuálna zmena voči priemeru aktívnych hráčov",
               lpars = list(col = "green", lwd = 5, lty = 7))
##########################################################
```

Na tomto grafe vidíme, že pri stúpaní počtu nakazených covidom, stúpa taktiež počet hráčov na platforme STEAM (% zmena oproti priemeru).



Záver

Vďaka tomuto projektu sme sa naučili ako analyzovať dáta pomocou jazyka R. V tomto zadaní sme si aj vyskúšali prácu s dvomi zaujímavými datasetmi a následne úpravu ich hodnôt. Po úprave ich hodnôt sme tieto datasety spojili a používali hodnoty na potvrdenie našej hypotézy. Po stanovení hypotézy a vytvorení modelov sme aplikovali cross validačný proces. Nakoniec sme zobrazili grafy a zistili sme, že nami zvolená hypotéza sa potvrdila.