RESPONSI_DS-(C)

Muhammad Dzaki_123200102

Intro

1.Kerjakan soal-soal yang ada! Jangan lupa AUTHOR diberi nama (pada bagian atas soal ini) 2. Boleh menggunakan PC lab / Laptop pribadi 3. Pengumpulan berupa hasil knit Rmd ke pdf dengan nama NIM_NAMA_RESPONSI_IF-C.pdf. 4. Durasi 2 Jam + 5 menit submit, > tidak bisa mengumpul. 5. Pengerjaan offline, pengumpulan di Spada (online). 6. Tidak boleh buka modul. 7. Tidak boleh membuka internet (googling, WhatsApp, ig, sosmed, dan media komunikasi lain). 8. Boleh bawa catatan 1 lembar A4. 9. Izin keluar maks. 1x dengan durasi 2 menit. 10. Tas dan HP diletakkan di depan. 11. Isi juga review/feedback/kritik/saran/masukan yang sudah disediakan di bagian paling bawah soal. ***WAJIB***

Persiapan

Load library apa saja yang kira-kira digunakan! Lalu load dataset 'googleplay.csv' dan 'google-play_user_review.csv'!

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.1.0     v tune         1.0.1
## v infer        1.0.4     v workflows    1.1.2
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.3     v yardstick    1.1.0
## v recipes      1.0.3
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
```

```
## x recipes::step()   masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(tidytext)
library(vroom)
library(here)
```

```
## here() starts at /Users/macpro/Documents/GitHub/Prak-DS
```

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
ggplaystore = vroom(
  here("/Users/macpro/Documents/GitHub/Prak-DS/Responsi/googleplaystore.csv")
  )
```

```
## Rows: 8196 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): App, Category, Size, Installs, Type, Price, Content Rating, Genres...
## dbl  (2): Rating, Reviews
##
## 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.
```

```
ggplaystore_review = vroom(here(
  "/Users/macpro/Documents/GitHub/Prak-DS/Responsi/googleplaystore_user_reviews.csv"
  ))
```

```
## Rows: 64295 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): App, Translated_Review, Sentiment
## dbl (2): Sentiment_Polarity, Sentiment_Subjectivity
##
## 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.
```

1. Tampilkan TOP 10 Aplikasi berdasarkan peringkat RATING yang diberikan user!

```
ggplaystore %>% arrange(desc(Rating)) %>% select(App, Rating) %>% head(10)
```

```
## # A tibble: 10 x 2
##    App                                          Rating
##    <chr>                                         <dbl>
##  1 Hojiboy Tojiboyev Life Hacks                      5
##  2 American Girls Mobile Numbers                     5
##  3 Awake Dating                                      5
##  4 Spine- The dating app                             5
##  5 Girls Live Talk - Free Text and Video Chat        5
##  6 Online Girls Chat Group                           5
##  7 Speeding Joyride & Car Meet App                   5
##  8 SUMMER SONIC app                                  5
##  9 Prosperity                                        5
## 10 Mindvalley U Tallinn 2018                         5
```

2. Tampilkan TOP 10 Aplikasi berdasarkan banyaknya REVIEWS secara urut dari yang terbesar!

```
ggplaystore %>% arrange(desc(Reviews)) %>% select(App, Reviews) %>% head(10)
```

```
## # A tibble: 10 x 2
##    App                                               Reviews
##    <chr>                                               <dbl>
##  1 Facebook                                         78158306
##  2 WhatsApp Messenger                               69119316
##  3 Instagram                                        66577313
##  4 Messenger <U+0096> Text and Video Chat for Free  56642847
##  5 Clash of Clans                                   44891723
##  6 Clean Master- Space Cleaner & Antivirus          42916526
##  7 Subway Surfers                                   27722264
##  8 YouTube                                          25655305
##  9 Security Master - Antivirus, VPN, AppLock, Booster 24900999
## 10 Clash Royale                                     23133508
```

3. Tampilkan TOP 10 Aplikasi berdasarkan banyaknya unduhan, dan tampilkan secara urut berdasarkan rating! Clue : data preprocessing

```
top = ggplaystore %>% arrange(Installs)
top10 = top[1:10,]
arrange(top10, desc(Rating)) %>% select(App, Installs, Rating) %>% head(10)
```

```
## # A tibble: 10 x 3
##    App                                               Installs       Rating
##    <chr>                                             <chr>           <dbl>
##  1 Subway Surfers                                    1,000,000,000+    4.5
##  2 WhatsApp Messenger                                1,000,000,000+    4.4
##  3 Google Chrome: Fast & Secure                      1,000,000,000+    4.3
##  4 Gmail                                             1,000,000,000+    4.3
##  5 Google Play Games                                 1,000,000,000+    4.3
##  6 Skype - free IM & video calls                     1,000,000,000+    4.1
##  7 Facebook                                          1,000,000,000+    4.1
##  8 Messenger <U+0096> Text and Video Chat for Free 1,000,000,000+    4
##  9 Hangouts                                          1,000,000,000+    4
## 10 Google Play Books                                 1,000,000,000+    3.9
```

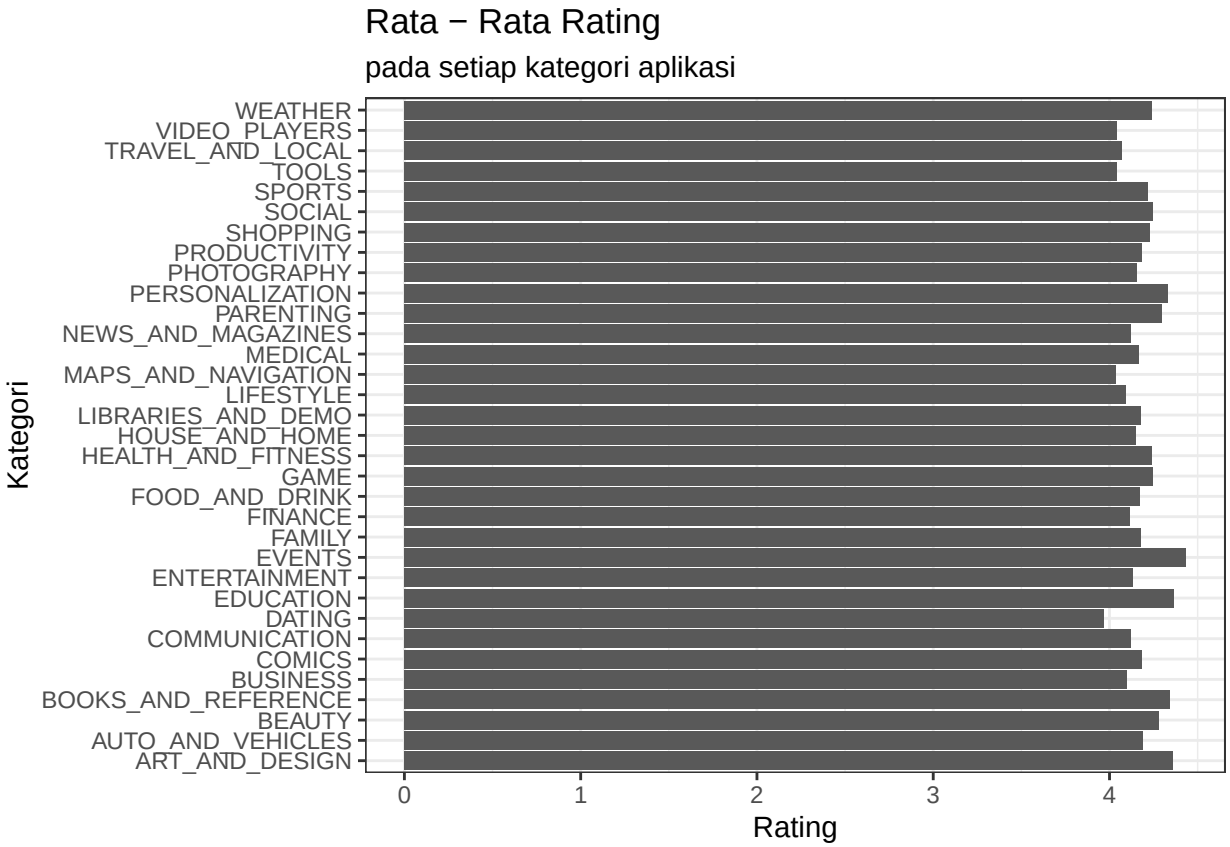
4. Tampilkan rata-rata RATING yang dihitung menggunakan fungsi buatan untuk setiap kategori aplikasi!

```
rerata = ggplaystore %>% group_by(Category) %>%
  summarize(mean_Rating = mean(Rating))
rerata
```

```
## # A tibble: 33 x 2
##    Category            mean_Rating
##    <chr>                     <dbl>
##  1 ART_AND_DESIGN             4.36
##  2 AUTO_AND_VEHICLES          4.19
##  3 BEAUTY                     4.28
##  4 BOOKS_AND_REFERENCE        4.34
##  5 BUSINESS                   4.10
##  6 COMICS                     4.18
##  7 COMMUNICATION              4.12
##  8 DATING                     3.97
##  9 EDUCATION                  4.36
## 10 ENTERTAINMENT              4.14
## # ... with 23 more rows
```

5. Berdasarkan soal nomor 4, buat plot untuk memvisualisasikan hasilnya! (Bentuk plot bebas)

```
ggplaystore %>% group_by(Category) %>%
  summarize(mean_Rating = mean(Rating)) %>%
  ggplot(aes(x = mean_Rating, y = Category)) + geom_col() + labs(
  x = "Rating",
  y = "Kategori",
  title = "Rata - Rata Rating",
  subtitle = "pada setiap kategori aplikasi"
) + theme_bw()
```



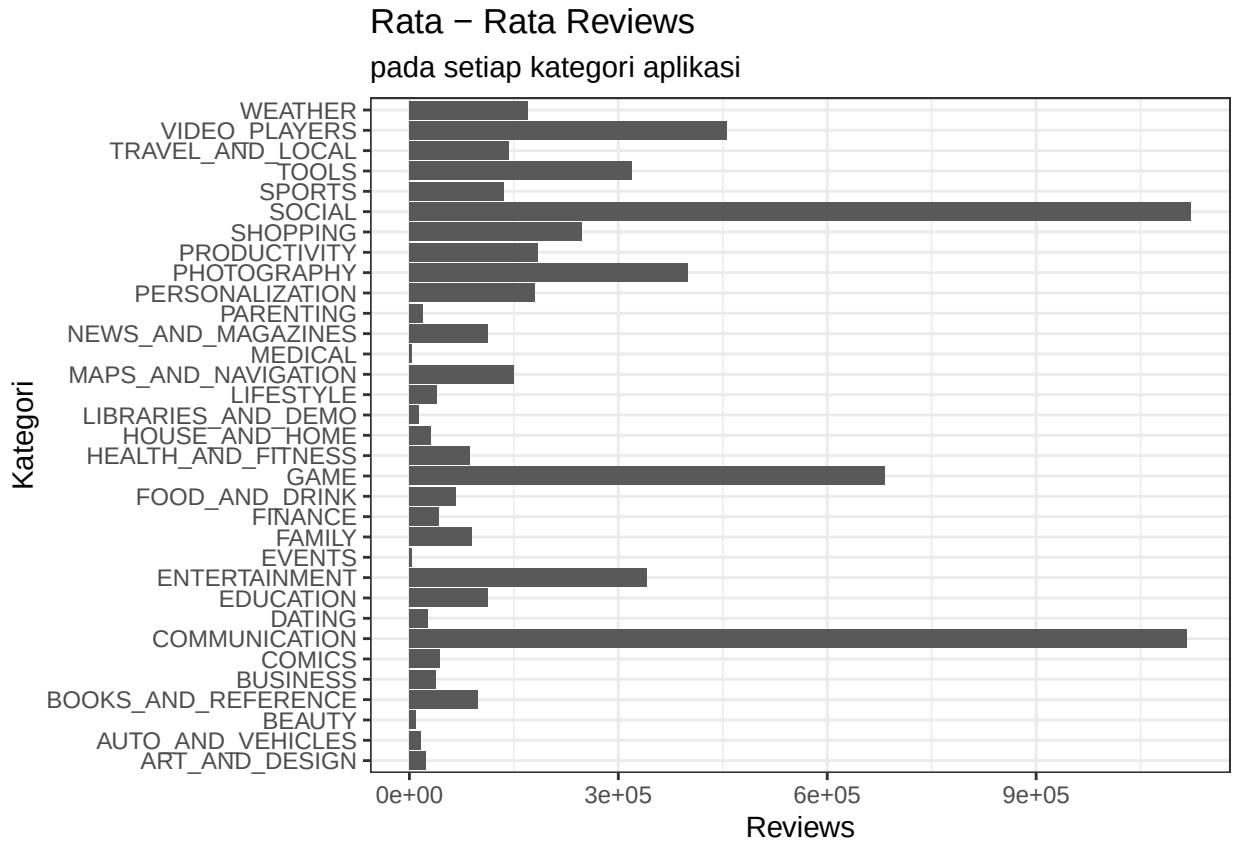
6. Tampilkan rata-rata REVIEWS yang dihitung menggunakan fungsi buatan untuk setiap kategori aplikasi!

```
rerata_review = ggplaystore %>% group_by(Category) %>%
  summarize(mean_Review = mean(Reviews))
rerata_review
```

```
## # A tibble: 33 x 2
##    Category            mean_Review
##    <chr>                     <dbl>
##  1 ART_AND_DESIGN           23265.
##  2 AUTO_AND_VEHICLES        15940.
##  3 BEAUTY                    9408.
##  4 BOOKS_AND_REFERENCE      98940.
##  5 BUSINESS                 37604.
##  6 COMICS                   43347.
##  7 COMMUNICATION          1116449.
##  8 DATING                   27040.
##  9 EDUCATION               113249.
## 10 ENTERTAINMENT           340810.
## # ... with 23 more rows
```

7. Berdasarkan soal nomor 6, buat plot untuk memvisualisasikan hasilnya! (Bentuk plot bebas)

```
ggplaystore %>% group_by(Category) %>%
  summarize(mean_Reviews = mean(Reviews)) %>%
  ggplot(aes(x = mean_Reviews, y = Category)) + geom_col() + labs(
  x = "Reviews",
  y = "Kategori",
  title = "Rata - Rata Reviews",
  subtitle = "pada setiap kategori aplikasi"
) + theme_bw()
```



Info untuk 2 soal 8-10:Terdapat dua dataset yang digunakan. Satu dataset untuk info aplikasi dan satu dataset lagi untuk kumpulan reviewnya.

8. Buat satu variable data baru yang isinya NAMA APLIKASI, RATING, dan JUMLAH REVIEW Positif! Tampilkan isi data tabel tersebut!

```
join_data = ggplaystore %>% inner_join(ggplaystore_review)
```

```
## Joining, by = "App"
```

```
join_data = join_data %>% filter(Translated_Review != "nan")

sentimen = join_data %>% group_by(App, Sentiment) %>% tally()
sentimen = dcast(sentimen, App~ Sentiment, fun.sum = length)
```

```
## Using n as value column: use value.var to override.
```

```
sentimen_positif = sentimen %>% inner_join(ggplaystore) %>%
  select(App, Rating, Positive)
```

```
## Joining, by = "App"
```

```
sentimen_positif %>% head(10)
```

```
##                                         App Rating Positive
## 1                       10 Best Foods for You    4.0      162
## 2                                       11st    3.8       23
## 3                    1800 Contacts - Lens Store    4.7       64
## 4                  21-Day Meditation Experience    4.4       68
## 5           2Date Dating App, Love and matching    4.4       26
## 6                    2GIS: directory & navigator    4.5       23
## 7                 2ndLine - Second Phone Number    4.2       17
## 8                                     2RedBeans    4.0       31
## 9  30 Day Fitness Challenge - Workout at Home    4.8       27
## 10                      365Scores - Live Scores    4.6        5
```

9. Buat satu variable data baru yang isinya NAMA APLIKASI, Total REVIEWS, JUMLAH REVIEW Positif, JUMLAH REVIEW Negatif, JUMLAH REVIEW Neutral! Lalu tampilkan isi data tabel tersebut!

```
sentimen[is.na(sentimen)] = 0
sentimen = mutate(sentimen,
                  TotalReviews = sentimen$Positive + sentimen$Negative +
                    sentimen$Neutral)

analisis_sentimen = sentimen %>% inner_join(ggplaystore, by="App") %>%
  select(App, TotalReviews, Positive, Negative, Neutral)
analisis_sentimen %>% head(10)
```

```
##                                         App TotalReviews Positive Negative
## 1                       10 Best Foods for You          194      162       10
## 2                                       11st           39       23        7
## 3                    1800 Contacts - Lens Store           80       64        6
## 4                  21-Day Meditation Experience           80       68       10
## 5           2Date Dating App, Love and matching           38       26        7
## 6                    2GIS: directory & navigator           40       23        6
## 7                 2ndLine - Second Phone Number           40       17        7
## 8                                     2RedBeans           39       31        2
## 9  30 Day Fitness Challenge - Workout at Home           31       27        2
## 10                      365Scores - Live Scores            7        5        0
##    Neutral
## 1       22
## 2        9
## 3       10
## 4        2
## 5        5
## 6       11
## 7       16
```

```
## 8        6
## 9        2
## 10       2
```

10. Dalam dunia data scientist, sebelum melakukan pemodelan ada baiknya data dilakukan preprocessing terlebih dahulu. Dengan dataset review yang sudah dimasukkan oleh user, lakukan sebuah preprocessing data SEDERHANA yang menurut kalian dapat dilakukan untuk dataset tersebut agar dataset bisa siap untuk dimodelkan (simpan hasil preprocessing dalam variabel baru)!

Clue : Clean, Tidy, no redundacy, no dupe, no null.

```
Cleaned_data = ggplaystore %>% inner_join(ggplaystore_review) %>%
  filter(Translated_Review != "nan") %>%
  unnest_tokens(word, Translated_Review) %>% anti_join(stop_words)
```

```
## Joining, by = "App"
## Joining, by = "word"
```

```
Cleaned_data
```

```
## # A tibble: 362,048 x 17
##    App   Categ~1 Rating Reviews Size  Insta~2 Type  Price Conte~3 Genres Last ~4
##    <chr> <chr>    <dbl>   <dbl> <chr> <chr>   <chr> <chr> <chr>   <chr>  <chr>
##  1 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  2 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  3 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  4 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  5 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  6 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  7 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  8 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
##  9 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
## 10 Colo~ ART_AN~    3.9     967 14M   500,00~ Free  0     Everyo~ Art &~ 15-Jan~
## # ... with 362,038 more rows, 6 more variables: `Current Ver` <chr>,
## #   `Android Ver` <chr>, Sentiment <chr>, Sentiment_Polarity <dbl>,
## #   Sentiment_Subjectivity <dbl>, word <chr>, and abbreviated variable names
## #   1: Category, 2: Installs, 3: `Content Rating`, 4: `Last Updated`
```

Kritik/saran/masukan/feedback/review/uneg-uneg: nggak tau, menurut ku udah bagus

========== SELESAI ==========