IMDB Scraping, Cleansing & Visualization

Oleh :
Filbert Utomo (15/383232/PA/16892)
Muammar Khadafi (15/378067/PA/16542)
Muhammad Nadhif Aswan (15/378068/PA/16543)
Muhammad Rizki Hakim (15/383247/PA/16907)
Ruli Sastra Putri (14/364140/PA/15914)
Tio Rahaditya Luthfitama (15/378076/PA/16551)

February 25, 2019

Link Notebook: https://rpubs.com/avosta/472642

Pre-requisites

Install package yang digunakan

```
#install.packages('rvest')
```

Library & Link IMDB

Menggunakan library rvest dan mengambil film yang rilis pada tahun 2018. https://www.imdb.com/search/title?count=250&release_date=2018,2018&title_type=feature

```
#Loading the rvest package
library('rvest')
```

```
## Loading required package: xml2
```

```
#Specifying the url for desired website to be scraped
url <- 'https://www.imdb.com/search/title?count=250&release_date=2018,2018&title_type=feature'


#Reading the HTML code from the website
webpage <- read_html(url)
```

Mengambil Deskripsi Film

```
#Using CSS selectors to scrap the description section
```

```
description_data_html <- html_nodes(webpage,'.ratings-bar+ .text-muted')

#Converting the description data to text
description_data <- html_text(description_data_html)
```

Mengambil Metascore

```
#Using CSS selectors to scrap the rankings section
metascore_html <- html_nodes(webpage,'.ratings-bar .rating-metascore')

#Converting the ranking data to text
metascore <- html_text(metascore_html)

length(metascore)
```

```
## [1] 0
```

Mengambil Artis

```
#Using CSS selectors to scrap the actors section
actors_data_html <- html_nodes(webpage,'.lister-item-content .ghost+ a')

#Converting the gross actors data to text
actors_data <- html_text(actors_data_html)
```

Mengambil Rating Penonton

Sebagai inputan rating batas dasar penonton.

```
rank_penonton_html <- html_nodes(webpage, '.text-muted .certificate')

rank_penonton <- html_text(rank_penonton_html)

length(rank_penonton)
```

```
## [1] 204
```

Only 204 data. So we are not using it.

Mengambil Ranknya

Sebagai inputan angka urutan pada data.

```
#Using CSS selectors to scrap the rankings section
rank_data_html <- html_nodes(webpage,'.text-primary')

#Converting the ranking data to text
rank_data <- html_text(rank_data_html)
```

Mengambil Judul

Bagian untuk scraping judul film.

```
#Using CSS selectors to scrap the title section
title_data_html <- html_nodes(webpage,'.lister-item-header a')
```

Mengambil Durasi

Bagian untuk scraping durasi film.

```
#Using CSS selectors to scrap the Movie runtime section
runtime_data_html <- html_nodes(webpage,'.runtime')
```

Mengambil Genre

Bagian untuk scraping genre.

```
#Using CSS selectors to scrap the Movie genre section
genre_data_html <- html_nodes(webpage,'.genre')

#Converting the genre data to text
genre_data <- html_text(genre_data_html)
```

Mengambil Rating Movie

Bagian untuk scraping rating.

```
#Using CSS selectors to scrap the IMDB rating section
rating_data_html <- html_nodes(webpage,'.ratings-imdb-rating strong')

```

```
#Converting the ratings data to text
rating_data <- html_text(rating_data_html)
```

Mengambil Votes Data

Bagian untuk scraping jumlah vote.

```
#Using CSS selectors to scrap the votes section
votes_data_html <- html_nodes(webpage,'.sort-num_votes-visible span:nth-child
(2)')


#Converting the votes data to text
votes_data <- html_text(votes_data_html)


#head(votes_data)
#length(votes_data)
```

Mengambil Directors Data

Bagian untuk scraping sutradara pada film tersebut.

```
#Using CSS selectors to scrap the directors section
directors_data_html <- html_nodes(webpage,'.text-muted+ p a:nth-child(1)')


#Converting the directors data to text
directors_data <- html_text(directors_data_html)
```

Mengambil Gross Data

Scraping pada bagian gross pada suatu film.

```
#Using CSS selectors to scrap the gross revenue section
#gross_data_html <- html_nodes(webpage,'.ghost~ .text-muted+ span')


#Converting the gross revenue data to text
#gross_data <- html_text(gross_data_html)


#length(gross_data)
```

Length yang didapatkan sangat jauh dari data lainnya (167 dengan 250)

Preprocessing

Preproc Artis

```
#Data-Preprocessing: converting actors data into factors

actors_data<-as.factor(actors_data)
```

Preproc Deskripsi Film

```
#Data-Preprocessing: removing '\n'

description_data<-gsub("\n","",description_data)


#Let's have another look at the description data

head(description_data)
```

```
## [1] "    The story of the legendary rock band Queen and lead singer Freddi
e Mercury, leading up to their famous performance at Live Aid (1985)."

## [2] "    A musician helps a young singer find fame, even as age and alcoho
lism send his own career into a downward spiral."

## [3] "    A working-class Italian-American bouncer becomes the driver of an
African-American classical pianist on a tour of venues through the 1960s Amer
ican South."

## [4] "    In early 18th century England, a frail Queen Anne occupies the th
rone and her close friend, Lady Sarah, governs the country in her stead. When
a new servant, Abigail, arrives, her charm endears her to Sarah."

## [5] "    Teen Miles Morales becomes Spider-Man of his reality, crossing hi
s path with five counterparts from other dimensions to stop a threat for all
realities."

## [6] "    A year in the life of a middle-class family's maid in Mexico City
in the early 1970s."
```

Mengubah ke numeric data rank.

```
class(rank_data)
```

```
## [1] "character"
```

```
#Data-Preprocessing: Converting rankings to numerical

rank_data<-as.numeric(rank_data)


#Let's have a look at the title

head(rank_data)
```

```
## [1] 1 2 3 4 5 6
```

Mengubah ke text data durasi.

```
#Converting the runtime data to text
runtime_data <- html_text(runtime_data_html)

#Let's have a look at the runtime
head(runtime_data)
## [1] "134 min" "136 min" "130 min" "119 min" "117 min" "135 min"
```

Menghilangkan “min” dan mengubah ke angka data durasi.

```
#Data-Preprocessing: removing mins and converting it to numerical

runtime_data<-gsub(" min","",runtime_data)
runtime_data<-as.numeric(runtime_data)

#Let's have another look at the runtime data
head(runtime_data)
## [1] 134 136 130 119 117 135
#Only 249 data!
length(runtime_data)
## [1] 249
#using mean for get last data
runtime_data.mean <- mean(runtime_data)
runtime_data.mean <-floor(runtime_data.mean)

runtime_data <- c(runtime_data,runtime_data.mean)
```

Mengubah ke text data judul.

```
#Converting the title data to text
title_data <- html_text(title_data_html)

#Let's have a look at the runtime
head(title_data)
## [1] "Bohemian Rhapsody"                 "A Star Is Born"
## [3] "Green Book"                        "The Favourite"
```

```
## [5] "Spider-Man: Into the Spider-Verse" "Roma"
```

Menghapus beberapa bagian.

```
#Data-Preprocessing: removing \n
genre_data<-gsub("\n","",genre_data)

#Data-Preprocessing: removing excess spaces
genre_data<-gsub(" ","",genre_data)

#taking only the first genre of each movie
genre_data<-gsub(",.*","",genre_data)

#Convering each genre from text to factor
genre_data<-as.factor(genre_data)

#Let's have another look at the genre data
head(genre_data)
```

```
## [1] Biography Drama     Biography Biography Animation Drama
## 12 Levels: Action Adventure Animation Biography Comedy Crime ... Thriller
```

Mengubah rating ke numerik.

```
#Data-Preprocessing: converting ratings to numerical
rating_data<-as.numeric(rating_data)
```

Menghapus “,” dan mengubah ke numerik data vote.

```
#Data-Preprocessing: removing commas
votes_data<-gsub(",","",votes_data)

#Data-Preprocessing: converting votes to numerical
votes_data<-as.numeric(votes_data)
```

mengubah data sutradara ke factor.

```
#Data-Preprocessing: converting directors data into factors
directors_data<-as.factor(directors_data)
```

Menghapus “$” & “M” dan mengambil data yang dibutuhkan.

```
#Data-Preprocessing: removing '$' and 'M' signs
#gross_data<-gsub("M","",gross_data)

#gross_data<-substring(gross_data,2,6)

#length(gross_data)
#anyNA(gross_data)
```

All List to form a data frame

Mengumpulkan kolom dan diubah ke data frame.

```
movies_df <- data.frame(Rank = rank_data,
                        Title = title_data,
                        Actor = actors_data,
                        Description = description_data,
                        Runtime = runtime_data,
                        Genre = genre_data,
                        Rating = rating_data,
                        Votes = votes_data,
                        Director = directors_data)
                        #Gross_Earning_in_Mil = gross_data)
```

Referensi dari: https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/ dengan perubahan sesuai kebutuhan.

DataFrameToCSV

```
write.csv(movies_df,"Movies-IMDB.csv")
```

Exploratory data analysis (EDA)

Library

Import Library.

```
library(ggplot2)
library(tidyr)
library(GGally)
```

OpenCSV

Membuka Data yang sudah disimpan diatas.

```
movies.csv <- read.csv("Movies-IMDB.csv")
```

Data Exploration

Melihat gambaran data yang ada.

```
str(movies.csv)
```

```
## 'data.frame':    250 obs. of  10 variables:
##  $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Title      : Factor w/ 250 levels "#SquadGoals",..: 40 7 83 194 177 163
104 130 150 51 ...
##  $ Actor      : Factor w/ 228 levels "Adam Collins",..: 178 135 218 172 19
4 224 155 192 124 162 ...
##  $ Description: Factor w/ 250 levels "    12 Strong tells the story of the
first Special Forces team deployed to Afghanistan after 9/11; under the le"|
__truncated__,..: 219 42 76 162 204 77 6 178 57 232 ...
##  $ Runtime    : int  134 136 130 119 117 135 118 124 110 130 ...
##  $ Genre      : Factor w/ 12 levels "Action","Adventure",..: 4 7 4 4 3 7 5
4 1 7 ...
##  $ Rating     : num  8.1 7.8 8.3 7.7 8.6 7.9 7.5 6.3 6.8 7.4 ...
##  $ Votes      : int  281581 202607 110318 80850 125672 89790 21908 15276 4
0485 50289 ...
##  $ Director   : Factor w/ 246 levels "Abby Kohn","Adam Collins",..: 41 37
191 246 33 9 213 139 142 220 ...
```

```
summary(movies.csv)
```

```
##        X                Rank                     Title
##  Min.   :  1.00   Min.   :  1.00   #SquadGoals   :  1
##  1st Qu.: 63.25   1st Qu.: 63.25   12 Strong     :  1
##  Median :125.50   Median :125.50   22 July       :  1
##  Mean   :125.50   Mean   :125.50   A Private War :  1
##  3rd Qu.:187.75   3rd Qu.:187.75   A Quiet Place :  1
##  Max.   :250.00   Max.   :250.00   A Simple Favor:  1
##                                    (Other)       :244
##                 Actor
##  Gerard Butler     :  3
##  John C. Reilly    :  3
##  Melissa McCarthy  :  3
##  Amandla Stenberg  :  2
##  Chloë Grace Moretz:  2
##  Christian Bale    :  2
##  (Other)           :235
##
Description
##      12 Strong tells the story of the first Special Forces team deployed t
o Afghanistan after 9/11; under the leadership of a new captain, the team mus
t work with an Afghan warlord to take down the Taliban.:  1
##      A 17-year-old girl suffers from a condition that prevents her from be
ing out in the sunlight.
:  1
##      A 90-year-old horticulturist and Korean War veteran is caught transpo
rting $3 million worth of cocaine through Illinois for a Mexican drug cartel.
:  1
##      A big box store worker reinvents her life and her life-story and show
s Madison Avenue what street smarts can do.
:  1
##      A biologist signs up for a dangerous, secret expedition into a myster
ious zone where the laws of nature don't apply.
:  1
##      A couple find themselves in over their heads when they foster three c
hildren.
:  1
##  (Other)
:244
```

```
##     Runtime              Genre        Rating          Votes
##  Min.   : 81.00   Action   :58   Min.   :3.200   Min.   :    19
##  1st Qu.: 97.25   Drama    :53   1st Qu.:5.900   1st Qu.:  4756
##  Median :107.50   Comedy   :38   Median :6.550   Median : 18451
##  Mean   :110.30   Biography:25   Mean   :6.447   Mean   : 43060
##  3rd Qu.:120.00   Crime    :21   3rd Qu.:7.100   3rd Qu.: 42313
##  Max.   :188.00   Adventure:17   Max.   :8.600   Max.   :593157
##                   (Other)  :38
##            Director
##  Ari Sandel    :  2
##  Clint Eastwood:  2
##  Eli Roth      :  2
##  Jason Reitman :  2
##  Abby Kohn     :  1
##  Adam Collins  :  1
##  (Other)       :240
```

```
head(movies.csv,10)
```

```
##     X Rank                                Title             Actor
## 1   1    1                  Bohemian Rhapsody        Rami Malek
## 2   2    2                     A Star Is Born         Lady Gaga
## 3   3    3                         Green Book   Viggo Mortensen
## 4   4    4                      The Favourite     Olivia Colman
## 5   5    5 Spider-Man: Into the Spider-Verse     Shameik Moore
## 6   6    6                               Roma  Yalitza Aparicio
## 7   7    7                     Instant Family     Mark Wahlberg
## 8   8    8               Mary Queen of Scots     Saoirse Ronan
## 9   9    9                           Overlord       Jovan Adepo
## 10 10   10                           Creed II Michael B. Jordan
##
Description
## 1
The story of the legendary rock band Queen and lead singer Freddie Mercury, l
eading up to their famous performance at Live Aid (1985).
## 2
A musician helps a young singer find fame, even as age and alcoholism send hi
s own career into a downward spiral.
```

```
## 3                                                                    A working-clas
s Italian-American bouncer becomes the driver of an African-American classica
l pianist on a tour of venues through the 1960s American South.

## 4      In early 18th century England, a frail Queen Anne occupies the thro
ne and her close friend, Lady Sarah, governs the country in her stead. When a
new servant, Abigail, arrives, her charm endears her to Sarah.

## 5                                                                  Teen Miles M
orales becomes Spider-Man of his reality, crossing his path with five counter
parts from other dimensions to stop a threat for all realities.

## 6
A year in the life of a middle-class family's maid in Mexico City in the earl
y 1970s.

## 7
A couple find themselves in over their heads when they foster three children.

## 8                                                                       Mary S
tuart's attempt to overthrow her cousin Elizabeth I, Queen of England, finds
her condemned to years of imprisonment before facing execution.

## 9
A small group of American soldiers find horror behind enemy lines on the eve
of D-Day.

## 10
Under the tutelage of Rocky Balboa, heavyweight contender Adonis Creed faces
off against Viktor Drago, son of Ivan Drago.

##    Runtime     Genre Rating  Votes         Director
## 1      134 Biography    8.1 281581     Bryan Singer
## 2      136     Drama    7.8 202607   Bradley Cooper
## 3      130 Biography    8.3 110318   Peter Farrelly
## 4      119 Biography    7.7  80850 Yorgos Lanthimos
## 5      117 Animation    8.6 125672  Bob Persichetti
## 6      135     Drama    7.9  89790   Alfonso Cuarón
## 7      118    Comedy    7.5  21908      Sean Anders
## 8      124 Biography    6.3  15276     Josie Rourke
## 9      110    Action    6.8  40485     Julius Avery
## 10     130     Drama    7.4  50289 Steven Caple Jr.
```

Data Cleaning

Menghapus Kolom “X”, karena tidak digunakan.

```
colSums(is.na(movies.csv))
```

```
##           X        Rank       Title       Actor Description     Runtime
```

```
##            0            0            0            0            0            0
##        Genre       Rating        Votes     Director
##            0            0            0            0
```

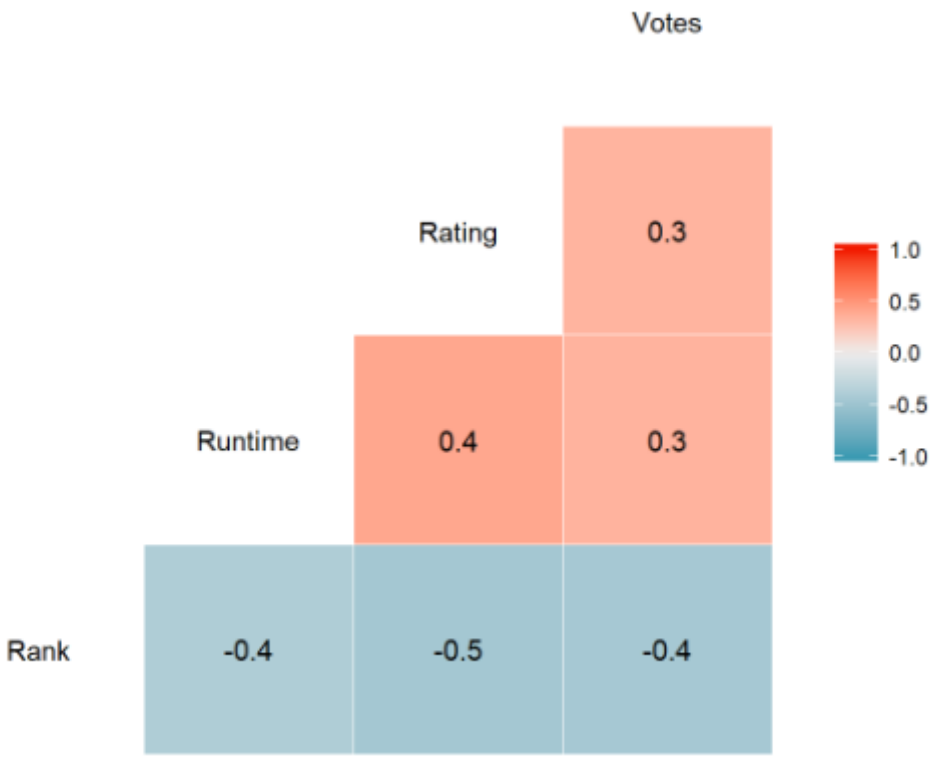
```
movies.csv$X = NULL
```

Model Building

Korelasi dari setiap data.

```
ggcorr(movies.csv, label = T)
```

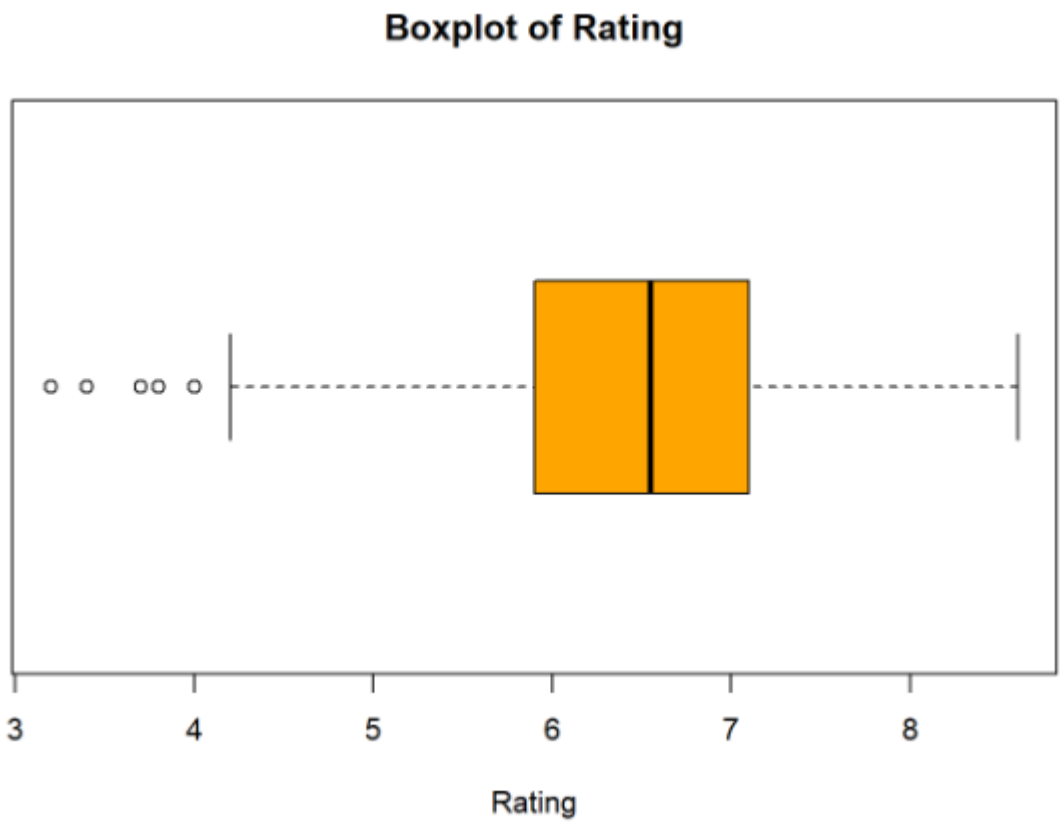
```
## Warning in ggcorr(movies.csv, label = T): data in column(s) 'Title',
## 'Actor', 'Description', 'Genre', 'Director' are not numeric and were
## ignored
```



Present Results

Visualisasi beberapa data.

```
boxplot(movies.csv$Rating, horizontal = T, main = "Boxplot of Rating", xlab=
"Rating", col = "orange")
```




```
qplot(data = movies.csv,Runtime,fill = Genre,bins = 25, main = "Histogram of
Runtime")
```

```
ggplot(movies.csv,aes(x=Runtime,y=Rating))+
geom_point(aes(size=Votes,col=Genre))+
ggtitle("Relasi Runtime dengan Rating")
```