UTS R

Danny Revaldo

2023-10-21

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyve
rse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_co
nflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to fo
rce all conflicts to become errors
```

1. DATA PREPROCESSING:

a) Import data CSV tersebut kedalam project kalian, Ada berapa kolom dan data didalam dataset tersebut?

```
df <- read.csv("dataset_UTS_bahasaR.csv")

jml_col <- ncol(df)
jml_row <- nrow(df)

paste("Jumlah kolomnya:",jml_col)
```

```
## [1] "Jumlah kolomnya: 20"
```

```
paste("Jumlah datanya:",jml_row)
```

```
## [1] "Jumlah datanya: 4803"
```

b) Buat dataframe baru dengan hanya mengambil kolom Judul (title), Durasi (runtime), Tanggal Rilis (release_date), Budget (budget), Pendapatan (revenue), Popularitas (popularity), dan Rating (vote_average) seperti tabel dibawah ini (dengan nama variabel datasetUTS)!

```
datasetUTS <- df%>%select(title,runtime,release_date,budget,revenue,pop
ularity,vote_average)

datasetUTS%>%head(10)
```

```
##                                         title runtime release_date
budget
## 1                                     Avatar     162   2009-12-10 237
000000
## 2  Pirates of the Caribbean: At World's End     169   2007-05-19 300
000000
## 3                                    Spectre     148   2015-10-26 245
000000
## 4                      The Dark Knight Rises     165   2012-07-16 250
000000
## 5                                John Carter     132   2012-03-07 260
000000
## 6                               Spider-Man 3     139   2007-05-01 258
000000
## 7                                    Tangled     100   2010-11-24 260
000000
## 8                    Avengers: Age of Ultron     141   2015-04-22 280
000000
## 9    Harry Potter and the Half-Blood Prince     153   2009-07-07 250
000000
## 10       Batman v Superman: Dawn of Justice     151   2016-03-23 250
000000
##       revenue popularity vote_average
## 1  2787965087  150.43758          7.2
## 2   961000000  139.08262          6.9
## 3   880674609  107.37679          6.3
## 4  1084939099  112.31295          7.6
## 5   284139100   43.92699          6.1
## 6   890871626  115.69981          5.9
## 7   591794936   48.68197          7.4
## 8  1405403694  134.27923          7.3
## 9   933959197   98.88564          7.4
## 10  873260194  155.79045          5.7
```

c) Film dikatakan sukses jika memiliki kriteria jika proyek film tersebut memiliki budget dibawah rata-rata namun memiliki pendapatan tinggi, sebaliknya film yang gagal memiliki kriteria jika proyek film tersebut memiliki budget diatas rata-rata namun memiliki pendapatan rendah.Carilah 10 film yang sukses (dengan nama variabel top10success) dan 10 film yang gagal (dengan nama variabel top10failure) di tahun 2015.

```
rata_budget <- datasetUTS%>%mutate(tahun_rilis=year(as.Date(release_dat
e)))

rata_budget <- rata_budget%>%filter(tahun_rilis==2015)%>%mutate(rata_ra
ta_budget=mean(budget))

top10success <-rata_budget%>%filter(budget<rata_rata_budget & revenue>b
udget)%>%arrange(desc(revenue))%>%head(10)
```

```
top10success
```

```
##                                    title runtime release_date   bud
get
## 1                          Pitch Perfect 2     115   2015-05-07 29000
000
## 2  Alvin and the Chipmunks: The Road Chip      92   2015-12-17
  0
## 3                   Straight Outta Compton     147   2015-08-13 28000
000
## 4                            The Big Short     130   2015-12-11 28000
000
## 5                          Magic Mike XXL     115   2015-07-01 14800
000
## 6                                Vacation      99   2015-07-28 31000
000
## 7                     Insidious: Chapter 3      97   2015-06-04 10000
000
## 8                                The Visit      94   2015-09-10  5000
000
## 9                                 Southpaw     123   2015-06-15 30000
000
## 10                               Spotlight     128   2015-11-06 20000
000
##      revenue popularity vote_average tahun_rilis rata_rata_budget
## 1  287506194   71.49689          6.8        2015         31132164
## 2  233755553   27.86737          5.8        2015         31132164
## 3  201634991   61.76233          7.7        2015         31132164
## 4  133346506   57.51847          7.3        2015         31132164
## 5  122513057   29.66099          6.3        2015         31132164
## 6  104384188   47.75360          6.2        2015         31132164
## 7  104303851   45.94652          6.2        2015         31132164
## 8   98450062   38.94708          6.0        2015         31132164
## 9   91709827   65.36445          7.3        2015         31132164
## 10  88346473   41.50359          7.8        2015         31132164
```

```
top10failure <- rata_budget%>%filter(budget>rata_rata_budget & revenue<
budget &tahun_rilis==2015)%>%arrange(desc(budget))%>%head(10)
```

```
top10failure
```

```
##                       title runtime release_date    budget   revenue
popularity
## 1                       Pan     111   2015-09-24 150000000 128388320
  48.03528
## 2  In the Heart of the Sea     122   2015-11-20 100000000  93820758
  50.76733
## 3                 Blackhat     133   2015-01-13  70000000  17752940
  46.83237
```

```
## 4                Mortdecai     106   2015-01-21  60000000  30418560
  52.41753
## 5         The Ridiculous 6     119   2015-12-11  60000000         0
  19.69469
## 6                 Child 44     137   2015-03-15  50000000   3324330
  40.37110
## 7      Victor Frankenstein     109   2015-11-10  40000000  34227298
  24.82114
## 8               The Gunman     115   2015-02-16  40000000  13644292
  26.93546
## 9                    Aloha     105   2015-05-27  37000000  26250020
  29.65254
## 10     Unfinished Business      91   2015-03-05  35000000  14431253
  16.83231
##    vote_average tahun_rilis rata_rata_budget
## 1           5.9        2015         31132164
## 2           6.5        2015         31132164
## 3           5.1        2015         31132164
## 4           5.4        2015         31132164
## 5           4.9        2015         31132164
## 6           6.1        2015         31132164
## 7           5.6        2015         31132164
## 8           5.5        2015         31132164
## 9           5.2        2015         31132164
## 10          5.0        2015         31132164
```

d) Dari dataset tersebut, terdapat film-film sequel, trilogi atau prequel dilihat dari judul film yang sama namun diikuti dengan angka seperti Avatar 1 dan Avatar 2 atau dengan "subtitle" berbeda seperti Harry Potter and The Goblet of Fire dan Harry Potter and The Prisoner of Azkaban.

```
franchise_data<-datasetUTS%>%mutate(franchise = gsub("^(\\w+\\s+\\w+).*
", "\\1", datasetUTS$title))

# Mengelompokkan judul film berdasarkan dua kata pertama (tanpa angka)
dan menghitung total revenue dan mean vote_average
franchise_data <- franchise_data %>%
  group_by(franchise) %>%
  summarise(total_revenue = sum(revenue),avg_rating = mean(vote_averag
e))%>%
  arrange(desc(total_revenue))
```

1. Carilah 5 franchise film besar yang memiliki sequel, trilogi, atau prequel dengan total pendapat yang tertinggi! (dengan nama variable top5franchise)

```
# Menampilkan hasil pengelompokkan dan total revenue
top5franchise <- franchise_data%>%head(5)
top5franchise%>%select(franchise,total_revenue)

## # A tibble: 5 × 2
##   franchise    total_revenue
##   <chr>                <dbl>
```

```
## 1 Harry Potter    5411061557
## 2 Pirates of      3727384838
## 3 Star Wars       3199113893
## 4 The Hunger      2944162634
## 5 The Hobbit      2935523356
```

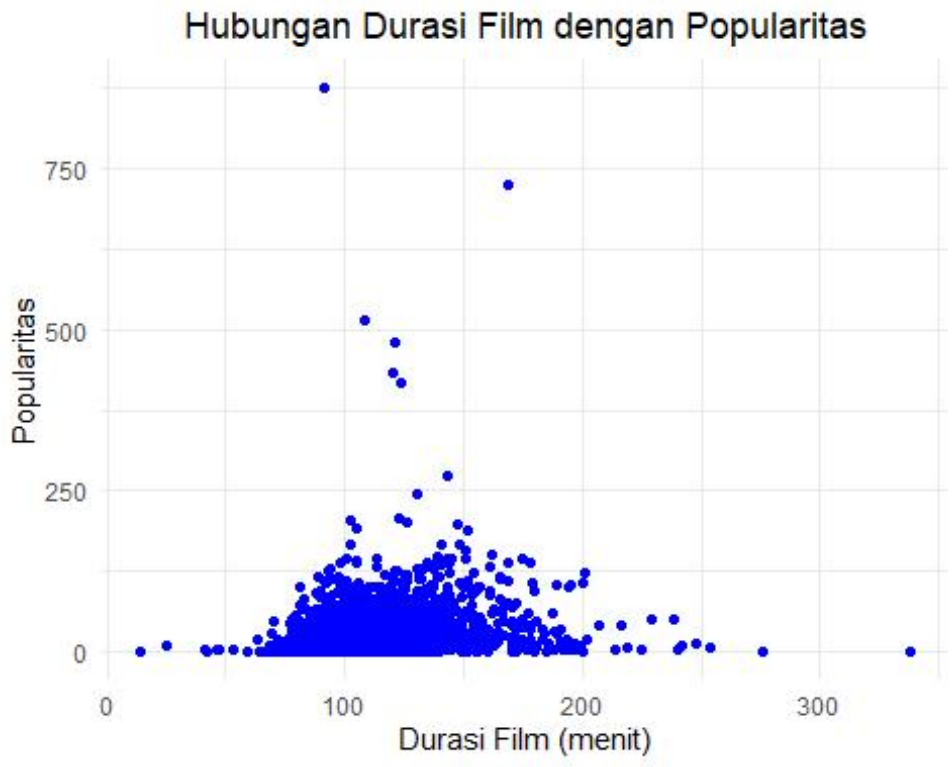
2. Dari 5 franchise tersebut, manakah yang memiliki rata-rata rating tertinggi? (dengan nama variabel bestfranchise)

```
# Menampilkan hasil pengelompokkan, total revenue, mean vote_average, d
an franchise dengan rating tertinggi
bestfranchise<-top5franchise%>%select(franchise,avg_rating)



bestfranchise%>%arrange(desc(avg_rating))%>%head(1)
```

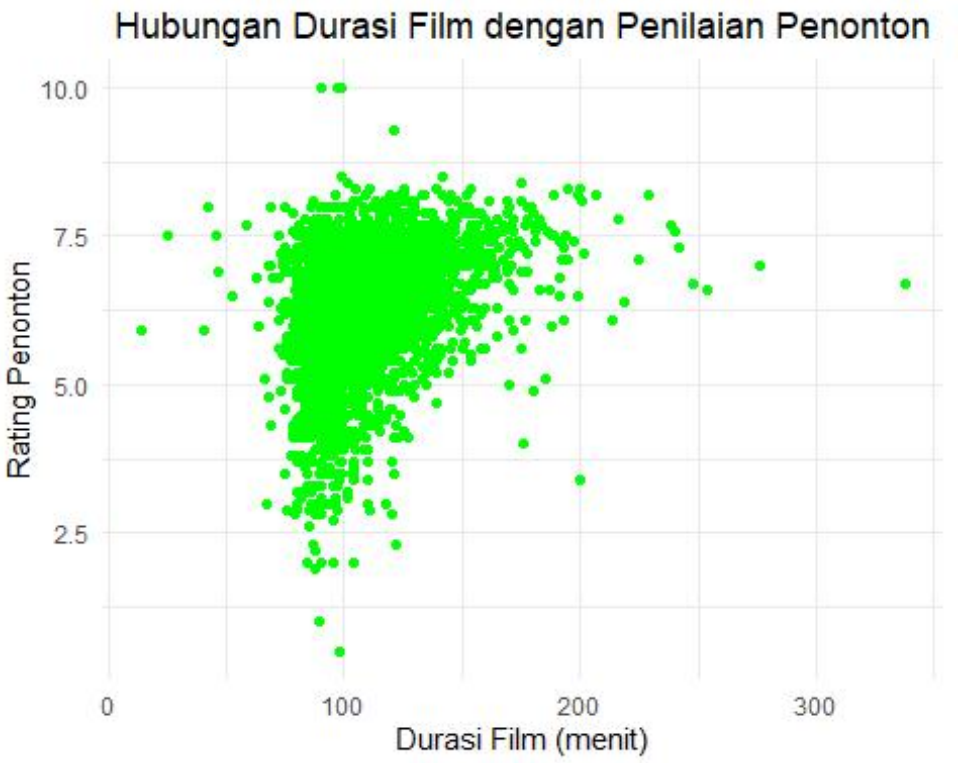
```
## # A tibble: 1 × 2
##   franchise    avg_rating
##   <chr>             <dbl>
## 1 Harry Potter       7.48
```

e) Carilah insight lain yang dapat diambil dari dataset tersebut! Jelaskan dengan tag komen (#)

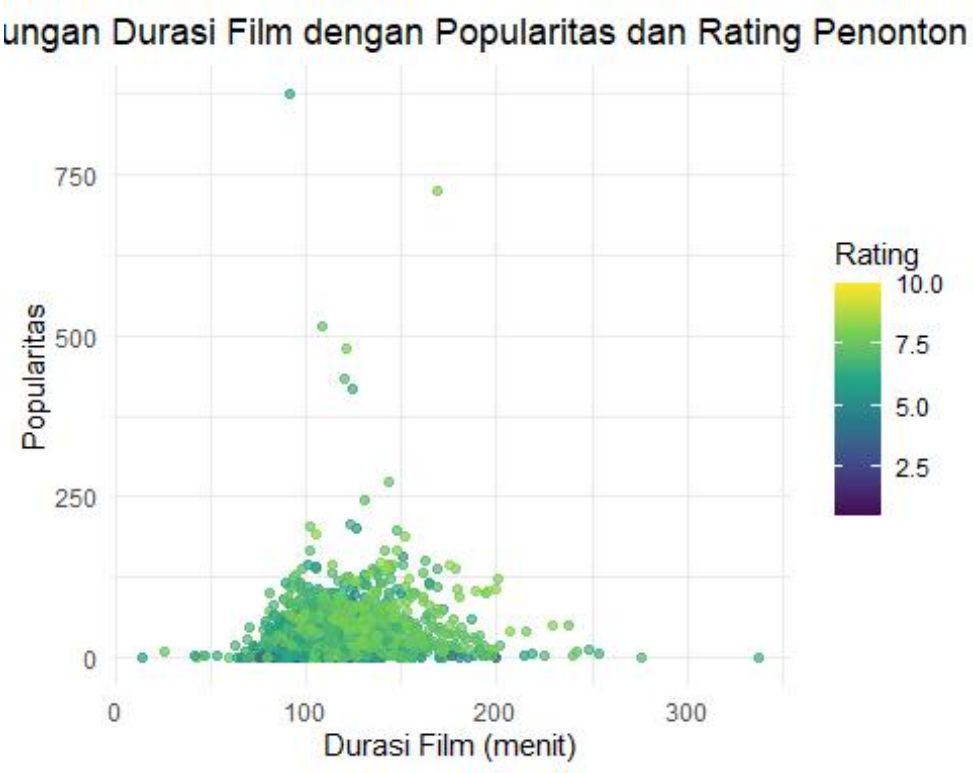
```
# filter nilai 0
hub_durasi_rating_popu <- datasetUTS%>%filter(popularity!=0&vote_averag
e!=0&runtime!=0)

# Membuat scatter plot untuk melihat hubungan antara durasi film dan po
pularitas
ggplot(hub_durasi_rating_popu, aes(x = runtime, y = popularity)) +
  geom_point(color = "blue") +
  labs(x = "Durasi Film (menit)", y = "Popularitas", title = "Hubungan
Durasi Film dengan Popularitas") +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))
```




```
# Membuat scatter plot untuk melihat hubungan antara durasi film dan ra
ting penonton
ggplot(hub_durasi_rating_popu, aes(x = runtime, y = vote_average)) +
  geom_point(color = "green") +
  labs(x = "Durasi Film (menit)", y = "Rating Penonton", title = "Hubun
gan Durasi Film dengan Penilaian Penonton") +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(hub_durasi_rating_popu, aes(x = runtime, y = popularity, color =
 vote_average)) +
  geom_point(alpha=0.7) +
  labs(x = "Durasi Film (menit)", y = "Popularitas", title = "Hubungan
Durasi Film dengan Popularitas dan Rating Penonton",color="Rating") +
  scale_color_viridis_c() + scale_size_continuous(range = c(3, 15)) +
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))
```

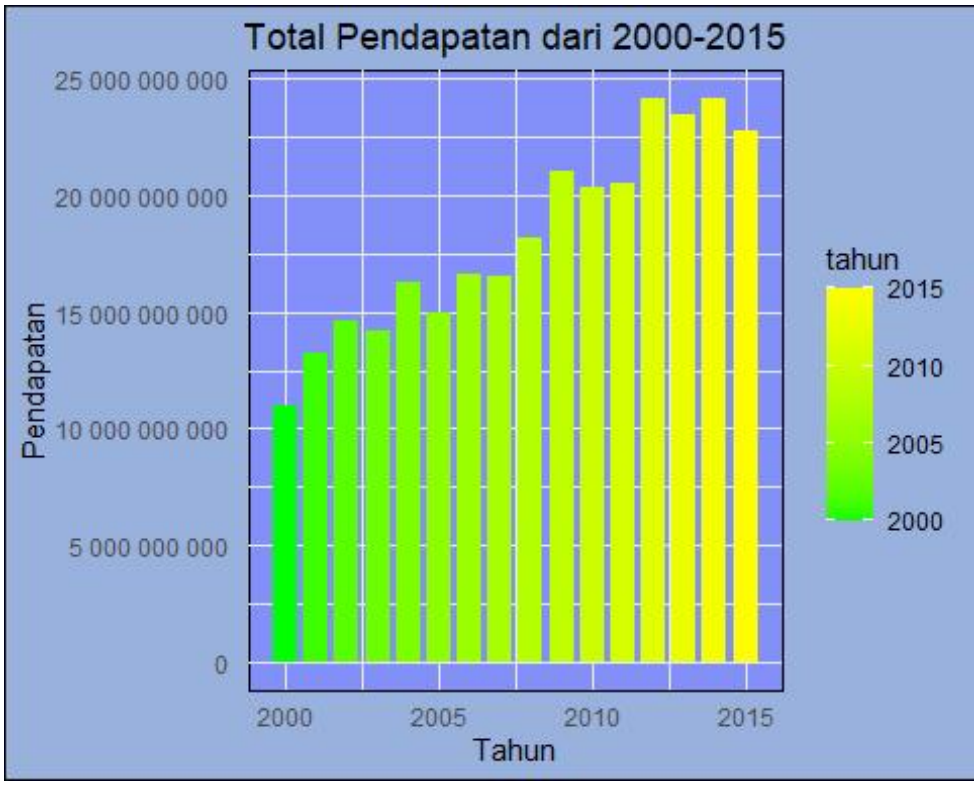


Dari hasil visualisasi diatas bisa kita ambil insight bahwa semakin lama durasi film dan tingginya popularitas tidak menjamin akan mendapat rating yang sempurna. Dari visualisasi data yang ketiga rata-rata film kebanyakan diantara 90 sampai 150 menit saja untuk popularitasnya dikisaran 0-250 sedikit film yang mencapai diatas 500 popularitasnya dan ratingnya dikisaran 5-7.5. Ada 1 film yang durasinya melebihi 300 menit yaitu “carlos”.

2. DATA VISUALIZATION

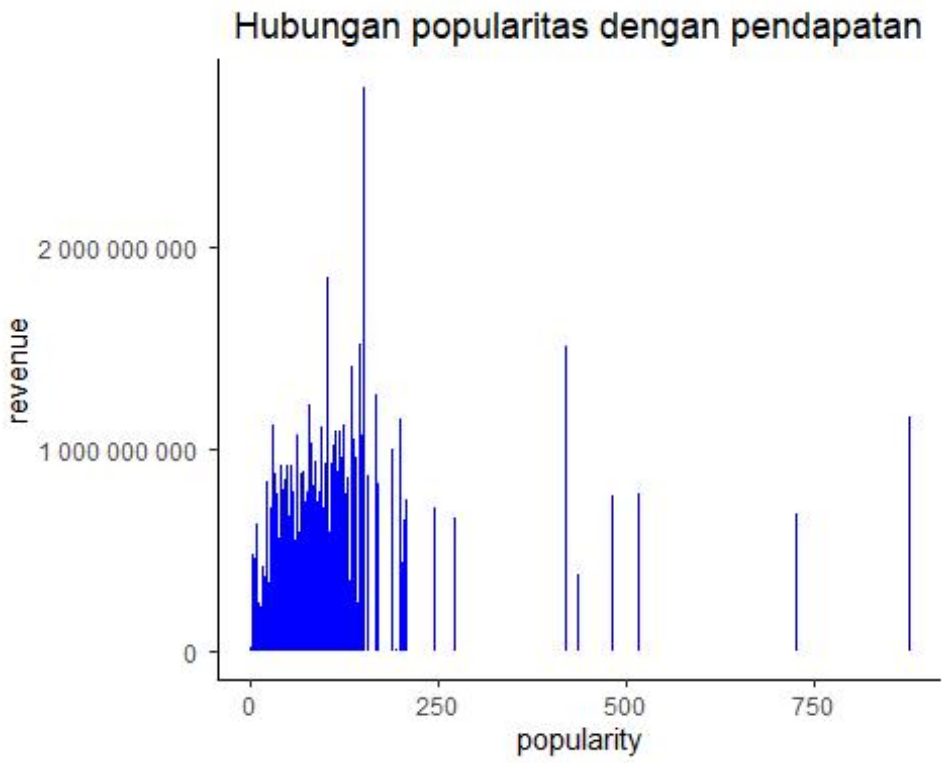
a) Buatlah Bar Chart untuk menunjukkan total pendapatan dari setiap tahun 2000-2015!

```
total_pendapatan_2015 <- datasetUTS%>%mutate(tahun=year(ymd(relea
se_date)))%>%filter(tahun<=2015&tahun>=2000)%>%group_by(tahun)%>%
summarise(jumlah= sum(revenue))

ggplot(total_pendapatan_2015, aes(x = tahun, y = jumlah, fill= ta
hun))+ geom_bar(stat = "identity",width = 0.8) + labs(title = "To
tal Pendapatan dari 2000-2015", x = "Tahun", y = "Pendapatan")+ s
cale_fill_gradient(low="green",high="yellow") + scale_y_continuou
s(labels = scales::unit_format(unit = "", scale = 1e-0)) +theme_m
inimal() +theme(plot.title = element_text(hjust = 0.5),panel.back
ground = element_rect(fill = "#8390fa"),plot.background = element
_rect(fill = "#99b2dd") )
```



b) Gunakan visualisasi untuk menunjukkan hubungan antara popularitas dan revenue Dan jabarkan dengan narasi dengan tag komen (#)!

```
hubungan_pop_pen <- datasetUTS%>%select(title,popularity,revenue)

ggplot(hubungan_pop_pen, aes(x=popularity , y=revenue)) + geom_ba
r(stat = "identity",color="blue") + scale_y_continuous(labels = s
cales::unit_format(unit = "", scale = 1e-0)) + labs(title = "Hubu
ngan popularitas dengan pendapatan")+theme_classic()+theme(plot.t
itle = element_text(hjust = 0.5))
```



Dari hasil visualisasi tentang tingkat popularitas dan pendapatan sebuah film bisa diambil kesimpulan bahwa popularitas tinggi tidak menjamin pendapatan yang tinggi. Tetapi popularitas yang dibawah 250 rata-rata pendapatannya diatas pendapatan dengan popularitas diatas 750 dan sekitar popularitas 170 keatas mencapai pendapatan hingga melebihi 2 juta USD dalam sekali tayangnya dan film itu adalah Avatar mencapai 2 Milyar USD. Lalu sedikit film yang mempunyai popularitas lebih dari 250 dan popularitas yang tertinggi dipegang oleh Minions sebesar 875.

Link drive:

https://tinyurl.com/UTSRDanny