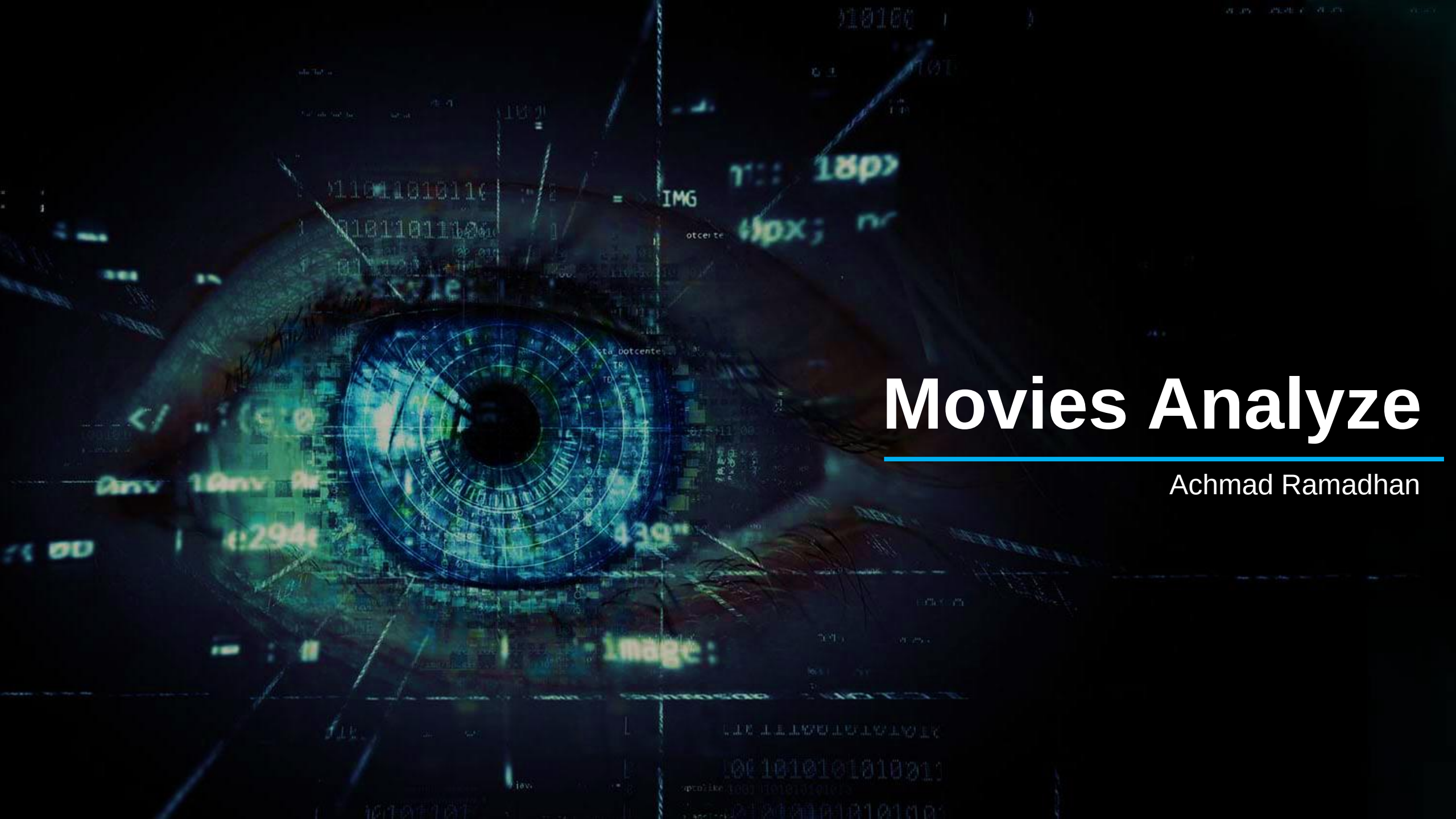


Movies Analyze

Achmad Ramadhan

Movies Analyze

Analisis ini digunakan sebagai personal project untuk mengidentifikasi trend-trend yang ada di data set movies. Data set ini diperoleh melalui Kaggle

01 Prepare
02 Process
03 Analyze
04 Insight

Prepare

Membuka Rstudio

```
library(dplyr)
library(janitor)
library(tidyverse)
library(ggplot2)
library(lubridate)
library(tidyr)
library(data.table)
library(scales)
library(grid)
library(gridExtra)
library(gghighlight)
```

kaggle

Mencari dan mengunduh data set melalui kaggle

Menyiapkan beberapa library untuk memudahkan dalam analisis di Rstudio

IMPORT DATA

Mengimport data yang telah di unduh melalui Kaggle menggunakan perintah read.csv

```
movies <- read.csv("movies.csv", na.strings = c('')) %>% clean_names()
glimpse(movies)
```

Menyimpan data yang diimport ke dalam data frame movies

Glimpse digunakan untuk melihat struktur data dari data frame movies

```
Rows: 739,048
Columns: 20
$ id                   <int> 663712, 732459, 436270, 675054, 42063…
$ title                <chr> "Terrifier 2", "Blade of the 47 Ronin…
$ genres               <chr> "Horror-Thriller", "Action-Fantasy", …
$ original_language    <chr> "en", "en", "en", "es", "en", "la", "…
$ overview             <chr> "After being resurrected by a siniste…
$ popularity           <dbl> 4608.567, 3821.739, 3772.253, 3401.45…
$ production_companies <chr> "Bloody Disgusting-Dark Age Cinema-Fu…
$ release_date         <chr> "2022-10-06", "2022-10-25", "2022-10-…
$ budget               <dbl> 2.5e+05, 0.0e+00, 2.0e+08, 0.0e+00, 1…
$ revenue              <dbl> 10155347, 0, 319000000, 0, 0, 0, 0, 6…
$ runtime              <dbl> 138, 106, 125, 93, 84, 0, 98, 149, 88…
$ status               <chr> "Released", "Released", "Released", "…
$ tagline              <chr> "Who's laughing now?", NA, "The world…
$ vote_average         <dbl> 7.095, 6.691, 6.854, 7.415, 6.555, 4.…
$ vote_count           <dbl> 496, 47, 939, 53, 908, 10, 132, 23, 4…
$ credits              <chr> "Lauren LaVera-David Howard Thornton-…
$ keywords             <chr> "clown-halloween-resurrection-sequel-…
$ poster_path          <chr> "/wRKHUqYGrp3PO91mZVQ18xlwYzW.jpg", "…
$ backdrop_path        <chr> "/y5ZOWesTjvn59jP6yo459eUsbli.jpg", "…
$ recommendations      <chr> "436270-732459-928123-575322-675054-4…
```

Data set Movies memiliki 739,048 baris dan 20 kolom

BIG DATA

Process

Mengubah tipe data kolom Release Date

Mengubahnya menjadi tipe data (Chr > Date

Memfilter Data

Melakukan filter data dengan rentang tanggal perilisan dari tahun “2000” sampai dengan “2022” dengan status “Released”

```
movies <- movies %>%
  select(-c(keywords, credits, poster_path, backdrop_path)) %>%
  mutate(year = year(release_date))%>%
  filter(year>=2000, year<=2022, status=='Released')
```

Menghapus Duplikat

Menghapus data duplikat berdasarkan id

```
```{r}
sum(duplicated(movies$id))
```

[1] 54182
```

Melihat isi dari kolom Status

```
Unique(movie$status)

[1] "Released" "Post Production" "In Production"
[4] "Planned" "Canceled" "Rumored"
```

Setelah data di filter dan menghapus data duplikat, terdapat 374978 Row

Split Genre Menjadi 2 Kolom

```
movies_genre <- as.data.frame(movies$genres, stringsAsFactors = FALSE)
movies_genre2 <- as.data.frame(tstrsplit(movies_genre[,1], '[-]', type.convert=TRUE),
                               stringsAsFactors=FALSE)
colnames (movies_genre2) <- c("genre1", "genre2")

movies_genre <-  movies_genre2 %>% select(c("genre1", "genre2"))

movies <- cbind (movies, movies_genre)
```

- Melakukan pengambilan kolom genre lalu disimpan ke data frame “movies_genre”
- Memisahkan data dengan pemisaah ‘-’ dan tiap pemisah akan disimpan di kolom genre1, genre2, genre3 dan selanjutnya
- Memilih kolom genre1 dan genre2 saja
- Menggabungkan kolom genre1 dan genre2 ke dalam data frame movies

Output

```
chr [1:374978] "Horror"
"Action" "Action"
"Horror" "Horror" ...
chr [1:374978]
"Thriller" "Fantasy"
"Fantasy" "Action" ...
```

Total Movies by Genre(1 & 2) Tahun 2000-2022

Genre 1

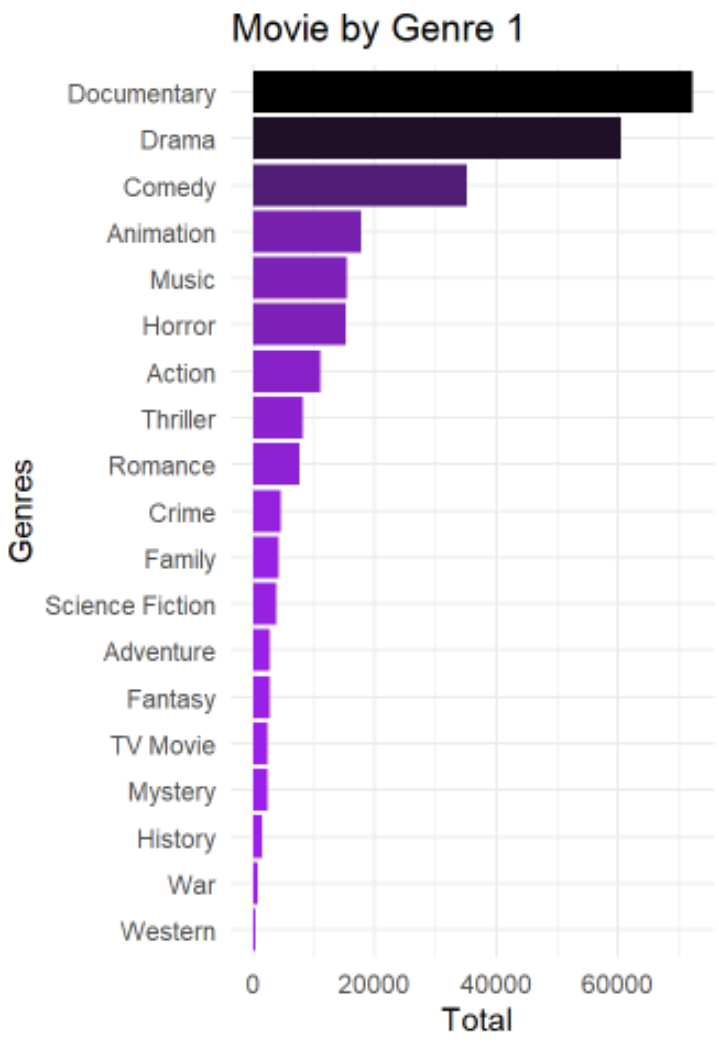
Documentary, Drama, Comedy merupakan 3 teratas total movie

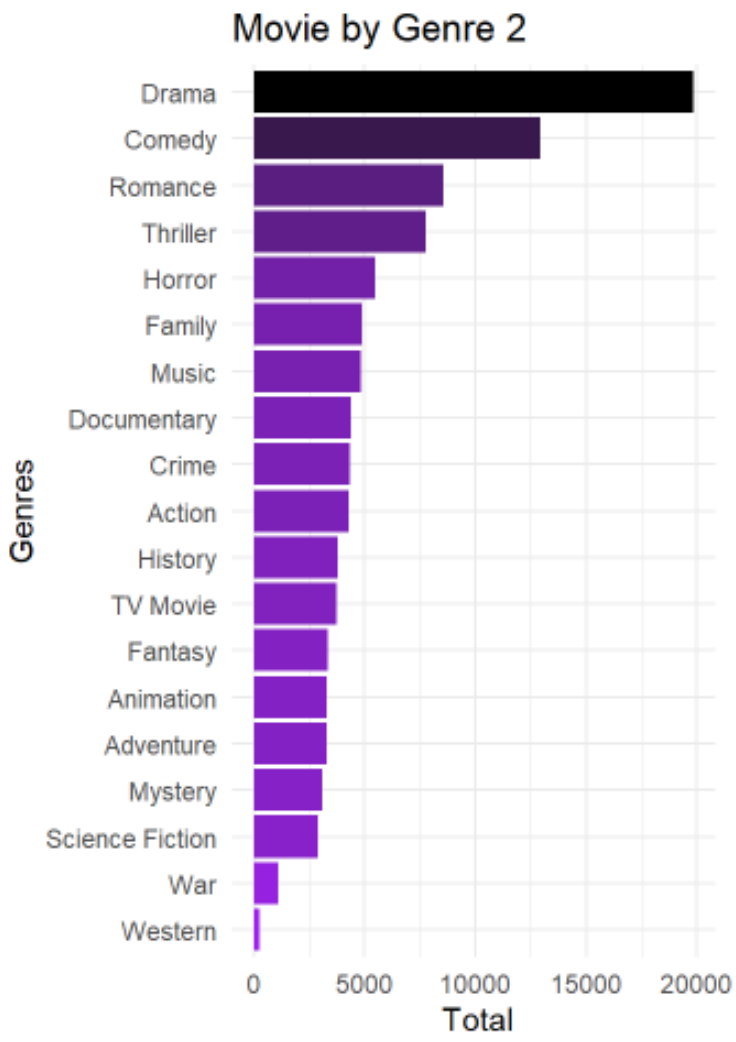
Western, War, History merupakan 3 terbawah total movie

Genre 2

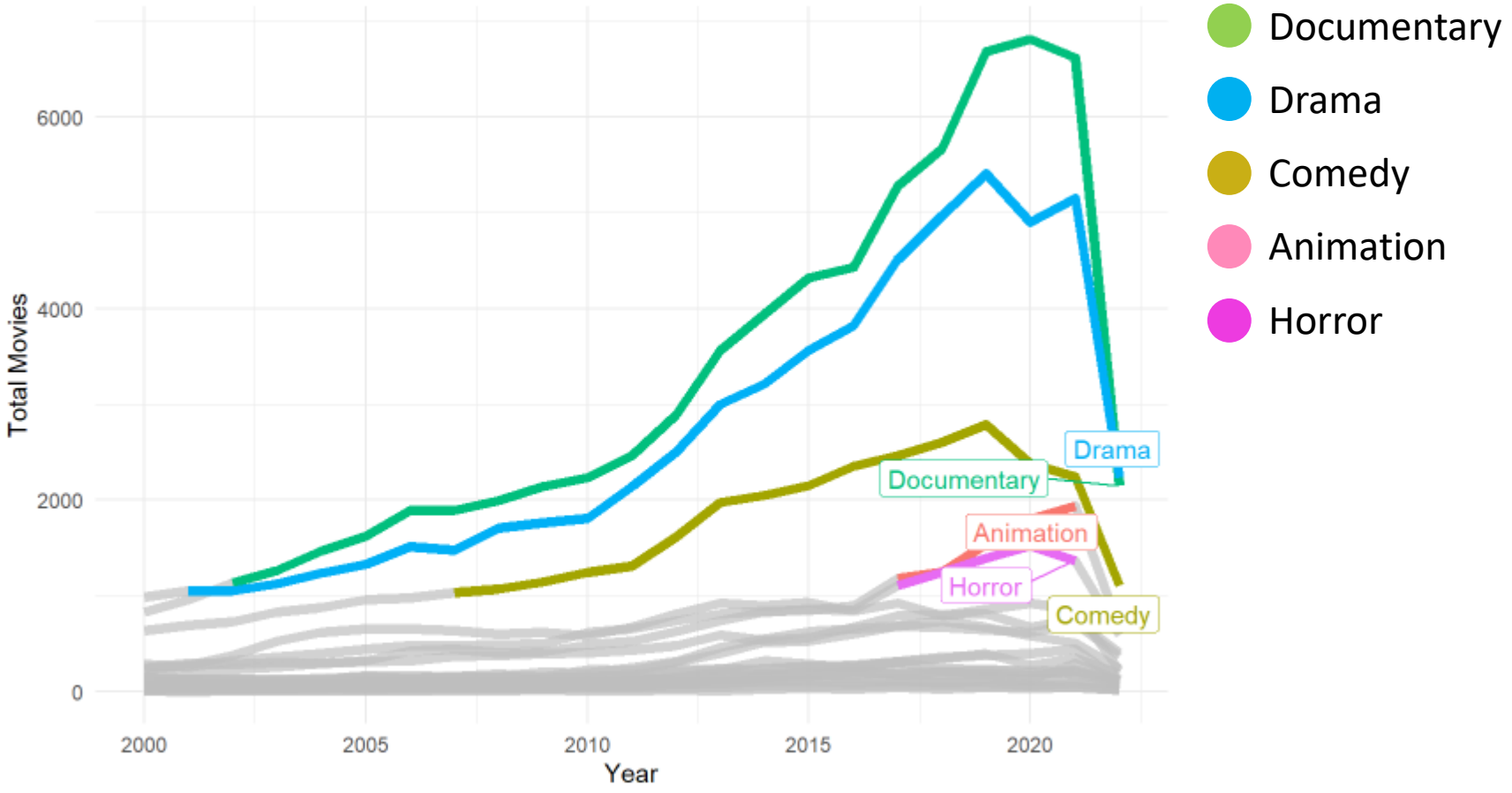
Drama, Comedy, Romance merupakan 3 teratas total movie

Western, War, Sci-Fi merupakan 3 terbawah total movie





Highlighted Movie More Than 1000 In A Year



Grafik disamping merupakan grafik yang menunjukkan perilisan movie di atas 1000 dalam 1 tahun

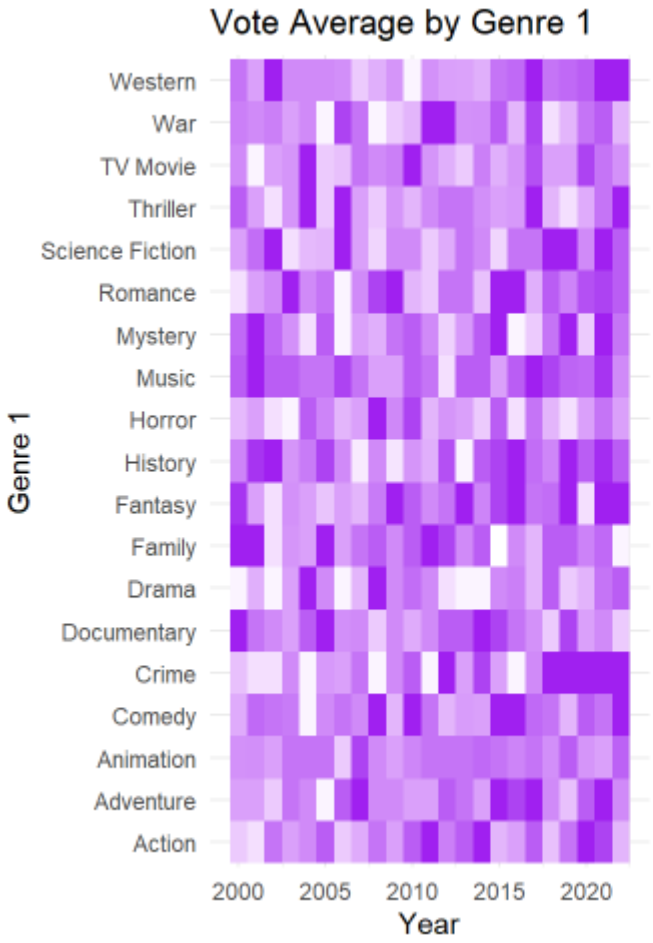
2001, Drama merilis lebih dari 1000 movie dalam setahun dan konsisten sampai 2022

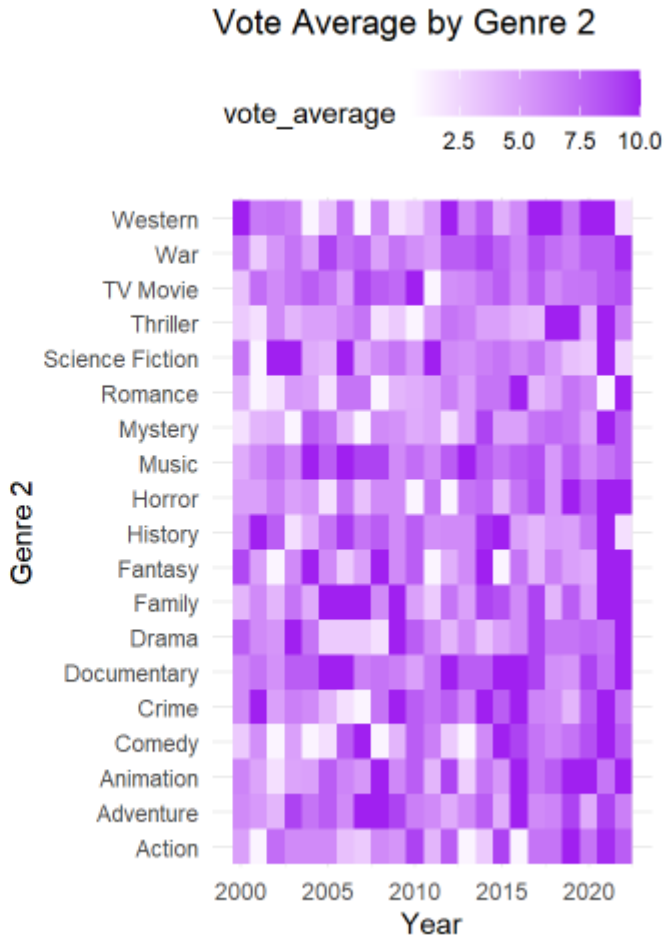
2002, documentary mulai merilis movie lebih dari 1000 sampai tahun 2022

2007, Comedy merilis lebih dari 1000 movie sampai tahun 2022

2017, Animation & Horror merilis movie lebih dari 1000, tetapi hanya sampai tahun 2021

Vote Average





Semakin ungu warna dari kotak menandakan semakin tinggi vote average terhadap suatu genre

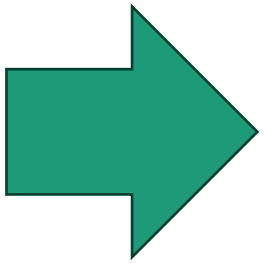
Semakin putih warna dari kotak menandakan semakin rendah vote average terhadap suatu genre

Prepare WordCloud

Documentary

Berdasarkan hasil dari diagram genre1, “Documentary” merupakan genre paling banyak perilisan movie dari tahun 2000 sampai dengan 2022

Melakukan Text Mining dengan kolom Overviews dalam genre “Documentary”



Tujuannya adalah untuk mengetahui kata apa yang sering muncul di dalam Overviews Movie dengan genre “Documentary”

Cleaning Text

```r
library(tidytext)
library(textclean)


documentary <- movies %>%
  select (genre1, overview) %>%
  filter (genre1 == "Documentary") %>%
  drop_NA()

overviews <- documentary$overview
```

Persiapan library dan data frame

Memilih kolom genre1, dan overview dengan memfilter data genre1 yaitu “Documentary” dan menghapus data yang kosong atau tidak memenuhi syarat

```r
# cleaning
overviews <- overviews %>%
  str_to_lower() %>% #Change words to lower case
  replace_contraction() %>%  #Replace contractions with both words (ex : i'm = i am)
  replace_word_elongation() %>%   #Replace word elongations with shortened form (ex : filmmm = film)
  strip() #Remove all non word characters
```

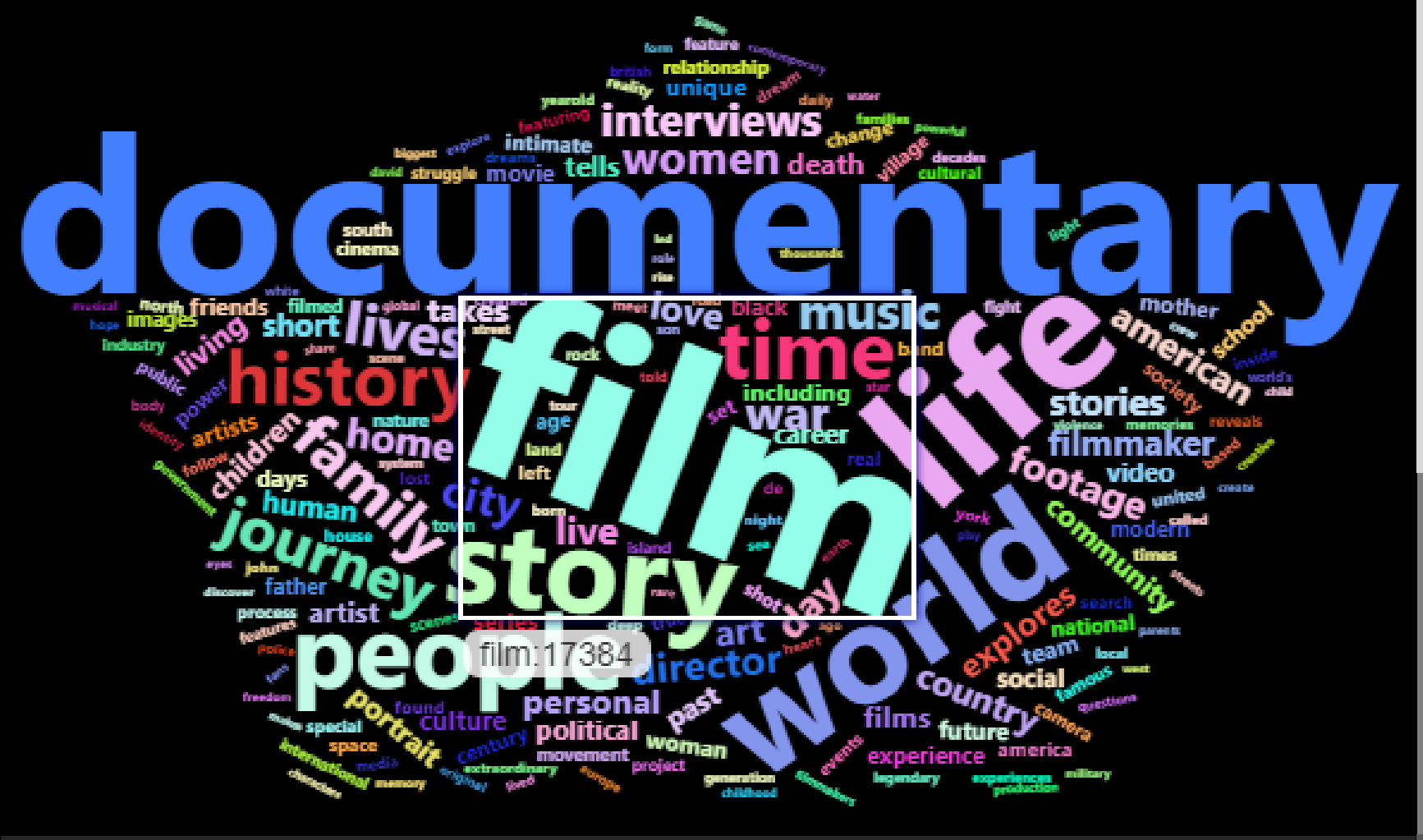
Proses Cleaning

- ❑ Mengubah seluruh kata menjadi huruf kecil
- ❑ Mengubah singkatan menjadi kata baku
- ❑ Menghilangkan huruf berlebihan
- ❑ Menghilangkan seluruh karakter yang bukan huruf

```r
#tokenize & remove stopwords
documentary <- enframe(overviews, value = "word", name=NULL) %>% #vector to data frame
  unnest_tokens(word, word) %>% #changing 1 word to 1 coloumn
  count(word, sort = T) %>% #counting word sorting by desc
  anti_join(stop_words) #anti join stop words (a, is, of, the..)
```

Tokenize dan Menghapus kata umum(stopwords)

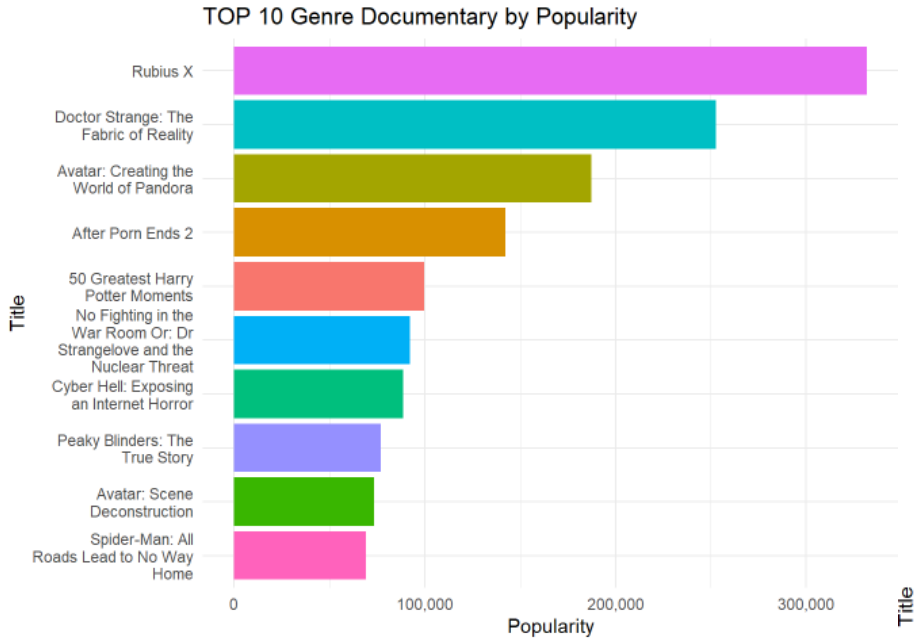
- ❑ Mengubah vector(overvies) menjadi dataframe(documentary)
- ❑ Mengubah 1 kata menjadi 1 kolom
- ❑ Menghitung kata yang sama dengan urutan desc
- ❑ Tidak memasukkan kata umum(stopwords)

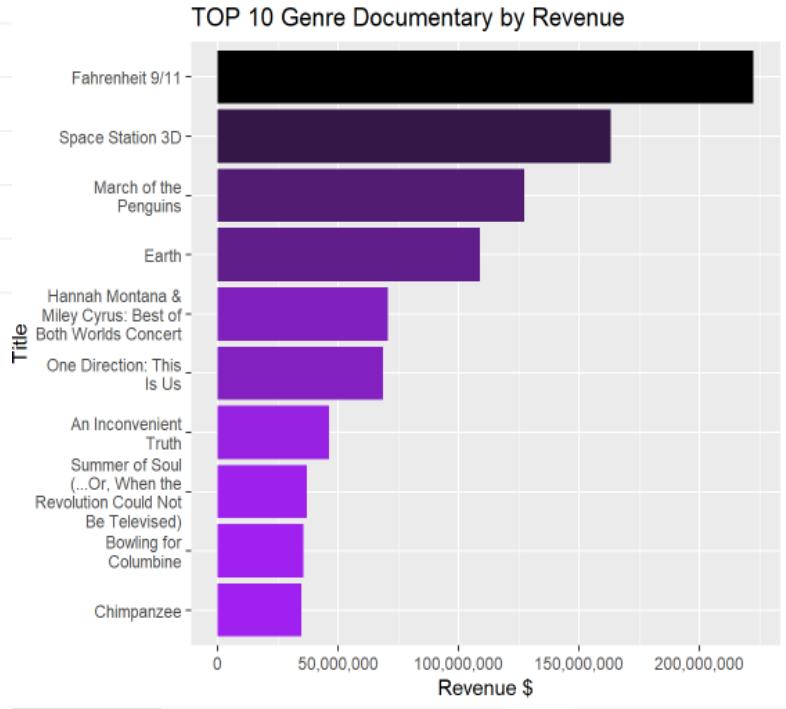


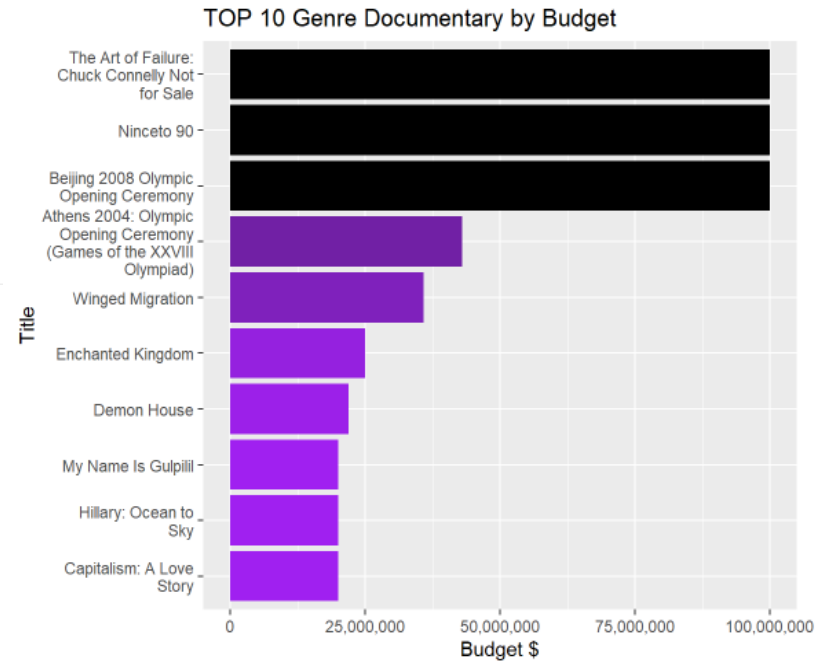
Semakin besar sebuah teks, semakin banyak penggunaan katanya. Sebaliknya semakin kecil huruf semakin sedikit kata yang digunakan

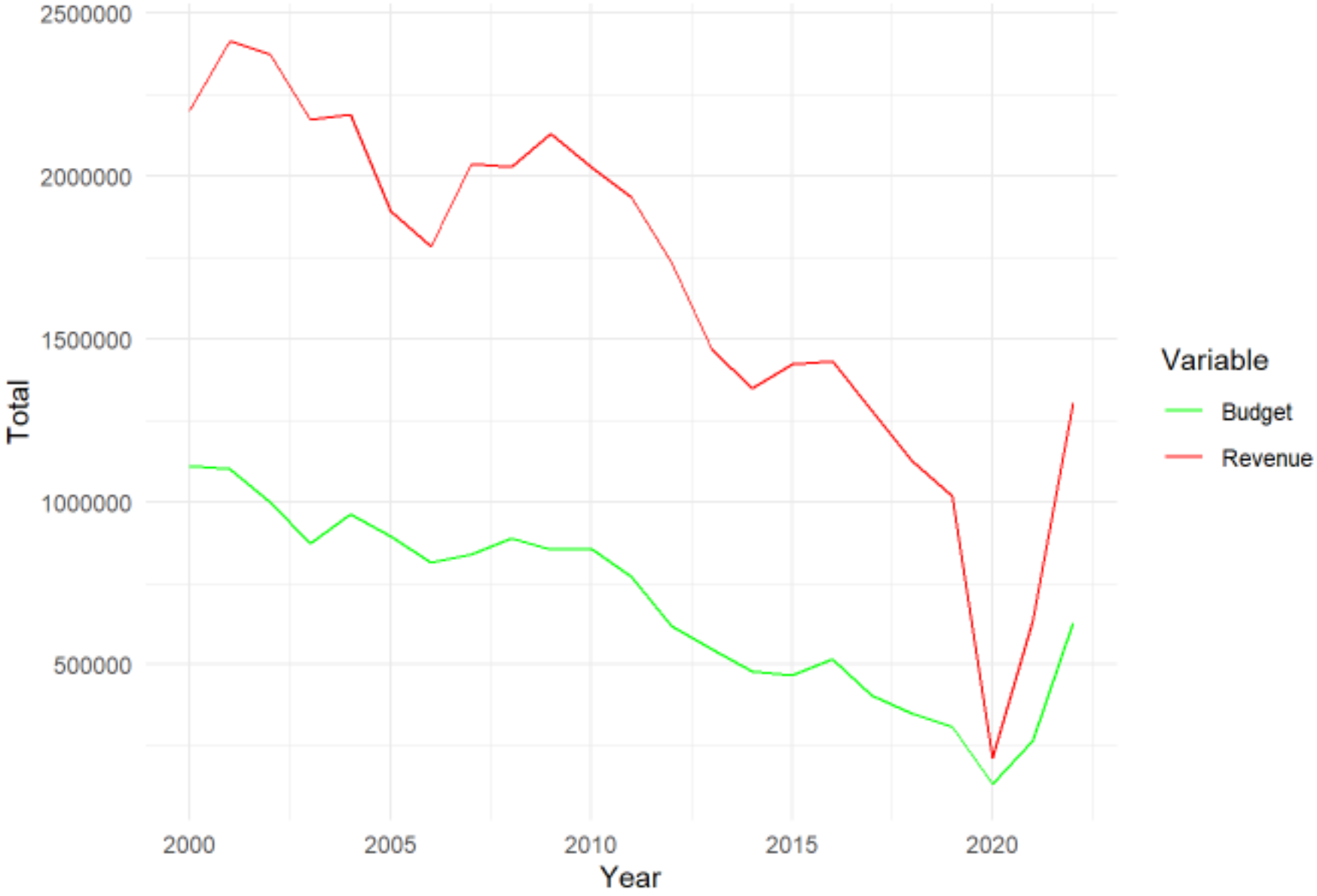
Film merupakan kata yang paling banyak ditemukan dalam overviews, sebanyak 17384

Output Text Mining dalam Overviews Movie dengan Genre “Documentary” Top 200 kata


TOP 10 Genre Documentary by Popularity
Rubius X
Doctor Strange: The Fabric of Reality
Avatar: Creating the World of Pandora
After Porn Ends 2
50 Greatest Harry Potter Moments
No Fighting in the War Room Or: Dr Strangelove and the Nuclear Threat
Cyber Hell: Exposing an Internet Horror
Peaky Blinders: The True Story
Avatar: Scene Deconstruction
Spider-Man: All Roads Lead to No Way Home
Title
0
100,000
200,000
300,000
Popularity

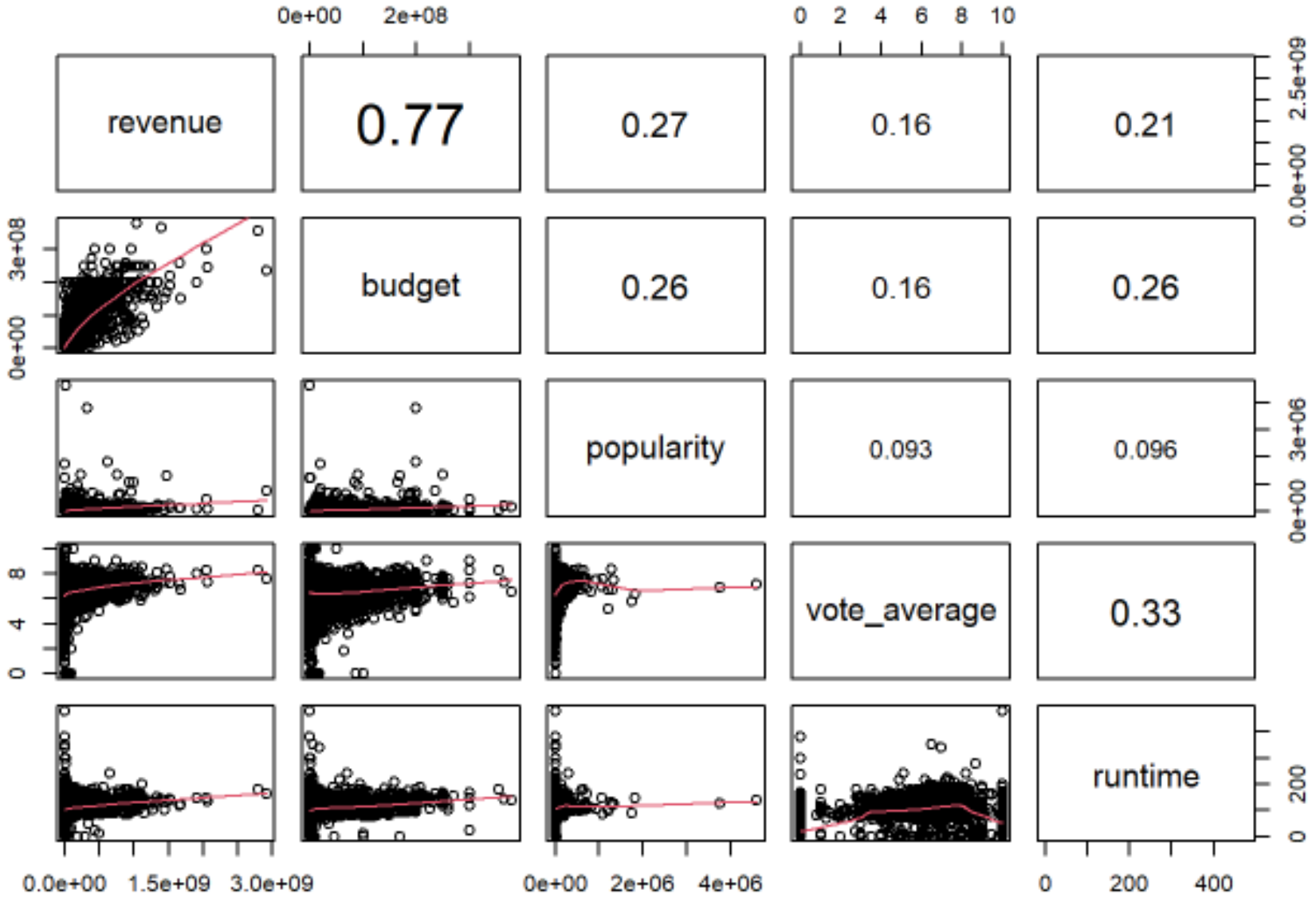

TOP 10 Genre Documentary by Revenue
Fahrenheit 9/11
Space Station 3D
March of the Penguins
Earth
Hannah Montana & Miley Cyrus: Best of Both Worlds Concert
One Direction: This Is Us
An Inconvenient Truth
Summer of Soul (...Or, When the Revolution Could Not Be Televised)
Bowling for Columbine
Chimpanzee
Title
0
50,000,000
100,000,000
150,000,000
200,000,000
Revenue $


TOP 10 Genre Documentary by Budget
The Art of Failure: Chuck Connelly Not for Sale
Ninceto 90
Beijing 2008 Olympic Opening Ceremony
Athens 2004: Olympic Opening Ceremony (Games of the XXVIII Olympiad)
Winged Migration
Enchanted Kingdom
Demon House
My Name Is Gulpilil
Hillary: Ocean to Sky
Capitalism: A Love Story
Title
0
25,000,000
50,000,000
75,000,000
100,000,000
Budget $



Grafik di samping adalah grafik rata-rata revenue dan budget dari tahun 2000 sampai tahun 2022

Rata-rata revenue dan budget terlihat tren menurun dari tahun 2001 sampai 2022



Disamping adalah korelasi antar kolom

Revenue dan Budget memiliki korelasi yang positif dengan nilai 0.77. dapat disimpulkan semakin tinggi budget semakin tinggi revenue yang didapat

INSIGHT

Perilisan movie dari tahun 2000 sampai 2022, Genre 1 “Documentary” adalah Genre yang merilis Movie terbanyak, sementara Genre 2 adalah “Drama”

Dalam Overviews Movie dengan Genre1 “Documentary” kata yang paling sering digunakan adalah “Film”

Revenue dan Budget memiliki korelasi dengan nilai 0.77, nilai tersebut menunjukkan nilai yang positif. Semakin tinggi nilai Budget, Semakin tinggi nilai Revenue yang di dapat