UTS Data Mining and Visualization

Data Exploration and Visualization using Netflix Dataset in R

2440094352 – Devin Augustin

A. Introduction

Dalam mengerjakan UTS Data Mining and Visualization, kita perlu menggunakan bahasa pemrograman R. Apa itu R? R merupakan bahasa pemrograman yang dapat kita gunakan untuk melakukan komputasi statistik dan presentasi grafis untuk melakukan analisis dan visualisasi terhadap data. R menyediakan banyak banyak teknik statistic dan memiliki banyak package untuk memecahkan masalah yang berbeda. R bersifat open-source dan juga gratis sehingga bisa dipakai oleh siapa saja.

B. Data Description

Dataset yang saya pilih adalah netflix_titles.csv yang terdiri dari 12 kolom dan 8807 baris. Deskripsi dari dataset Netflix:

- show_id (char) = ID dari tiap acara.
- type (char) = Tipe acara yaitu movie dan tv show.
- title (char) = Judul dari acara.
- director (char) = Nama dari director acara.
- cast (char) = Nama dari pemeran karakter di acara.
- country (char) = Negara di mana acara tersebut tayang.
- date_added (char) = Tanggal ditambahkannya acara di platform netflix.
- release_year (int) = Tahun rilis acara tersebut.
- rating (char) = Rating dari acara atau klasifikasi film.
- duration (char) = Durasi dari setiap acara.
- listed_in (char) = Genre dari masing-masing acara.
- description (char) = Deskripsi singkat dari acara.

C. Data Exploration and Visualization

1. Missing Data

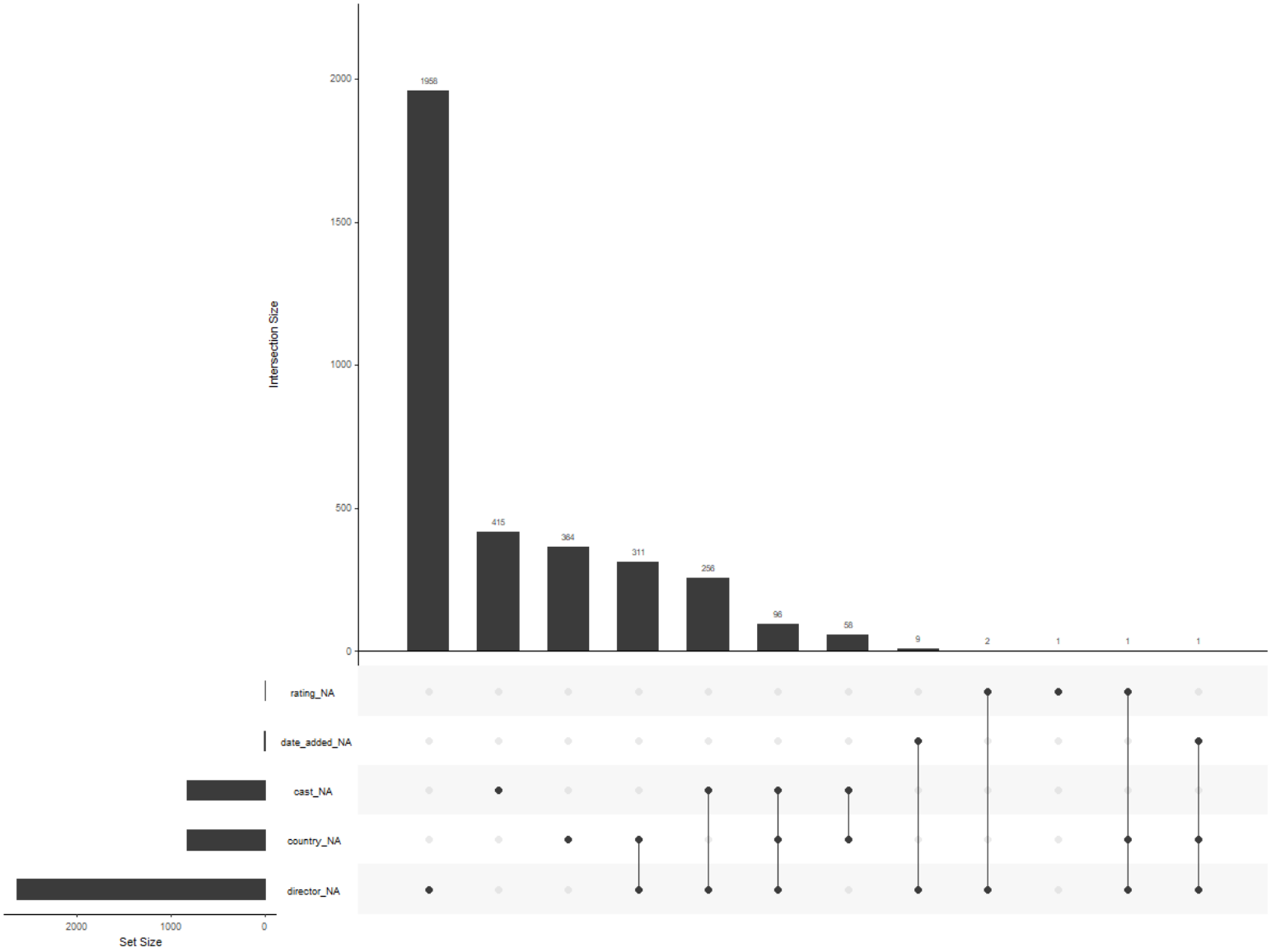
Terdapat beberapa missing data yang dapat kita lihat dibawah

```
data.frame("Variable"=c(colnames(netflix)),
           "missing values count"=sapply(netflix, function(x)sum(is.na(x))),
           row.names = NULL)
```

	variable	missing.values.count
1	type	0
2	title	0
3	director	2634
4	cast	825
5	country	831
6	date_added	10
7	release_year	0
8	rating	4
9	duration	0
10	listed_in	0
11	description	0

atau melalui grafik dibawah

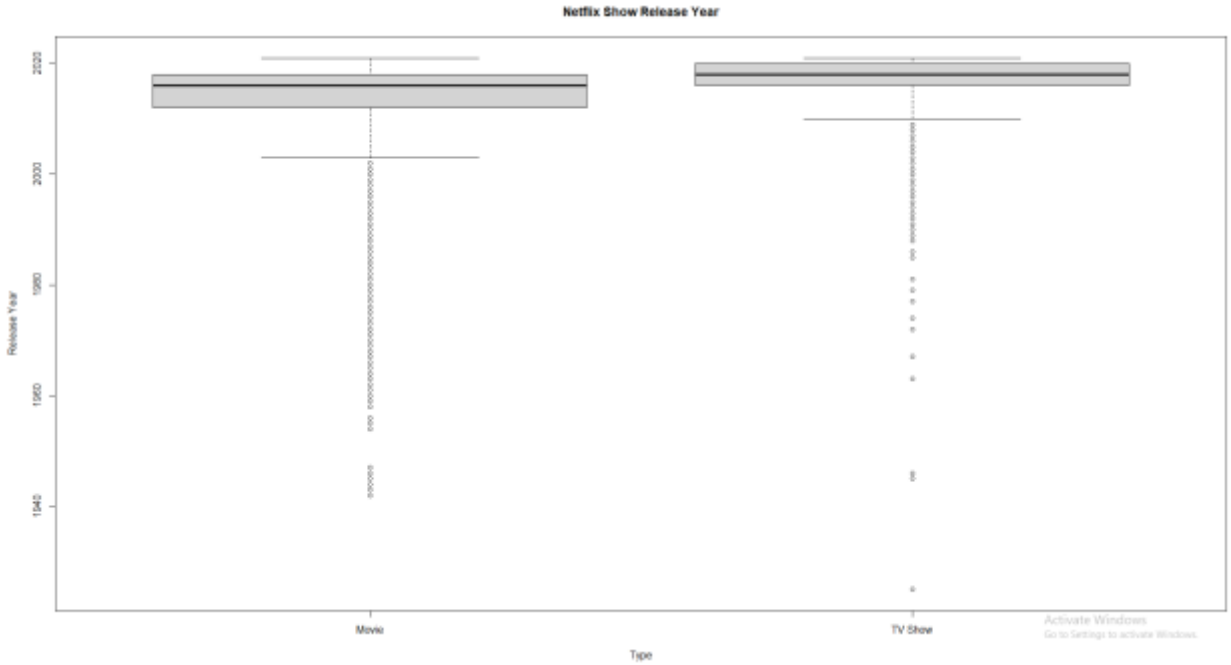
```
gg_miss_upset(netflix)
```



seperti kita lihat dari grafik diatas, terdapat lima variable yang memiliki missing value, yaitu rating, date_added, cast, country, dan director.

2. Outliers

```
boxplot(netflix$release_year~netflix$type, data=netflix,
        xlab = "Type", ylab = "Release Year",
        main = "Netflix Show Release Year")
```

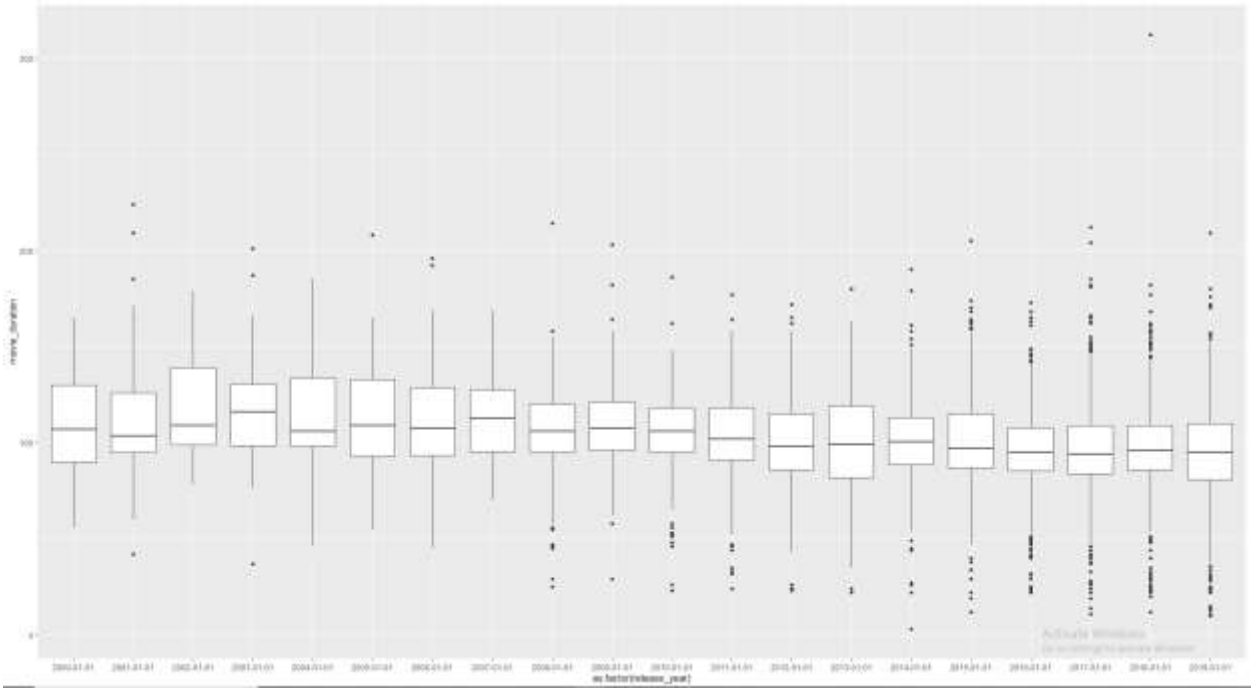


Dari grafik di atas, kita bisa melihat adanya beberapa outliers dan yang paling mencolok adalah outlier pada TV Show yang terbit pada tahun 1925, di mana rata-rata dari Movie dan TV Show di Netflix merupakan acara keluaran tahun 2000an.

3. Shape of the Distribution

Visualisasi Release Year terhadap Movie Duration

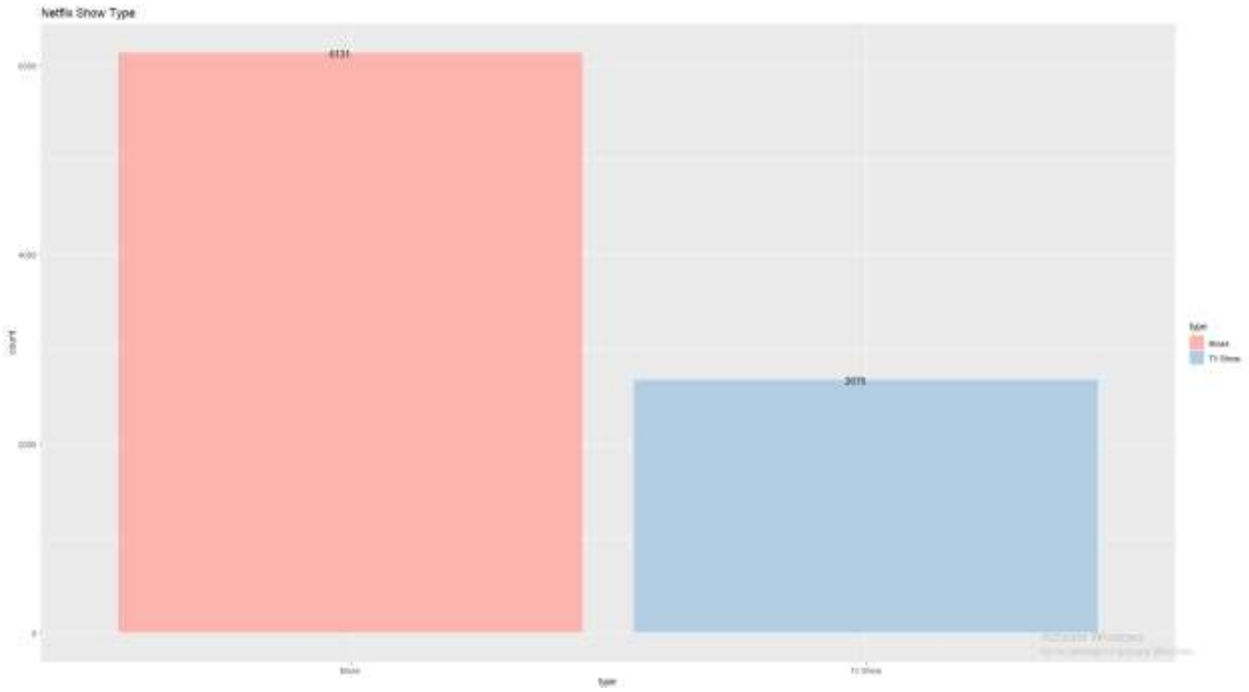
```
netflix %>%
  filter(type=='Movie'& release_year<="2020-01-01" & release_year>="1999-01-01") %>%
  mutate(movie_duration=substr(duration,1,nchar(as.character(duration))-4)) %>%
  mutate(movie_duration = as.integer(movie_duration)) %>%
  select(release_year,movie_duration) %>%
  ggplot() +
  geom_boxplot(aes(x=as.factor(release_year),y=movie_duration))
```



Kita bisa melihat bahwa rata-rata durasi acara di Netflix tidak berbeda jauh dari tahun ke tahunnya untuk sebagian besar acara dari tahun 2000 - 2019

Visualisasi Netflix Show Type

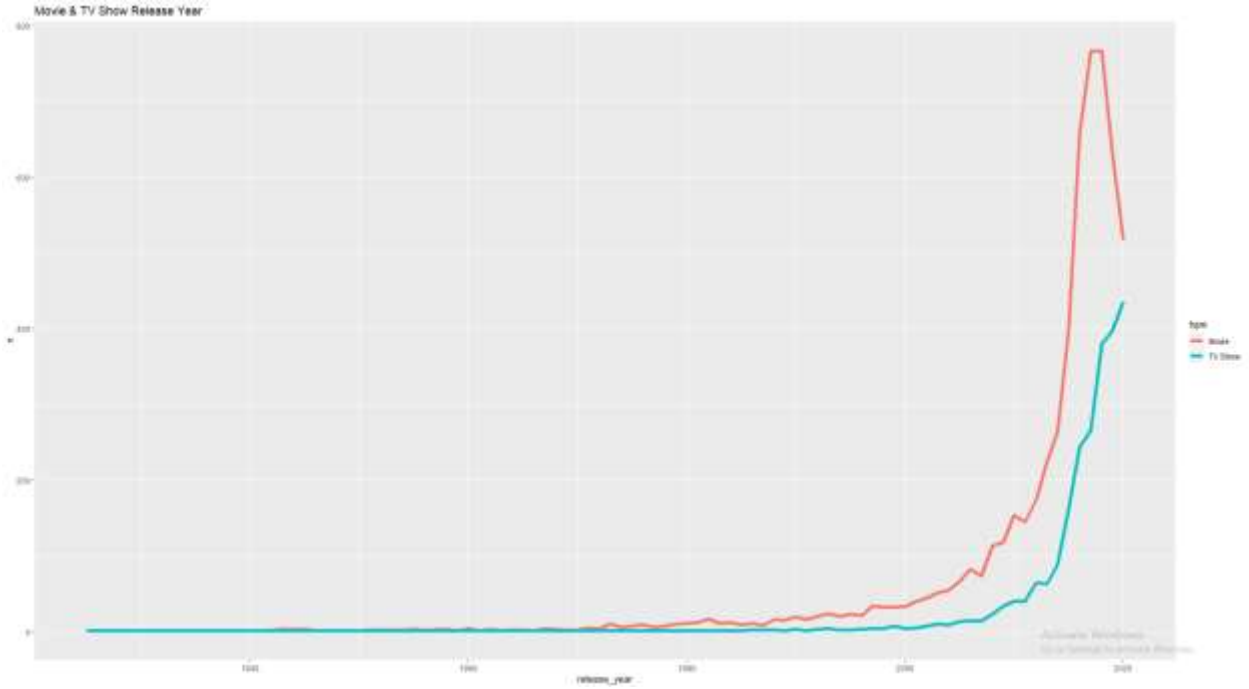
```
netflix %>%
  ggplot(aes(type,fill=type)) +
  geom_bar()+
  scale_fill_brewer(palette='Pastel1')+
  ggtitle("Netflix Show Type") +
  geom_text(stat='count',aes(label=..count..),hjust=1)
```



Dari sini kita bisa melihat bahwa Netflix show tipe Movie lebih banyak dibandingkan TV Show di Netflix hingga dua kali lipat.

Visualisasi Acara Netflix Berdasarkan Tahun Rilisnya

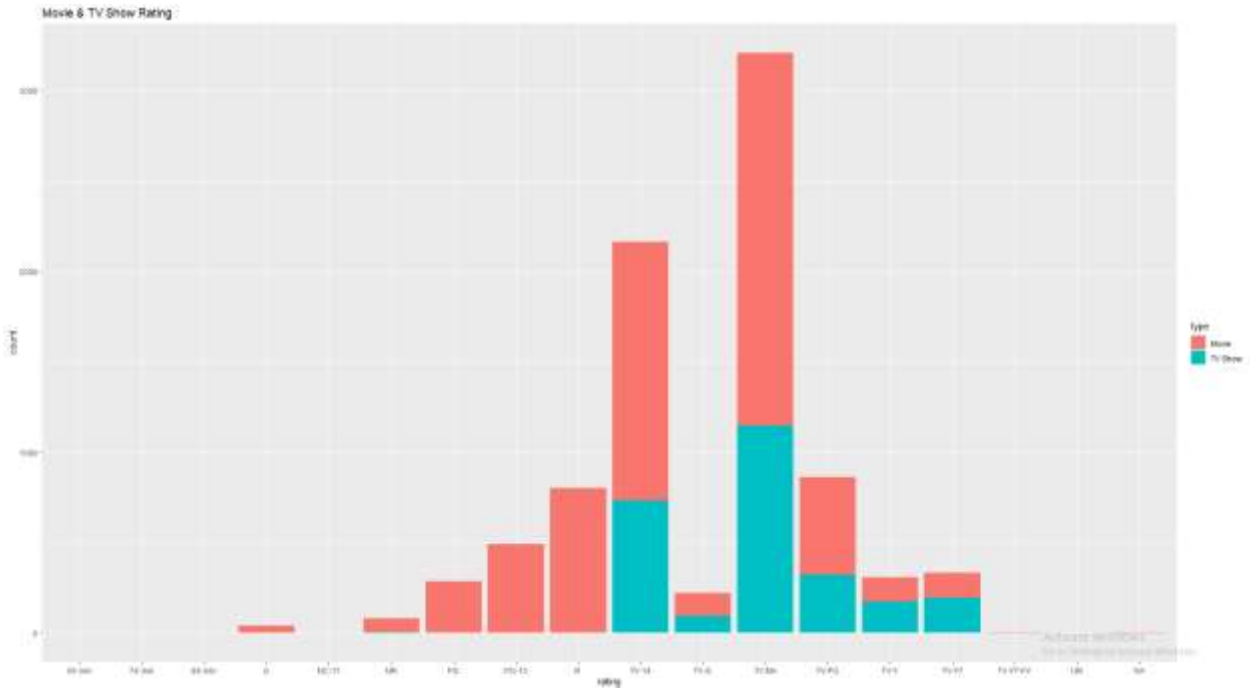
```
netflix %>%
  filter(release_year<"2021-01-01" & release_year>="1900-01-01") %>%
  group_by(release_year,type) %>%
  count() %>%
  ggplot(aes(x=release_year,y=n,fill=type),) +
  ggtitle("Movie & TV Show Release Year")+
  geom_line(aes(color=type), size = 2)
```



Dari grafik diatas, jumlah Movie dan TV Show di awal tahun 1900an hingga 1980an tidak berbeda jauh dan cenderung sedikit. Tetapi setelah tahun 2000, Movie dan TV Show lebih banyak diproduksi dengan Movie mencapai lebih dari 700 movies.

Visualisasi rating acara Netflix

```
netflix %>%
  ggplot(aes(x=rating,fill=type))+
  ggtitle("Movie & TV Show Rating")+
  geom_bar()
```



Berdasarkan grafik di atas, rating Movie & TV Show di Netflix didominasi oleh TV-14 (Konten untuk anak usia 14 tahun keatas) dan TV-MA (Konten untuk orang dewasa atau 18 keatas)

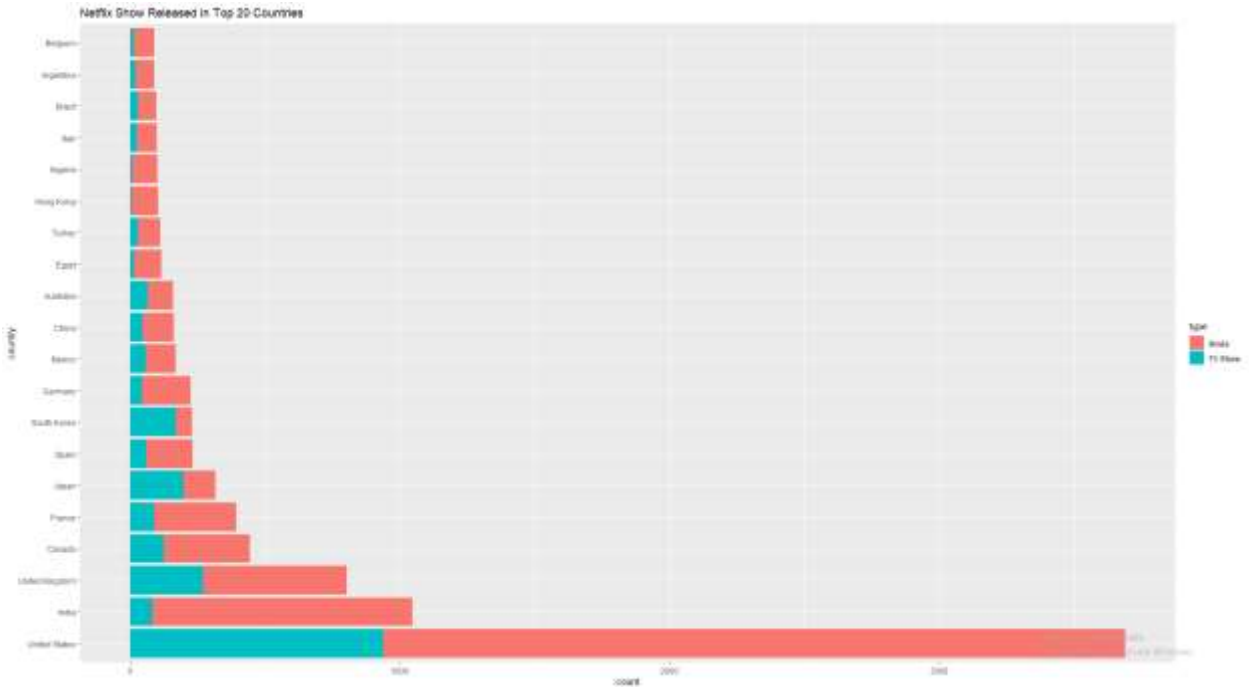
Visualisasi netflix show yang dirilis di 20 negara terbanyak

```
netflix %>%
  group_by(country) %>%
  count()

netflix %>%
  separate_rows(country,sep=", ") %>%
  mutate(country=fct_infreq(country)) %>%
  group_by(type) %>%
  ggplot(aes(country,fill=type))+
  geom_histogram(stat = 'count')+
  coord_cartesian(xlim = c(1,10))

top_countries <- netflix %>%
  separate_rows(country,sep=", ") %>%
  group_by(country) %>%
  summarise(countt=n()) %>%
  remove_missing() %>%
  arrange(desc(countt))
tc <- as.vector(top_countries$country[1:20])
```

```
netflix %>%
  separate_rows(country,sep=", ") %>%
  filter(country %in% tc) %>%
  mutate(country=factor(country,levels = tc)) %>%
  group_by(country,type) %>%
  summarise(count=n()) %>%
  ggplot(aes(x=country,y=count,fill=type))+
  ggtitle("Netflix Show Released in Top 20 Countries")+
  geom_bar(stat='identity')+
  coord_flip()
```



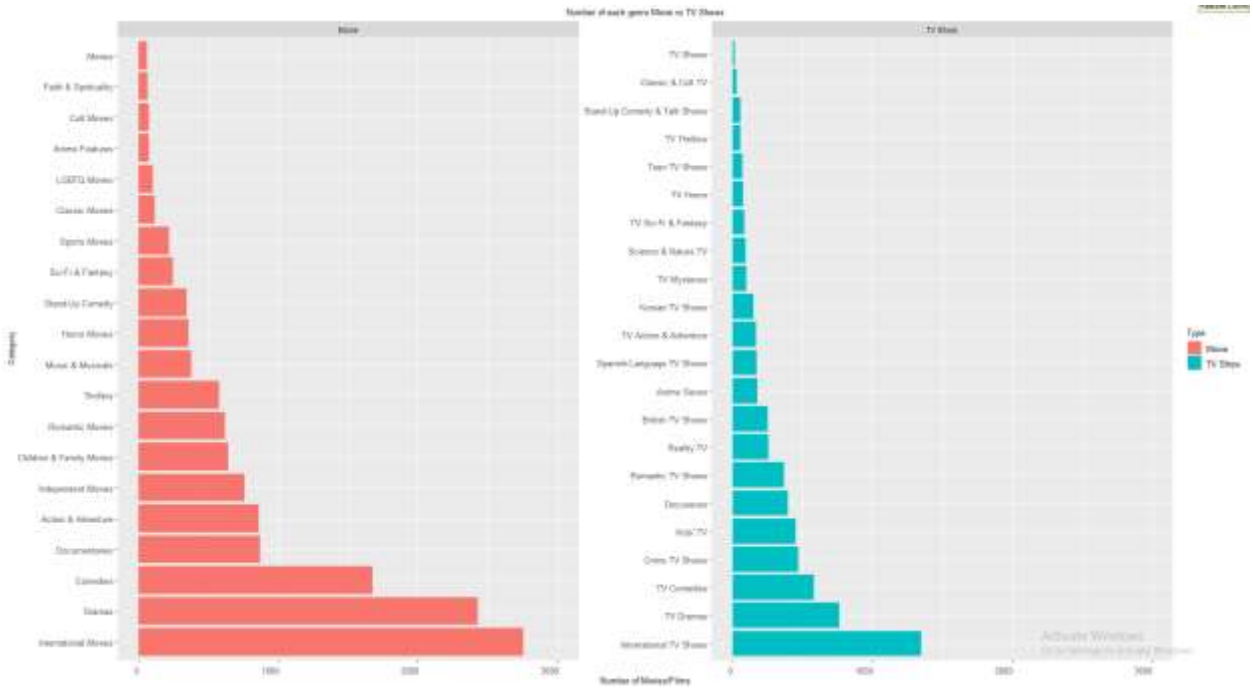
Berdasarkan grafik di atas, United Stated menjadi negara dengan acara Netflix terbanyak di dunia dengan total lebih dari 3000 acara, India di posisi kedua, dan United Kingdom di posisi ketiga.

Visualisasi Genre Movie & TV Show di Netflix

```
genre_number <- netflix %>%
  drop_na(listed_in) %>%
  select(c('type','listed_in')) %>%
  separate_rows(listed_in, sep = ',') %>%
  rename(Category = listed_in)

genre_number$Category <- trimws(genre_number$Category)
```

```
options(repr.plot.width = 14, repr.plot.height = 8)
genre_number %>%
  mutate(Category = fct_infreq(Category)) %>%
  ggplot(genre_number,
         mapping = aes(x = Category,fill=type)) +
  geom_bar() +
  facet_wrap(~type, scales = 'free_y') +
  theme(plot.title = element_text(hjust=0.5,size=10),
        axis.text.x = element_text(size=10,hjust=0.95,vjust=0.2),
        axis.text.y = element_text(size=10),
        axis.title = element_text(size=10),
        legend.text=element_text(size=10),
        legend.title=element_text(size=10)) +
  labs(x='Category',
       y='Number of Movies/Films',
       title = 'Number of each genre Movie vs TV Shows')  +
  scale_fill_discrete(name='Type') +
  coord_flip()+
  ylim(0,3000)
```



Berdasarkan grafik diatas, Movie dan TV Show di Netflix memiliki urutan genre terbanyak di tiga besar yang sama, yaitu comedy, drama, dan International movies/TV show.

D. Discussion/Analysis

Dari data-data di atas, kita bisa melihat berbagai macam perbandingan dari variable-variable dataset netflix. Selain itu, kita juga bisa melihat distribusi data-data netflix dengan menggunakan R, baik itu genre, type, rating, maupun negara. Secara sekilas kita bisa melihat bahwa Movie mendominasi sebagian besar acara di Netflix. Rating acara terbanyak di Netflix merupakan

rating TV-MA (konten acara untuk 18 tahun ke atas). United States menjadi negara dengan acara Netlfix terbanyak di dunia, diikuti oleh India, lalu United Kingdom. Di genre acara Netflix sendiri, International Movie dan International TV Show menjadi genre terbanyak yang ada di Netflix.