

# STAT2604 Assignment 1

Chan Hou Long, Guyver

10/11/2021

1.

```
Fibonacci <- numeric(20)
Fibonacci[1] <- 0
Fibonacci[2] <- 1
for (i in 3:20) Fibonacci[i] <- Fibonacci[i - 2] + Fibonacci[i - 1]
print("First 20 Fibonacci numbers:")
```

```
## [1] "First 20 Fibonacci numbers:"
```

```
print(Fibonacci)
```

```
## [1] 0 1 1 2 3 5 8 13 21 34 55 89 144 233 377
## [16] 610 987 1597 2584 4181
```

2.

```
newton <- function(f, tol=1e-7, x0=2, N=100) {
  h <- 1e-7
  i <- 1; x1 <- x0
  p <- numeric(N)

  while (i <= N) {
    df.dx <- (f(x0+h)-f(x0))/h
    x1 <- (x0 - (f(x0)/df.dx))
    p[i] <- x1
    i <- i + 1
    if (abs(x1-x0) < tol) break
    x0 <- x1
  }
  return(p[1:(i-1)])
}

f <- function(x) {
  x^3 - 3*x - 3
}

newton(f, x0=2)
```

```
## [1] 2.111111 2.103836 2.103803 2.103803
```

### 3.

Netflix has become my favourite streaming content platform replacing traditional TV channels. I watch movies and TV shows, especially documentaries almost every day. Hence, I would like to look into this entertainment platform and wonder about the statistic about it. Are movies more than TV shows? Do the contents add more after popular? What regions provides the most content?

```
library(tidyverse)
library(ggplot2)

# Read csv file
netflix <- read.csv("/Users/guyverchan/Documents/HKU/STAT2604/netflix_titles.csv")

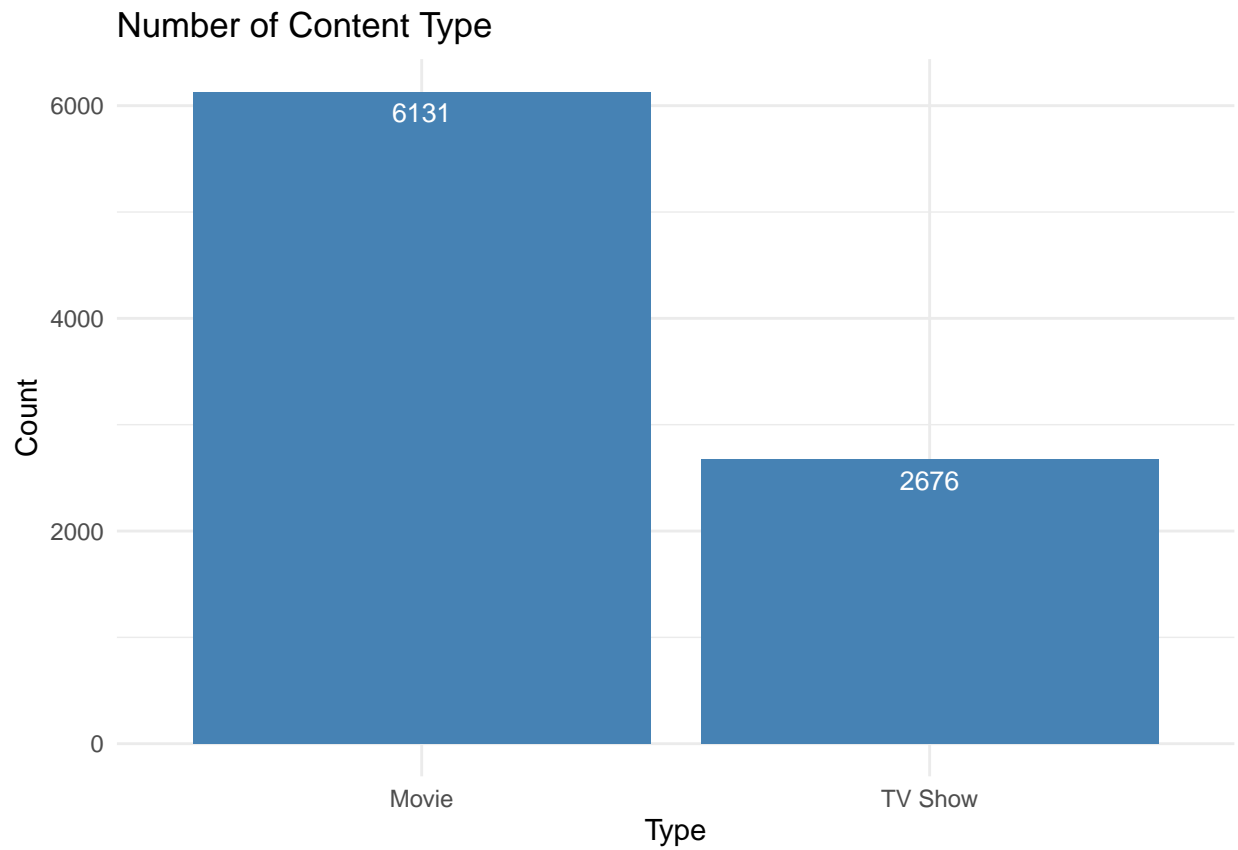
# First I would like to remove descriptive columns.
netflix <- netflix %>% select( -c(description,cast,director,listed_in))
# Then I separate data to ensure only single value in a cell.
netflix <- separate( netflix , country , into = c( "region" ) , sep = ",")
netflix <- netflix %>% separate ( date_added , into = c("month_added" , "year_added" ) , sep=", " )
netflix <- netflix %>% separate ( month_added , into = c("month_added" ) , sep= " ")

# First I would like to remove all empty values to increase the accuracy of the test.
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
netflix$region <- ifelse( is.na(netflix$region),
                        getmode(netflix$region),
                        netflix$region)
netflix$rating <- ifelse( is.na(netflix$rating),
                        getmode(netflix$rating),
                        netflix$rating)
netflix$year_added <- ifelse( is.na(netflix$year_added),
                        getmode(netflix$year_added),
                        netflix$year_added )
```

From the bar chart, it can be seen the movie is triple the number of the TV show.

```
# Find type total
## I group the data by category "type", then count the number of each type.
type_total <- netflix %>% group_by(type) %>% summarise(total=n())

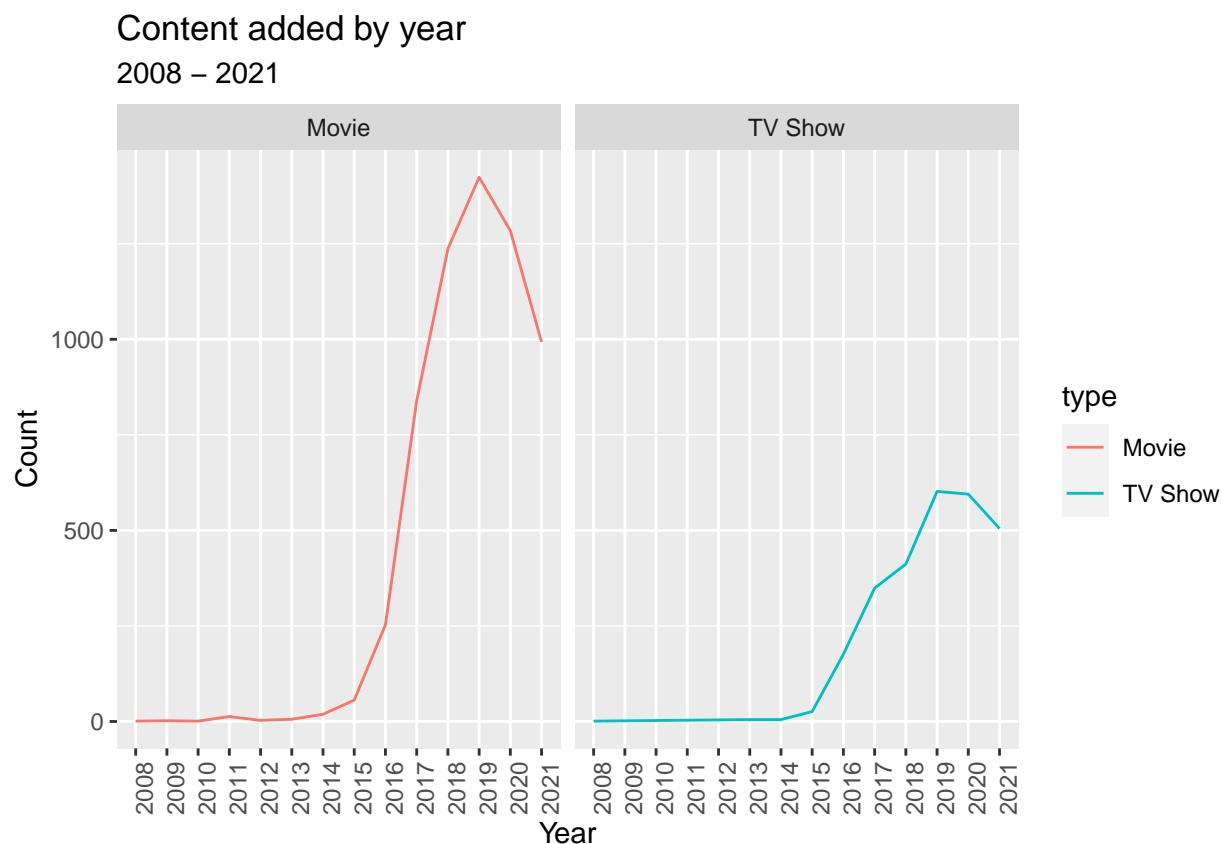
ggplot(data=type_total, aes(x=type, y=total)) +
  geom_bar(stat="identity", fill="steelblue") +
  geom_text(aes(label=total), vjust=1.6, color="white", size=3.5) +
  theme_minimal() +
  labs(title="Number of Content Type", x="Type",y="Count")
```



From the line chart, it can be seen that Netflix presented more films than TV shows starting from 2015 each year and kept increasing steadily. 2019 provided the most number of movies and TV shows. However, due to the pandemic, the productions dropped each year afterwards.

```
# Content added by year
## I group the data by category "type" and year that added into the platform, then count the number of
c_year <- netflix %>% group_by(type,year_added) %>% summarise(count=n())

ggplot(data=c_year,aes(x=year_added,y=count,group=1,color=type)) +
  geom_line()+facet_wrap(~type)+
  theme(axis.text.x=element_text(angle=90))+
  labs(title="Content added by year", subtitle="2008 - 2021", x="Year",y="Count")
```



By selecting the top 10 of each type, it can be seen that the US provides the most content in both films and TV shows. I am surprised that India is the second big provider of movies. UK is also an essential content marker that contributes third in movies and second in TV shows. After that, it is Japan and South Korea, which is understandable as people especially in HK are addicted to their series.

```
# Regions that provided the most
## I group the data by category "type" and "region", then count the number of each region by type.
type_movie <- netflix %>% group_by(type, region) %>% summarise(count=n())
type_movie[type_movie==""] <- NA
type_movie <- type_movie[rowSums(is.na(type_movie)) == 0,]

ggplot(data = top_n(type_movie, n=10, count), aes(x = region, y = count)) +
  geom_bar(stat="identity", fill="steelblue") +
  scale_x_discrete() + facet_wrap(~type, scales = 'free_x') +
  theme(axis.text.x=element_text(angle=90))
```

