Module 6 Project

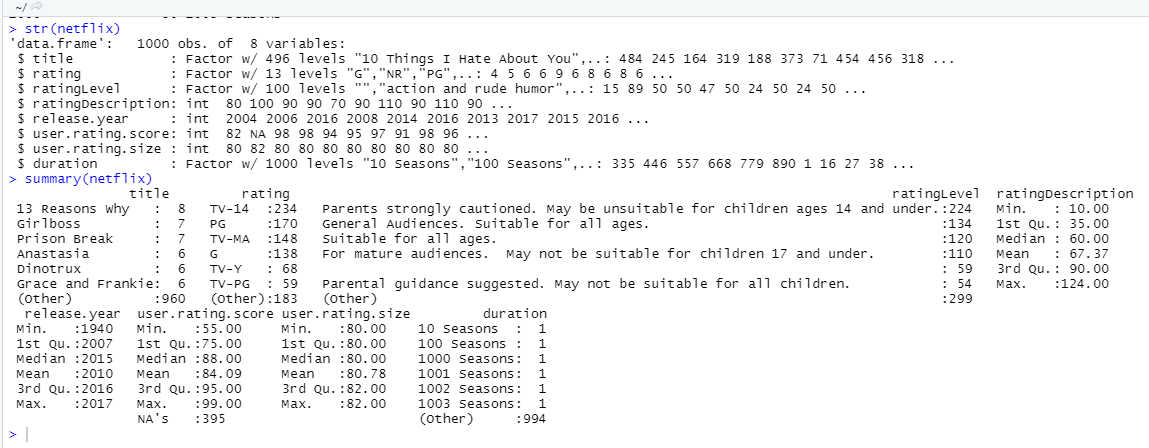
**Executive Summary Report 6**

**Name**: Rupali S. Metkari

**Date**: 27th February 2021

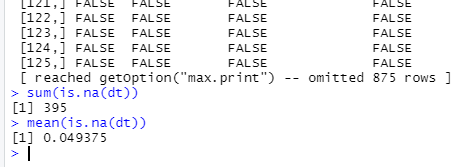
**Data Key Points**

We downloaded Netflix shows data from the Kaggle. In this data have 1000 records with 8 variables.



The given dataset consists of Shows and Movies from the year 1940 to 2017, available on Netflix, a worldwide popular video-streaming application. The company’s primary business is its subscription-based streaming service, which offers online streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries, and more on thousands of internet-connected devices. Netflix is a popular entertainment service used by people around the world.

Details included about each show are rating, rating description, rating level, release year, user rating score, user rating score, user rating size and duration.



**Problems found in the dataset:**

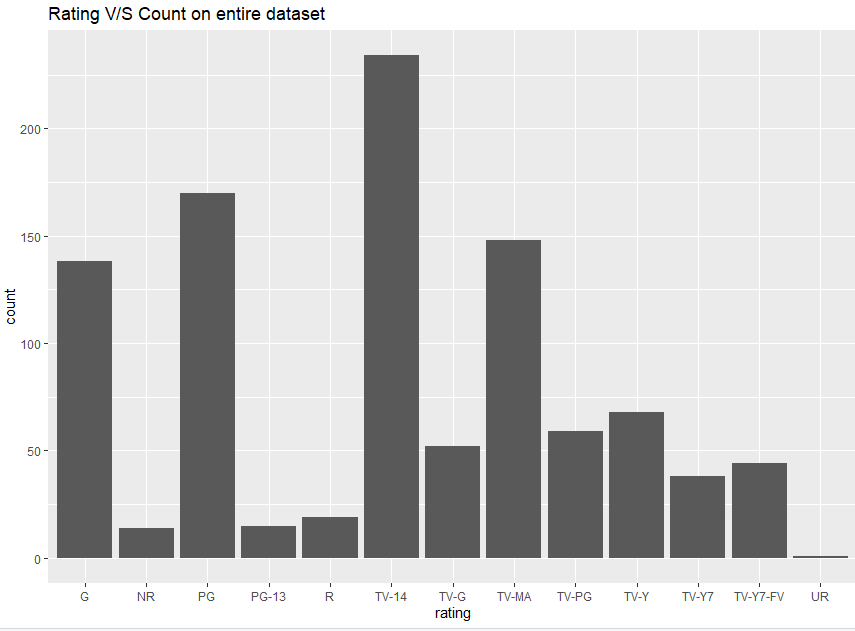
There are several missing values in some of the rows for user.rating.score. I initially found how many values were missing and found only one column had missing values. Hence I replaced the missing values with mode of the entire column.

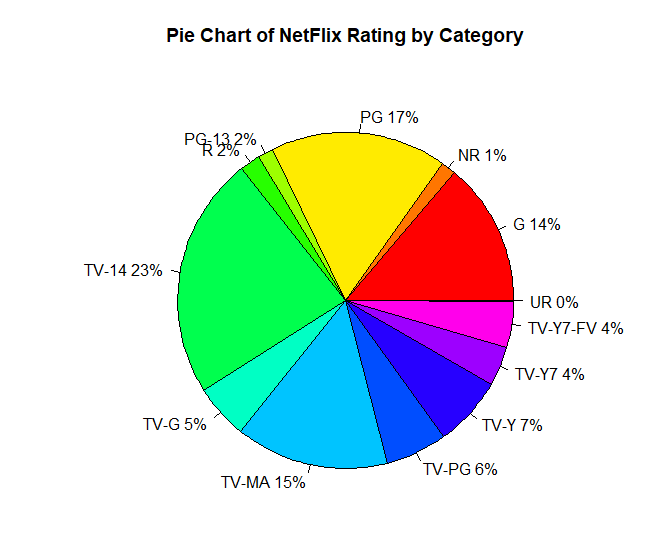
is.na(netflix)

netflix[is.na(netflix)] <- names(sort(-table(netflix$user.rating.score)))[1]

is.na(netflix)

**Rating System:**

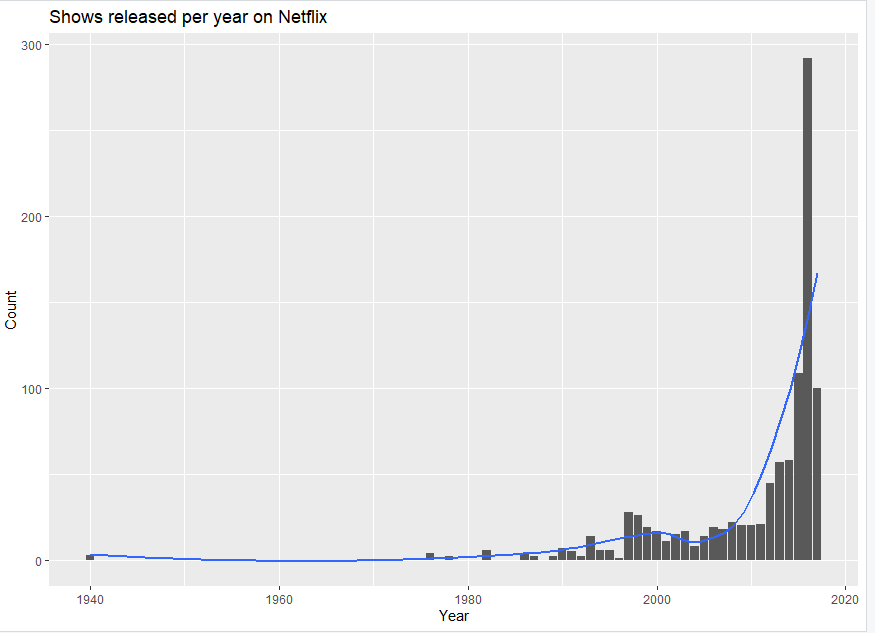




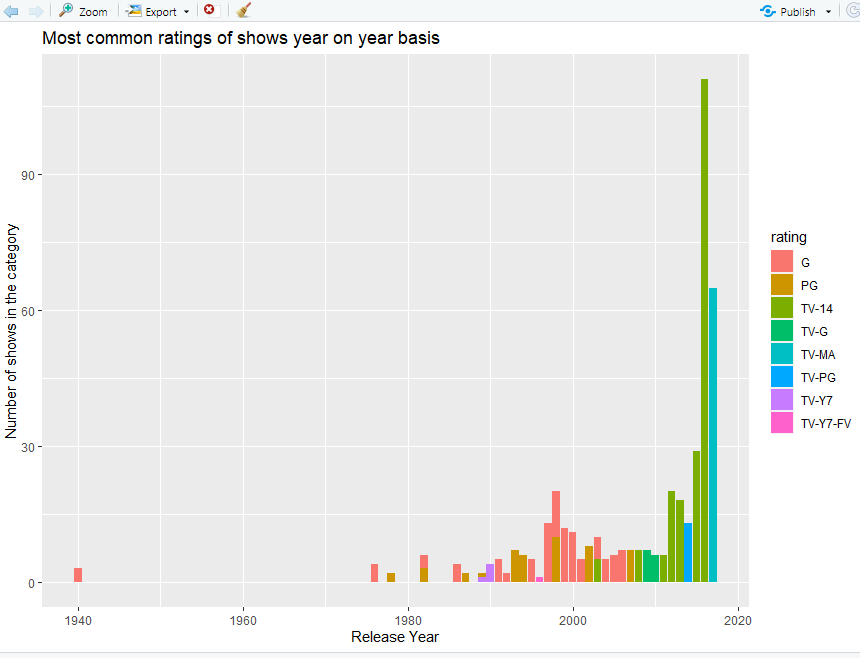


Here we can see there are total 13 types of rating types with frequency. Like “G” (General Audiences) , “R”(strong violence, sexual content and adult language).

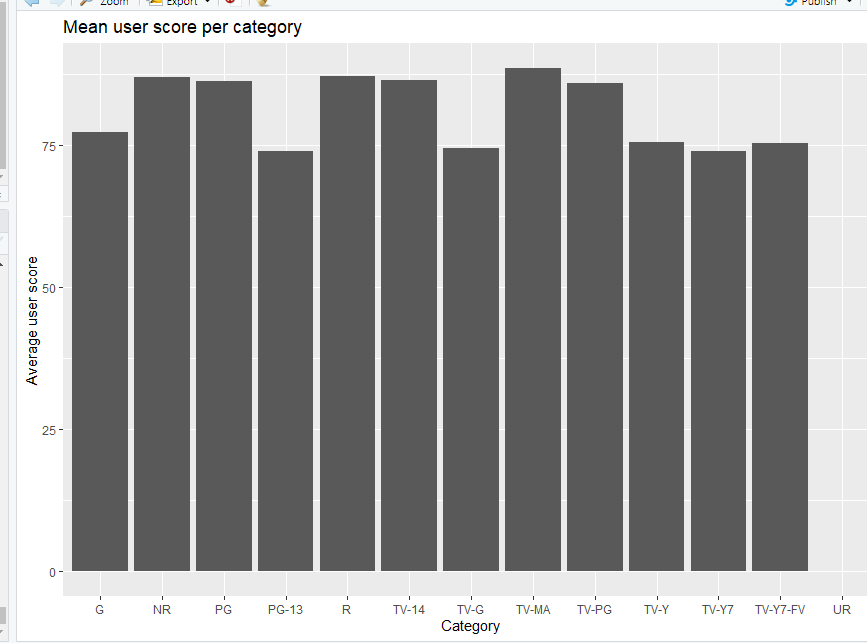
TV-14 rating is highest rating, so TV-14 is the most common type of show.



Mostly the shows / movies available are those released after 1990. After 1990 released shows frequency is increasing. Between 2015-to-2017-year highest rate of release shows.



We can see that up until 2010 ‘G’ type shows was mostly released after that ‘TV – 14’ type shows is become popular.



Here graph shows avg user rating score per rating category, So we found that NR, PG, R, TV-14, TV-MA, TV-PG these category shows are high rating score. These score finds from the consumers. User interested these 6 types of categories shows.

Conclusion:

Here we found that peoples are so much interested to watching tv series, movies, streaming video on Netflix. After year 1990 Netflix has been made more popular and keep going. Netflix have multiple categories shows available so that people have many choices to watching shows with family, children. When user search specific category on Netflix then Netflix recommended same type shows with same rating.

**Bibliography**

1. Robert I. Kabacoff (2011). *R in action Data analysis and graphics with R*. Shelter Island, New York: Publisher.
2. Bluman Alan(2009). *Elementary Statistics A Step by Step Approach, 7th edition.* New York: Publisher.
3. Kaggle Data Source: *https://www.kaggle.com/chasewillden/netflix-shows*.
4. Environmental Computing : Publisher[*http://environmentalcomputing.net/plotting-with-ggplot-colours-and-symbols/*](http://environmentalcomputing.net/plotting-with-ggplot-colours-and-symbols/)
5. The Analysis factor [*https://www.theanalysisfactor.com/linear-models-r-plotting-regression-lines/*](https://www.theanalysisfactor.com/linear-models-r-plotting-regression-lines/). The Analysis factor: Publisher
6. Stack overflow: https://stackoverflow.com/questions/44429398/how-to-plot-pie-chart-in-r/44429741
7. R plot Symbols. [*http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r*](http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r)
8. Youtube : https://www.youtube.com/watch?v=D\_CNmYkGRUc&list=WL&index=12
9. Youtube *https://www.youtube.com/watch?v=DPAsYtUHkkM&list=WL&index=11*
10. STHDA*.* [*http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization*](http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization)
11. Cookbook of R. [*http://www.cookbook-r.com/Graphs/Bar\_and\_line\_graphs\_(ggplot2)*](http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2))

**Appendix**

#import library

library(ggplot2)

library(plyr)

library(dplyr)

library(magrittr)

library(tidyverse)

library(scales)

#loaded netflix.csv

netflix <- read.csv("D:/Assignmentas/corona/Netflix Shows.csv")

#Display the head, tail and structure of <bio>

head(netflix)

tail(netflix)

str(netflix)

summary(netflix)

table(netflix$rating)

a<-unique(netflix$rating)

length(a)

#find null values

is.na(netflix)

netflix[is.na(netflix)] <- names(sort(-table(netflix$user.rating.score)))[1]

is.na(netflix)

#plot histogram for rating

ggplot(netflix, aes(x = rating )) + geom\_histogram(stat = "count") + ggtitle("Rating V/S Count on entire dataset")

# draw pie chart

slices <- table(netflix$rating)

lbls <- levels(netflix$rating)

pct <- round(slices/sum(slices)\*100)

lbls <- paste(lbls, pct)

lbls <- paste(lbls,"%",sep="")

pie(slices,labels = lbls, col=rainbow(length(lbls)),main="Pie Chart of NetFlix Rating by Category")

#Lets see when the movies / shows are released on netflix

releases\_per\_year <- netflix %>% select(title, release.year) %>% group\_by(release.year) %>% summarise( movies\_per\_year = n()) %>% arrange(desc(movies\_per\_year))

ggplot(releases\_per\_year, aes(x = (release.year), y = movies\_per\_year)) + geom\_bar(stat = "identity") + geom\_smooth( se = F) + xlab("Year") + ylab("Count") + ggtitle("Shows released per year on Netflix")

#Which type of show / movie is mostly common year on year basis.

ratings\_per\_year <- netflix %>% group\_by(rating, release.year) %>% summarise(count = n())

ratings\_per\_year <- ratings\_per\_year %>% group\_by(release.year) %>% filter(count == max(count))

ggplot(ratings\_per\_year, aes(x= release.year, y = count, fill = rating)) + geom\_bar(stat = "identity") + xlab("Release Year") + ylab("Number of shows in the category") + ggtitle("Most common ratings of shows year on year basis")

#avg user rating score per rating category

avg\_score <- netflix %>% group\_by(rating) %>% summarise(avg\_score = mean(user.rating.score, na.rm = T))

ggplot(avg\_score, aes(x = as.factor(rating), y = avg\_score)) + geom\_bar(stat = "identity") + xlab("Category") + ylab("Average user score") + ggtitle("Mean user score per category")

ggplot(netflix,aes(x=release.year)) + geom\_bar(stat='count') + scale\_x\_continuous(limits=c(1995,2017))

ggplot(netflix,aes(x=release.year,y=user.rating.size)) + geom\_bar(stat='identity') + scale\_x\_continuous(limits=c(1995,2017))