Assignment 1

Exploratory data analysis

(Dataset: Best Show by Year Netflix, Data Source: Kaggle.com)

Submitted by,

Varshini.S

Submitted to,

Dr.V.Bhuvaneshwari

ASSUMPTION:

- The dataset explains about the Best show by year in Netflix.
- Score is the most important thing in the Netflix, based on the reviews we watch movies .
- There are different types of genre in Netflix ,each person has various perspective , and they express their opinions.
- The more you watch a movie ,the better you review movies.
- Score and genre plays a important role in the Netflix, because they can be used to calculate the best show in Netflix.

```r
library(lattice)
library(ggplot2)
library(dplyr)

library(MASS)

library(tidyverse)

df2 <- read.csv('Best Show by Year Netflix.csv')


#summary
summary(df2)
```

```
##      index         TITLE            RELEASE_YEAR      SCORE
##  Min.   : 0.0   Length:31          Min.   :1969   Min.   :6.700
##  1st Qu.: 7.5   Class :character   1st Qu.:2000   1st Qu.:8.500
##  Median :15.0   Mode  :character   Median :2007   Median :8.800
##  Mean   :15.0                      Mean   :2006   Mean   :8.606
##  3rd Qu.:22.5                      3rd Qu.:2014   3rd Qu.:8.950
##  Max.   :30.0                      Max.   :2022   Max.   :9.500
##  NUMBER_OF_SEASONS  MAIN_GENRE        MAIN_PRODUCTION
##  Min.   : 1.000    Length:31          Length:31
##  1st Qu.: 3.000    Class :character   Class :character
##  Median : 5.000    Mode  :character   Mode  :character
##  Mean   : 5.323
##  3rd Qu.: 6.000
##  Max.   :21.000
```

```r
str(df2)
```

```
## 'data.frame':    31 obs. of  7 variables:
##  $ index            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ TITLE            : chr  "Monty Python's Flying Circus" "Knight Rider" "
Seinfeld" "Star Trek: Deep Space Nine" ...
##  $ RELEASE_YEAR     : int  1969 1982 1989 1993 1995 1997 1998 1999 2000 20
01 ...
##  $ SCORE            : num  8.8 6.9 8.9 8.1 8.5 8.4 8.9 8.8 8.2 8.6 ...
##  $ NUMBER_OF_SEASONS: int  4 4 9 7 1 10 1 21 8 12 ...
##  $ MAIN_GENRE       : chr  "comedy" "action" "comedy" "scifi" ...
##  $ MAIN_PRODUCTION  : chr  "GB" "US" "US" "US" ...
```

```r
#charater to categorical
df2$MAIN_GENRE=as.factor(df2$MAIN_GENRE)
df2$MAIN_PRODUCTION=as.factor(df2$MAIN_PRODUCTION)
summary(df2)
```

```
##      index         TITLE            RELEASE_YEAR      SCORE
##  Min.   : 0.0   Length:31          Min.   :1969   Min.   :6.700
##  1st Qu.: 7.5   Class :character   1st Qu.:2000   1st Qu.:8.500
##  Median :15.0   Mode  :character   Median :2007   Median :8.800
```

```
##  Mean   :15.0                      Mean   :2006   Mean   :8.606
##  3rd Qu.:22.5                      3rd Qu.:2014   3rd Qu.:8.950
##  Max.   :30.0                      Max.   :2022   Max.   :9.500
##  NUMBER_OF_SEASONS       MAIN_GENRE MAIN_PRODUCTION
##  Min.   : 1.000    action     :4    CA: 2
##  1st Qu.: 3.000    comedy     :8    GB: 4
##  Median : 5.000    documentary:1    IN: 1
##  Mean   : 5.323    drama      :8    JP: 7
##  3rd Qu.: 6.000    scifi      :9    US:17
##  Max.   :21.000    western    :1
```

```
str(df2)
```

```
## 'data.frame':    31 obs. of  7 variables:
##  $ index            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ TITLE            : chr  "Monty Python's Flying Circus" "Knight Rider" "
Seinfeld" "Star Trek: Deep Space Nine" ...
##  $ RELEASE_YEAR     : int  1969 1982 1989 1993 1995 1997 1998 1999 2000 20
01 ...
##  $ SCORE            : num  8.8 6.9 8.9 8.1 8.5 8.4 8.9 8.8 8.2 8.6 ...
##  $ NUMBER_OF_SEASONS: int  4 4 9 7 1 10 1 21 8 12 ...
##  $ MAIN_GENRE       : Factor w/ 6 levels "action","comedy",..: 2 1 2 5 5 5
6 1 2 2 ...
##  $ MAIN_PRODUCTION  : Factor w/ 5 levels "CA","GB","IN",..: 2 5 5 5 4 5 4
4 5 1 ...
```

```
#checking null values
colSums(is.na(df2))
```

```
##             index             TITLE      RELEASE_YEAR             SCORE
##                 0                 0                 0                 0
## NUMBER_OF_SEASONS        MAIN_GENRE   MAIN_PRODUCTION
##                 0                 0                 0
```

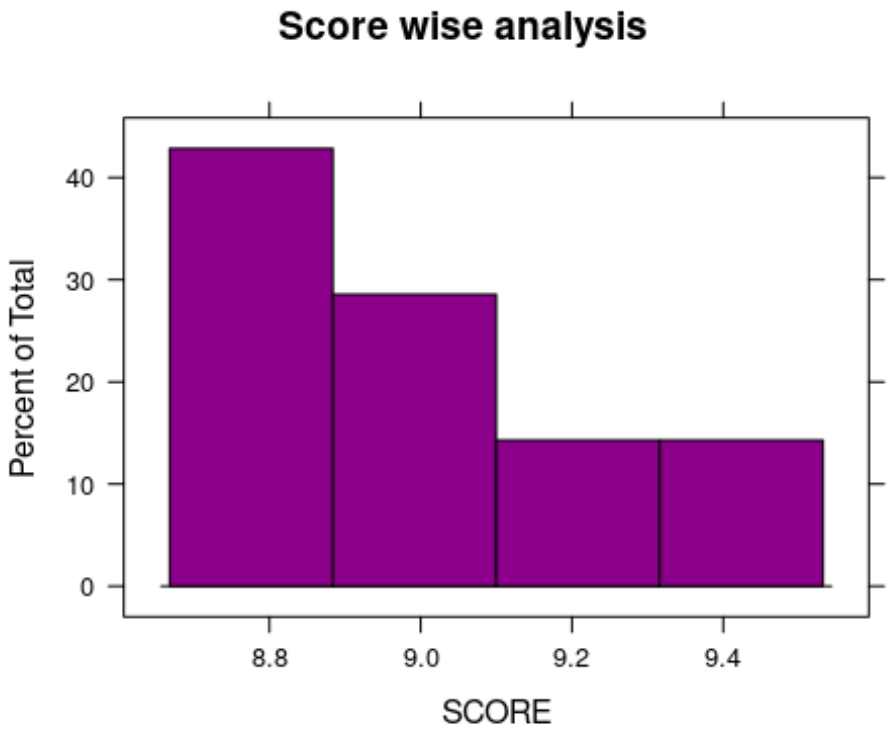
```
#dimensions
dim(df2)
```

```
## [1] 31  7
```

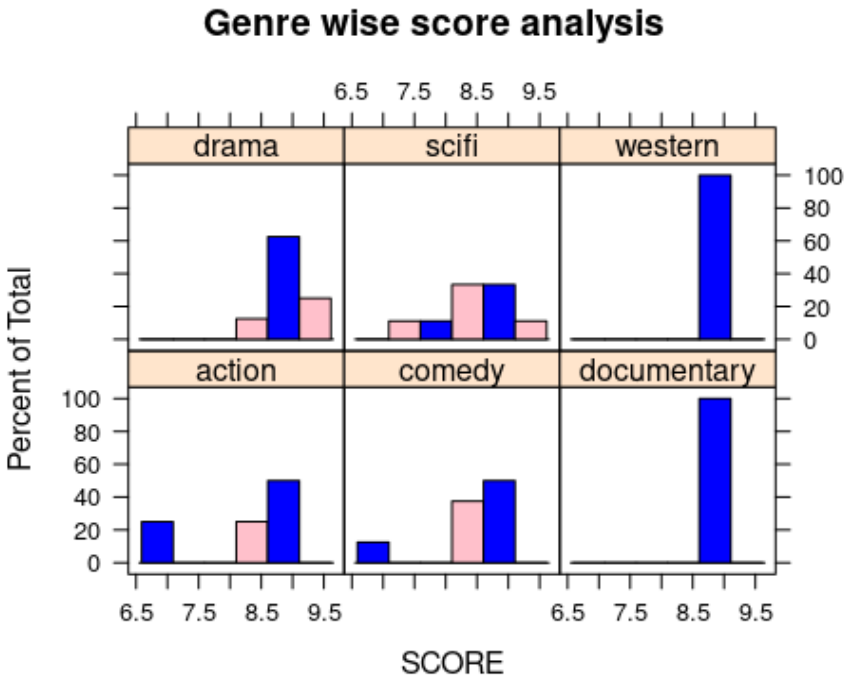
```
#Subsetting
df3<-subset(df2,MAIN_GENRE=='drama'&SCORE>8.5,select=c(MAIN_GENRE,SCORE))
head(df3)
```

```
##    MAIN_GENRE SCORE
## 17      drama   9.5
## 19      drama   8.7
## 20      drama   9.0
## 23      drama   8.8
## 26      drama   8.7
## 28      drama   9.3
```

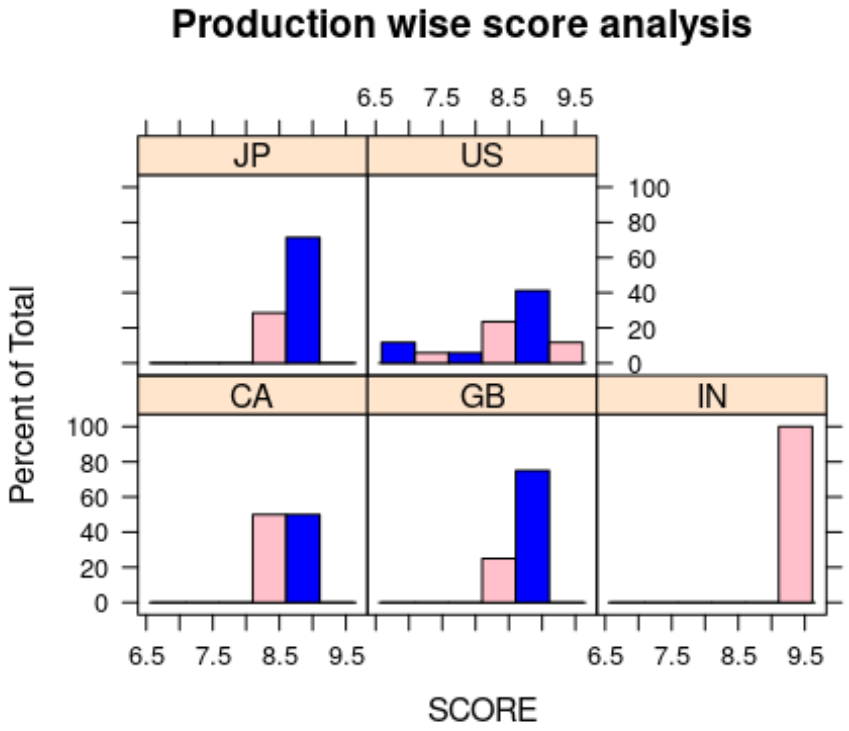
```
df4<-subset(df2,MAIN_GENRE=='scifi'&SCORE<8.5,select=c(MAIN_GENRE,SCORE))
head(df4)

##    MAIN_GENRE SCORE
## 4       scifi   8.1
## 6       scifi   8.4
## 11      scifi   8.4
## 13      scifi   7.3

df5<-subset(df2,MAIN_PRODUCTION=='US'&SCORE<8.5,select=c(MAIN_PRODUCTION,SCOR
E))
head(df5)

##    MAIN_PRODUCTION SCORE
## 2               US   6.9
## 4               US   8.1
## 6               US   8.4
## 9               US   8.2
## 13              US   7.3
## 16              US   6.7

#histogram
histogram(~SCORE,data=df3,col="darkmagenta",main="Score wise analysis")
```



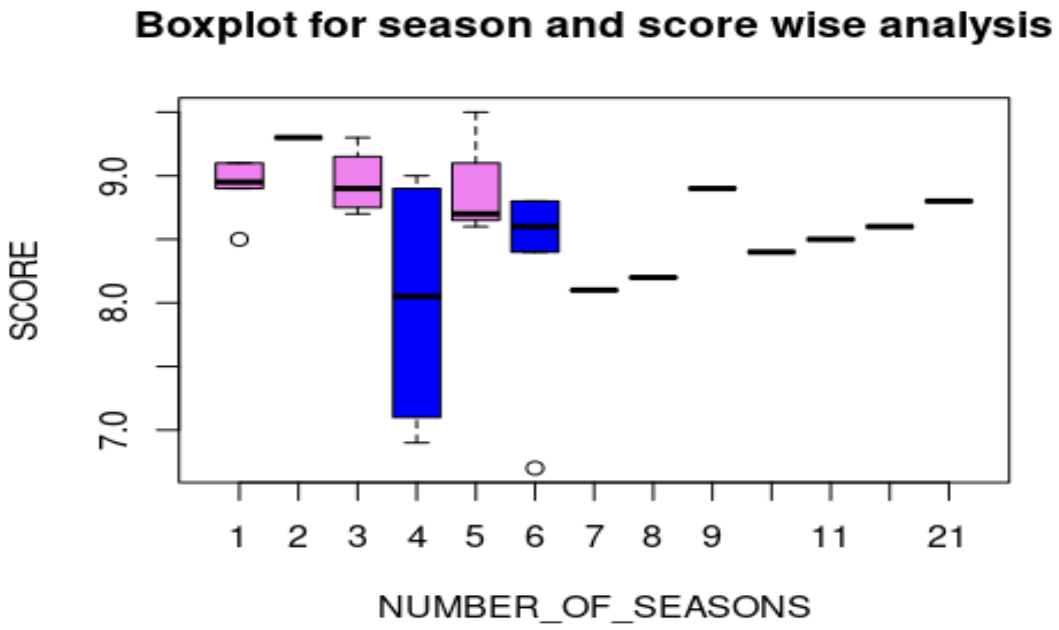
```
histogram(~SCORE|MAIN_GENRE,data=df2,col=c("blue","pink"),main="Genre wise sc
ore analysis")
```



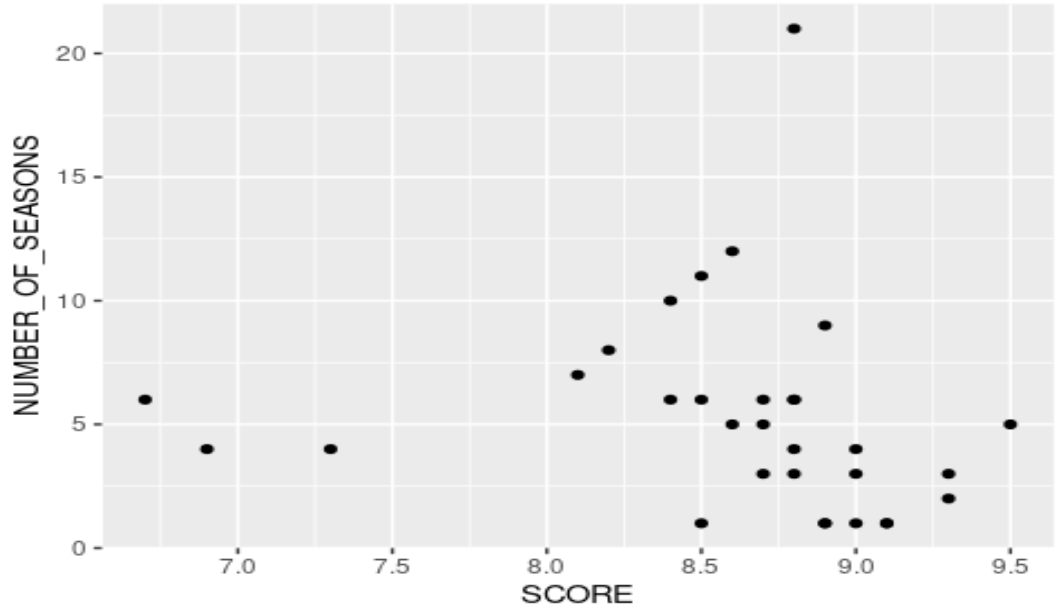

```
histogram(~SCORE|MAIN_PRODUCTION,data=df2,col=c("blue","pink"),main="Producti
on wise score analysis")
```



```r
#boxplot
boxplot(SCORE~NUMBER_OF_SEASONS,data=df2,col=c("violet","blue"),main="Boxplot
for season and score wise analysis")
```




```r
#scatterplot
ggplot(df2, aes(x = SCORE, y = NUMBER_OF_SEASONS)) +
  geom_point()
```



ATTRIBUTE DESCRIPTION :

- The "Best Show By Year Netflix" dataset has 7 variables and 31 observations.
- The dataset is filled with integer , numerical ,and character data types.
- There are 31 titles and it is released in different years, each movie is reviewed at a rate of 6.5-9.5 rating .
- Number_of_seasons vary for every movies.
- Main_genre of the movies are drama , scifi , western , documentary , comedy and action.
- Main_production of the movies are US , JP , GP, CA , IN.

INFERENCE:

HISTOGRAM

- Starting the analysis with 'Genre and Score' attribute . Western and the Documentary are the only movies with 100 percentage and rated with 8.6-9.1 compared with other movies. In scifi the movies are spread in the range of 7.0-9.5, action and comedy movies almost have same number of ratings and drama is rated above 8.0
- In productions, CA has produced with 50 percentage , US is the only movie production having the movies rated from 6.0-9.5 and IN has 100 percentage and rated above 8.6.

BOXPLOT

- We are analyzing Number_of_seasons and Score, in season 1 there is no maximum and minimum values the median lies near the Q1 quartile so the movies rated are above 9. It is positively skewed.

- In season 3 there are maximum and minimum values ,the median lies near the minimum so it is positively skewed.
- In season 4 the median lies in the center so it has no skewness.

SCATTER PLOT

- X axis is plotted as the Score and the Y axis is plotted as the Number_of_seasons .
- NO CORRELATION because the points are scattered all over the plot it is difficult to conclude whether it is increasing or decreasing.

INSIGHTS:

- Western and documentary genre are the highly rated movies.
- All other genre are moderate rated movies
- US production has produced movies rated from moderate to high.
- Movies produced in other production are ranged above moderate.
- In season 3 and 5 it gradually increases the score of the movies
- In season 4 and 6 the movies rated is equally.