lab3.R

sathvik

2021-01-31

```r
# 171EC146
# Sathvik S Prabhu

# Loading the dataset into R
library(readxl)
d<-read_excel("/home/sathvik/EC8/ML/Lab/Lab3/dataset.xlsx",sheet=2)
d
```

```
## # A tibble: 150 x 10
##    `S No` `Release Date`      `Movie Name` `Release Date (~ `Genre - Define~
##     <dbl> <dttm>              <chr>        <chr>            <chr>
##  1      1 2014-04-18 00:00:00 2 States     LW               Romance
##  2      2 2013-01-04 00:00:00 Table No. 21 N                Thriller
##  3      3 2014-07-18 00:00:00 Amit Sahni ~ N                Comedy
##  4      4 2013-01-04 00:00:00 Rajdhani Ex~ N                Drama
##  5      5 2014-07-04 00:00:00 Bobby Jasoos N                Comedy
##  6      6 2014-05-30 00:00:00 Citylights   HS               Drama
##  7      7 2014-09-19 00:00:00 Daawat-E-Is~ N                Comedy
##  8      8 2013-01-11 00:00:00 Matru Ki Bi~ N                Comedy
##  9      9 2014-01-10 00:00:00 Dedh Ishqiya LW               Comedy
## 10     10 2013-01-11 00:00:00 Gangoobai    N                Drama
## # ... with 140 more rows, and 5 more variables: Budget <dbl>, `Box Office
## #   Collection` <dbl>, `Youtube Views` <dbl>, `Youtube Likes` <dbl>, `Youtube
## #   Dislikes` <dbl>
```

```r
str(d) # gives data structure
```

```
## tibble [150 x 10] (S3: tbl_df/tbl/data.frame)
##  $ S No                          : num [1:150] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Release Date                  : POSIXct[1:150], format: "2014-04-18" "2013-01-04" ...
##  $ Movie Name                    : chr [1:150] "2 States" "Table No. 21" "Amit Sahni Ki List" "Rajd
##  $ Release Date (N / LW / Festive): chr [1:150] "LW" "N" "N" "N" ...
##  $ Genre - Defined               : chr [1:150] "Romance" "Thriller" "Comedy" "Drama" ...
##  $ Budget                        : num [1:150] 36 10 10 7 18 7 30 33 31 1.8 ...
##  $ Box Office Collection         : num [1:150] 104 12 4 0.35 10.8 35 24.6 40 27 0.01 ...
##  $ Youtube Views                 : num [1:150] 8576361 1087320 572336 42626 3113427 ...
##  $ Youtube Likes                 : num [1:150] 26622 1129 586 86 4512 ...
##  $ Youtube Dislikes              : num [1:150] 2527 137 54 19 1224 ...
```

```r
summary(d) # gives minimum, Q1, median, mean, Q3, maximum
```

```
##       S No      Release Date                  Movie Name
##  Min.   :  1   Min.   :2013-01-04 00:00:00   Length:150
##  1st Qu.: 38   1st Qu.:2013-06-28 00:00:00   Class :character
##  Median : 75   Median :2014-02-07 00:00:00   Mode  :character
##  Mean   : 75   Mean   :2014-01-11 08:41:52
##  3rd Qu.:112   3rd Qu.:2014-07-04 00:00:00
##  Max.   :149   Max.   :2015-03-20 00:00:00
##  NA's   :1     NA's   :1
##  Release Date (N / LW / Festive) Genre - Defined        Budget
##  Length:150                      Length:150         Min.   :  1.80
##  Class :character                Class :character   1st Qu.: 11.00
##  Mode  :character                Mode  :character   Median : 21.00
##                                                     Mean   : 29.43
##                                                     3rd Qu.: 35.00
##                                                     Max.   :150.00
##                                                     NA's   :1
##  Box Office Collection Youtube Views      Youtube Likes    Youtube Dislikes
##  Min.   :  0.010       Min.   :    4354   Min.   :     1   Min.   :    1
##  1st Qu.:  9.085       1st Qu.: 1076591   1st Qu.:  1377   1st Qu.:  189
##  Median : 28.100       Median : 2375050   Median :  4111   Median :  614
##  Mean   : 60.196       Mean   : 3337920   Mean   :  7878   Mean   : 1208
##  3rd Qu.: 57.862       3rd Qu.: 4550051   3rd Qu.:  9100   3rd Qu.: 1419
##  Max.   :735.000       Max.   :23171067   Max.   :101275   Max.   :11888
##                        NA's   :1          NA's   :1        NA's   :1
```

```r
# A few variables have NA's. As Sl No. ranges from 1 to 149, the last row is excluded.

# Reading without NA's
d<-read_excel("/home/sathvik/EC8/ML/Lab/Lab3/dataset.xlsx",sheet=2, n_max=149)
d<-data.frame(d,stringsAsFactors = T)

str(d) # gives data structure
```

```
## 'data.frame':    149 obs. of  10 variables:
##  $ S.No                           : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Release.Date                   : POSIXct, format: "2014-04-18" "2013-01-04" ...
##  $ Movie.Name                     : chr  "2 States" "Table No. 21" "Amit Sahni Ki List" "Rajdhani Ex
##  $ Release.Date..N...LW...Festive.: chr  "LW" "N" "N" "N" ...
##  $ Genre...Defined                : chr  "Romance" "Thriller" "Comedy" "Drama" ...
##  $ Budget                         : num  36 10 10 7 18 7 30 33 31 1.8 ...
##  $ Box.Office.Collection          : num  104 12 4 0.35 10.8 35 24.6 40 27 0.01 ...
##  $ Youtube.Views                  : num  8576361 1087320 572336 42626 3113427 ...
##  $ Youtube.Likes                  : num  26622 1129 586 86 4512 ...
##  $ Youtube.Dislikes               : num  2527 137 54 19 1224 ...
```

```r
# 10 variables, each with 149 observations

summary(d[c("Budget","Box.Office.Collection","Youtube.Views","Youtube.Likes","Youtube.Dislikes")]) # gi
```

```
##      Budget       Box.Office.Collection Youtube.Views      Youtube.Likes
##  Min.   :  1.80   Min.   :  0.01        Min.   :    4354   Min.   :     1
##  1st Qu.: 11.00   1st Qu.:  8.78        1st Qu.: 1076591   1st Qu.:  1377
##  Median : 21.00   Median : 28.00        Median : 2375050   Median :  4111
##  Mean   : 29.43   Mean   : 55.67        Mean   : 3337920   Mean   :  7878
##  3rd Qu.: 35.00   3rd Qu.: 57.45        3rd Qu.: 4550051   3rd Qu.:  9100
##  Max.   :150.00   Max.   :735.00        Max.   :23171067   Max.   :101275
##  Youtube.Dislikes
##  Min.   :    1
##  1st Qu.:  189
##  Median :  614
##  Mean   : 1208
##  3rd Qu.: 1419
##  Max.   :11888
```

```r
IQR(d$Box.Office.Collection) # Gives interquartile range Q3-Q1
```

```
## [1] 48.67
```

```r
factor(d$Release.Date..N...LW...Festive.)
```

```
##   [1] LW N  N  N  N  HS N  N  LW N  N  N  HS N  N  N  LW HS N  HS N  N  N  FS N 
##  [26] N  LW N  N  HS N  FS HS N  N  N  HS N  N  HS FS N  N  N  N  N  N  N  HS HS
##  [51] N  N  LW N  N  HS N  N  N  LW HS LW FS N  N  N  N  HS N  N  N  N  N  N  N 
##  [76] N  N  N  LW N  LW N  N  LW N  HS N  N  N  HS N  N  FS N  N  N  LW N  N  HS
## [101] N  N  N  N  N  N  FS N  N  N  N  N  N  N  LW FS FS N  N  FS N  FS FS N  N 
## [126] FS FS FS FS N  LW LW LW HS N  N  N  N  N  N  N  N  FS FS N  N  N  N  HS
## Levels: FS HS LW N
```

```r
# There are 4 levels under the release date type: FS, HS, LW and N
table(d$Release.Date..N...LW...Festive.) # Number of movies under each Release date type
```

```
## 
## FS HS LW  N 
## 17 18 15 99
```

```r
release_table<-table(d$Release.Date..N...LW...Festive.)
round(prop.table(release_table)*100) # Approx. percentage of movies under each release date type
```

```
## 
## FS HS LW  N 
## 11 12 10 66
```

```r
factor(d$Genre...Defined)
```

```
##   [1] Romance  Thriller Comedy   Drama    Comedy   Drama    Comedy   Comedy  
##   [9] Comedy   Drama    Action   Romance  Romance  Action   Comedy   Action  
##  [17] Thriller Comedy   Comedy   Comedy   Thriller Action   Action   Drama   
##  [25] Romance  Drama    Drama    Drama    Thriller Drama    Thriller Thriller
##  [33] Romance  Drama    Drama    Action   Action   Romance  Thriller Comedy  
##  [41] Drama    Action   Romance  Action   Thriller Romance  Comedy   Comedy  
##  [49] Action   Drama    Romance  Thriller Comedy   Thriller Action   Drama   
##  [57] Drama    Comedy   Drama    Comedy   Comedy   Action   Thriller Drama   
##  [65] Romance  Comedy   Romance  Romance  Thriller Drama    Drama    Thriller
##  [73] Comedy   Thriller Drama    Comedy   Drama    Action   Action   Comedy  
##  [81] Romance  Drama    Romance  Romance  Comedy   Comedy   Drama    Comedy  
##  [89] Thriller Drama    Romance  Action   Action   Thriller Thriller Comedy  
##  [97] Comedy   Romance  Thriller Thriller Action   Drama    Drama    Thriller
## [105] Drama    Romance  Romance  Action   Thriller Romance  Romance  Comedy  
## [113] Comedy   Thriller Thriller Comedy   Thriller Thriller Drama    Action  
## [121] Drama    Thriller Romance  Romance  Comedy   Comedy   Comedy   Drama   
## [129] Drama    Comedy   Action   Romance  Comedy   Drama    Drama    Drama   
## [137] Action   Thriller Action   Drama    Thriller Drama    Romance  Action  
## [145] Comedy   Thriller Comedy   Comedy   Action  
## Levels: Action Comedy Drama Romance Thriller
```

```
# There are 5 levels under Genre: Action, Comedy, Drama, Romance, Thriller

table(d$Genre...Defined) # Number of movies under each genre
```

```
##
##   Action   Comedy    Drama  Romance Thriller
##       24       36       35       25       29
```

```
action_p=24/149 # Proportion of movies under action
action_p
```

```
## [1] 0.1610738
```

```
Genre_table<-table(d$Genre...Defined)
round(prop.table(Genre_table)*100) # Approx. percentage of movies under each genre
```

```
##
##   Action   Comedy    Drama  Romance Thriller
##       16       24       23       17       19
```

```
# Comedy has the highest proportion of movies

quantile(d$Budget, seq(from=0, to=1, by=0.2))
```

```
##    0%   20%   40%   60%   80%  100%
##   1.8  10.0  15.0  27.0  40.0 150.0
```

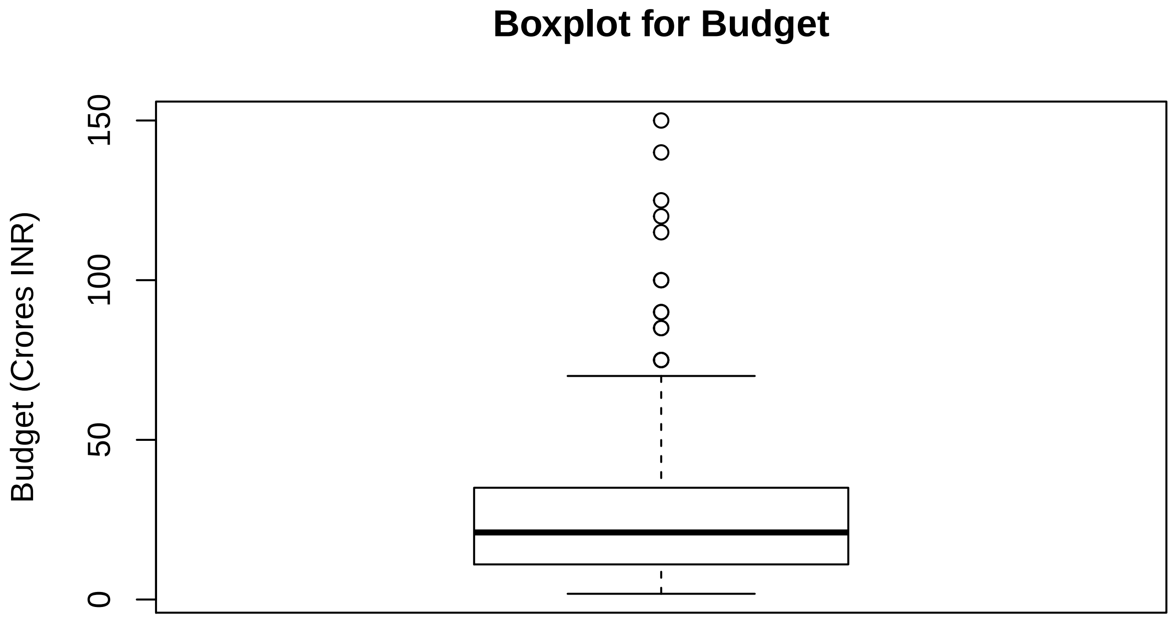
```
# Gives budget values at the 0th,20th,40th,60th,80th and 100th percentiles
var(d$Budget) # variance in the Budget
```

```
## [1] 798.0849
```

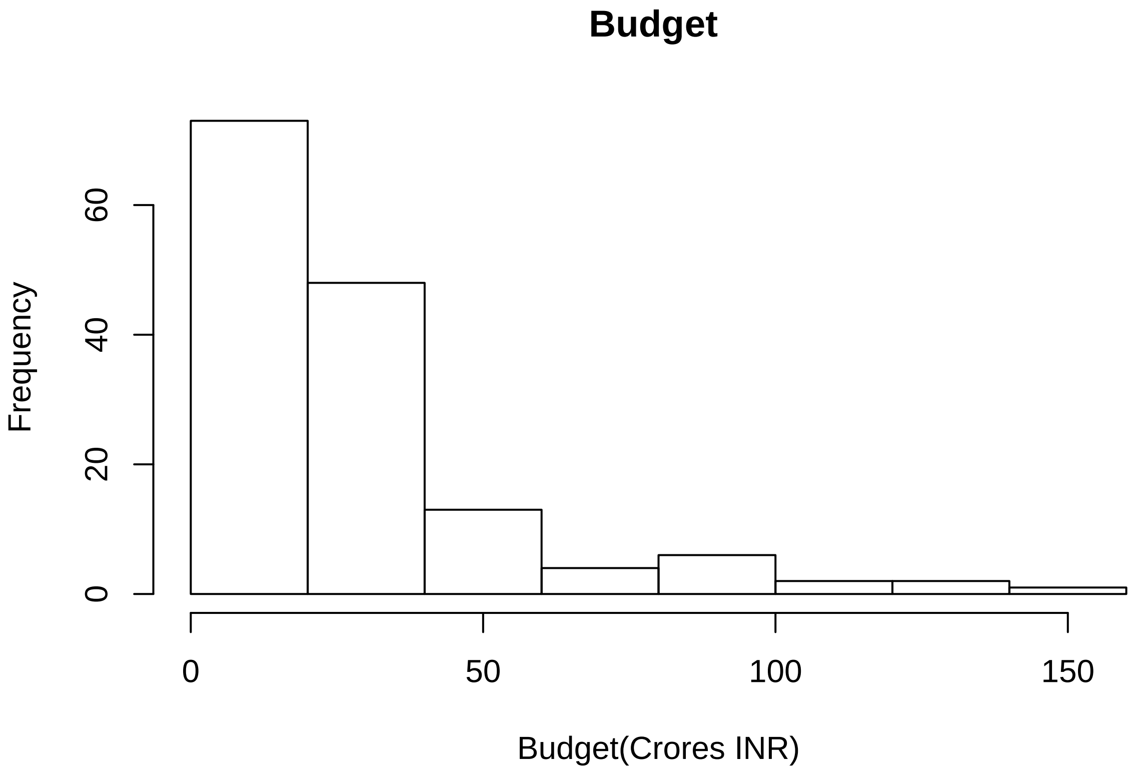
```
sd(d$Budget) # standard deviation in the Budget
```

```
## [1] 28.2504
```

```
boxplot(d$Budget,main="Boxplot for Budget",ylab="Budget (Crores INR)")
```



```
# Many outliers are present. Some movies have exceptionally high budgets at their disposal.
hist(d$Budget,main="Budget ",xlab="Budget(Crores INR)")
```



```
# Skewed when compared to the normal distribution. Most movies spent in the range 0-40 Crores INR.

quantile(d$Box.Office.Collection, seq(from=0, to=1, by=0.2))
```

```
##      0%     20%     40%     60%     80%    100%
##   0.010   5.868  18.560  35.900  66.600 735.000
```

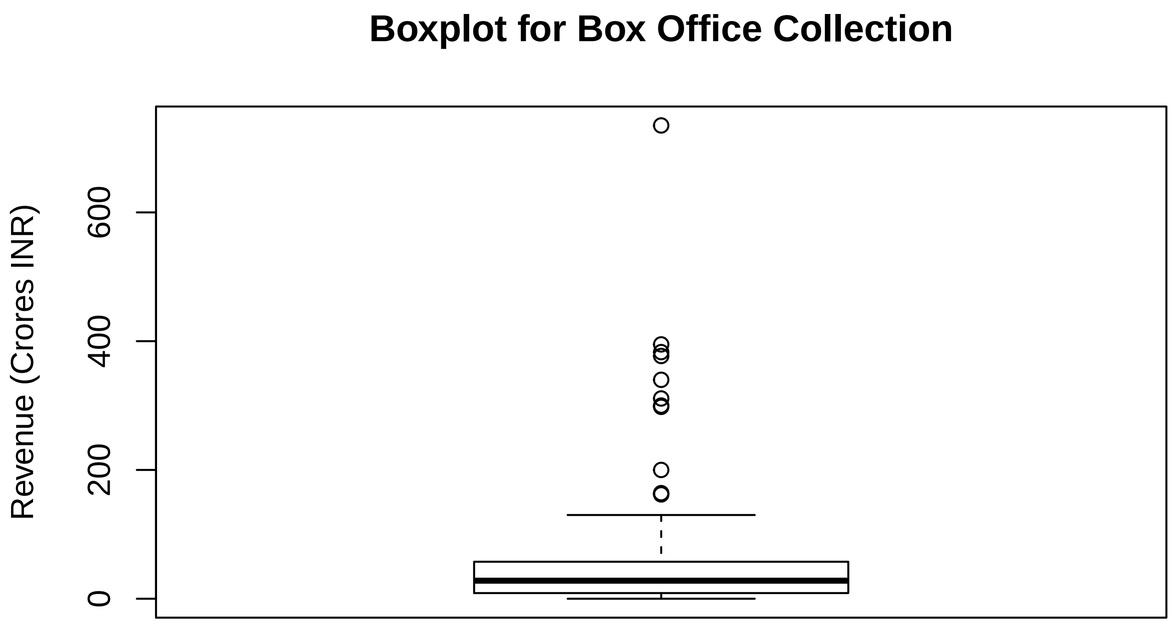
```
# Gives Box office Collection values at the 0th,20th,40th,60th,80th and 100th percentiles
var(d$Box.Office.Collection) # variance in the Box office Collection
```

```
## [1] 8929.216
```

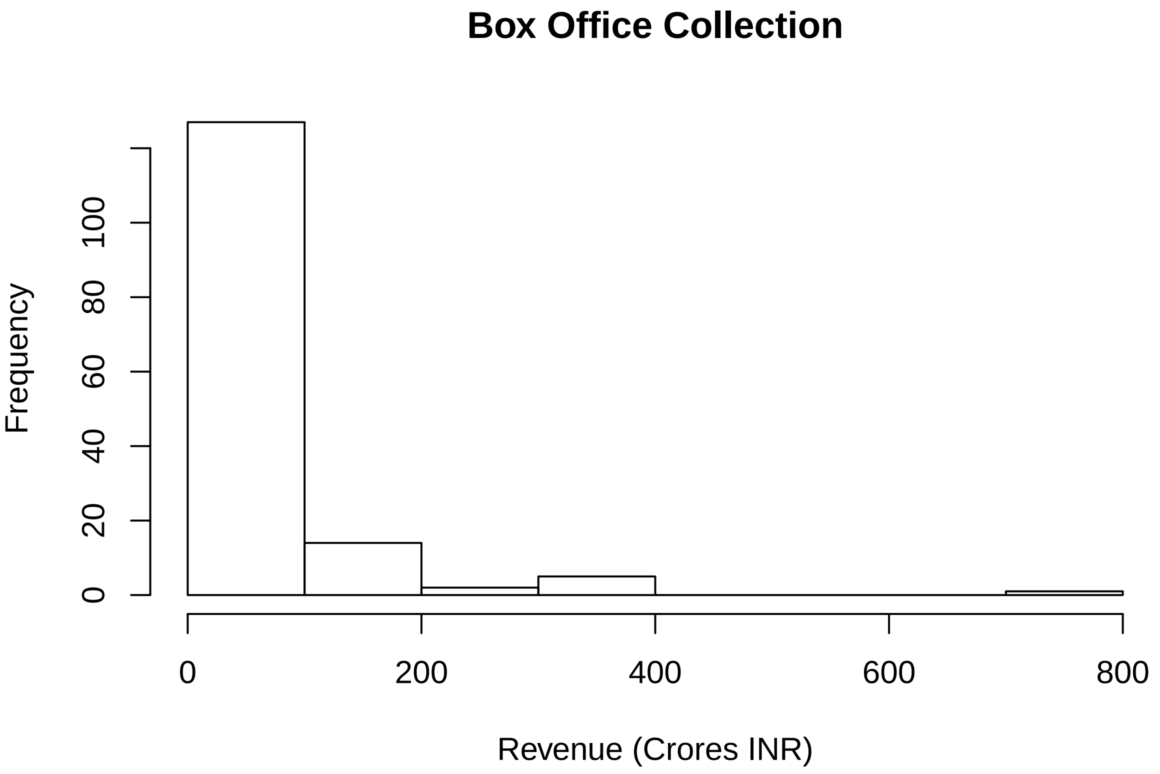
```
sd(d$Box.Office.Collection) # standard deviation in the Box office Collection
```

```
## [1] 94.49453
```

```
boxplot(d$Box.Office.Collection,main="Boxplot for Box Office Collection",ylab="Revenue (Crores INR)")
```

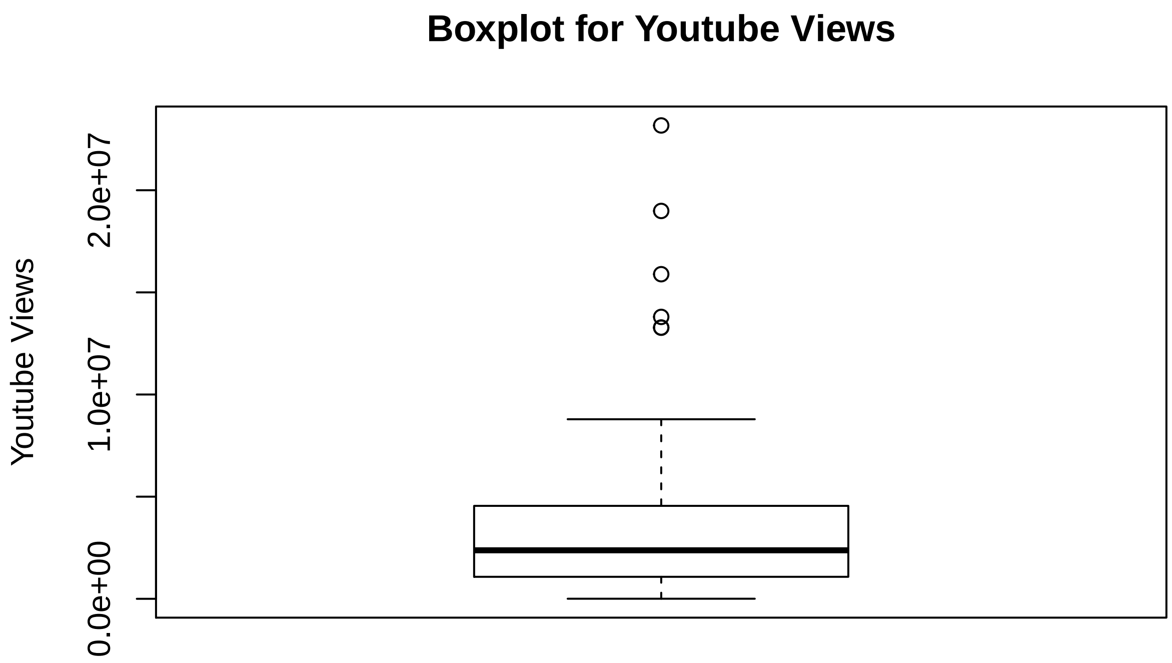


```r
# Boxplot shows that many outliers are present. Some movies have performed exceptionally well.
hist(d$Box.Office.Collection,main="Box Office Collection",xlab="Revenue (Crores INR)")
```

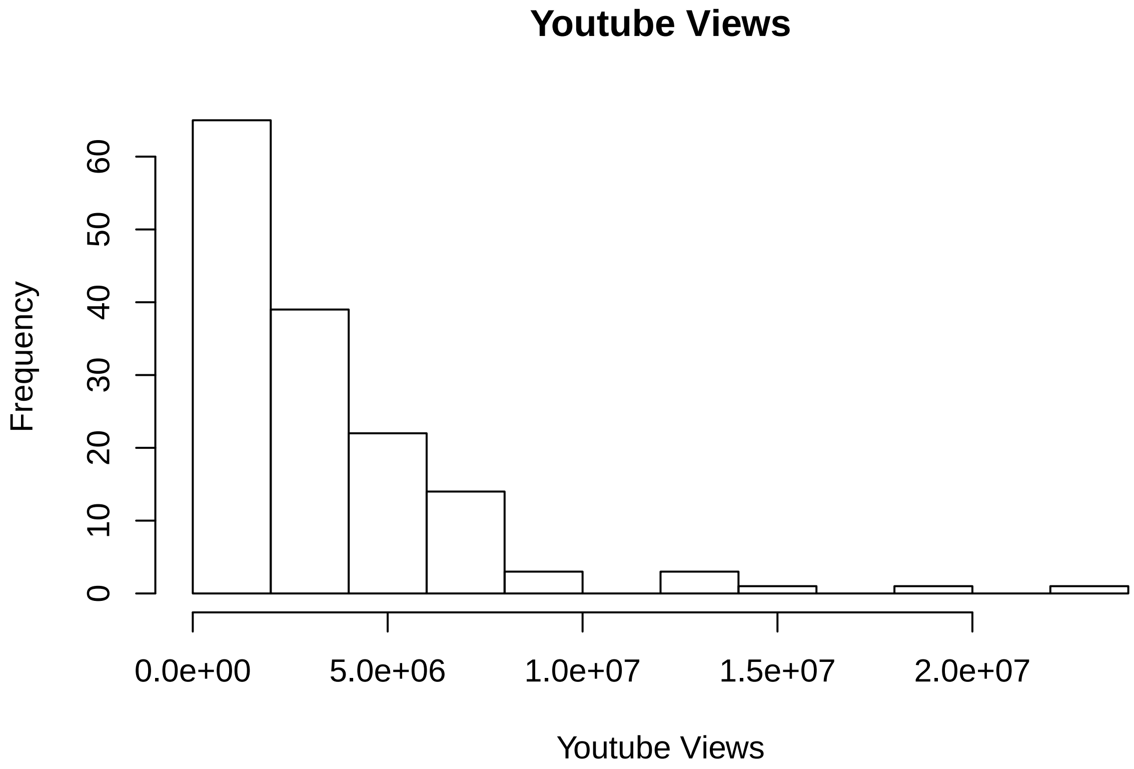


```r
# Histogram shows that most movies have earned in the range 0-100 Crores INR
# Both these plots show that the data is skewed when compared to a normal distribution

boxplot(d$Youtube.Views,main="Boxplot for Youtube Views",ylab="Youtube Views")
```

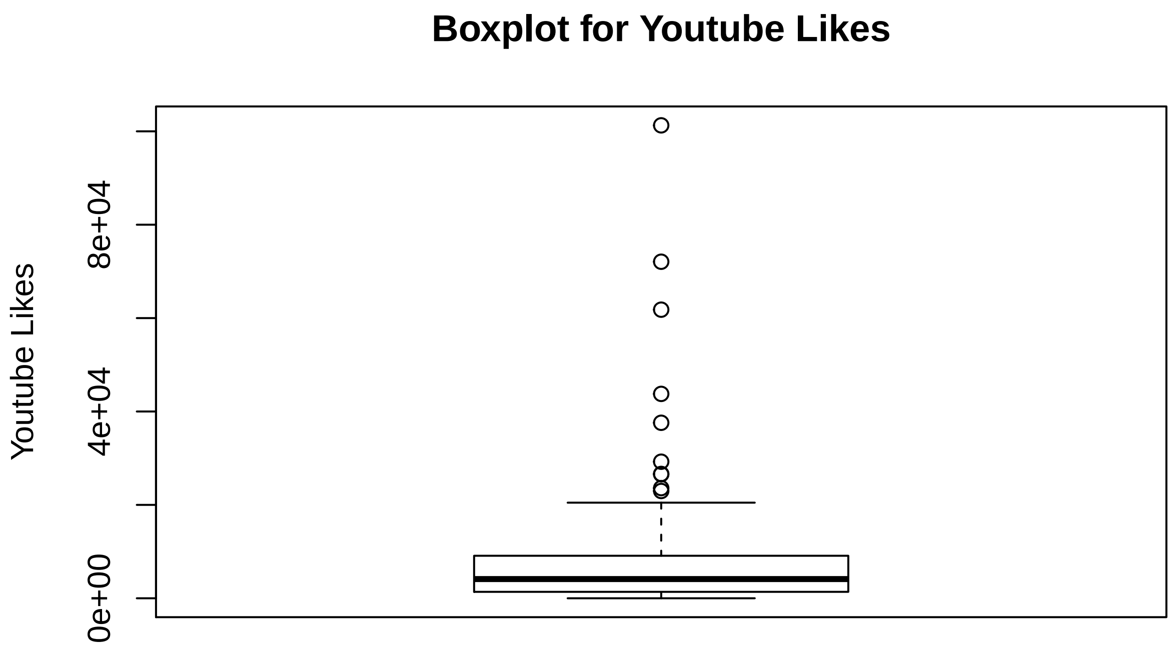


```r
hist(d$Youtube.Views,main="Youtube Views",xlab="Youtube Views")
```

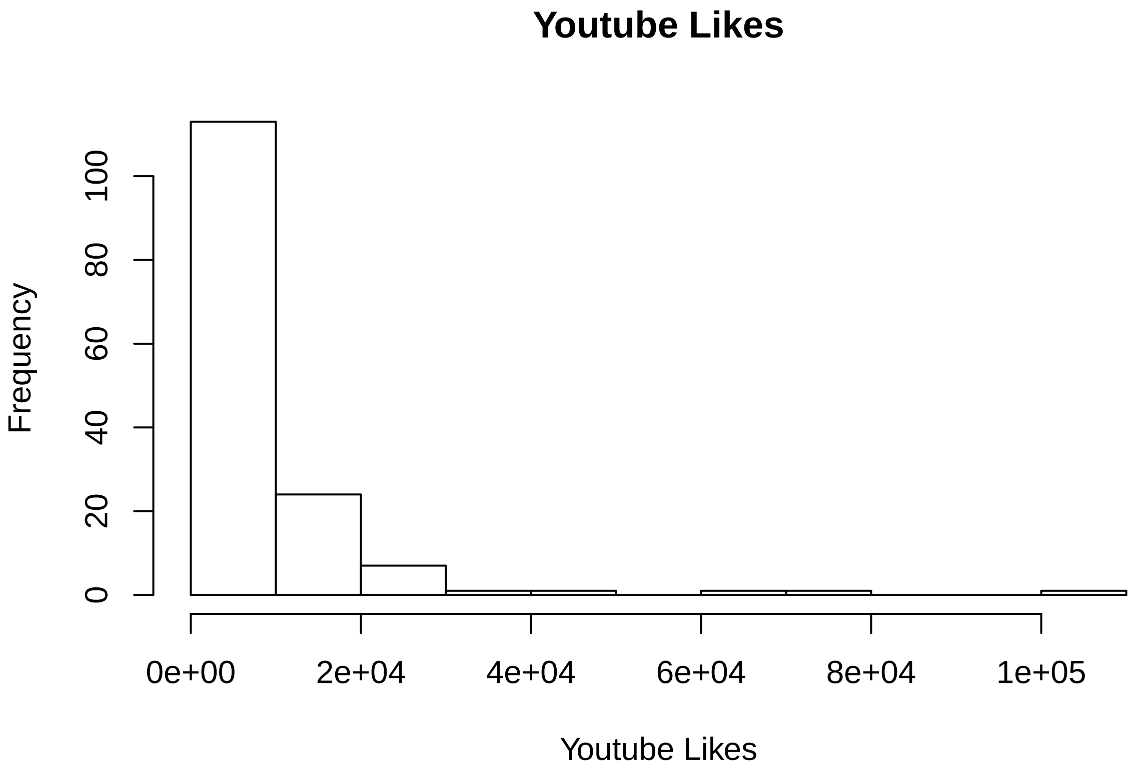


```r
# Few outliers are present, on the upper end. Skewed distribution.

boxplot(d$Youtube.Likes,main="Boxplot for Youtube Likes",ylab="Youtube Likes")
```

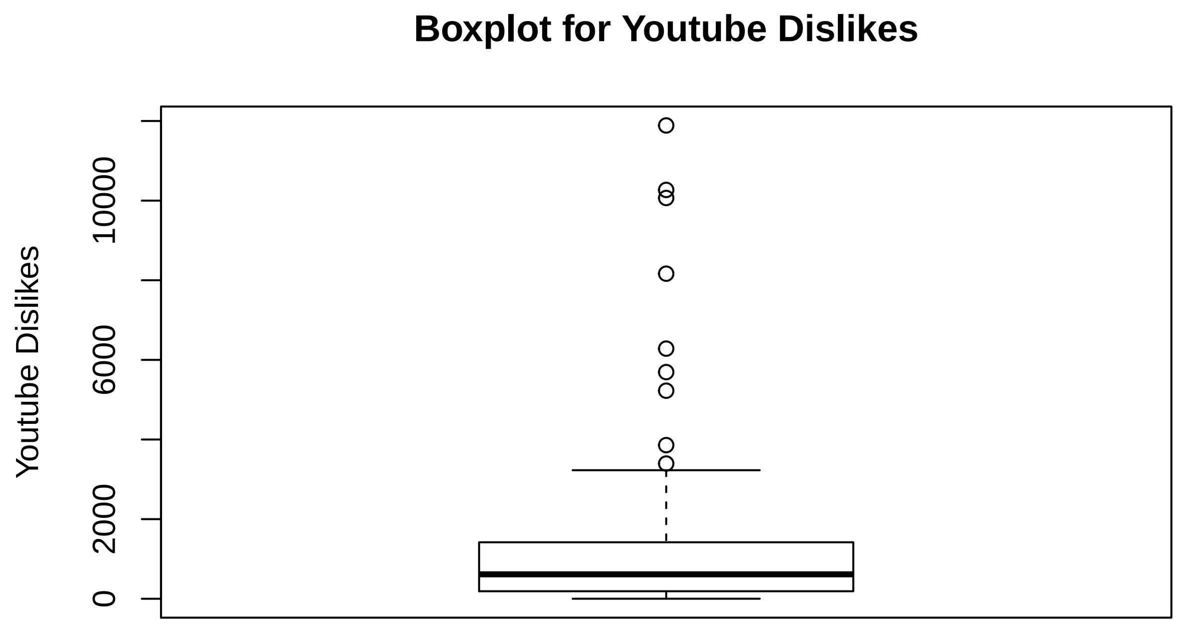


```
# Many outliers are present, on the upper end.
hist(d$Youtube.Likes,main="Youtube Likes",xlab="Youtube Likes")
```

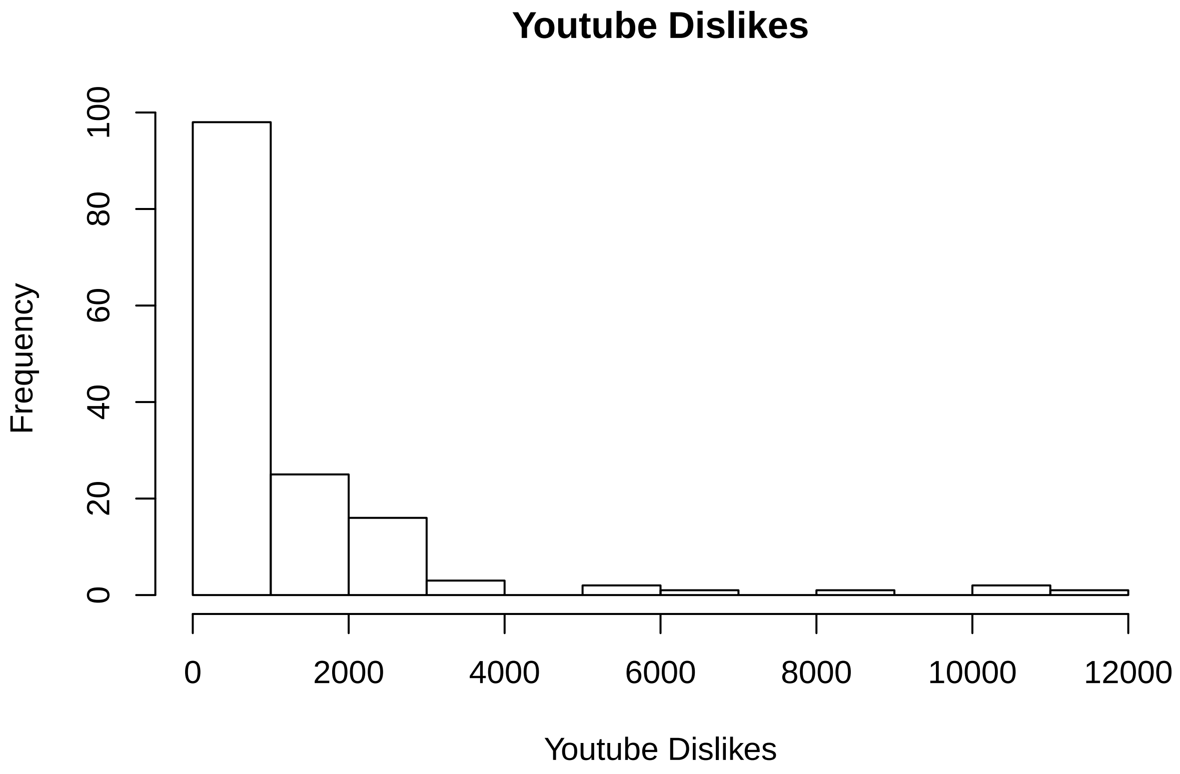


```
# Skewed distribution. A large proportion is in the 0-10K range.

boxplot(d$Youtube.Dislikes,main="Boxplot for Youtube Dislikes",ylab="Youtube Dislikes")
```

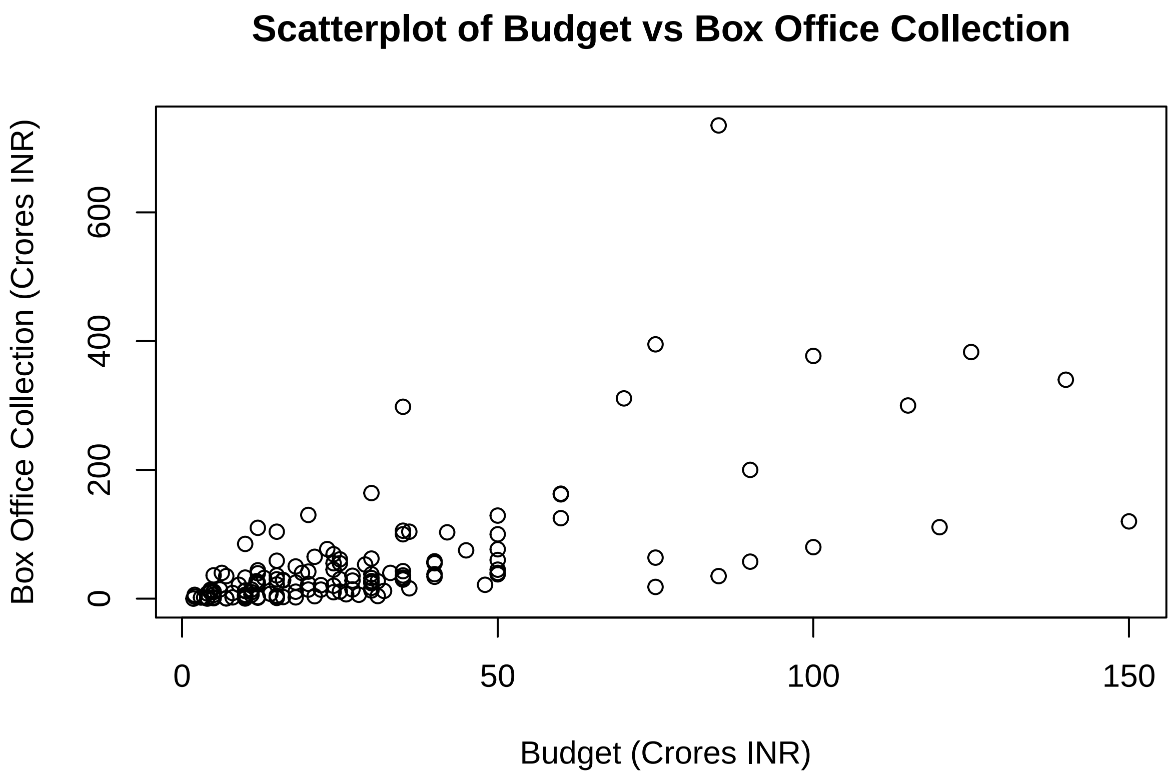


```
# Many outliers are present, on the upper end.
hist(d$Youtube.Dislikes,main="Youtube Dislikes",xlab="Youtube Dislikes")
```



```
# Skewed distribution. A large proportion is in the 0-1K range.

# Relationship between Budget and Box Office Collection
plot(x=d$Budget, y=d$Box.Office.Collection, main="Scatterplot of Budget vs Box Office Collection", xlab=
```

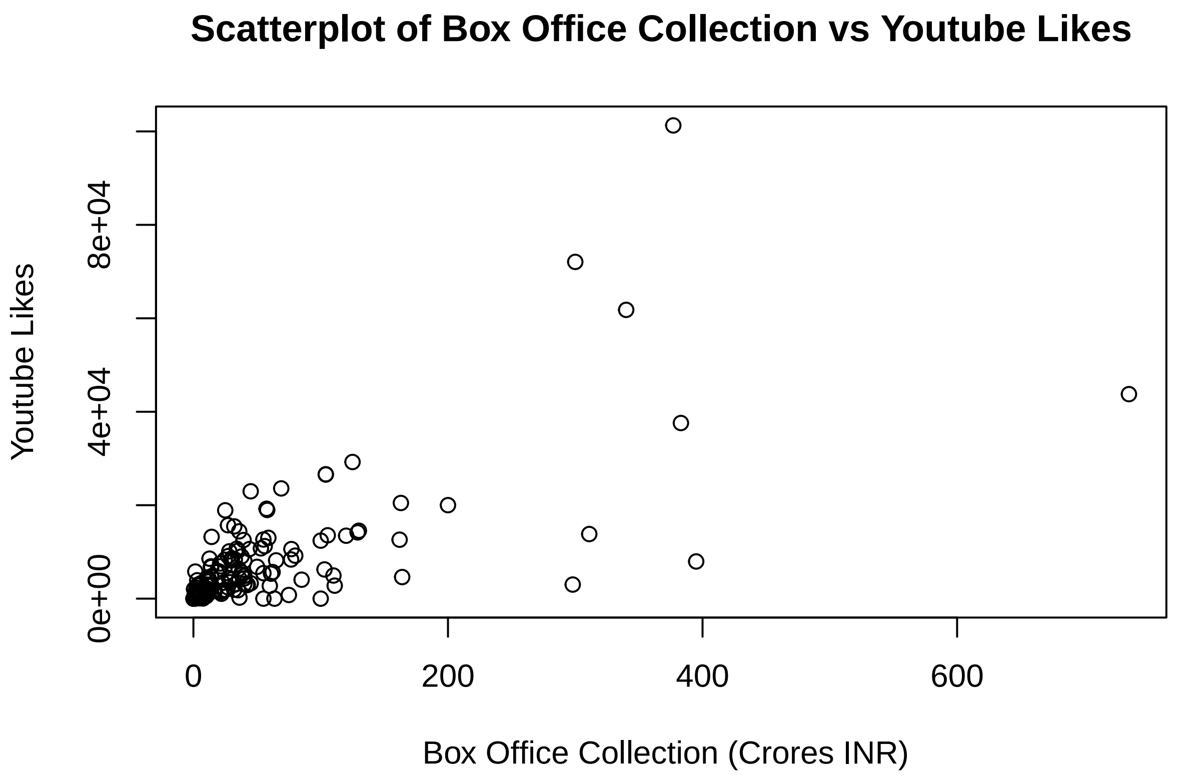


```
cor(d$Budget,d$Box.Office.Collection)
```

```
## [1] 0.6503803
```

```
# a correlation coefficient of 0.65, moderately strong positive correlation

# Relationship between Box Office Collection and Youtube Likes
plot(x=d$Box.Office.Collection, y=d$Youtube.Likes, main="Scatterplot of Box Office Collection vs Youtube
```

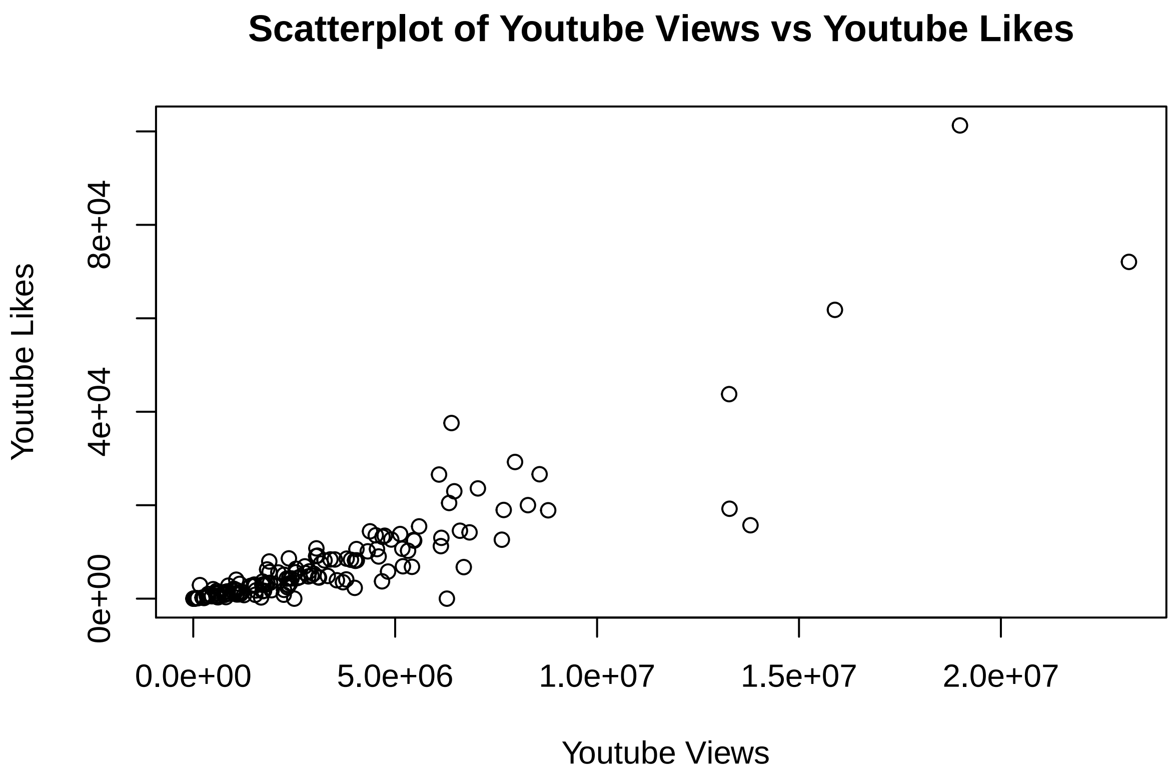


```
cor(d$Box.Office.Collection,d$Youtube.Likes)
```

```
## [1] 0.6825166
```

```
# a correlation coefficient of 0.68, moderately strong positive correlation

# Relationship between Youtube Views and Youtube Likes
plot(x=d$Youtube.Views, y=d$Youtube.Likes, main="Scatterplot of Youtube Views vs Youtube Likes", xlab="
```



Scatterplot of Youtube Views vs Youtube Likes

```
cor(d$Youtube.Views,d$Youtube.Likes)
```

```
## [1] 0.8840548
```

```
# As expected, these two are strongly correlated. rho=0.88

library(gmodels)
genre_popular<-d$Genre...Defined %in% c("Comedy","Drama")
genre_popular # The two most popular genre
```

```
##   [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
##  [13] FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE
##  [25] FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE
##  [37] FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
##  [49] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [61]  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE
##  [73]  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
##  [85]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
##  [97]  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
## [121]  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
## [133]  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
## [145]  TRUE FALSE  TRUE  TRUE FALSE
```

```
table(genre_popular)
```

```
## genre_popular
## FALSE  TRUE
##    78    71
```

```
# A cross table between two categorical variables
CrossTable(x=d$Release.Date..N...LW...Festive., y=genre_popular)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  149
##
##
##                                   | genre_popular
## d$Release.Date..N...LW...Festive. |     FALSE |      TRUE | Row Total |
## ----------------------------------|-----------|-----------|-----------|
##                                FS |        10 |         7 |        17 |
##                                   |     0.136 |     0.150 |           |
##                                   |     0.588 |     0.412 |     0.114 |
##                                   |     0.128 |     0.099 |           |
##                                   |     0.067 |     0.047 |           |
## ----------------------------------|-----------|-----------|-----------|
##                                HS |         7 |        11 |        18 |
##                                   |     0.623 |     0.684 |           |
##                                   |     0.389 |     0.611 |     0.121 |
##                                   |     0.090 |     0.155 |           |
##                                   |     0.047 |     0.074 |           |
## ----------------------------------|-----------|-----------|-----------|
##                                LW |         9 |         6 |        15 |
##                                   |     0.168 |     0.184 |           |
##                                   |     0.600 |     0.400 |     0.101 |
##                                   |     0.115 |     0.085 |           |
##                                   |     0.060 |     0.040 |           |
## ----------------------------------|-----------|-----------|-----------|
##                                 N |        52 |        47 |        99 |
##                                   |     0.001 |     0.001 |           |
##                                   |     0.525 |     0.475 |     0.664 |
##                                   |     0.667 |     0.662 |           |
##                                   |     0.349 |     0.315 |           |
## ----------------------------------|-----------|-----------|-----------|
##                      Column Total |        78 |        71 |       149 |
##                                   |     0.523 |     0.477 |           |
## ----------------------------------|-----------|-----------|-----------|
##
##
```

```
# More popular genres have a release date type HS than other genres. The opposite is true for FS.
```