```
# 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
str(d) # gives data structure
summary(d) # gives minimum, Q1, median, mean, Q3, maximum
# 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
# 10 variables, each with 149 observations

summary(d[c("Budget","Box.Office.Collection","Youtube.Views","Youtube.Likes","Youtube.Dislik
es")]) # gives minimum, Q1, median, mean, Q3, maximum
IQR(d$Box.Office.Collection) # Gives interquartile range Q3-Q1

factor(d$Release.Date..N...LW...Festive.)
# There are 4 levels under the release date type: FS, HS, LW and N
table(d$Release.Date..N...LW...Festive.) # Number of movies under each Release date type
release_table<-table(d$Release.Date..N...LW...Festive.)
round(prop.table(release_table)*100) # Approx. percentage of movies under each release date type

factor(d$Genre...Defined)
# There are 5 levels under Genre: Action, Comedy, Drama, Romance, Thriller

table(d$Genre...Defined) # Number of movies under each genre
action_p=24/149 # Proportion of movies under action
action_p

Genre_table<-table(d$Genre...Defined)
round(prop.table(Genre_table)*100) # Approx. percentage of movies under each genre
# Comedy has the highest proportion of movies

quantile(d$Budget, seq(from=0, to=1, by=0.2))
# Gives budget values at the 0th,20th,40th,60th,80th and 100th percentiles
var(d$Budget) # variance in the Budget
sd(d$Budget) # standard deviation in the Budget

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))
# Gives Box office Collection values at the 0th,20th,40th,60th,80th and 100th percentiles
```

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

boxplot(d$Box.Office.Collection,main="Boxplot for Box Office Collection",ylab="Revenue
(Crores INR)")
# Boxplot shows that many outliers are present. Some movies have performed exceptionally well.
hist(d$Box.Office.Collection,main="Box Office Collection",xlab="Revenue (Crores INR)")
# 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$Youtube.Views,main="Boxplot for Youtube Views",ylab="Youtube Views")
hist(d$Youtube.Views,main="Youtube Views",xlab="Youtube Views")
# 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=" Budget (Crores INR)", ylab="Box Office Collection (Crores INR)")
cor(d$Budget,d$Box.Office.Collection)
# 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 Likes", xlab=" Box Office Collection (Crores INR)", ylab="Youtube Likes")
cor(d$Box.Office.Collection,d$Youtube.Likes)
# 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=" Youtube Views", ylab="Youtube Likes")
cor(d$Youtube.Views,d$Youtube.Likes)
# As expected, these two are strongly correlated. rho=0.88

library(gmodels)
genre_popular<-d$Genre...Defined %in% c("Comedy","Drama")
genre_popular # The two most popular genre
table(genre_popular)
# A cross table between two categorical variables
CrossTable(x=d$Release.Date..N...LW...Festive., y=genre_popular)
# More popular genres have a release date type HS than other genres. The opposite is true for FS.
```