Data Analysis

Dataset:Movie Industry | Kaggle

```r
## Required libraries
library(GGally)

library(dplyr)

library(nnet)
library(gridExtra)
library(ggplot2)
library(caret)

library(lattice)
library(MASS)

library(klaR)
library(zoo)

library(clue)
```

1.Importing Dataset

```
data=read.csv('C:/Users/ANIL/Downloads/movies.csv')
dim(data)
```

```
## [1] 6820   15
```

```
summary(data)
```

```
data_type=sapply(data, class)  #to get data type of each column
fact_data=data[data_type=='factor']  #to get only factors from all colu
mns
num_data=data[data_type!='factor']  #to get only numeric  from all colu
mns
```

2. Data Preprocessing &Exploratory Analysis:

```
#Data-preprocessing
#1)Checking missing values
colnames(num_data)
```

```
## [1] "budget"  "gross"   "runtime" "score"   "votes"   "year"
```

```
any(num_data$budget==0)  #check whether budget columns are zero or not
```

```
## [1] TRUE
```

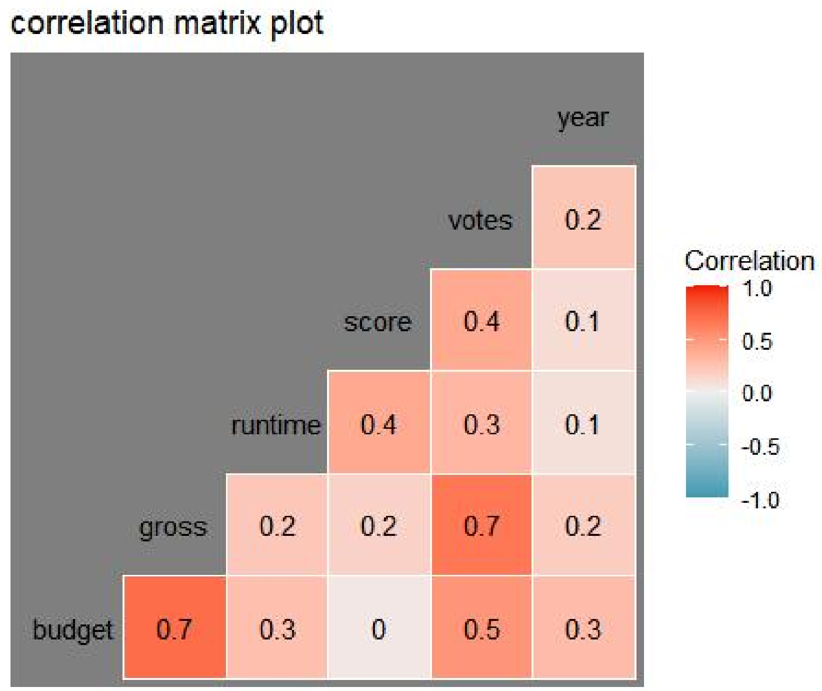
```
#Building Correlation plot
ggcorr(num_data, name = "Correlation", label = TRUE, alpha = TRUE, pale
tte = "PuOr") +
  ggtitle("correlation matrix plot") + theme_dark()
```

Interpretation:

1) In above see that the ‘budget’ column contains null values.Hence first we have Impute those null values

2) To Impute null values we first plot correlation plot and by using we have to find out which variable is most significant with respect to ‘budget’ column

3) Now we see that below correlation between ‘budget’ and ‘gross’ is 0.7.Hence we use ‘gross’ feature to impute missing values of ‘budget column.



I]. Impute missing values

Hence to impute missing values we fit linear regression model.Here we consider ***budget be response and gross be the regressor***

1)separate out data

```
data_nonzero=subset(num_data, budget!= 0)
data_zero=subset(num_data, budget==0)
dim(num_data);dim(data_nonzero);dim(data_zero)
```

```
## [1] 6820    6

## [1] 4638    6

## [1] 2182    6

#we see that 2182 values of budget columns are missing
```

2) Fitting model:

```
model=lm(data_nonzero$budget~data_nonzero$gross,data=data_nonzero)
df1=data.frame(model$fitted.values,data_nonzero$budget)
colnames(df1)=c('Fitted_values','Actual_values')
```

3) To check Accuracy

```
d = data_nonzero$budget-model$fitted.values
d=scale(data_nonzero$budget)-scale(model$fitted.values)  #scale data
mse = mean((d)^2)
mae = mean(abs(d))
```

```
rmse = sqrt(mse)
mse;mae;rmse

## [1] 0.6397969

## [1] 0.5025675

## [1] 0.7998731

Comment:-here we see that mean squared error(mse),mean absolute error
(mae),   root mean squared error(rmse) are moderate,hence we use linear
 regression to impute missing values
```

4)Prediction of Missing values

```
prdf=as.data.frame(data_zero$gross)
miss_budgt_pred=predict(model,newdata=prdf)

fill_miss_data=num_data
for (i in 1:nrow(data_zero)) {
  if(data_zero$budget[i]==0)
  {
    fill_miss_data$budget[i]=miss_budgt_pred[i]
  }
}

Total_data=cbind(fill_miss_data,fact_data,by = 0)[,-16]
```

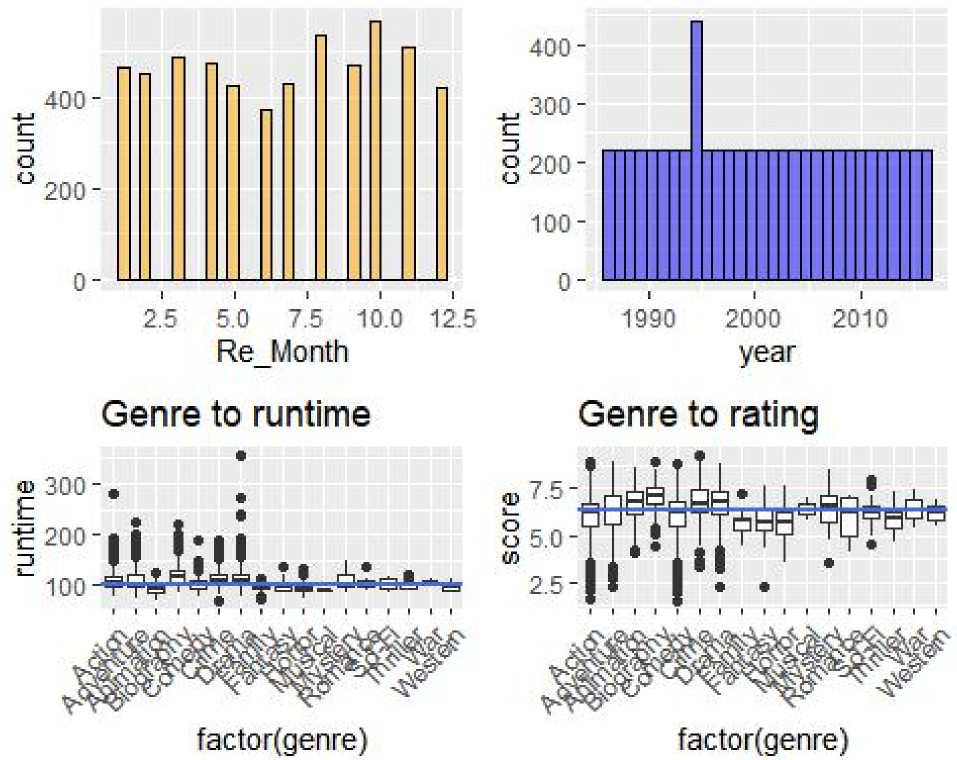
5)Checking Missing values of Rating column

```
unique(Total_data$rating)

##  [1] R             PG-13         PG            UNRATED       Not spe
cified
##  [6] G             NC-17         NOT RATED     TV-PG         TV-MA

## [11] B             B15           TV-14
## 13 Levels: B B15 G NC-17 NOT RATED Not specified PG PG-13 R TV-14 ...
 UNRATED

Not_specifi_rating=subset(Total_data,rating=='Not specified')
specifi_rating=subset(Total_data,rating!='Not specified')
dim(Not_specifi_rating);dim(specifi_rating)

## [1] 63 15

## [1] 6757   15

Comment:-1)Here we see that some movies rating are Not specified and
these are only 63 values are missing hence we remove those observation
from data
```

```
6)Get Final Dataset
Final_data=subset(Total_data,rating!='Not specified')
Final_data=subset(Final_data,budget!=0)
Final_data['Re_Month']=as.numeric(format(as.Date(Final_data$released),
"%m"))
Final_data=subset(Final_data,select = -c(name,released))
dim(Final_data)

## [1] 5684   14
```

Exploratory Analysis:

```
1) Box plot model for Genere wise movie runtime in minutes
p_genrerun = ggplot(Final_data, aes(x=factor(genre), y=runtime)) +
  geom_boxplot() + theme(axis.text.x = element_text(angle = 45, hjust =
 1)) +
  ggtitle("Genre to runtime") +
  geom_hline(yintercept =median(Final_data$runtime,na.rm = TRUE), col =
 "royalblue",lwd = 1)

2) Box plot model for Genere wise score
p_genrerating = ggplot(Final_data, aes(x=factor(genre), y=score)) +
  geom_boxplot() + theme(axis.text.x = element_text(angle = 45, hjust =
 1))+
  ggtitle("Genre to rating") +
  geom_hline(yintercept=median(Final_data$score, na.rm = TRUE), col = "
royalblue",lwd = 1)

3)Histogram along with Scatterplot showing theater release month and ye
ar data
g1=ggplot(data = na.omit(Final_data), aes(x = Re_Month)) + geom_histogr
am(colour = "black", fill =
                                                                     "o
range", alpha = 0.5)
g2=ggplot(data = na.omit(Total_data), aes(x = year)) + geom_histogram(c
olour = "black", fill =
                                                                      "
blue", alpha = 0.5)
grid.arrange(g1, g2,p_genrerun,p_genrerating,nrow = 2, ncol = 2)
```



Interpretation:-

1) In the First plot of Released month vs count we see that highest number of Movies are Released in the 10th and 9th month and lowest is in 6th month

2) In second plot of year vs count we see that most of data are uniformly distributed expect year 1995.Hence we say average number of movies are Released in each year is approximately 220

3) In third box plot runtime of ‘Drama’ genre are high among the all and median runtime is approximately 100 minutes.There is lot of outliers in the ‘Drama’

4) In fourth plot we see that median rating is 6.5 and lot of outliers is in data mostly in ‘comedy’ and ‘Action’ Genre.

3.Problem statements:

Q.1

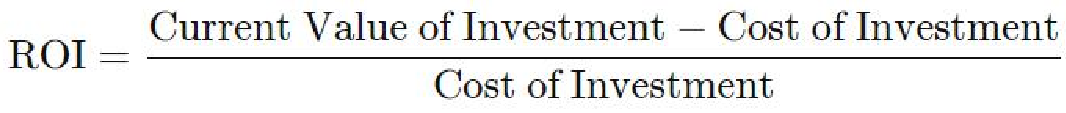
If you want to produce a movie to get high return on investment (ROI), what would be your recipe for success? (You may want to first state your definition of ROI with justification)

What is (Return on Investment)ROI?

Return on Investment (ROI) is a performance measure used to evaluate the efficiency of an investment or compare the efficiency of a number of different investments. ROI tries to directly measure the amount of return on a particular investment, relative to the investment's cost. To calculate ROI, the benefit (or return) of an investment is divided by the cost of the investment. The result is expressed as a percentage or a ratio.

How to Calculate ROI

The return on investment formula is as follows:

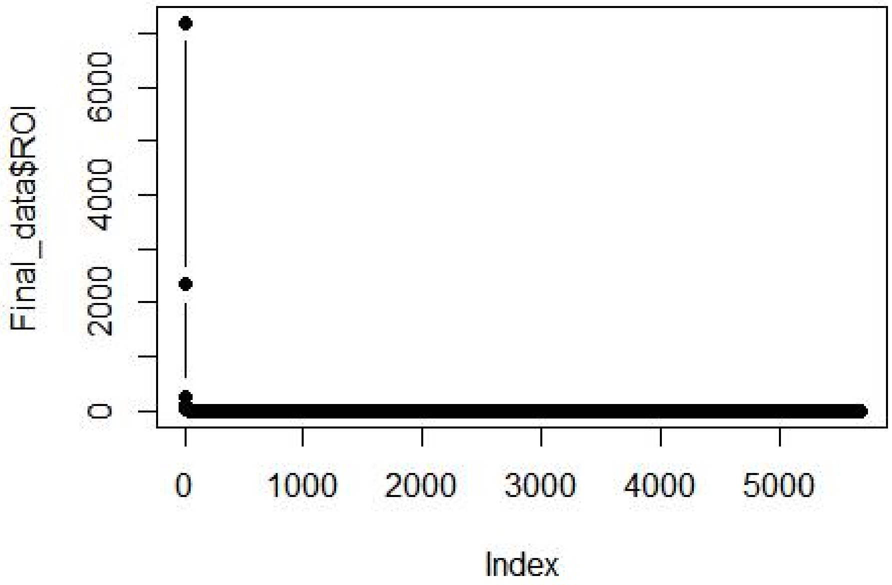


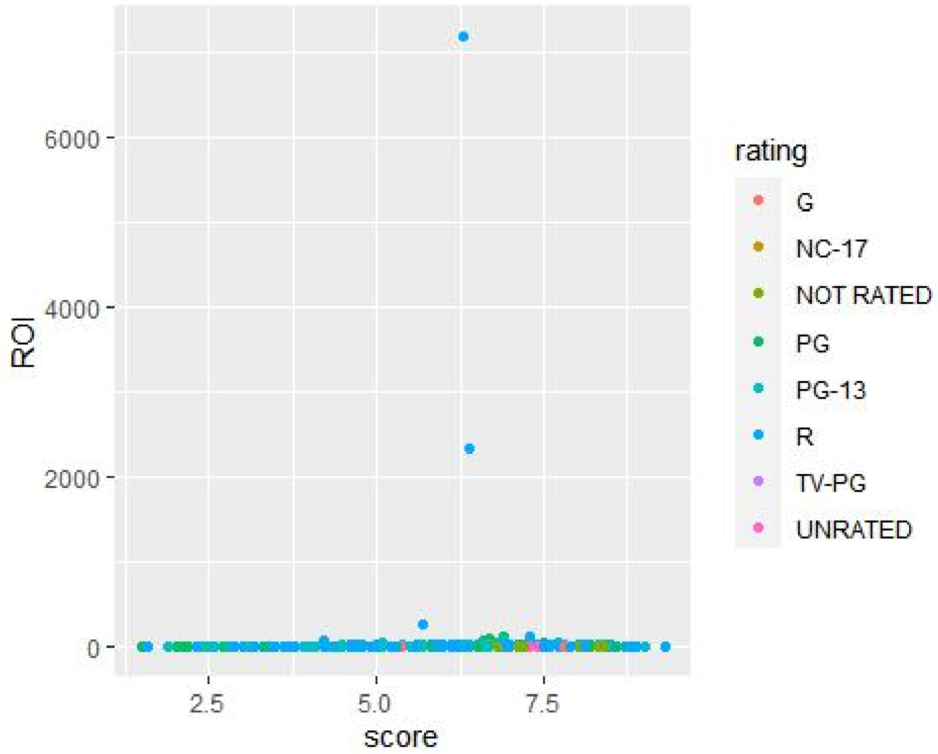
$$\text{ROI} = \frac{\text{Current Value of Investment} - \text{Cost of Investment}}{\text{Cost of Investment}}$$

Here Now our problem we consider as follow:

Current Value of Investment=gross
Cost of Investment=budget

```
#ROI=(gross-budget)/budget
ROI=(Final_data$gross-Final_data$budget)/Final_data$budget
Final_data['ROI']=ROI
Final_data=arrange(Final_data,desc(ROI))
boxplot(Final_data$ROI,pch=19)

scatter.smooth(Final_data$ROI,type='b',pch=19)
Arr_data=arrange(Final_data,desc(ROI))
ggplot(data=Arr_data,aes(y=ROI,x=score,colour=rating))+
  geom_point()

Comment:-Here we see that first three points are two large i.e. outlier hence we remove it and In second plot we cant say anything about Rating and ROI
```


Final_data$ROI
0
2000
4000
6000
0
1000
2000
3000
4000
5000
Index


ROI
0
2000
4000
6000
2.5
5.0
7.5
score
rating
G
NC-17
NOT RATED
PG
PG-13
R
TV-PG
UNRATED

1)Toward the problem statement

A)For categorical variables selection,which are related to ROI

```
s=c(15,8,9,10,11,12,13,14)
Cat_Data=Arr_data[,s]
summary(Cat_Data)
```

```
y=Cat_Data$ROI
x=Cat_Data$star
v=c(unique(as.character(x))
SMean=c()
for(i in 1:length(v)){
  ystar=y[which(Cat_Data$star==v[i])]
  SMean[i]=mean(ystar)
}
Choice_Of_Star=v[which(SMean==max(SMean))];Choice_Of_Star
```

```
## Katie   Featherston
```

```
D=Cat_Data$director
v=c(unique(as.character(D)))
SMean=c()
for(i in 1:length(v)){
  ystar=y[which(Cat_Data$director==v[i])]
  SMean[i]=mean(ystar)
}
Choice_Of_Director=v[which(SMean==max(SMean))];Choice_Of_Director
```

```
## Oren Peli
```

```
W=Cat_Data$writer
v=c(unique(as.character(W)))
SMean=c()
for(i in 1:length(v)){
  ystar=y[which(Cat_Data$writer==v[i])]
  SMean[i]=mean(ystar)
}
Choice_Of_Writer=v[which(SMean==max(SMean))];Choice_Of_Writer
```

```
## Oren Peli
```

```
c=Cat_Data$country
v=c(unique(as.character(c)))
SMean=c()
for(i in 1:length(v)){
  ystar=y[which(Cat_Data$country==v[i])]
  SMean[i]=mean(ystar)
}
Choice_Of_Contry=v[which(SMean==max(SMean))]

g=Cat_Data$genre
v=c(unique(as.character(g)))
```

```
SMean=c()
for(i in 1:length(v)){
  ystar=y[which(Cat_Data$genre==v[i])]
  SMean[i]=mean(ystar)
}
Choice_Of_Genre=v[which(SMean==max(SMean))];Choice_Of_Genre
```

```
##  Horror
```

```
r=Cat_Data$rating
v=c(unique(as.character(r)))
SMean=c()
for(i in 1:length(v)){
  ystar=y[which(Cat_Data$rating==v[i])]
  SMean[i]=mean(ystar)
}
Choice_Of_Rating=v[which(SMean==max(SMean))];Choice_Of_Rating
```

```
## R
```

```
re=Cat_Data$Re_Month
v=c(unique(as.character(re)))
length(which(is.na(re))) # only 57 observation is missing which we can neglect it.
```

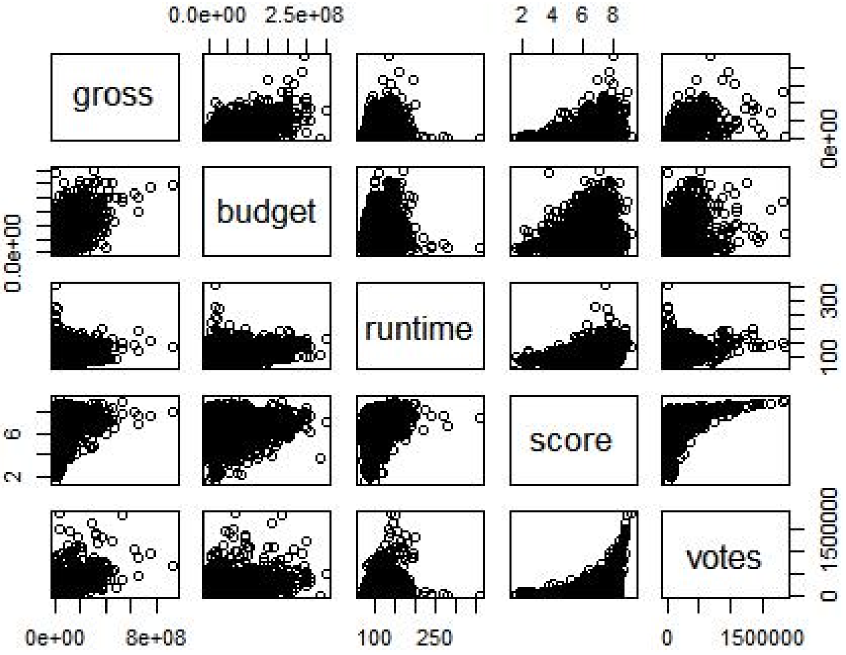
```
## [1] 57
```

```
v=c(na.omit(v))
SMean=c()
for(i in 1:length(v)){
  ystar=y[which(Cat_Data$Re_Month==v[i])]
  SMean[i]=mean(ystar)
}
Choice_Of_Re_month=v[which(SMean==max(SMean))];Choice_Of_Re_month
```

```
## [1] 10
```

```
Variables=c("Choice_Of_Star","Choice_Of_Director","Choice_Of_Writer","Choice_Of_Genre","Choice_Of_Re_month","Choice_Of_Rating")
Choice=c(Choice_Of_Star,Choice_Of_Director,Choice_Of_Writer,Choice_Of_Genre,Choice_Of_Re_month,Choice_Of_Rating)
Cat_summary=data.frame(Variables,Choice);Cat_summary

aa=Cat_summary$Choice
Hroi=c(as.character(Arr_data$star[aa[1]]),as.character(Arr_data$director[aa[2]]),as.character(Arr_data$writer[aa[3]]),as.character(Arr_data$genre[aa[4]]),as.character(Arr_data$Re_Month[aa[5]]),as.character(Arr_data$rating[aa[6]]))
Cat_summary['info']=Hroi
Cat_summary
```

```
##            Variables Choice             info
## 1      Choice_Of_Star     10          Katie   Featherston
## 2 Choice_Of_Director     10          Oren Peli
## 3    Choice_Of_Writer     10          Oren Peli
## 4     Choice_Of_Genre     10          Horror
## 5 Choice_Of_Re_month     10            10
## 6    Choice_Of_Rating     10             R

Comment:1)In above table we all have the information about the choice
of the features
2)Hence to produce movie to get high return we have to use above inform
ation,that is we may cast the ‘Katie Featherston’ as lead star,Director
as ‘Oren Peli’,writer as ‘Oren Peli’ and we release movie in the month
of ‘october’

B)For contineous variables selection,which are related to ROI
s=c(2,1,3,4,5)
K=Arr_data[,s]
ContinuousData=Arr_data[,s]
plot(ContinuousData)
```



```
cor(ContinuousData)

##              gross     budget   runtime      score     votes
## gross   1.0000000 0.61517365 0.2550457 0.21856130 0.6601184
## budget  0.6151737 1.00000000 0.2584234 0.06485805 0.4032798
## runtime 0.2550457 0.25842337 1.0000000 0.40853683 0.3498912
## score   0.2185613 0.06485805 0.4085368 1.00000000 0.4369699
## votes   0.6601184 0.40327982 0.3498912 0.43696992 1.0000000
```

Comment:

1)Here we see that correlation between ‘gross’ and ‘votes’ is 0.66.hence we predict ‘gross’ from ‘votes’.we fit linear regression model.

```
1)Model Fitting
model=lm(ContinuousData$gross~ContinuousData$votes+ContinuousData$budget,ContinuousData)
Cont_S=summary(model)
```

```
Catagorical_Data_summary=Cat_summary
```

```
Continuous_Data_summary=Cont_S;Continuous_Data_summary
```

```
##
## Call:
## lm(formula = ContinuousData$gross ~ ContinuousData$votes + ContinuousData$budget,
##     data = ContinuousData)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -420635897  -15358126   -4769664    9973371  620216491
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -4.888e+06  7.694e+05  -6.354 2.27e-10 ***
## ContinuousData$votes   2.182e+02  4.161e+00  52.441  < 2e-16 ***
## ContinuousData$budget  6.996e-01  1.575e-02  44.414  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40060000 on 5681 degrees of freedom
## Multiple R-squared:  0.5812, Adjusted R-squared:  0.581
## F-statistic:  3942 on 2 and 5681 DF,  p-value: < 2.2e-1
```

Interpretation:

1) Hence from case A we say that to produce movie to get high return we have to we use above information,that is we may cast the ‘Katie Featherston’ *as lead star,Directoras as* ‘Oren Peli’*,writer as* ‘Oren Peli’ *and we release movie in the month of ‘october’.If movie has Rating is ‘R’ then we say that movie is hit.*

```
2) In case B we fit linear model.we see that p-value is less then 0.05
Hence we say that Model is significant that is ‘Gross’ is well explaine
d by two features namely votes and budget.In above summary from coeffic
ient of ‘budget’ we say that if we have 100 budget to make movie then w
e get approximately 66% increase in ‘gross’ collection.

3) From all these we also say to get Good ROI we have to increase movie
 ‘budget’ also.
```

Q.2

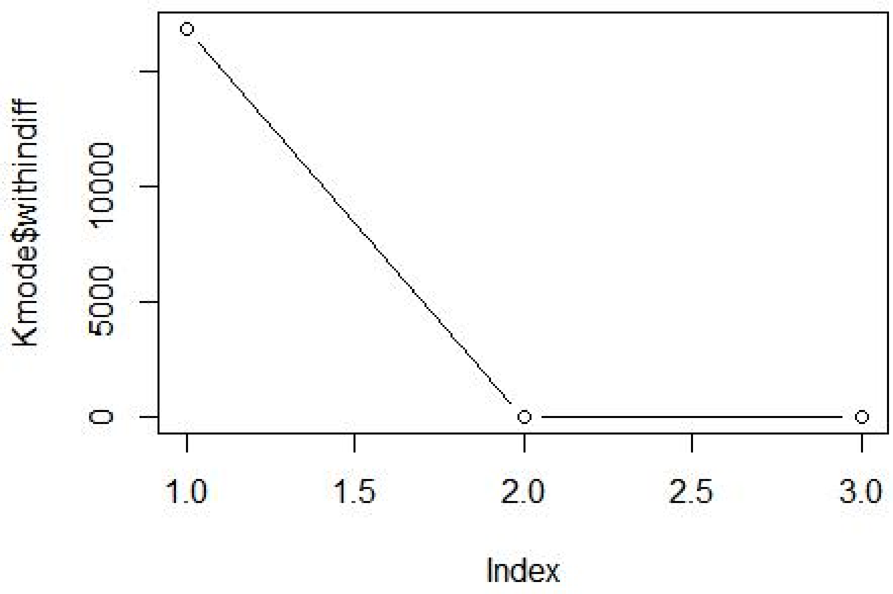
If suppose the actors you want to cast or the directors or the writers you want to hire,are not ready / available to work with you, how would you think of replacement actors/ directors / writers?

1)Towards the problem statement

A)Here First we use the k-mode clustering to find out problem

a)*in case of actor*

```
set.seed(12)
rep_data=subset(Final_data,select= c(star,director,writer))
Kmode=kmodes(rep_data,modes=3,iter.max = 10, weighted = FALSE, fast = T
RUE)
plot(Kmode$withindiff,type='b')
```



```
Kmode

## K-modes clustering with 3 clusters of sizes 5657, 11, 16
##
## Cluster modes:
##                 star       director       writer
## 1       Nicolas Cage    Woody Allen  Woody Allen
## 2 Patricia Arquette   Roland Joffé  Alex Lasker
## 3 Reese Witherspoon Matthew Bright Amanda Brown
##
## Within cluster simple-matching distance by cluster:
## [1] 16870    20    29
```

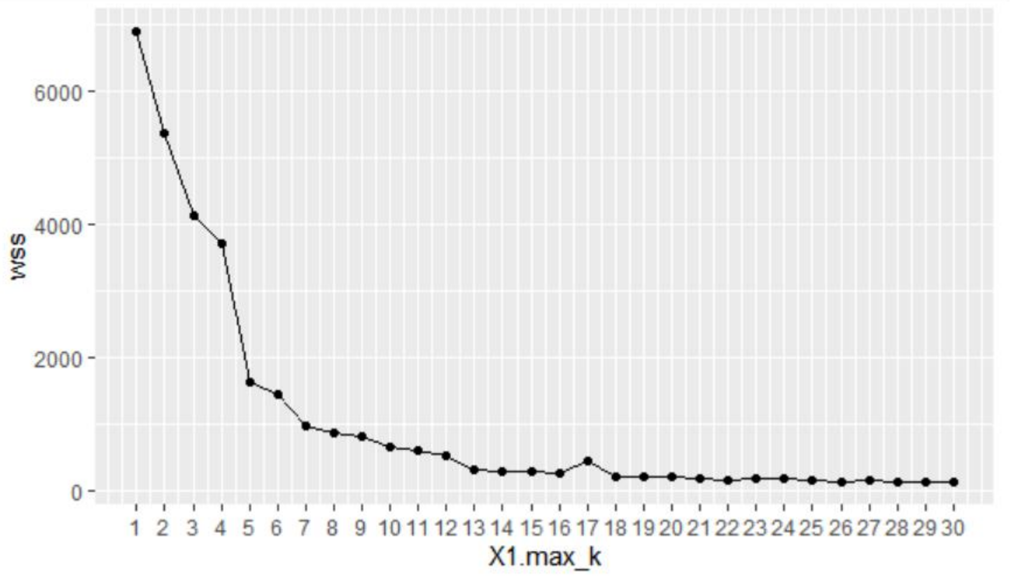
```
##
## Available components:
## [1] "cluster"    "size"       "modes"      "withindiff" "iterations"
## [6] "weighted"
```

```
Comment:1)In above we see that above data is cluster into three part.fr
om elbow method we compute number of optimal cluster that are 3
2)but we see that approimately 95% data is belongs to first cluster and
 remaining two cluster are vary less percentage of observations.Hence w
e not use these method to figure out our problem.
```

B) Here we use K-means clustering

1) In below we first consider ‘aggfunc’ which will gives the data in sorted .I.e it will gives groupy data.
2)Then ‘kmean_withinss’ function is used to compute within cluster sum of square to find out optimal number of cluster.
3)’Rep_direct’ is used to replace required feature
4)’elbow_method’ is used to compute optimal number of cluster
5)Here we fit K-means clustering algoritham to each of feature(director,star,writer) with respect to the ‘votes’,’score’,and ‘ROI’
6)And from each of these we separately compute required replace of feature

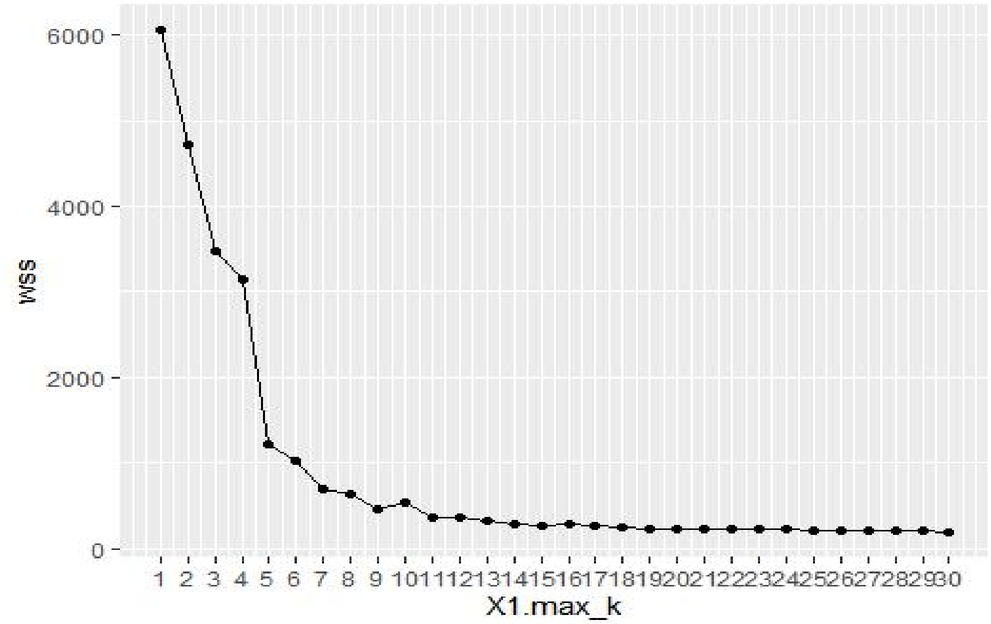
```
aggfunc=function(coln='director-star-writer'){
  a=aggregate(Final_data$ROI, by=list(Final_data[,coln]), FUN=mean)
  b=aggregate(Final_data$votes, by=list(Final_data[,coln]), FUN=sum)
  c=aggregate(Final_data$score, by=list(Final_data[,coln]), FUN=mean)
  Group_m=data.frame(b,a[,2],c[,2])
  colnames(Group_m)=c(coln,'votes','ROI','score')
  return(Group_m)
}

#
kmean_withinss = function(k,arrd) {
  cluster = kmeans(scale(arrd), k,iter.max = 50)  #remove first two row
  return (cluster$tot.withinss)
}


#Replace for perticular
Rep_direct=function(da='data',anyrep='director-star-writer name'){
  if(anyrep %in% da[,1])
  {
    temp1=which(da[,1]==anyrep)
    df=subset(da,Cluster=da[temp1,]$Cluster)
    temp=''
    maxi=sort(df$ROI,decreasing = TRUE)
    temp=subset(df,ROI==maxi[1])[,1]
    if(temp==anyrep){
      temp=subset(df,ROI==maxi[2])[,1]
    }else{
      temp=temp
    }
  }else{
    temp='Given Name Not exist'
  }
  return(temp)
}

#
elbow_method=function(arrd){
  wss=c(0)
  max_k=30
  for (i in 1:max_k) {
    wss[i]=kmean_withinss(i,arrd)
  }
  elbow=data.frame(1:max_k, wss)
  library(ggplot2)
  ggplot(elbow, aes(x = X1.max_k, y = wss)) +
    geom_point() +
    geom_line() +
    scale_x_continuous(breaks = seq(1, max_k, by = 1))
}

#a)#replace director
Group_m=aggfunc('director')
arrd=Group_m[,-1]
elbow_method(arrd)
```



```r
Kmean=kmeans(scale(arrd),9,iter.max = 20)
#data with respect to cluster
Newd=arrange(Group_m,desc(ROI))
Newd['Cluster']=Kmean$cluster
#prediction
Rep_direct(Newd,'Doug Liman') #replace that director
```

```
## [1] Oren Peli
## 2759 Levels: A.R. Murugadoss Aamir Khan Aaron Blaise ... Zoya Akhtar
```

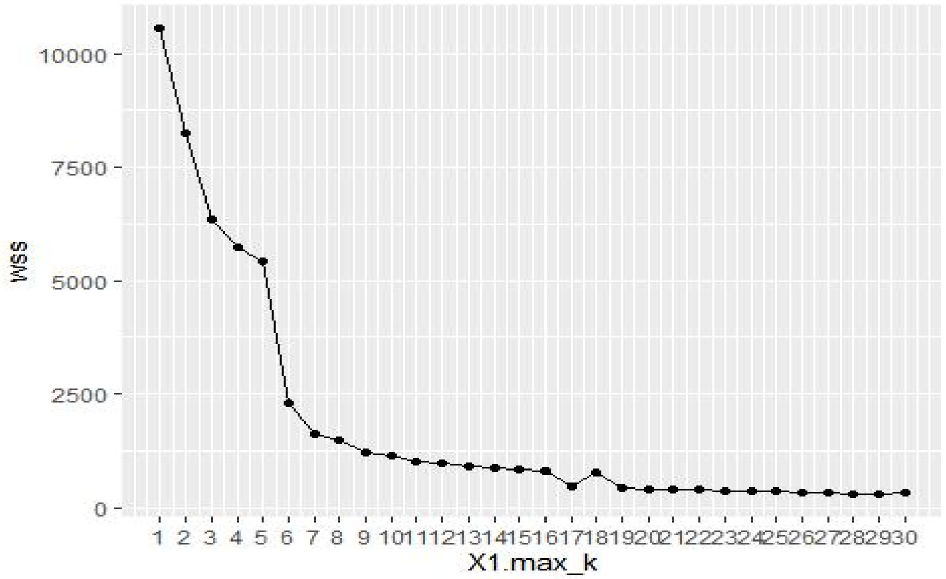
```r
#b)
#a)#replace star
Group_m=aggfunc('star')
arrd=Group_m[,-1]
elbow_method(arrd)
```



```r
Kmean=kmeans(scale(arrd),9,iter.max = 20)
#data with respect to cluster
Newd=arrange(Group_m,desc(ROI))
Newd['Cluster']=Kmean$cluster
#prediction
Rep_direct(Newd,'Sean Gullette')
```

```
## [1] Katie Featherston
## 2504 Levels: 'Weird Al' Yankovic 50 Cent A.J. Cook Aaliyah ... Zooey
 Deschanel
```

```r
#c)#writer
set.seed(12)
Group_m=aggfunc('writer')
arrd=Group_m[,-1]
elbow_method(arrd)
```



```
Kmean=kmeans(scale(arrd),9,iter.max = 20)
#data with respect to cluster
Newd=arrange(Group_m,desc(ROI))
Newd['Cluster']=Kmean$cluster
#prediction
Rep_direct(Newd,'Jared Hess')
```

```
## [1] Oren Peli
```

Interpretation:

1) From above three plot we see that optimal number of clusters are 9.Hence to process further computation we use number of cluster is 9

2) By using above function we may replace any of director,stars or writer.Here we are using the K-means clustering

3) First we fit algoritham and then we arrange data accordingly clusters

4) While replacing any feature we first check the which cluster that feature is belongs ,then based on that cluster only we arrange data with respect highest ‘ROI’ and then we replace that feature which highest ‘ROI

Q.3

As a producer, if you want to choose a country where you would like to settle in order to be successful, which country would you choose? (Assume there are no other constraints like language, nationality etc. You may want to first state your definition of success with justification). Justify your answer.

1) Towards problem statement

Definition of success:

A) I)According to my point of view definition of success would be ,we choose that ‘country’ which having highest number of ‘gross’ collection and having higest number of ‘votes’.

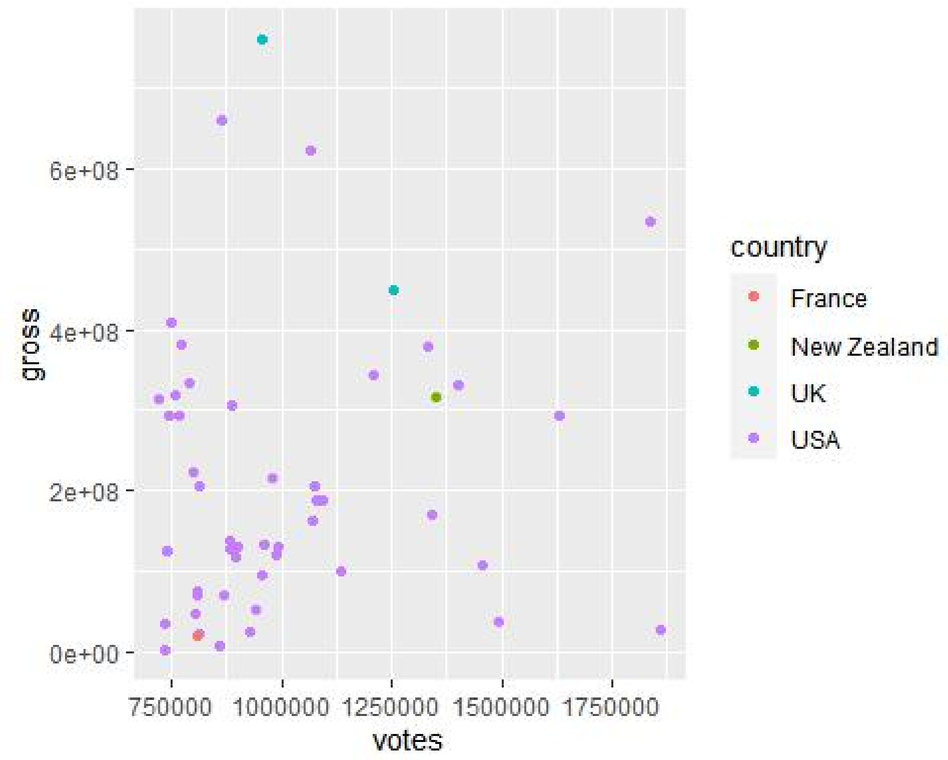
Ii)That is because if any country having more ‘gross’ collection then there is more number of peoples who watch the movies. And if more number of ‘votes’ then there will be more peoples are interested in watching the movies.Hence we prefer that country which having more ‘gross’ collection and more ‘votes’

:R-code

```
suc_data1=subset(Final_data,select = c(votes,gross,country,star,director))
head(suc_data1)
```

```
##    votes     gross country               star      director
## 1 195668 107918810     USA Katie Featherston     Oren Peli
## 2 202691 140539099     USA   Heather Donahue Daniel Myrick
## 3  44989  30610863     USA    Blanchard Ryan   Chris Kentis
## 4  11992   2856622  Canada     Aaron Eckhart    Neil LaBute
## 5 171007  44540956     USA         Jon Heder     Jared Hess
## 6  13291  10178331     USA     Alex Kendrick Alex Kendrick
```

```
a1=aggregate(Final_data$gross, by=list(Final_data$country), FUN=mean)
b1=aggregate(Final_data$votes, by=list(Final_data$country), FUN=sum)
df=data.frame(a1,b1[,2])
colnames(df)=c('Country','gross','votes')

#Plot with respect to votes
ggplot(data=arrange(suc_data1,desc(votes))[1:50,],aes(y=gross,x=votes,colour=country))+
  geom_point()
```



```
ggplot(data=arrange(df,desc(votes)),aes(y=gross,x=votes,colour=Country))
+
  geom_point()
```

```
Comment:Here we plot top 50 values of gross collection after arranging
data with respect to country.we see that in top only four country's
having highest number of gross collection and having high number votes
Hence we say that USA is prefered city for producer
```

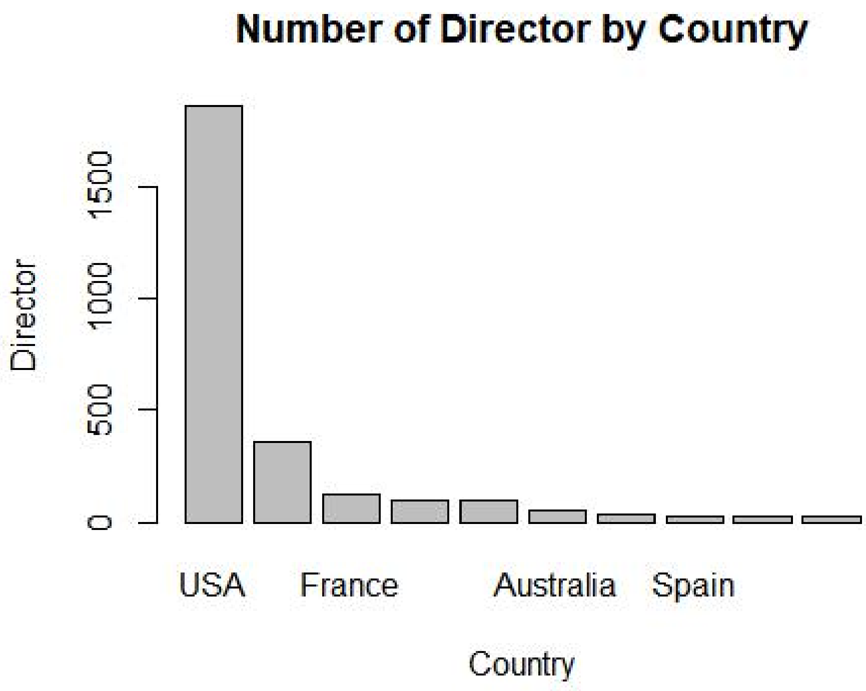
B) I)Here choose that city where more number of stars and directors are living

:R-code

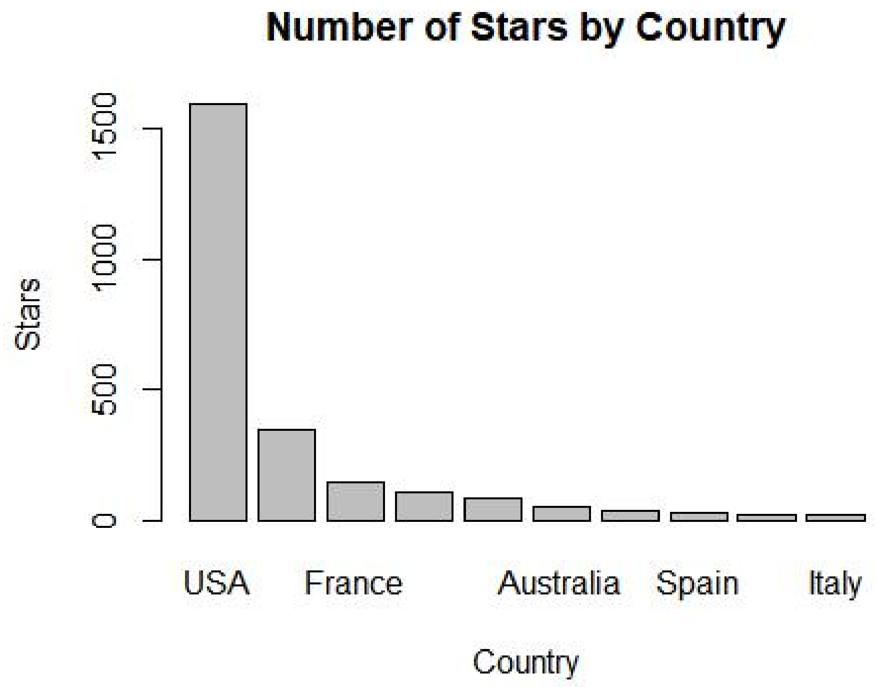
```
s=suc_data1$star
C=suc_data1$country
d=suc_data1$director
v=as.character(unique(C))
no_star=c(0)
no_dir=c(0)
for (i in 1:length(v)){
  no_star[i]=length(unique(s[c(which(C==v[i]))]))
  no_dir[i]=length(unique(d[c(which(C==v[i]))]))
}
dd=data.frame('Country'=v,no_star,no_dir)
head(dd)
```

```
##     Country no_star no_dir
## 1       USA    1592   1861
## 2    Canada     102     98
## 3    France     141    125
## 4 Australia      52     49
## 5     Japan      36     31
## 6      Iran       5      4
```

```
#for director
arr_dir=arrange(dd,desc(no_dir))
arr_dir=arr_dir[1:10,]
barplot(arr_dir$no_dir,
        main = "Number of Director by Country",
        xlab = "Country",
        ylab = "Director",
        names = arr_dir$Country)
```



```
#for stars
arr_str=arrange(dd,desc(no_star))
arr_str=arr_str[1:10,]
barplot(arr_str$no_star,
        main = "Number of Stars by Country",
        xlab = "Country",
        ylab = "Stars",
        names = arr_str$Country)
```



Interpretation:

1) In above two plots we see that in 'USA' most of directors and stars are living.

2) To become successful producer we have to settle in "USA'. That Is because if the 'country ' has more number of 'directors' or more number of 'stars',then we are easily approach any 'star' or 'director' to make any movie.

3) In case A we also see that more revenue are generated from 'USA'.Hence it is better to settle in that country which has more population . 'USA' has highest 'votes'.It shows that more number of peoples are interested to watch movies.

4) Hence we say that 'USA' is the best city to be settle & become successful producer.