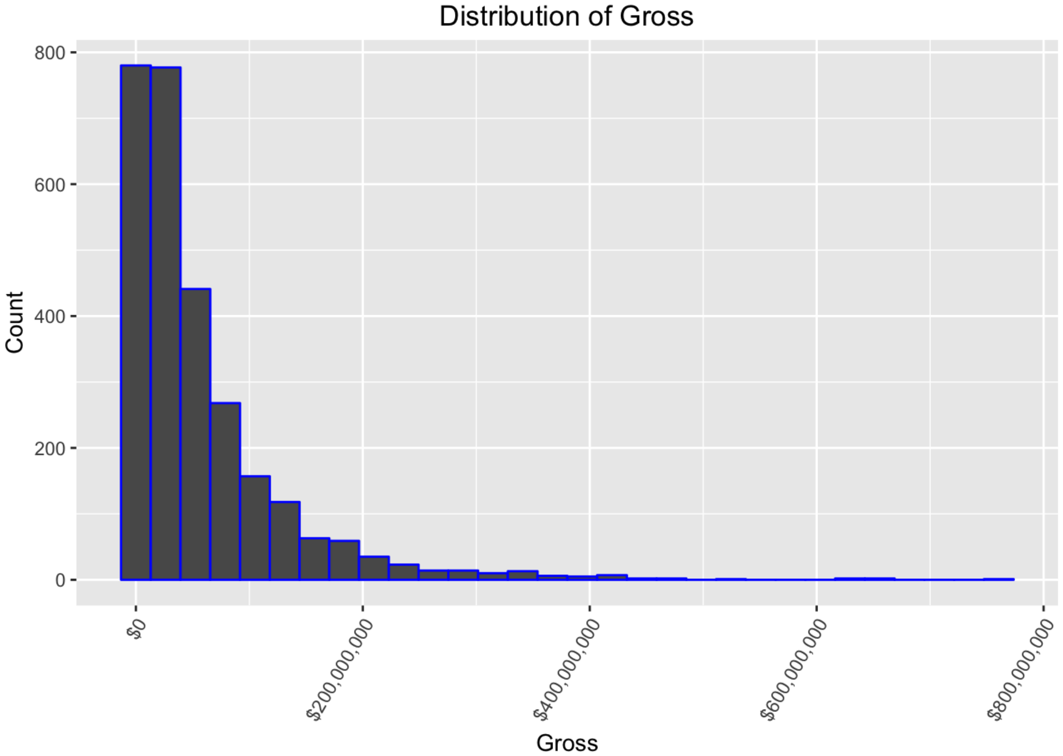
ANALYSING IMDB DATA

Introduction

The aim of this report is to analyze what are the characteristics that have a relationship with gross amount of a movie and build a model to show how changes gross amount with them.For this purpose Imdb 5000 Movie Dataset was chosen to analyze.It is a scrapping data from Imdb website one of the most popular online database of movies where movies can be voted 0 to 10.Data is provided by Kaggle platform.Dataset contains 28 variables for 5043 movies from 66 countries between the years 1920 and 2016.For this report movie title,duration,genres,title year,imdb score,country,gross and budget amounts were taken and movies with no gross information was excluded from the dataset.Moreover although foreign movies budgets and grosses are not in US dollars there is no information about currency in the dataset.Because of that only USA movies was considered to investigate.And it was assumed that inflation factors is fixed over years.Also grouping data by decades showed that there are unblanced sample groups for instance between 1920 and 1930 only 2 movies are in the dataset that’s why movies since 1990 studied for a meaningful interpretation when taking movie decades into account as explanatory variable.It would be interesting to understand what kind of movie more related to gross amount but genres attribute has 914 levels.One movie belongs different kind of genres of movies.For this reason instead of considering all genres to understand behavior of gross,it is decided to add an indicator variable to dataset which split data two groups whether a movie belongs to drama genre or not.Last of all,2800 USA movies since 1990 is inspected in this analysis.

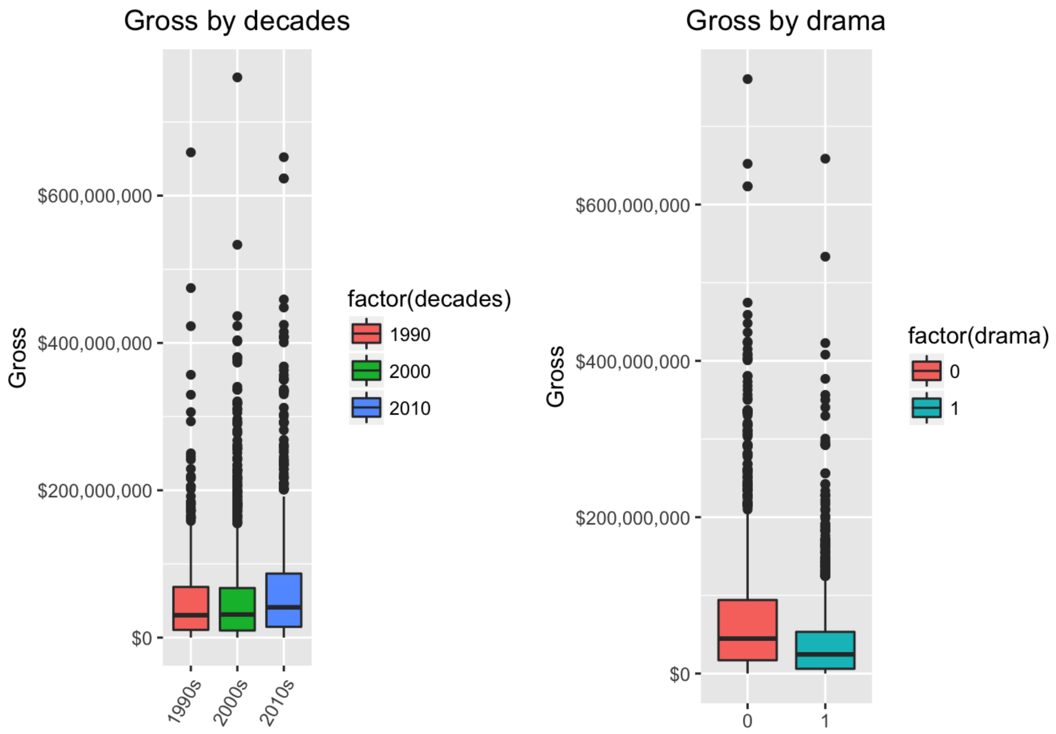
Visualisation and Analysis

Distribution of gross amount highly right skewed with a large range from nearly 700 USD to 800.000.000 USD as shown in Figure1.The average of gross is about 57.000.000 USD with a median nearly 34.000.000 USD.Most of movies have grosses under 100.000.000 USD and movies grossed over 200.000.000 USD are the least frequented ones.Movie budgets are also skewed right with a large range and mostly under 50.000.000 USD.Budget average and median value are around 43.000.000 USD and 28.000.000 USD respectively.They are less than gross average and median.There are many outliers in terms of gross and budget.Gross and budget is much more normalized after transformation with square-root function.Furthermore the majority of imdb scores is around 6 and movie durations are mainly around 100 minutes.(See Appendix 1-Appendix 2)



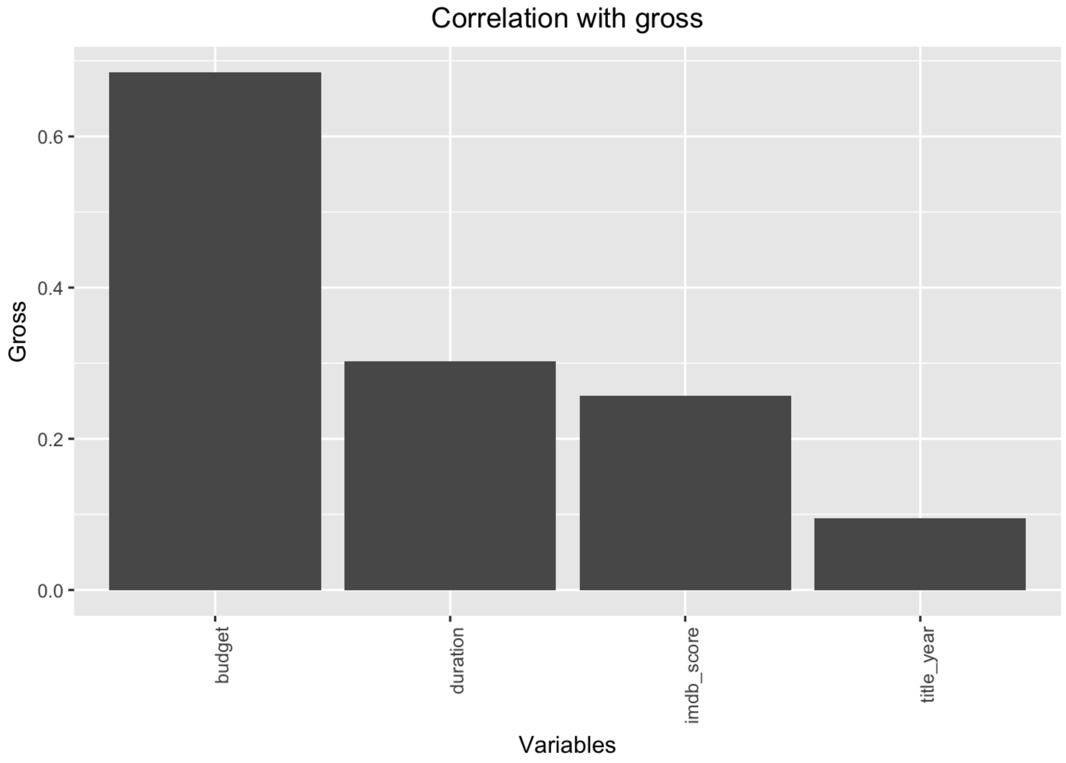
#Figure1

Figure 2 shows gross by movie decades 1990s,2000s and 2010s and gross by drama factor in which 1 indicates movies belong to drama genre and 0 indicates movies don’t belong to drama genre.It can be seen that highest grossing movie is non-drama and belongs to 2000s.Although graphics does not show clearly the differences in terms of gross by factors,Kruskal-Wallis test results significantly showed that grosses of drama movies and non-drama movies are non-identical populations following different distributions.Moreover results indicated that gross amounts of movies belong to 2010s forms a population not same as 2000s and 1990s.However no evidence was found to say populations of movie gross amounts of 2000s and 1990s are non-identical.Kruskal-Wallis test was performed since distributions couldn't be assumed normal by graphs in Figure 2.



#Figure 2

Strength of correlation graph in terms of gross can be seen in Figure 3.Correlation analysis indicated that gross variable strongly possitive correlated with budget and linear relationship between gross amount and the other variables imdb score,duration and title year is weak.Furthermore,multiple correlation was inspected in order not to choose two strongly correlated variables as a predictor variables considering independence of predictor variables when building a linear regression.There isn't any strong correlation between variables except budget and gross amount.(See Appendix 3).

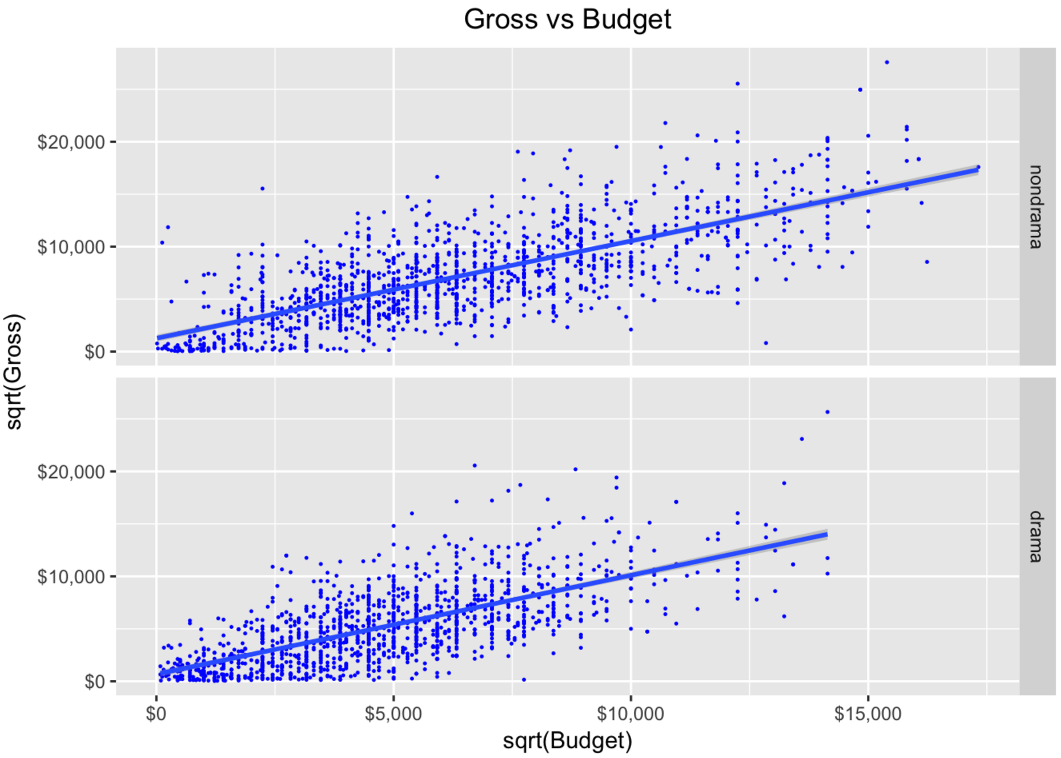


#Figure 3

When building a linear regression model considering gross as a response variable although statistically significant variables were obtained and unsignificant variables were excluded from the model, assumptions of linear regression couldn't met since regression residuals were not normally distributed and not randomly scatter around zero.For this reason square root transformation were applied to the variables in the model.After that residual assumptions were ensured roughly.(See Appendix 4).Linear regression equation are shown below:

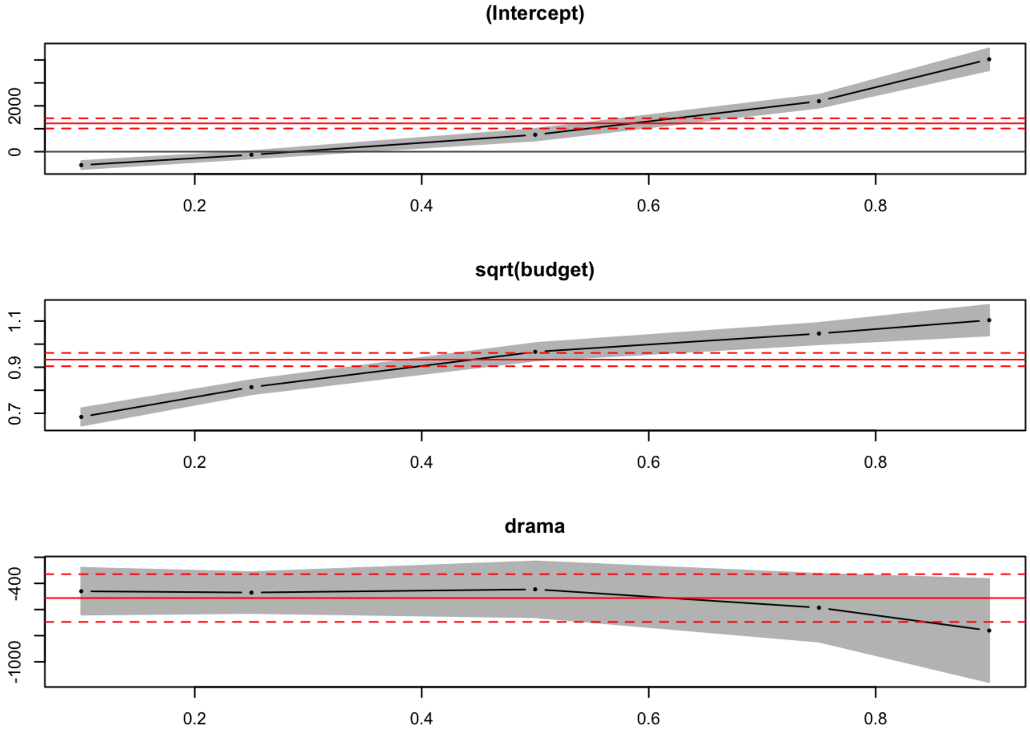
```
        sqrt(gross)  =  0.93293* sqrt(budget) + -512.16964 *drama
```

In linear regression R-squared value is 0.53 and F statistic is very big.As gross has a positive relationship with budget,there is a negative relationship with drama factor.After all,multicollinearity was tested by obtaining VIF for each budget and drama.The results showed that there is no multicollinearity between predictor variables in the model since VIF values were around 1.Visualization of linear regression model is shown in Figure 4.



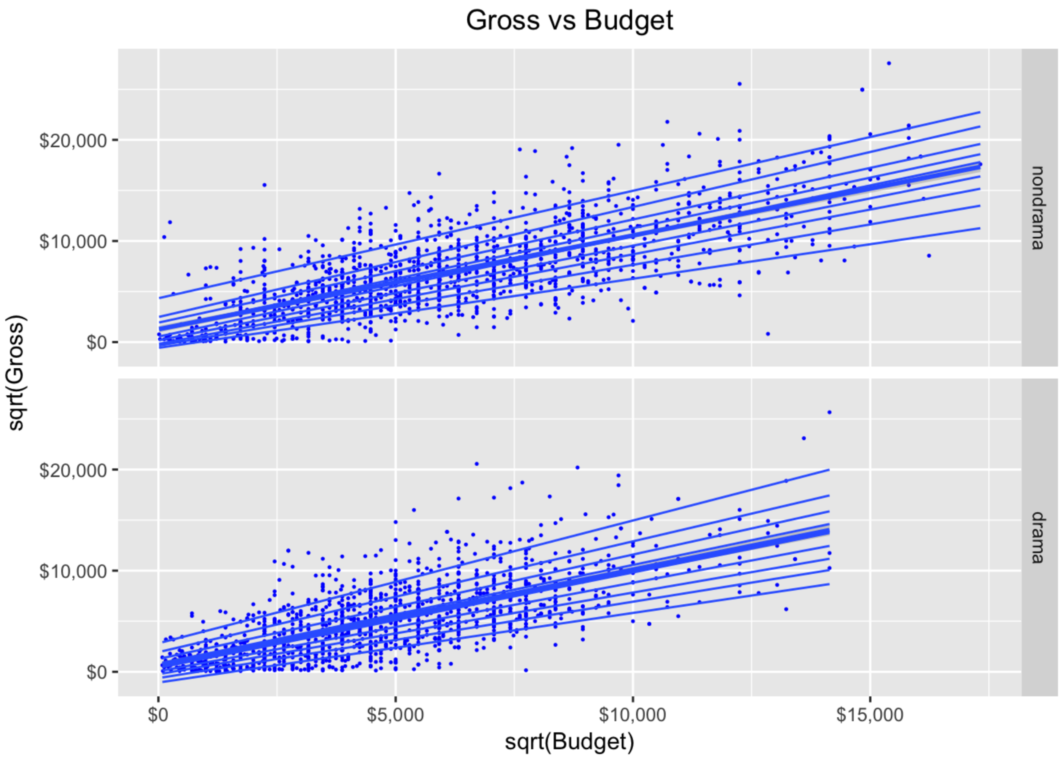
#Figure 4

Additionally quantile regression was performed based on 0.1,0.25,0.5 0.75 and 0.9 quantiles.Figure 5 shows coefficients of quantile regression and confidence for ordinary least squares coefficients with horizontal red lines.As can be seen on graph there are differences in coefficients between linear regression and quantile regression.The intercept and budget coefficient in linear regression are higher than intercepts and budget coefficients of quantile regression for lower quantiles however for upper quantiles it is opposite.Besides drama coefficients of quantile regression are nearly same as linear regression coefficient of drama.Furthermore anova test significantly indicated that the coefficients of quantile regression in four quantiles 0.1,0.25,0.5 0.75 are different.This shows that using quantile regression model instead of ordinary least squares model is more proper to understand the behavior of gross amount.Ordinary least squares regression might not be optimal.



#Figure 5

The quantile regression model is shown in Figure 6.

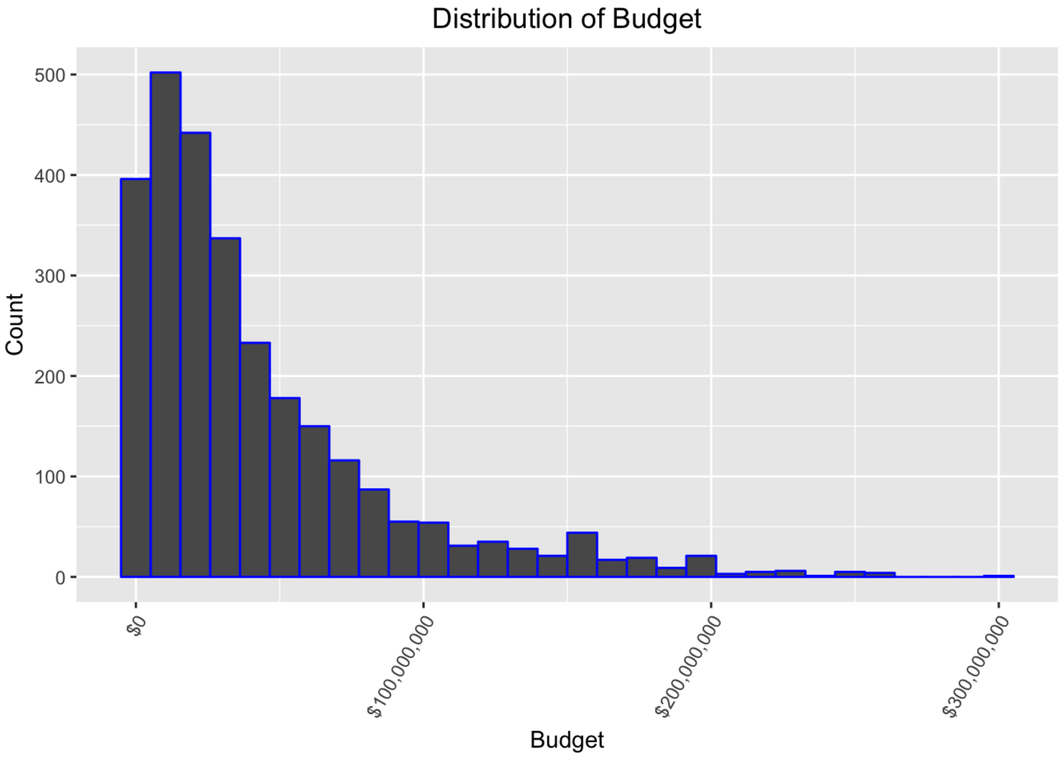


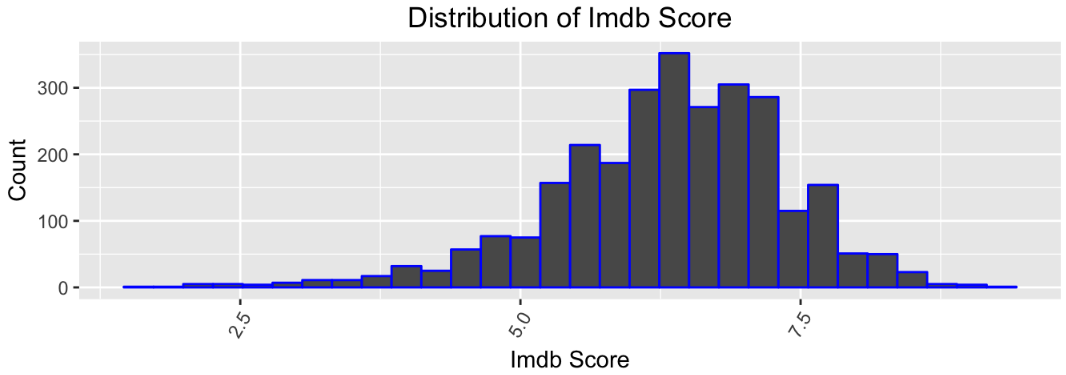
#Figure 6

Conclusion

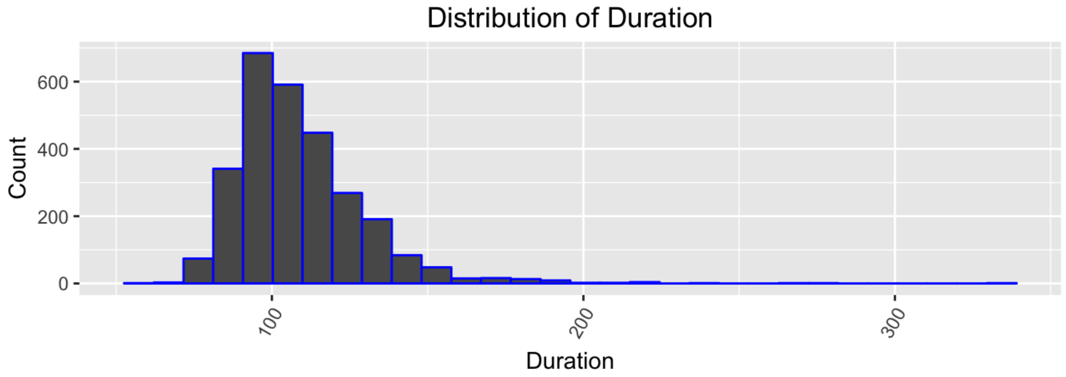
In closing,this analysis suggests that movie gross amount strongly correlated only with budget amount when taking movie characteristics such as duration,title year,imdb score and budget into consideration.Moreover,it shows gross amounts of drama movies and non-drama movies originate from different populations in terms of their distributions and only gross amounts of 2010s movies population is non-identical than the others 1990s and 2000s but there is no reason to declare that movie grosses of 1990s and 2000s is different from each other.Furthermore linear model is provided to interpret the behavior of gross amount based on drama factor and budget amount which shows that generally when budget amount increasing gross amount increasing as well and we can usually expect that non-drama movie makes larger gross than the drama one if they have same budgets.Besides when gross is increasing effect of budget on gross is getting bigger as drama effect on gross is staying nearly same.

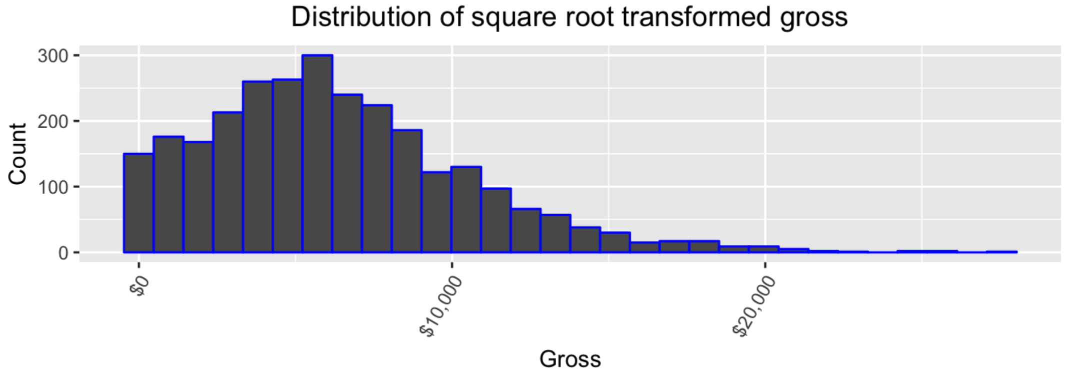
Appendix



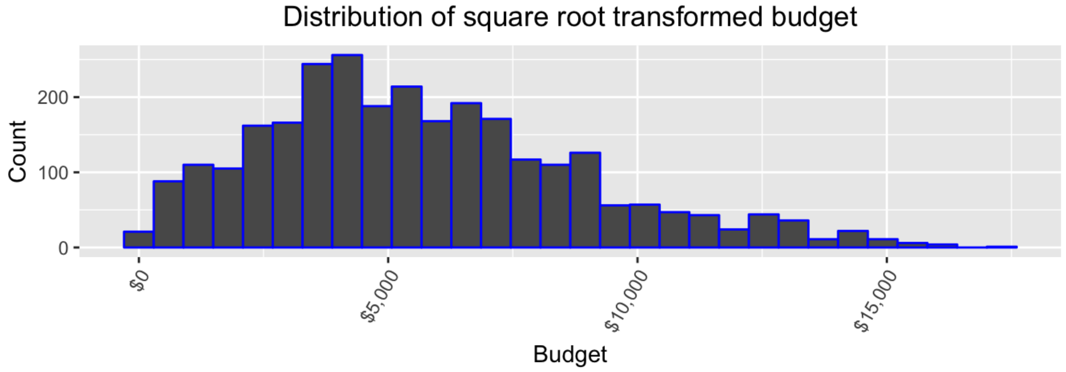


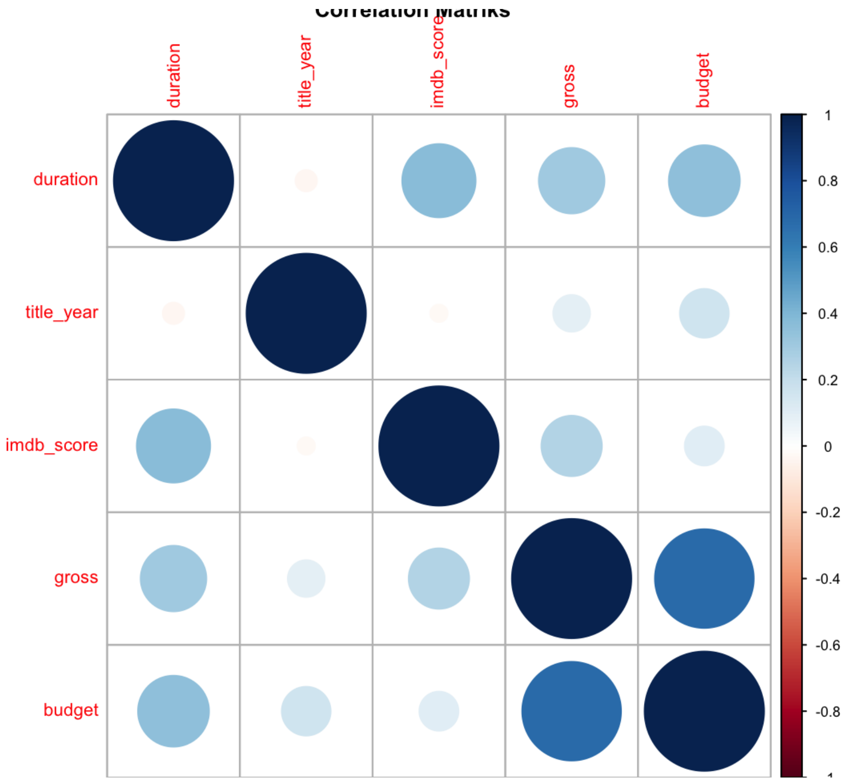
#APPENDIX 1



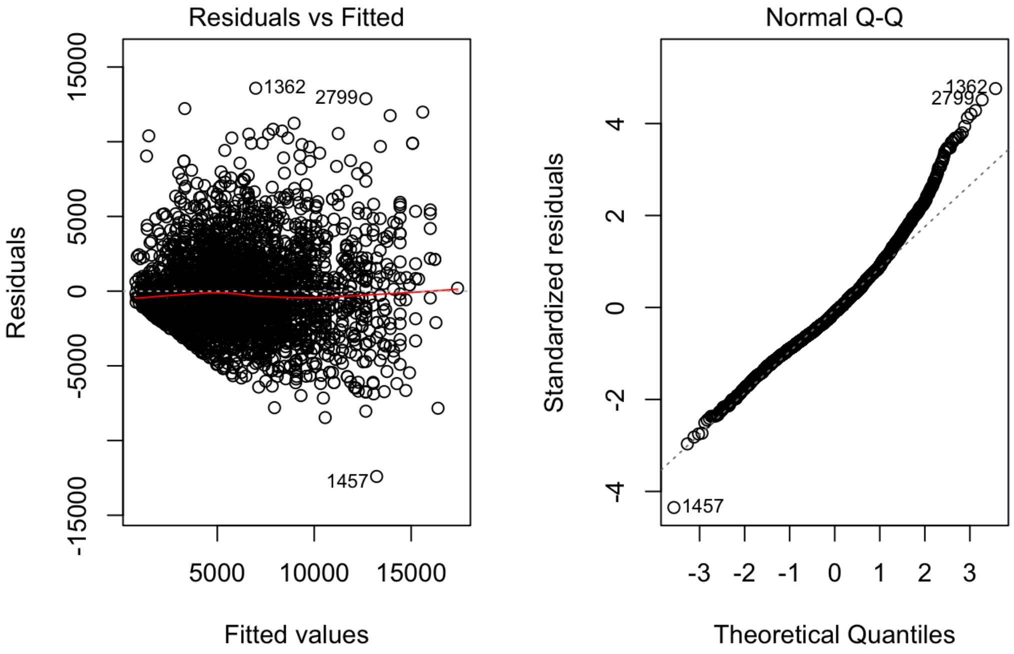


#APPENDIX 2





#APPENDIX 3



#APPENDIX 4

Code

```
library(dplyr)

library(stats)

library(ggplot2)

library(gridExtra)

library(corrplot)

library(sqldf)

library(corrr)

library(VIF)
library(scales)

library(psych)

library(plyr)

library(quantreg)

setwd("/Users/ozgesahin/RProgramming")

movies = read.csv("movie_metadata.csv")

str(movies)

head(movies)

summary(movies)

attach(movies)

count(movies[which(country == "USA"),])

movies %>% group_by(factor(genres)) %>% summarise(count(genres))

movies = dplyr::select(movies,duration,genres,title_year,imdb_score,country,gross,budget,movie_title)

attach(movies)

movies = subset(movies,country =="USA"&gross>0&budget>0&title_year>1990)

attach(movies)

movies = na.omit(movies)

attach(movies)

movies = movies %>% mutate(decades= title_year - (title_year %% 10))

attach(movies)

movies = movies[order(gross),]

attach(movies)

movies %>% group_by(decades) %>% summarise(count(decades))

movies[1:2,"genres"]

head(genres)

#######putting drama factor

dramaset = sqldf("select * from movies where genres  LIKE '%Drama%'")

notdramaset = sqldf("select * from movies where genres  NOT LIKE '%Drama%'")

notdramaset = mutate(notdramaset,drama =0)

dramaset = mutate(dramaset,drama =1)

#######merging

movies<- rbind(dramaset,notdramaset)

attach(movies)

str(movies)

summary(movies)

range(gross)
range(budget)

########Distribution of Gross

ggplot(movies, aes(x=gross))+ geom_histogram(color ="blue")+
  ggtitle("Distribution of Gross")+
  xlab("Gross")+theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=60, hjust=1
))+
  ylab("Count")+scale_x_continuous(labels = dollar)
```

```
########Gross by decades and drama

c = c("1990"="1990s","2000"="2000s","2010"="2010s")

source("multiplot.R")

h1 =ggplot(movies, aes(factor(decades,labels=c), gross)) +
  geom_boxplot(aes(fill=factor(decades)))+
  theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=60, hjust=1))+
  scale_y_continuous(labels = dollar)+
  ggtitle("Gross by decades")+ylab("Gross")+xlab("")

h2 =ggplot(movies, aes(factor(drama),gross)) +
  geom_boxplot(aes(fill=factor(drama)))+
  theme(plot.title = element_text(hjust = 0.5))+scale_y_continuous(labels = dollar)+
  ggtitle("Gross by drama")+ylab("Gross")+xlab("")

multiplot(h1,h2, cols=2)
```

```
########kruskal test

kruskal.test(list(notdramaset$gross,dramaset$gross))#non identical

kruskal.test(list(movies[which(decades ==2000),"gross"],
                  movies[which(decades==1990),"gross"], movies[which(decades ==2010),"gross"]))#non ident
ical

kruskal.test(list(movies[which(decades ==2000),"gross"],movies[which(decades==1990),"gross"]))# identical

kruskal.test(list(movies[which(decades ==2000),"gross"],movies[which(decades==2010),"gross"]))#non identi
cal

kruskal.test(list(movies[which(decades ==1990),"gross"],movies[which(decades==2010),"gross"]))#non identi
cal

########focus correlation with gross

dplyr::select(movies,duration,title_year,gross,budget,imdb_score) %>% correlate() %>% focus(gross)%>%
  ggplot(aes(x = rowname, y =gross)) +
  geom_bar(stat = "identity") +
  ylab("Gross") +ggtitle("Correlation with gross")+
  xlab("Variables")+theme(axis.text.x = element_text(angle = 90, hjust = 1)) +theme(plot.title = element_
text(hjust = 0.5))
```

```
#######linear regression

gross.reg<-lm(gross~budget+drama+duration+decades+imdb_score,data=movies)

#######exclude unsignificant predicted values

gross.reg<-lm(gross~budget+imdb_score+drama,data =movies)

########with log transformation

log <- function(x) ifelse(x <= 0, 0, base::log(x))  # redefine log to be zero #when x<0

gross.reg<-lm(log(gross)~log(budget)+drama+imdb_score,data =movies)#0.49

########with sqrt root transformation

gross.reg<-lm((sqrt(gross))~sqrt(budget)+drama+duration+imdb_score,data =movies)

gross.reg<-lm((sqrt(gross))~sqrt(budget)+drama,data =movies)

summary(gross.reg)

#######Outliers Test

car::outlierTest(gross.reg)

#######test multicollinearity in regression

vif(gross.reg)

#######Gross vs Budget method lm

ggplot(movies, aes(x=sqrt(budget),y=sqrt(gross)))+ geom_point(color ="blue",size =0.2)+
  ggtitle("Gross vs Budget")+
  xlab("sqrt(Budget)")+theme(plot.title = element_text(hjust = 0.5))+
  ylab("sqrt(Gross)")+scale_y_continuous(labels = dollar)+
  scale_x_continuous(labels = dollar)+
  facet_grid(factor(drama,labels = c("0"="nondrama","1"="drama"))~ .)+geom_smooth(method = lm)
```

```
#######quantile regression

gross.quantile =rq(formula = sqrt(gross)~sqrt(budget)+drama,
                   tau =c(0.1,0.25,0.5,0.75,0.9), data =movies)

plot(summary(gross.quantile,parm="x"))
```

```r
#######quantile regression

gross.quantile1 =rq(formula = sqrt(gross)~sqrt(budget)+drama, tau =0.1, data =movies)

gross.quantile2 =rq(formula = sqrt(gross)~sqrt(budget)+drama, tau =0.25, data =movies)

gross.quantile3 =rq(formula = sqrt(gross)~sqrt(budget)+drama, tau =0.5, data =movies)

gross.quantile4 =rq(formula = sqrt(gross)~sqrt(budget)+drama, tau =0.75, data =movies)

gross.quantile5 =rq(formula = sqrt(gross)~sqrt(budget)+drama, tau =0.9, data =movies)

summary(gross.quantile)

coef(gross.quantile)

#######anova

anova(gross.quantile1,gross.quantile2,gross.quantile3,gross.quantile4,gross.quantile5)

#######reject null hypothesis the coefficients are different

#######quantile regression graph

qs =1:9/10

ggplot(movies, aes(x=sqrt(budget),y=sqrt(gross)))+ geom_point(color ="blue",size =0.2)+
  ggtitle("Gross vs Budget")+
  xlab("sqrt(Budget)")+theme(plot.title = element_text(hjust = 0.5))+
  ylab("sqrt(Gross)")+scale_y_continuous(labels = dollar)+
  scale_x_continuous(labels = dollar)+
  facet_grid(factor(drama,labels = c("0"="nondrama","1"="drama"))~ .)+
  geom_smooth(method = lm)+geom_quantile(quantiles=qs)
```

```r
#######Appendix

#######Distributions

g1 =ggplot(movies, aes(x=imdb_score))+ geom_histogram(color ="blue")+
  ggtitle("Distribution of Imdb Score")+
  xlab("Imdb Score")+
  theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=60, hjust=1))+
  ylab("Count")

g2 =ggplot(movies, aes(x=duration))+ geom_histogram(color ="blue")+
  ggtitle("Distribution of Duration")+
  xlab("Duration")+
  theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=60, hjust=1))+
  ylab("Count")

ggplot(movies, aes(x=budget))+ geom_histogram(color ="blue")+
  ggtitle("Distribution of Budget")+
  xlab("Budget")+
  theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=60, hjust=1))+
  ylab("Count")+scale_x_continuous(labels = dollar)
```

```r
source("multiplot.R")

multiplot(g1,g2, cols=1)
```

```
#######Distribution of square root transformed variables

t1 =ggplot(movies, aes(x=sqrt(gross)))+ geom_histogram(color ="blue")+
  ggtitle("Distribution of square root transformed gross")+
  xlab("Gross")+theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=60, hjust=1
))+
  ylab("Count")+scale_x_continuous(labels = dollar)

t2 =ggplot(movies, aes(x=sqrt(budget)))+ geom_histogram(color ="blue")+
  ggtitle("Distribution of square root transformed budget")+
  xlab("Budget")+theme(plot.title = element_text(hjust = 0.5),axis.text.x = element_text(angle=60, hjust=
1))+
  ylab("Count")+scale_x_continuous(labels = dollar)

multiplot(t1,t2, cols=1)
```

```
#######multiple correlation

cex.before <- par("cex")

par(cex = 0.7)

corrd =cor(dplyr::select(movies,duration,title_year,imdb_score,gross,budget))

corrplot(corrd, method = "circle",main = "Correlation Matriks")
```

```
#######linear regression residuals

gross.reg<-lm((sqrt(gross))~sqrt(budget)+drama,data =movies)#0.53

par(mfrow=c(1,2))

plot(gross.reg,which = 1)

plot(gross.reg,which = 2)
```

Sources

https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

http://data.library.virginia.edu/getting-started-with-quantile-regression/

http://www.statisticalanalysisconsulting.com/wp-content/uploads/2011/06/SA04.pdf

http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

http://www.math.smith.edu/r/excerpt-4.pdf

https://rpubs.com/ibn_abdullah/rquantile

https://www.cscu.cornell.edu/news/statnews/stnews70.pdf