Exploratory Data Analysis of Movies' Revenue, Runtime, and Rating

Arbind Shrestha

#Big Idea: How does a movie's runtime and rating affect its revenue?

#Installing required library and packages. Importing a csv file.

```
suppressMessages(library(readr))
suppressMessages(Movies <- read_csv("Movies.csv"))
suppressMessages(install.packages("randomcoloR"))
suppressMessages(library(randomcoloR))
suppressMessages(install.packages("plotly"))
suppressMessages(library(plotly))
suppressMessages(library(dplyr))
suppressMessages(library(ggplot2))
suppressMessages(install.packages("ggcorrplot"))
suppressMessages(library(ggcorrplot))
suppressMessages(install.packages("ggridges"))
suppressMessages(library(ggridges))
```

#The data is gathered from IMDb.

#Printing the header of the Movies data frame

```
head(Movies)
```

Title	Genre
<chr>	<chr>
Guardians of the Galaxy	Action,Adventure,Sci-Fi
Prometheus	Adventure,Mystery,Sci-Fi
Split	Horror,Thriller
Sing	Animation,Comedy,Family
Suicide Squad	Action,Adventure,Fantasy
The Great Wall	Action,Adventure,Fantasy

6 rows | 1-2 of 11 columns

#Adding and Deleting columns and Filtering items in the columns.

##Removing Despcription, Director, Actors, Year, Metascore, Genre, and Title columns

```
movies_df <- select(Movies,-Description,-Director,-Actors,-Year, -Metascore,-Genre,-Title)
```

##Converting numerical Rating into categorial value. Movies rated above 8 are the 'best', between 7 and 8 are 'good', between 6 and 7 are 'fair', and below 6 are 'poor'.

```
movies_df<-mutate(movies_df,
               Grade = ifelse(Rating>=8, "Best",
                               ifelse(Rating>=7&Rating<8, "Good",
                                      ifelse(Rating>=6&Rating<7,"Fair",
                                             "Poor"))))
```

##Removing null values in the Revenue and Runtime columns. The new data fram is named as 'movies_new

```
movies_new=filter(movies_df, movies_df$Revenue!='NA' & movies_df$Runtime!='NA')
head(movies_new)
```

Runtime	Rating	Votes	Revenue	Grade
<dbl>	<dbl>	<dbl>	<dbl>	<chr>
121	8.1	757074	333.13	Best
124	7.0	485820	126.46	Good
117	7.3	157606	138.12	Good
108	7.2	60545	270.32	Good
123	6.2	393727	325.02	Fair
103	6.1	56036	45.13	Fair

6 rows

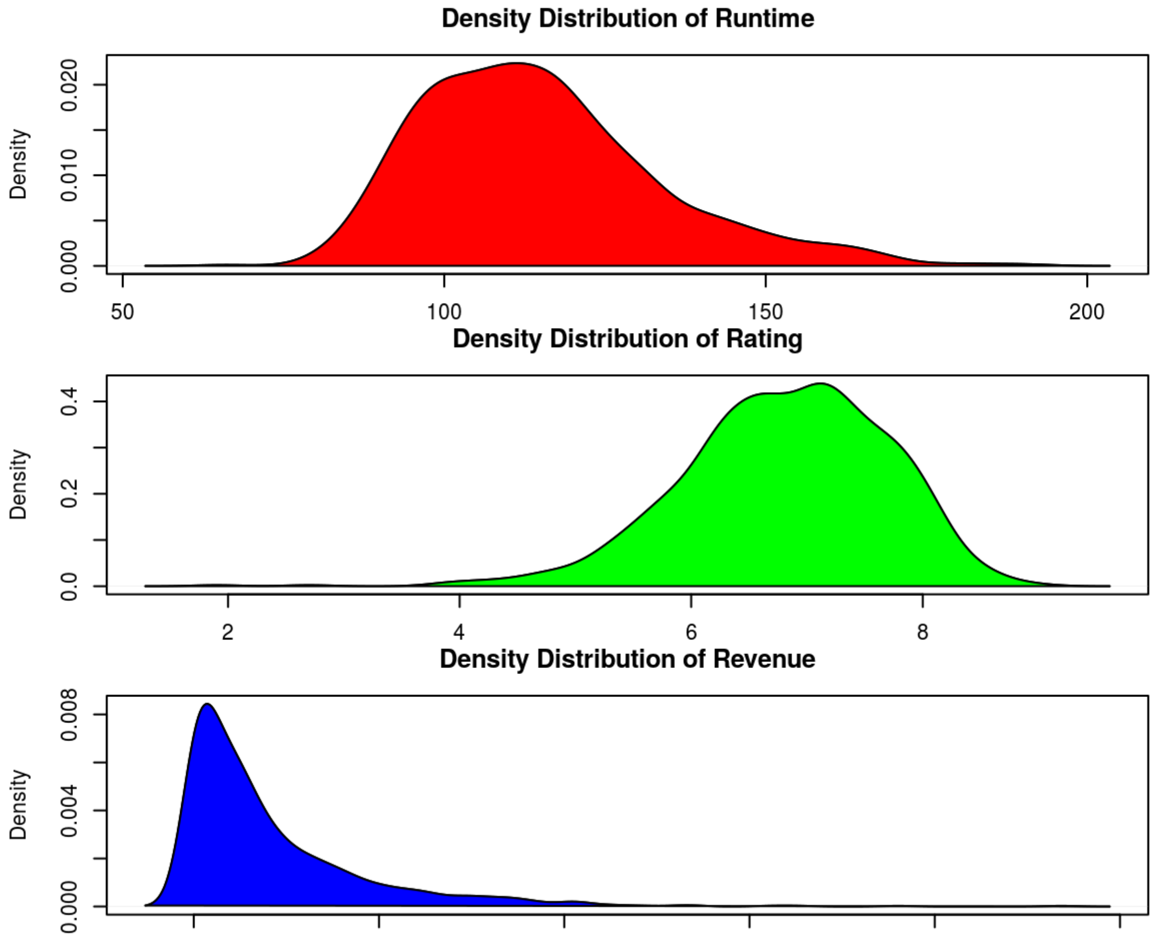
#Checking the summary of the data frame movies_new

```
summary(movies_new)
```

```
##     Runtime          Rating          Votes            Revenue
##  Min.   : 66.0   Min.   :1.900   Min.   :    178   Min.   :  0.00
##  1st Qu.:101.0   1st Qu.:6.300   1st Qu.:  60628   1st Qu.: 13.27
##  Median :112.0   Median :6.900   Median : 134654   Median : 47.98
##  Mean   :114.8   Mean   :6.814   Mean   : 190970   Mean   : 82.96
##  3rd Qu.:125.0   3rd Qu.:7.500   3rd Qu.: 267833   3rd Qu.:113.72
##  Max.   :191.0   Max.   :9.000   Max.   :1791916   Max.   :936.63
##     Grade
##  Length:872
##  Class :character
##  Mode  :character
##
##
##
```

#Checking the density distributin of Runtime, Revenue, and Rating

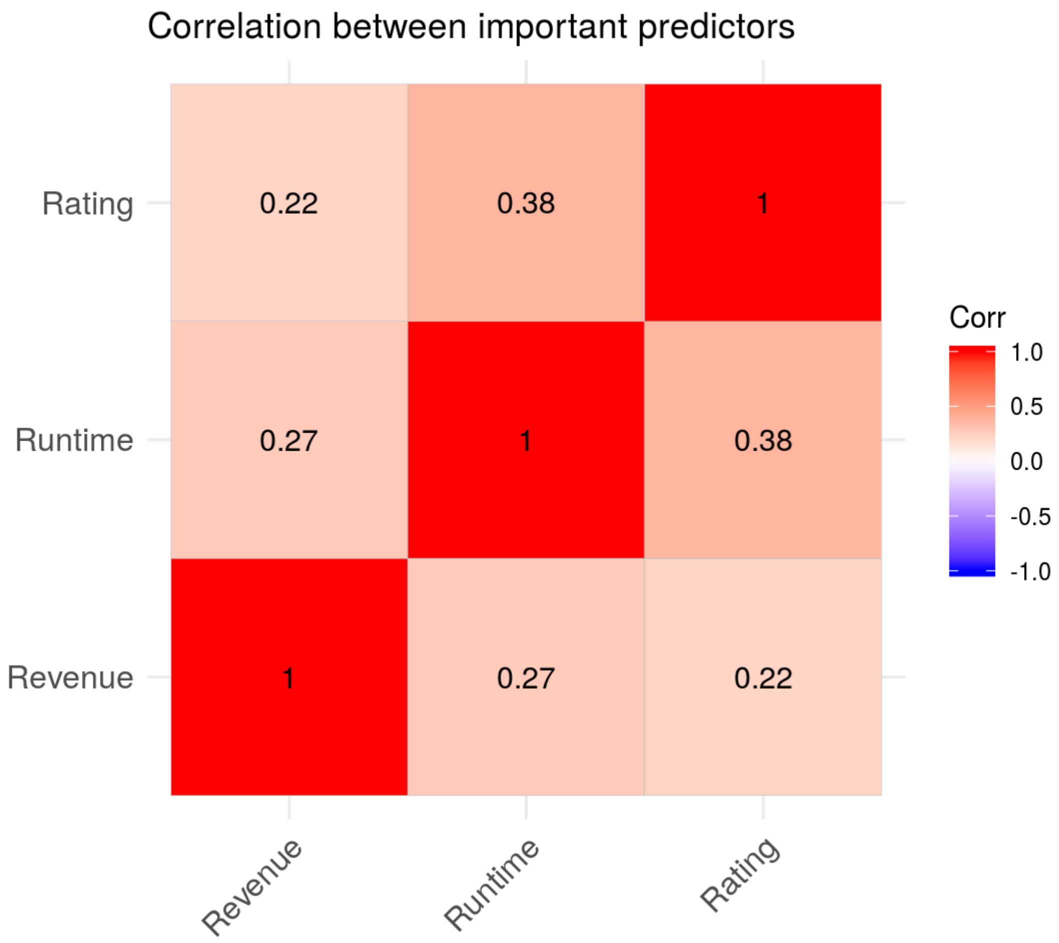
```
par(mfrow=c(3,1),mar=c(1, 6, 3, 6))
h1 <- density(movies_new$Runtime)
plot(h1, main="Density Distribution of Runtime")
polygon(h1, col="red")

h2 <- density(movies_new$Rating)
plot(h2, main="Density Distribution of Rating")
polygon(h2, col="green")

h3 <- density(movies_new$Revenue)
plot(h3, main="Density Distribution of Revenue")
polygon(h3, col="blue")
```



###The runtime and rating seems closer to normal distribution. The revenue is positively skewed. Thus, the mean is greater than the median.

#Question: How does the runtime, revenue, rating correlate to each other?

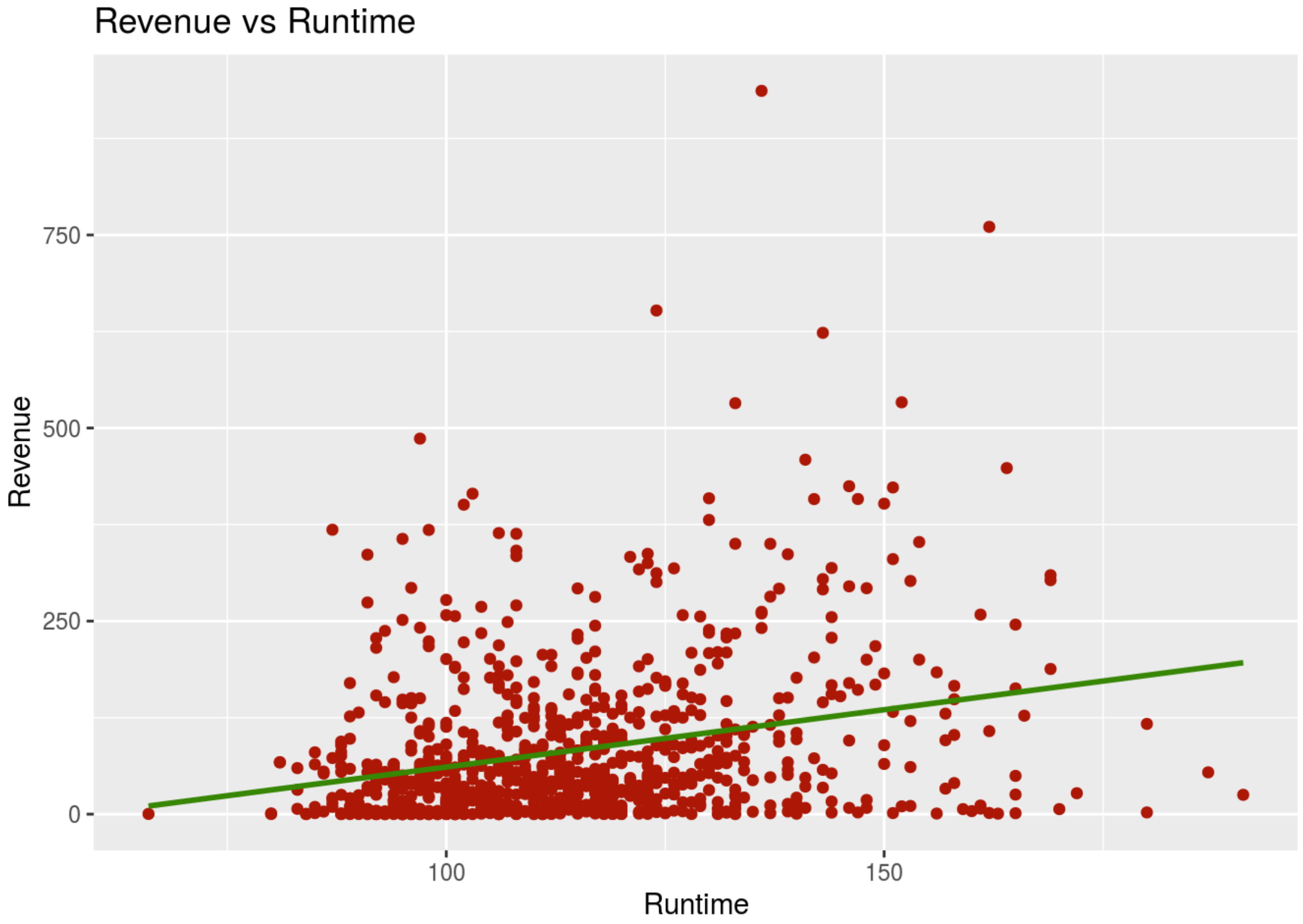
```
cor.movies <- cor(movies_new[,c('Runtime','Revenue','Rating')])
ggcorrplot(cor.movies, hc.order = TRUE,lab=TRUE) + ggtitle("Correlation between importan
t predictors")
```



###It looks like the revenue is weakly and positively correated to both the runtime and rating. There is slightly better correlation between the rating and the runtime.

##Plotting a linear model of the revenue versus runtime.

```
ggplot(movies_new, aes(x=Runtime, y=Revenue))+
    geom_point(alpha=1, col=randomColor(luminosity = 'dark'))+
    geom_smooth(se=FALSE, method='lm',col=randomColor(luminosity = 'dark'))+
    xlab('Runtime')+
    ylab("Revenue")+
    ggtitle("Revenue vs Runtime")
```



##Looking statistically.

```
mod1=lm(Revenue~Runtime, data=movies_new)
summary(mod1)
```

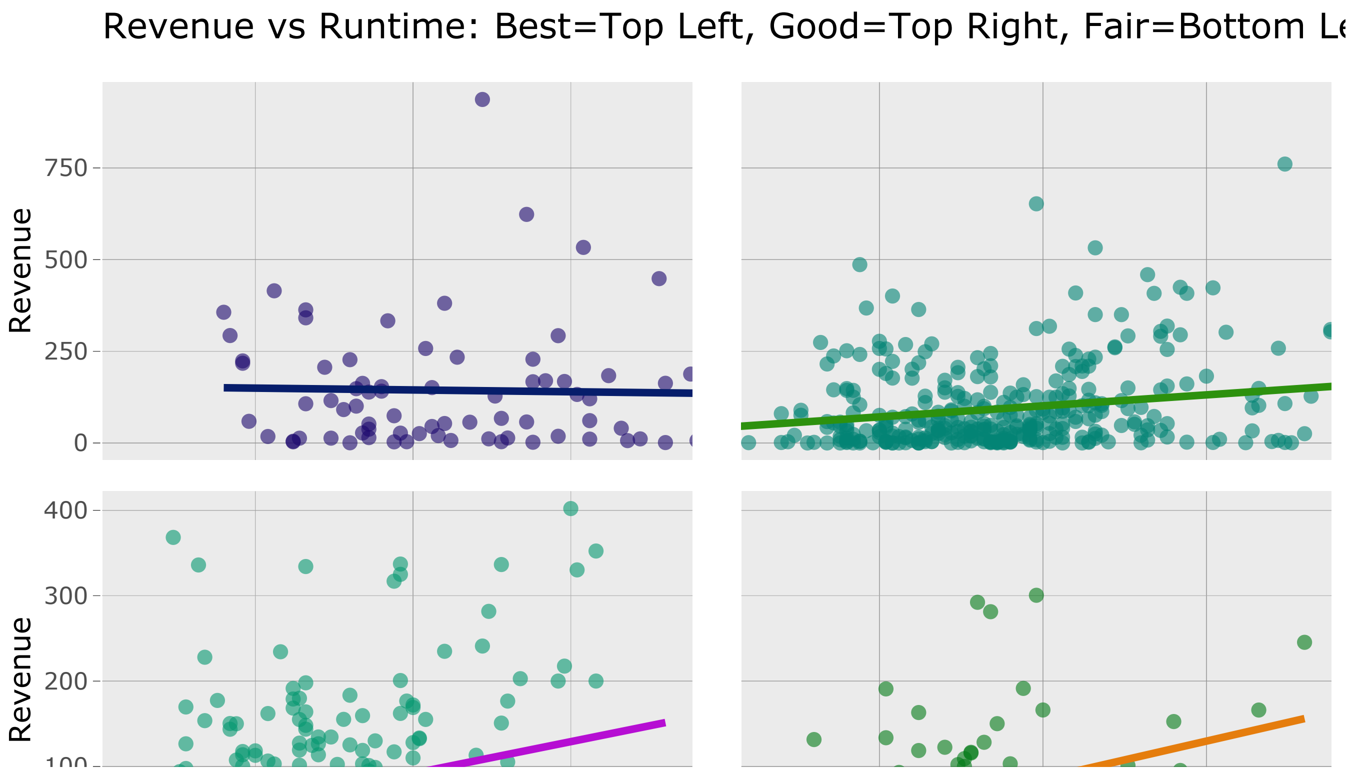
```
## 
## Call:
## lm(formula = Revenue ~ Runtime, data = movies_new)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -177.59  -58.76  -28.88   29.83  822.17 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -87.477     21.047  -4.156 3.56e-05 ***
## Runtime        1.485      0.181   8.203 8.36e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 99.53 on 870 degrees of freedom
## Multiple R-squared:  0.0718, Adjusted R-squared:  0.07073 
## F-statistic:  67.3 on 1 and 870 DF,  p-value: 8.359e-16
```

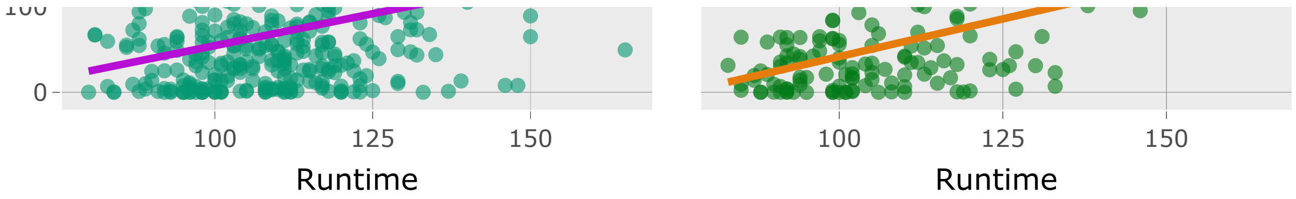
##The runtime is very significant in our model.

#.....

##Next, plotting different linear models of the revenue versus runtime based the categorical rating of the movies.

```
#Creating a list of the categorical rating and assigning to grade1.
grade1<-c('Best', 'Good', 'Fair', 'Poor')

#Creating an empty list.
sublist<-list()

#Using a for loop to loop over the data frame movies_new.
for (i in grade1)
  {
#Filtering the dataframe based on the string rating.
p<-filter(movies_new, Grade==i)

#GGploting the graph of the filtered dataframe.

g<-ggplot(p, aes(x=Runtime, y=Revenue))+
  geom_point(alpha=.6,col=randomColor(luminosity = 'dark'))+
  geom_smooth(se=FALSE,method='lm',col=randomColor(luminosity = 'dark'))+
  ylab("Revenue")+
  xlab("Runtime") +
  ggtitle("Revenue vs Runtime: Best=Top Left, Good=Top Right, Fair=Bottom Left, Poor=Bot
tom Right")
#print(g)

sublist[[i]] <- g
  }

#Combining all the graphs into one using subplot function.
subplot(sublist,nrows = 2, widths = NULL, heights = NULL, shareX = TRUE, shareY = TRUE)
```





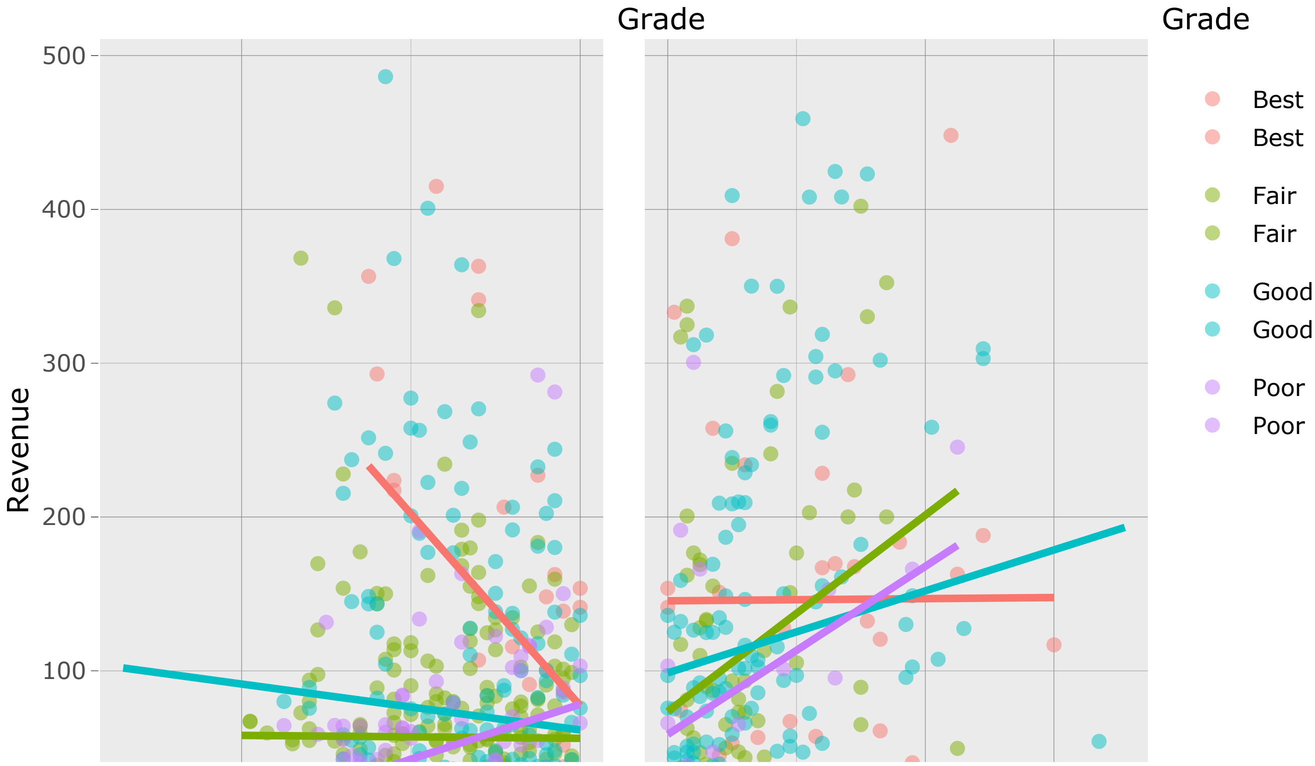
#It looks there is a strong positive correlation between the runtime and revenue for the movies rated poor and fair. There is weak relation for good movies and slight negative or zero for best movies.

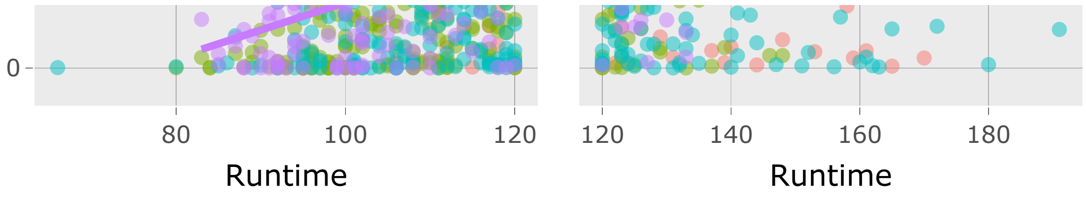
##Lets further investigate how runtime affects revenue by dividing the runtime into two catagories. One with lenth greater than 120 and another less then 120min

```
#Using filtere function to divide.
Ltime<-filter(movies_new, Runtime>=120)
Stime<-filter(movies_new, Runtime<=120)
```

#Ploting linear models with the different runtimes and Grades

```
#short movies with Grades
q1<-ggplot(Stime, aes(x=Runtime, y=Revenue, col=Grade))+
  geom_point(alpha=.5)+
  geom_smooth(se=FALSE, method = "lm")

#long movies with Grades
q2<-ggplot(Ltime, aes(x=Runtime, y=Revenue, col=Grade))+
  geom_point(alpha=.5)+
  geom_smooth(se=FALSE, method = "lm")


#Combining both graphs
subplot(q1,q2, shareY = TRUE, shareX = TRUE)
```





#It appears best rated movies make less revenue as the runtime increases.

#Looking the above graph statistically.

```
#Regression analysis with grades and revenue
mod2=lm(Revenue~Runtime+Grade, data=movies_new)
mod2
```

```
##
## Call:
## lm(formula = Revenue ~ Runtime + Grade, data = movies_new)
##
## Coefficients:
## (Intercept)      Runtime    GradeFair    GradeGood    GradePoor
##      -9.624        1.176      -50.250      -36.256      -62.533
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = Revenue ~ Runtime + Grade, data = movies_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -183.70  -57.38  -24.59   28.89  786.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.6235    27.6792  -0.348  0.72816
## Runtime       1.1756     0.1933   6.081 1.78e-09 ***
## GradeFair   -50.2499    13.3083  -3.776  0.00017 ***
## GradeGood   -36.2560    12.8166  -2.829  0.00478 **
## GradePoor   -62.5330    15.0405  -4.158 3.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.6 on 867 degrees of freedom
## Multiple R-squared:  0.09235,    Adjusted R-squared:  0.08816
## F-statistic: 22.05 on 4 and 867 DF,  p-value: < 2.2e-16
```

```
#Regression analysis with grades and revenue, with an interaction
mod3=lm(Revenue~Runtime+Grade+Runtime*Grade, data=movies_new)
mod3
```

```
##
## Call:
## lm(formula = Revenue ~ Runtime + Grade + Runtime * Grade, data = movies_new)
##
## Coefficients:
##       (Intercept)            Runtime          GradeFair
##          169.3769            -0.1978          -263.8603
##         GradeGood          GradePoor  Runtime:GradeFair
##         -218.2540          -303.4629             1.6882
## Runtime:GradeGood  Runtime:GradePoor
##            1.3985             1.9564
```

```
summary(mod3)
```

```
##
## Call:
## lm(formula = Revenue ~ Runtime + Grade + Runtime * Grade, data = movies_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -165.05  -57.62  -23.02   30.09  794.15
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        169.3769    73.4453   2.306  0.02134 *
## Runtime             -0.1978     0.5566  -0.355  0.72242
## GradeFair         -263.8603    83.9704  -3.142  0.00173 **
## GradeGood         -218.2540    80.6085  -2.708  0.00691 **
## GradePoor         -303.4629    95.9275  -3.163  0.00161 **
## Runtime:GradeFair    1.6882     0.6667   2.532  0.01151 *
## Runtime:GradeGood    1.3985     0.6208   2.253  0.02453 *
## Runtime:GradePoor    1.9564     0.8006   2.444  0.01474 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.32 on 864 degrees of freedom
## Multiple R-squared:  0.1005, Adjusted R-squared:  0.09324
## F-statistic: 13.79 on 7 and 864 DF,  p-value: < 2.2e-16
```

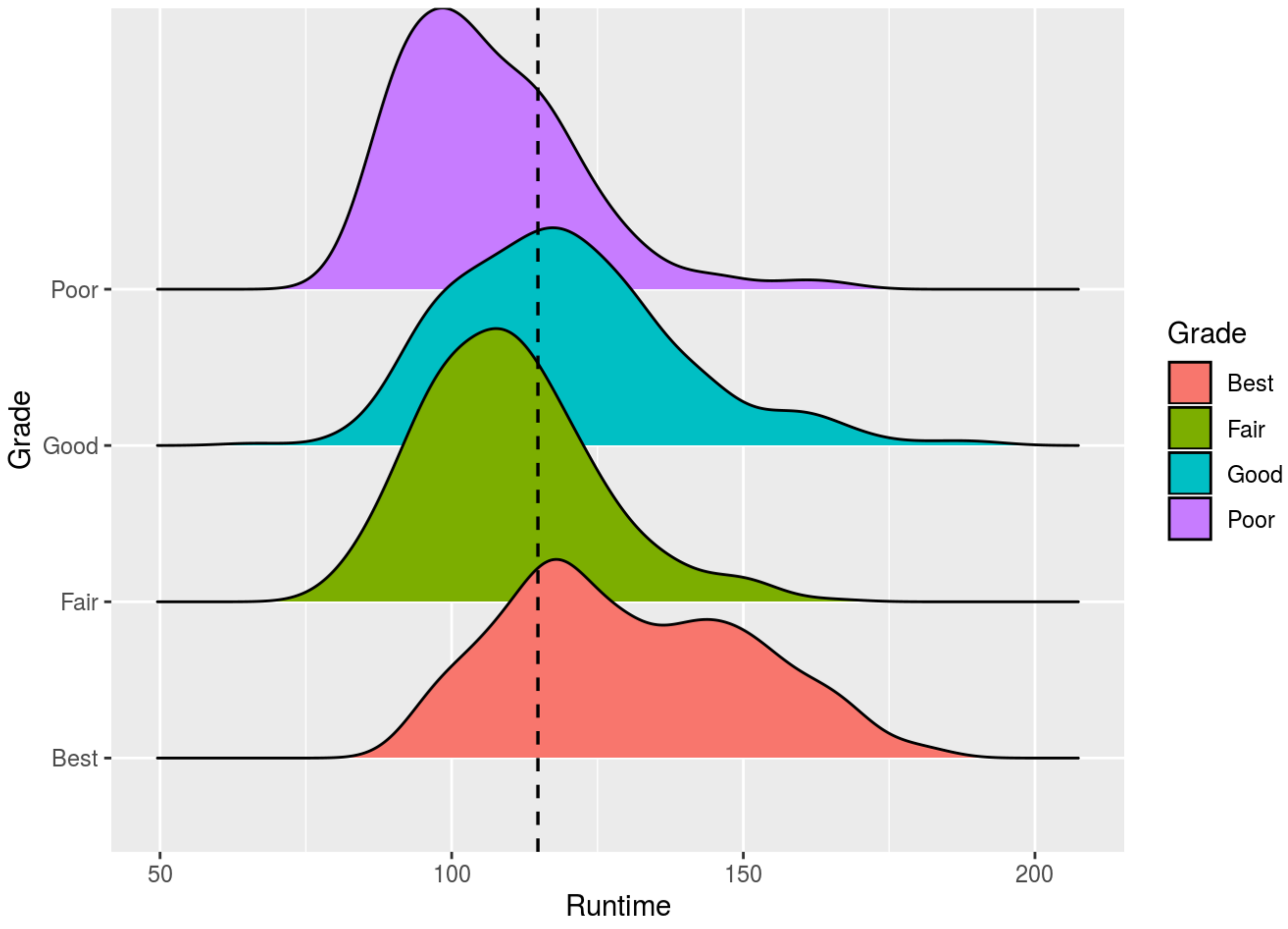
##It is clear that the revenue decreases with the decrease in the rating of a movie. But, the poor, fair, and even good rated movies earn more revenue with the increase of their runtime. The case is opposite for the best movies.

#Question: What is going on with the best movies?

#Separating the density distribution of different rated movies and the runtime.

```
#Using density_ridges funtion to creat density distribution.
den1<-ggplot(movies_new,aes(x = Runtime,y=Grade))+
  geom_density_ridges(aes(fill = Grade))+
  geom_vline(aes(xintercept = mean(movies_new$Runtime)),  linetype = "dashed", size = 0.
6)
den1
```

```
## Picking joint bandwidth of 5.48
```

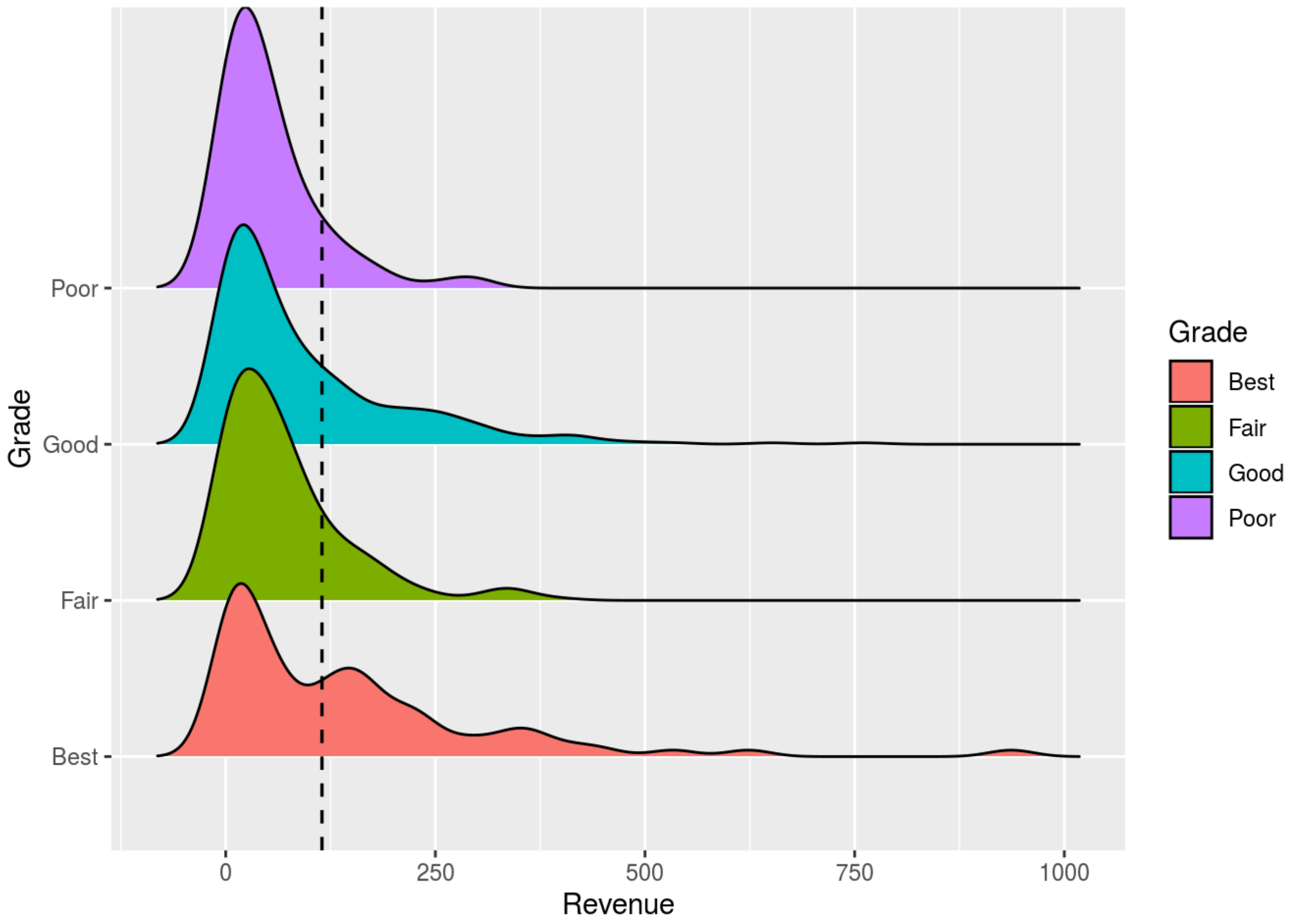


##Majority of the poor movies have short runtime. In contrast, majority of the best rated movies are lenghty.

##Separating the density distribution of different rated movies and the revenue.

```
#Using density_ridges funtion to creat density distribution.
den2<-ggplot(movies_new,aes(x = Revenue, y=Grade))+
  geom_density_ridges(aes(fill = Grade))+
  geom_vline(aes(xintercept = mean(movies_new$Runtime)),  linetype = "dashed", size = 0.
6)

den2
```

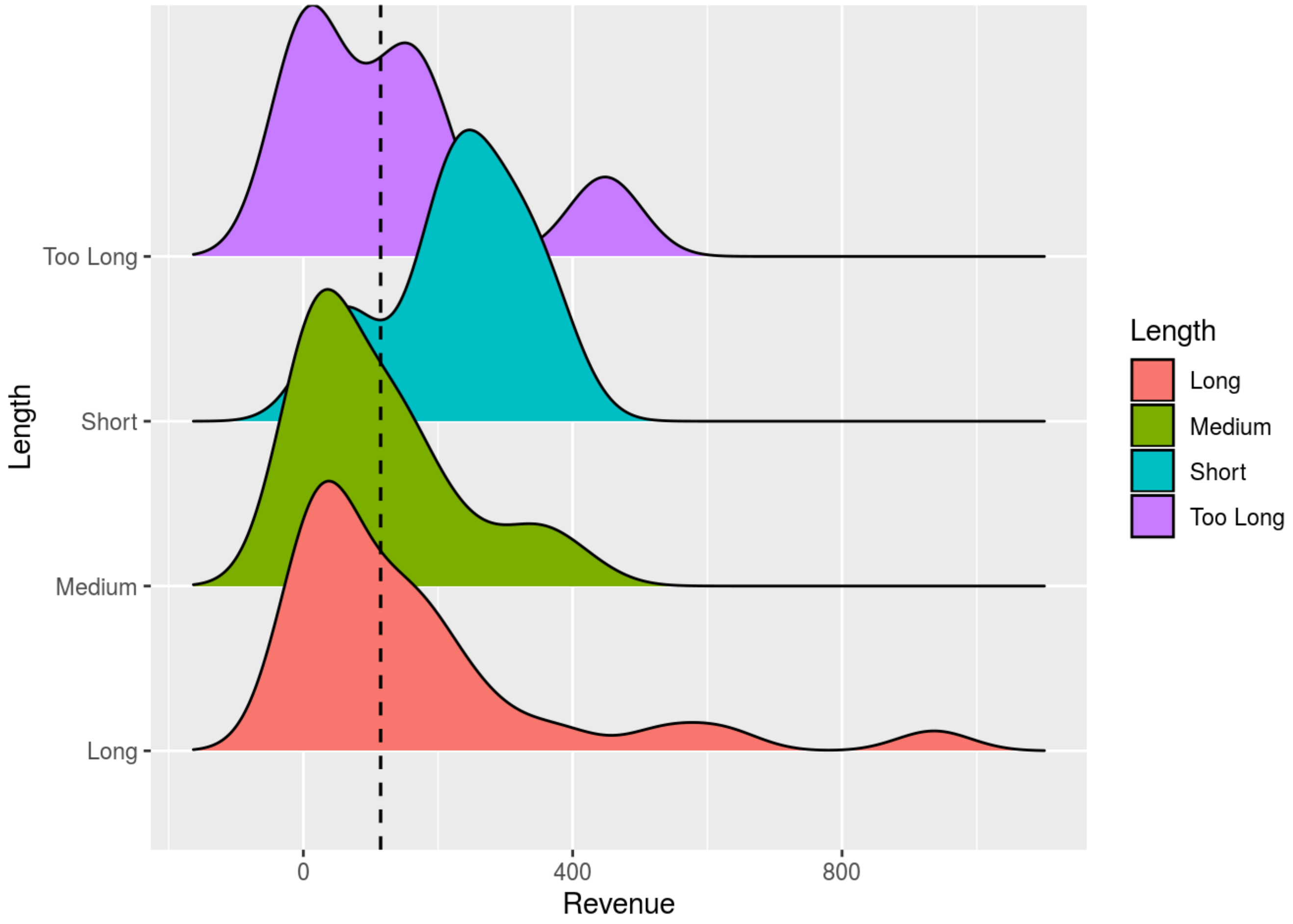
```
## Picking joint bandwidth of 27.1
```



###It appears that best rate movies are very unevenly distributed interms of revenue generation. To understand better, we need to separate the length and grade of the best reated movies to compare with the revenue.

```
#Filtering best movies from the movies_new data frame and assinging to Bestmovies data f
rame.
Bestmovies<-filter(movies_new, movies_new$Grade=='Best')

#Seperating the length of the best movies into four categories. 'Too long' for movies lo
nger than 160min, 'Long' for movies between 130 to 160 min, 'Medium' for movies between
 100 and 130, and 'Short' for the rest.
Bestlength<-mutate(Bestmovies,
               Length = ifelse(Runtime>=160, "Too Long",
                               ifelse(Runtime>=130&Runtime<160, "Long",
                                      ifelse(Runtime>=100&Runtime<130,"Medium","Short"
))))

#Graphing the density distribution.

density1<-ggplot(Bestlength,aes(x = Revenue, y=Length))+
  #geom_density(aes(color = Grade))
  geom_density_ridges(aes(fill = Length))+
  geom_vline(aes(xintercept = mean(movies_new$Runtime)),  linetype = "dashed", size = 0.
6)


density1
```

```
## Picking joint bandwidth of 54.6
```



##Short best rated movies generat more revenue than too long. This explains why the best movies with long runtime generate lesser revenue than best movies with short runtime.

##There are several limitation to this model.

##1. This model excludes the genre, votings, actors, and release year of the movies. Thus, the conclusion might not be 100% accurate.
##2. This model does not explain any causation.