Assignment 1, Question 2(b)

The IMDB data was taken from Kaggle website. The data was recorded for 1000 movies for a span of 10 years from 2006-2016. The data had attributes named Rank, Title, Genre, Description, Director, Actors, Years, Runtime, Ratings, Votes, Revenue and Metascore.

First of all, to perform data analysis on this data I performed a summary operation on each of the columns present in the dataset. Then I found that there were some NA values in the data in Revenue and Metascore column. I wanted to replace the NA values but first I worked to find the relation of Revenue and Metascore with other variables. I found that it is not a linear relation but a complex function of other parameters. Hence, finally I decided to ignore NA values as substituting them with other values would not give a correct data analysis as approx. 192 values were NA which accounted to a major portion of data.

Then, I decided to make categorization of different variables based on their quantiles. There categories were for variables- Votes, Rating, Revenue and Metascore.

Thereafter, I used the conditional plotting to find the relation of each variable with other variables. The main use of conditional plotting was that it considered the set points in that range only which helped to study the relationship in detail. Moreover, it also reduced the number of set points that were analyzed in a singular plot.

The following are the inferences which I derived from the plots-

# It can be seen that for Low\_votes a majority of movies lie in the rating range of 5-8. While on the other hand there is a large variation in rank of movies of approximately same rating. Perhaps we will have to explore the plot on more variables to find the reason for variation.

# The plot is more compressed for medium\_votes range as ratings on Y-axis vary from 6-8 for a majority of movies. While on the other hand the disparity of rank remains the same.

# Like the plot for Medium\_Votes, it can be seen that a majority of movies lie in the rating range of 6-8. While on the other hand there is a large variation in rank of movies of approximately same rating. Perhaps we will have to explore the plot on more variables to find the reason for variation.

# The plot is compressed for highest\_votes range as ratings on Y-axis vary from 6.5-8 for a majority of movies. Also it can be seen that there is a huge concentration of movies from rank 100-450.

# It can be seen that for low\_revenue category, the ratings vary for 5.5-8 for a majority of movies. But there is a vast divergence in terms of rank. So there is a great uncertainity to predict rank.

# For Medium\_Revenue category, the rating vary from 6-8 from a majority of movies while the rank varies from 10-1000. So ratings can be predicted pretty accurately as compared to rank.

# For High\_Revenue category, it can be seen that ratings vary from 6-8 and also there is a huge concentration of points from 0-400 rank. Hence we can conclude that movies which generate a high revenue generally have a rating above 6 and rank below 400 with some exceptions.

# It can be seen that for lowest\_metascore category, ratings usually vary in the range of 5-7, while the rank is usually above 350. Hence we can conclude that metascore depends on rank as well as ratings for lowest\_metascore categories.

# It can be seen that for low\_metascore categories ratings usually vary in the range of 6-8, while the rank is varying from 10-1000.

# It can be seen that for medium\_metascore categories ratings usually vary in the range of 6.5-8, while the rank has a large variance.

# It can be seen that for high\_metascore categories ratings usually vary in the range of 7-8, while rank has a large variance.

# It can be concluded that metascore has a direct relation with the ratings.

# From rank vs runtime graph, it can be concluded that a majority of movies have a runtime between 90-130 minutes.

# It can be concluded that log(IMDB\_data$votes) varies between 10-14 for a majority of movies that is there are movies of every rank category with a certain range in votes, so the number of votes cannot be a governing factor for rank.

# It can be seen that in low\_rating category, majority of movies generate a revenue below 20 million dollars. Moreover their metascore is also below 55. It can be concluded that revenue generation and metascore of a movie may have a direct relation.

# It can be seen that for high\_ratings category, movies have a metascore varying from 30 to 90 but a majority of movies have a generated revenue below 150 million dollars. So a high/low metascore movie in high rating category cannot necessarily impy that it would be a high revenue generating movie.

# It can be seen that for a low\_metascore category, ratings are usually in between 5.5-7.5, while revenue is below 100 million dollars. It can be hence concluded that revenue generation can definitely affect metascore for a movie.

# It can be said that as the ratings increase the metascore also increases for a majority of low\_revenue category movies.

# It can be seen that for as runtime increases, ratings also increase in a certain range for a majority of movies.

# It can be seen that qqplot for Ratings vs Metascore shows that set points have a normal distribution.