****

**DATA DRIVEN AND DECISION MAKING**

**\*\*\*ASSIGNMENT # 1\*\*\***

**Submitted By:**

**+ Minahil Mubasher 21L-7063**

**+ Bilal Nadeem 21L-7062**

**+ Aiza Aftab 21L-7064**

**+ Maha Ehtesham 21L-7061**

**+ Khadija Alam 21L-7039**

**Submitted To:**

**Dr. Komal Nadeem**

R Code file is attached for the reference **.**

**Question # 1:**

**What does Pre-processing steps are needed to prepare to run linear regression to predict movie gross and measure the regression performance?**

Firstly, we need to identify the movies data that need to be pre-processed for running the regression measures.

***-*** The Score (field/variable) have an outlier that will disturb the regression analysis so those outliers must be removed. Due to those outliers in Score variable, the distribution becomes left skewed.

***-*** The runtime field/variable also have an outlier in data-set that need to removed from data for further regression analysis. So they can’t make distribution right skewed.

***-*** The Budget variable also have outliers in data that also need to be removed from movies data-set.Because these outliers make distribution right-skewed.

***-*** The Response variable (gross) also have outliers and it’s distribution is rightly skewed.

***-*** Rating variable is a categorical variable that is needed to create the dummies of rating variable for regression.

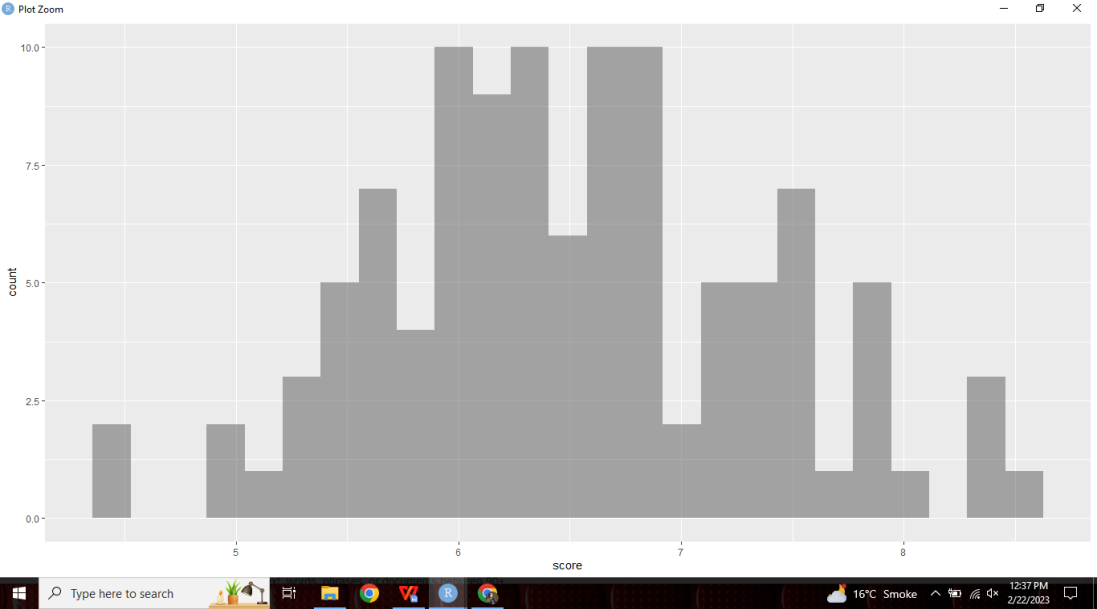
So, the Movies data-set have neither any missing values nor any duplication in data, they only have **outliers** in data. To measure the regression performance, we need to remove those outliers and “Name” variable/field from Movies data.

**Question # 2:**

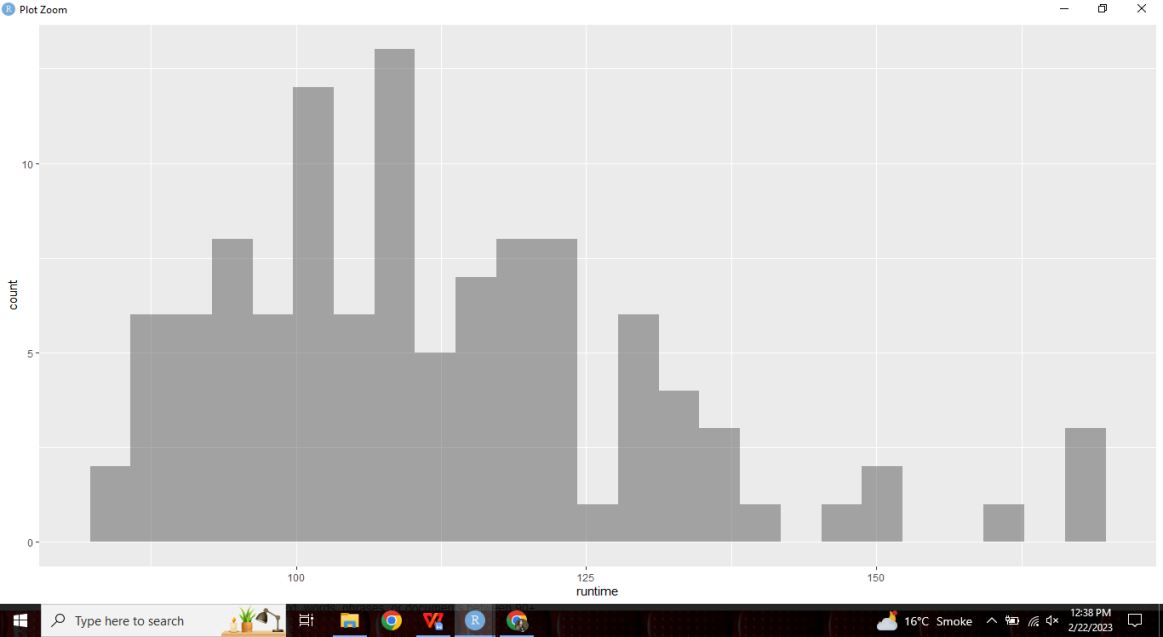
**Carry out the required data pre-processing steps and run a regression that uses the movies IMDB score (“score”) and its budget to predict the movies gross.What is the root mean squared error for this model in the training data?**

As in the previous answer we concluded that this data have only outliers in the numeric variable, so we changed the outliers value with the nearest data point because the data-set is already small so for the regression analysis we can’t remove those outliers from data.

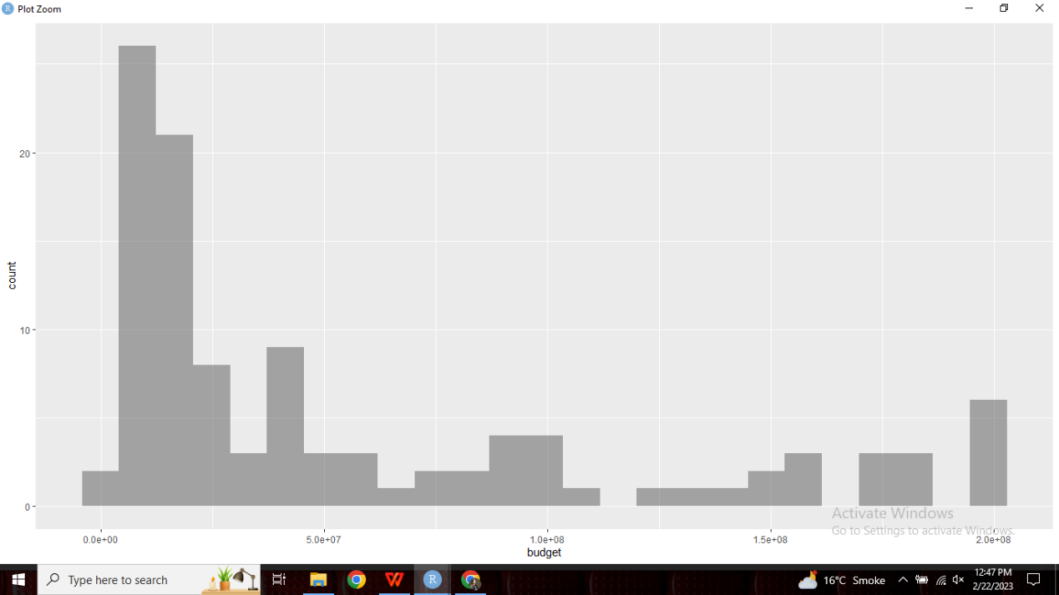
First, we changed the **score** variable outlier point to the most nearest point , after that the data become normal and the distribution become normal.



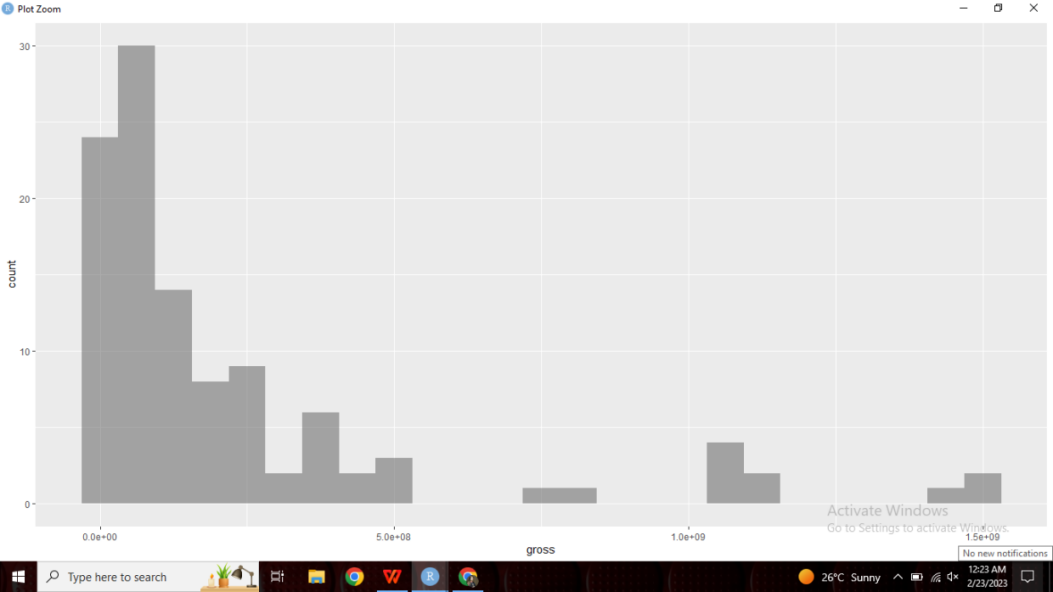
Then,we changed the outlier point of **Runtime** variable with the most nearest point and the distribution become right skewed but have no outlier remaining .



The budget variable also have few outliers point so those points are replaced with the most nearest point and the distribution became right skewed because every person have different budget for spending on movies.



The gross variable had outliers and all those outliers were replaced with the nearest point.But the distribution remained same and skewed in nature, but variable have no serious outlier or problem, so for the regression analysis we moved with this.



After replacing outliers points, we removed Name variable from movies data, created dummy variable of Rating variable. Now, the Pre\_Processing of this movie data is almost done. As in this movies data the variables are on different scale so normalizing this movies data required to apply normalization, but on this step we didn't perform normalization.

**Pre\_processing Steps:**

* Changed outliers points
* Removed Name variable from data.
* Produced Dummies of Rating variable

**Regression** to predict the Movies gross with the IMDB Score and Budget. Firstly, we splitted the movies data (partition) into train and test that the 70% is training data and 30% is testing data.

So, after concluding the result of regression to predict the Movies Gross with the IMDB score and budget, the **RMSE** became **196629658** (1.9e+08)

**Question # 3:**

**Now run a Stepwise regression to predict movies gross using the other variables in the data with the same partition.Which independent variables are included in this model? Why are they included?**

With the Stepwise regression AIC became 2985.11 on step 3. The Stepwise regression select the same variable as in linear regression model (regress in the above part)

Stepwise Regression Model selects the **Budget** and **Score** variable from all independent variables.

Because those variables are statistically significant (it’s p-value is less than 5% ) and appropriate model to predict the Movies gross.

**Question # 4:**

**According to the stepwise regression, which factors appear to increase the movies gross? Which appear decrease them?**

Both factor (Budget and Score) will increase the movies gross because both coefficient are positively effect the gross of movies. If the movies average audience score will increase then Gross movies will also increase (direct relation) and if budget will increase then gross movies also increase. So the result conclude that the budget and score will increase the gross movies.

**Question # 5:**

**What gross reveune does the stepwise model predict for a movie with a budget of $100,000,000 , an IMDB score of 10 , a runtime of 100 minutes , and PG-13 rating.**

Gross= (-3.998e+08) + 2.830e+00 Budget + 6.742e+07 Score.

**Gross = 557400000**

If budget is 100000000 and the score is 10 then Gross Movies will become 5.5e+08

**Question # 6:**

**What is the root mean squared error for the stepwise regression model in the training data?**

Stepwise regression RMSE for training data is **196629658** (1.9e+08)

**Question # 7:**

**Which of these two models has a lower root mean squared error in the training data? Why is this?**

Both RMSE are same because both models are same (in factors) to predict the gross movies. In the linear regression model , we regress the gross on the budget and score. And in the stepwise regression the result also conclude the both variable because these are significant. So in the training data RMSE are same for both models.

**Question # 8:**

**Find the root mean squared error of the stepwise regression model in the test data. Is it larger or smaller than the root mean squared error in the train data larger? Why?**

The RMSE of stepwise regression with the test data is **199077423**. That is greater than the RMSE of the training data. So all these conclusion conclude that the regression model (or stepwise model ; score & budget) with the training data is better than to have model with testing data. Because lower will be RMSE, model will be better.

**Question # 9:**

**About how large should we expect the average error to be in any predictions made with our stepwise regression model?**

In the Stepwise regression model, the model explains 43.4% of the total variation in the gross (dependent variable), which means that 56.6% of the total variation is still unexplained and can be considered the average error of the model.