A MINI PROJECT REPORT

on

New York State Gender Ratio Prediction

Submitted By-

Name: Shivaansh Agarwal

Roll No: 161500514 (24)

To-

Mr. Ankit Dwivedi

Department of Computer Engineering & Applications

Institute of Engineering & Technology

GLA University

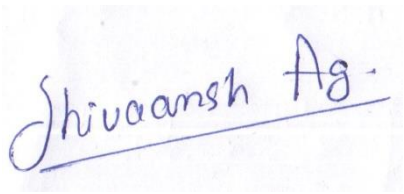
Mathura- 281406, INDIA

December, 2018

Department of Computer Engineering and Applications
GLA University, Mathura
17 km. Stone NH#2, Mathura-Delhi Road, P.O. – Chaumuha,
Mathura – 281406

Declaration

*We hereby declare that the work which is being presented in the Mini Project "**Title: New York State Gender Ratio Prediction using Multiple Regression",** in partial fulfilment of the requirements for Mini-Project LAB, is an authentic record of our own work carried under the supervision of **Ankit Dwivedi Sir, Technical Trainer, GLA University, Mathura**.*



Shivaansh Agarwal

Department of Computer Engineering and Applications
GLA University, Mathura
17 km. Stone NH#2, Mathura-Delhi Road, P.O. – Chaumuha,
Mathura – 281406

CERTIFICATE

*This is to certify that the project entitled **"New York State Gender Ratio Prediction"** carried out in Mini Project – I Lab is a bonafide work done by **Shivaansh Agarwal (161500514)** and is submitted in partial fulfilment of the requirements for the award of the degree Bachelor of Technology (Computer Science & Engineering).*

Signature of Supervisor:

Name of Supervisor:

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Mini Project undertaken during B. Tech. Third Year. This project in itself is an acknowledgement to the inspiration, drive and technical assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and guidance that we have received.

Our heartiest thanks to ***Dr. (Prof). Anand Singh Jalal,*** *Head of Dept., Department of CEA for providing us with an encouraging platform to develop this project, which thus helped us in shaping our abilities towards a constructive goal.*

We owe special debt of gratitude to ***Mr. Ankit Dwivedi,*** *Technical Trainer, Department of CEA, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. He has showered us with all his extensively experienced ideas and insightful comments at virtually all stages of the project & has also taught us about the latest industry-oriented technologies.*

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind guidance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Shivaansh Agarwal

ABSTRACT

The project involved working on the 'New York State Demographics' dataset. The data in the dataset was of the 1990 census conducted in USA and contained information about 790 different places (villages, towns, Census Designated Place(CDP)).

The information included Name of the Place, Total Population of that place, Percentage of people below the age of 18, Percentage of people between 18 to 65 years of age, Percentage of people above 65 years of age, and the Male to Female Ratio, i.e. The number of males per 100 females in that region.

The problem was to build a Machine Learning Model using the available Data to predict the Male to Female ratio which can be used by different Governmental as well as Private Firms for various purposes like making policies or business strategies for that particular region.

The Machine Learning Algorithm used for this purpose was Multiple Regression, and the tool used to complete the task was R.

The model used for the completion of the product was CRISP-DM (Cross Industry Standard Process for Data Mining) which consisted of 6 phases which are: Business Understanding Phase, Data Understanding Phase, Data pre-processing Phase, Modelling Phase, Evaluation Phase, Deployment Phase.

All the phases required considerable amount of time, among which the maximum time was required in the Data Pre-processing phase as the Dataset had many outliers which had to be dealt with in a proper way in order to proceed forward.

Overall working on this project was a great learning experience as a lot of challenges had to faced, for which to overcome a lot of research and understanding of the topic was required, and we're successfully able to complete the project within the allotted time duration.

Table of Contents

CHAPTER-1

Introduction

1.1 Overview and Motivation

Before the use of technologies like Machine Learning and Data Mining & the use of statistical tools like R, different government as well as the business organisations in order to make policies or any business strategies for a specific region or a particular group of people relied primarily on intuition or by doing some sort of surveys. But using this method for making policies and business decisions was not that effective and the success ratio was also quite low.

But as technology grew and more and more data got generated and stored a need was felt to use that data for gaining actionable insights and patterns and then using that information to make predictions and estimates about future outcomes.

So nowadays when a business organisation wants to open a new store in a region or bring out a new discount sale in order to attract more customers, or when a Government Body wants to do some progress in a region they first decide to get some background knowledge of that region before doing or rolling out new policies or services with the help of available data and by using Machine Learning and Data Mining Processes.

The use of Demographics Data in such cases becomes very important in order to gain information about the behaviour, interests of the people of a particular region for these Government bodies and Business Firms.

1.2 Objective

This project involved working on the 'New York State Demographics' dataset. The data in the dataset was of the 1990 census conducted in USA and contained information about 790 different places (villages, towns, Census Designated Place(CDP)).

The information included Name of the Place, Total Population of that place, Percentage of people below the age of 18, Percentage of people between 18 to 65 years of age, Percentage of people above 65 years of age, and the Male to Female Ratio, i.e. The number of females per 100 males in that region.

The problem was to build a Machine Learning Model using the available Data to predict the Male to Female ratio which can be used by different Governmental as well as Private Firms for various purposes like making policies or business strategies for that particular region.

CHAPTER-2

Project Design

The various steps involved in the completion of the project are:

1. Business Understanding: In this the business problem and the motivation of that problem had to be understood in order to act accordingly further in the project.
   This business problem was provided. Then further research on the particular domain was done using different resources over the internet.

2. Data Understanding: In this step the roles and meanings of different variables present in the Dataset had to be understood. During the step univariate analysis and multivariate analysis is done in order to gain insights about different variables and then decide which variable is important and how much data cleaning has to be performed on a particular variable.

3. Data Preparation: In this step data cleaning is performed in order to prepare data for further analysis and for building the machine learning model.

4. Modeling: In this step the appropriate machine learning model is selected, then dataset is divided into the training and test sets, and then finally the model is trained on the training data in order to get the best results.

5. Evaluation: In this step the model is evaluated against the test set using various performance metrics like looking at R-squared, Adjusted R-squared, Residual Plots, etc.

6. Deployment: The model is then finalized after getting satisfied results in the previous step and finally delivered to the client for their use.

CHAPTER-3

Project Implementation

3.1 Understanding The Dataset

	PLACE	TOT_POP	PCT_U18	PC_18_65	PCT_O65	MALE_FEM
1	Adams village	1753	25.3	56.2	18.5	79.0
2	Adams Center CDP	1675	30.4	58.7	10.9	87.0
3	Addison village	1842	28.8	56.0	15.2	84.0
4	Airmont CDP	7835	25.0	65.3	9.7	89.0
5	Akron village	2906	24.7	57.1	18.2	81.1
6	Albany city	101082	17.8	66.8	15.3	84.2
7	Albertson CDP	5166	18.9	61.5	19.6	91.8
8	Albion village	5863	26.1	56.8	17.1	85.2
9	Alden village	2457	25.0	60.6	14.4	88.3
10	Alexandria Bay village	1194	20.8	56.2	23.0	83.3
11	Alfred village	4559	3.2	94.0	2.9	136.8
12	Allegany village	1980	20.8	62.7	16.5	89.1
13	Altamont village	1519	28.1	62.1	9.8	89.9
14	Altona CDP	1003	9.9	84.6	5.5	527.8
15	Amenia CDP	1057	24.8	60.0	15.2	86.2
16	Amityville village	9286	20.3	61.8	17.9	87.3
17	Amsterdam city	20714	22.4	54.1	23.5	79.9
18	Andover village	1125	27.8	56.4	15.8	88.4
19	Angola village	2231	27.8	58.9	13.3	86.3

Figure 1: Snippet of the Dataset

The above figure depicts a snippet of the New York State Demographics Dataset which was created by taking data from the 1990 US census.

The Dataset consists of 6 columns of which 5 will be used as the predictor (independent) variables which are PLACE, TOT_POP, PCT_U18, PC_18_65, PCT_O65 and 1 will be used as the target (dependent) variable which is MALE_FEM.

The description of the columns is as follows:

1. PLACE: This column contains the name of towns/villages/cities in New York State.
2. TOT_POP: This column contains number of people living in that particular region.
3. PCT_U18: This column contains the percentage of people less than 18 years of age.
4. PC_18_65: This column contains the percentage of people between 18 to 65 years of age.
5. PCT_O65: This column contains the percentage of people older than 65 years of age.

6. MALE_FEM: This column contains the ratio of the number of males for every 100 females.

3.2 Exploratory Data Analysis

Before building the model, in order to better understand the flaws or importance of different variables Exploratory Data Analysis is performed. First an overview of the dataset is taken, to get knowledge about if there are any missing values in the dataset.

The following figure confirms that there are no missing values in any of the variables.

```
PLACE   TOT_POP   PCT_U18 PC_18_65   PCT_O65 MALE_FEM
    0         0         0        0         0        0
```

The following figure shows that there are no empty strings in our dataset.

```
PLACE   TOT_POP   PCT_U18 PC_18_65   PCT_O65 MALE_FEM
    0         0         0        0         0        0
```

Hence we can say that all entries in the dataset are present.

Now let's check whether the dataset contains any outliers or any other irregularities.

1. **PLACE:**

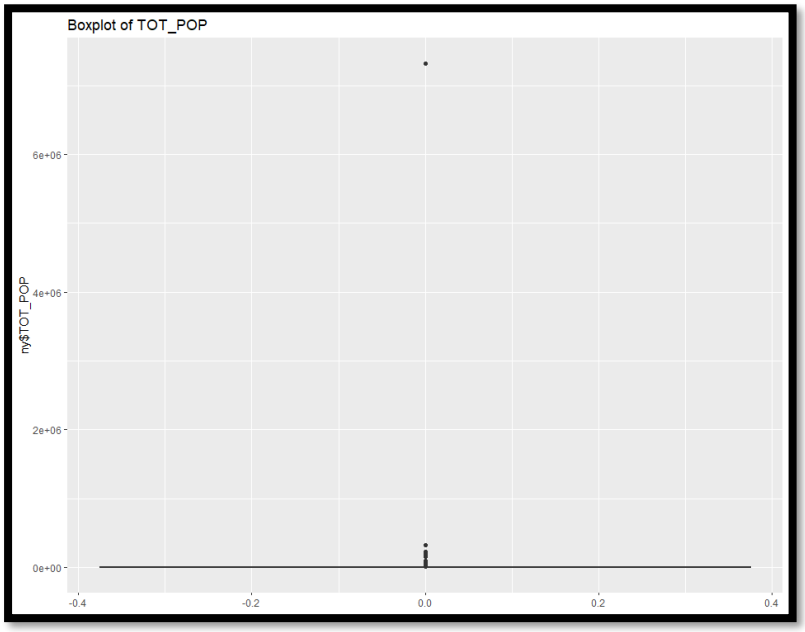
   Since every value in this column is unique as well as categorical therefore this variable is not useful for building the model, hence this column will be dropped.

2. **TOT_POP:**

   After looking at the summary of this variable, it can be seen that this column might contain some outliers as the maximum value is quite high, moreover the difference between the mean and median is also very high.

```
    TOT_POP
Min.   :   1000
1st Qu.:   1904
Median :   4013
Mean   :  18305
3rd Qu.:   9057
Max.   :7322564
```

The presence of outliers can be confirmed by making the boxplot of this variable.



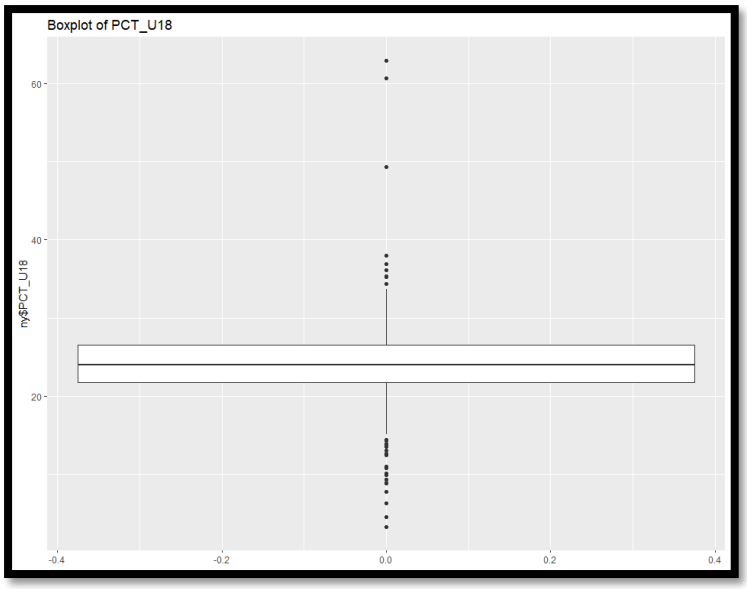
As it can be seen that some values are very large as compared to others, so while cleaning outliers have to be removed from this variable.

3. **PCT_U18:**

After looking at the summary of this variable following results were obtained:

```
    PCT_U18
Min.   : 3.20
1st Qu.:21.70
Median :24.00
Mean   :23.96
3rd Qu.:26.50
Max.   :63.00
```

Here since the Mean and the Median are quite close therefore it might have less outliers as compared to TOT_POP variable. The boxplot of this variable is as below:



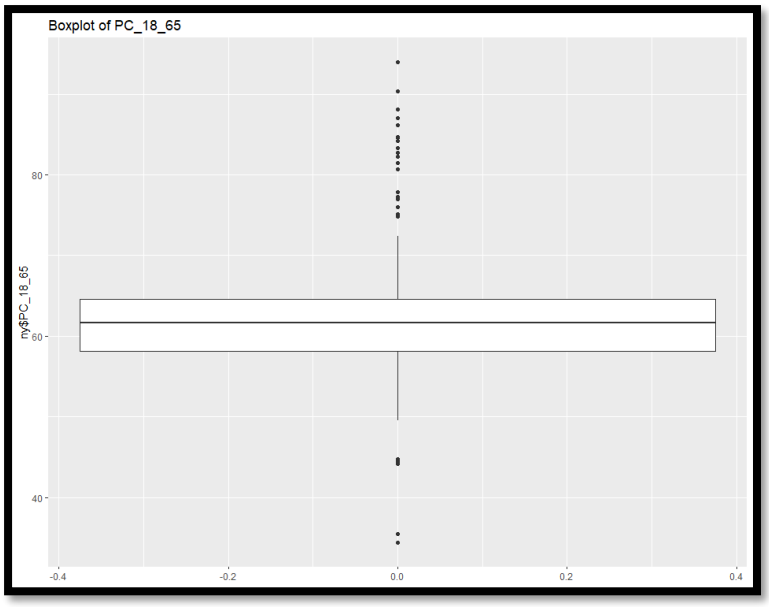
Looking at the above boxplot it can be said that the this variable also contains some outliers which are dealt with during Data Cleaning.

4. **PC_18_65:**

After looking at the summary of this variable following results were obtained:

```
    PC_18_65
Min.   :34.40
1st Qu.:58.12
Median :61.60
Mean   :61.67
3rd Qu.:64.60
Max.   :94.00
```

Looking at the above summary statistics, since mean and median are pretty close to each other it can be inferred that this column might contain less number of extreme values and less or no outliers.



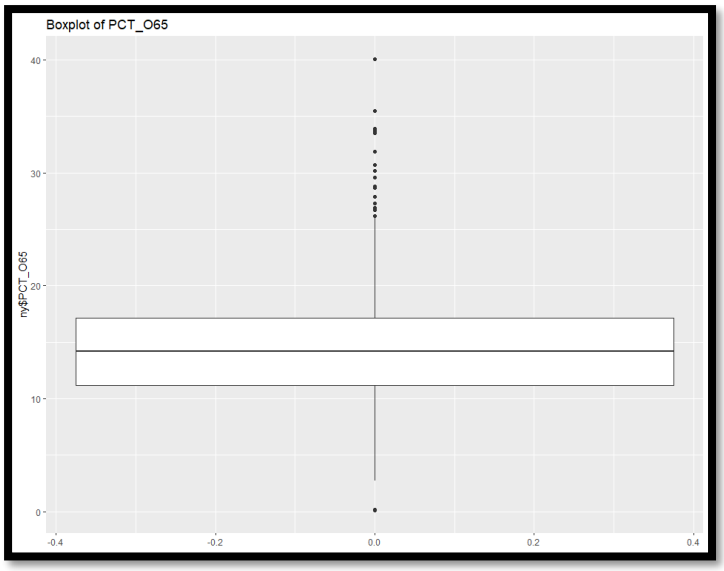
It can be seen from the above boxplot that some outliers are also present in this variable which have to removed before model building.

5. **PCT_O65:**

This is also one of the predictor variables which might also have to be used for model building. Having a look at the summary statistics of this variable following results were obtained.

```
    PCT_O65
Min.   : 0.10
1st Qu.:11.20
Median :14.20
Mean   :14.37
3rd Qu.:17.18
Max.   :40.10
```

By looking at the summary statistics not much information about the outliers can be gained. The boxplot of the same variable is shown below:



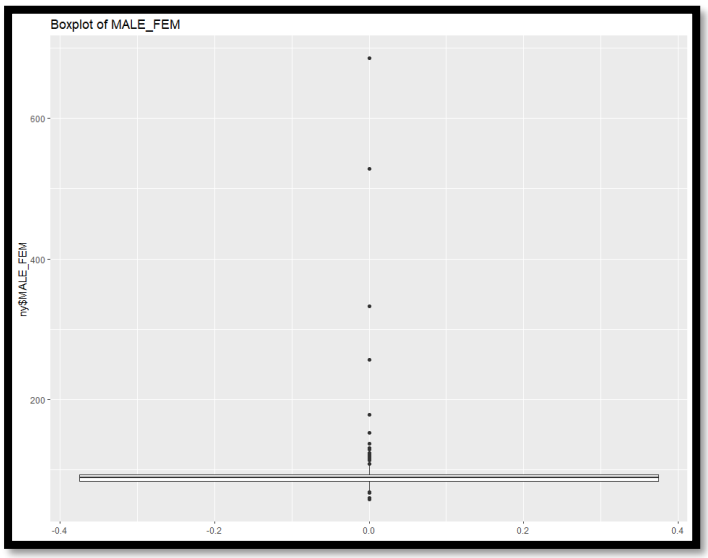
In this variable also it can be seen that there are some outliers present.

6. **MALE_FEM:**

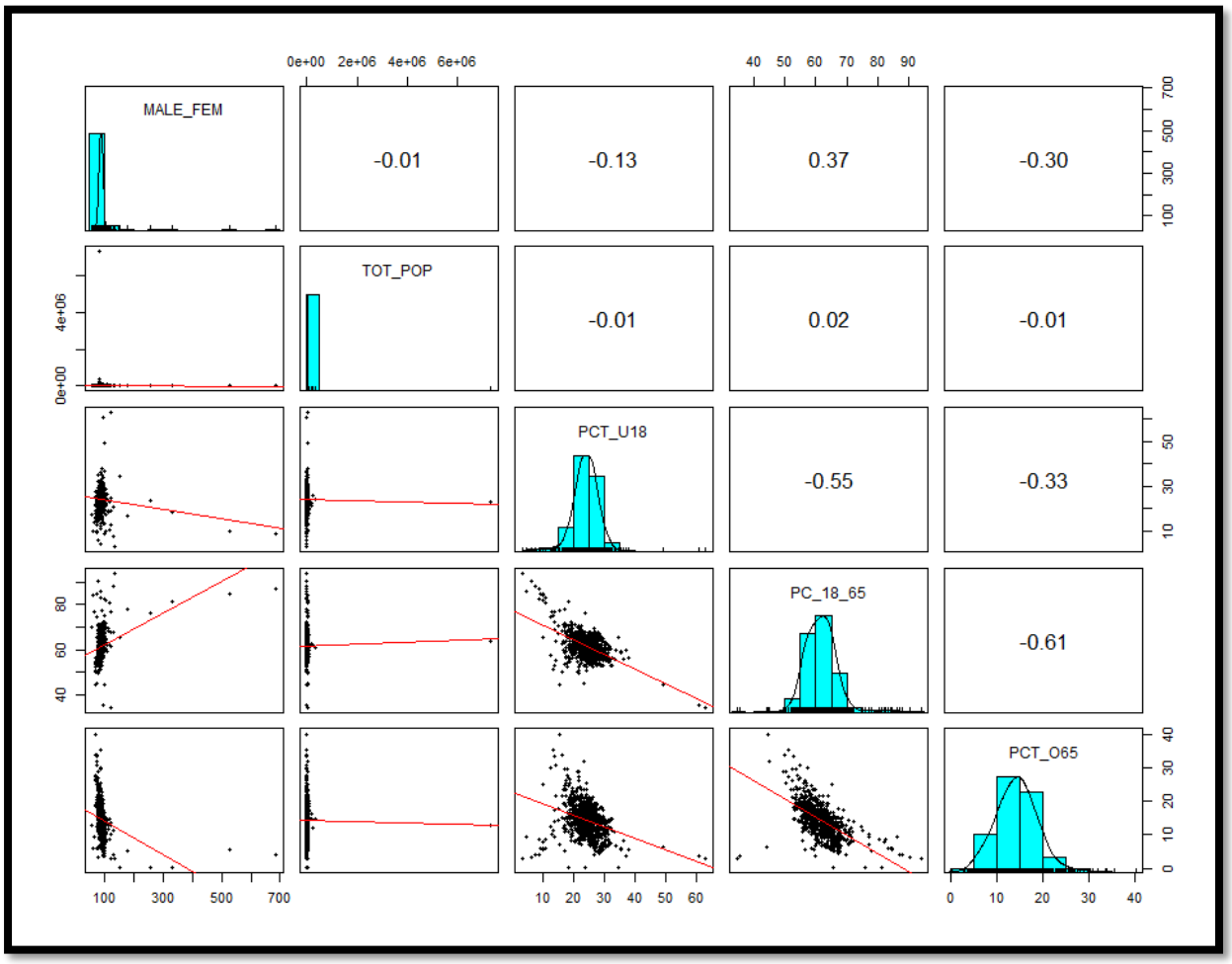
This is the dependent variable that has to be predicted. Looking at the summary statistics of this variable:

```
   MALE_FEM
Min.   : 58.00
1st Qu.: 83.80
Median : 88.50
Mean   : 90.75
3rd Qu.: 93.30
Max.   :686.00
```

It can be clearly inferred by looking at the maximum value that there might be some outliers in this dataset. Having a look at the boxplot of the same variable to have a better understanding of the outliers in this variable:



Before performing any Data Cleaning operations the following results can be obtained.



In the above figure:

- The left diagonal shows the distribution of the different variables in the dataset.
- The cells below the diagonal shows the scatterplots of the different variables with respect to each other.
- The cells above the diagonal shows the correlation of one variable with respect to another variable.

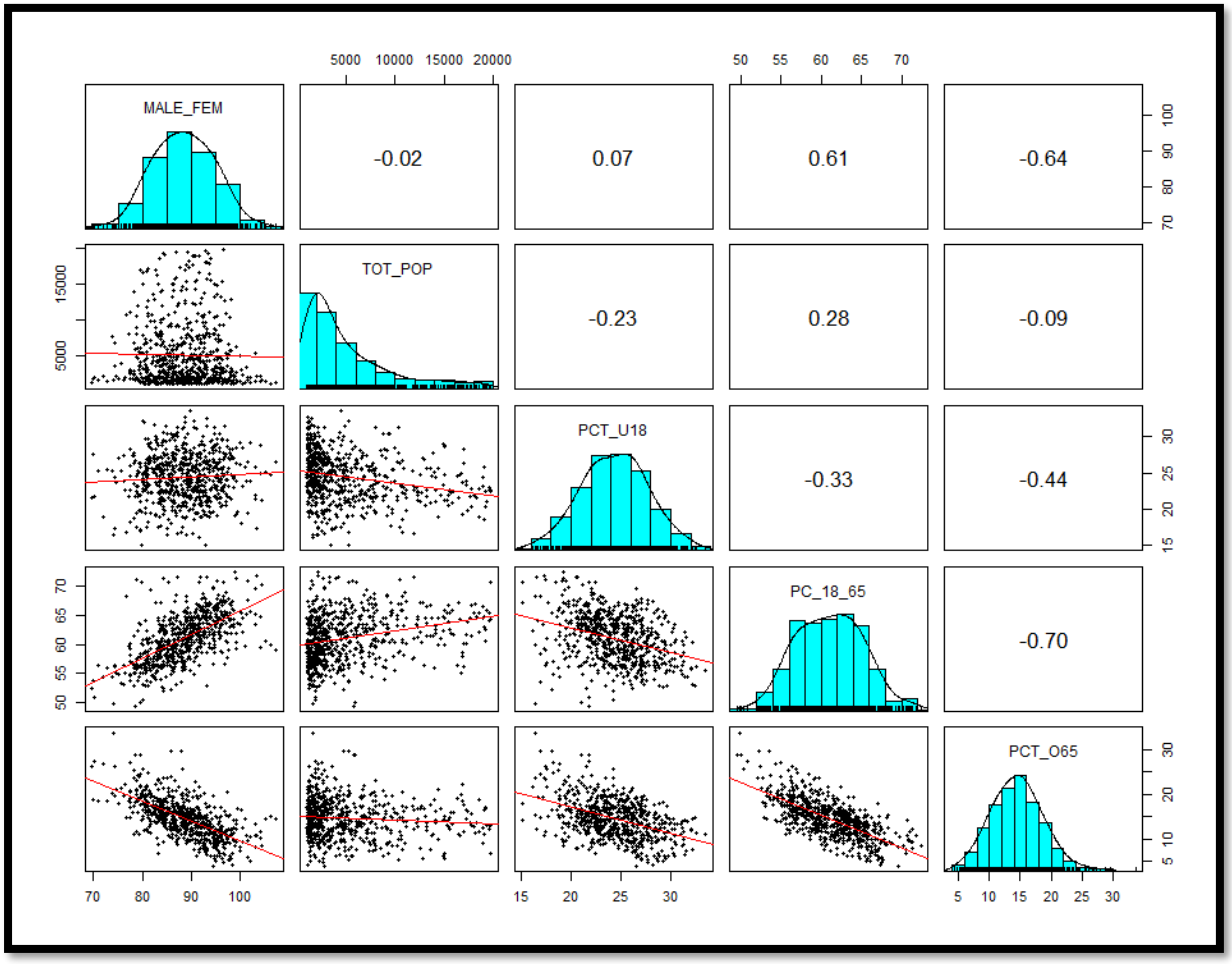
Now since it can be seen that the correlation between some variables is quite high, i.e. it can be inferred that there might be some multicollinearity present in the dataset which is not good for the model.

3.3 Data Cleaning

Since the dataset didn't contain any missing values only outliers have to be dealt with because of the following reasons:

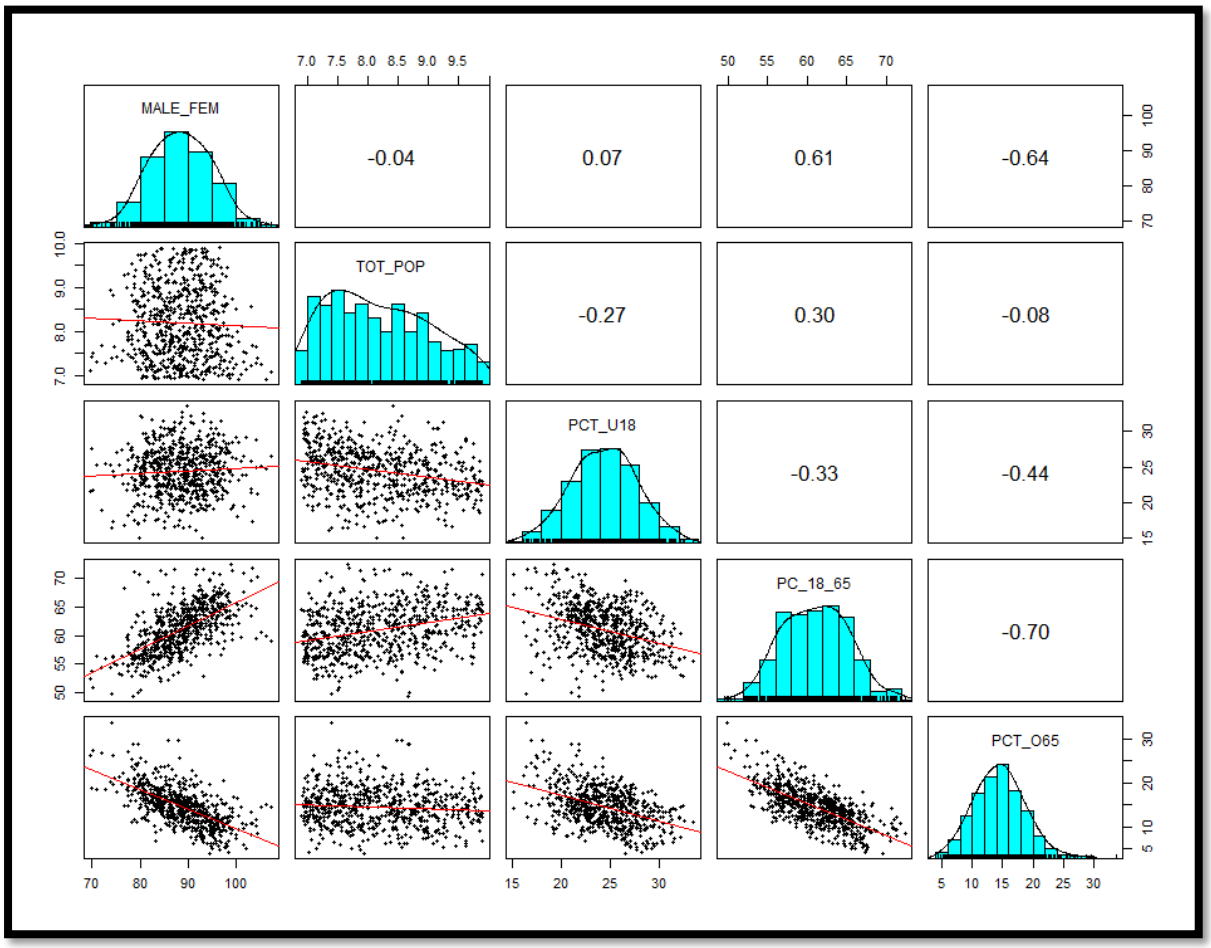
- Outliers are data points that deviate significantly from the normal objects as if it were generated by different mechanism.
- Effect of outliers on model is quite ambiguous, they might be important for model so it is important to make a model with them and another model without them.
- Outliers can be detected using threshold method in a Box and Whisker Plot. It is a standardized way of displaying the distribution of data based on the five-number summary: minimum, first quartile, median, third quartile, and maximum.
- The values which are greater than the value of 1.5*(Q3-Q1)+Q3 and values that are less than 1.5*(Q3-Q1)-Q1 are considered as outliers using this method.

After removing the outliers using the above specified Threshold method following combination of Scatterplots, Overlay Histograms and Correlation Matrix can be obtained:



From the above figure it can be seen that except TOT_POP all other variables have become almost normally distributed.

Now in order to make TOT_POP normally distributed we've to take log of that variable. So after performing that following result is obtained.



Here it can be seen that the TOT_POP variable has become almost normally distributed.

Now we've to figure out which Machine Learning is to be used in order to predict the value of MALE_FEM.

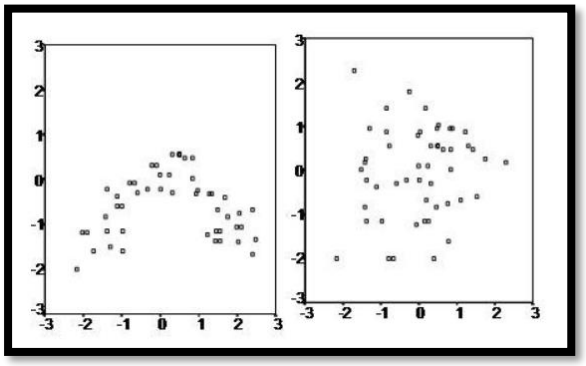
3.4 Why Multiple Regression?

In the dataset it can be seen that the target variable MALE_FEM is a continuous variable so a Regression Algorithm can be applied here and moreover since there are more than one predictor variables i.e. TOT_POP, PCT_U18, PC_18_65, PCT_O65, so Multiple Linear Regression will be applied.

Now before applying the Multiple Linear Regression it's assumptions have to be verified.

3.5 Assumptions of Multiple Regression

1. Linear Relationship between the Independent and the Dependent Variables:

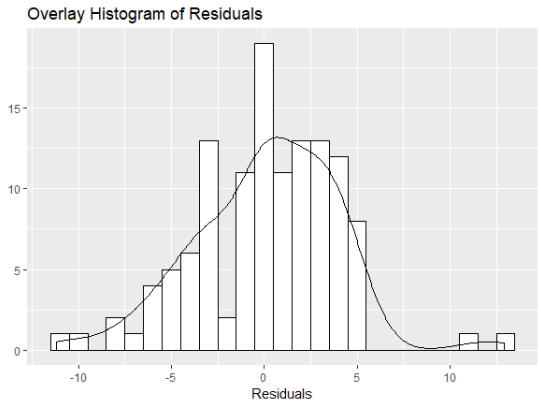


The above figure is used to depict in what case we can consider that the relationship between the Dependent and the Independent variables to be Linear. So if the scatterplot between is like the one shown on the right it can be said that there is a linear relationship between the Dependent and the Independent Variables.

Hence the first assumption is valid for our dataset.

2. Multivariate Normality: It means that the residuals are normally distributed.

This can be verified after the model has been made. So after creating the model the following graph was obtained:



From this it can be seen that the Distribution is roughly a Normal Distribution which is enough to verify the second condition of Multivariate Normality.

3. No Multicollinearity:

Since it is known because of the correlation matrix above that there is multicollinearity in the dataset, it has to be removed for the model to work properly.

Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be checked multiple ways:

- Correlation matrix – When computing a matrix of Pearson's bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.
- Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem. Variance Inflation factor is a famous method to detect multicollinearity, a Variance inflation factor is a measure of the amount of multicollinearity in a set of multiple regression variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression. When significant multicollinearity issues exist, the variance inflation factor will be very large for the variables involved.

  Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

  A rule of thumb for interpreting the variance inflation factor:

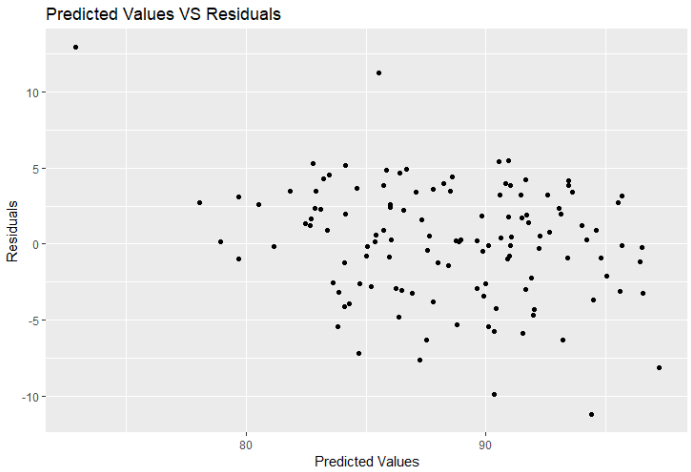
  - 1 = not correlated.
  - Between 1 and 5 = moderately correlated.
  - Greater than 5 = highly correlated.

If multicollinearity is found in the data, one possible solution is to centre the data. To centre the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing

multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.

So multicollinearity will be removed after a model is built with all variables and then using the Variation Inflation Factor we'll remove the undesired column.

4. Homoscedasticity: According to this when a scatterplot of residuals vs predicted values is made there should be no clear pattern in the distribution. So clearly there is no clear pattern in this scatterplot.



Hence it can be seen that the all assumptions have been satisfied except Multicollinearity (Assumption-3) which will be satisfied in the following section.

3.6 Model Building and Evaluation

After splitting the dataset into Training set and Test Set the multiple regression models are built by using Backward Selection in which the first model is made by including all the variables then gradually in subsequent models variables are excluded based on the level of significance and p values and then the model with the best accuracy is chosen for Deployment purposes.

In this project after building the 1st and looking at it's summary it can be seen that the TOT_POP is the most significant variable, so it cannot be removed, and moreover since the VIF is showing the maximum values for PC_18_65 and PCT_O65 one of the two have to be removed.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.4219     0.2017 438.405  < 2e-16 ***
TOT_POP      -1.4285     0.2141  -6.672 6.34e-11 ***
PCT_U18      -4.0689    13.4246  -0.303    0.762
PC_18_65     -2.3830    17.0234  -0.140    0.889
PCT_O65      -7.7648    17.7858  -0.437    0.663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure : Summary of model with all Variables included

```
  TOT_POP      PCT_U18     PC_18_65      PCT_O65
 1.124921 4422.120018 7110.809964 7761.939088
```

Figure : Variation Inflation Factor of Variables

So PCT_O65 is removed the model with the variables TOT_POP, PCT_U18, PC_18_65 is the best model and will be used for deployment.

The following figure shows the summary of the Final Model.

```
Call:
lm(formula = MALE_FEM ~ TOT_POP + PCT_U18 + PC_18_65, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0818  -2.7303   0.0127   2.7841  18.2764

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.4219     0.2015 438.737  < 2e-16 ***
TOT_POP      -1.4328     0.2137  -6.704 5.17e-11 ***
PCT_U18       1.7912     0.2177   8.226 1.48e-15 ***
PC_18_65      5.0484     0.2193  23.021  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.679 on 535 degrees of freedom
Multiple R-squared:  0.5016,    Adjusted R-squared:  0.4988
F-statistic: 179.5 on 3 and 535 DF,  p-value: < 2.2e-16
```

Figure: Summary of the Final Model

```
  TOT_POP   PCT_U18 PC_18_65
1.122564 1.165111 1.181760
```

Figure: Variation Inflation Factor of variables in Final Model

CHAPTER-4

Conclusions

The following conclusions can be drawn from the analysis on the dataset:

- After cleaning the data and then applying the Multiple Regression algorithm the model is made which can be helpful in predicting the Male to Female ratio of the town in New York state with 50% accuracy, given the Total Population of the region, Percentage of people below 18 years of age, percentage of people between 18 to 65 years of age, percentage of people above 65 years of age.
- The Residual Standard Error is 4.679 with Error Rate of 5.29%. The residual standard error (RSE) gives a measure of error of prediction, the lower the RSE the more accurate the model. The error rate can be calculated by dividing the RSE by mean(dependent variable).
- Higher accuracy is not the parameter of judging the goodness of many models like in this case, because of the lack of availability of more columns.

CHAPTER-5

Appendix

This chapter contains the code of the analysis our project.

Setting The Working Directory

```
setwd("D:/01. B.TECH 3rd YEAR/04. Mini Project/MINI")
```

Reding The Dataset

```
ny <- read.delim("New_York.dat")
```

EXPLORATORY DATA ANALYSIS

 # Count of NA values

```
sapply(ny, function(x) sum(is.na(x)))
```

 # Count of empty strings

```
sapply(ny, function(x) length(which(x==")))
```

 # Counting the Count of Number of Unique values in every Column

```
sapply(ny, function(x) length(unique(x)))
```

```
str(ny)
summary(ny)
```

 # Here we find that the median population is 4013 but Max population is coming very high i.e 7322563.

 # Therefore we can sense that there might be outliers present in this column.

 # Now we'll see which place has such a high value

```
new_york[which(new_york$TOT_POP==7322564),]
```

 # Hence we come to know that this population is of New York City

 # After Cross Checking on the Internet, in 2015 which was when the book in which this dataset is given was published,

 # the population of New York City was 8.2 million but since here we have the population as 7.32 million which

 # was the population of New York City in 1990. So maybe this dataset is of 1990.

```
# install.packages("ggplot2")

library(ggplot2)

ggplot(data = ny, aes(y=ny$TOT_POP)) + geom_boxplot() + ggtitle("Boxplot of
TOT_POP")

ggplot(data = ny, aes(y=ny$PCT_U18)) + geom_boxplot() + ggtitle("Boxplot of PCT_U18")

ggplot(data = ny, aes(y=ny$PC_18_65)) + geom_boxplot() + ggtitle("Boxplot of PC_18_65")

ggplot(data = ny, aes(y=ny$PCT_O65)) + geom_boxplot() + ggtitle("Boxplot of PCT_O65")

ggplot(data = ny, aes(y=ny$MALE_FEM)) + geom_boxplot() + ggtitle("Boxplot of
MALE_FEM")


# install.packages("psych")

library(psych)

pairs.panels(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")],meth
od = "pearson",lm=TRUE, ellipses = FALSE)


#### Data Cleaning ####

# TOT_POP

boxplot(ny$TOT_POP,main="TOT_POP")$stats[c(1,5),] #1000 #19750

length(which(ny$TOT_POP>19750)) #81

length(which(ny$TOT_POP<1000))  #0

ny <- ny[-which(ny$TOT_POP>19750),]

summary(ny$TOT_POP)


# PCT_U18

boxplot(ny$PCT_U18,main="PCT_U18")$stats[c(1,5),]

length(which(ny$PCT_U18>33.7)) #9

length(which(ny$PCT_U18<15.1)) #23

ny <- ny[-which(ny$PCT_U18>33.7),]

ny <- ny[-which(ny$PCT_U18<15.1),]


# PC_18_65
```

```
boxplot(ny$PC_18_65,main="PC_18_65")$stats[c(1,5),]
length(which(ny$PC_18_65>72.4)) #4
length(which(ny$PC_18_65<49.5)) #2
ny <- ny[-which(ny$PC_18_65>72.4),]
ny <- ny[-which(ny$PC_18_65<49.5),]
```

 # PCT_O65

```
boxplot(ny$PCT_O65,main="PCT_O65")$stats[c(1,5),]
length(which(ny$PCT_O65>26.2)) #10
length(which(ny$PCT_O65<2.7))
ny <- ny[-which(ny$PCT_O65>26.2)]
```

 # MALE_FEM

```
boxplot(ny$MALE_FEM,main="MALE_FEM")$stats[c(1,5),]
length(which(ny$MALE_FEM<69.7)) #3
length(which(ny$MALE_FEM>107.4)) #5
ny <- ny[-which(ny$MALE_FEM<69.7),]
ny <- ny[-which(ny$MALE_FEM>107.4),]
```

 # Scatterplot Matrix

```
pairs(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")])
# Correlation Matrix (Pearson)
cor(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")])

pairs.panels(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")],meth
od = "pearson",lm=TRUE, ellipses = FALSE)
#
pairs.panels(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")],meth
od = "spearman",lm=TRUE)
```

 # Making the TOT_POP normally distributed.

```
ny$TOT_POP <- log(ny$TOT_POP)

pairs.panels(ny[c("MALE_FEM","TOT_POP","PCT_U18","PC_18_65","PCT_O65")],meth
od = "pearson",lm=TRUE, ellipses = FALSE)


#### Splitting The Dataset ####

ny1 <- ny[,c(6,2,3,4,5)]


#install.packages('caTools')

library(caTools)

set.seed(123)

#Split of dataset into training dataset and test dataset

split <- sample.split(ny1$MALE_FEM,SplitRatio = 0.8)

training_set <- subset(ny1, split==TRUE)

test_set <- subset(ny1, split == FALSE)


#6. Feature Scaling

##(i) Standardisation

##(ii) Normalisation

#For putting variables in same scale

training_set[,2:5] <- scale(training_set[,2:5])

test_set[,2:5] <- scale(test_set[,2:5])


#### Model Building ####

# Using all variables

model1 <- lm(MALE_FEM ~ TOT_POP + PCT_U18 + PC_18_65 + PCT_O65,data =
training_set)

summary(model1)

y_pred1 = predict(model1,newdata=test_set)

cor(y_pred1,test_set$MALE_FEM)*cor(y_pred1,test_set$MALE_FEM)

# 48.08
```

```
# Eliminating PCT_O65
## FINAL MODEL
model2 <- lm(MALE_FEM ~ TOT_POP + PCT_U18 + PC_18_65,data = training_set)
summary(model2)
y_pred2 = predict(model2,newdata=test_set)
cor(y_pred2,test_set$MALE_FEM)*cor(y_pred2,test_set$MALE_FEM)
# 48.48

# Removing the PCT_U18 in model2
model22 <- lm(MALE_FEM ~ PCT_U18 + PC_18_65,data = training_set)
summary(model22)

# Calculating Variance Inflation Factor
library(caret)
varImp(model1)

# Error Rate
sigma(model2)/mean(training_set$MALE_FEM)*100

# Calculating the No of rows in training set
nrow(training_set)

# Eliminating PC_18_65
model3 <- lm(MALE_FEM ~ TOT_POP + PCT_U18 + PCT_O65,data = training_set)
summary(model3)
y_pred3 = predict(model3,newdata=test_set)
cor(y_pred3,test_set$MALE_FEM)*cor(y_pred3,test_set$MALE_FEM)
# 48.24

# Eliminating PC_18_65 and PCT_O65
```

```
model4 <- lm(MALE_FEM ~ TOT_POP + PCT_U18,data = training_set)
summary(model4)
y_pred4 = predict(model4,newdata=test_set)
cor(y_pred4,test_set$MALE_FEM)*cor(y_pred4,test_set$MALE_FEM)
# 0.1

#### Different Plots ####

# Histogram of residuals
resid = test_set$MALE_FEM-y_pred2

ggplot() + aes(resid)+ geom_histogram(binwidth=1, colour="black", fill="white") +
  geom_density(aes(y=1*..count..)) +
  ggtitle("Overlay Histogram of Residuals") +
  xlab("Residuals") +
  ylab("")

# Homoscedasticity
ggplot(data=NULL,aes(x=y_pred2,y=resid))+geom_point() +
  ggtitle("Predicted Values VS Residuals") +
  ylab("Residuals") +
  xlab("Predicted Values")
```