**285 San-Francisco Employee Data Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| First Name | Last Name | Email (hawk.iit.edu) | Student ID |
| Utkarsha | Vidhale | uvidhale@hawk.iit.edu | A20448993 |
| Vaidehi | Rathkanthiwar | [vrathkanthiwar@hawk.iit.edu](mailto:vrathkanthiwar@hawk.iit.edu) | A20450703 |
| Prashant | Raina | praina@hawk.iit.edu | A20444481 |

Table of Contents

[**1. Introduction** 2](#_Toc529694108)

[**2. Data** 2](#_Toc529694109)

[**3. Problems to be Solved** 3](#_Toc529694110)

[**4. Solutions** 3](#_Toc529694111)

[**5. Experiments and Results** 3](#_Toc529694112)

[5.1. Methods and Process 4](#_Toc529694113)

[5.2. Evaluations and Results 16](#_Toc529694114)

[5.3. Findings 40](#_Toc529694115)

[**6. Conclusions and Future Work** 41](#_Toc529694116)

[6.1. Conclusions 41](#_Toc529694117)

[6.2. Limitations 42](#_Toc529694118)

[6.3. Potential Improvements or Future Work 42](#_Toc529694119)

# **1. Introduction**

Our application is based on San Francisco Employee compensation data which describes the various features related to employee department, organization, job profile, salary and benefits.

In a corporate structure, employees are the integral part of the organization. No matter your company size, your people are your most important asset. They are the backbone of your business. So, one of the most important aspects of running your business is keeping your employees happy by offering them high-quality employee benefits and compensation.

Employee benefits refer to all non-wage compensation or bonus provided to employees in addition to their salaries. The type of benefits your company decides to offer will vary based on the organization and job profile.

But, sometimes many companies don’t realize how much time and money ineffective HR processes are costing them. Providing benefits to those type of job profile which have a very low productivity has often come into wrong consideration. So, there must be some solution in which company can know in advance about the compensation structure based on job profile and organization. This provided us the opportunity to develop model which can predict compensation and benefits based on different factors. Employers can use this model to imbibe some knowledge regarding the compensation factors and employees can use it to decide which job profiles are receiving maximum benefits

# **2. Data**

Briefly introduce your data sets, such as which application or domain the data belongs to, where did you collect it, how large it is, how many features there are, and so forth.

* The dataset hosted by the city of San Francisco. The organization has an open data platform and they update their information according the amount of data that is brought in. The San Francisco Controller's Office maintains a database of the salary and benefits paid to City employees since fiscal year 2013.
* This dataset is updated annually. New data is added on a bi-annual basis when available for each fiscal and calendar year. It has been collected from kaggle.com (<https://www.kaggle.com/san-francisco/sf-employee-compensation>) and is available in csv format (170 MB). There are 8,35,308 instances(records) and 22 attributes(columns) in the dataset. Out of 22 attributes, 13 are numerical variables and 9 are categorical variables.

Following are the attributes in this dataset:

* + - Year Type: (Nominal/Categorical variable)
    - Year: (Numerical)
    - Organization Group Code: (Numerical)
    - Organization Group: (Nominal/Categorical variable)
    - Department Code: (Nominal/Categorical variable)
    - Department: (Nominal/Categorical variable)
    - Union Code: (Numerical)
    - Union: (Nominal/Categorical variable)
    - Job Family Code: (Nominal/Categorical variable)
    - Job Family: (Nominal/Categorical variable)
    - Job Code: (Nominal/Categorical variable)
    - Job (Nominal/Categorical variable)
    - Employee Identifier: (Numerical)
    - Salaries: (Numerical)
    - Overtime: (Numerical)
    - Other Salaries: (Numerical)
    - Total Salary: (Numerical)
    - Retirement: (Numerical)
    - Health/Dental: (Numerical)
    - Other Benefits: (Numerical)
    - Total Benefits: (Numerical)
    - Total Compensation: (Numerical)

# **3. Problems to be Solved**

List your research problems, that is, what kinds of the problems you want to solve.  
You cannot simply say I want to explore the data and find the patterns  
You should provide finer-grained research problems that can be solved by statistical techniques.

* Based on the dataset, the interested research problems are:

1. Predicting the total compensation of the employee based on various factors that will help the employers to decide what compensation should be given to employee in advance in order to keep tabs on their financial section.
2. Predicting the salaries of the employee based on benefits, compensation and job profile that will help the employees to aim for better job profiles based on high benefits.
3. Are the average salaries of all the employees same or different for various

organizations or job profiles?

# **4. Solutions**

For each problem you list above, figure out feasible solutions, and introduce your plan to perform experiments

* Feasible solutions:

1. We will use Multiple linear regression to predict the compensation and benefits given to the employee based on salary, organization and job profile.
2. We will use Multiple linear regression to predict the total salary given to the employee based on organization and job profile and other factors.
3. We will use ANOVA to compare average salaries of different employees based on job profiles and organization.

# **5. Experiments and Results**

## 5.1. Methods and Process

1. Preprocessing:

1.1 Checking and Removal of Negative values from the following numerical variables



* + Salaries
  + Overtime
  + Other Salaries
  + Retirement
  + Other Benefits

1.2 Replacement of missing values in the following nominal variables

* + Department Code

The blanks were replaced by the not applicable instead of DPH. The not applicable were not replaced by DPH.



* + Union

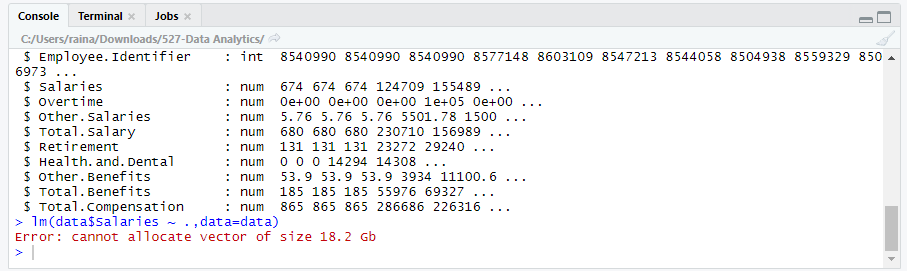
1.3 Removal of unnecessary columns from the dataset

* Employee Identifier
* Department
* Job family code
* Union code
* Organization group code
* Job code

**Normalization is not performed as per the changes told.**

**The dataset was being used without normalizing the numerical variables.**

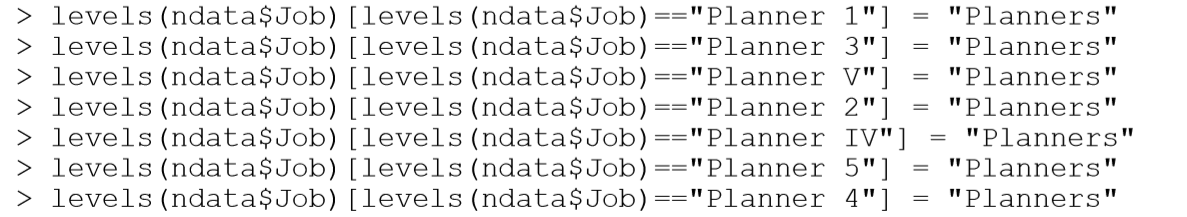
2. Resolving issues while loading the dataset



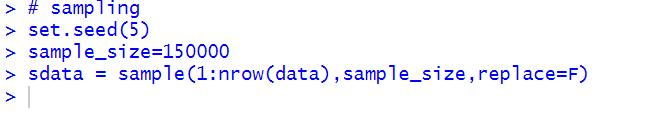
Following solutions were performed in order to reduce the size of the dataset :

2.1 Grouping was performed on the following columns

* Job
* Job Family
* Union



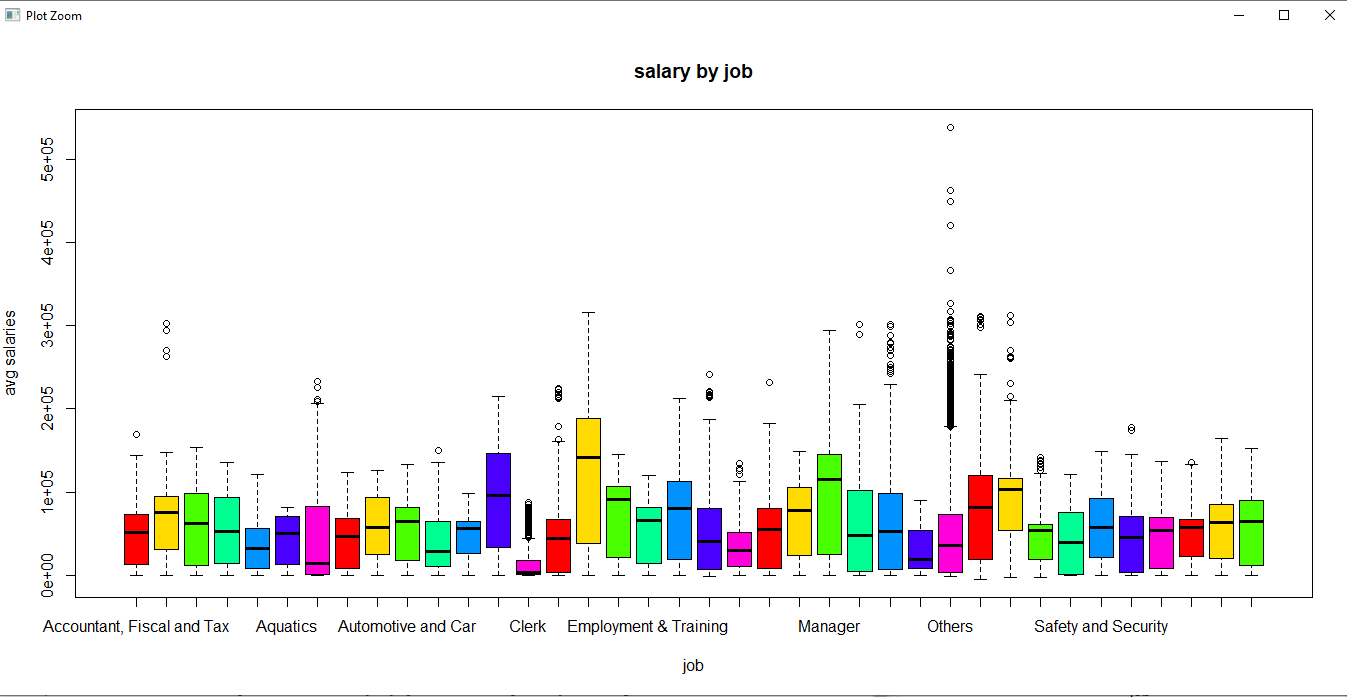
2.2 Sampling the dataset to 150000



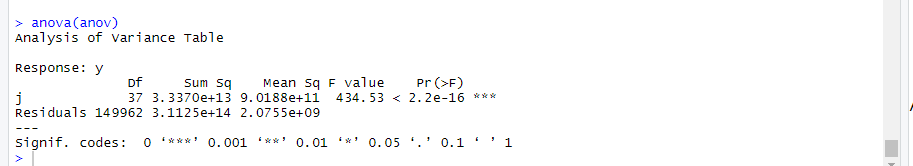
3 ANOVA and Hypothesis Testing for Job

ANOVA is used to compare average salaries of different employees based on job profiles.

* Boxplot for Salaries vs Job

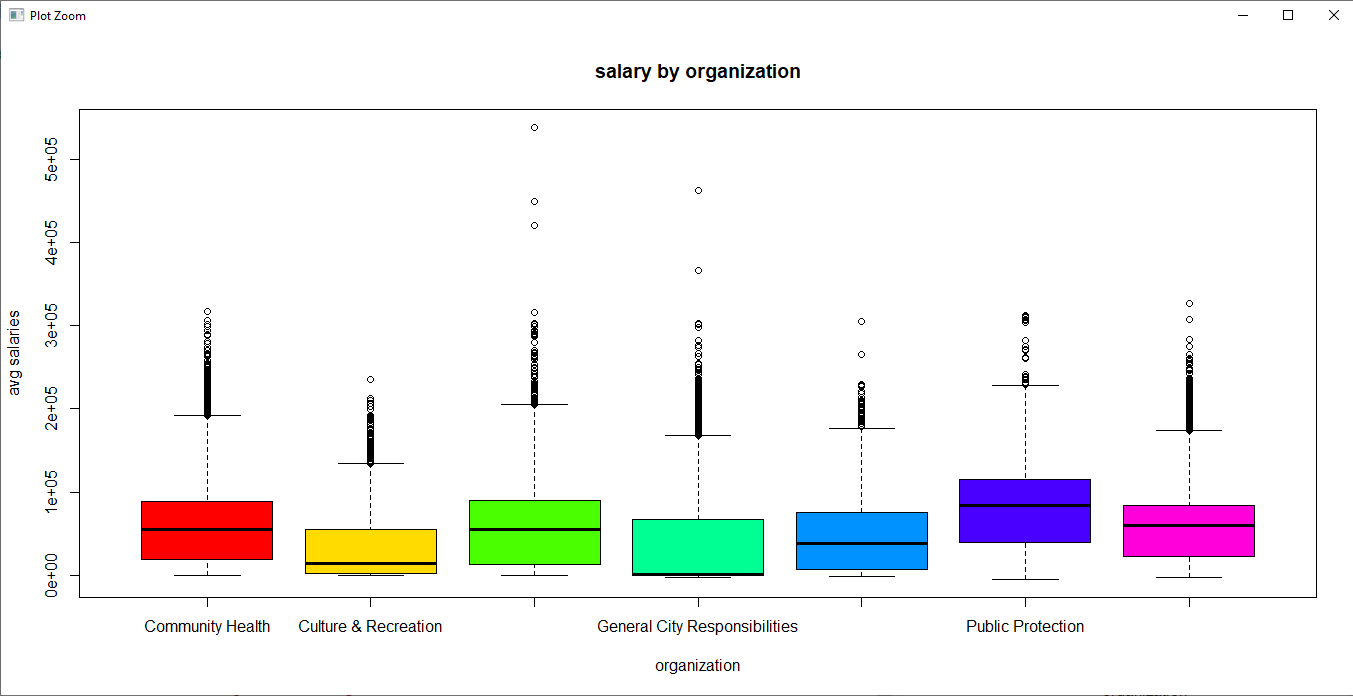


* Null Hypothesis : All the average salaries for jobs are equal
* Alternate hypothesis : Not all the average salaries for jobs are equal



* At 95% confidence level, p-value is less than 0.05,  we can reject null- hypothesis. Hence, the avg salaries are not equal for all jobs.
  1. Mean comparison for Organization\_group
* Boxplot for Salaries vs Organization\_group

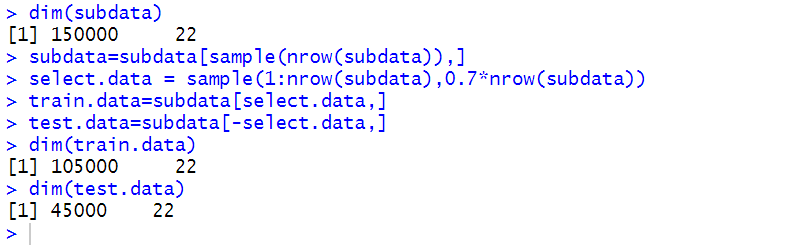
Boxplot is used to compare average salaries of different employees based on the organization group.



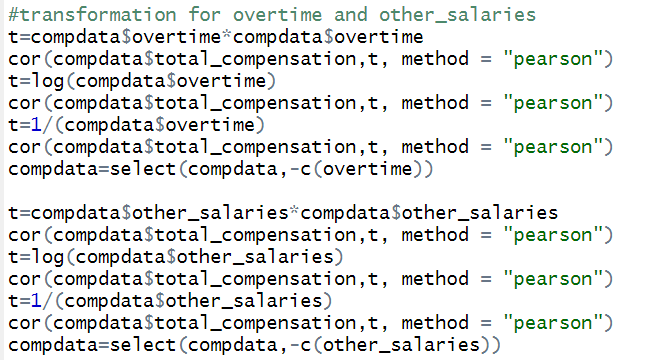
* 1. Building Predictive Models
* Creating Dummy variables



* Hold Out Evaluation



* Weak Co relations and Transformation



As we can observe that the variables overtime and other salaries have a have a weak co relation even after performing transformation.

Thus, we would remove these variables.

* 1. Predicting Total Compensation

**Search Algorithm – Backward Elimination, Feature Selection Criteria – AIC**

Model 1:

The first model was being built here.

Residual analysis was being performed on the model -

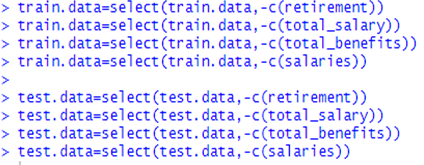
Following steps were performed –

* Checking the normality test
* Checking the variance
* Jarque-Bera Test
* Calculation of the RMSE

Check the multicollinearity using the VIF.



From the VIF calculated and after checking the co-relations we can give a conclusion that some columns can be removed having higher multi collinearity.



Model after resolving the multi-collinearity-

A screenshot of a cell phone

Description automatically generated

A picture containing sky

Description automatically generated

A screenshot of a cell phone

Description automatically generated

After resolving the multi collinearity we can observe that the adjusted R2 changes to 0.837.

Residual analysis was being performed again on the new data after removing the variables overtime and other salaries.

Following steps were performed –

* Checking the normality test
* Checking the variance
* Jarque Bera Test
* Calculation of the RMSE
  1. Predicting Total Compensation

**Search Algorithm –Forward Elimination, Feature Selection Criteria – AIC**

Model 1 –

The first model was being built here.

Residual analysis was being performed again on the data again using forward elimination with AIC.

Following steps were performed –

* Checking the normality test
* Checking the variance
* Jarque Bera Test
* Calculation of the RMSE

After this we checked the multi collinearity (certain columns with VIF more than 4 were removed after rechecking the correlations.)

Model 2:

The second model was built here.

Residual analysis was being performed again on the model again using forward elimination with AIC.

Following steps were performed:

* Checking the normality test
* Checking the variance
* Jarque Bera Test
* Calculation of the RMSE

Thus, after building the forward and backward models for total compensation, a comparison was being done between both of them.

The model with the lowest RMSE was chosen as a better model amongst both of them.

* 1. Predicting Salary

**Search Algorithm –Backward Elimination, Feature Selection Criteria – AIC**

Model 1:

The first model was being built here.

Residual analysis was being performed again on the data again using forward elimination with AIC.

Following steps were performed –

* Checking the normality test
* Checking the variance
* Jarque Bera Test
* Calculation of the RMSE

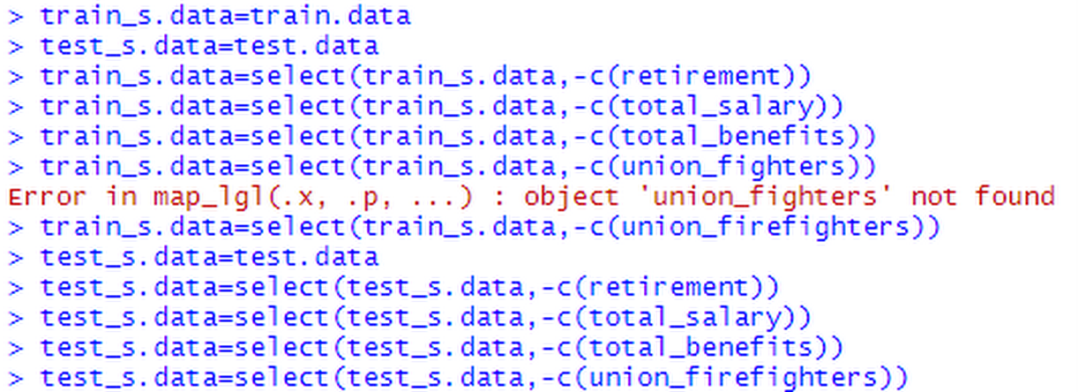
After this we checked the multi collinearity (certain columns with VIF more than 4 were removed after rechecking the correlations.)

VIF calculation:

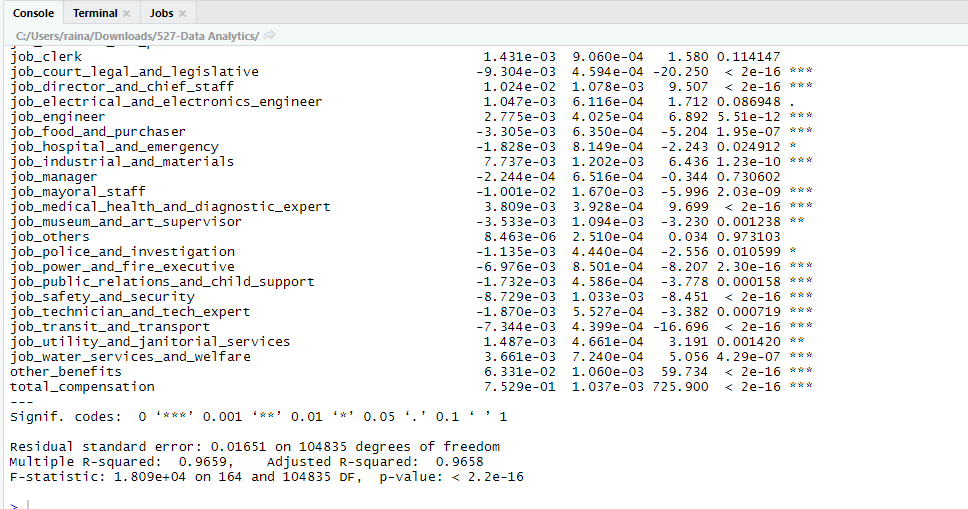
2ND MODEL:



From the VIF calculated and after checking the co-relations we can give a conclusion that some columns can be removed having higher multi co linearity.



Model after resolving the multi-collinearity-



Model 2 –

The second model was built here

Residual analysis was being performed again on the data using forward elimination with AIC.

Following steps were performed:

* Checking the normality test
* Checking the variance
* Jarque Bera Test
* Calculation of the RMSE
  1. Predicting Salary

**Search Algorithm –Forward Elimination, Feature Selection Criteria – AIC**

Model 1:

The first model was being built here.

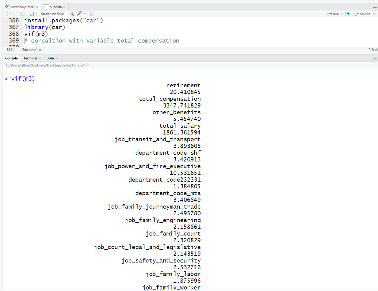
Residual analysis was being performed again on the data again using forward elimination with AIC.

Following steps were performed:

* Checking the normality test
* Checking the variance
* Jarque Bera Test
* Calculation of the RMSE

After this we checked the multi collinearity (certain columns with VIF more than 4 were removed after rechecking the correlations.)

VIF calculation



Model 2:

The second model was built here

Residual analysis was being performed again on the data using forward elimination with AIC.

Following steps were performed:

* Checking the normality test
* Checking the variance
* Jarque Bera Test
* Calculation of the RMSE

Thus, after building the forward and backward models for total compensation, a comparison was being done between both.

The model with the lowest RMSE was chosen as a better model amongst both.

5.2. Evaluations and Results

Given a same problem, you may have several solutions or build several models

Evaluate your solutions based on selected metrics and compare them

To evaluate which model is the best, we need to test all the models against the test data.

For total compensation we have built two models including the forward elimination and the backward elimination.

A conclusion would be given based on comparing their RMSE values.

**5.2.1 Search Algorithm – Backward Elimination, Feature Selection Criteria – AIC**

Model 1:

A screenshot of a cell phone

Description automatically generated

AIC, ADJUSTED R2 AND RMSE VALUES –

A close up of a persons face

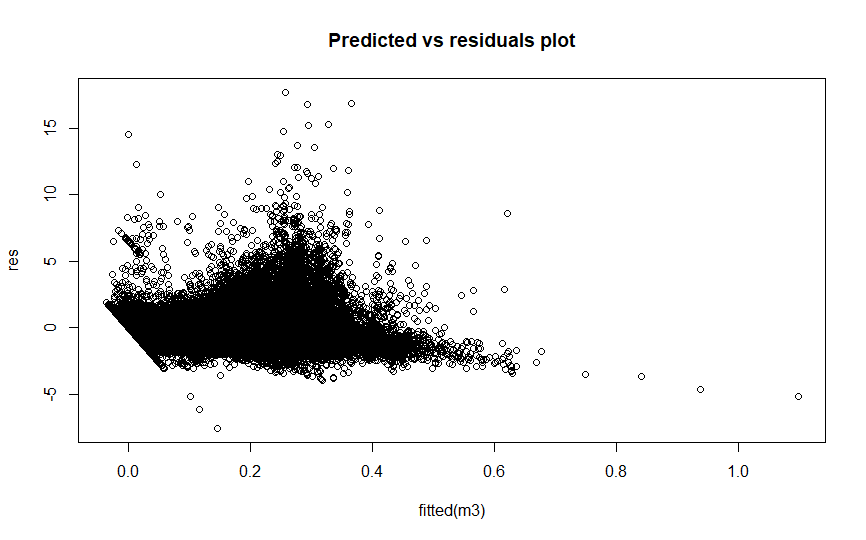
Description automatically generated

A screenshot of a cell phone

Description automatically generated

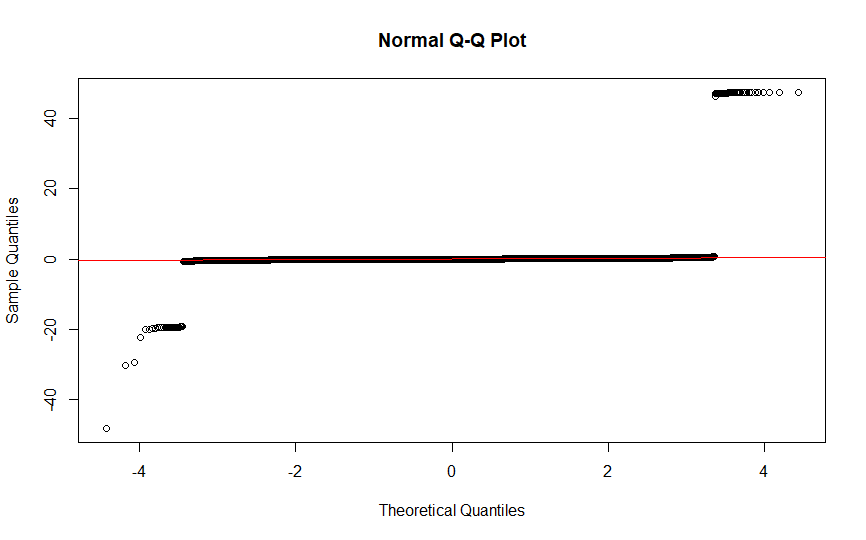
Residual Analysis:

Check the variance:



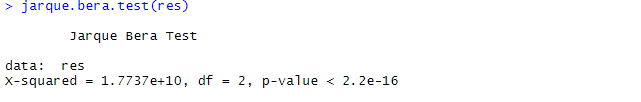
Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

Normality Test by QQ plot:



Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Jarque-Bera Test:



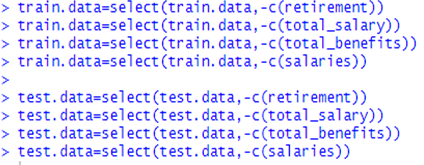
From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

In order to check the multicollinearity using the VIF.

After checking the multicollinearity and removing the variables having correlation greater than +-0.9 (around +- 1), we will built a new model.



From the VIF calculated and after checking the correlations we can give a conclusion that some columns can be removed having higher multi collinearity.



Model after resolving the multi-collinearity-

A screenshot of a cell phone

Description automatically generated

A picture containing sky

Description automatically generated

Model 2:

A screenshot of a cell phone

Description automatically generated

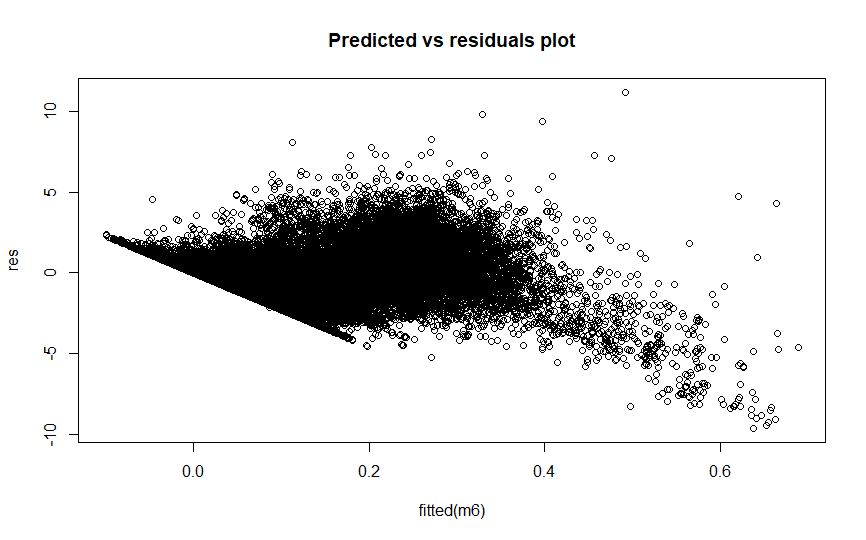
Normality Test

A close up of a map

Description automatically generated

Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Residual Plot



Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

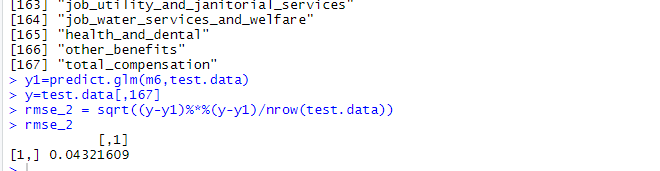
* Jarque-Bera Test

A close up of a logo

Description automatically generated

From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

* RMSE

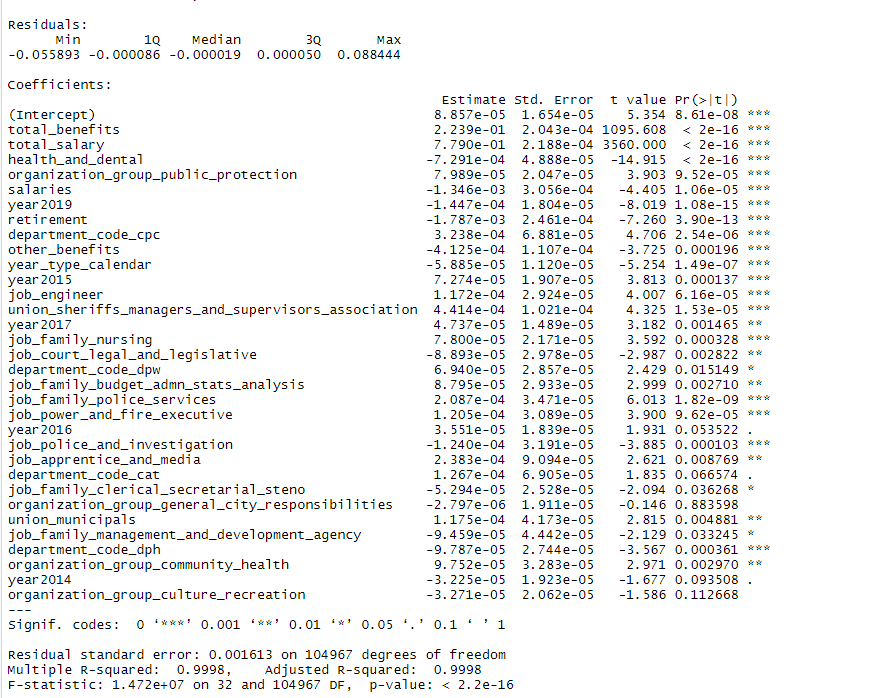


**5.2.1** **Search Algorithm – Forward Selection -Feature Selection Criteria – AIC**

**Creating Model**

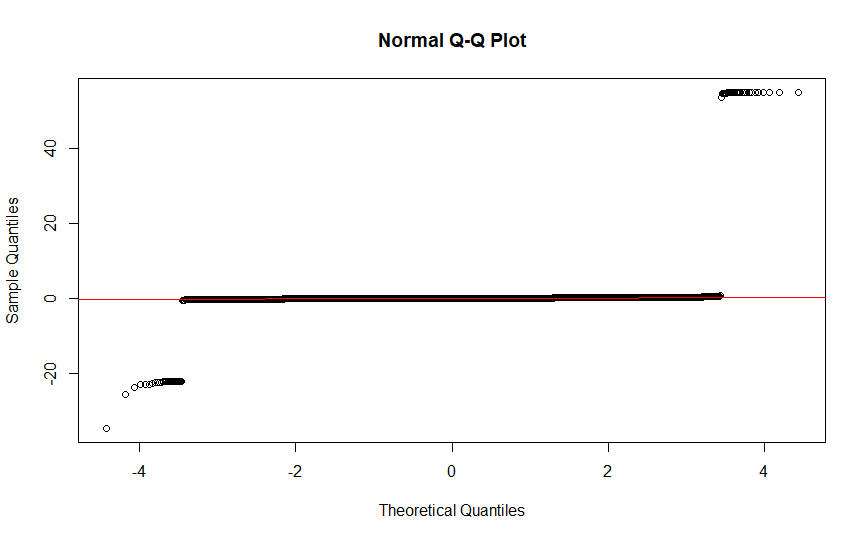
Model 1:

Calculating AIC, ADJUSTED R2 AND RMSE VALUES



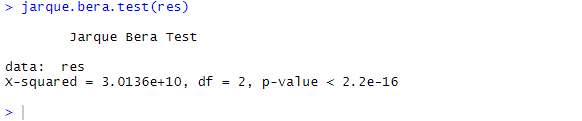
Residual Analysis-

Normality Test-



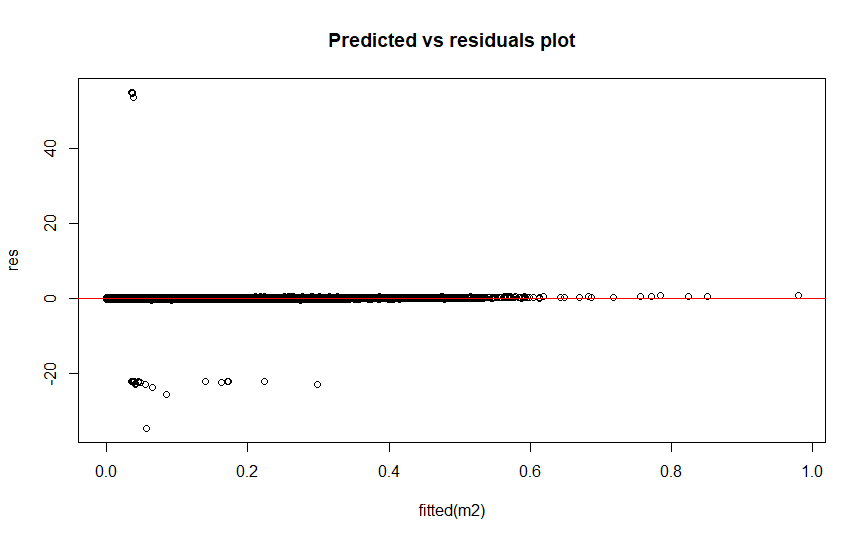
Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Jarque Bera Test-



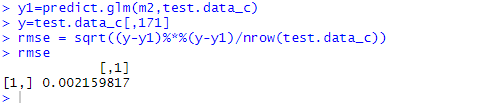
From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

Check the variance:



Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

In order to calculate the value of RMSE –



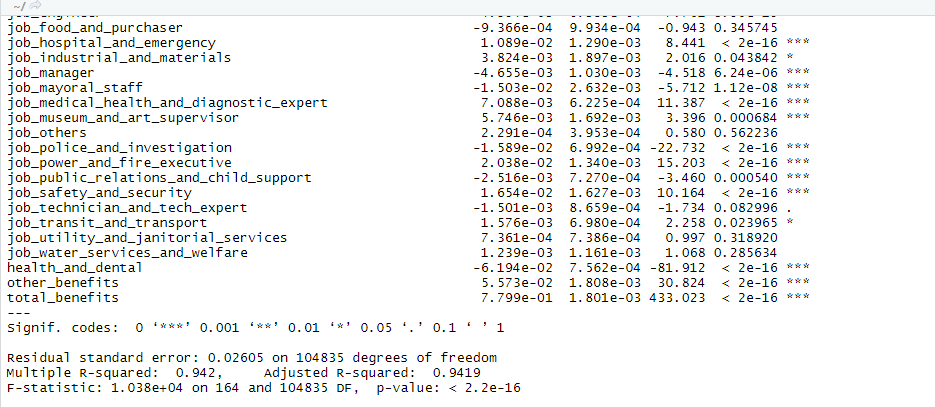
In order to check the multicollinearity using the VIF.

After checking the multicollinearity and removing the variables having correlation greater than +-0.9 (around +- 1), we will built a new model.

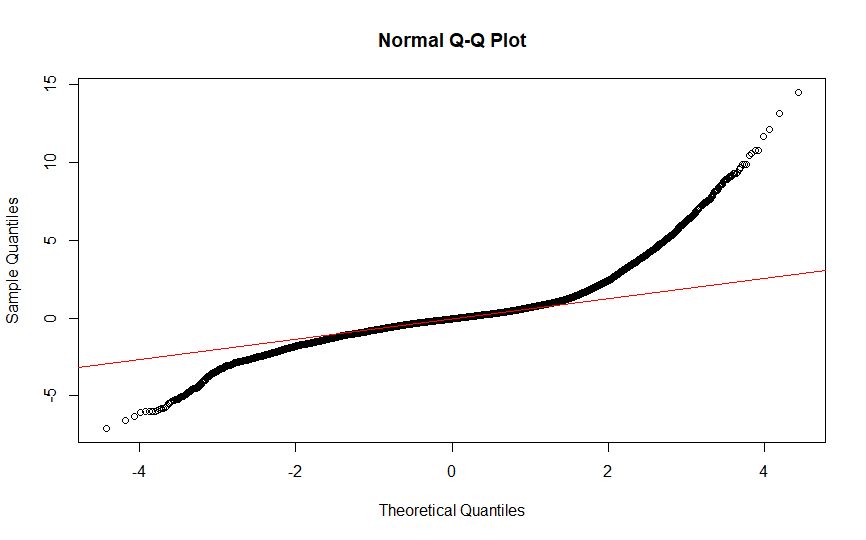
From the VIF calculated and after checking the co-relations we can give a conclusion that some columns can be removed having higher multi collinearity.

Model after resolving the multi-collinearity:

Model 2:

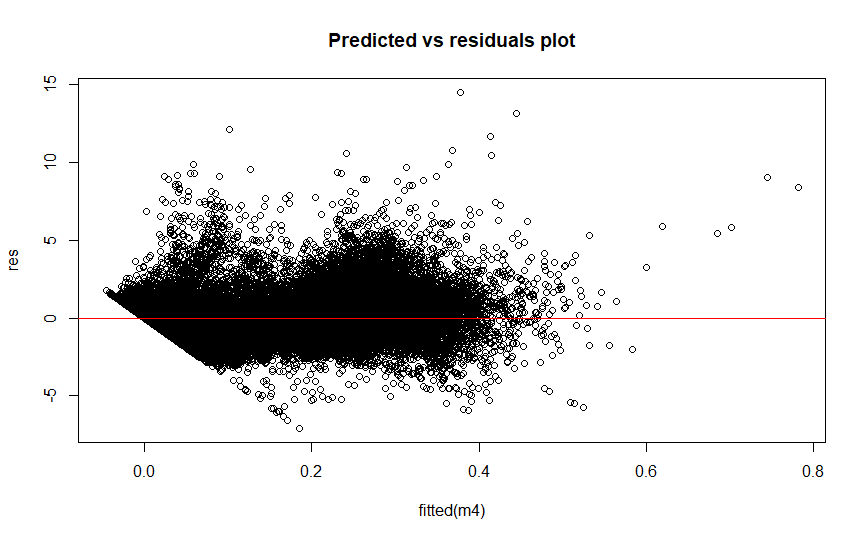


Normality Test



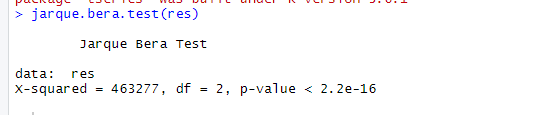
Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Residual Plot



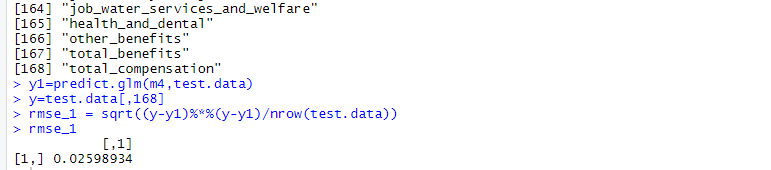
Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

* JarqueBera Test



From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

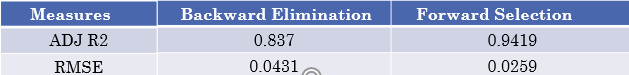
* RMSE



Here, we can observe that the variance is constant as the spread is constant and points are not scattered.

A final conclusion can be given based on the comparison of the two models built on total compensation including the backward and the forward .

Comparing the backward and forward models for total compensation -

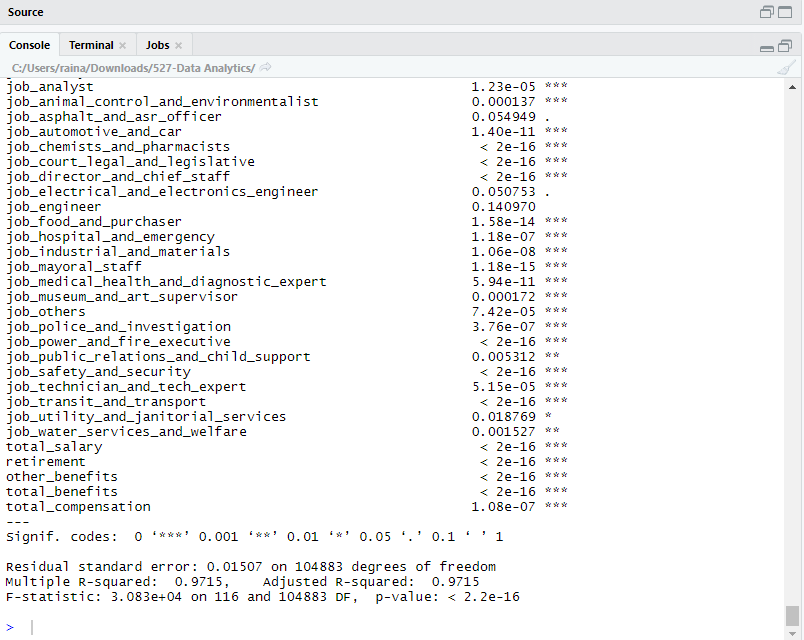


As we can observe that the RMSE for the forward selection is less as compared to the backward one. Thus, we would prefer the forward model for total compensation instead of the backward one.

* 1. Salary using backward elimination

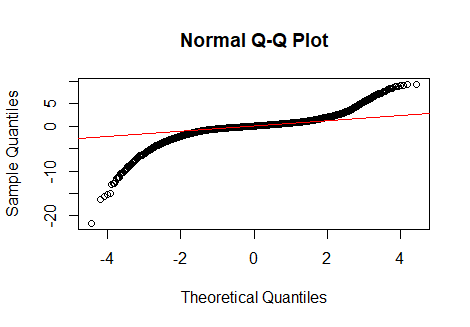
Model 1 –

Step 2 – Calculating AIC , ADJUSTED R2 AND RMSE VALUES –



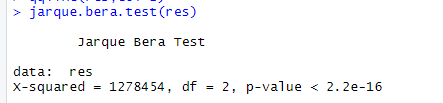
Residual Analysis-

Normality Test-



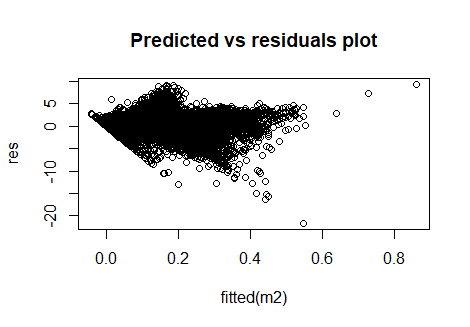
Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Jarque Bera Test-



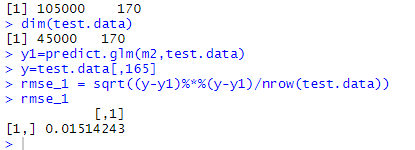
From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

In order to check the variance, we draw residual plots –



Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

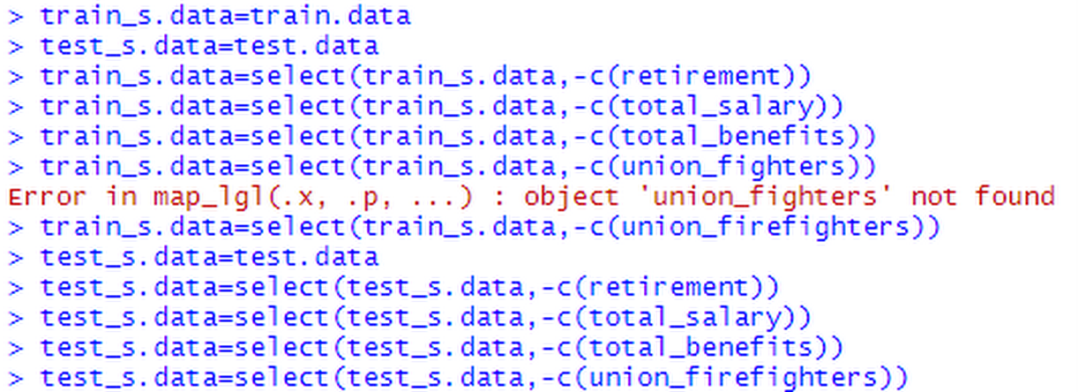
Calculate the value of RMSE –



Check the multicollinearity using the VIF.

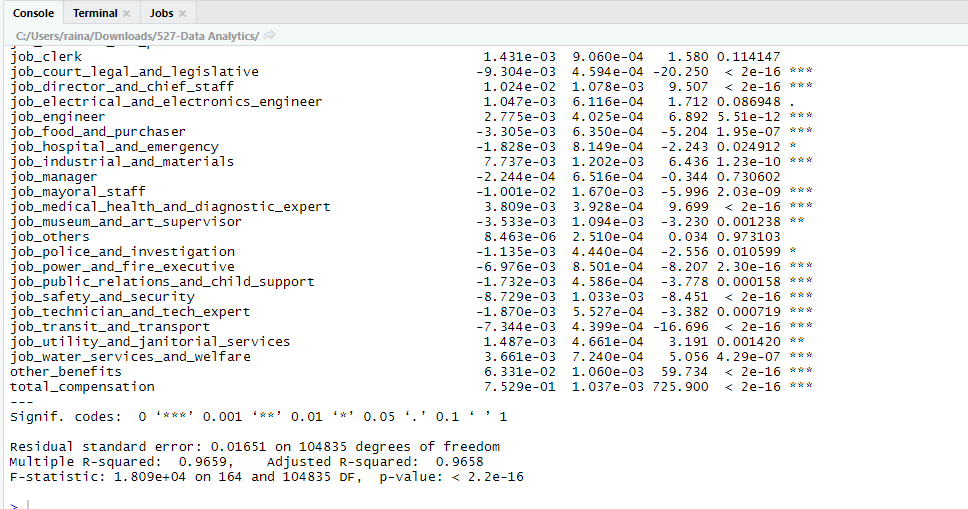
After checking the multicollinearity and removing the variables having co linearity greater than 0.09 a new model was being built.

From the VIF calculated and after checking the co-relations we can give a conclusion that some columns can be removed having higher multi co l

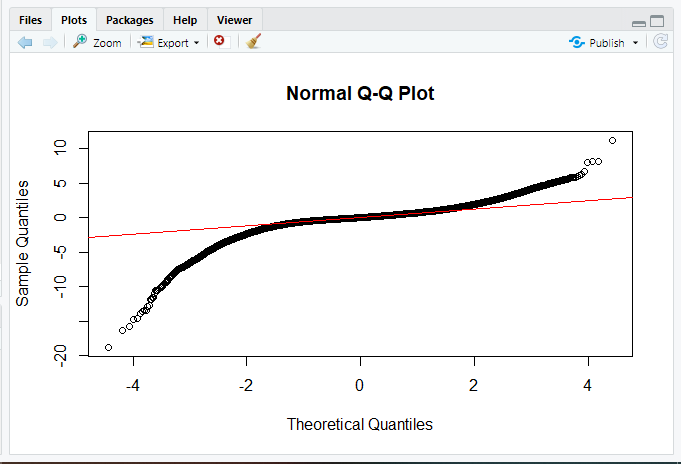


Model after resolving the multi-collinearity-

Model 2 -

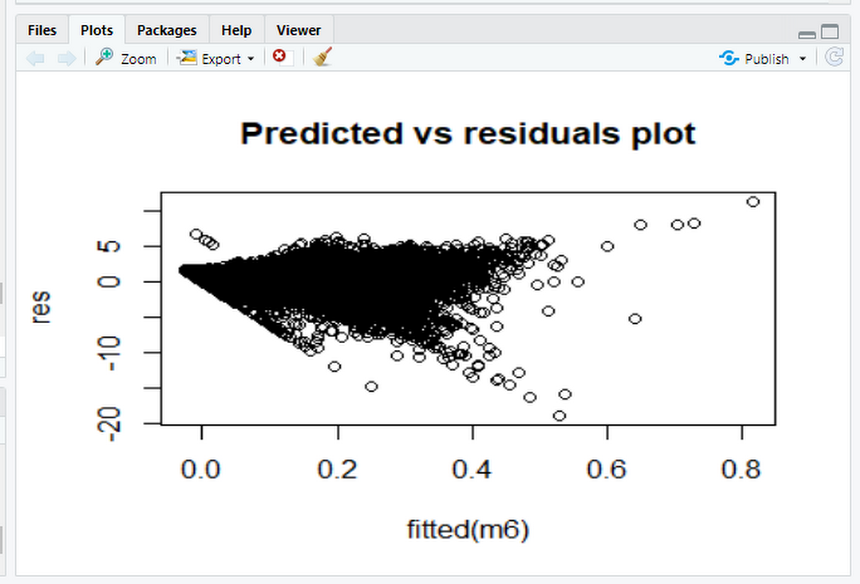


Normality Test



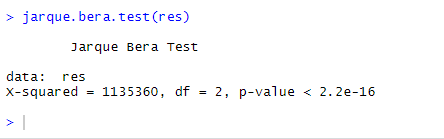
Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Residual Plot



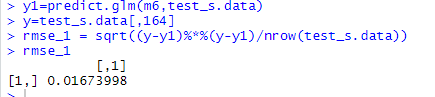
Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

* JarqueBera Test



From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

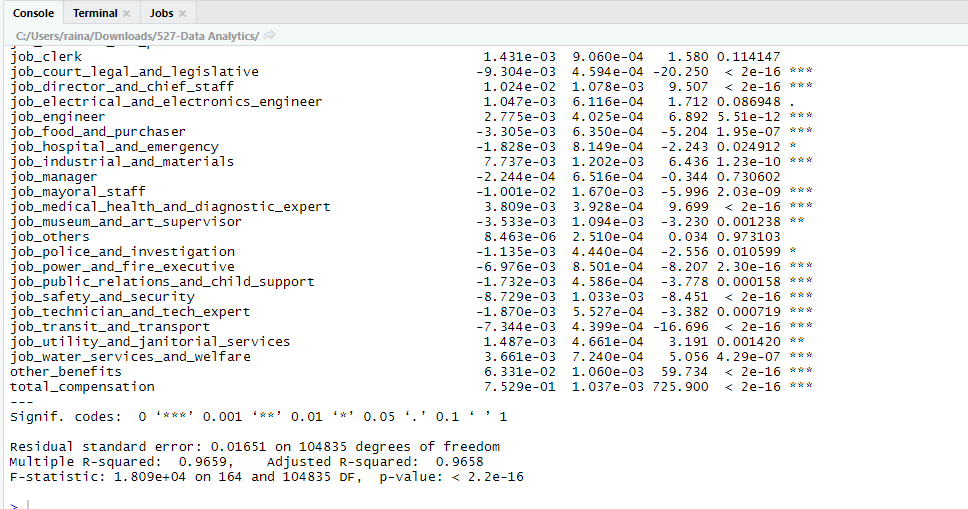
* RMSE



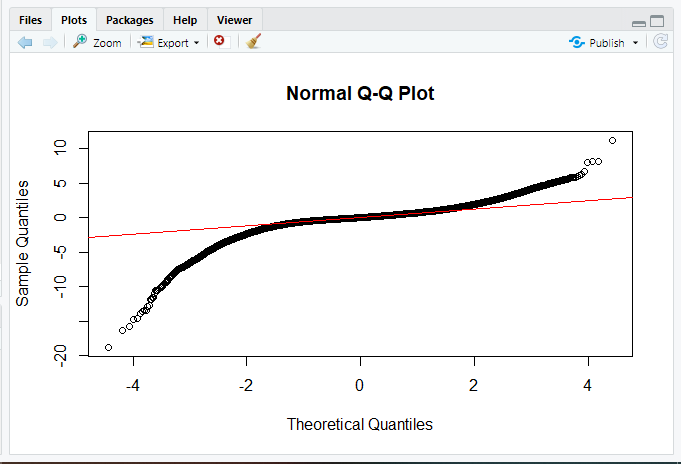
A final conclusion can be given based on the comparison of the two models built on total compensation including the backward and the forward model.

Model 2

Model after resolving the multi-collinearity-



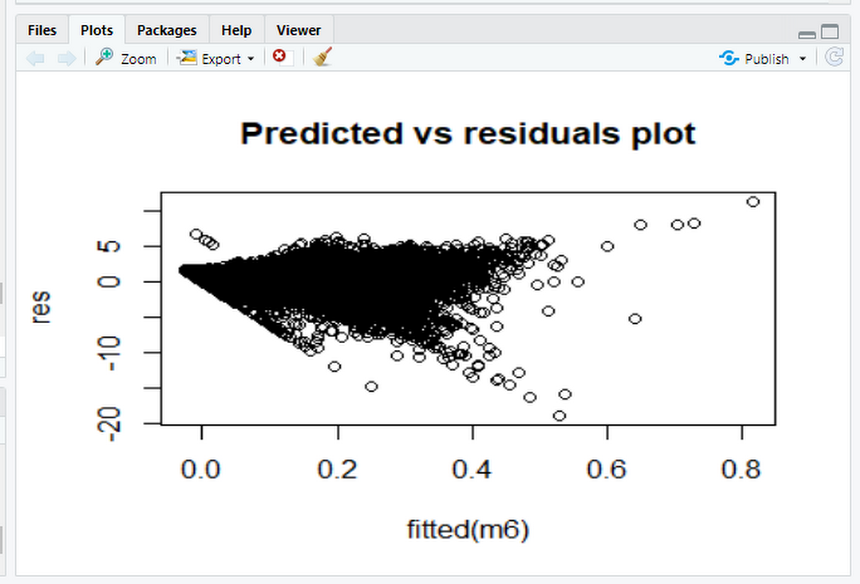
Normality Test



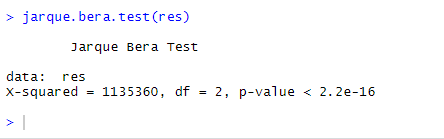
Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Residual Plot

Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

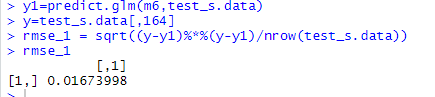


* Jarque-Bera Test



From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

* RMSE

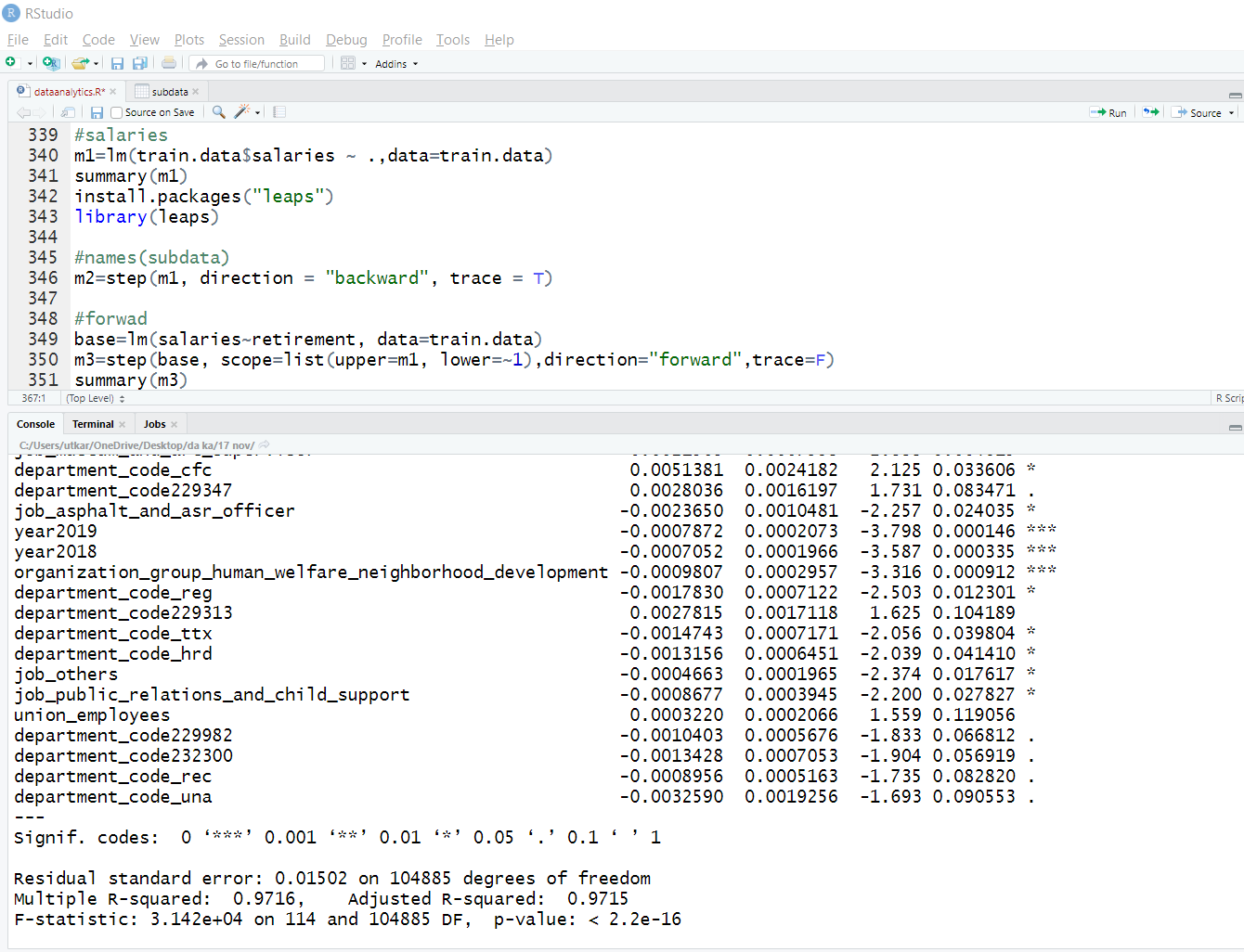


A final conclusion can be given based on the comparison of the two models built on total compensation including the backward and the forward.

* 1. Salary using forward elimination

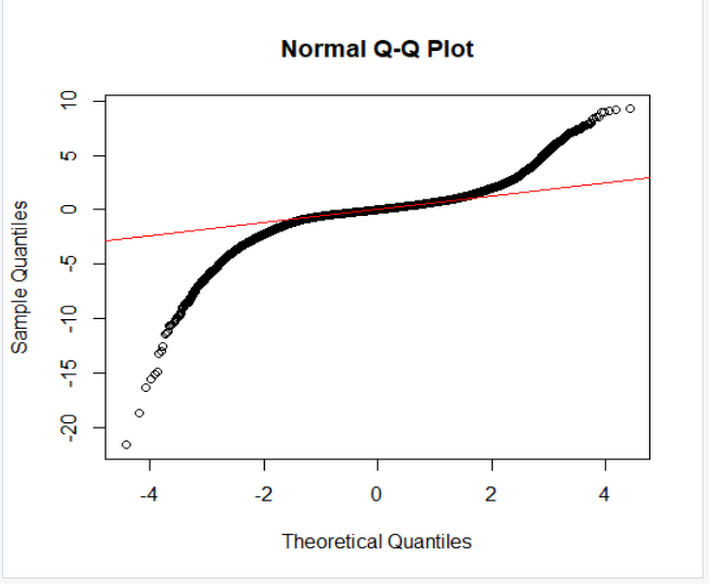
Model 1 –

Step 2 – Calculating AIC , ADJUSTED R2 AND RMSE VALUES –



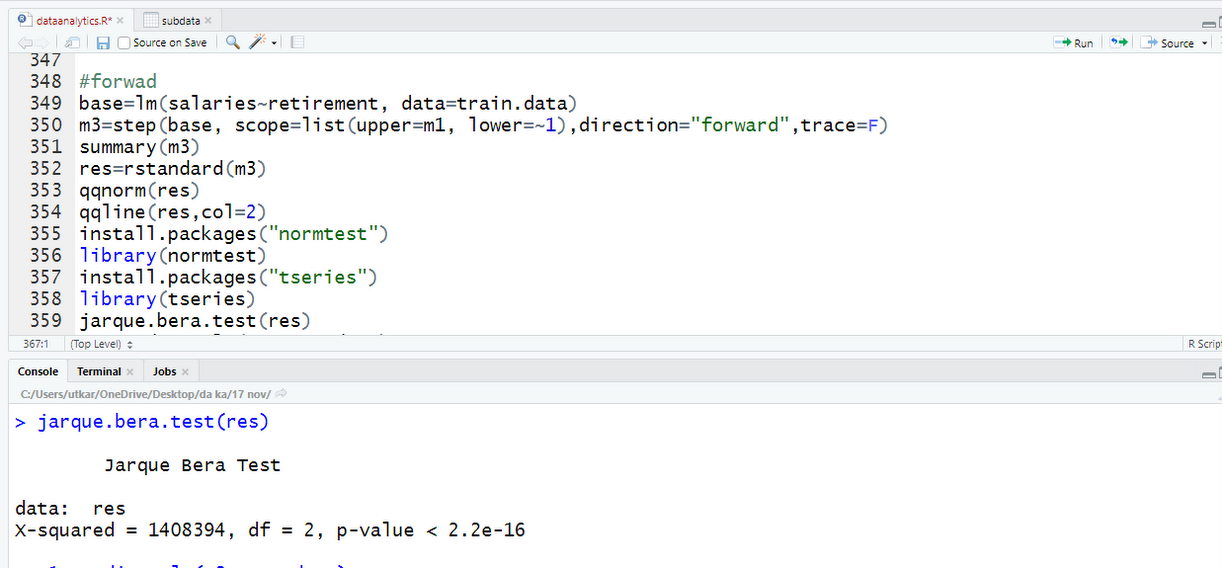
Residual Analysis-

Normality Test-



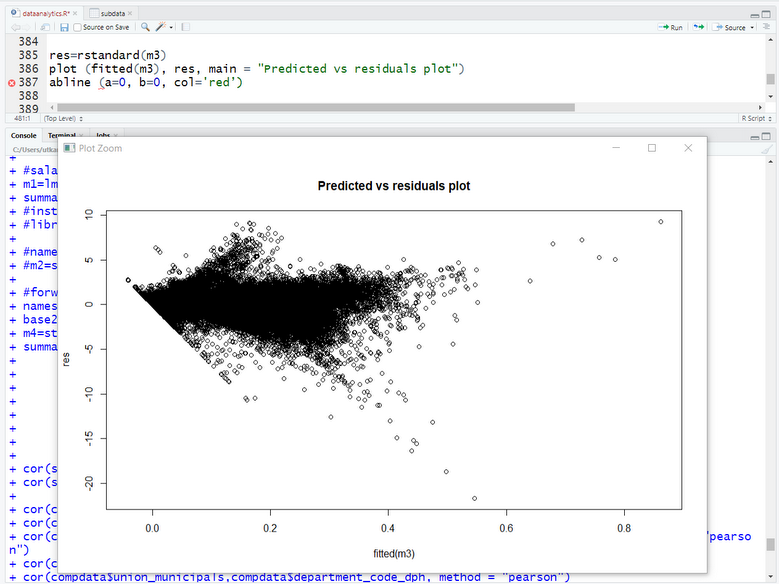
Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Jarque Bera Test-



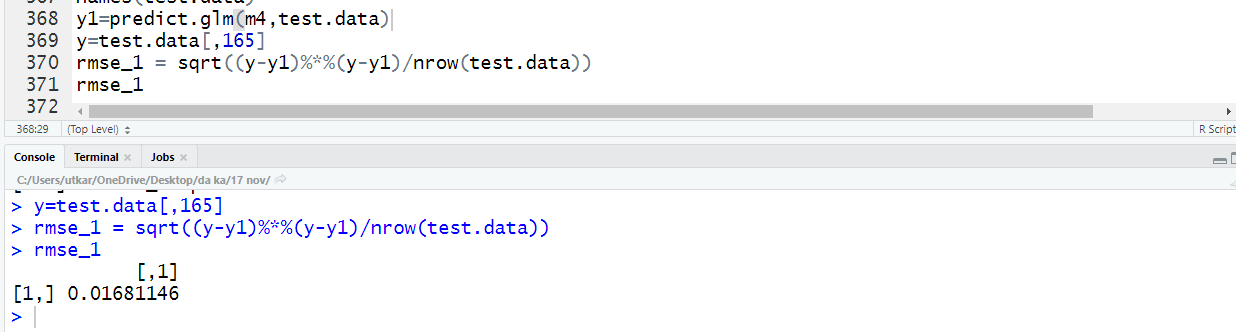
From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

Check the variance:



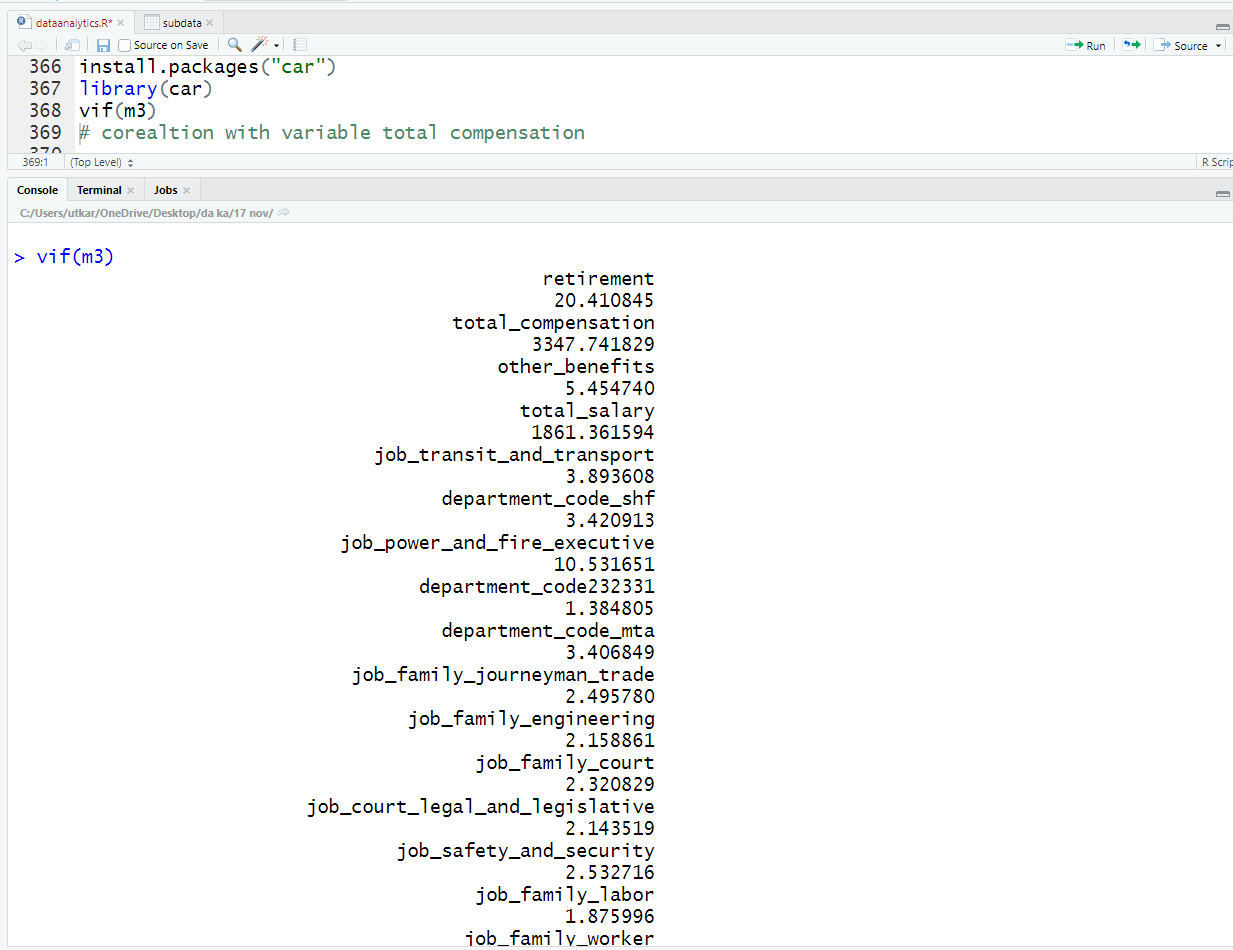
Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

In order to calculate the value of RMSE –



In order to check the multicollinearity using the VIF.

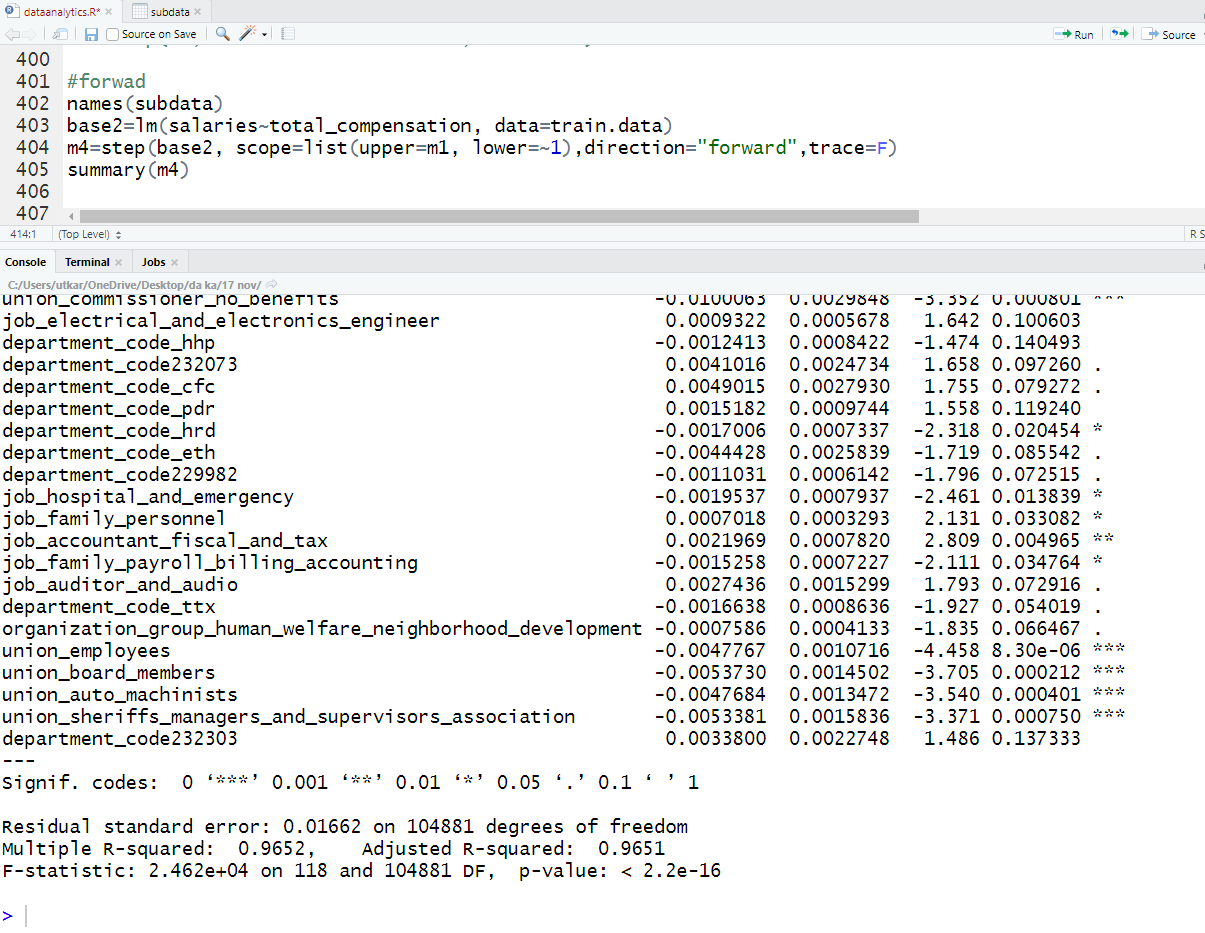
After checking the multicollinearity and removing the variables having collinearity greater than +-0.9 a new model was being built.



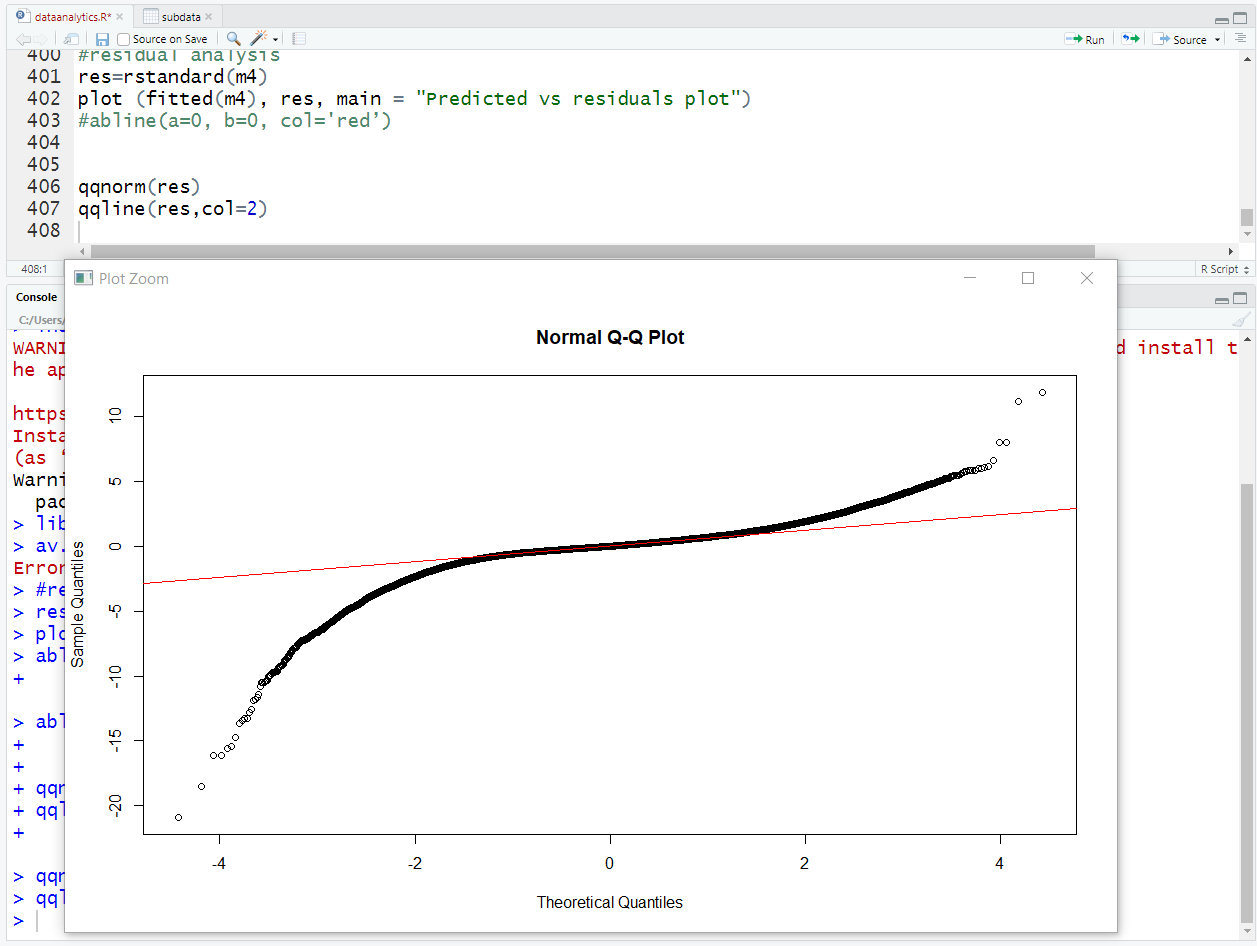
From the VIF calculated and after checking the co-relations we can give a conclusion that some columns can be removed having higher multi collinearity.

Model after resolving the multi-collinearity-

Model 2 -

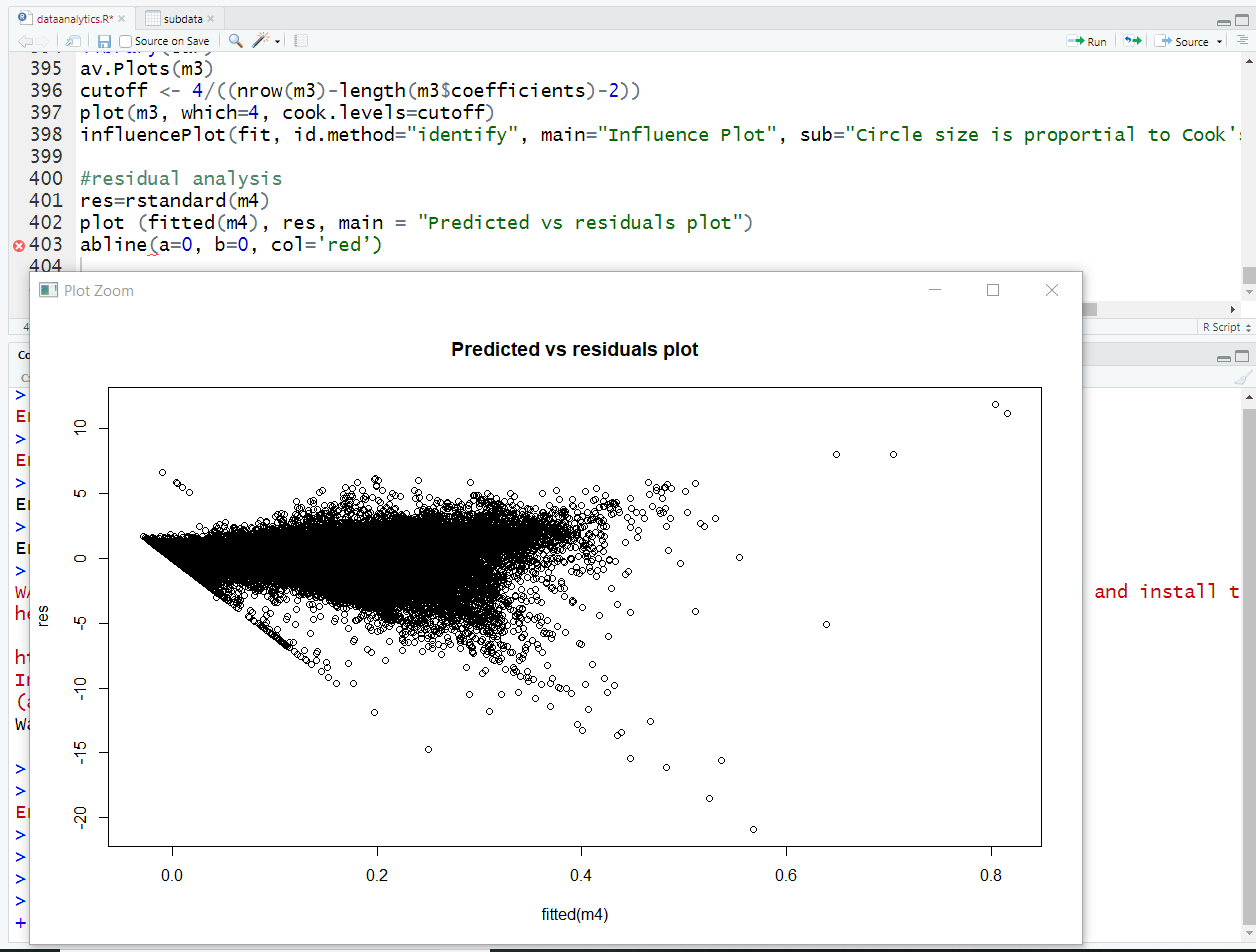


Normality Test



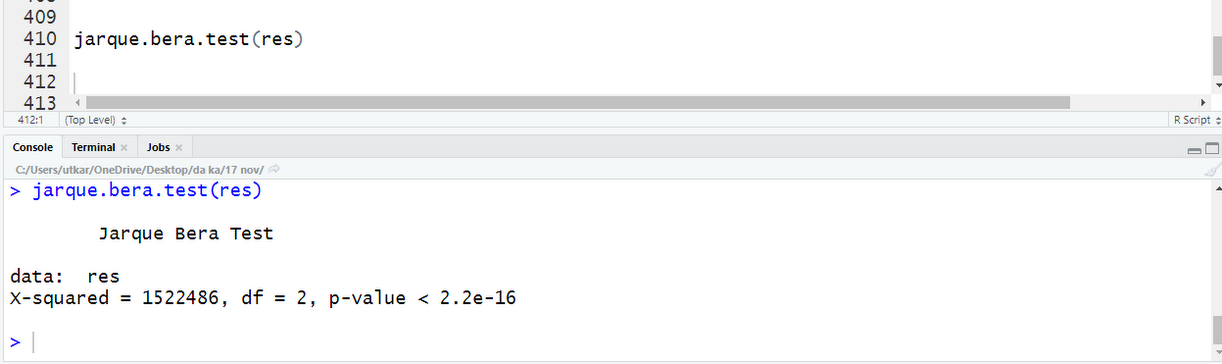
Here we can observe that the residual is normally distributed as most of the points are lying near the straight line.

Residual Plot



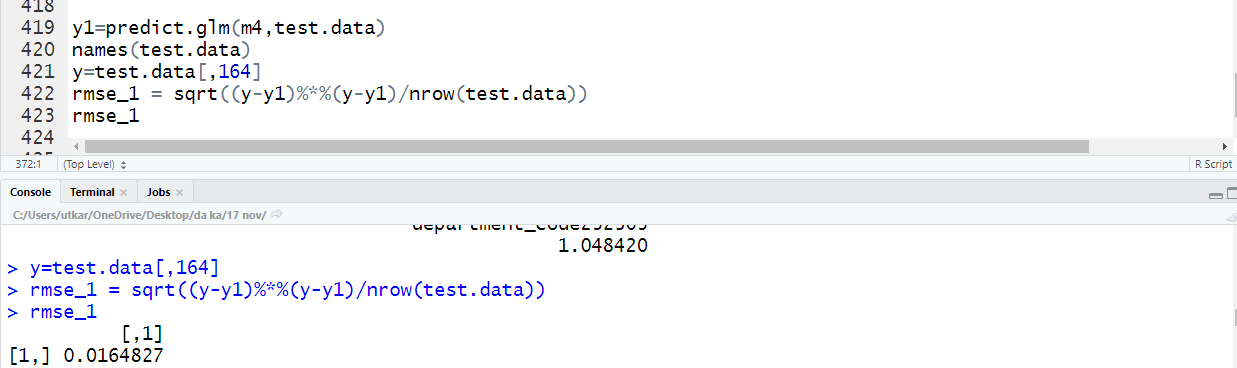
Here we can observe that the variance is constant as the spread is constant and points are randomly scattered. There is no pattern here.

* Jarque-Bera Test



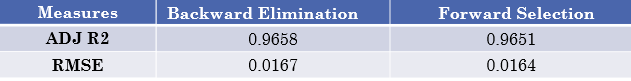
From the above test, we can conclude that p-value is less than 0.05, so the model follows normality.

* RMSE



A final conclusion can be given based on the comparison of the two models built on total compensation including the backward and the forward.

Comparing the backward and forward models for total compensation -



As we can observe that the RMSE for the forward selection is less as compared to the backward one.

Thus, we would prefer the forward model for salary prediction instead of the backward one.

## 5.3. Findings

# **For Total Compensation**

We observed that the RMSE for the backward elimination model is less as compared to the forward selection model.

Thus, we would prefer the backward model for Total compensation prediction instead of the forward one.

# **For Salary**

We observed that the RMSE for the forward selection is less as compared to the backward one.

Thus, we would prefer the forward model for salary prediction instead of the backward one.

# **ANOVA testing for job**

At 95% confidence level, p-value is less than 0.05, we can reject null- hypothesis. Hence, the average salaries are not equal for all jobs.

# **BOX PLOT for organization**

Using box plot, we can compare salaries between different organizations as variance is smaller and we can conclude that **Public Protection** has the highest average salary

# **6. Conclusions and Future Work**

## 6.1. Conclusions

* We wanted to predict the total compensation of the employee based on various factors that will help the employers to decide what compensation should be given to employee in advance in order to keep tabs on their financial section.
* We wanted to predict the salaries of the employee based on benefits, compensation and job profile that will help the employees to aim for better job profiles based on high benefits.
* We wanted to tell if all the employees are same or different for various organizations or job profiles.
* We used Multiple linear regression to predict the compensation and benefits given to the employee based on salary, organization and job profile.
* We used Multiple linear regression to predict the salaries given to the employee based on organization and job profile and other factors.
* We will use ANOVA to compare average salaries of different employees based on job profiles and organization.
* In total compensation as we can observe that the RMSE for the forward selection is less as compared to the backward one.
* Thus, we would prefer the forward model for total compensation instead of the backward one.
* In salary we can observe that the RMSE for the backward selection is less as compared to the forward one.
* Thus, we would prefer the backward model for total compensation instead of the forward one.

## 6.2. Limitations

* Due to large dataset, we had to sample our data and then build the model as it was showing system limitations.
* There was issue in calculating Influence measures as it was not showing proper results, so we had to omit that part in our case.

## 6.3. Potential Improvements or Future Work

* Grouping of the job profiles in a better way in order to provide best association.
* Individual parameter test for each job profile in ANOVA testing in order to build better prediction model.
* Treatment of influential points; due to large dataset, influence measures wasn’t giving proper results for influence points, so we can do it better on proper systems with enhanced specifications.
* Employees can use the predictive model to imply better strategies in terms of better job search which can provide better compensation and salary.
* Similarly, Employers can decide what compensation and salary should be given to the job seeker based on job and other factors in order to optimize their financial status.