

Contents

[Executive Summary 3](#_Toc500479058)

[Introduction 5](#_Toc500479059)

[About New York City jobs 5](#_Toc500479060)

[About the dataset 5](#_Toc500479061)

[Characteristics 5](#_Toc500479062)

[Steps undertaken for data cleaning 8](#_Toc500479063)

[Research Question 10](#_Toc500479064)

[Business Proposal 10](#_Toc500479065)

[Hypothesis 10](#_Toc500479066)

[Limitations 10](#_Toc500479067)

[Analysis 11](#_Toc500479068)

[Univariate Analysis 11](#_Toc500479069)

[AVERAGE Annual salary Distribution 11](#_Toc500479070)

[Job categories 12](#_Toc500479071)

[ExecJobs 13](#_Toc500479072)

[Degree Level 14](#_Toc500479073)

[Seasons 15](#_Toc500479074)

[NYC Residency Criteria 16](#_Toc500479075)

[Bivariate Analysis 17](#_Toc500479076)

[Job category 17](#_Toc500479077)

[ExecJob 19](#_Toc500479078)

[Degree levels 20](#_Toc500479079)

[Season 21](#_Toc500479080)

[NYC Residency Criteria 22](#_Toc500479081)

[Hypothesis testing 23](#_Toc500479082)

Null [Hypothesis 1 23](#_Toc500479083)

Null [Hypothesis 2 24](#_Toc500479084)

Null [Hypothesis 3 24](#_Toc500479085)

Null [Hypothesis 4 25](#_Toc500479086)

Null [Hypothesis 5 25](#_Toc500479087)

[ANOVA Testing 26](#_Toc500479088)

[Executive jobs 26](#_Toc500479089)

[Season 27](#_Toc500479090)

[Residency requirement 27](#_Toc500479091)

[Degree 28](#_Toc500479092)

[Job category 29](#_Toc500479093)

[Linear regression 30](#_Toc500479094)

[Model 1 30](#_Toc500479095)

[Model 2 31](#_Toc500479096)

[Model 3 31](#_Toc500479097)

[Multiple regression 32](#_Toc500479098)

[Model 4 32](#_Toc500479099)

[Model 5 33](#_Toc500479100)

[Removing the outliers 33](#_Toc500479101)

[Remodelling 33](#_Toc500479102)

[Variance Inflation factor 34](#_Toc500479103)

[Conclusion 34](#_Toc500479104)

[Future scope 34](#_Toc500479105)

# Executive Summary

This Report summarizes the statistical modeling and results of salaries offered for various kinds of jobs in New York city. The purpose of this report is to document the process of cleaning, transforming and modeling data with the goal of developing a regression model and achieving high overall accuracy.

The initial dataset for designing our model has been taken from New York city website and is an example of Time-Series data. The link can be found [*here*](https://data.cityofnewyork.us/City-Government/NYC-Jobs/kpav-sd4t/data). The original dataset has 3244 observations and 28 variables. After intense data cleaning and adding extra columns derived from the original dataset, the final dataset contains 1726 observations and 51 variables. The now cleansed dataset has many independent variables of which Season, Exec Job, Degree Requirement, NYC residency requirement and 17 job category variables are the chosen independent variables used for the analysis to model our dependent variable ‘Average Salary’ which is derived as the average from the variables ‘Salary Range From’ and ‘Salary Range To’.

The univariate analyses summarize the distribution of variables across the data and helps us find patterns in the data. The density plot of the ‘Average salary’ variable is a bit right skewed. Therefore, square root of Average Salary was used in the regression model which is almost normal. Other independent variables have also been modified and adjusted to model the Average Salary.

The bivariate analyses on dependent and independent variables have been conducted. Welch two samples T-test and ANOVA test are used to determine the effect of independent variables on the dependent variable i.e. AverageSalary. Each required independent variable was analyzed using the diagnostic plot.

After modeling several regression models and comparing those with each other to determine an acceptable R-squared value, a linear regression model is proposed for modeling the Average Salary.

# introduction

## About NYC jobs

As citied by Investopedia-The state of New York has the third-largest economy in the United States, trailing only Texas and California. New York's economy is so large, it would rank as the 15th or 16th largest economy in the world if it were its own country; the exact ranking depends on the source.

Its signature metropolitan center, New York City, is the single largest regional urban economy in the country. New York City is [the leading job hub](https://www.investopedia.com/articles/financialcareers/top-10-financial-career-cities.asp) for banking, information technology, finance and communication in the US. New York is also a major manufacturing center and shipping port, and it has a thriving technological sector. There are more books, magazines and newspapers published in New York than in any other state in the country. In short, the leading industries in New York are not just driving the state's economy; they are leading the charge on a national and global scale.

## About the dataset

### Characteristics

The [Mayor’s Office of Data Analytics (MODA)](http://www1.nyc.gov/site/analytics/index.page) and the [Department of Information Technology and Telecommunications (DoITT)](http://www1.nyc.gov/site/doitt/index.page) partner to form the Open Data team. As a hub of analytics in the City, MODA advocates for the use of Open Data in citywide data analytics and in the community. The data is updated in real time. We have just captured the snap of the data with the below descriptions.

The data has 3244 observations and 28 variables.

Below is the description of the dataset, after converting the variables into appropriate type and classifying them into independent, dependent and control variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Type | Classification | Description |
| Job ID | Character | Independent | Job ID is a Unique number given to the Job Postings |
| Agency | Character | Independent | Agency is the title of the Department in the Job Posting. There are 51 Agencies in total. |
| Posting Type | Character | Independent | Posting Type can be internal or external |
| # Of Positions | Integer | Independent | No of positions available in a job posting |
| Business Title | Character | Independent | There are 1,368 different business titles |
| Civil Service Title | Character | Independent | There are 325 different Civil Service titles |
| Title Code No | Character | Independent | Title codes associated with each Civil Service Title. |
| Level | Factor | Independent | There are 18 different levels of salary |
| Salary range from and Salary range to | Numeric | Dependent | Salary range; we computed average salary |
| Salary Frequency | Factor | Independent | The salary can be paid hourly, weekly or annually. |
| Work Location | Character | Independent | There are 198 different work locations |
| Division/Work Unit | Character | Independent | There are 638 different divisions/work units |
| Job Description | Character | Independent | Description of the job |
| Minimum Qual Requirements | Character; its conversion is discussed later in the report | Control | Minimum qualification requirement for the job |
| Preferred Skills | Character | Independent | Preferred skills for the job |
| Additional Information | Character | Independent | Addition information about the job |
| To Apply | Character | Independent | How to apply for the job |
| Hours/Shift | Character; its calculation has been discussed later in the report | Independent | Information about salary/hour |
| Work Location 1 | Character | Independent | Work Location |
| Recruitment Contact | Character | Independent | Recruitment Contact |
| Residency Requirement | Factor | Independent | Residency required in the New York City or not |
| Posting Date | Date | Independent | Posting Date |
| Post Until | Date | Independent | Post Until |
| Posting Updated | Date | Independent | Posting Updated |
| Process Date | Date | Independent | Process Date |

### steps undertaken for data cleaning

#### ComputING the average salary from “salary range from” and “salary range to” columns

The mean of ‘Salary Range From’ and ‘Salary Range To’ is computed and used as the ‘Average Salary’ for the jobs in New York City.

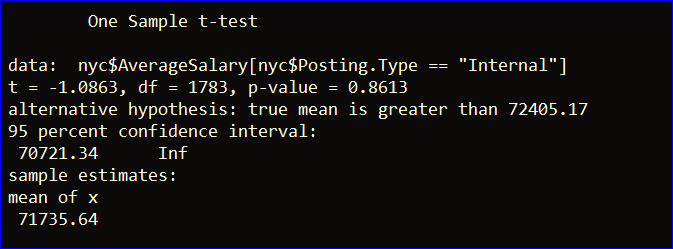
#### CalculatING hourly and weekly salaries

To uniform the ‘Average Salary’ across all jobs, hourly and weekly salaries are converted into annual salary. The Assumptions used are that total permissible working hours is 7 hours a day, business days are 5 days a week and 250 days in a year.

#### DeletING the duplicate rows

#### 

The dataset contains duplicate observations (highlighted in red) hence we removed all such rows. When we have closer look at those observations in green highlight, we see that all the values including Job ID are same except for posting type. We have a total of 3244 observations i.e. Job entries, out of which 230 observations are unique and 1507 observations are the Job entries which are posted in both internal and external medium. As 1507 observations have duplicates values, we have removed those observations. To support our cleaning, we performed hypothesis testing for average salary of Internal and External posting types. As expected, the p-value is greater than 0.05 which means we fail to reject the null hypothesis which explains that the means of average salary for internal posting type jobs and external posting type jobs are almost equal. The difference in mean is due to 230 Unique observations in the dataset which are either internal or external.



#### Changing the date format from character to date for all the date variables using Lubridate library.

Initially, date columns were ‘Character’ variable. With the help of ‘lubridate’ function, the date columns are converted to Year:Month:Date format.

#### SegregatiNG Business Title to Executive and Non-Executive

With the help of ‘grepl’ function, we have categorized the business title into ‘Executive’ and ‘Non Executive’ jobs. ‘Executive’ jobs are the ones which have ‘Director’ and ‘Executive’ in the business title and the rest of them are ‘Non Executive’ jobs.

#### AddING columns ‘Degree Level’, ‘Residency Requirement’,’SEASON’ and 17 ‘Job Categories’

* Derived ‘Degree Level’ column from ‘Minimum qual requirement’ using the ‘grepl’ function. ‘Degree Level’ is a factor with three levels namely Master’s, Bachelor’s and None
* ‘NYC Residency Requirement’ column gives information regarding whether the job requires the candidate to have residency in New York city. The data have been extracted using ‘grepl’ function from ‘NYC Residency Requirement’ column to check if there’s difference between jobs being offered to candidates with and without residency
* Based on ‘Job Posted’ date, each posting is categorized into seasons. Based on Solstice and Equinoxes,

Dec 16th -Mar 15th corresponds to Winter

Mar 16th -Jun 15th corresponds to Spring

Jun 16th -Sep 15th corresponds to Summer

Sep 16th -Dec 15th corresponds to Fall

* Depending on industry segment, Job categories have been classified into 17 categorical factors with the use of ‘grepl’ function on the ‘Job category’ column.

## Research Question

What are the factors that drive the salary of jobs offered in New York City?

## Business Proposal

Identifying a candidate who will fit a position in the firm is one part of the hiring process, offering the right amount of compensation which will match the market offering and the expectation of the candidate is the other. The compensation offered for a position is quite influential on the decision made by the candidate. Either the firms lose an excellent candidate because he got a better offer in the market or the firms offered too much as they were anxious to have the candidate on-board. Offering the right pay is crucial from both the company’s and candidate’s perspective. Keeping these factors in mind, we have developed this model which will help the business in decision making.

## Null hypotheSES

* [The average salary for executive jobs is same as non-executive jobs.](#_Toc500455569)
* [The average salary for jobs offered in summer is same as winter.](#_Toc500455570)
* [The average salary for jobs offered to resident is same as that of non-resident.](#_Toc500455571)
* [The average salary for the jobs with master’s degree is same as that of jobs with no master’s degree.](#_Toc500455572)
* [The average salary for engineering jobs is same as that of finance jobs](#_Toc500455573).

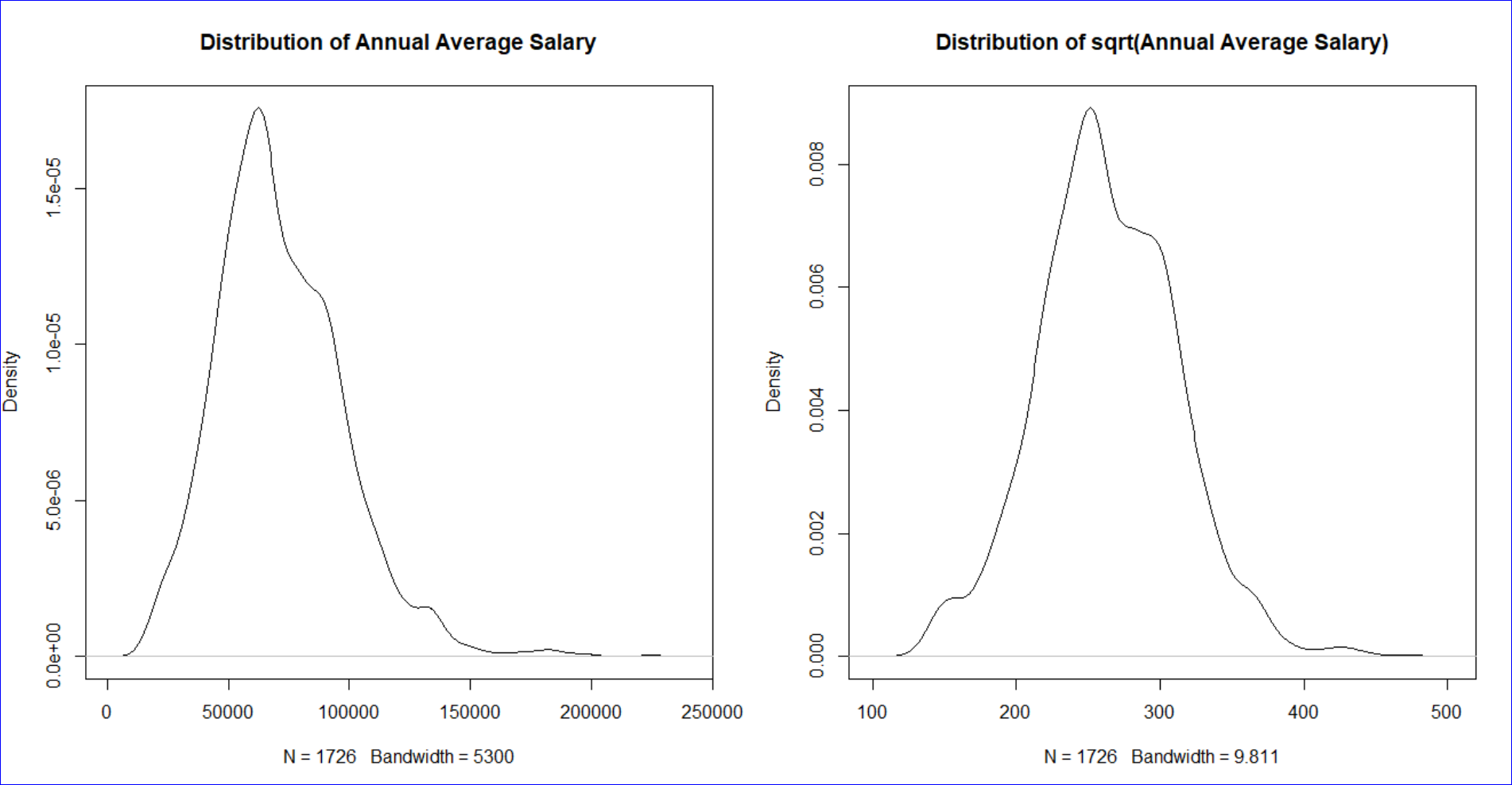
## LIMITATIONS

* Annual average salary is an estimate based on the range of the salary provided in the data set for each job ID.
* There was no exact information to estimate the total working hours of hourly and daily observations.

# Analysis

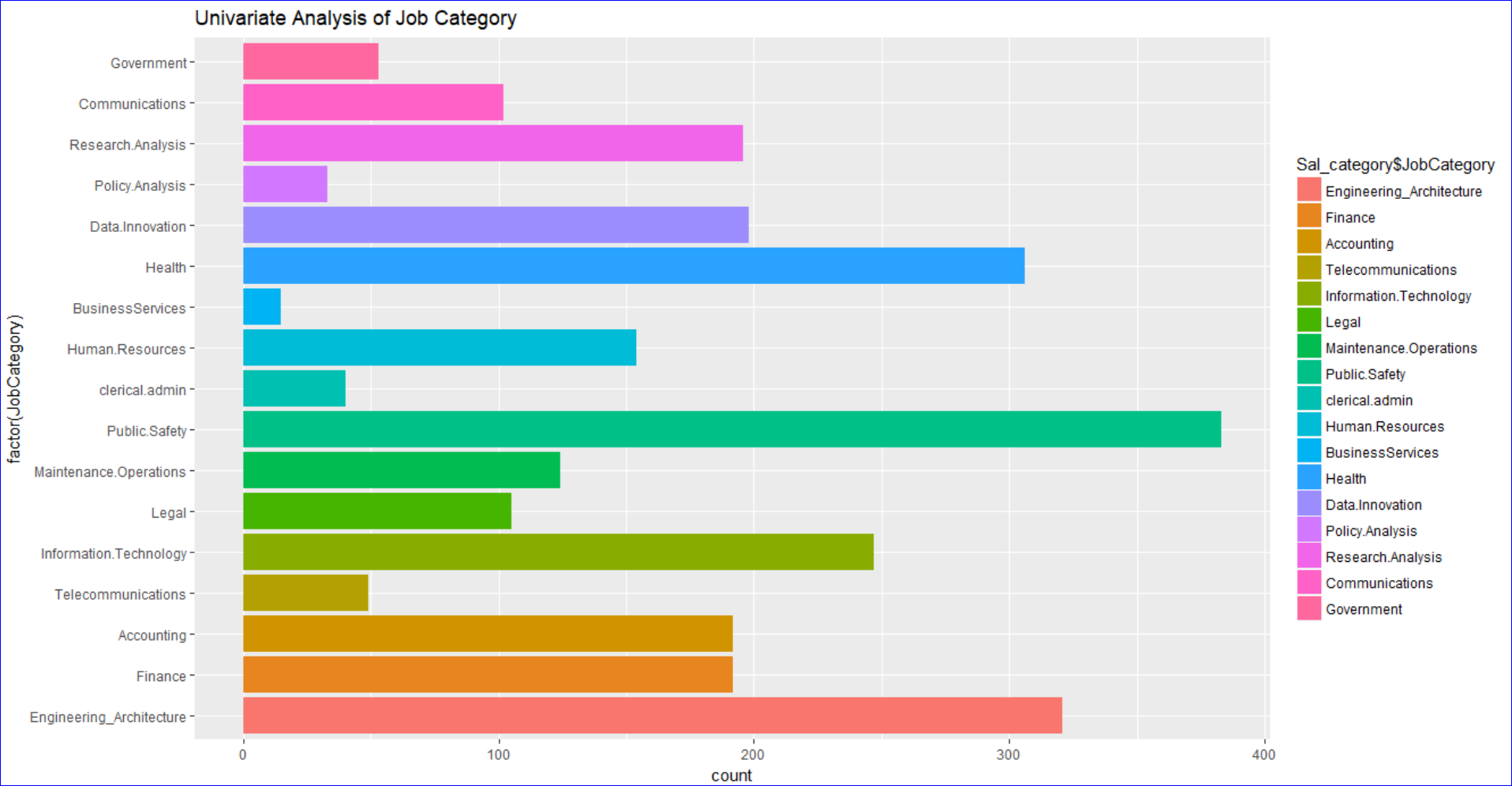
## Univariate Analysis

### AVERAGE Annual salary Distribution

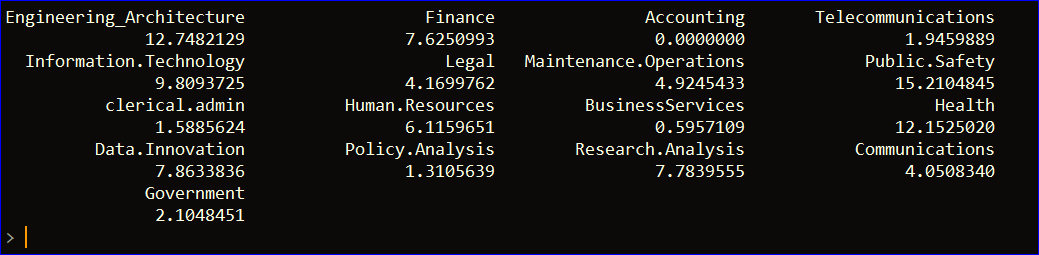


The plots describe the distribution of average salary and square root of average salary respectively. The density plot of the ‘Average salary’ variable is a bit right skewed as depicted in the first plot. Therefore, square root of Average Salary was used in the regression model which is almost normal.

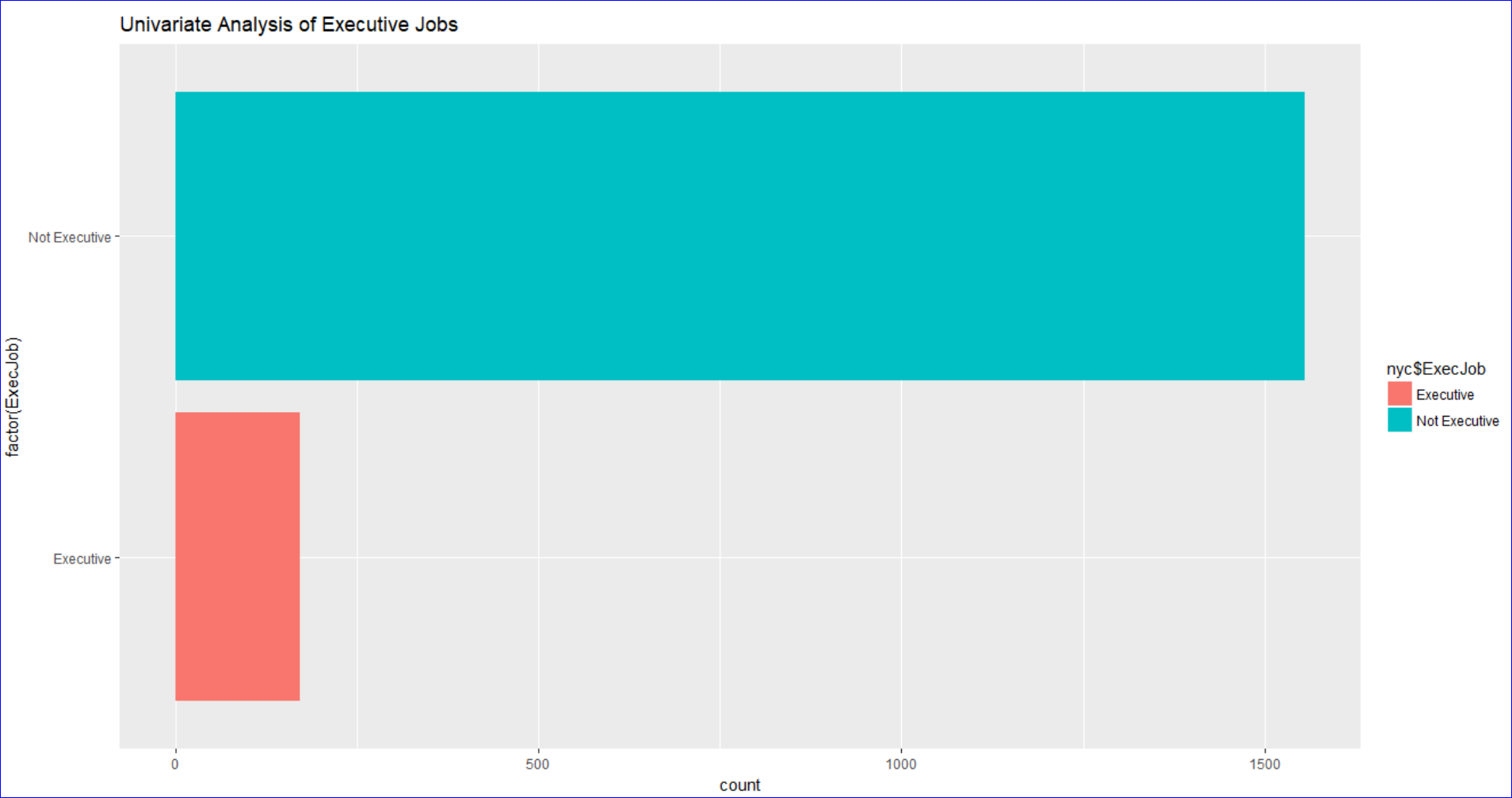
### Job categories



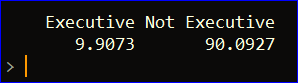
The graph shows the count of each job category in the data set. We can observe that the maximum number of jobs are being offered in the field of ‘Public Safety’ while the least number of jobs are being offered in the field of Business Services. The proportion table below provides the same insight.



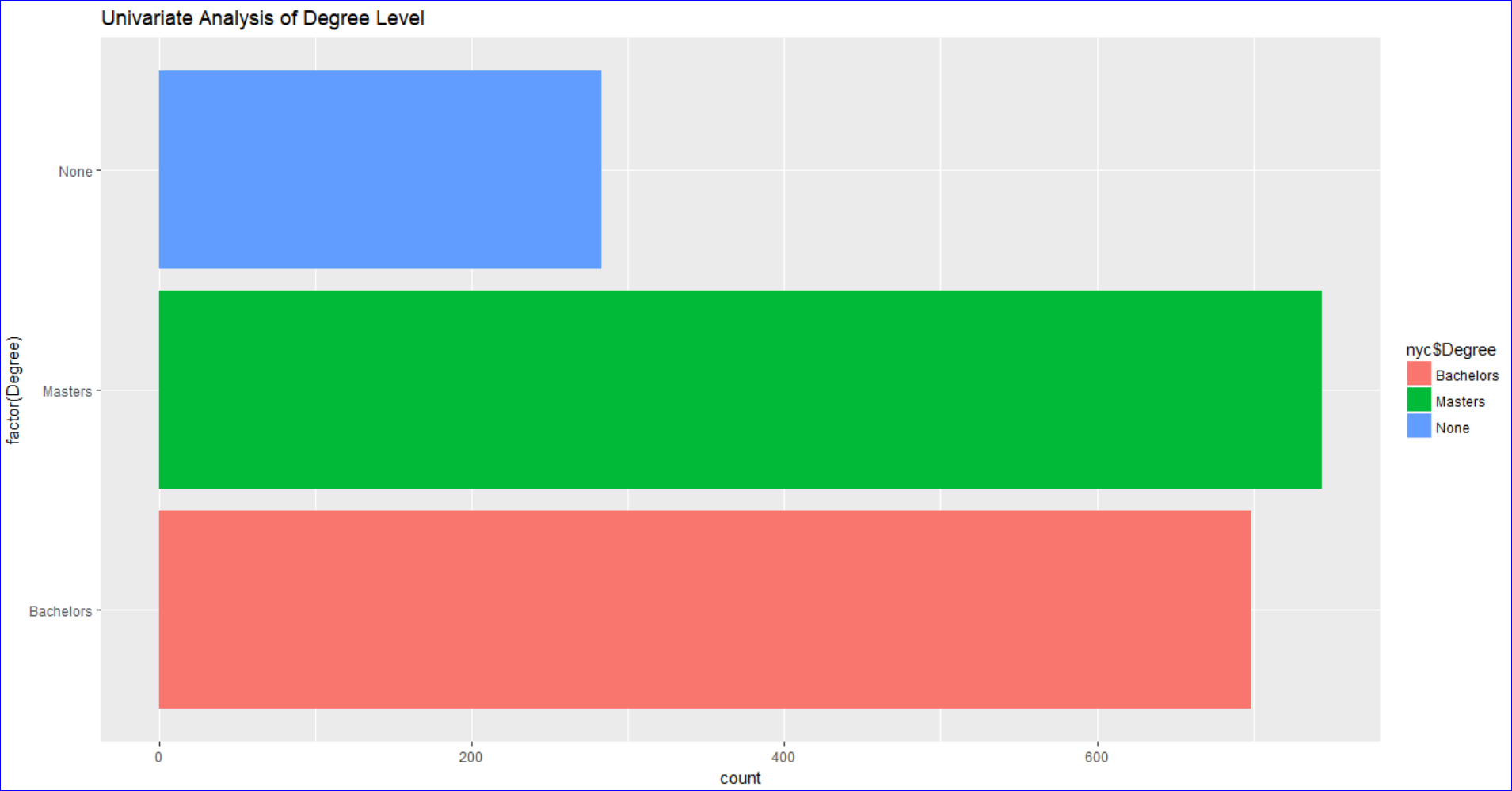
### ExecJobs



The graph shows the count of the number of jobs offered for Executive and Non-Executive jobs. From the proportion table below, we can observe that number of non-executive jobs is almost 10 times the number of executive jobs.



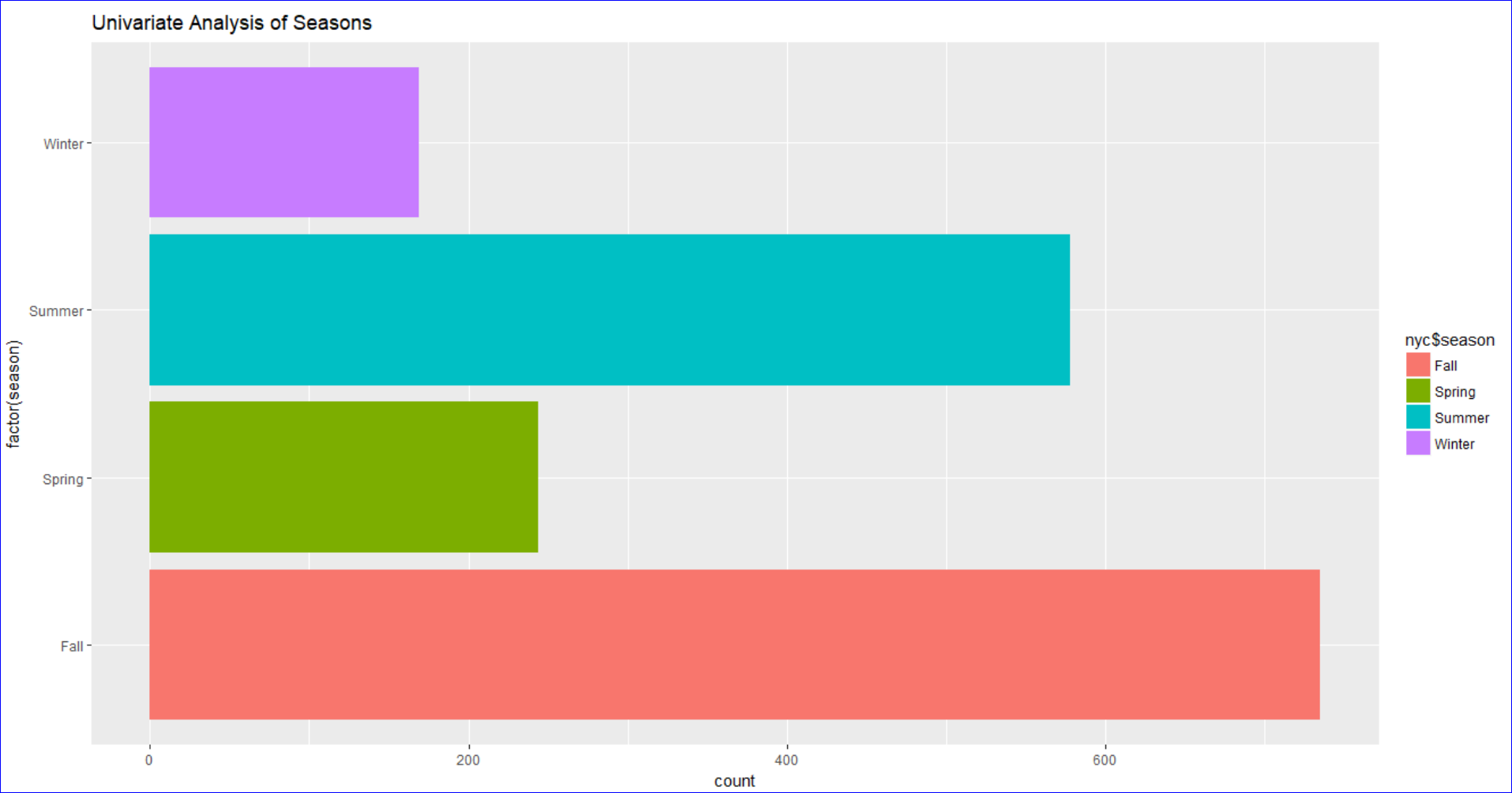
### Degree Level



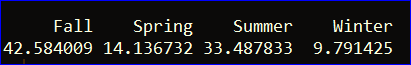
The graph shows the count of the number of jobs offered to the people with minimum degree requirement as bachelor’s degree, master’s degree and none. From the proportion table below, we can observe that number of jobs offered to people with Master’s degree is maximum followed by Bachelor’s degree and no degree at all.



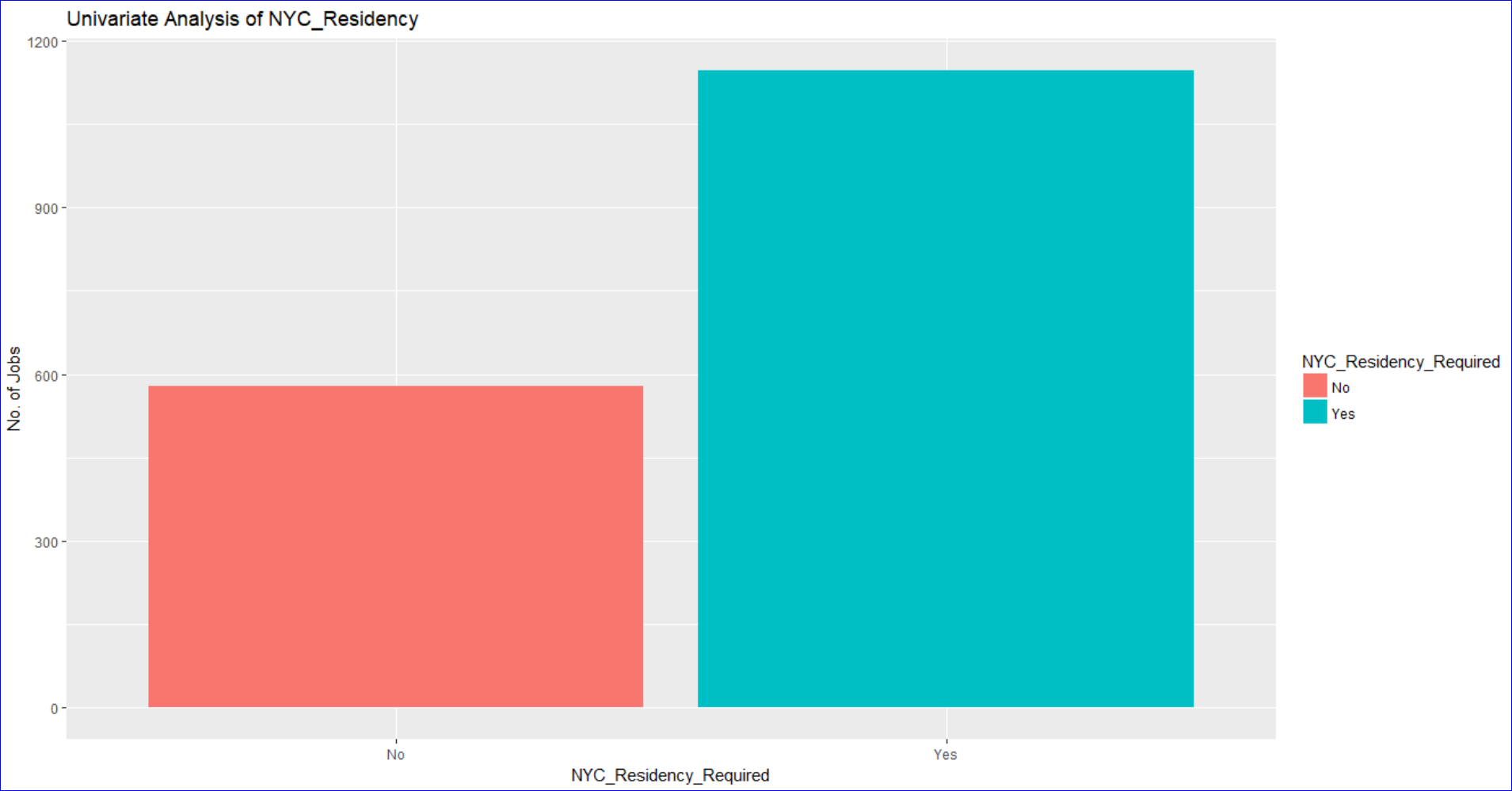
### Seasons



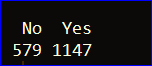
The graph shows the count of the number of jobs offered during each season based on equinoxes and solstice. From the proportion table below, we can observe that number of jobs offered in Fall is maximum followed by summer, spring and winter being the least.



### NYC Residency Criteria

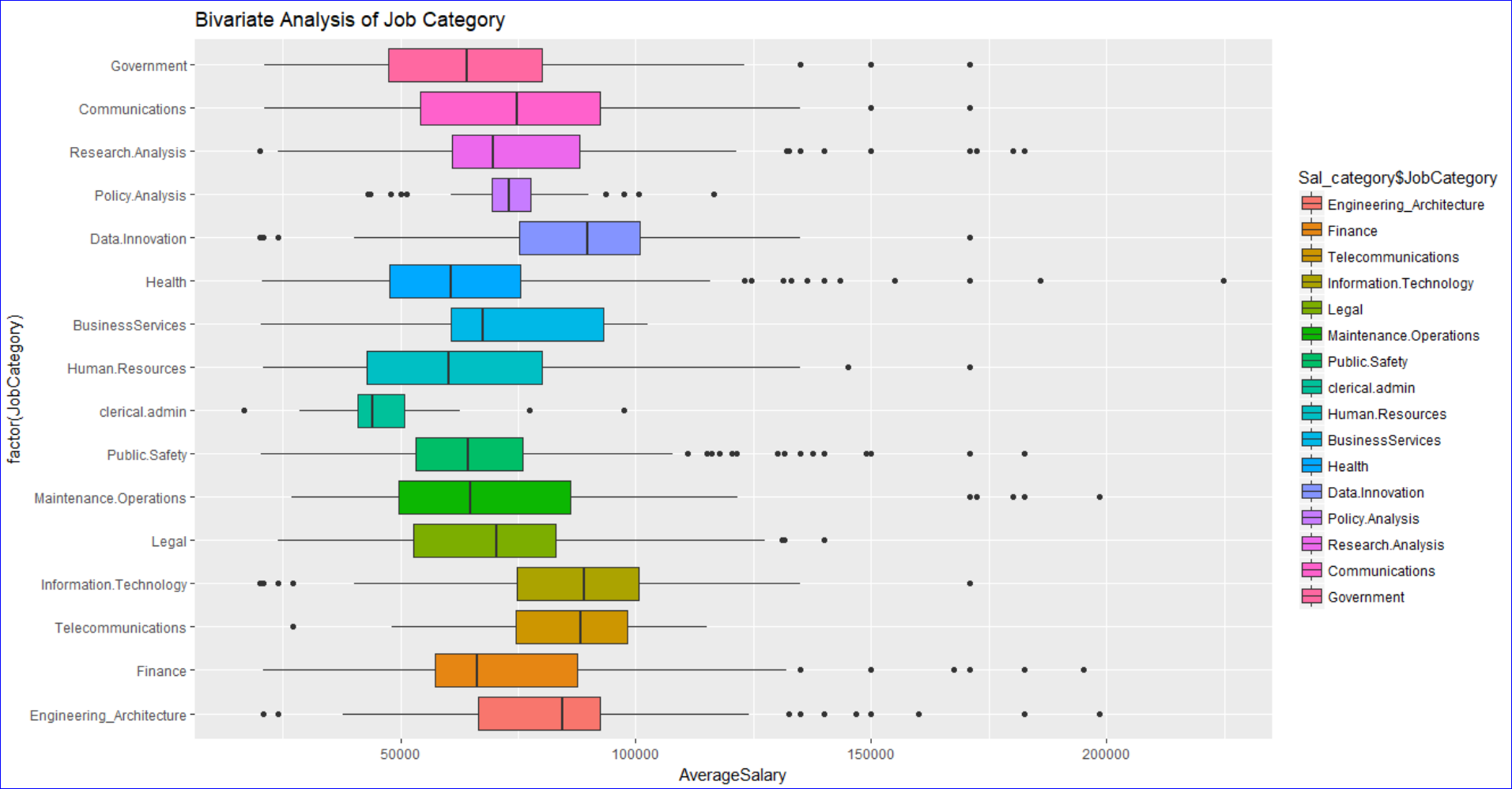


The graph shows the count of the number of jobs offered based on the residency requirement criteria. From the proportion table, we can observe that jobs offered to people with NYC residency is almost double as compared to people who don’t have NYC residency.

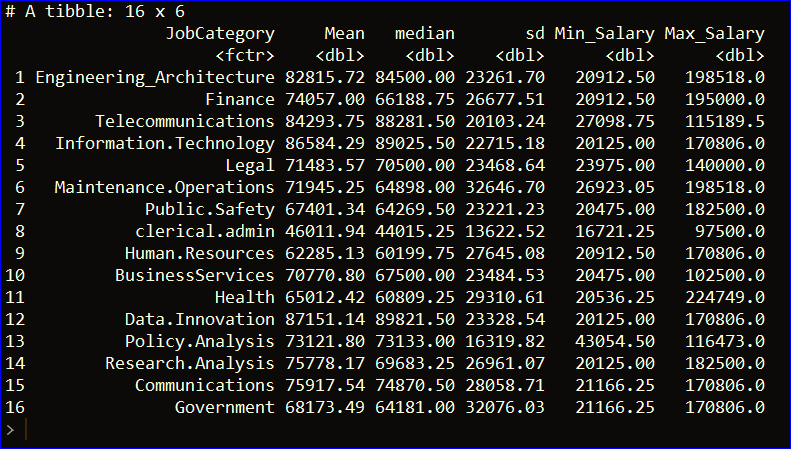


## Bi-Variate Analysis

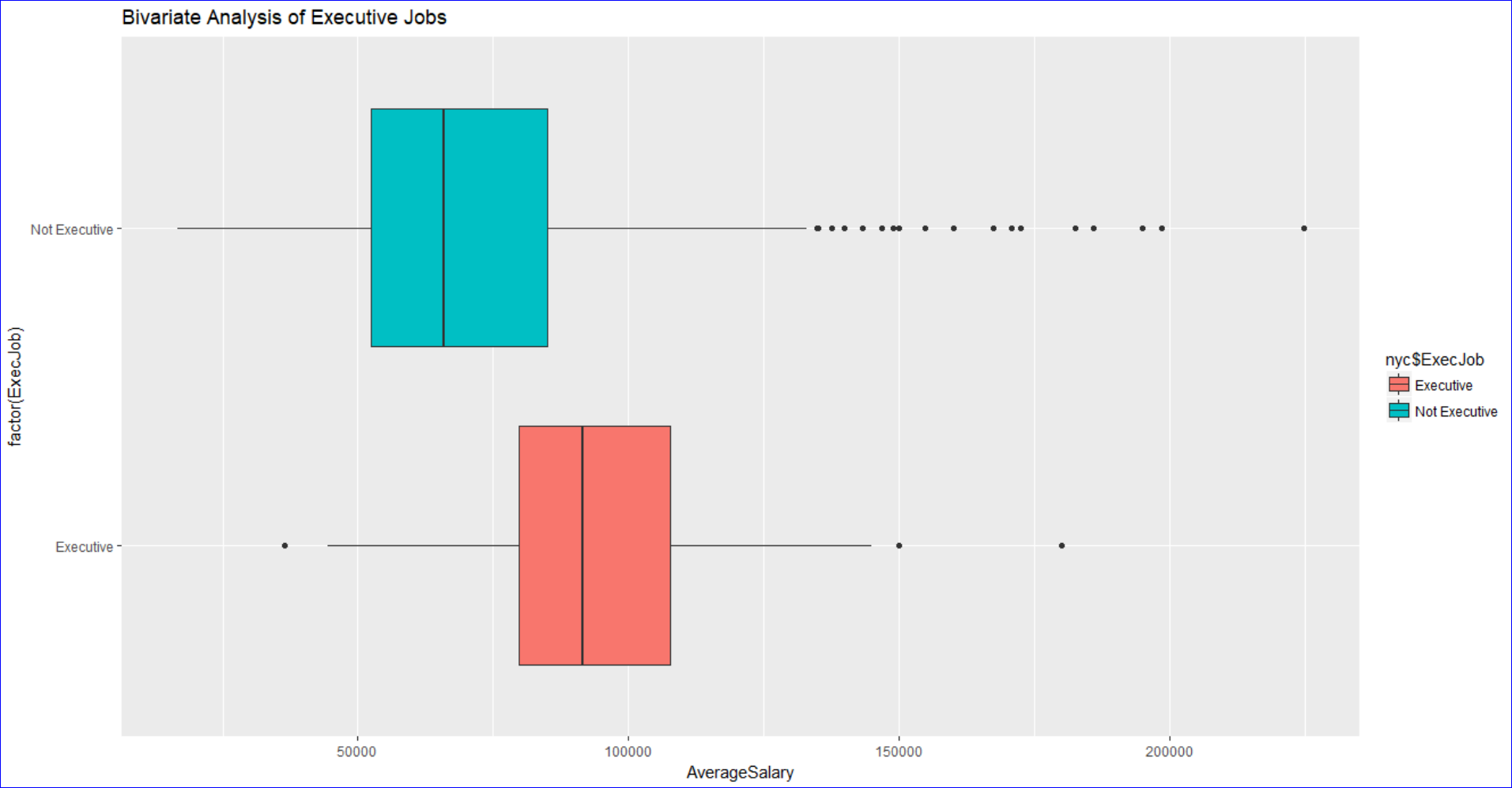
### Job category



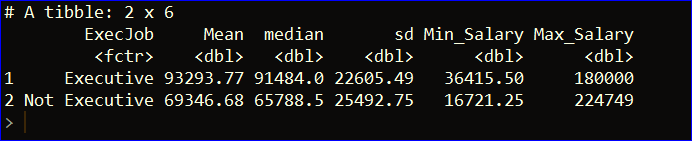
Above figure depicts the distribution of each job category. We can infer that Telecommunications, Information Technology and data innovations jobs offer higher median average salary as compared to other job categories. Policy Analysis jobs has the least range suggesting that overall salaries for Policy Analysis jobs are almost same while Human resources, maintenance operations and communications seem to have dominant ranges suggesting that different amount of salary is being offered in these areas. We can also observe the patterns and outliers for the job categories.



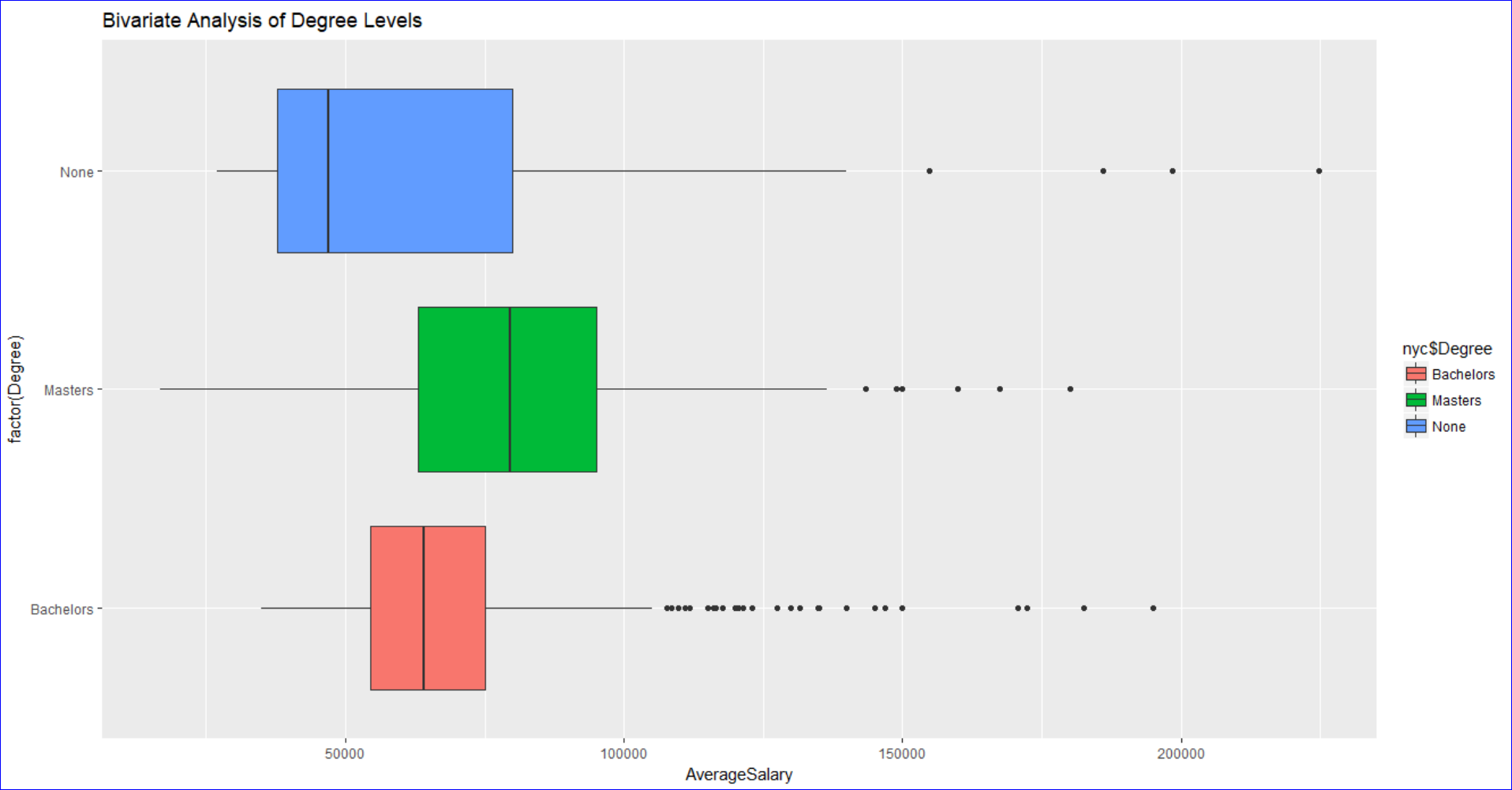
### ExecJob



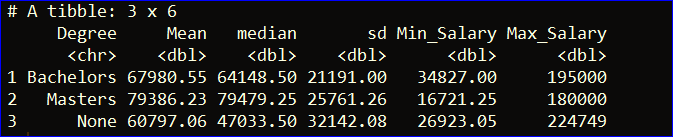
Above plot shows the distribution of average salary of Executive and Non-Executive. We can infer from the above plot that the median Average salary of Executive Jobs is higher than the Median average salary of the non-executive jobs. We can also observe that the range of Non-Executive jobs is a bit more as compared to the Executive jobs. We see some outliers for Executive jobs and the number of outliers for Non-Executive jobs is significantly more than the Executive jobs. Below is the statistical analysis of both Executive and Non-Executive jobs.



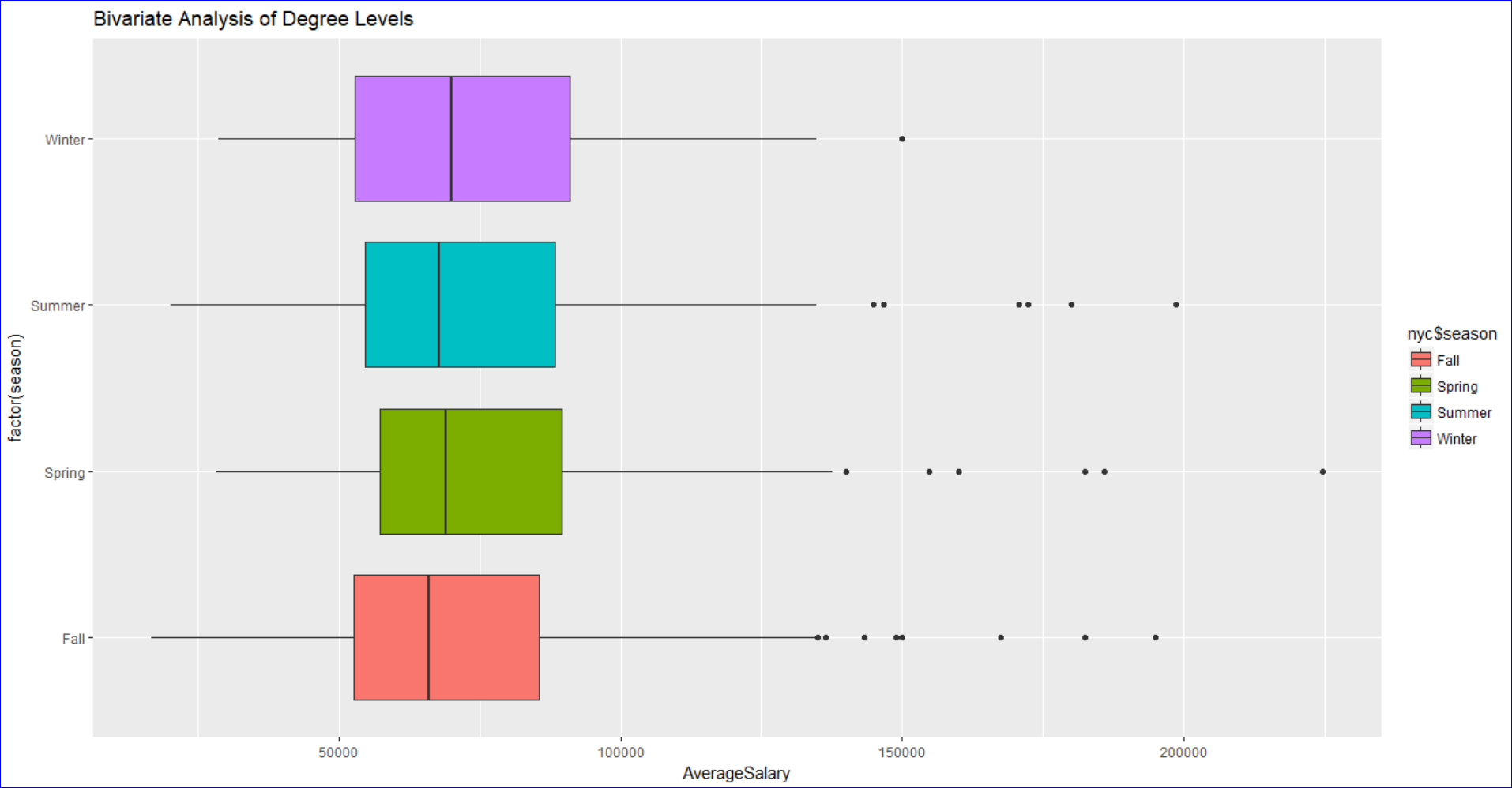
### Degree levels



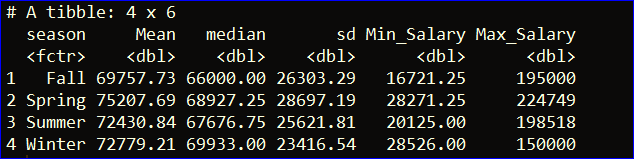
Above is the distribution of Average salary of the Degree variable. We have Bachelors, Masters and none group under the Degree variable. We can observe that the range of jobs offered for candidates with no degree is more followed by jobs offered for candidates with master’s and bachelor’s degree. We can see that the median of the average salary offered for the jobs with minimum degree requirement of master’s is higher as compared to bachelors and candidates with no degree at all. Below is the statistical analysis for each category.



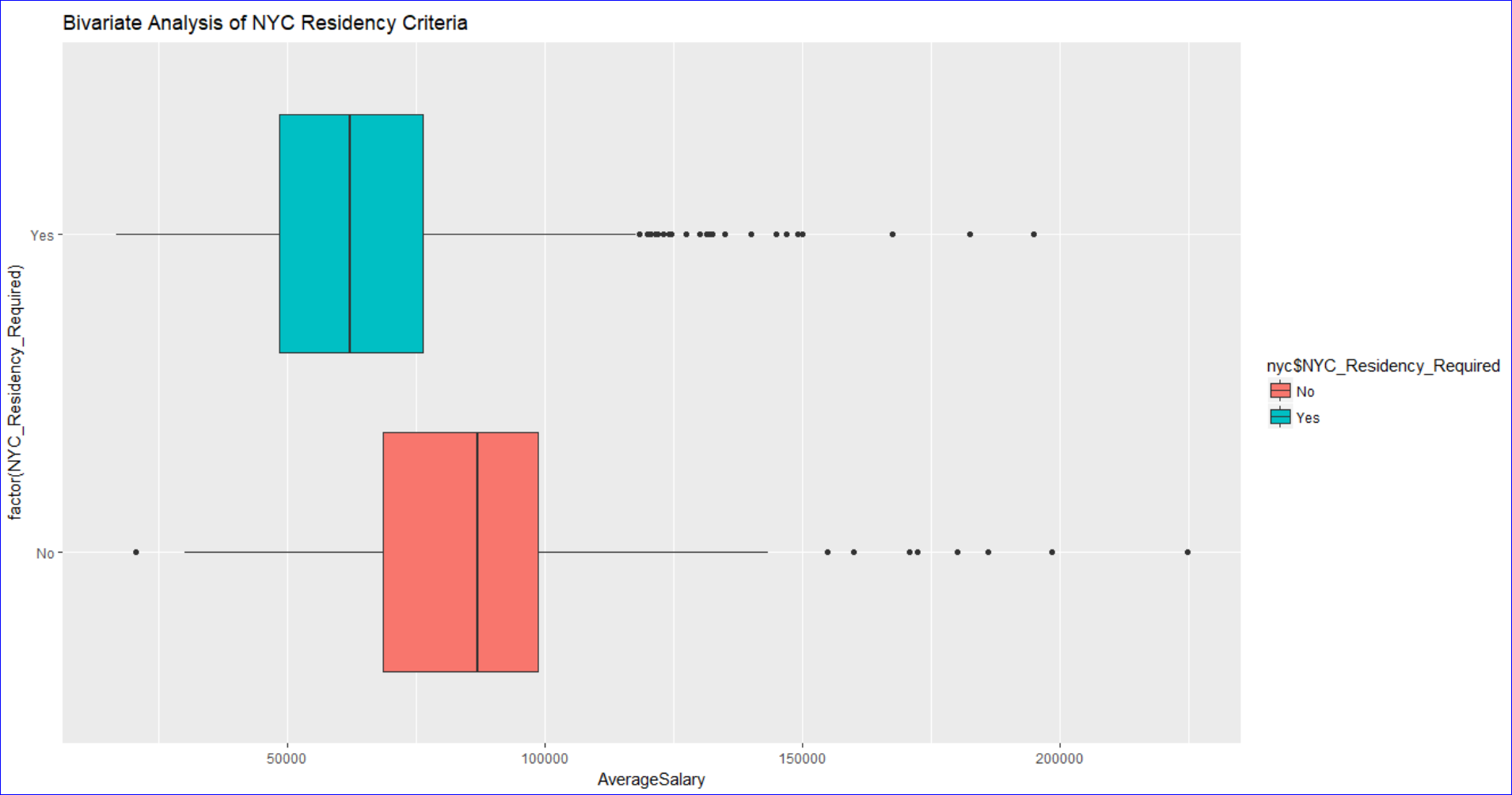
### Season



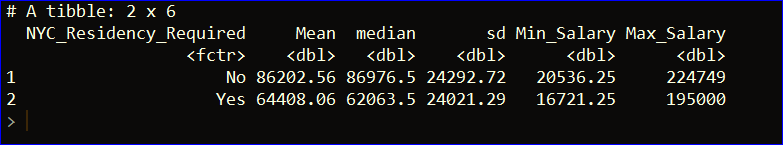
Above is the distribution of average salary offered in each season. We have Fall, Summer, Spring and winter as levels under season variable. We can observe that median of average salary for each of these season is almost the same. However, we see some outliers for each season. Below is the statistical analysis for each group of the season variable.



### NYC Residency Criteria



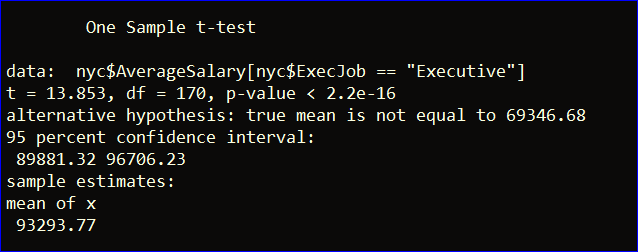
Above is the distribution of average salary for jobs which have the NYC residency requirement. We can observe that the median of average salary for jobs which doesn’t have NYC residency requirement is more as compared to the jobs which require the candidate to be a NYC resident. We also observe more outliers for the jobs which require the candidate to be a NYC resident. There is no significant difference in the range of average salary of both these groups. Below is the statistical analysis for each group for comparison.



## Hypothesis testing

### NULL HYPOTHESIS 1

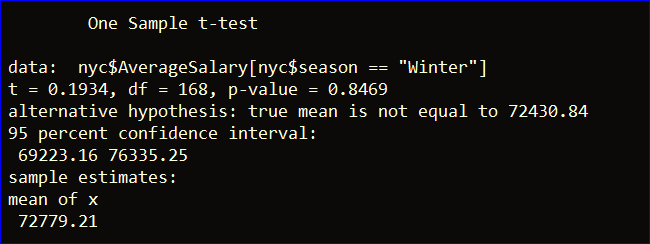
The mean of ‘Average Salary’ offered for Executive Jobs Is Same as that of Non-Executive Jobs.



Based on the above t-test finding, since the p value (2.2e-16) is less than 0.05 for 95% confidence interval, we can reject the null hypothesis. Thus, we can conclude that there is a significant difference between the means of average salary of executive and non-executive jobs in New York City.

### NULL HYPOTHESIS 2

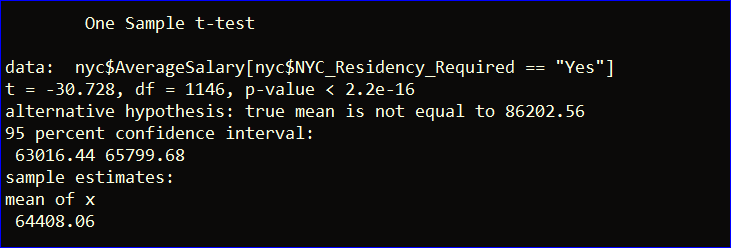
The mean of ‘Average Salary’ for jobs offered in Summer is Same as that of Winter.



Based on the above t-test finding, since the p value (0.8469) is greater than 0.05 for 95% confidence interval, we fail to reject the null hypothesis. Thus, we can conclude that there is no statistical substantial difference between seasons in New York City job postings.

### NULL HYPOTHESIS 3

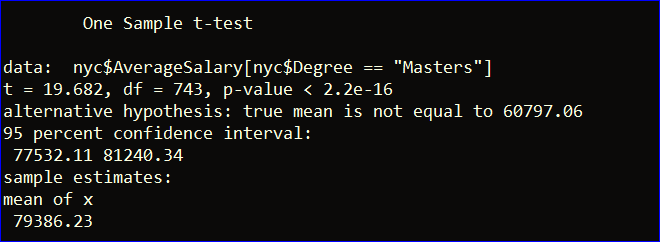
The mean of ‘Average Salary’ for jobs offered to NYC Residents is same as that of Non-Residents.



Based on the above t-test finding, since the p value (2.2e-16) is less than 0.05 at 95% confidence interval, we can reject the null hypothesis over alternative hypothesis. Thus, we can conclude that there is statistical significance in residency requirement for jobs in New York City.

### NULL HYPOTHESIS 4

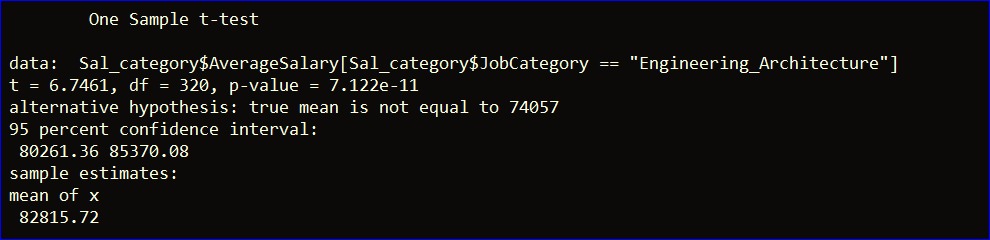
The mean of ‘Average Salary’ for the jobs with Master’s Degree is same as that of jobs with no Master’s Degree.



Based on the above t-test finding, since the p value (2.2e-16) is less than 0.05 at 95% confidence interval, we can reject the null hypothesis over alternative hypothesis. Thus, we can conclude that there is statistical significance between degree level for jobs in New York City.

### NULL HYPOTHESIS 5

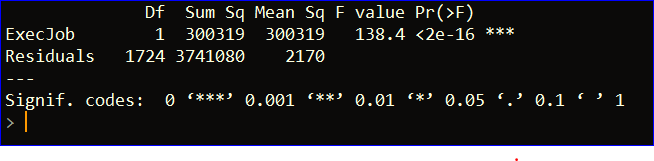
The mean of ‘Average Salary’ for Engineering jobs is same as that of Finance jobs.



Based on the above t-test finding, since the p value (7.1e-11) is less than 0.05 at 95% confidence interval, we can reject the null hypothesis over alternative hypothesis. Thus, we can conclude that there is statistical significance between job categories in New York City.

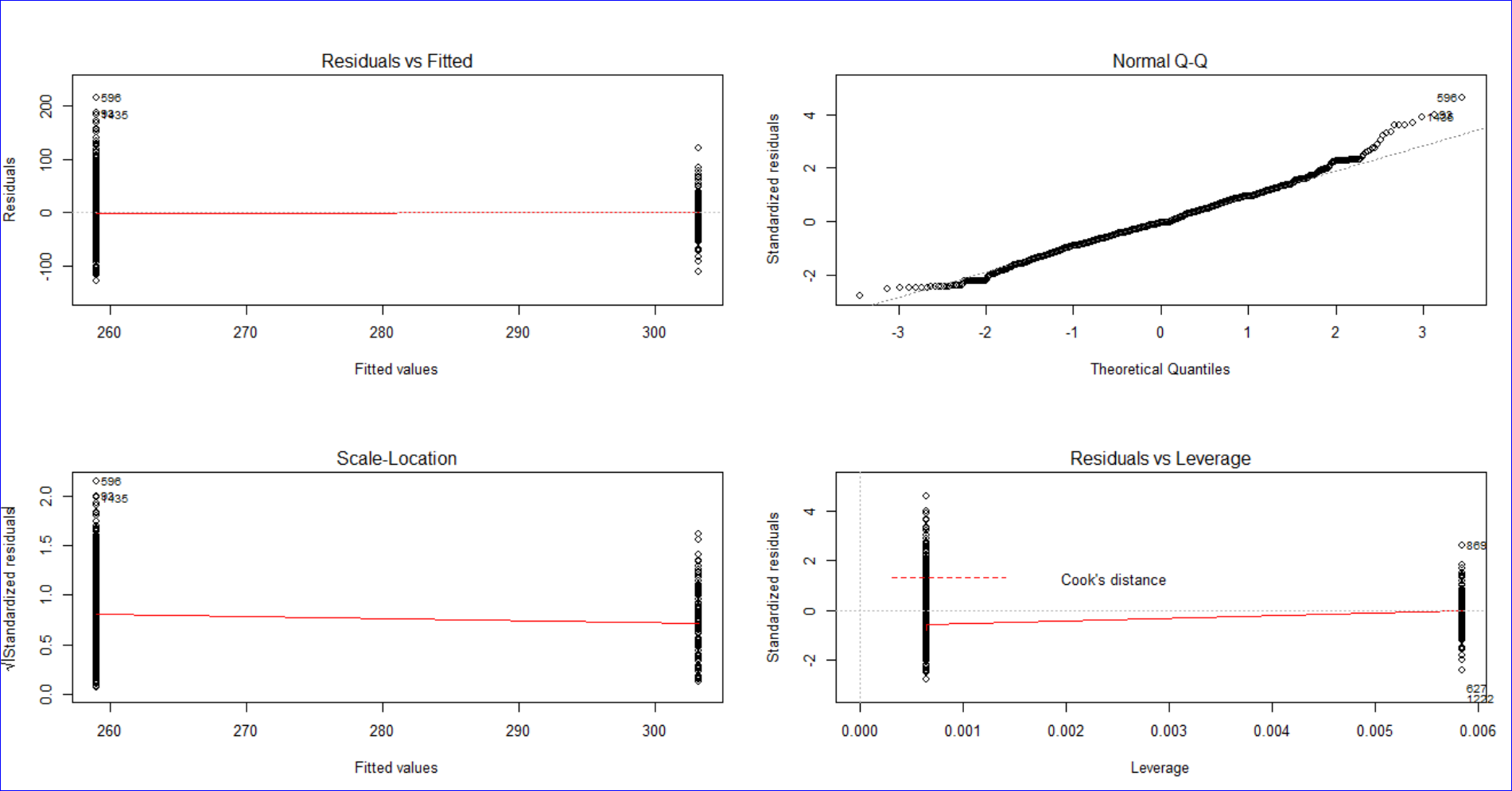
## ANOVA TESTING

### EXECUTIVE JOBS

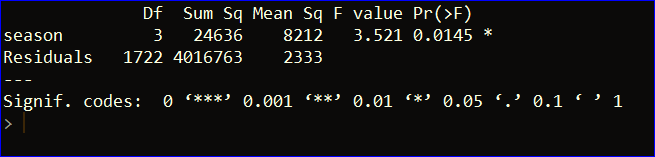


The above snippet shows that the p-value obtained is less than 2e-16 which is really low and hence signifies that the Exec job variable is statistically significant and can be included in the linear regression model.

#### Diagnostic plots

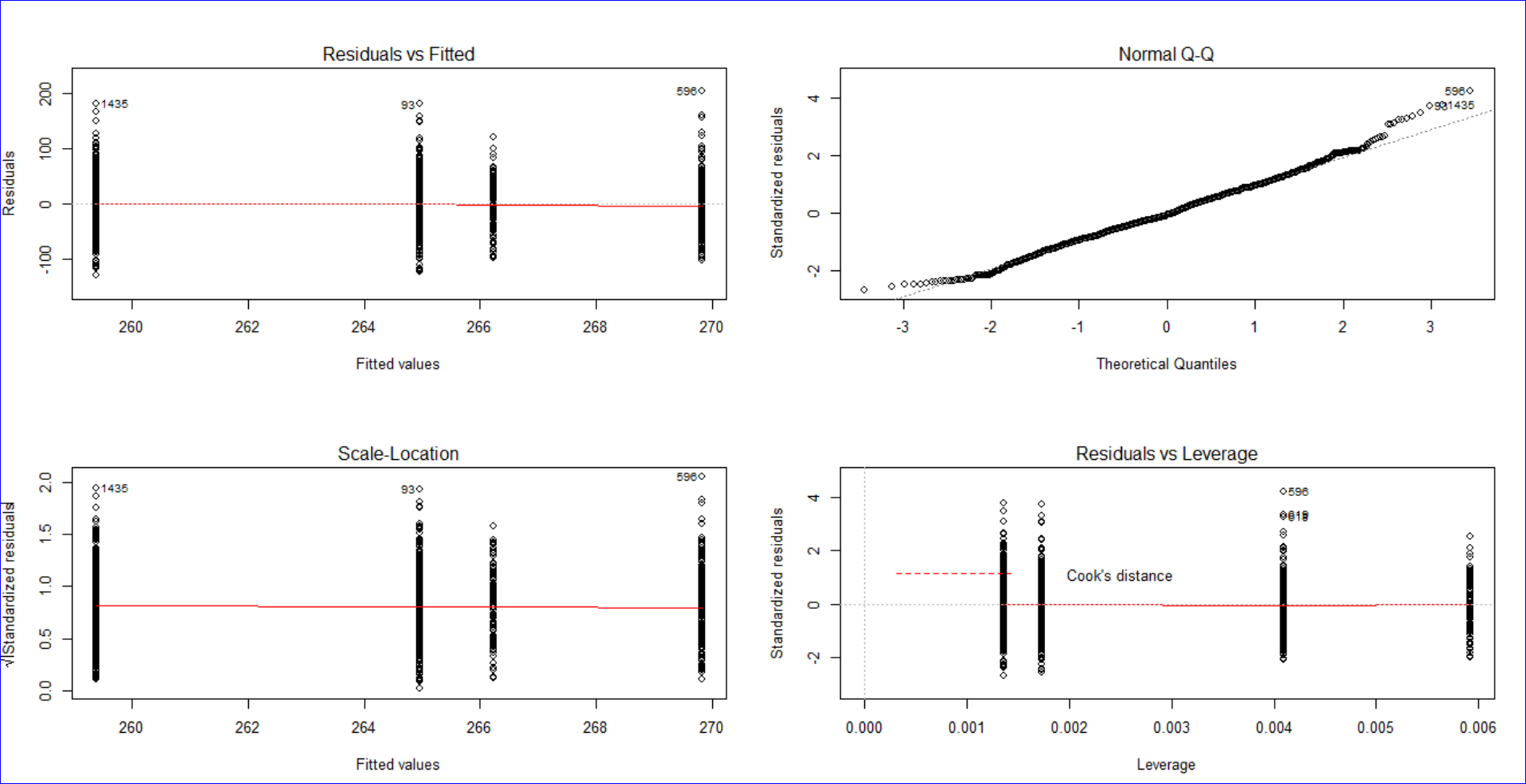


### SEASON

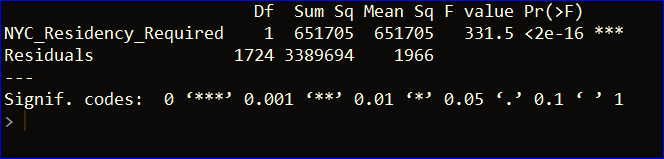


The above snippet shows that the p-value obtained is 0.0145 which is quite low and hence signifies that the variable season is not statistically significant and cannot be included in the linear regression model.

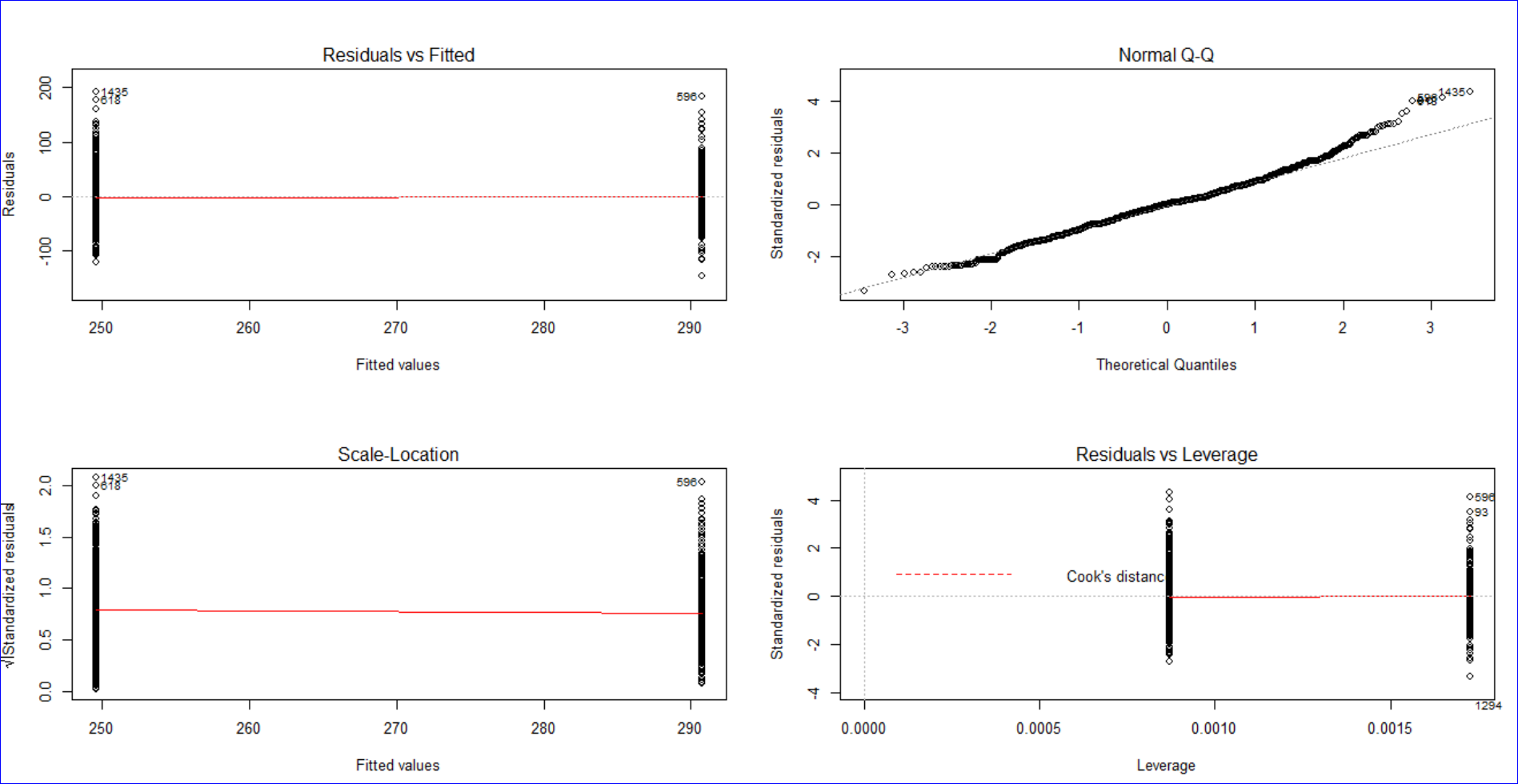
#### Diagnostic plots



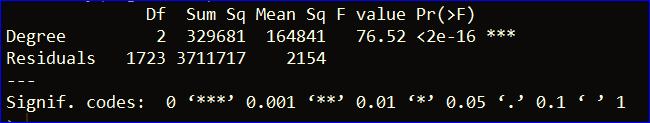
### RESIDENCY REQUIREMENT

  
The above snippet shows that the p-value obtained is less than 2e-16 which is really low and hence signifies that the NYC residency required variable is statistically significant and can be included in the linear regression model.

#### Diagnostic plots

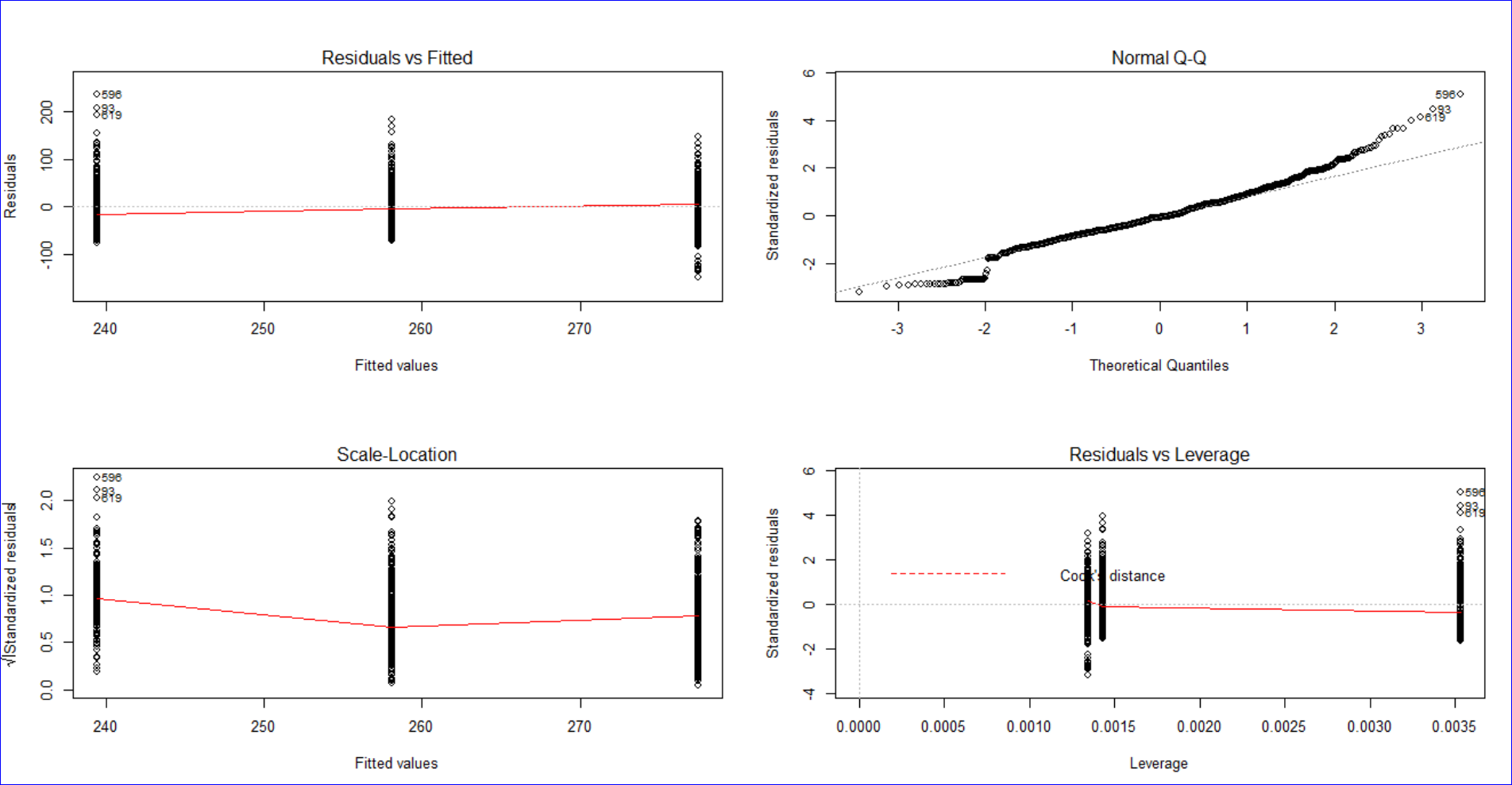


### Degree

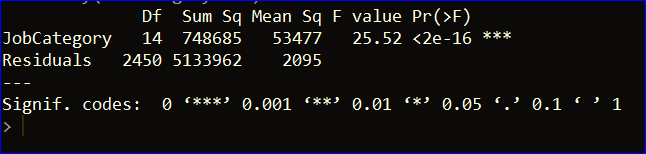


The above snippet shows that the p-value obtained is less than 2e-16 which is really low and hence signifies that the variable Degree is statistically significant and can be included in the linear regression model.

#### Diagnostic plots

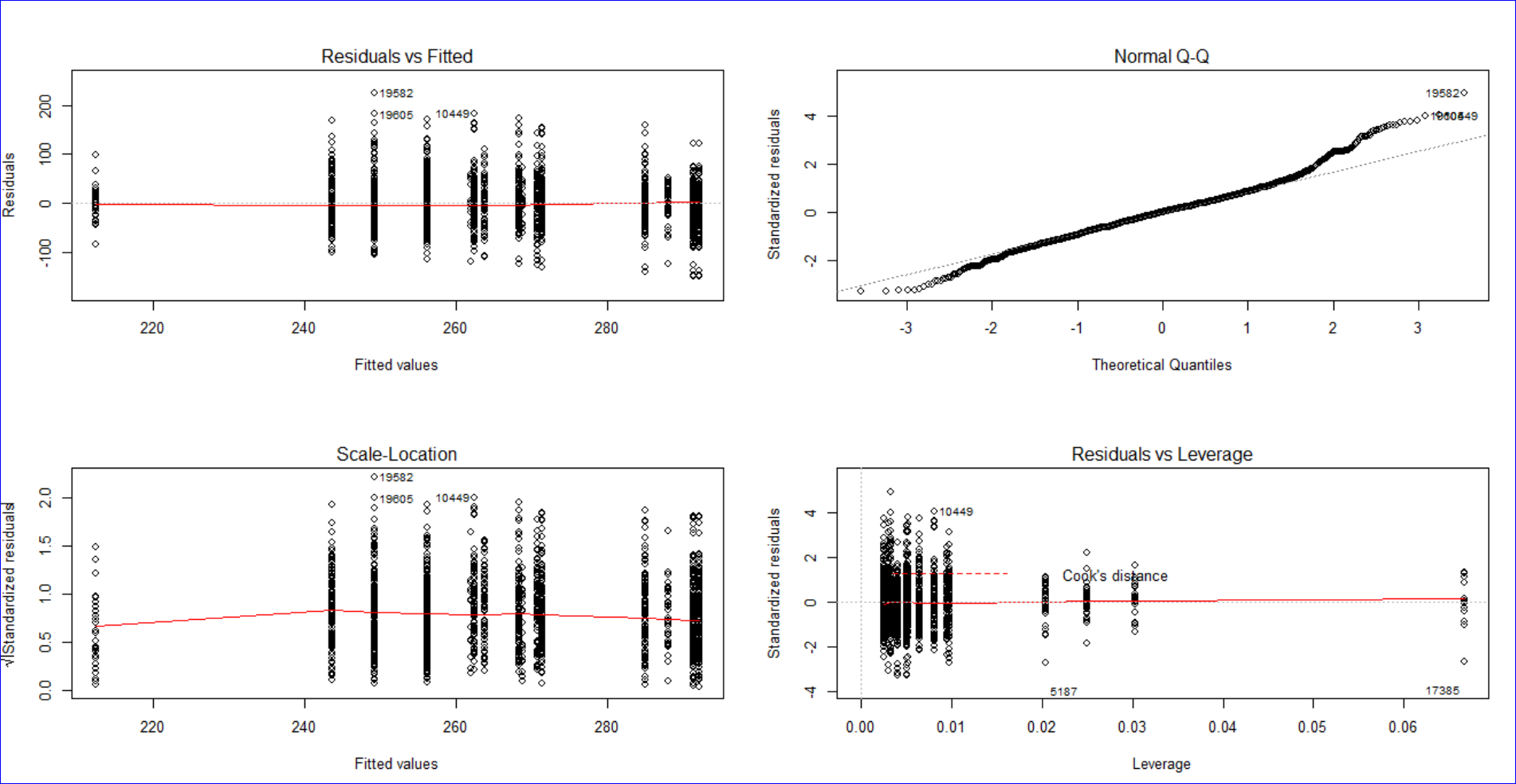


### JOB CATEGORY



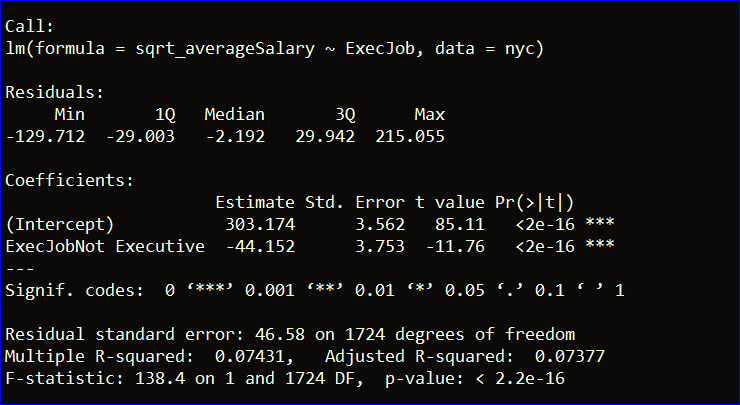
The above snippet shows that the p-value obtained is less than 2e-16 which is really low and hence signifies that the variable Job Category is statistically significant and can be included in the linear regression model.

#### Diagnostic PLOTS



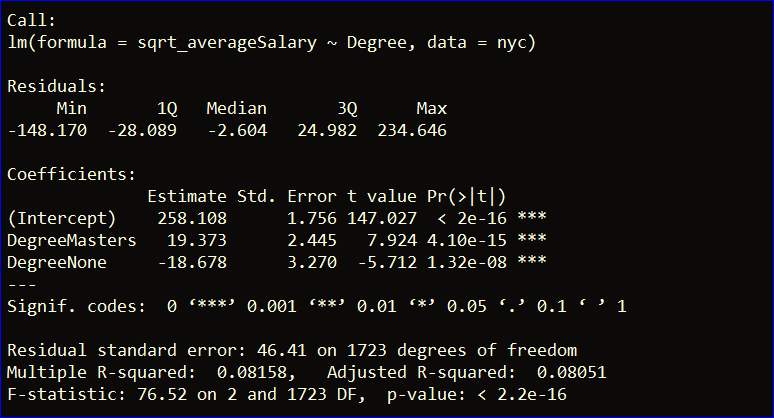
## Linear regression

### MODEL 1



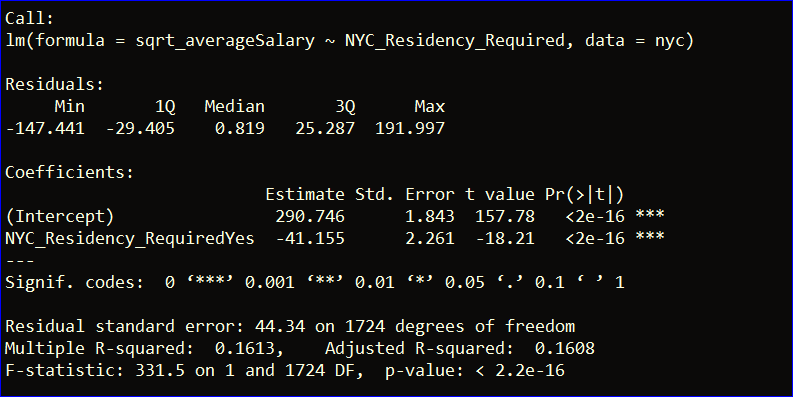
‘ExecJob’ variable indicates whether the job is an executive job or non-executive job which can help influence the average salary in a significant way. Adding this to the linear regression model has resulted in the accuracy of 7.3% which is quite low. Hence, we are proceeding to determine if other variables help to achieve higher accuracy.

### MODEL 2



‘Degree’ variable indicates the degree requirement for the job – Bachelor’s, Master’s, None. Adding this variable to the linear regression model has resulted in the adjusted R-squared value of 0.08051, which means this model can explain only 8% of the variability around its mean.

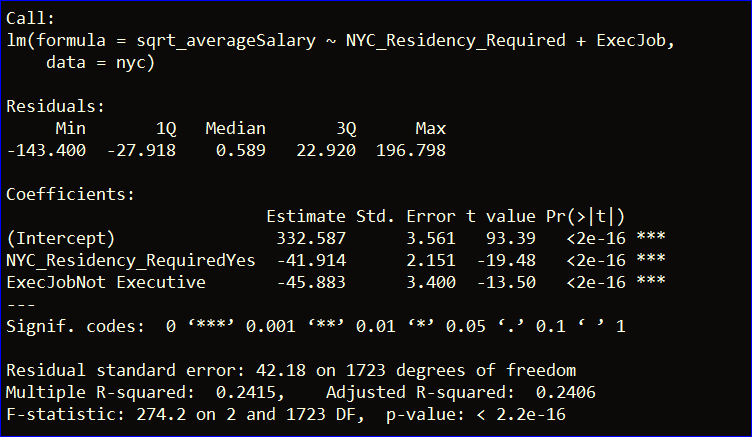
### MODEL 3



‘NYC\_Residency\_Required’ variable has been created to check the differences between the jobs being offered to residents and nonresidents. Adding this variable to the linear regression model has resulted in the adjusted R-squared value of 0.1608, which means this model can explain 16.08% of the variability around its mean.

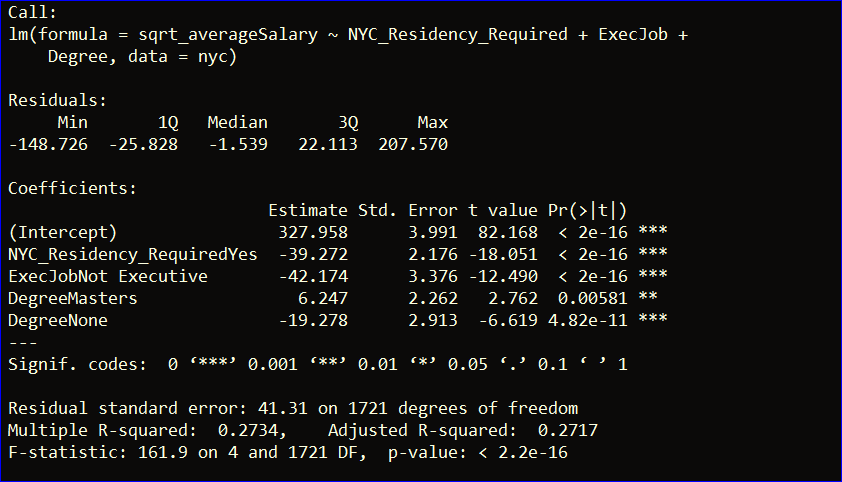
### MULTIPLE REgression

### MODEL 4



As stated in Model 1, Model 2 and Model 3, the adjusted R-squared value is quite low, which leads us to developing a multiple regression model. In this model, ‘NYC\_Residency\_Required’ and ‘ExecJob’ are being added together to build a model, resulting in an adjusted R-squared value of 0.2406, which means this model can explain 24.06 % of the variability around its mean. The highest adjusted R-squared value from linear regression is 16.08%. Switching to multiple regression has resulted in an increase of 50%.

### MODEL 5



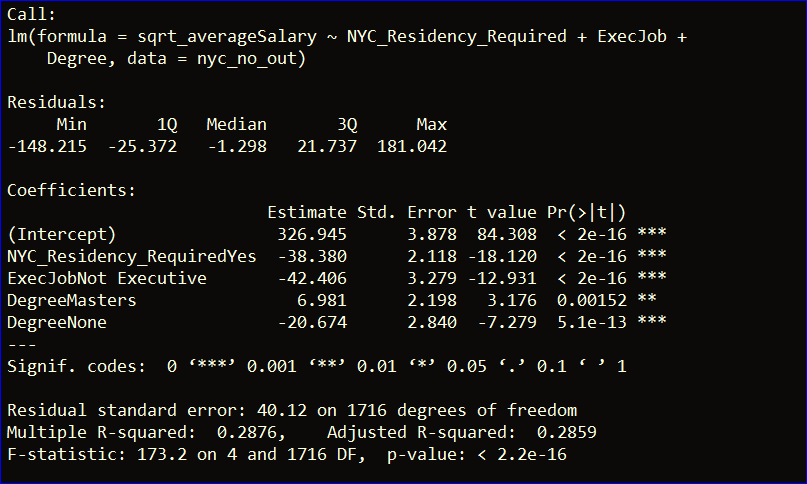
Adding two variables in Model 4 has resulted in adjusted R-squared value of 24.06%. In Model 5, the third variable ‘Degree’ is being added to the model. The model has shown a slight improvement, resulting in adjusted R-squared value of 0.2717. However, there’s still one implication that we can consider to improve the model i.e. removing the outliers.

## Removing the outliers

We have used cook’s distance to find the outliers values. We chose the top 5 observations with higher cook’s distance for model 5.

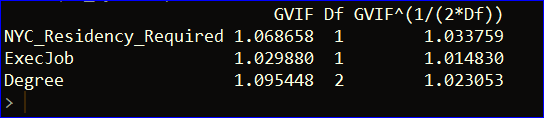
Outliers removed are: 620,162,619,93,596.

### REMODELLING



After removing the outliers, adjusted R-squared value has increased from 0.2717 to 0.2859 i.e. the increased of 5%.

## Variance Inflation factor



To check for multicollinearity, variance inflation factor has been computed. Generally, VIF should be less than 1, however there’s no need to be overly concerned if VIF if it deviated by 0.05, like in our case. The VIF computation in the above table proves that there’s no multicollinearity in the variables that we have used to build the model.

# Conclusion

After analyzing the data and building various regression models, it can be concluded that ‘NYC\_Residency\_Required’, ‘ExecJob’, ‘Degree’ drive the salary of jobs offered in New York City. With the adjusted R-squared value of 0.2859, the model explains 28.59% of the variability of the average salary around its mean.

# Future scope

* The jobs can be ranked based on how well they pay and how much they are projected to grow in the coming decade.
* The research can be extended to include the comparison of average salaries for job listings in different cities across the USA. Furthermore, we can also include what skills are in demand for specific job categories.