
# EMPLOYEE SATISFACTION SURVEY: 2010 DATASET

**BANA 6043: Statistical Computing**

**Submitted by:**

Sai Armandala
Nidhi Yaduvashi
Yiqui Wang
Sriya Boyapalli

# Table of Contents

# Abstract

The identification of Job Satisfaction Index and Employee Engagement Index is very crucial for any organization. Employee engagement is the level of commitment and involvement an employee has towards his/her organization and its values. Job satisfaction helps to know if an employee likes his/her job or individual aspects of work such as nature of work, his/her team at work. The indices together help an organization to know about their employees and in turn develop different strategies to improve the relationship with them. However, the question remains how to measure the two indices and what to infer from them? This report intends to analyze, measure and find if the two mentioned indexes are correlated. The project report has considered top 5 (by employee proportion) governmental agencies. The two indexes were developed mathematically by considering the responses to few questions from the Employee Satisfaction Survey 2010 data. The report also discusses univariate analysis of all the factor variables in the survey data, bivariate analysis between the two built indexes and factor variables and the hypothesis testing

# Introduction

Employee Satisfaction Survey was conducted in the year of 2010. Over 260,000 federal government employees participated in the survey. A number of questions were asked to each employee of various government departments. Employees were required to give his/her view by selecting the magnitude of agreement or disagreement on the scale of 1 to 5 in the options given in the questions. The survey also accepts certain general information of the employee, like the sex of the employee or the supervisory status of the employee etc. The survey helps in understanding various trends that are prevalent in different departments. The survey answers myriad questions but we have chosen few particular variables in calculating Employee Engagement Index and Job Satisfaction Index.

- **Employee Engagement Index** tells the company the extent to which the employees are passionate about their job and how enthusiastic are they about coming back to their office the next day.
- **Job Satisfaction Index** tells the company how satisfied are employees with their job. It helps the company to know if the employee is planning to leave the company or stick with the company for a longer time.

Depending upon the calculated indexes, we answer the two questions that the government wants to know. Below are the research questions:

- If there is any relationship between Job Satisfaction and Employee Engagement Index?
- Does the Employee Engagement Index depend upon Location, Supervisory Status, Sex and Age Group?

To answer the above two research questions, we have formulated the below hypothesis.

- Job satisfaction index (JSI) is positively related with Employee Engagement Index (EMPINDEX).
- There is a difference in the average EMPINDEX between Males and Females.
- There is a difference in the average EMPINDEX between Age Groups.
- There is a difference in the average EMPINDEX between employee located in Headquarters and Field.
- There is a difference in the average EMPINDEX between employees of different supervisory status.

# Data Manipulation

## 1. Exploring the Dataset and Selecting the Variables

From the 72 odd questions asked in the survey we have to select the questions that could help us with our analysis. We have selected 20 questions from the set of 72 questions that helped us with the formulation of two indices.
To formulate the Employee Engagement Index we analyzed the data and chose 8 questions.

Below were the questions for Employment Engagement Index:
Q3. I feel encouraged to come up with new and better ways of doing things.
Q4. My work gives me a feeling of personal accomplishment.
Q6. I know what is expected of me on the job.
Q11. My talents are used well in the workplace.
Q47. Supervisors/team leaders in my work unit support employee development.
Q48. My supervisor/team leader listens to what I have to say.
Q53. In my organization, leaders generate high levels of motivation and commitment in the workforce.
Q56. Managers communicate the goals and priorities of the organization.

To formulate the Job Satisfaction Index we selected 7 questions.

Below were the questions used for Job Satisfaction Index formulation:
Q4. My work gives me a feeling of personal accomplishment.
Q5. I like the kind of work I do.
Q13. The work I do is important.
Q63. How satisfied are you with your involvement in decisions that affect your work?
Q67. How satisfied are you with your opportunity to get a better job in your organization?
Q69. Considering everything, how satisfied are you with your job?
Q70. Considering everything, how satisfied are you with your pay?

## 2. Data Cleansing

- **Removing Invalid Value**
  After eliminating the invalid values, the data set we are considering for analysis has 263475 observations and 20 variables, but this data set requires cleansing as there are many values denoted by "X" which will hamper our analysis, We worked on the same and after eliminating "X", we got 71482 observations and 20 variables.

- **Changing values to correct data type:**
  Our Analysis required all the variables to be in integer format, however few of the variable namely Q47, Q11, Q13, Q53, Q56 were in factor data format, so we converted them to integer format.

- **Giving meaningful names to the variable entries:**
  Almost all the entries in the variables namely DLOC, DSEX, DSUPER, DAGEGRP, DLEAVING were assigned values like "A", "B", "C" etc. which were not very

informative as to what the entities denoted, so we substituted the below mentioned values.

DLOC variable containing value “A” substituted to “HQ”
DLOC variable containing value “B” substituted to “F”
DSEX variable containing value “A” substituted to “M”
DSEX variable containing value “B” substituted to “F”
DSUPER variable containing value “A” substituted to “Non Supervisor/Team Leader”
DSUPER variable containing value “B” substituted to “SUPERVISOR”
DSUPER variable containing value “C” substituted to “MANAGER/EXECUTIVE”
DAGEGRP variable containing value “B” substituted to “29 and Under”
DAGEGRP variable containing value “C” substituted to “30-39”
DAGEGRP variable containing value “D” substituted to “40-49”
DAGEGRP variable containing value “E” substituted to “50-59”
DAGEGRP variable containing value “F” substituted to “60 or Above”
DLEAVING variable containing value “A” substituted to “N” as in No
DLEAVING variable containing value “B” substituted to “R” as in retired
DLEAVING variable containing value “C” substituted to “Y” as in Yes, to take another job within the Federal Government
DLEAVING variable containing value “D” substituted to “Y” as in Yes, to take another job outside the Federal Government
DLEAVING variable containing value “E” substituted to “Y” as in Yes, other
All the unused levels, which were created while converting variables to integer, were dropped

## 3. Data Selection & Addition

Based on the maximum employee count, we selected top 5 departments and began our analysis.
The top 5 departments with the maximum number of employees are:

1. Army (Army)
2. Agriculture (AG)
3. Commerce (CM)
4. Justice (DJ)
5. Treasury (TR)

After selecting these Departments we got 71,482 observations and 20 variables.
Our next important step is to add suitable variables to our Data set, Hence the addition of Job Satisfaction Index (JSI) & Employment Engagement Index (EmpIndex). Our analysis is based on these two indices.

# Data Analysis

## 1. Univariate Analysis

**a) Agency**

We identified the top 5 governmental agencies by employee proportion. The top 5 agencies were Army, Agriculture, Commerce, Justice and Treasury. On plotting a 3D pie chart, we found that Army is the biggest agency (23.5% equivalent to 16816) out of the top 5 agencies. Agriculture had the minimum number of employees (17.2% equivalent to 12263) out of the top 5 agencies.

Ref. Plot 1: % Distribution by top 5 agencies

**b) Location**

We identified that an employee can be working at either a Field location or Headquarters. On having a summary statistics on the location variable, we found that there are more number of employees working in Field (40158) as compared to Headquarters (31324).

Ref Plot 2: % Distribution by Location

**c) Supervisory Status**

We identified that the survey data has 3 supervisory status. There are Supervisor, Non-supervisor/Team-leader and Manager/Executive. When we checked the summary statistics for this variable, we found that maximum proportion was of non-supervisor/Team-leader and there were least number of Managers/Executive.

Ref Plot 3: % Distribution by Supervisory status

**d) Sex**

On plotting a bar plot, we found that Males dominated the top 5 governmental agencies in terms of their proportion. There were 54.9% of males as compared to 45.1% of females.

Ref Plot 4: % Distribution by Sex

**e) Leaving Status**

We identified all the reason codes in the variable DLEAVING, and accordingly made three new leaving status: Y for Yes, the employee is leaving; N for No, the employee is not leaving; and R for Retire, the employee is about to retire. Further, when we plotted the bar plot we found that most proportion of employees (73%) are going to stay with the governmental agencies, 20.8% are about to leave the top 5 agencies and 6.2% of employees are about to retire.

Ref Plot 5: % Distribution by Leaving Status

**f) Age Group**

We saw that the employees in agencies had age which varies widely. Therefore, we subdivided the age into 5 ranges: 29 and under, 30-39, 40-49, 50-59 and 60 or above. From the pie chart we plotted this variable, we found that 37% of people are in the range of 50-59, which is the maximum, and only 5.6% of people are in the range of 30-39, which is the lowest.

Ref Plot 6: % Distribution by Age Group

**g) Job Satisfaction Index (JSI)**

The scaled index is an integer variable with a range from 0 to 1. JSI has a mean of 0.77 and a median of 0.8. The trimmed median is 0.78 implying less number of outliers. The standard deviation is 0.14 indicating the deviation is less. The skewness is -0.65. The skewness is high and the distribution should be left skewed. Histogram and lines plot prove the above statistics. We squared the JSI variable to reduce the skewness and improve the distribution.

Ref Plot 7: Histogram and boxplot of EmpIndex

**h) Employee Engagement Index (EmpIndex)**

The scaled index is an integer variable with a range from 0 to 1. We find that EmpIndex has a mean of 0.75 and a median of 0.78. The trimmed median is 0.76 implying less number of outliers. Additionally, the standard deviation is 0.16 indicates that the deviation is less. The skewness is -0.73 indicating high skewness and the distribution should be left skewed. Histogram and lines plot prove the above statistics. We also squared the EmpIndex variable to reduce the skewness and improve the distribution.

Ref Plot 8: Distribution of Squared EmpIndex

# 2. Bivariate Analysis

**a) JSI and Agency**

The distribution plot shows that in all the departments the maximum number of employee have JSI value around 0.77 and the least number of employees have JSI around 0.25, the plots in all the departments are slightly left skewed

Ref Plot 9: Distribution of Agency by JSI

The Plot mean graph shows the department that has highest average JSI value is Department of Justice and the department, which has the lowest JSI value, is Agriculture

Ref Plot 10: Mean JSI value by Agency

**b) JSI and Age Group**

The distribution plot shows that in all the age groups the maximum number of employees have JSI value around 0.77 and the least number of employees have JSI around 0.26 the plots in all the departments are slightly left skewed

Ref Plot 11: Distribution of Agency by Age group

The Plot mean graph shows the age group which has highest average JSI value is group of 60 and Above and the lowest JSI value is of group 29 and Under

Ref Plot 12: Mean Agency Value by Age Group

**c) JSI and Sex**

The distribution plot shows that among the two sexes the maximum number of employees have JSI value around 0.77 and the least number of employees have JSI around 0.26 the plots in all the departments are slightly left skewed.

Ref Plot 13: Distribution of JSI by Sex

The Plot mean graph shows that between the two sexes males have higher JSI value than the females.

Ref Plot 14: Mean JSI value by Sex

**d) JSI and Supervisory Status**

The distribution plot shows that in all the supervisory statuses groups the maximum number of employees have JSI value around 0.77 and the least number of employees have JSI around 0.26 the plots in all the departments are slightly left skewed
Ref Plot 15: Distribution of JSI by Supervisory status

The Plot mean graph shows the supervisory status group which has highest average JSI value is of group of Manager/Executive and the lowest JSI value is of Non-Supervisory/Team Leader
Plot 16: Mean JSI value by Supervisory Status

**e) JSI and Leaving Status**

The distribution plot shows that in all the Leaving reasons groups the maximum number of employees have JSI value around 0.77 except in "Y" where the highest is around 0.69 and the least number of employees have JSI around 0.26 the plots in all the departments are slightly left skewed
Ref Plot 17: Distribution of JSI by DLeaving factor

The Plot mean graph shows the leaving status group which has highest average JSI value is of the group No and the lowest JSI value is of the group Yes
Plot 18: Mean JSI value vs DLeaving factor

**f) JSI and EmpIndex by Pay Satisfaction levels**

The plots shows that there is high correlation between JSI and EmpIndex when satisfaction with pay is high, this means if the employees are paid better or are satisfied by their pay they are more likely to be engaged in their jobs.
Ref Plot 19: JSI and EmpIndex by Pay Satisfaction Level

**g) Agency and likeness to work with Supervisory Status**

The plot shows that the likeness to work is similar and high between the top 5 agencies in various supervisory statues.
Ref Plot 20: Agency and interest towards work with Supervisory status

**h) Location and Job satisfaction levels with supervisory status**

The plot shows the Job Satisfaction levels is similar and high between the locations in various supervisory statues.
Ref Plot 21: Location and Job Satisfaction levels with Supervisory status

# 3. Hypothesis Testing

**a) Job satisfaction index (JSI) is positively related with Employee Engagement Index (EmpIndex)**

We used correlation test to find the relation between JSI and EmpIndex. We used the cor function and got the value of 0.8411. This result indicates that JSI and EmpIndex are highly positive correlated.

**b) There is a difference in the average EmpIndex between Males and Females**

Since the scaled indexes are skewed, we used Wilcox test to identify the difference in average EmpIndex of males and females. The p value turned out to be 0.006987, which suggests that we are confident in our alternate hypothesis. The null hypothesis, which says that EmpIndex is no different for male or female, can be rejected at 99% confidence. We conclude that our alternate hypothesis statement is correct and justifiable.Ref Plot 22: Mean EmpIndex value by Sex

**c) There is a difference in the average EmpIndex between Age Groups**

We ran Anova on EmpIndex and different age groups to understand whether or not EmpIndex varies with different age. P value from the Anova test came out to be 1.37e-13 which suggests that the we are confident with our alternate hypothesis. The null hypothesis, which says that mean values in EmpIndex is no different between different age groups, can be rejected at 99% confidence. Additionally, we ran Tukey test on our Anova result to understand the EmpIndex difference between different age groups. We find that mean values for EmpIndex are not different between age groups 50-59 and 29 and under, 60 or above and 29 and under, 40-49 and 30-39, and 50-59 and 30-39. Otherwise, EmpIndex is different between other age groups.

Ref Plot 23: Mean EmpIndex value by Age groups

**d) There is a difference in the average EmpIndex between employee located in Headquarters and Field**

Since the scaled indexes are skewed, we used Wilcox test to identify the difference in average EmpIndex for employees working at Headquarters and Field location. The p value turned out to be 2.924e-15 which suggests that we are confident in our alternate hypothesis. The null hypothesis, which says that average EmpIndex is no different for employees located in Headquarters or at Field, can be rejected at 99% confidence. We conclude that our alternate hypothesis statement is correct and justifiable.

Ref Plot 24: Mean EmpIndex value by Location

**e) There is a difference in the average EmpIndex between employees of different supervisory status**

We ran Anova on EmpIndex and different supervisory status to understand whether or not EmpIndex varies with different supervisory status. P value from the Anova test came out to be 2e-16 which suggests thatwe are confident with our alternate hypothesis. The null hypothesis, which says that there is no difference in the average EmpIndex between different supervisory statuses, can be rejected at 99% confidence. Additionally, we ran Tukey test on our Anova result to understand the EmpIndex difference between different supervisory statuses. We conclude that mean values for EmpIndex are different for different supervisory statuses.

Ref Plot 25: Mean EmpIndex value by Supervisory Status

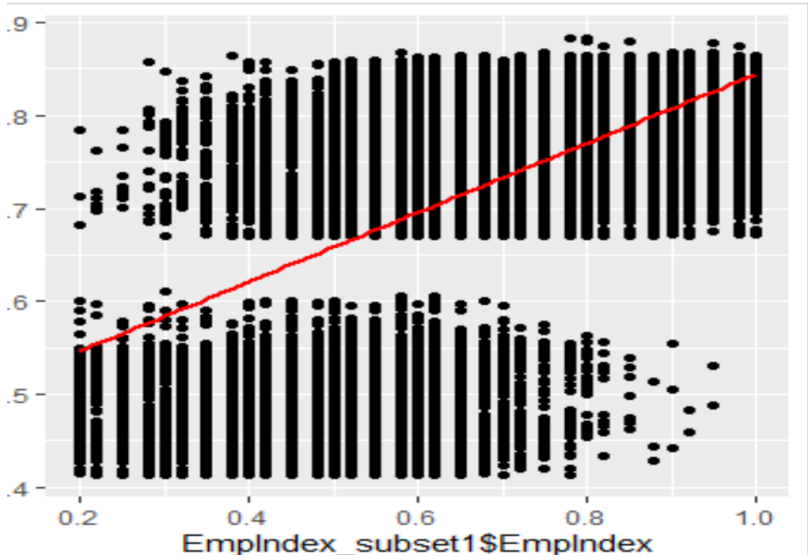
**f) There is a difference in the average EmpIndex between employees of different leaving statuses**

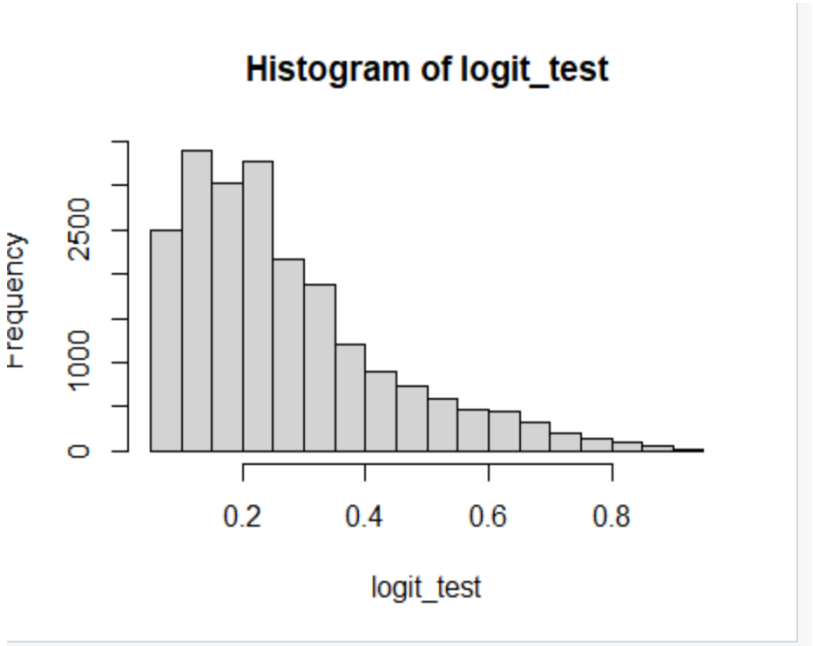
We ran Anova on EmpIndex and different leaving status to understand whether or not EmpIndex varies with different leaving status. P value from the Anova test came out to be 2e-16 which suggests that we are confident with our alternate hypothesis. The null hypothesis, which says that there is no difference in the average EmpIndex between different leaving statuses, can be rejected at 99% confidence. Additionally, we ran Tukey test on our Anova result to understand the EmpIndex difference between different leaving statuses. We conclude that mean values for EmpIndex are different for different leaving statuses.

Ref Plot 26: Mean EmpIndex value by Leaving statuses

# LINEAR REGRESSION

We have built a linear regression model on EmpIndex_subset1 and EmpIndex. We can see that the estimated coefficient for EmpIndex_subset1 is statistically significant, as the associated p-value is greater than our threshold of 0.05.

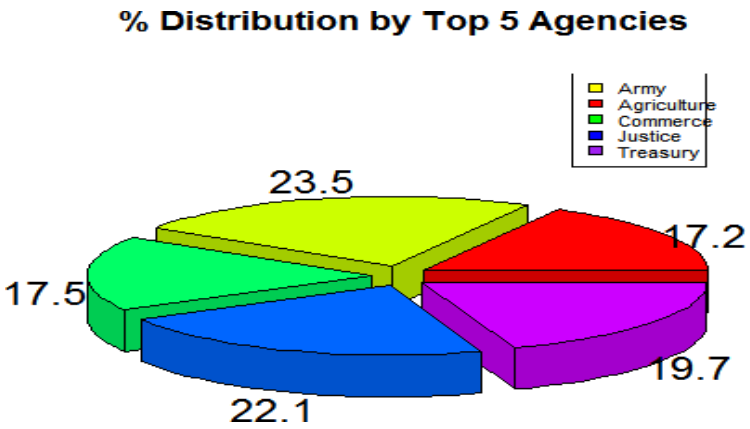


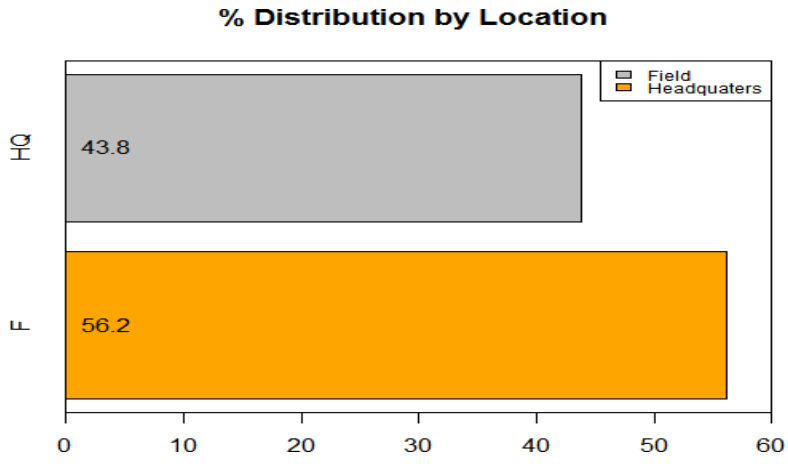
# Conclusion

The Employee Engagement Index and Job Satisfaction Index are positively correlated as proved through our correlation test result. Higher the job satisfaction, higher the employee engagement. The average EmpIndex differs with Gender, Age groups, Location, Supervisory status and Leaving status. We conclude that every alternative hypothesis is justifiable and null hypothesis can be rejected at 99% confidence. Additionally, through our bivariate analysis, we find that job satisfaction is most in department of Justice. However, Employee Engagement is highest and similar in Department of Agriculture, Justice and Treasury. Employee Engagement and Job satisfaction is also high when an employee is promoted to higher levels like Manager as compared to Non-supervisor or Team lead position. Additionally, the job satisfaction and employee engagement has been higher in case of male employee as compared to females. We also conclude that when the satisfaction with pay is high, there is a major correlation between JSI and EmpIndex. Moreover, the Job Satisfaction levels is similar and high between the locations in various supervisory statues.

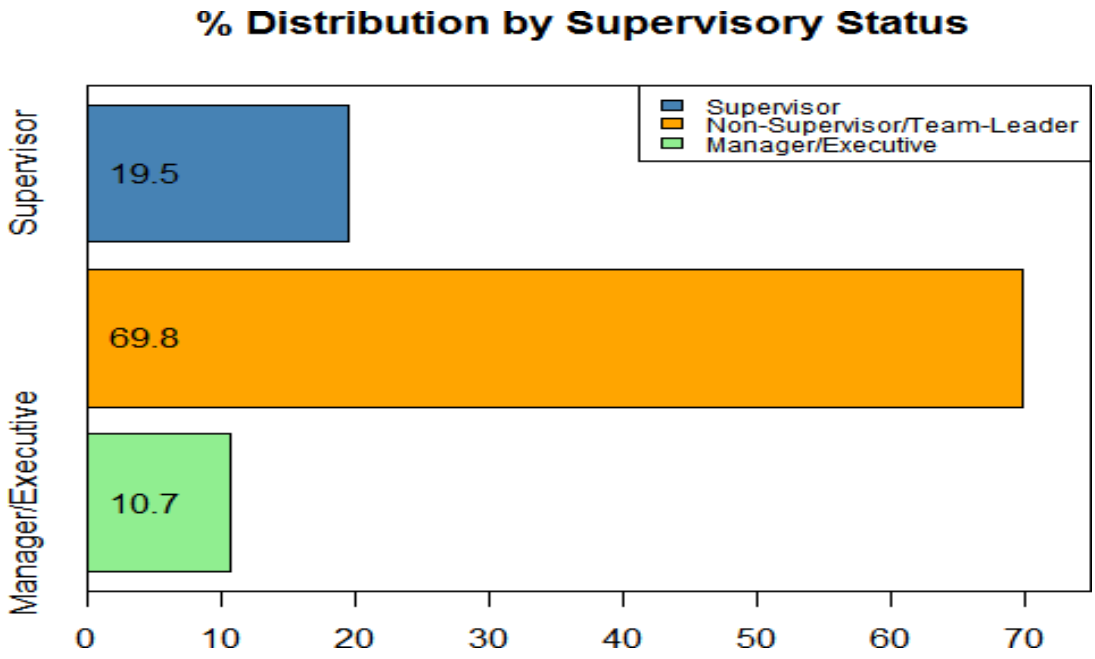
# Appendix



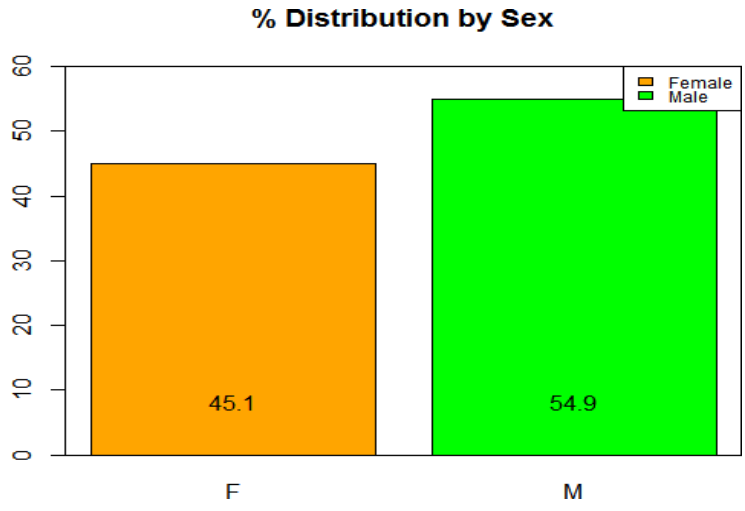

Plot 1: % Distribution by top 5 agencies



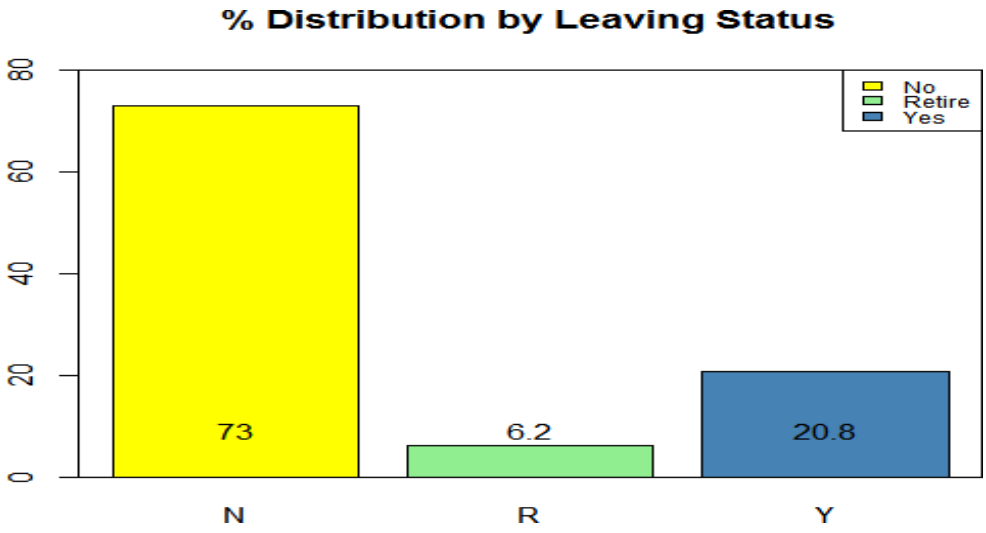

Plot 2: % Distribution by Location



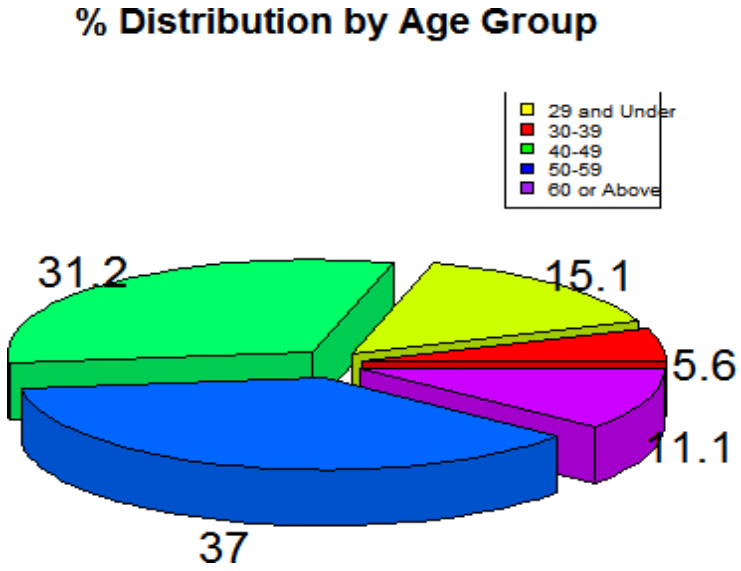

Plot 3: % Distribution by Supervisory status



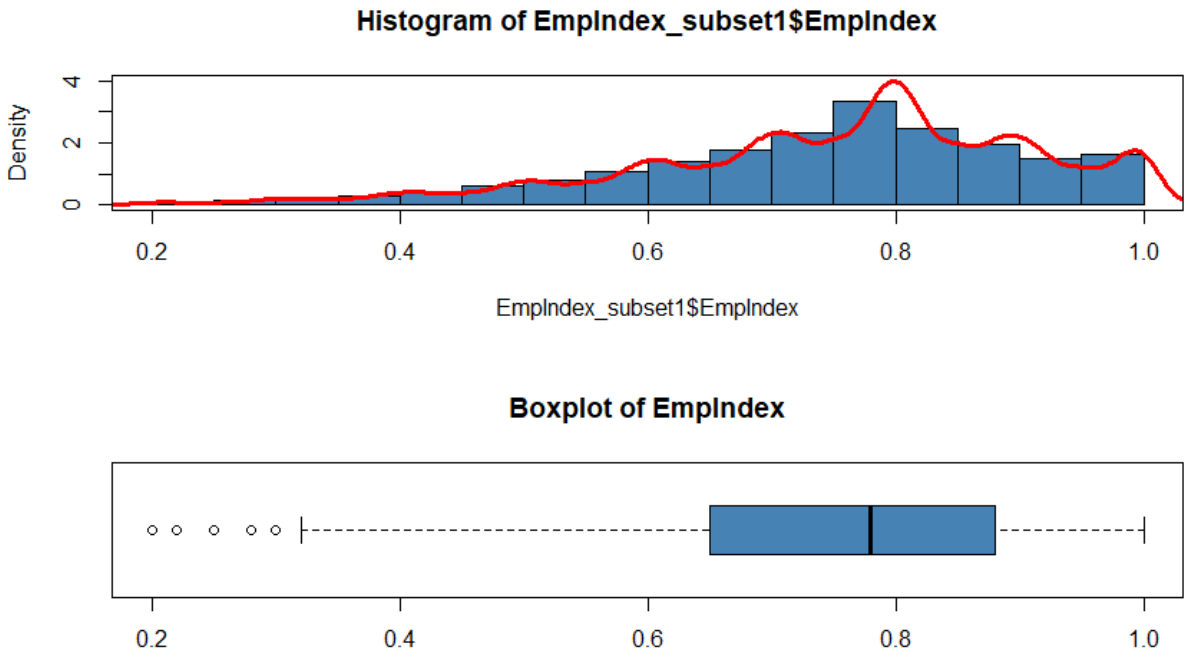

Plot 4: % Distribution by Sex




Plot 5: % Distribution by Leaving Status



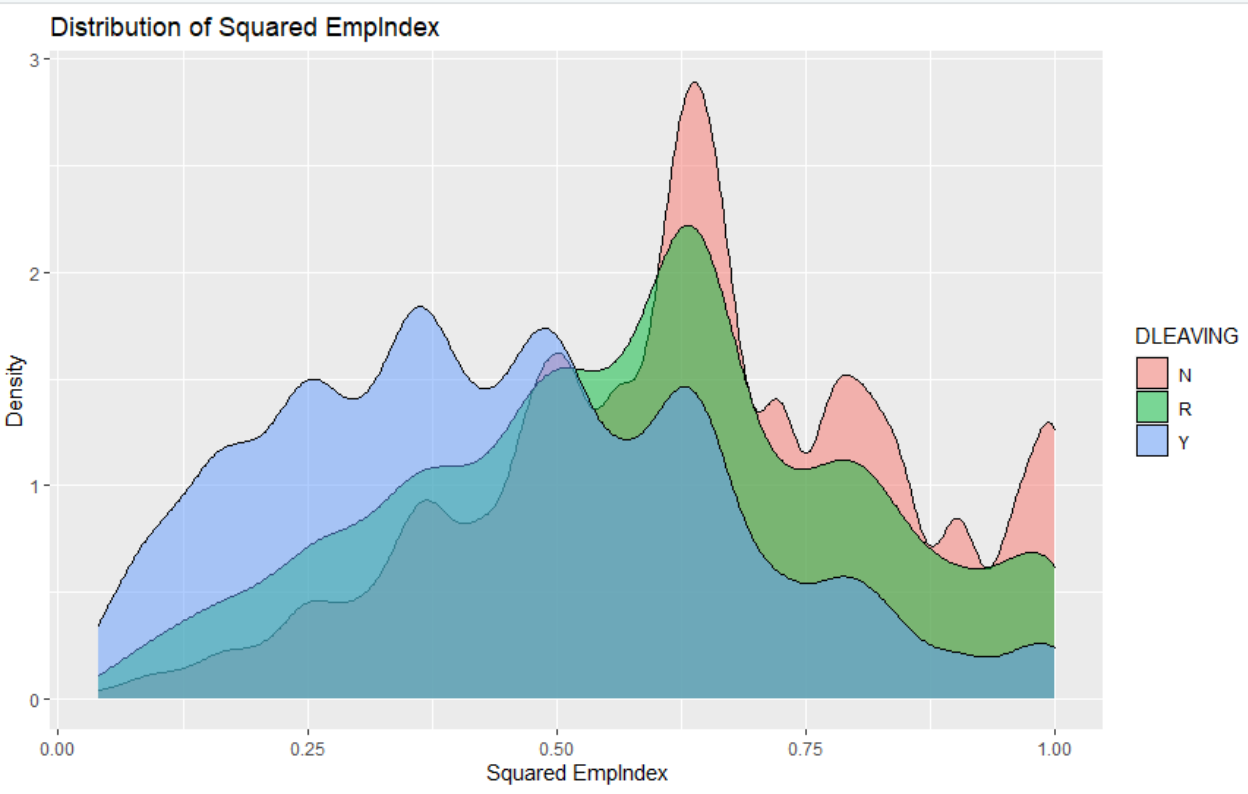

Plot 6: % Distribution by Age Group



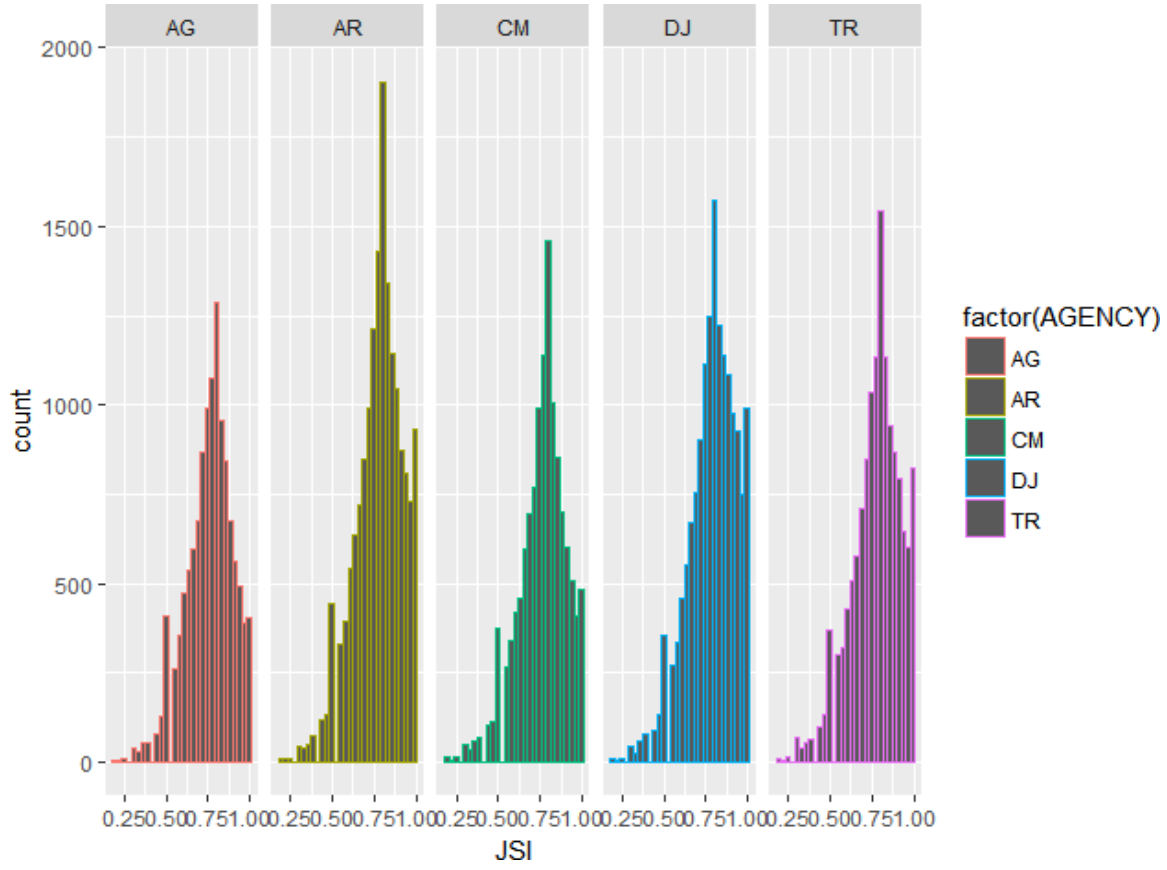




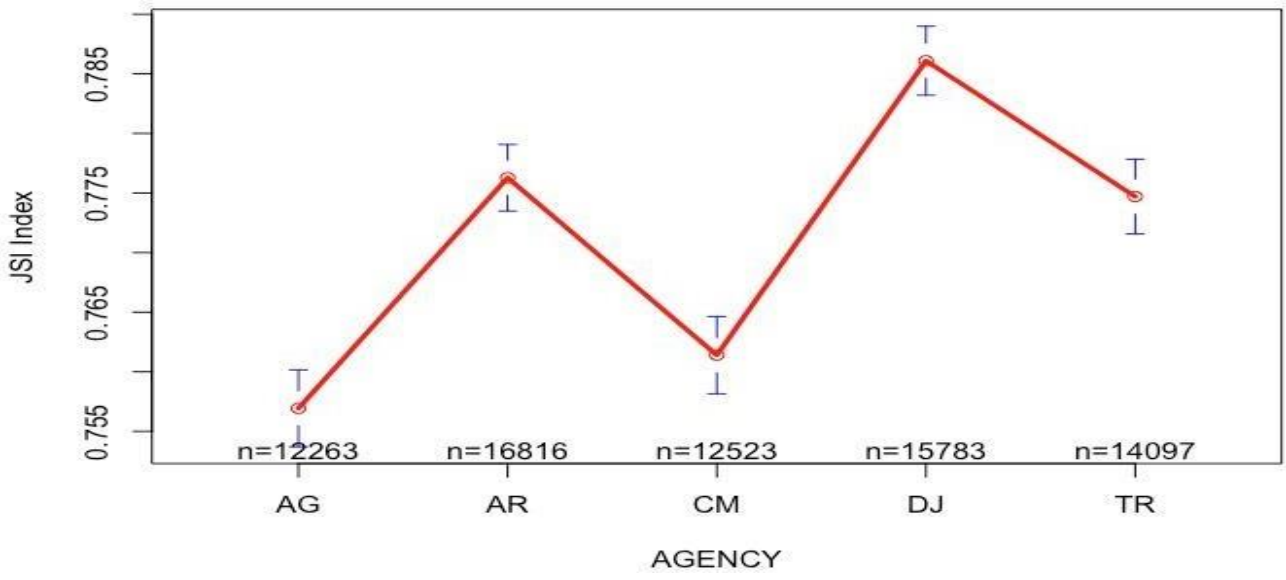
Plot 7: Histogram and boxplot of EMPINDEX



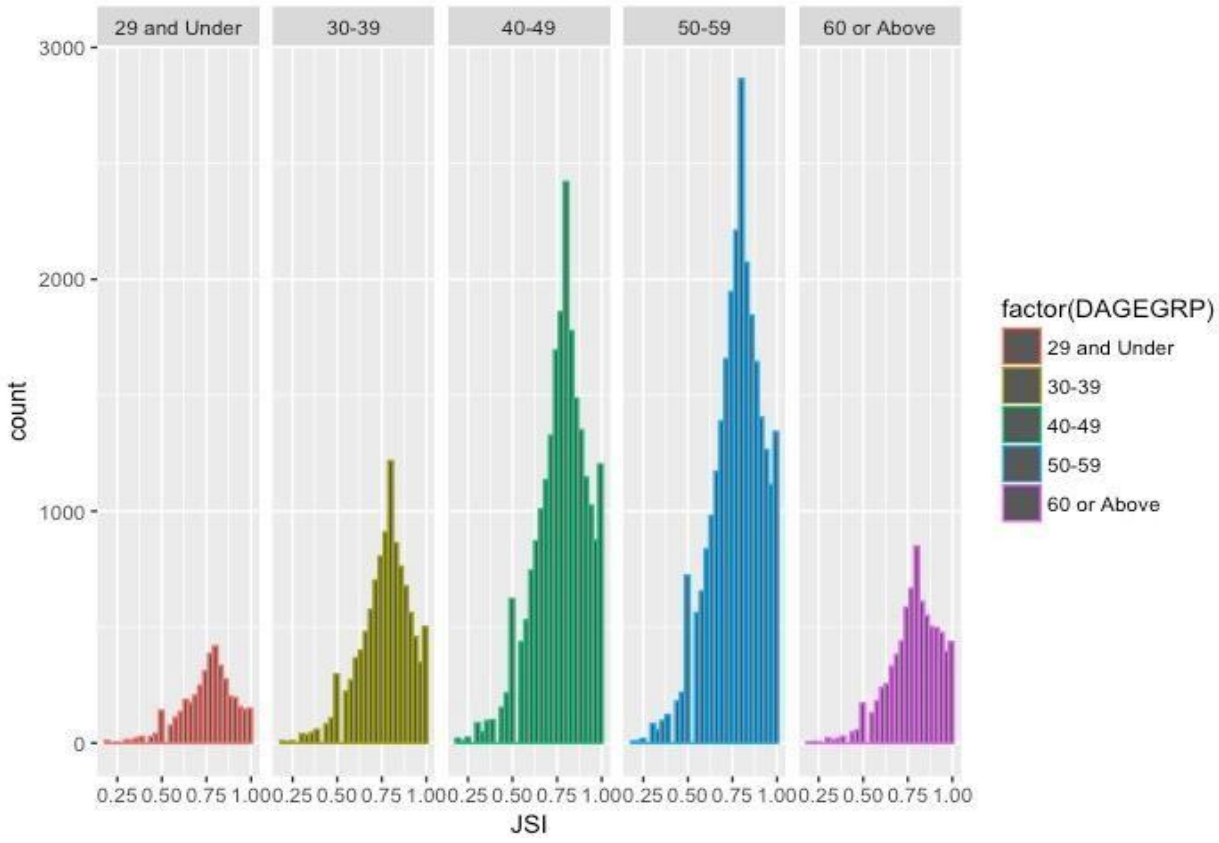

Plot 8: Distribution of Squared EMPINDEX



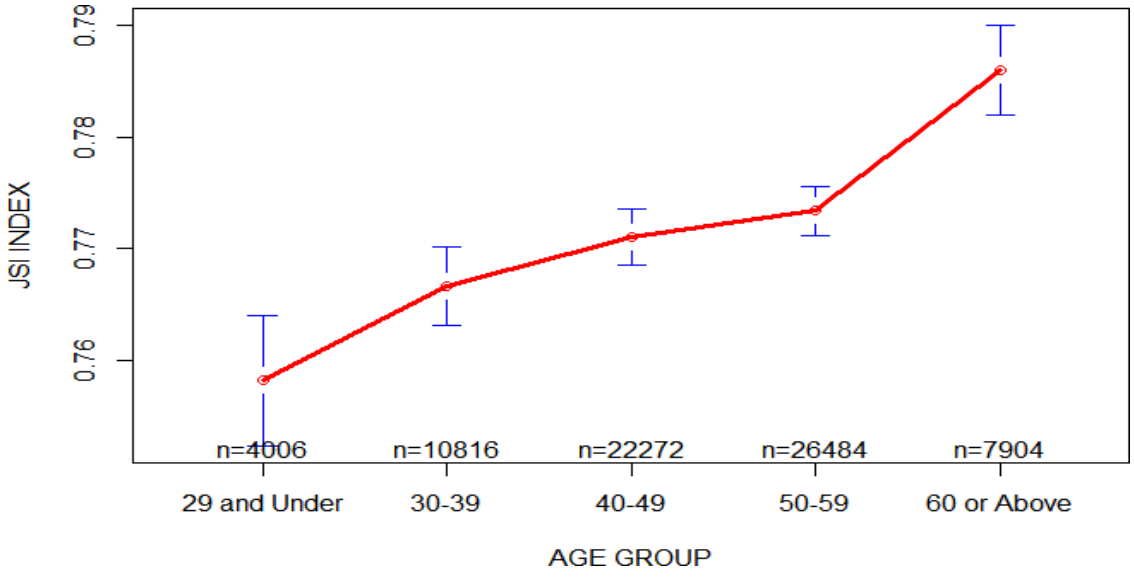

Plot 9: Distribution of Agency by JSI



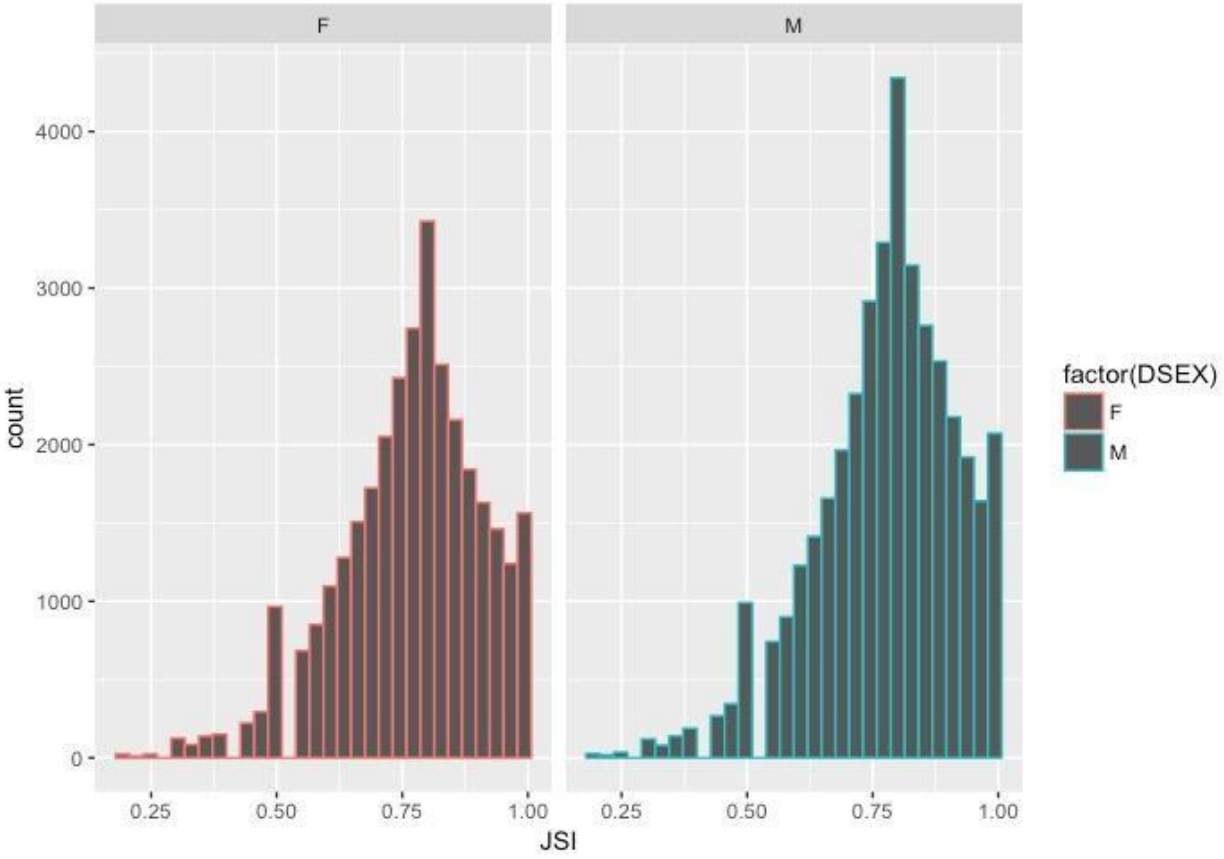

Plot 10: Mean JSI value by Agency



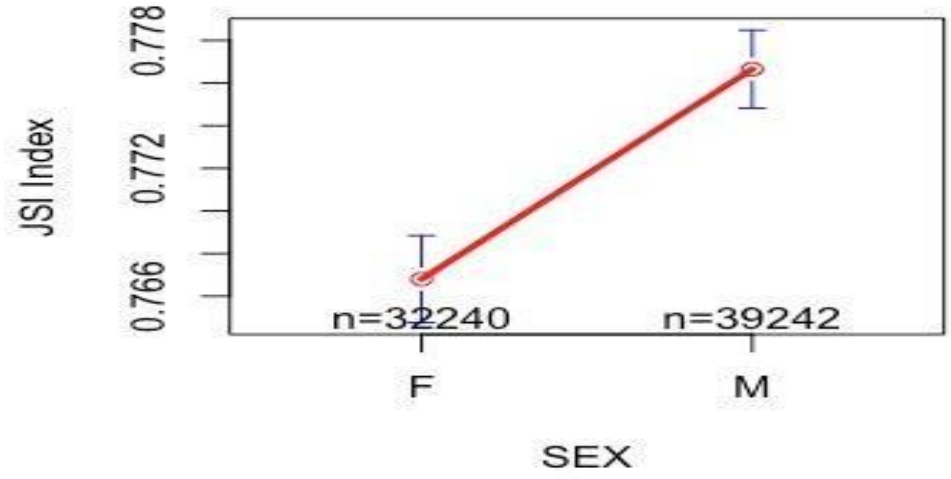

Plot 11: Distribution of Agency by Age group



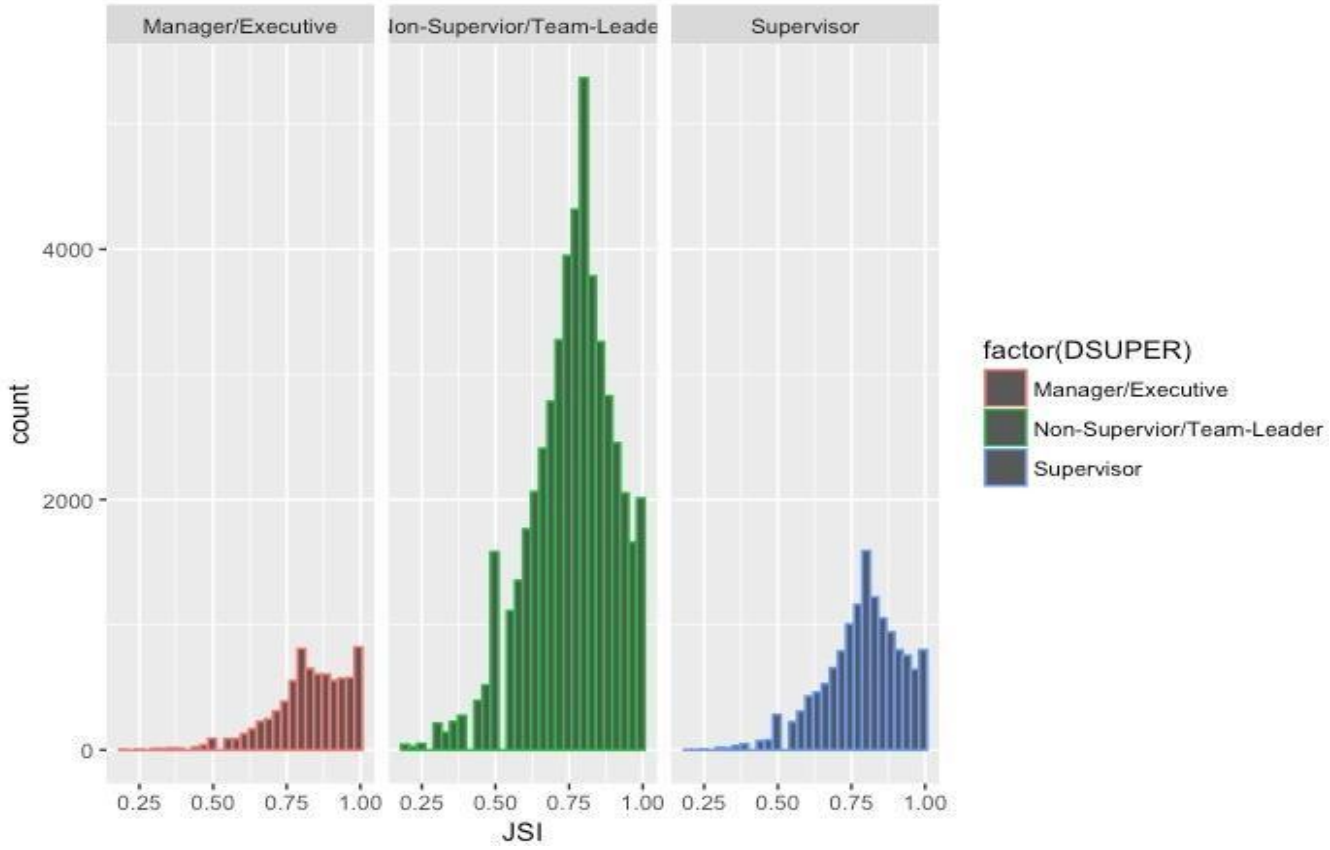

Plot 12: Mean Agency Value by Age Group



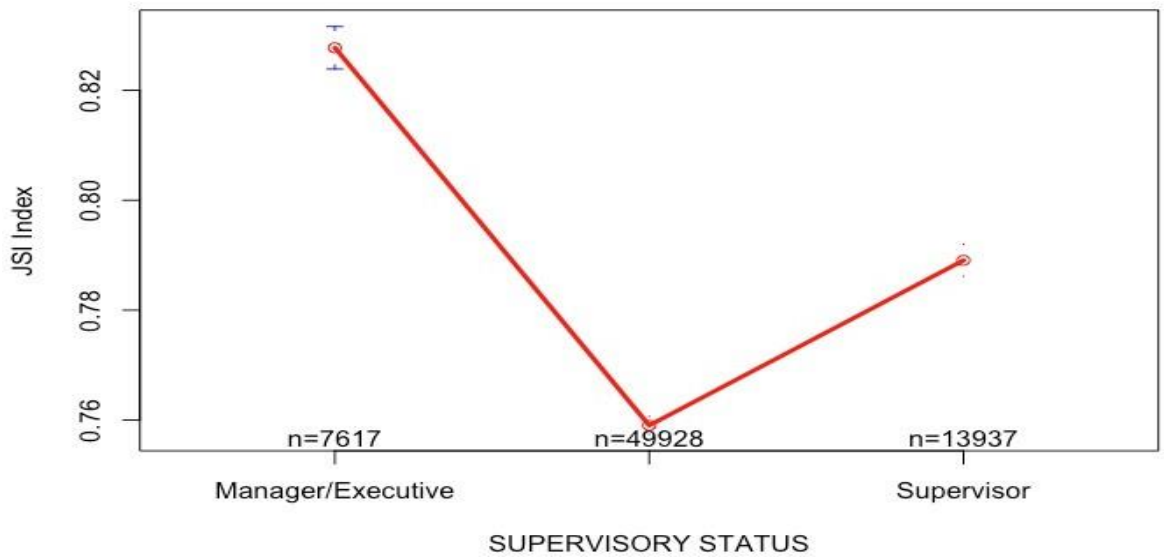

Plot 13: Distribution of JSI by Sex



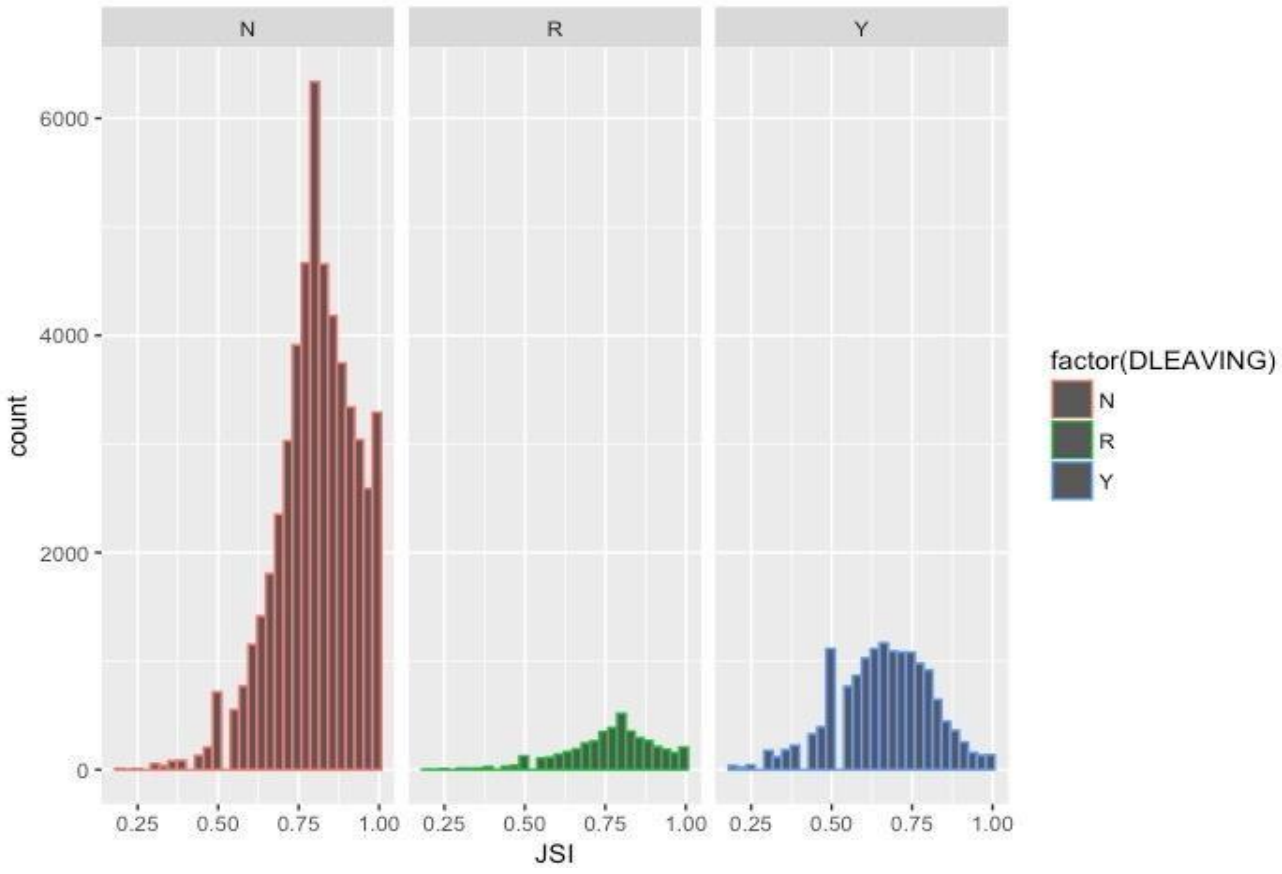

Plot 14: Mean JSI value by Sex




Plot 15: Distribution of JSI by Supervisory status

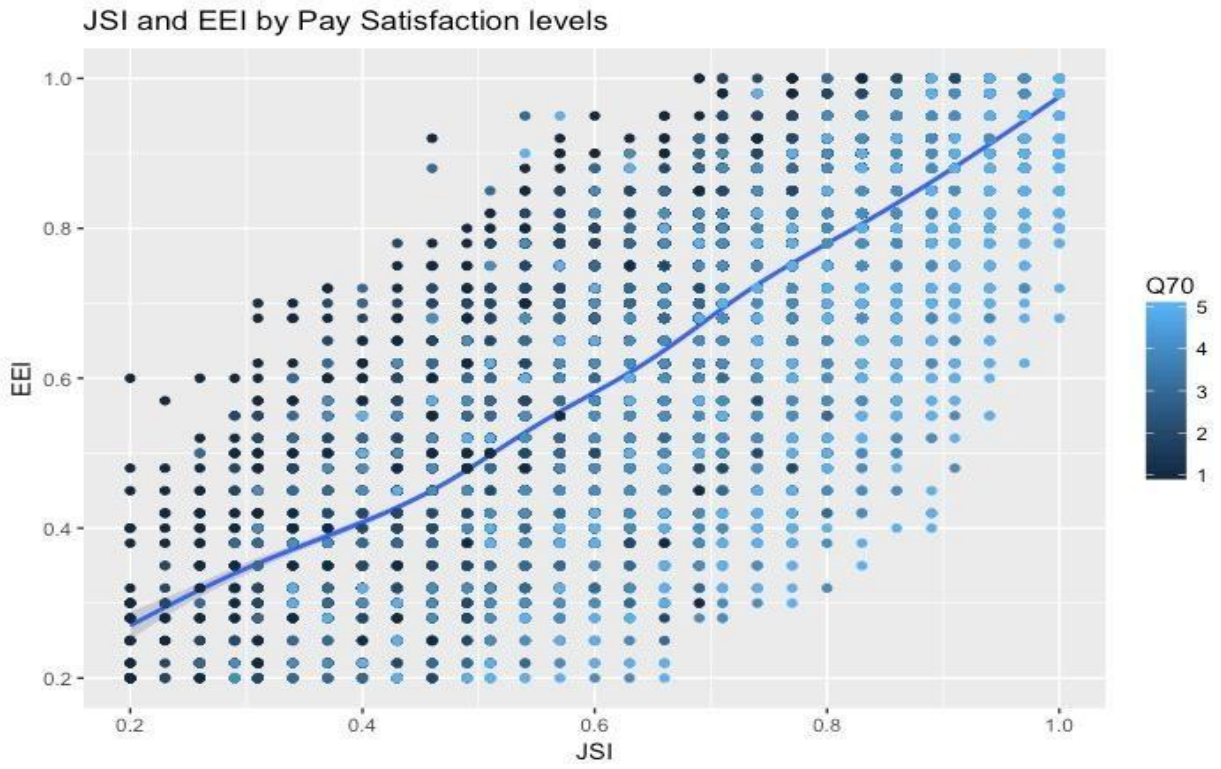



Plot 16: Mean JSI value by Supervisory Status

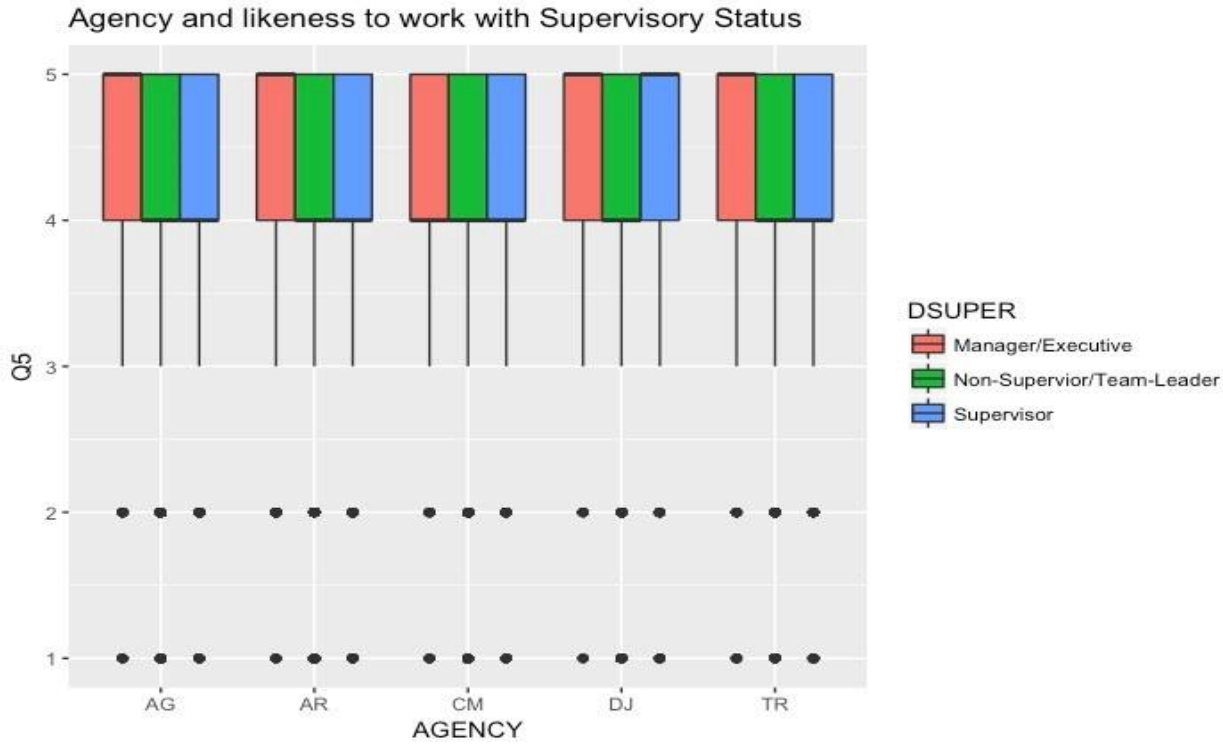



Plot 17: Distribution of JSI by DLeaving status

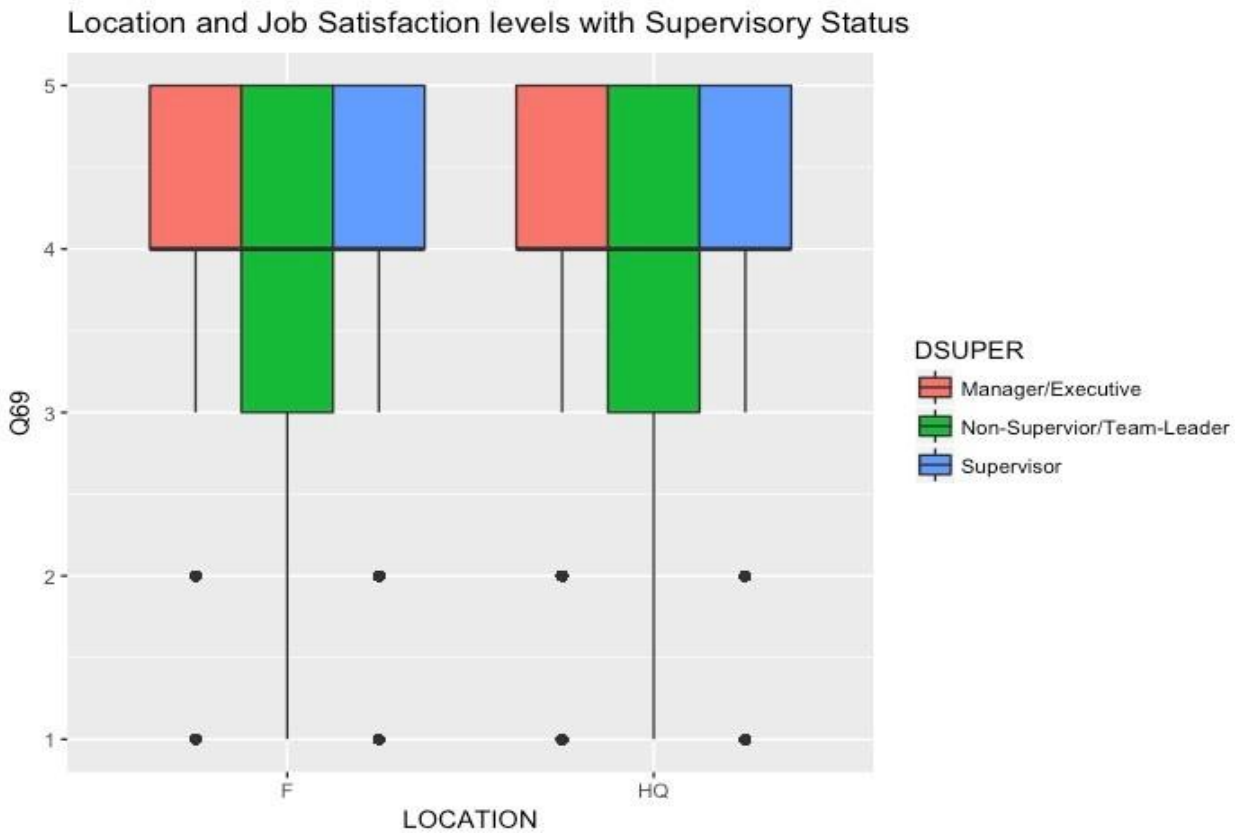

Plot 18: Mean JSI value vs Leaving status



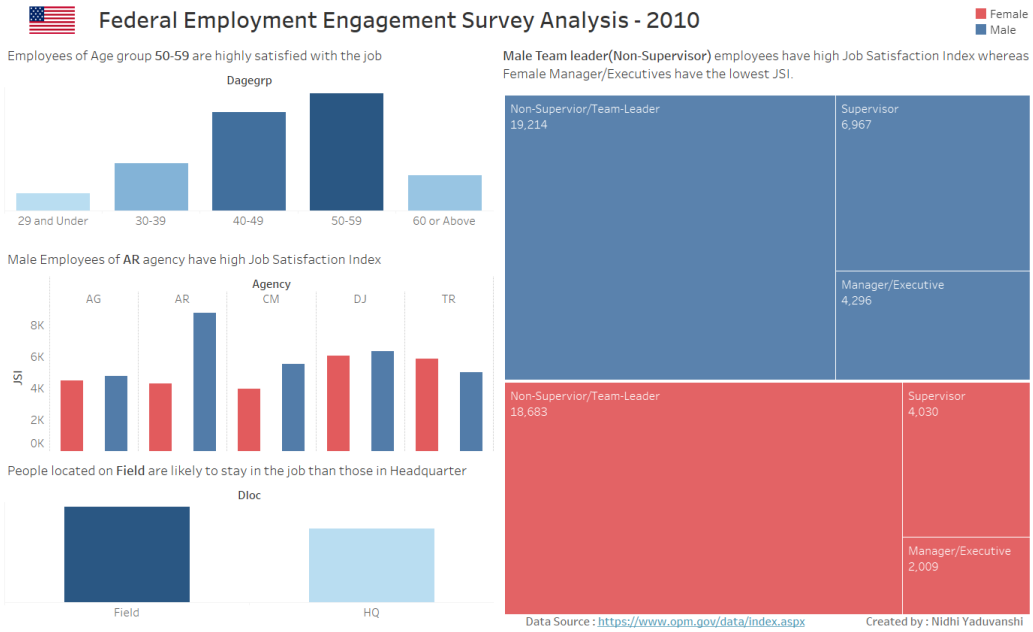

Plot 19: JSI and EMPINDEX by Pay Satisfaction Level



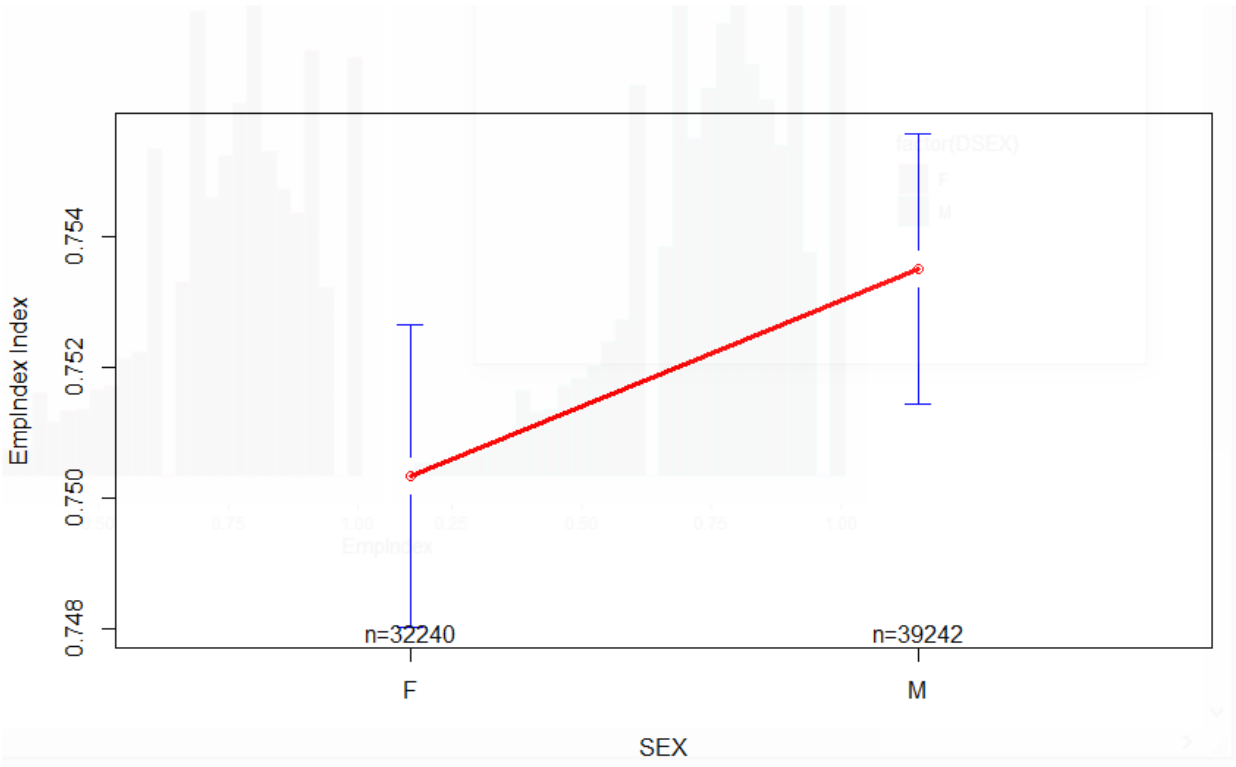

Plot 20: Agency and interest towards work with Supervisory status



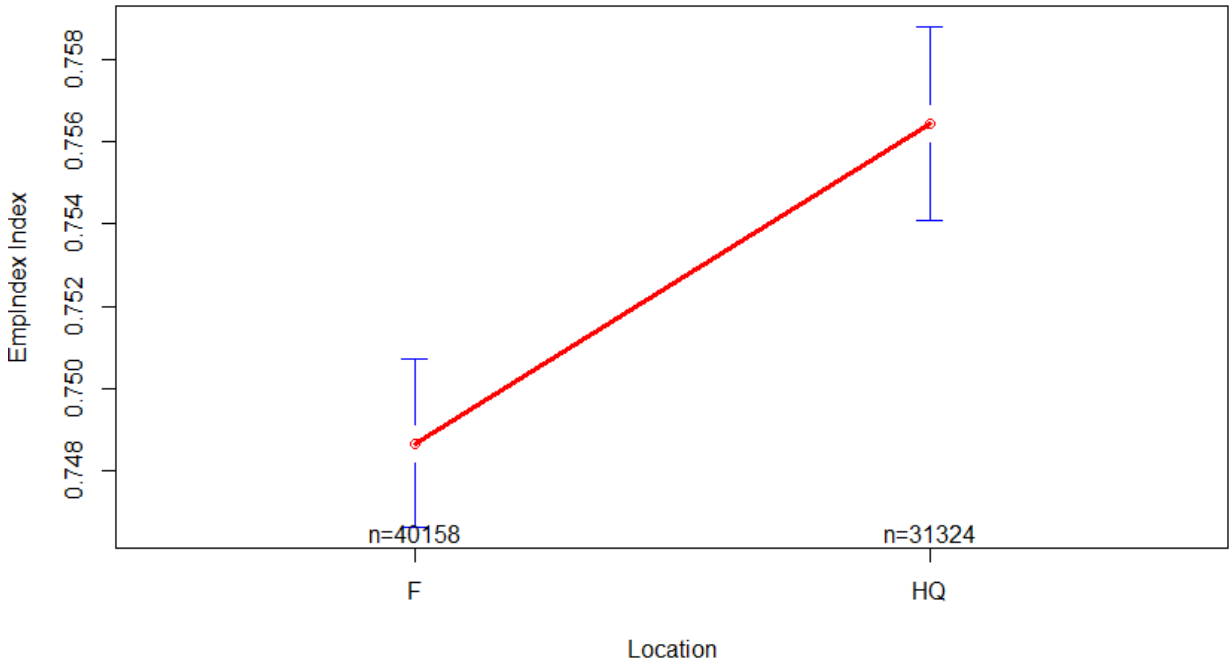

Plot 21: Location and Job Satisfaction levels with Supervisory status



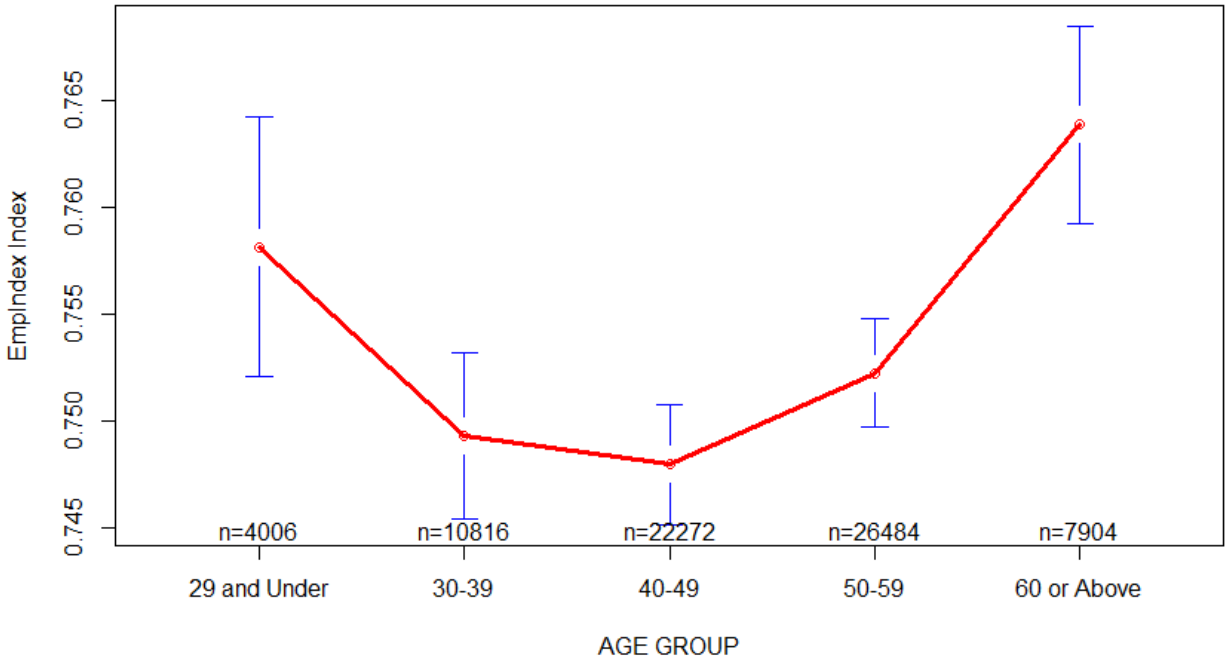

Employment Survey Analysis dashboard on Tableau.



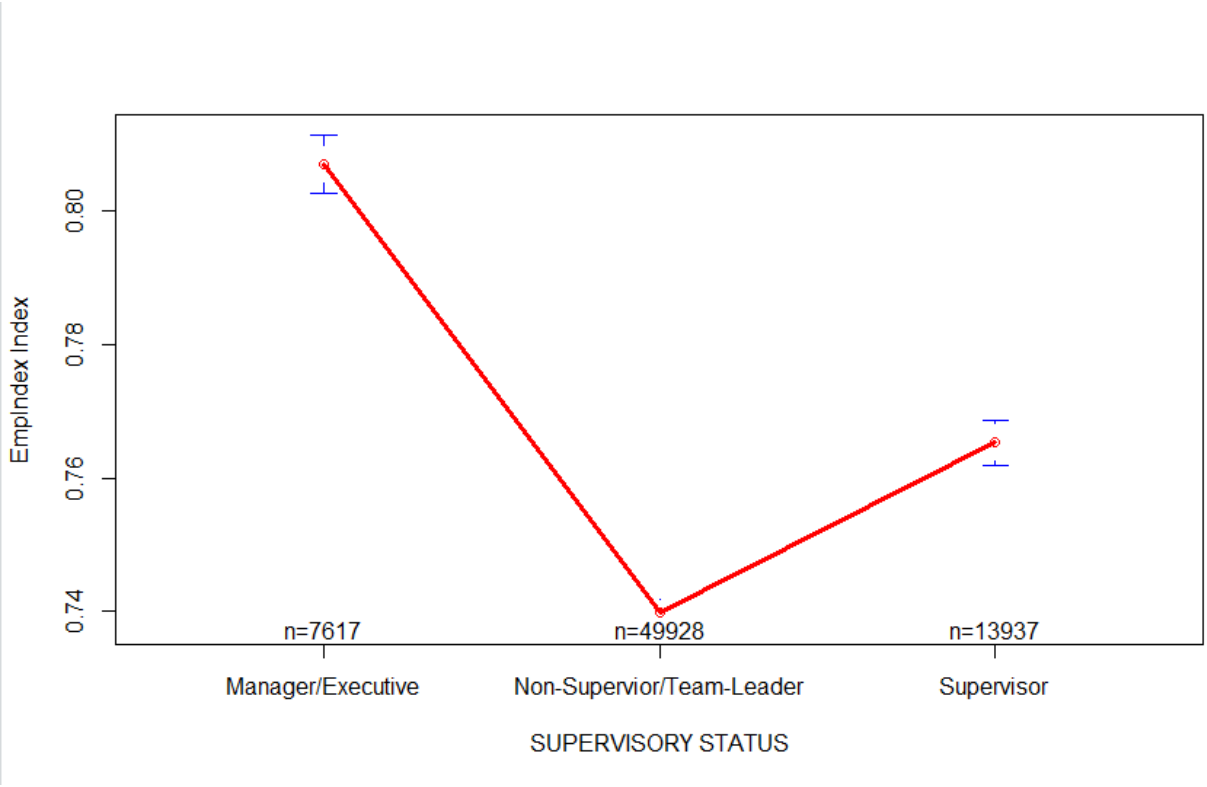

Plot 22: Mean EMPINDEX value by SEX



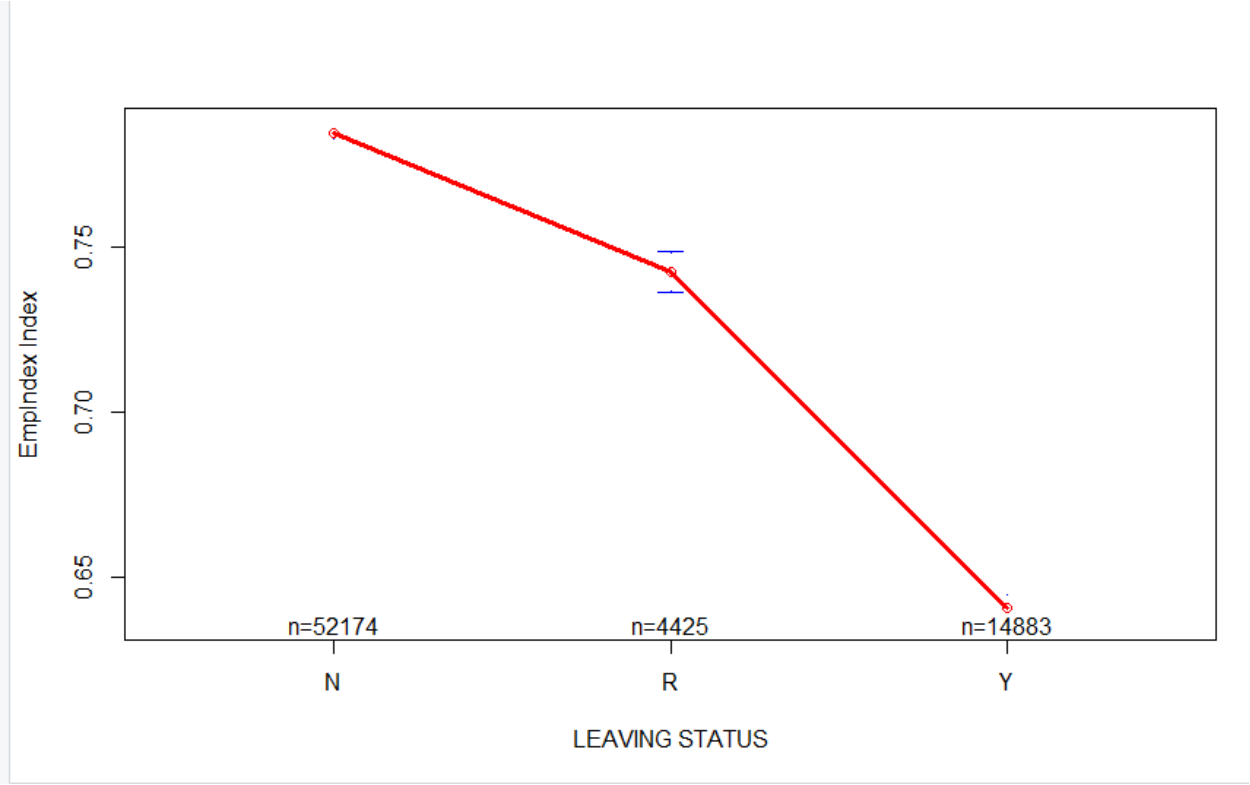

Plot 23: Mean EMPINDEX value by Location




Plot 24: Mean EMPINDEX value by Age Group




Plot 25: Mean EMPINDEX value by Supervisory Status




Plot 26: Mean EMPINDEX value by Leaving Status