The University of Illinois at Chicago

IDS 570 – Statistics for Management

(Spring-2017)

Project Report

on

Job Satisfaction Factors

Federal Human Capital Survey 2006

Iram Habiba Tarique (673504301)

Navya Hanumantharao (672346768)

Prashansa Nande (655111131)

Praveen Kasturi (658272335)

Renuka Ramachandran (679618738)

Contents

Data Description

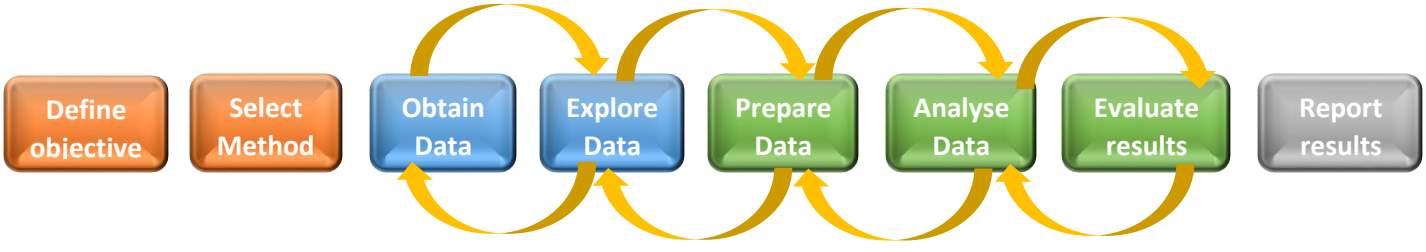
Data for this project is taken from – '*The Federal Human Capital Survey', which is conducted to understand employee perceptions about conditions pertaining to the work in their agencies*

Responses from 221479 federal agency employees were collected in the year 2006.

The data contained responses to a total of 73 questions along with other information (some demographic information like – gender, age group, location, etc.,) of the employees culminating to 88 columns in the dataset.

On a brief note, following actions were performed as an approach for analysis of given dataset-

- First, a high-level study of the given dataset was performed and understood the structure of data and some idea about different variables in the data.
- After getting an overall idea about the data, the immediate step was to come up with a research question that could perfectly define our purpose for the analysis and few hypothesis to support our claim.
- Then, the focus was on deciding the dependent and independent variable(s).
- Then data exploration was done to help us understand the data that could help us in the next step of data cleaning.
- The very important step in the entire process was to prepare the data for analysis. This included, removal of any blanks or "NA" values from the dataset. And to deal with some unknown values like "X" and "0" present in the question columns.
- After the data was ready for analysis, we performed univariate and bivariate analysis of some variables to better understand the data and find dependencies between variables if any.
- Then we performed indexing by clubbing questions under same theme into separate groups. And performed correlation test over those indexes.
- Then we proceeded to Hypothesis testing followed by Linear Regression.
- It was an iterative process and the steps were repeated till we got the best possible model to prove our research question.
- As per the final regression output, appropriate conclusions were derived which supported proved our hypotheses and thus the research question right to a good extent.



Introduction

Research Question

What are the factors that contribute to overall job satisfaction of employees in federal agencies?

Hypotheses

1. **Employees who share positive relationship with their supervisors are satisfied with their job.**
2. **Employees who believe their organization provides them good training and scope for personal growth are satisfied with their job.**
3. **Employees who feel their performance evaluation is a true reflection of their work are satisfied with their job.**
4. **Employees who are satisfied with the compensation and benefit they receive are more satisfied with their job .**
5. **Employees who observe that their organization manages diversity at workplace are satisfied with their job**
6. **Employees who regard their organization to be fair are satisfied with their job**

Units of Analysis

Dependent Variable

Question	*Variable Description*
Q56	How satisfied are you with the recognition you receive for doing a good job?
Q57	How satisfied are you with the policies and practices of your senior leaders?
Q58	How satisfied are you with your opportunity to get a better job in your organization?
Q60	Considering everything, how satisfied are you with your job?
Q61	Considering everything, how satisfied are you with your pay?
Q62	Considering everything, how satisfied are you with your organization?

Independent Variables

	Question	Variable Description
Relationship with Supervisor/ Leadership	Q8	I recommend my organization as a good place to work.
	Q37	In my organization, leaders generate high levels of motivation and commitment in the workforce.
	Q38	My organization's leaders maintain high standards of honesty and integrity.
	Q39	Managers communicate the goals and priorities of the organization.
	Q40	Managers review and evaluate the organization's progress toward meeting its goals and objectives.
	Q41	Employees are protected from health and safety hazards on the job.
	Q48	Supervisors/team leaders in my work unit support employee development.

	Q56	How satisfied are you with the recognition you receive for doing a good job?
	Q58	How satisfied are you with your opportunity to get a better job in your organization?
Organization Fairness	Q44	Arbitrary action, personal favouritism and coercion for partisan political purposes are not tolerated.
	Q45	Prohibited Personnel Practices (for example, illegally discriminating for or against any employee/applicant, obstructing a person's right to compete for employment, knowingly violating veterans' preference requirements) are not tolerated.
	Q46	I can disclose a suspected violation of any law, rule or regulation without fear of reprisal.
	Q47	Supervisors/team leaders provide employees with constructive suggestions to improve their job performance.
Training and Personal Growth	Q3	I have enough information to do my job well.
	Q6	I like the kind of work I do.
	Q14	My work unit is able to recruit people with the right skills.
	Q16	I have sufficient resources (for example, people, materials, budget) to get my job done.
	Q19	I know how my work relates to the agency's goals and priorities.
	Q21	Physical conditions (for example, noise level, temperature, lighting, cleanliness in the workplace) allow employees to perform their jobs well.
	Q25	Employees are rewarded for providing high quality products and services to customers.
	Q49	Employees have electronic access to learning and training programs readily available at their desk.
	Q51	Managers promote communication among different work units (for example, about projects, goals, needed resources).
	Q53	Employees use information technology (for example, intranet, shared networks) to perform work.
	Q59	How satisfied are you with the training you receive for your present job?
	Q60	Considering everything, how satisfied are you with your job?
Performance Evaluation	Q23	In my work unit, steps are taken to deal with a poor performer who cannot or will not improve.
	Q27	Pay raises depend on how well employees perform their jobs.
	Q28	Awards in my work unit depend on how well employees perform their jobs.
	Q29	In my work unit, differences in performance are recognized in a meaningful way.
	Q30	My performance appraisal is a fair reflection of my performance.
	Q31	Discussions with my supervisor/team leader about my performance are worthwhile.
	Q56	How satisfied are you with the recognition you receive for doing a good job?

Diversity Management	Q34	Policies and programs promote diversity in the workplace (for example, recruiting minorities and women, training in awareness of diversity issues, mentoring).
	Q35	Managers/supervisors/team leaders work well with employees of different backgrounds.
	Q36	I have a high level of respect for my organization's senior leaders.
Compensation	Q62	Considering everything, how satisfied are you with your organization?
	Q64	How satisfied are you with health insurance benefits?
	Q65	How satisfied are you with life insurance benefits?
	Q66	How satisfied are you with long term care insurance benefits?
	Q67	How satisfied are you with the flexible spending account (FSA) program?
	Q69	How satisfied are you with paid leave for illness (for example, personal), including family care situations (for example, childbirth/adoption or eldercare?
	Q70	How satisfied are you with child care subsidies?

Control Variables

Some of the control variables in the dataset are as follows:

Variable	*Variable Description*
DLOC	This variable shows the location of work of the employees' i.e Headquarters or Field.
DAGEGRP	This variable describes the various age groups of the employees working in the organization.
DSEX	This variable shows the gender of each employee working in the organization.
DAGYTEN	This variable describes the tenure of each employee.
DRETIRE	This variable keeps records of the employees who plan to retire in the coming years.
DPAYCAT	This variable describes the different pay scales of the employees.

Workflow Activity

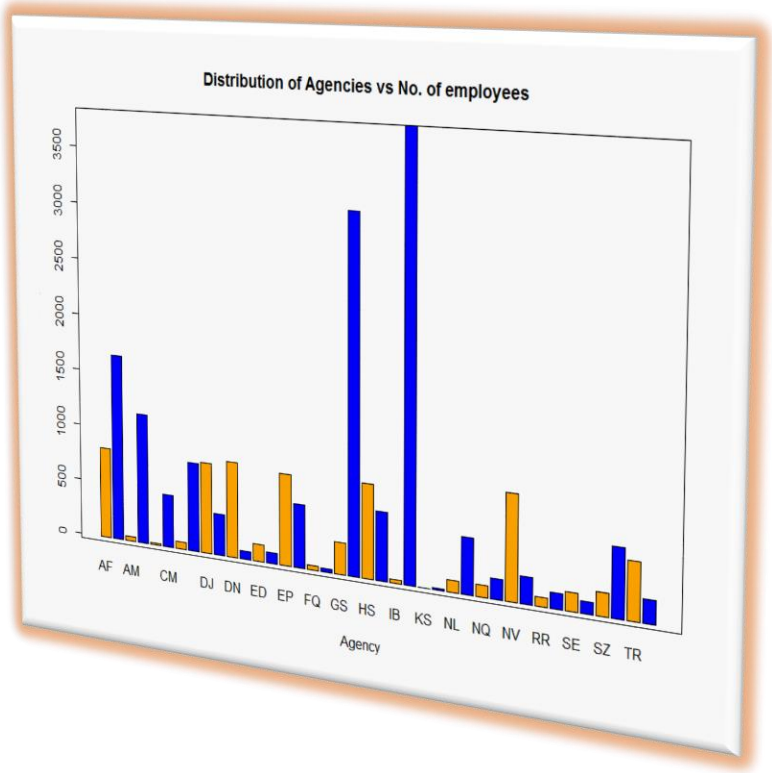
Data Cleaning

Rows containing missing data was omitted, which reduced the dataset to 81198 records
Further, questions that had certain records with 'X' as the response were replaced with 0
The data type of the columns containing response to survey questions were converted to numeric.

Exploratory Data Analysis

As the data was extremely large, we decided to filter it based on the Agency. On exploring the data, we found 39 agencies in the organization. The chart below shows the total number of employees in each agency. We decided to consider the top four agencies with maximum number of employees.

Scope of Analysis



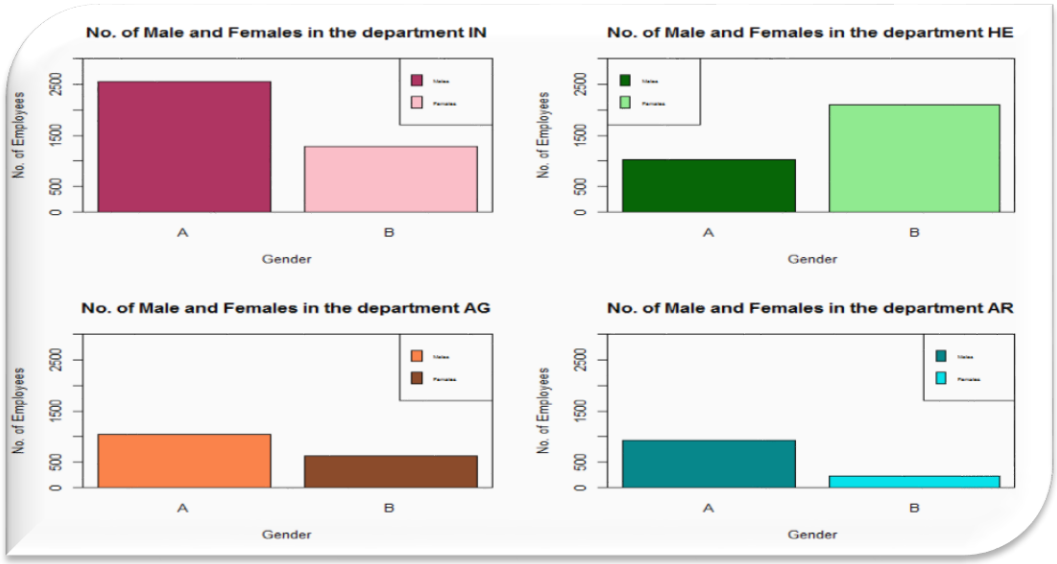
The top 4 departments, which are chosen for further analysis are:

- Department of Health and Human Services (HE)
- Department of the Interior (IN)
- Department of Agriculture (AG)
- Department of the Army (AR)

1. Univariate Analysis

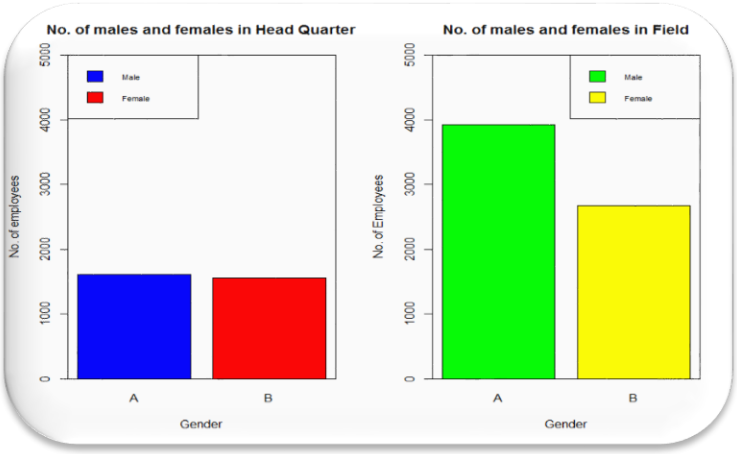
Gender (DSEX)

We further explored the four agencies based on gender. The charts below show that the number of males are more in IN, AG and AR agencies whereas the number of females are more in the HE agency. As seen below, In IN agency, there are 2549 males and 1286 females. In AG, there are 1046 males and 613 females. In AR, there are 919 males and 236 females. In HE, 1017 males and 2090 females.



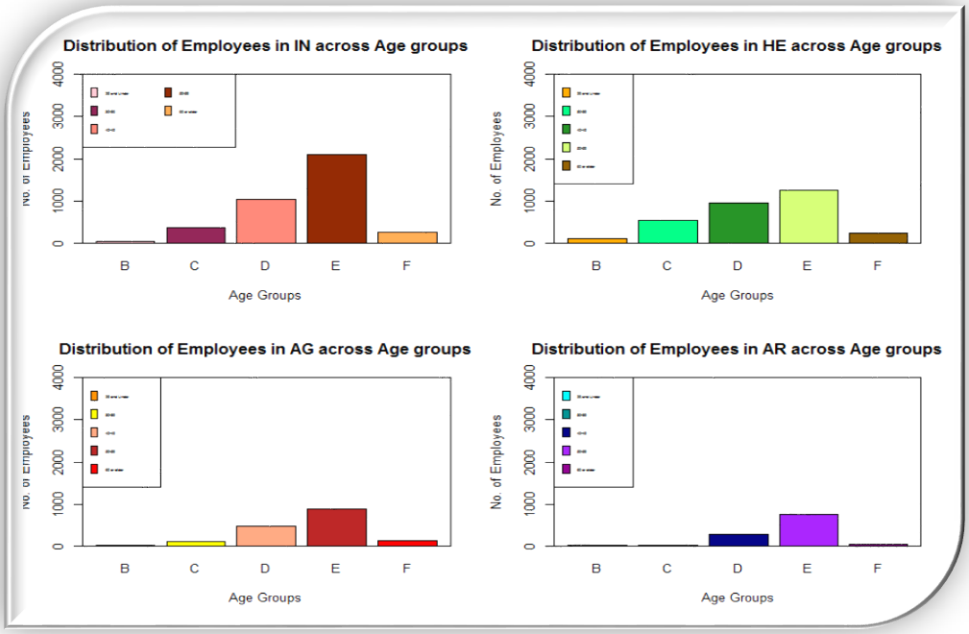
Job Location (DLOC)

We then explored the number of employees based on their job location. The chart below shows that Headquarter (HQ) has more number of female employees whereas there are more number of males in the Field Work.



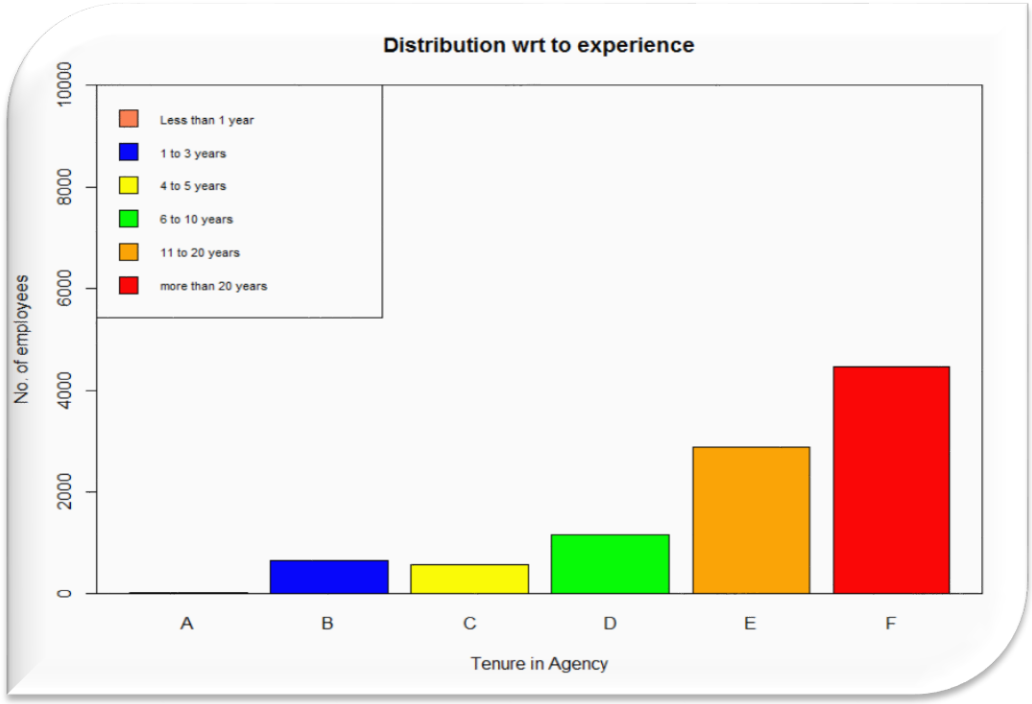
Age Group (DAGEGRP)

We explored the DAGEGROUP variable and found that most of the employees belong to the age group of 50-59 in all the four agencies.



Tenure (DFEDTEN)

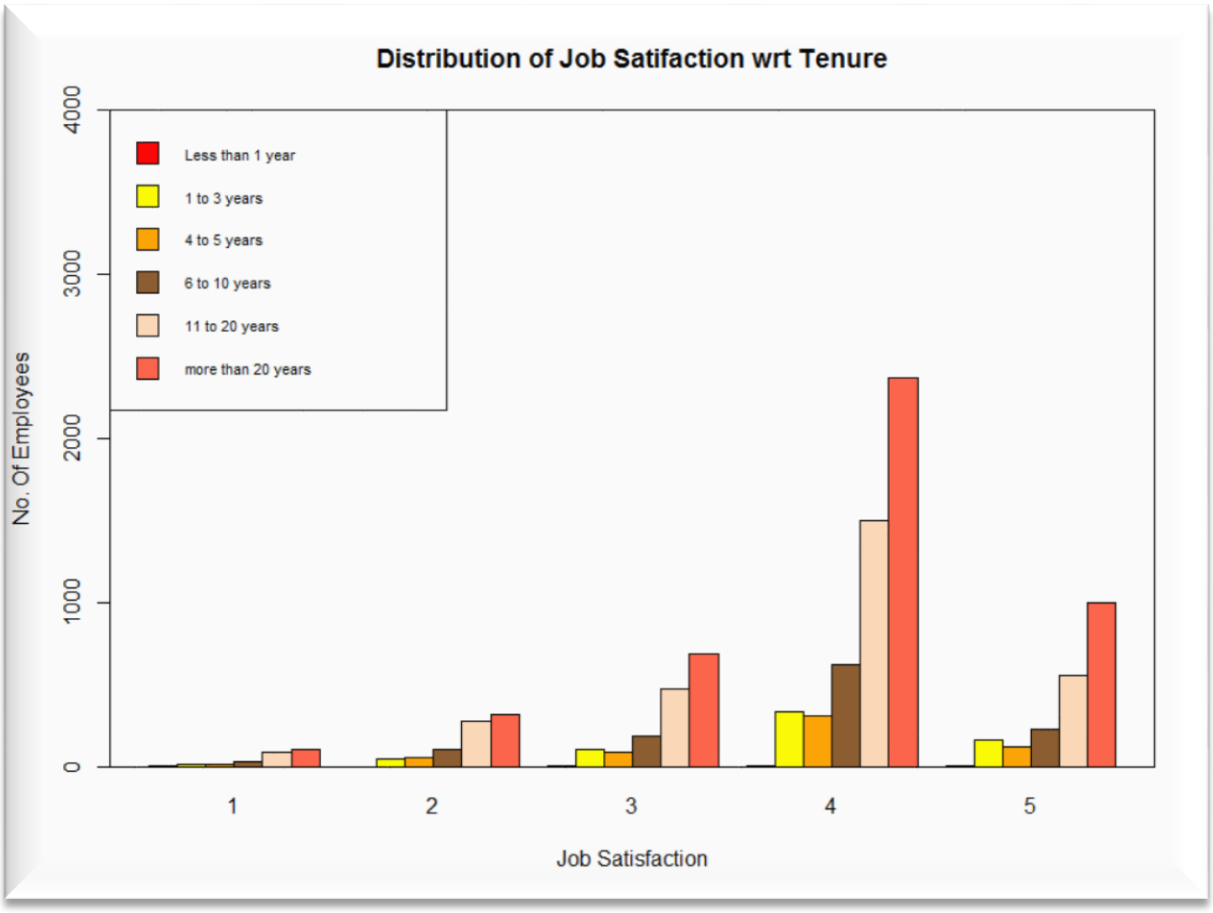
On exploring the Tenure, we see that most of the employees have more than twenty years of experience and the departments have very few freshers.



2. Bi-Variate Analysis

Tenure vs Job Satisfaction

On analysing with job satisfaction, we see that most of the employees are satisfied with their job irrespective of their tenure.

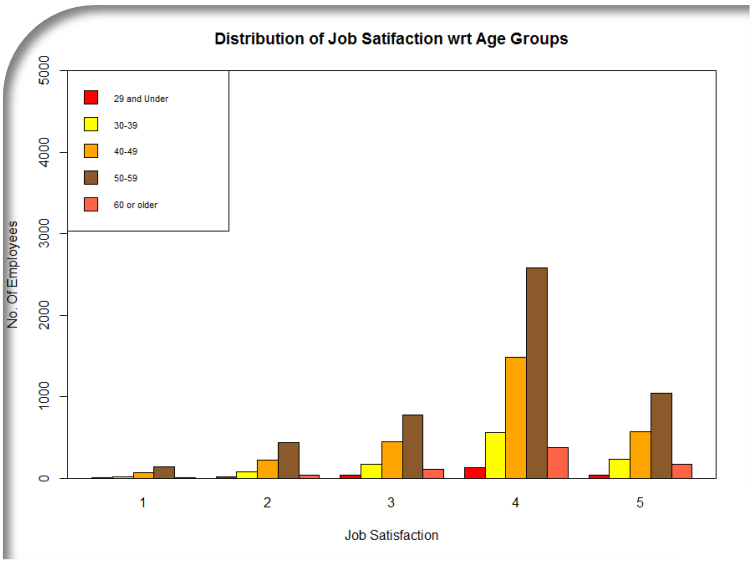


Location vs Job Satisfaction

We analysed the job satisfaction with respect to the job location. We observe from the chart below that most of the employees in the field are more satisfied with their job.

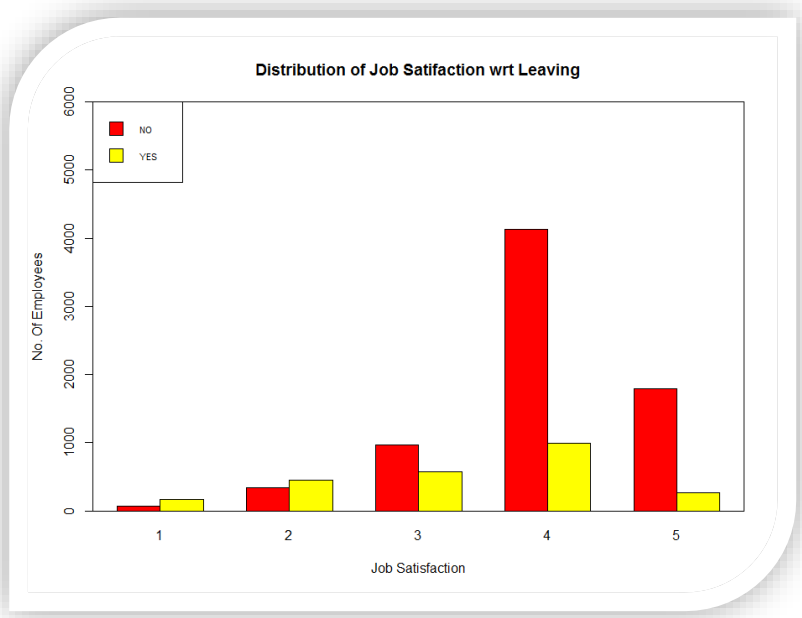
Tenure vs Job Satisfaction

When it comes to the age group, there is no particular relationship with job satisfaction as most of the employees of all the age groups are satisfied.



Leaving vs Job Satisfaction

We further analysed the employees' response to leaving their agency or organization with in the next year. As we observe from the graph below, most of the employees are satisfied with their job and hence do not wish to leave the organization.



We have now completed Data Exploration, Uni variate and Bi Variate Analysis of the final dataset. In Next section, we will be discussing Variable Index Creation, Hypothesis Testing, Linear Regression, and Regression Diagnostics.

Variable Creation

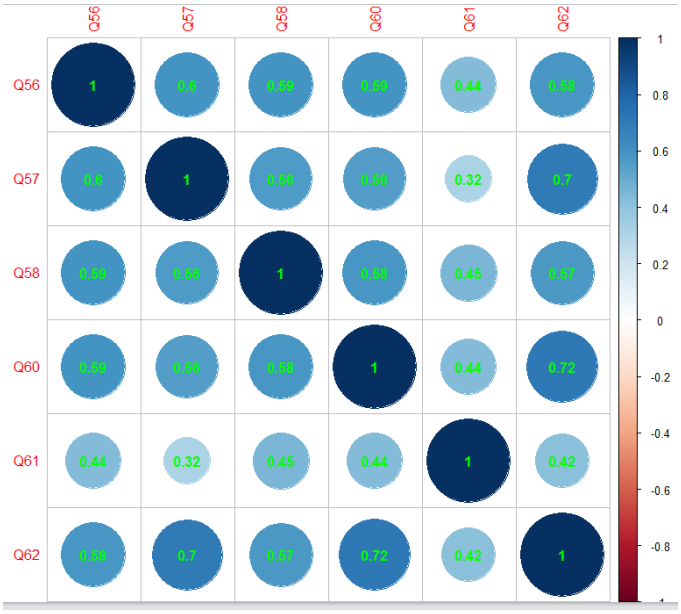
Dependent Variable

Overall Job Satisfaction:

Our Dependent variable in the model is the "Overall Job Satisfaction" Index of employees. For this, we develop an index with multiple-item questionnaire. We combined 6 questions to create this multiple-item measure. In each question, Respondents answered the item on 5-point scales (5 = "strongly agree" to 1 = "strongly disagree"). We had to combine these and create an index because these are categorical values, and we need continuous as dependent variable for Linear Regression. These items measure satisfaction with the job itself, with opportunities for recognition and advancement, with pay, with co-workers, and with supervisors. Before creating an Index, we check the correlation among variables and selected significantly high correlated variables into account. We didn't add the variables that has low correlation index value, in other words, the variables that are not contributing much required information. Here is the step by step procedure we followed to create Overall Job satisfaction index, Dependent variable.

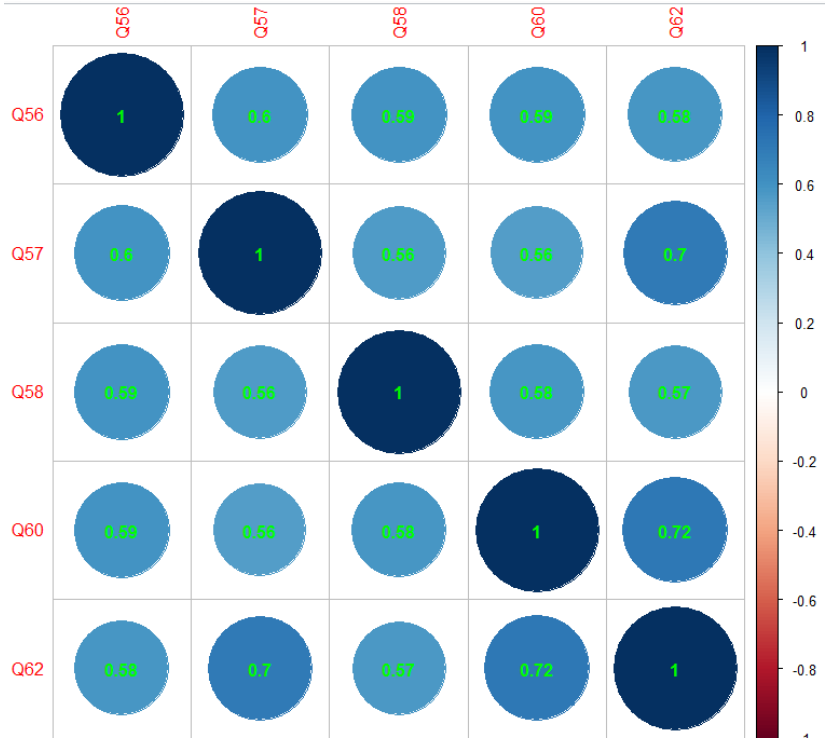
Step 1:

- For the reasons mentioned above, we have decided to combine Q56, Q57, Q58, Q60, Q61, and Q62.
- Identified the correlation among these 6 questions.



Step 2:

- Identified low contributing variable – we can observe 'Q61' have very low correlation values with other variables. Hence, dropped and created an index.
- Here is the correlation after removing Q61.
- We have taken ***0.5*** as significant correlation value to be considered to add to the Index.



Step 3:

- At this stage, we have required Questions ready to create an Index. We have created a new column with the variable name 'Overall_Job_Satisfaction' by adding values from these 5 questions.

```
Overall_Job_Satisfaction <- (survey_required$Q56 + survey_required$Q5

          +survey_required$Q58+survey_required$Q60+survey_required$Q62)
```

Note: Description of questions had given at the beginning of the document.

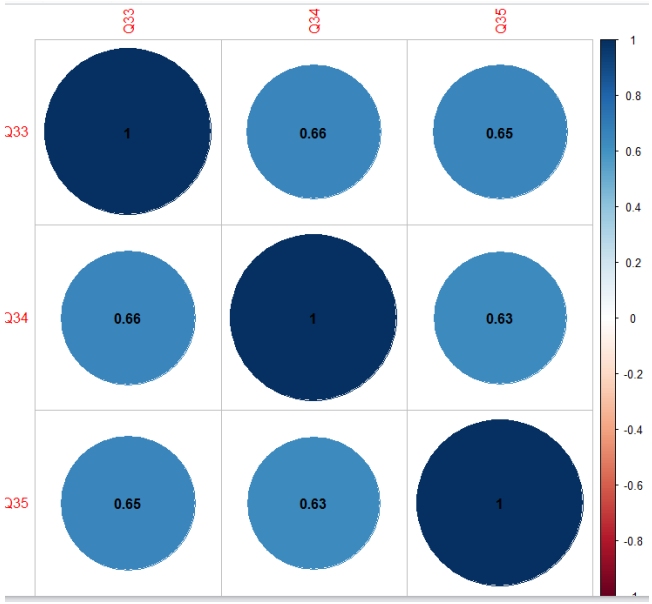
we have followed same procedure to create dependent variables too. We didn't explain the in-detailed procedure as it will be repetitive. But, we outlined high-level summary in creating dependent variables.

Independent Variables

Diversity Management:

The diversity management measure includes three questions that inquire about employees' perceptions of diversity management programs and policies of their agencies. The questions assess leaders' commitment to diversity and policies and practices to promote diversity.

- We haven Q33, Q34, and Q35 to sum up and make index to create a new Independent variable *'Diversity'*.
- `survey_required$Diversity<-(survey_required$Q33+survey_required$Q34+survey_required$Q35)`

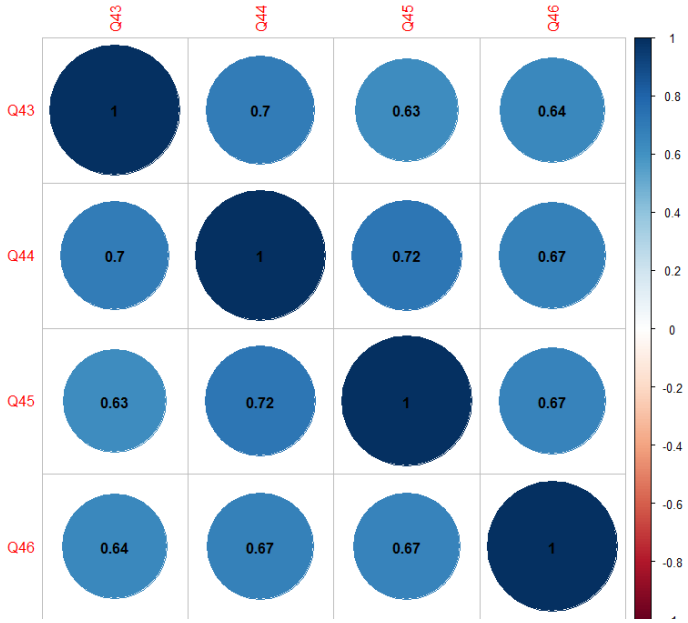


Organizational Fairness:

The measure evaluates respondents' perception of fairness in important organizational or personnel procedures in their agencies. The four survey questions (Q43, Q44, Q45, Q46) presented in the Appendix make up the measure. Our measure of perceived organizational fairness is akin to procedural justice. The survey does not include questions that support direct measurement of all these criteria. Instead, we aim to measure the overall perception of fairness in important organizational or personnel procedures. All the items used in this study measure employees' perception on fairness in personnel procedures and decision-making processes.

Our questions for organizational fairness are combined into a single index variable *"OrgFairness"* by using correlation matrix with the procedures discussed above. Here is the Correlation matrix for this variable.

```
survey_required$OrgFairness <-(survey_required$Q43 + survey_required$Q44 +
survey_required$Q45 + survey_required$Q46)
```

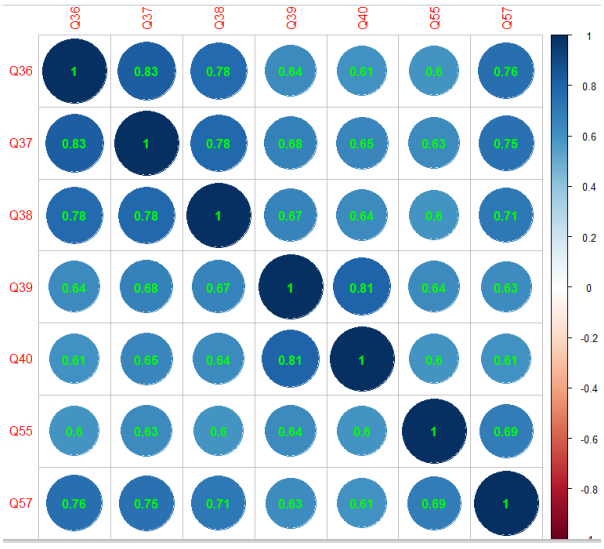


Relationship with supervisor:

Supervisors work well with different people, value representation of different segments of society, and there are diversity policies. The employee job satisfaction might increase if he/she satisfied with his/her Supervisor.

- We have taken Q7, Q38, Q39, Q40, Q41, Q42, Q49, Q57, and Q59. But after checking correlation matrix, we dropped Q7 and Q47 as they are not fitting into the group other set of questions.
- We have created a new variable "*Leadership*" and added the following questions.

```
survey_required$Leadership <-(survey_required$Q36+survey_required$Q37+
survey_required$Q38+survey_required$Q39+survey_required$Q40+survey_required
$Q55+ survey_required$Q57)
```

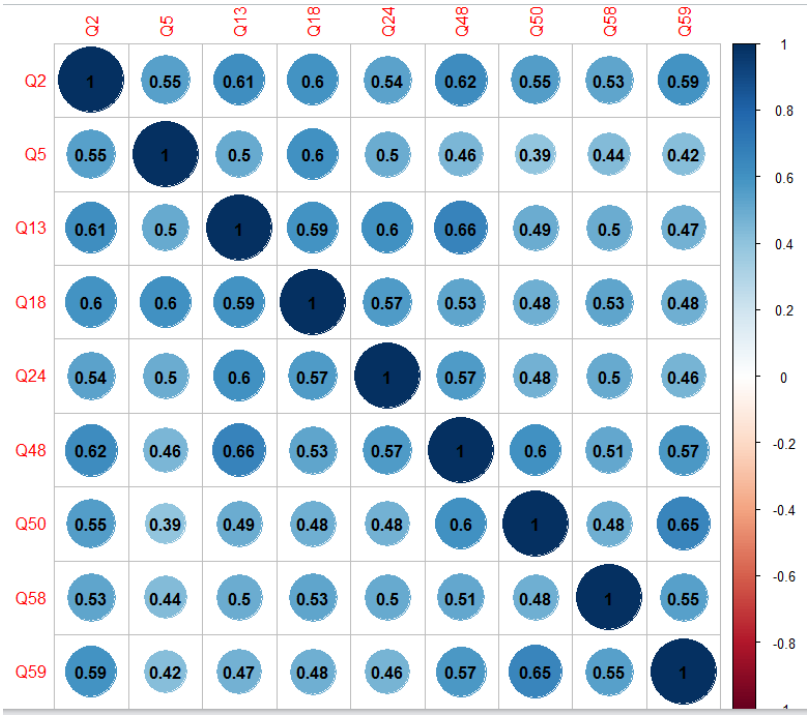


Training and Personal Growth:

The training and Personal growth parameter includes set of questionaries' to determine his/her satisfaction towards his achievements. Employees and management alike believe the workforce has the skills needed to get the job done, and a majority believes their talents are used well in the workplace. However, respondents are less optimistic about recruiting efforts and skill-level improvements in their work units.

- Here are the questions we have used to create an Index "Training_PGrowth".

```
survey_required$Training_PGrowth <-
survey_required$Q2+survey_required$Q5+survey_required$Q13+survey_required$Q
18+survey_required$Q24+survey_required$Q48+survey_required$Q50+survey_requi
red$Q58+survey_required$Q59)
```

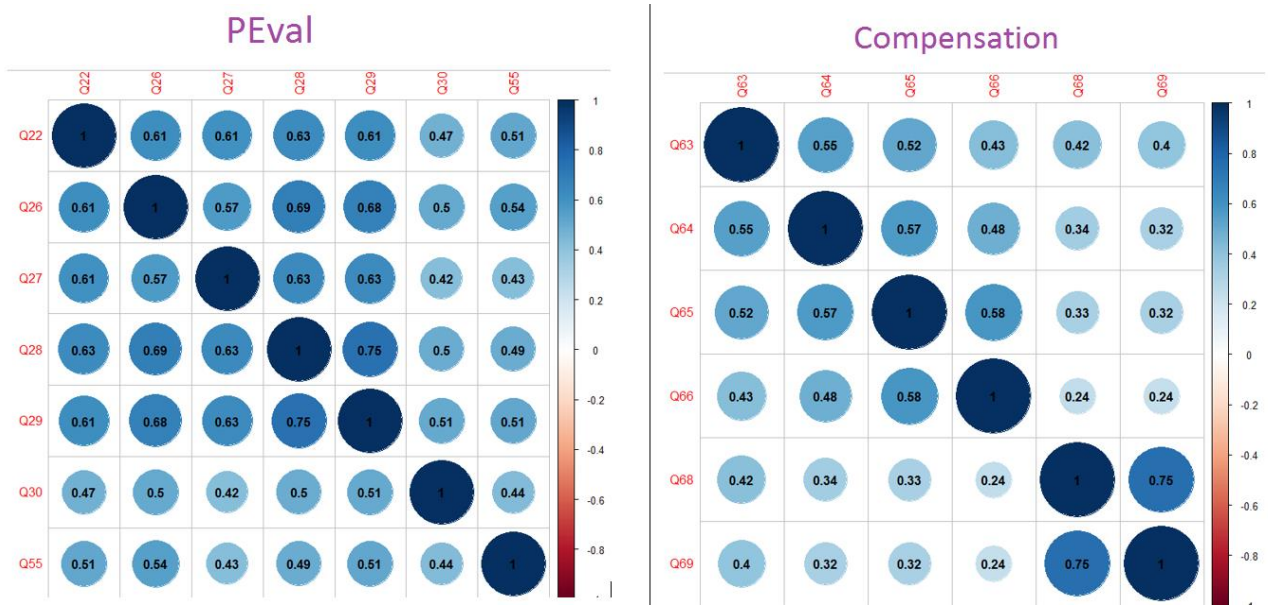


Performance Evaluation:

Generally, the greatest differences occur in perceptions of promotions, awards/rewards, and performance management. Another area of discrepancy revolves around employee ratings of perceived fairness. If an employee is not directly involved in a dispute, etc., they generally are not as familiar with the process. Generally, these issues are supposed to be confidential between the individuals involved, so lower ratings by employees are not completely unexpected. However, clear communication from top management about agency policy and expected behaviour in these areas may help to boost the ratings on these items and reduce the discrepancies between employee and executive views in these zero tolerance areas.

- Here are the questions we have considered to create a variable 'PEval' and respective correlation matrix has also shown.

```
survey_required$PEval<-
(survey_required$Q22+survey_required$Q26+survey_required$Q27+
survey_required$Q28+survey_required$Q29+survey_required$Q30+survey_required
$Q56)
```



Compensation:

The survey contains ratings of all respondents, non-supervisors, supervisors, managers, and executives. Pay for performance is an important factor in determining job satisfactory of any employee. Hence, we have chosen few set of questions that depicts their satisfaction towards 'Pay for Performance', discussions with managers regrading pay, appraisal as a fair reflection of performance etc.,

- Here are the final set of questions we have considered to create 'PEval' variable.

- We kept Q68, Q69 as they are highly correlated to each other.

```
survey_required$Compensation <- (survey_required$Q63+survey_required$Q64+
survey_required$Q65+survey_required$Q66+survey_required$Q68+survey_required
$Q69)
```

Hypothesis Testing

After creating the different indices, the next step was to test the hypotheses using these various indices. By conducting the test, we wanted to understand whether the newly created indices have some dependency on Job Satisfaction as such. This was conducted by doing a correlation test between the independent variable Over Job Satisfaction and individual dependent variables i.e. the indices created.

Correlation test was carried out using the following sample R code:

```
#Performance Evaluation vs Job Satisfaction

jsat_peval <-
cor.test(survey_required$Overall_Job_Satisfaction,survey_required$PEval)

jsat_peval
```

Hypothesis	Test Result

H1: Employees who share positive relationship with their supervisors are more satisfied with their job.	`> jsat_leadership <- cor.test(survey_required$Overall_Job_Satisfaction,survey_required$Leadership)`
	`> jsat_leadership`
	
	`Pearson's product-moment correlation`
	
	`data: survey_required$Overall_Job_Satisfaction and survey_required$Leadership`
	`t = 150.46, df = 9754, p-value < 2.2e-16`
	`alternative hypothesis: true correlation is not equal to 0`
	`95 percent confidence interval:`
	`0.8299184 0.8418721`
	`sample estimates:`
	`cor`
	`0.8359944`

H2: Employees who believe their organization provides them good training and scope for personal growth are more satisfied with their job.	`> jsat_training`
	
	`Pearson's product-moment correlation`
	
	`data: survey_required$Overall_Job_Satisfaction and survey_required$Training_PGrowth`
	`t = 159.24, df = 9754, p-value < 2.2e-16`
	`alternative hypothesis: true correlation is not equal to 0`
	`95 percent confidence interval:`
	`0.8442136 0.8552421`
	`sample estimates:`
	`cor`
	`0.8498208`

H3: Employees who feel their performance evaluation is a true reflection of their work are more satisfied with their job.	`> jsat_peval`
	
	`Pearson's product-moment correlation`
	
	`data: survey_required$Overall_Job_Satisfaction and survey_required$PEval`
	`t = 128.74, df = 9754, p-value < 2.2e-16`
	`alternative hypothesis: true correlation is not equal to 0`
	`95 percent confidence interval:`
	`0.7859557 0.8006626`
	`sample estimates:`
	`cor`
	`0.7934249`

H4: Employees who are satisfied with the compensation and benefit they receive are more satisfied with their job.	`> jsat_compensation`
	
	`Pearson's product-moment correlation`
	
	`data: survey_required$Overall_Job_Satisfaction and survey_required$Compensation`
	`t = 35.283, df = 9754, p-value < 2.2e-16`
	`alternative hypothesis: true correlation is not equal to 0`
	`95 percent confidence interval:`
	`0.3187087 0.3539057`
	`sample estimates:`
	`cor`
	`0.3364247`

H5: Employees who observe that their organization manages diversity at workplace are more satisfied with their job	> jsat_diversity Pearson's product-moment correlation data: survey_required$Overall_Job_Satisfaction and survey_required$Diversity t = 73.408, df = 9754, p-value < 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.5836083 0.6091759 sample estimates: cor 0.5965434
H6: Employees who regard their organization to be fair are more satisfied with their job.	> jsat_orgfair Pearson's product-moment correlation data: survey_required$Overall_Job_Satisfaction and survey_required$OrgFairness t = 98.75, df = 9754, p-value < 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.6969971 0.7168472 sample estimates: cor 0.7070614

As we can see in the results, for each of the hypothesis that was tested the p-value is a very small number <2e-16. *This indicates that the distinct factors such as Relationship with Supervisor/Leadership, Performance Evaluation, Training & Personal Growth, Organization Fairness, Diversity Management and Compensation are all dependent on Overall Job Satisfactions to some extent.*

Two other additional tests were conducted as part of data exploration.

Chi-Square Test

Null Hypothesis: There is no association between Gender and Work Location.

Alternate Hypothesis: The work location of an employee is associated with the employee's gender.

```
> chisq.test(tab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab
X-squared = 65.326, df = 1, p-value = 6.347e-16
```

Analysis: We reject the Null Hypothesis since the p-value is low. This implies that there is an association between Gender and Work Location.

Linear regression

At this juncture, we are familiar with our Dependent and Independent variables. Also, we did Hypothesis testing on few variables. Now, we are good to go and design a predictive model - Linear Regression.

Linear Regression is an approach for modelling relationship between a scalar independent and one or more explanatory variables.

Now, we will perform linear regression of independent variables against our dependent variable 'Overall Job satisfaction". Here, we have run the model against all IVs and few demographic variables. Then, we have experimented by removing few other variables and got a better model.

Before going into the prediction model, we have checked and make sure few Linear Regression assumptions.

- DV should be normally distributed.

Data Transformations:

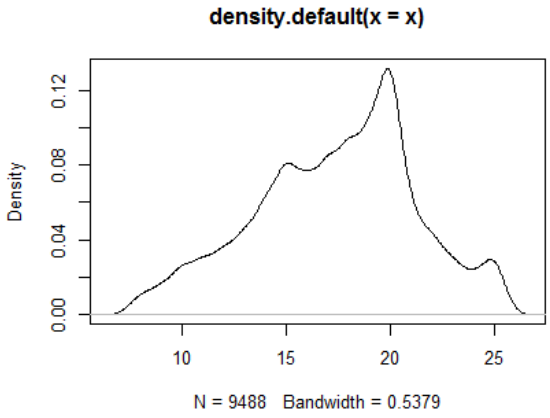
We have verified our DV (Overall Job satisfaction) for normal distribution. We have also tried to check data transformation techniques to verify if any transformation like 'log', 'square' doing better than default one.

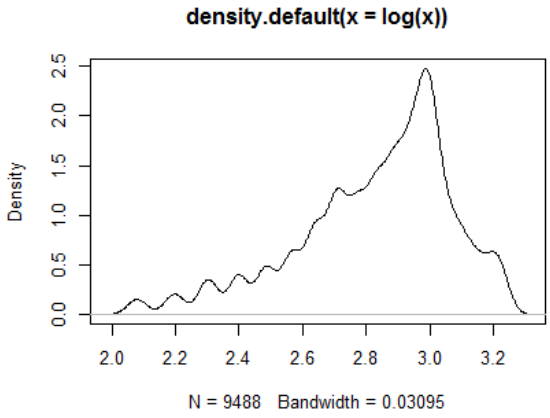
Default one did better than the transformations in this case. Here is the function we have written for transformations and the code we used to write the function also has been provided.

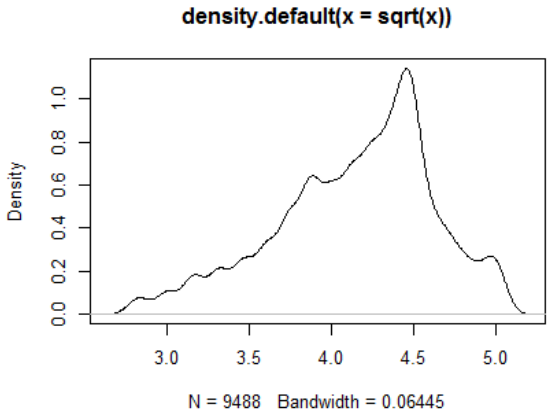
```
transforms <- function(x)
{
 par(mfrow=c(2,2))
 plot(density(x))
 plot(density(log(x)))
 plot(density(sqrt(x)))
}
```

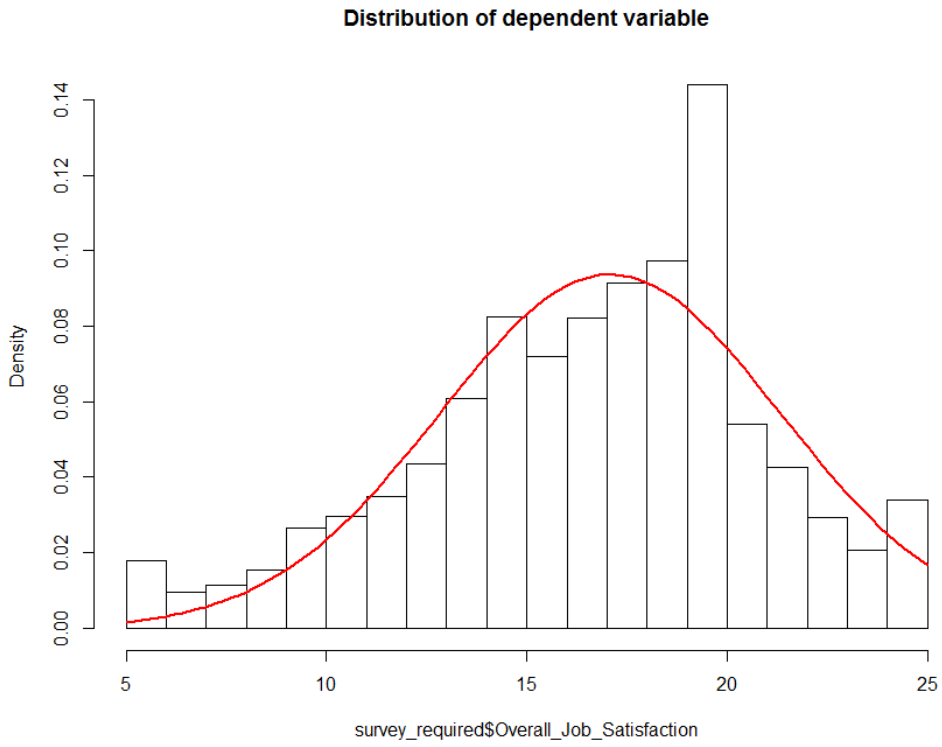
```
transforms(final_data$Overall_Job_Satisfaction)
```

- The 'transforms' function we have written will perform 3 kinds of transformations and gives us the density plots. Accordingly, we can choose the best out of one.









Now, let's build an initial model for linear regression.

Model1:

We run the linear regression model with independent variables that had been created along with few demographic variables DLOC and DESX. Here is the output and had written our observations.

```
Call:
lm(formula = Overall_Job_Satisfaction ~ Leadership + PEval +
    Training_PGrowth + OrgFairness + Compensation + DLOC + DSEX,
    data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3944 -1.1765  0.0577  1.2115  8.6687

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.276522   0.156625 -14.535  < 2e-16 ***
Leadership        0.253060   0.005224  48.444  < 2e-16 ***
PEval             0.164139   0.006116  26.836  < 2e-16 ***
Training_PGrowth  0.238509   0.005701  41.838  < 2e-16 ***
OrgFairness      -0.022645   0.009486  -2.387    0.017 *
Compensation      0.192354   0.005705  33.716  < 2e-16 ***
DLOCB             0.177076   0.044507   3.979 6.98e-05 ***
DSEXB             0.045260   0.042669   1.061    0.289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.03 on 9606 degrees of freedom
Multiple R-squared:  0.8032,    Adjusted R-squared:  0.803
F-statistic:  5600 on 7 and 9606 DF,  p-value: < 2.2e-16
```

- Job satisfaction Index in my response variable and others are my predictor variables.
- Intercept is a mandatory parameter 'R' shows unless we explicitly tell R not to have.
- Call gives the formulae for linear regression model.
- Estimate gives us estimated coefficients. It will give us coefficients.
- Intercept estimate is the average value when IV values' =0

Intercept Coefficient:

Intercept coefficient = -2.27 means that, when all the other coefficients become zero, the overall job satisfaction index score Is -2.276522

Estimated Value:

Leadership Estimate Value = 0.253 means that, for each unit increase in Leadership variable, job satisfaction Index will increase by '1' unit.

RSE:

Standard error is the estimated variability in a coefficient due to SAMPLING variability. i.,e different samples may result in different values.

Std Error

This is standard variability. Variability for the leadership is 0.005.

t-value

t-value = Estimate / Std Error. The more the Estimate coefficient value, the more the t-value. The more Std error, the less the t-value.

P-value

These are the p-values for t values on each coefficient, given n-2 degrees of freedom.it tells us how statistically significant each variable is.

Residual Standard Error

RSE value we for is 2.03 on 9606 degrees of freedom.

It quantifies how well or purely the model does predicting 'Y' values in the data, on average. Thus, its like average error for the model.

Mathematically, RSE = Squared Root (sum of squared errors/degrees of freedom)

R^2 Value

We have got 0.803, it's 80% in other words. It is interpreted as the percentage of variation in the response variable that is explained by variation in explanatory variable.

It is calculated as the percentage of total sums of squares (SST) that is composed of sums of squares from the model (SSM).

Mathematically, R^2 = SSM/SST = 1- SSE/SST

- Adjusted R^2 considers how many explanatory variables in our model. If we add multiple variables, our R^2 value will go up. If the values we are adding statistically significant, then adjusted R^2 takes this into account, and presents a much lower than value than R^2.

F-Value

It is the ratio of how well the model is doing and how the error is doing. In general, Higher F-value means, the model is doing good as it explains more of the model than error. We have got 5600 in this scenario.

The first term (7) is Numerator degree of freedom and the second one (9606) is Error degrees of freedom. Mathematically, it is ratio of Mean square model to Mean square error.

Overall P-Value

It is very familiar to us. We either accept or reject null hypothesis based on the P-value. P value less than 0.05 indicates overall model is significant. we have got much less P-value; Hence, our model is much significant.

Now, let's try out another model with some other variables.

Model2:

```
Call:
lm(formula = Overall_Job_Satisfaction ~ Leadership + PEval +
    Training_PGrowth + Diversity + +Compensation + DLOC, data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4541 -1.1624  0.0504  1.2131  8.7739

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.756925   0.158687 -11.072  < 2e-16 ***
Leadership      0.262121   0.004879  53.720  < 2e-16 ***
PEval           0.176046   0.006027  29.211  < 2e-16 ***
Training_PGrowth  0.250315   0.005683  44.049  < 2e-16 ***
Diversity      -0.161125   0.012170 -13.240  < 2e-16 ***
Compensation    0.195624   0.005644  34.659  < 2e-16 ***
DLOCB           0.157968   0.043970   3.593 0.000329 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.012 on 9607 degrees of freedom
Multiple R-squared:  0.8065,    Adjusted R-squared:  0.8064
F-statistic:  6675 on 6 and 9607 DF,  p-value: < 2.2e-16
```

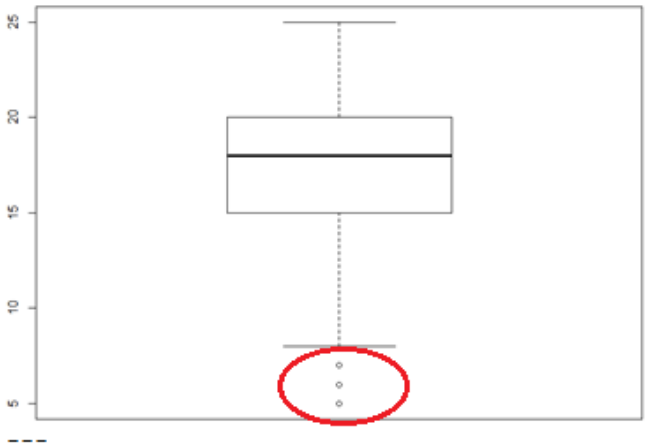
In this model, we have eliminated a control variable DSEX and tested the linear regression model. We have observed *slight increase in Adjusted R^2 value and a significant increase in F-static value*, which indicates this as a better model than model1.

Regression Diagnostics

Outlier Treatment:

At this juncture, we have decided to eliminate some outliers and check if it gives us a better regression model. Here are the steps we have followed to eliminate outliers.

```
b1 <- boxplot(survey_required$Overall_Job_Satisfaction)

outliers <- b1$out

outliersdf <- survey_required[survey_required$Overall_Job_Satisfaction %in%
outliers,]

# Use the %in% function to match each row to these values

# In the above, you've got an entire data frame. If you only need row
#names, use this function

final_no_outliers_overall <-
subset(survey_required,!(survey_required$Overall_Job_Satisfaction %in%
outliers))

final_data <- final_no_outliers_overall
```



At this juncture, we have eliminated the outliers (shown in red block in above image). Now, we will check for Multi-collinearity.

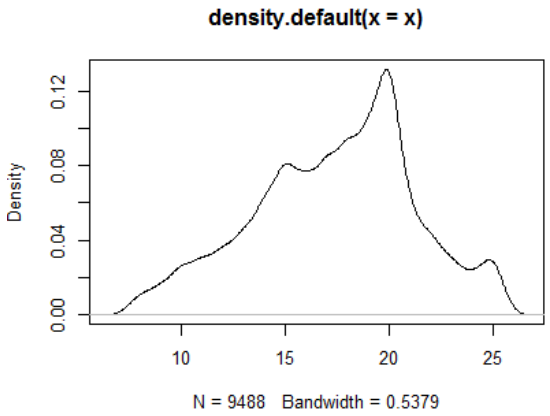
Multi collinearity:

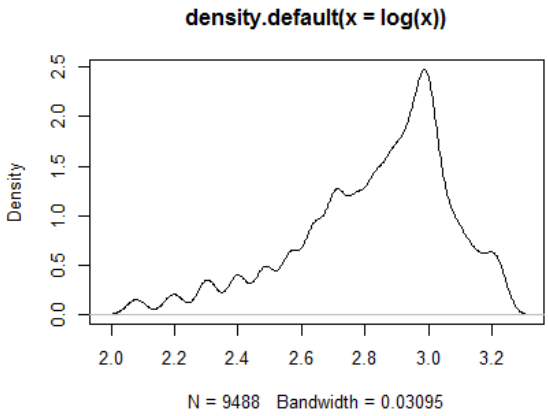
We have re-visited and did check for Multi collinearity through 'Variance Influence Factor' analysis, and we found no variable need to be eliminated.

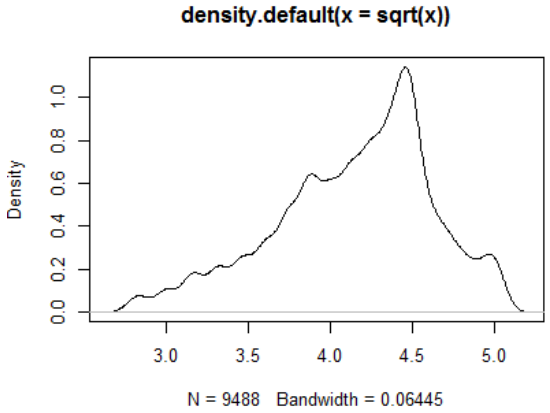
```
#Checking for Multicollinearity
cormat<- cor(final_data[,c(90:95)])
cormat # highly correlated
corrplot(cormat, method="circle")
corrplot(cormat, method="circle", addCoef.col="black")
dev.off()
# A more precise analysis is using VIF (Variance Inflation Factor)
# It looks for multicollinearity in a model as a whole
vif(mod) #Anything less than 4 is okay!
sqrt(vif(mod)) > 2 # if any variable is true, we would need to drop it
#None of the variables need to be dropped based on VIF
```

```
> sqrt(vif(mod)) > 2 # if any variable is true, we would need to drop it
     Leadership           PEval Training_PGrowth       Diversity     OrgFairness     Compensation            DLOC
          FALSE           FALSE           FALSE           FALSE           FALSE           FALSE           FALSE
           DSEX
          FALSE
> #None of the variables need to be dropped based on VIF
> transforms(final_no_outliers_overall$Overall_Job_Satisfaction)
>
```

Now, let's look back if there is any change in the distribution of Dependent variable after removing the outliers.







Dependent variable Transformations

Transformations after dropping outliers.

A 'little' better distribution then before

Model 3:

```
mod_2 <- lm(Overall_Job_Satisfaction ~ Leadership+PEval+Training_PGrowth+Diversity+
      +Compensation+DLOC, data=final_data)
summary(mod_2)
```

This is final model, which we got the best output among every other model.

In this model, after dropping outliers, we have seen significant *improve in decrease of Residual error value, which makes us a better fit model and also, increase in adjusted R^2 value*.

```
Call:
lm(formula = Overall_Job_Satisfaction ~ Leadership + PEval +
    Training_PGrowth + Diversity + +Compensation + DLOC, data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0745 -0.9742  0.0182  1.0041  7.6872

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.457158   0.134821  -3.391 0.000700 ***
Leadership      0.279244   0.004236  65.927  < 2e-16 ***
PEval           0.144799   0.005027  28.805  < 2e-16 ***
Training_PGrowth  0.245921   0.004726  52.040  < 2e-16 ***
Diversity      -0.117798   0.010295 -11.442  < 2e-16 ***
Compensation    0.052201   0.005288   9.871  < 2e-16 ***
DLOCB           0.124049   0.036933   3.359 0.000786 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.678 on 9481 degrees of freedom
Multiple R-squared:  0.8126,    Adjusted R-squared:  0.8125
F-statistic:  6852 on 6 and 9481 DF,  p-value: < 2.2e-16
```

Linear regression model interpretation:

- When all the IVs are zero, the overall job satisfaction index value is -0.457
- Percentage of variation in the response variable is 81.25%
- For 4 units increase in Leadership index, 1 unit of overall job satisfaction increases.
- For 7 units increase in Performance Evaluation Index, 1 unit of Job satisfaction increases.
- For 20 units increase in Compensation, 1 unit job satisfaction index increases.
- Location has similar effect as Evaluation does, and Training has similar effect as Leadership does.
- Diversity factor has Negative impact on job satisfaction Index.
- For 10 units increase in Diversity index, 1 unit of job satisfaction index decreases.
- for any additional change of values, it depicts 13% standard error rate.
- As F-static is the ratio of How well the model is doing to error values, 6852 values indicate that the model doing well.
- as Adjusted ‘R square’ value increases and Residual error value decreases over the three models, this is the best we got among all three.
- DSEX variable has less significant effect on overall model.

Performance Evaluation

We have performed performance evaluation of the 3 models we have constructed. Here is the comparison and the summary of these.

Model Comparison	Adjusted R square value	Residual Error value	F-static value	Estimate Intercept
Model 1	0.803	2.03	5600	-2.2765
Model 2	0.8064	2.012	6675	-1.7569
Model 3 (Eliminating outliers)	0.8125	1.678	6852	-0.4572

Best Values

- We have got 1.2% increase in Adjusted R square value in model 3 when compared to model 1.
- Residual error had been reduced by 17.5% in model 3.
- 22% increase in F-static value over model 1.
- Intercept value decreased by 80% value in model 3 when compared to model 1.

Summary

- Satisfactory with Leadership, Training and Personal Growth are major factors to contribute Highest Overall Job satisfaction rate in Top 4 agencies.
- People are not satisfied when the organization has maintained diversity management.
- Compensation has less significant effect on overall job satisfaction when compared to Leadership and Training parameters.

"In Top 4 Agencies, to increase overall Job satisfaction, Organization needs to put efforts on Employee- Manager satisfaction and organizations should focus on Employee Training and their personal Growth, professionally."


THANK YOU!!