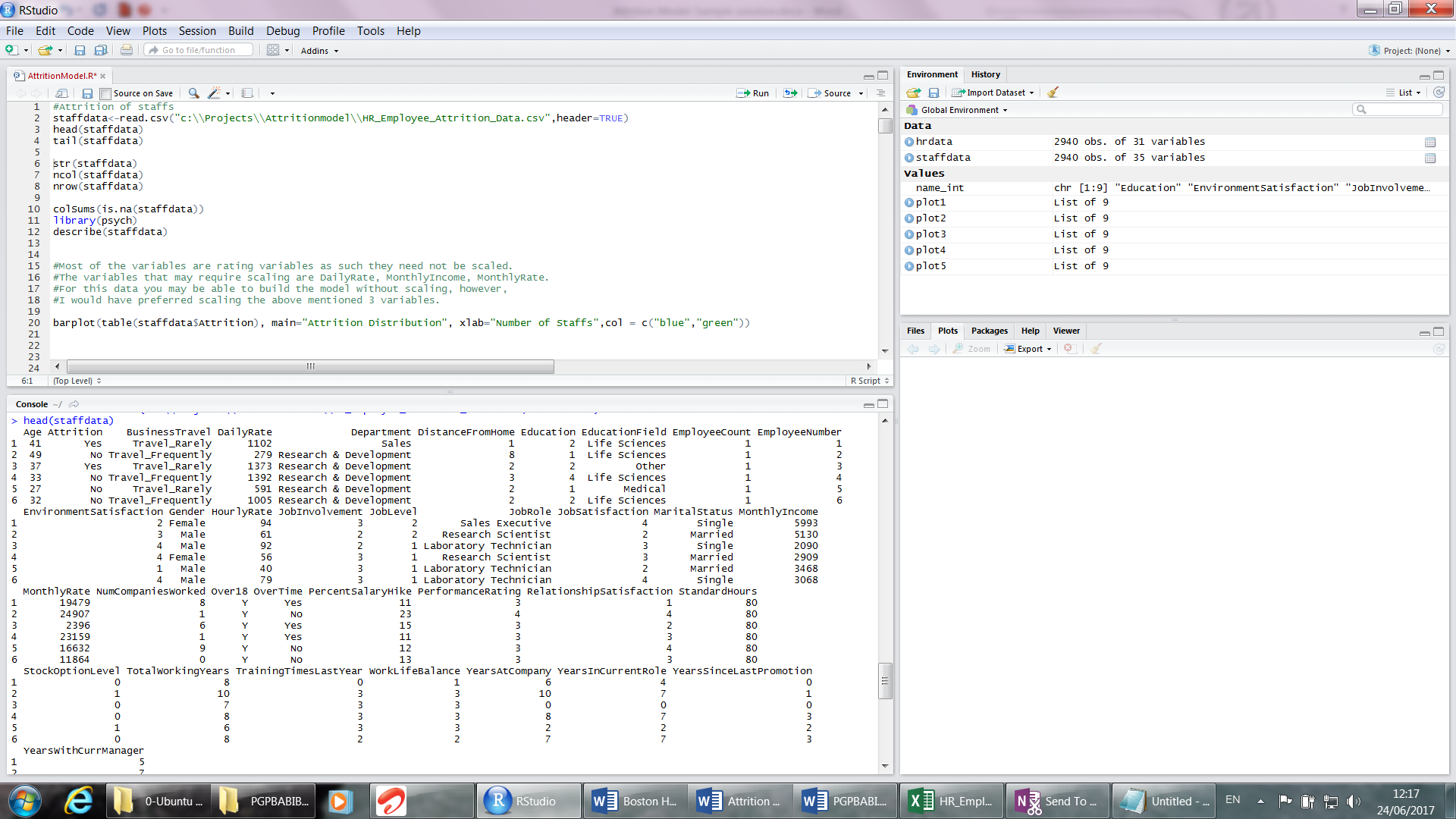
**HR Employee Attrition Model in R**

Attrition dataset contain Employee data and a number of predictors determining the Atrrition.Let us load the dataset into R environment from CSV file as "staffdata".

> staffdata<-read.csv("c:\\Projects\\Attritionmodel\\HR\_Employee\_Attrition\_Data.csv",header=TRUE)

> head(staffdata)



The variables description are below.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age | age of employee (continuous variable) |
| Attrition | describe status of attrition (dichotomous variable, values – Yes, No) |
| BusinessTravel | Describe official travelling of employees (categorical variable having 3 Labels Non-Travel, Travel\_Frequently, Travel\_Rarely) |
| DailyRate | Daily rate of employee (continuous variable) |
| Department | Department of employee (categorical variable having 3 labels - Human Resources, Research & Development, Sales) |
| DistanceFromHome | distance of office from home in km (continuous variable) |
| Education | Education of employee in rating (categorical variables ,ordinal data, 5 Ratings 1, 2, 3, 4, 5) |
| EducationField | Education Stream of employee (categorical variable having 6 labels-Human Resources, Life Sciences, Marketing, Medical, Other, Technical Degree) |
| EmployeeCount | Count of employee (Continuous variable and only 1 value present like a Constant) |
| EmployeeNumber | Employee identification number an ID variable |
| EnvironmentSatisfaction | Job Environment Stratification of employee (rating scale data 1, 2, 3, 4 categorical variable nominal/ordinal data) |
| Gender | Describe gender of employee Categorical variable (2 labels Female, Male) |
| HourlyRate | describe hourly rate of employee (Continuous variable) |
| JobInvolvement | Engagement in the jog (Rating data having 4 ratings 1,2,3,4 categorical variable) |
| JobLevel | Level of job in organization structure (Rating data having 5 ratings 1, 2,3,4,5 categorical variable) |
| JobRole | Healthcare Representative,Human Resources,Laboratory Technician,Manager,Manufacturing Director,Research Director,Research Scientist,Sales Executive,Sales Representative |
| JobSatisfaction | Stratification of employee within job (rating scale data 1, 2, 3, 4 categorical variable nominal/ordinal data) |
| MaritalStatus | Marital Status(Categorical Variable having 3 labels - Divorced,Married,Single) |
| MonthlyIncome | monthly income of employee(Continuous variable) |
| MonthlyRate | Monthly rate of employee(Continuous variable) |
| NumCompaniesWorked | Number of company employee worked for(Count data having 0,1,2,3,4,5,6,7,8,9, continuous variable) |
| Over18 | Status of adult employee(categorical variable having 1 category - Y) |
| OverTime | OverTime status(categorical variable having 2 category - Yes, No) |
| PercentSalaryHike | Salary hike %age (continuous variable Percentage figures) |
| PerformanceRating | Performance rating (Rating data ,Categorical nominal/ordinal data- 3,4) |
| RelationshipSatisfaction | Shows status of Relationship with Managers(Rating data 4 labels ,Categorical nominal/ordinal data-1,23,4) |
| StandardHours | Standard working hours (Numerical data , 80 hours for everybody, almost constant data) |
| StockOptionLevel | stock option(Rating data 4 labels ,Categorical nominal/ordinal data-1,23,4) |
| TotalWorkingYears | Total year of experience(Continuous variable) |
| TrainingTimesLastYear | Training Time(Continuous variable) |
| WorkLifeBalance | WorkLifeBalance : rating 1,2,3,4 |
| YearsAtCompany | Years in current company(Continuous variable) |
| YearsInCurrentRole | Years in current Role(Continuous variable) |
| YearsSinceLastPromotion | Years Since last Promotion(Continuous variable) |
| YearsWithCurrManager | Years in current Managerial role(Continuous variable) |

Here our target variable is Attrition which is categorical in nature.So we require a classification technique to predict the attrition.All other are predictor variables where as Employeee number is Id Variable.

Let us see the dataset structure and no of variables in dataset, no of observations and missing values.

> str(staffdata)

'data.frame': 2940 obs. of 35 variables:

$ Age : int 41 49 37 33 27 32 59 30 38 36 ...

$ Attrition : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...

$ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 3 2 3 2 3 2 3 3 2 3 ...

$ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...

$ Department : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...

$ DistanceFromHome : int 1 8 2 3 2 2 3 24 23 27 ...

$ Education : int 2 1 2 4 1 2 3 1 3 3 ...

$ EducationField : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...

$ EmployeeCount : int 1 1 1 1 1 1 1 1 1 1 ...

$ EmployeeNumber : int 1 2 3 4 5 6 7 8 9 10 ...

$ EnvironmentSatisfaction : int 2 3 4 4 1 4 3 4 4 3 ...

> ncol(staffdata)

[1] 35

> nrow(staffdata)

[1] 2940

> head(staffdata)

Age Attrition BusinessTravel DailyRate Department DistanceFromHome Education EducationField EmployeeCount EmployeeNumber

1 41 Yes Travel\_Rarely 1102 Sales 1 2 Life Sciences 1 1

2 49 No Travel\_Frequently 279 Research & Development 8 1 Life Sciences 1 2

> tail(staffdata)

Age Attrition BusinessTravel DailyRate Department DistanceFromHome Education EducationField EmployeeCount

2935 26 No Travel\_Rarely 1167 Sales 5 3 Other 1

In our dataset there are 2940 obseravtions and 35 variables.

Summary() function gives the min,1st quartile,median,3rd quartile,max and also report missing Observations.Describe() function gives some more statistics.Variablename\* are the categorical variables.Min,max,count(n) statistics are more relevant for categorical variables.

> library(psych)

> describe(staffdata)

vars n mean sd median trimmed mad min max range skew kurtosis se

Age 1 2940 36.92 9.13 36.0 36.47 8.90 18 60 42 0.41 -0.41 0.17

Attrition\* 2 2940 1.16 0.37 1.0 1.08 0.00 1 2 1 1.84 1.39 0.01

BusinessTravel\* 3 2940 2.61 0.67 3.0 2.76 0.00 1 3 2 -1.44 0.69 0.01

DailyRate 4 2940 802.49 403.44 802.0 803.83 510.01 102 1499 1397 0.00 -1.21 7.44

Department\* 5 2940 2.26 0.53 2.0 2.25 0.00 1 3 2 0.17 -0.40 0.01

DistanceFromHome 6 2940 9.19 8.11 7.0 8.08 7.41 1 29 28 0.96 -0.23 0.15

Education 7 2940 2.91 1.02 3.0 2.98 1.48 1 5 4 -0.29 -0.56 0.02

EducationField\* 8 2940 3.25 1.33 3.0 3.10 1.48 1 6 5 0.55 -0.69 0.02

EmployeeCount 9 2940 1.00 0.00 1.0 1.00 0.00 1 1 0 NaN NaN 0.00

EmployeeNumber 10 2940 1470.50 848.85 1470.5 1470.50 1089.71 1 2940 2939 0.00 -1.20 15.66

EnvironmentSatisfaction 11 2940 2.72 1.09 3.0 2.78 1.48 1 4 3 -0.32 -1.20 0.02

Gender\* 12 2940 1.60 0.49 2.0 1.62 0.00 1 2 1 -0.41 -1.83 0.01

HourlyRate 13 2940 65.89 20.33 66.0 66.02 26.69 30 100 70 -0.03 -1.20 0.37

JobInvolvement 14 2940 2.73 0.71 3.0 2.74 0.00 1 4 3 -0.50 0.26 0.01

JobLevel 15 2940 2.06 1.11 2.0 1.90 1.48 1 5 4 1.02 0.39 0.02

JobRole\* 16 2940 5.46 2.46 6.0 5.61 2.97 1 9 8 -0.36 -1.19 0.05

JobSatisfaction 17 2940 2.73 1.10 3.0 2.79 1.48 1 4 3 -0.33 -1.22 0.02

MaritalStatus\* 18 2940 2.10 0.73 2.0 2.12 1.48 1 3 2 -0.15 -1.12 0.01

MonthlyIncome 19 2940 6502.93 4707.16 4919.0 5667.24 3260.24 1009 19999 18990 1.37 1.00 86.81

MonthlyRate 20 2940 14313.10 7116.58 14235.5 14286.48 9201.76 2094 26999 24905 0.02 -1.22 131.25

NumCompaniesWorked 21 2940 2.69 2.50 2.0 2.36 1.48 0 9 9 1.02 0.00 0.05

Over18\* 22 2940 1.00 0.00 1.0 1.00 0.00 1 1 0 NaN NaN 0.00

OverTime\* 23 2940 1.28 0.45 1.0 1.23 0.00 1 2 1 0.96 -1.07 0.01

PercentSalaryHike 24 2940 15.21 3.66 14.0 14.80 2.97 11 25 14 0.82 -0.31 0.07

PerformanceRating 25 2940 3.15 0.36 3.0 3.07 0.00 3 4 1 1.92 1.68 0.01

RelationshipSatisfaction 26 2940 2.71 1.08 3.0 2.77 1.48 1 4 3 -0.30 -1.19 0.02

StandardHours 27 2940 80.00 0.00 80.0 80.00 0.00 80 80 0 NaN NaN 0.00

StockOptionLevel 28 2940 0.79 0.85 1.0 0.67 1.48 0 3 3 0.97 0.36 0.02

TotalWorkingYears 29 2940 11.28 7.78 10.0 10.37 5.93 0 40 40 1.12 0.91 0.14

TrainingTimesLastYear 30 2940 2.80 1.29 3.0 2.72 1.48 0 6 6 0.55 0.49 0.02

WorkLifeBalance 31 2940 2.76 0.71 3.0 2.77 0.00 1 4 3 -0.55 0.41 0.01

YearsAtCompany 32 2940 7.01 6.13 5.0 5.99 4.45 0 40 40 1.76 3.91 0.11

YearsInCurrentRole 33 2940 4.23 3.62 3.0 3.85 4.45 0 18 18 0.92 0.47 0.07

YearsSinceLastPromotion 34 2940 2.19 3.22 1.0 1.48 1.48 0 15 15 1.98 3.59 0.06

YearsWithCurrManager 35 2940 4.12 3.57 3.0 3.77 4.45 0 17 17 0.83 0.16 0.07

> colSums(is.na(staffdata))

Age Attrition BusinessTravel DailyRate Department

0 0 0 0 0

DistanceFromHome Education EducationField EmployeeCount EmployeeNumber

0 0 0 0 0

EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel

0 0 0 0 0

JobRole JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate

0 0 0 0 0

NumCompaniesWorked Over18 OverTime PercentSalaryHike PerformanceRating

0 0 0 0 0

RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear

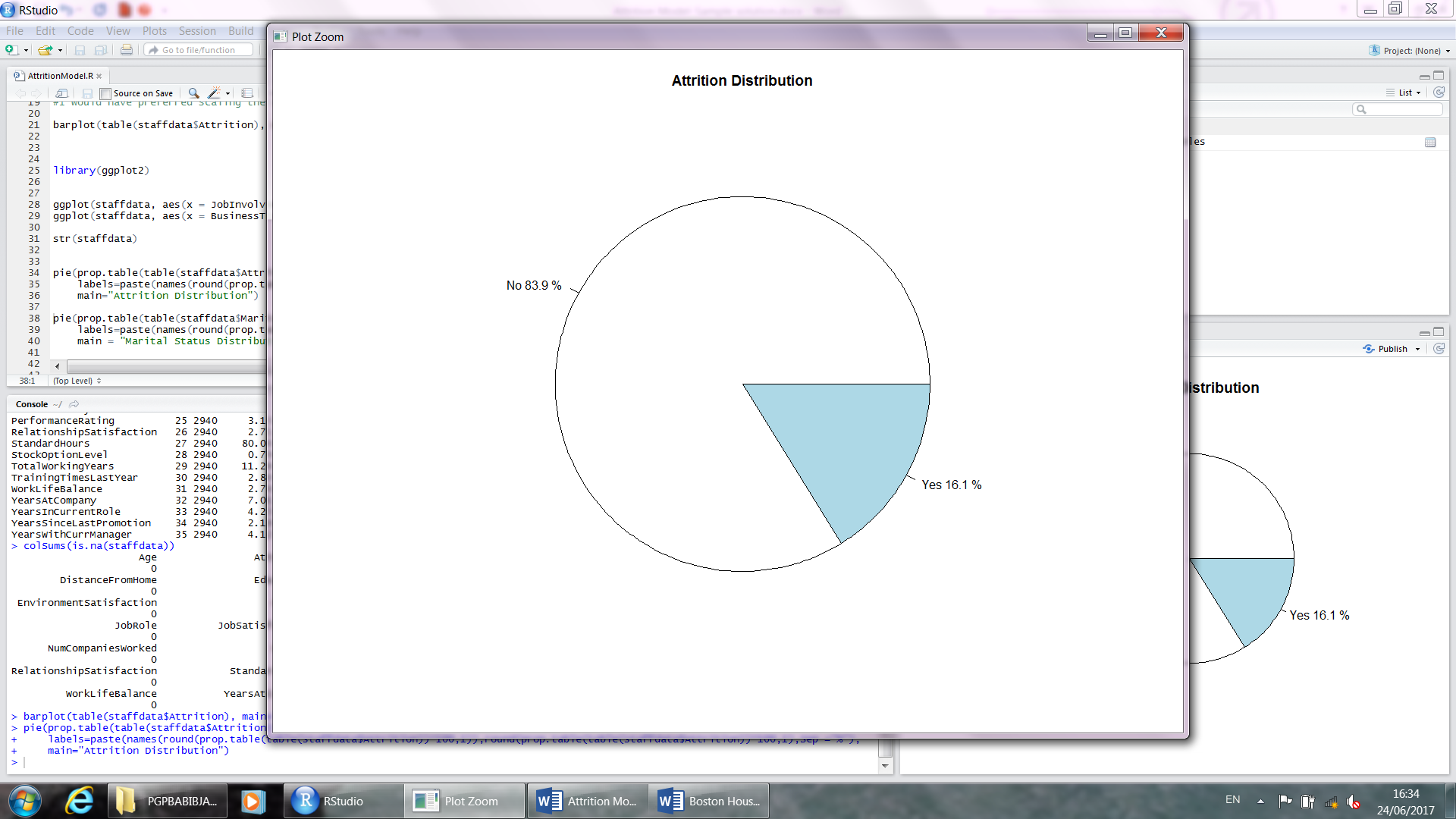
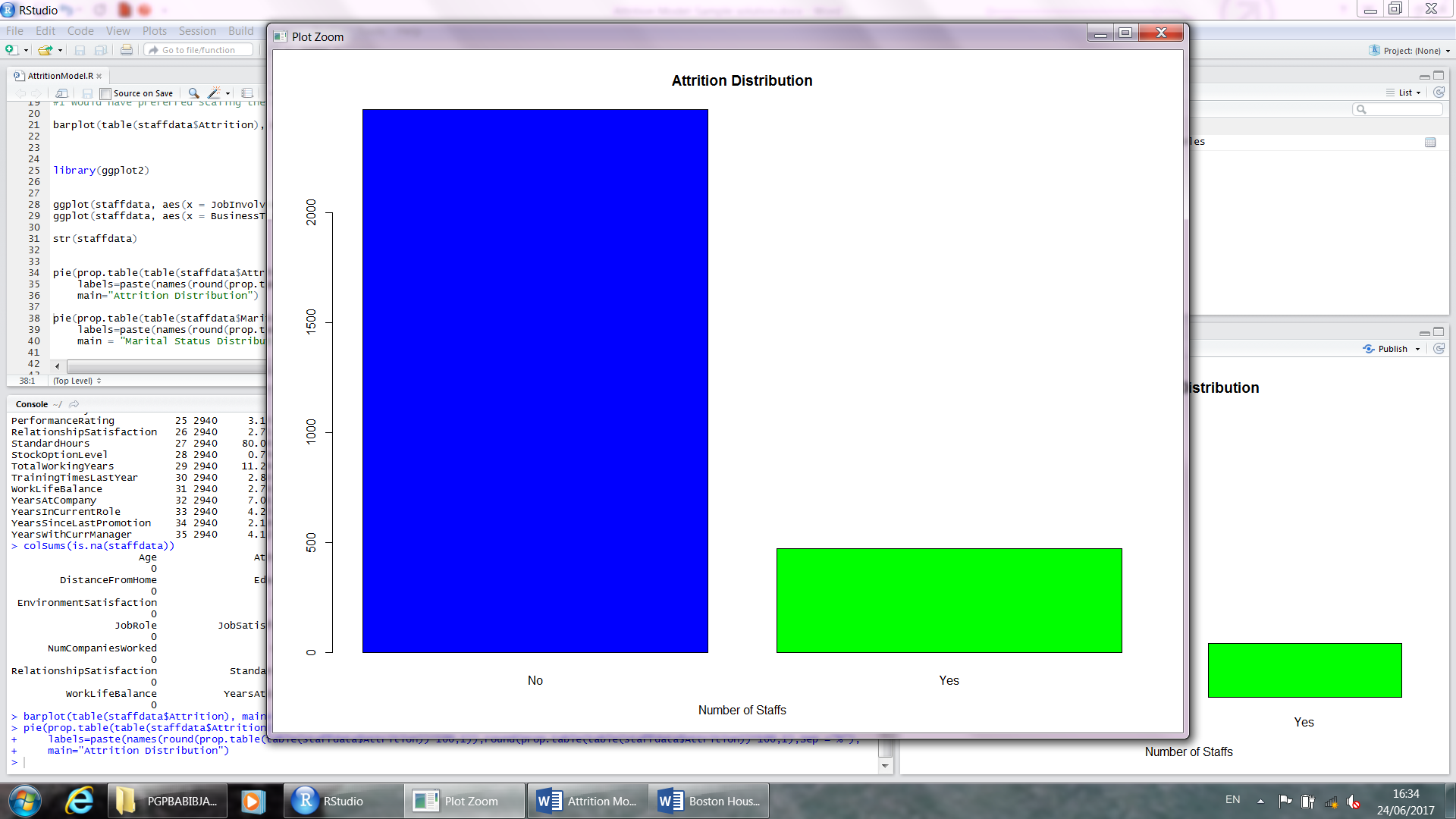
0 0 0 0 0

WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager

0 0 0 0 0

There is no missing data present. If any missing data present then you treat the missing data. If target variable missing then remove the observation. Observation remove is not allowed for predictors missing. For predictors missing impute the value. For categorical predictor, impute mode value .For continuous vaiable,impute median value as median is not influences by outliers. You may use mean imputation, knn imputation, regression imputation based upon actual data. It is little subjective based upon actual data. But you must check missing values and impute the missing values. You must not built any model with missing data.

Let us check Target variable distribution.



> prop.table(table(staffdata$Attrition))

No Yes

0.8387755 0.1612245

> table(staffdata$Attrition)

No Yes

2466 474

There are 474 observation of attarition and 2466 no of non-attrition observations.in our dataset 16.1% observations are attrition. Our event is attrition and we will predict this event by a classification technique.

Let us make some Hypothesis.

1. Employee having low increment tends to higher attrition rate.
2. Bachelors having higher attarition rate than Married staff.
3. Frequently travelling and non-traveling employee having higher attrition rate than rarely-traveling employee.
4. High qualified employee are more attired than low qualified employee.
5. Male employee perform more than female employee.
6. Higher is the salary lower is the attrition.
7. Stable jobs and job stratified employee are lesser attired.

If some employee got less increment than other employee then he/she get dissatisfied with the increment and then she/he try to change job. In reality this is true .bachelors can move from one city to other city and they can change a job easily.

Let us test the hypothesis male employee perform more than female.

The null hypothesis there is that the difference in mean performance of male and female is zero. Alternative there is significance difference in mean performance of male and female.

Do a 2 sample t-test. PerformanceRating already coded 3,4 in our dataset.We can run ttest to test mean.

> t.test(subset(staffdata, Gender=="Male", select = c(PerformanceRating)),subset(staffdata, Gender=="Female", select = c(PerformanceRating)) )

Welch Two Sample t-test

data: subset(staffdata, Gender == "Male", select = c(PerformanceRating)) and subset(staffdata, Gender == "Female", select = c(PerformanceRating))

t = -0.74722, df = 2470.9, p-value = 0.455

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.03698266 0.01657450

sample estimates:

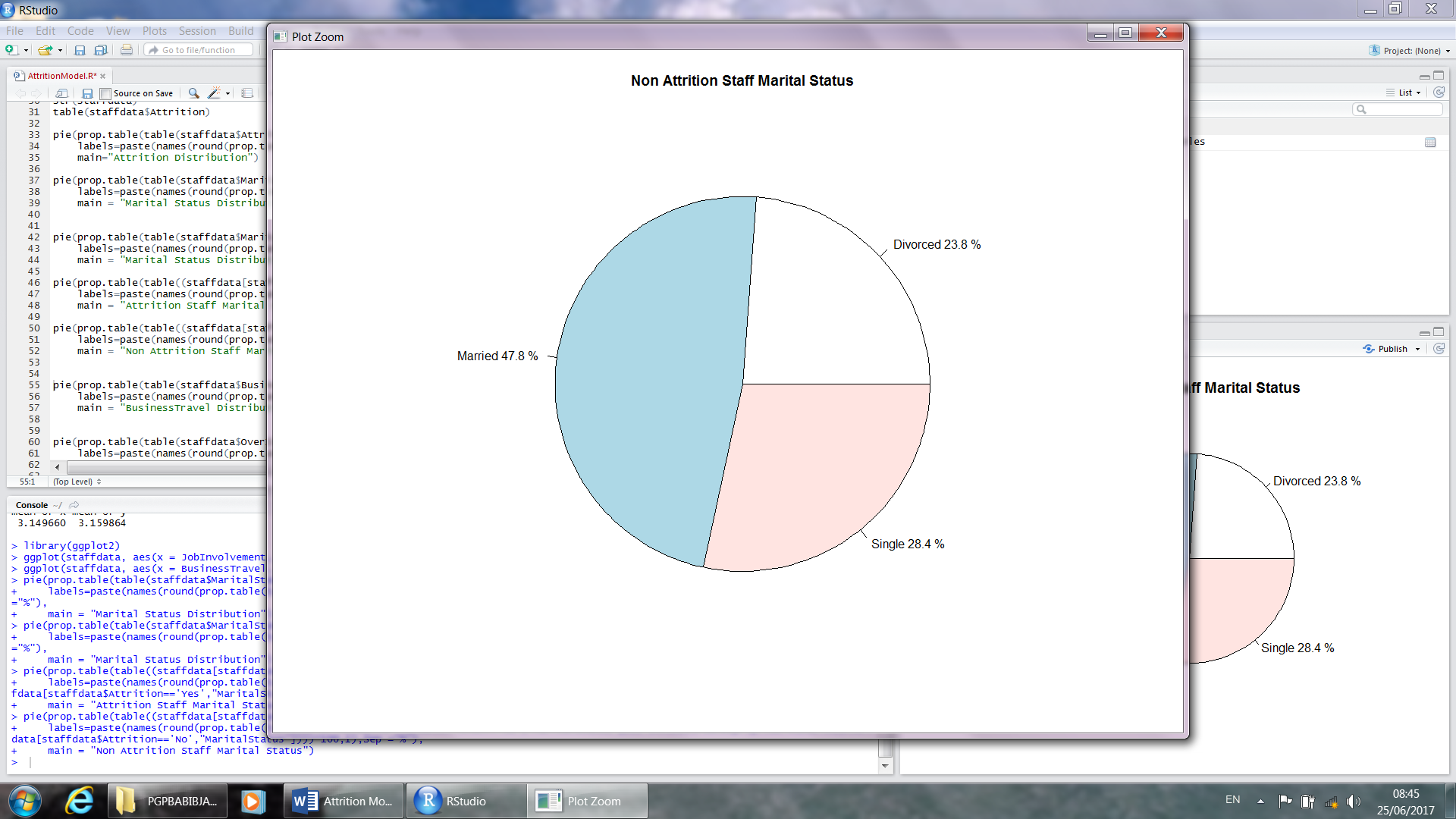
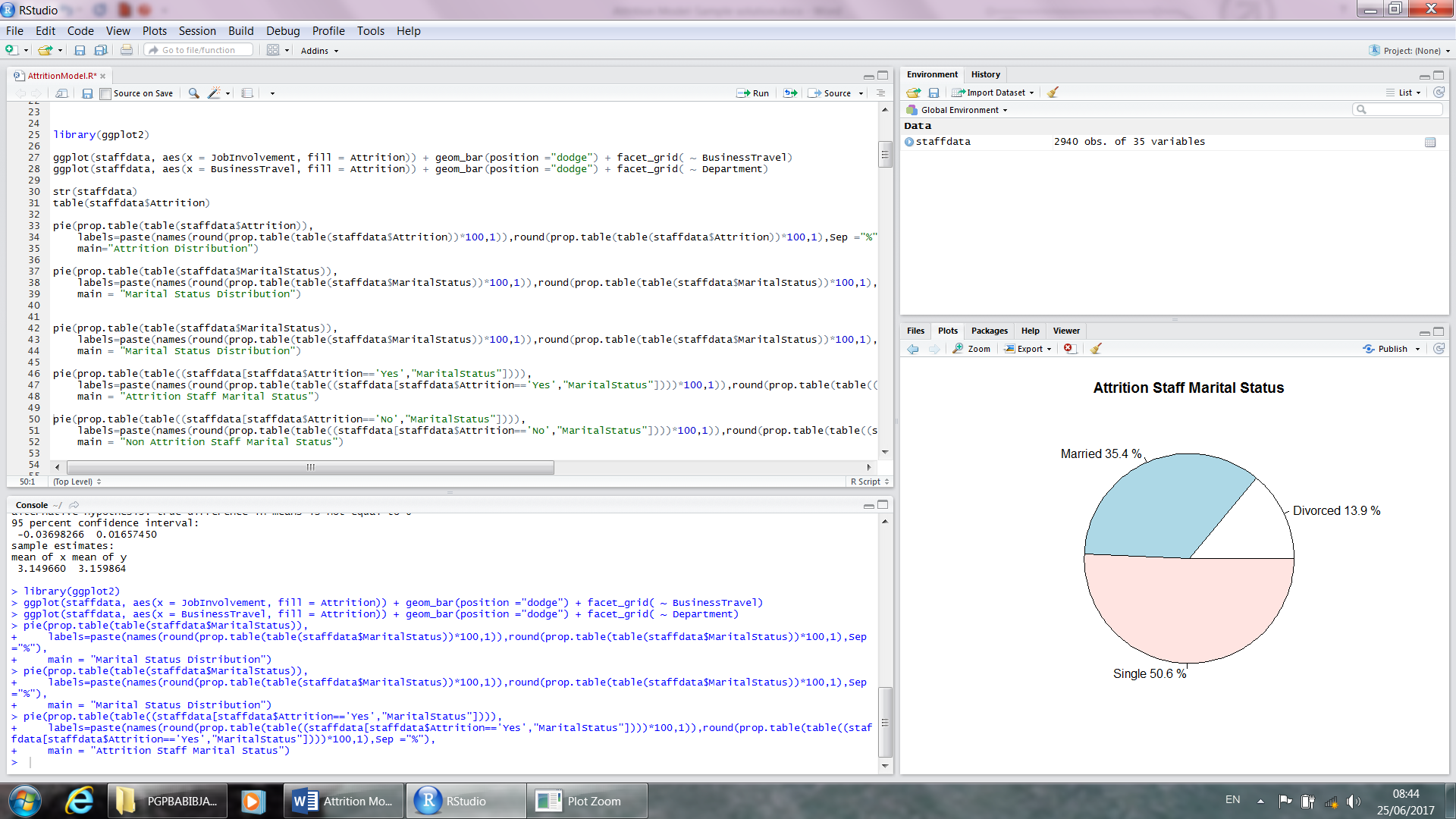
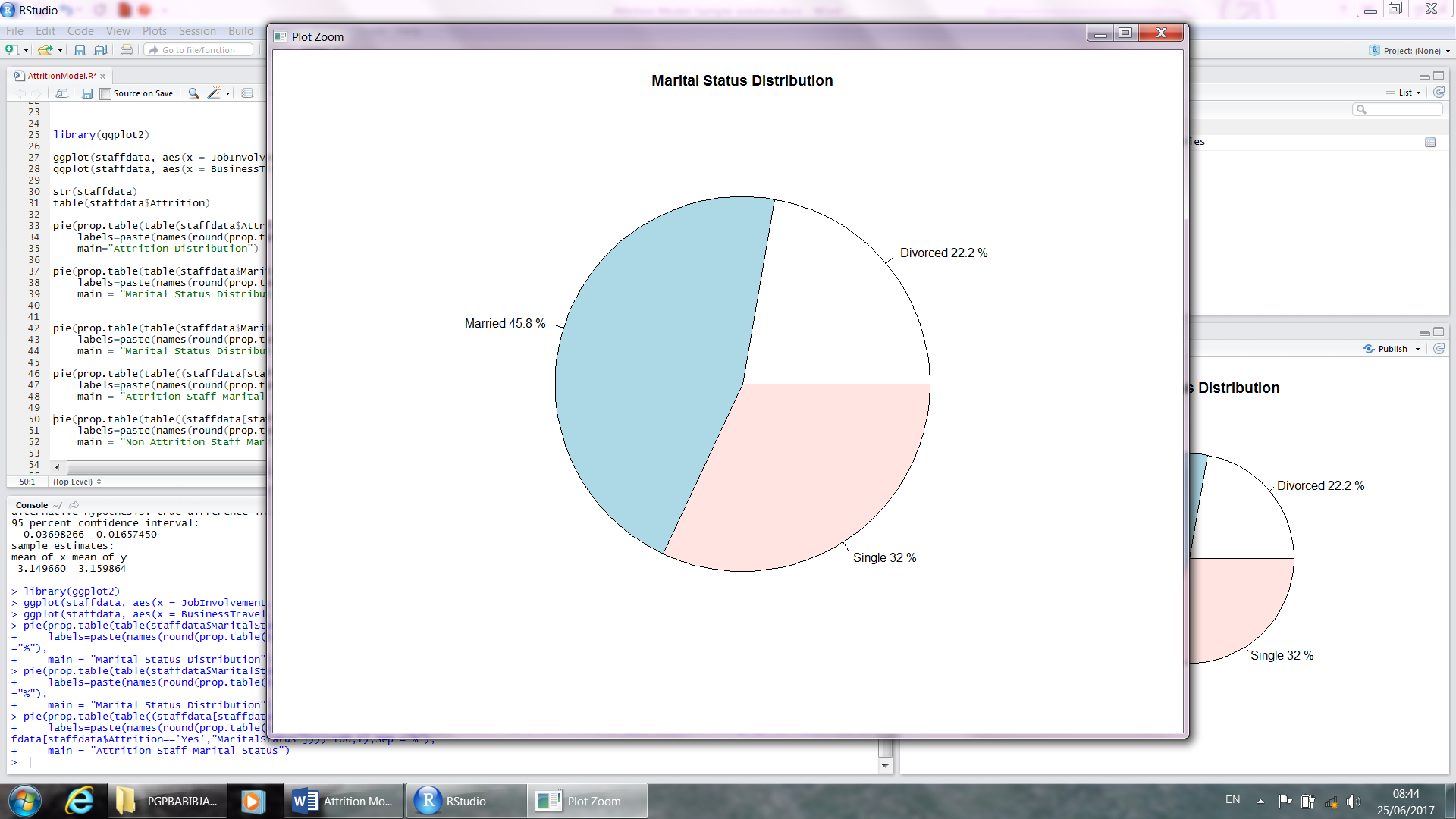
mean of x mean of y

3.149660 3.159864

Above ttest is insignificant. We conclude that the male and female employee perform equally.

Similarly we can test other hypothesis.

Let us explore marital status variable.



In our dataset 45.8% are married,27.2% are divorced,32% are single. In Attired staff 50.6% are single,35.4%are married and 13.9% are divorced. Single staff are attired more than married and divorced

> library(ggplot2)

> ggplot(staffdata, aes(x = JobInvolvement, fill = Attrition)) + geom\_bar(position ="dodge") + facet\_grid( ~ BusinessTravel)

> ggplot(staffdata, aes(x = BusinessTravel, fill = Attrition)) + geom\_bar(position ="dodge") + facet\_grid( ~ Department)

> pie(prop.table(table(staffdata$MaritalStatus)),

+ labels=paste(names(round(prop.table(table(staffdata$MaritalStatus))\*100,1)),round(prop.table(table(staffdata$MaritalStatus))\*100,1),Sep ="%"),

+ main = "Marital Status Distribution")

> pie(prop.table(table(staffdata$MaritalStatus)),

+ labels=paste(names(round(prop.table(table(staffdata$MaritalStatus))\*100,1)),round(prop.table(table(staffdata$MaritalStatus))\*100,1),Sep ="%"),

+ main = "Marital Status Distribution")

> pie(prop.table(table((staffdata[staffdata$Attrition=='Yes',"MaritalStatus"]))),

+ labels=paste(names(round(prop.table(table((staffdata[staffdata$Attrition=='Yes',"MaritalStatus"])))\*100,1)),round(prop.table(table((staffdata[staffdata$Attrition=='Yes',"MaritalStatus"])))\*100,1),Sep ="%"),

+ main = "Attrition Staff Marital Status")

> pie(prop.table(table((staffdata[staffdata$Attrition=='No',"MaritalStatus"]))),

+ labels=paste(names(round(prop.table(table((staffdata[staffdata$Attrition=='No',"MaritalStatus"])))\*100,1)),round(prop.table(table((staffdata[staffdata$Attrition=='No',"MaritalStatus"])))\*100,1),Sep ="%"),

+ main = "Non Attrition Staff Marital Status")

> pie(prop.table(table(staffdata$BusinessTravel)),

+ labels=paste(names(round(prop.table(table(staffdata$BusinessTravel))\*100,1)),round(prop.table(table(staffdata$BusinessTravel))\*100,1),Sep ="%"),

+ main = "BusinessTravel Distribution")

> pie(prop.table(table((staffdata[staffdata$Attrition=='Yes',"MaritalStatus"]))),

+ labels=paste(names(round(prop.table(table((staffdata[staffdata$Attrition=='Yes',"BusinessTravel"])))\*100,1)),round(prop.table(table((staffdata[staffdata$Attrition=='Yes',"BusinessTravel"])))\*100,1),Sep ="%"),

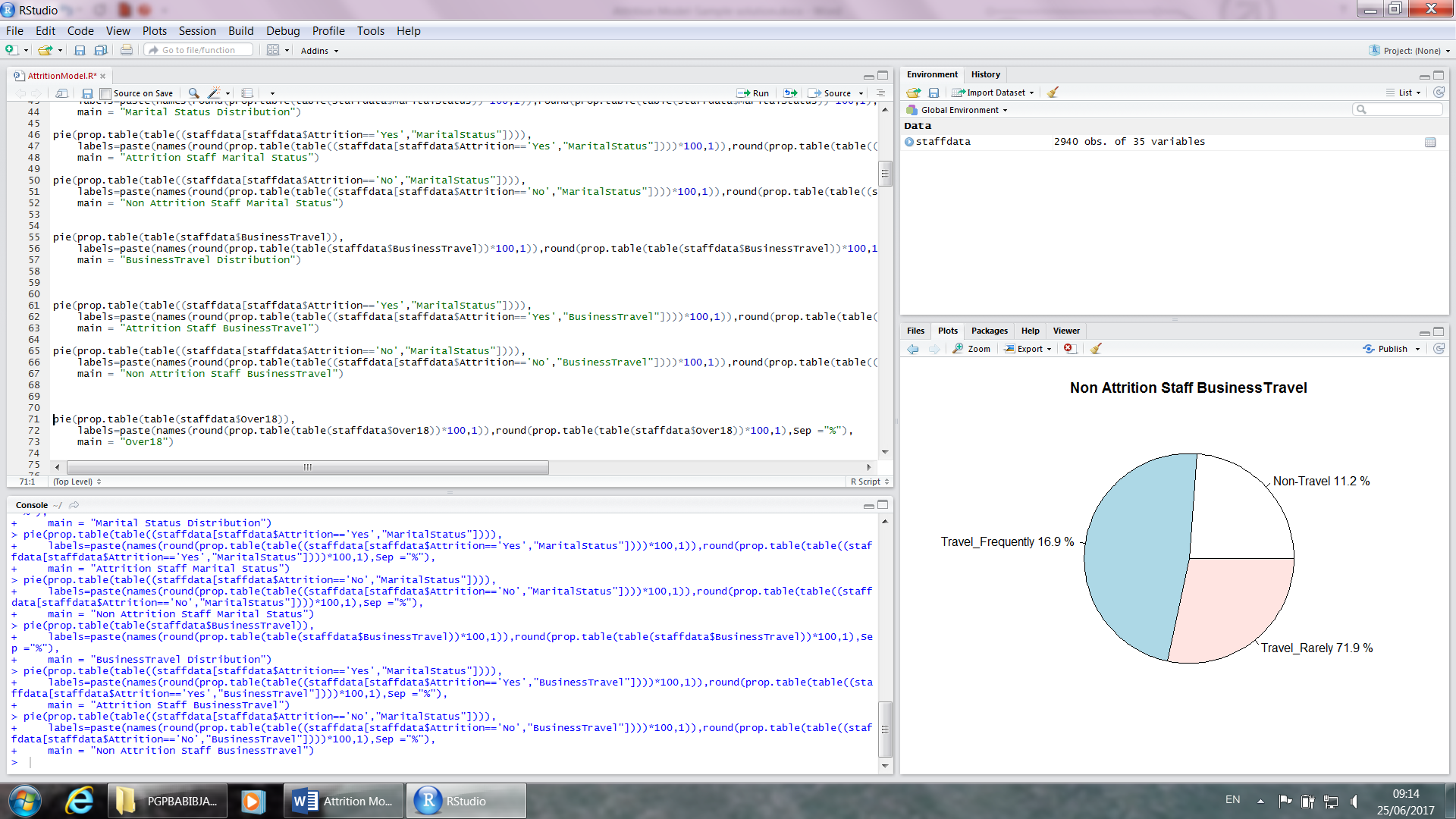
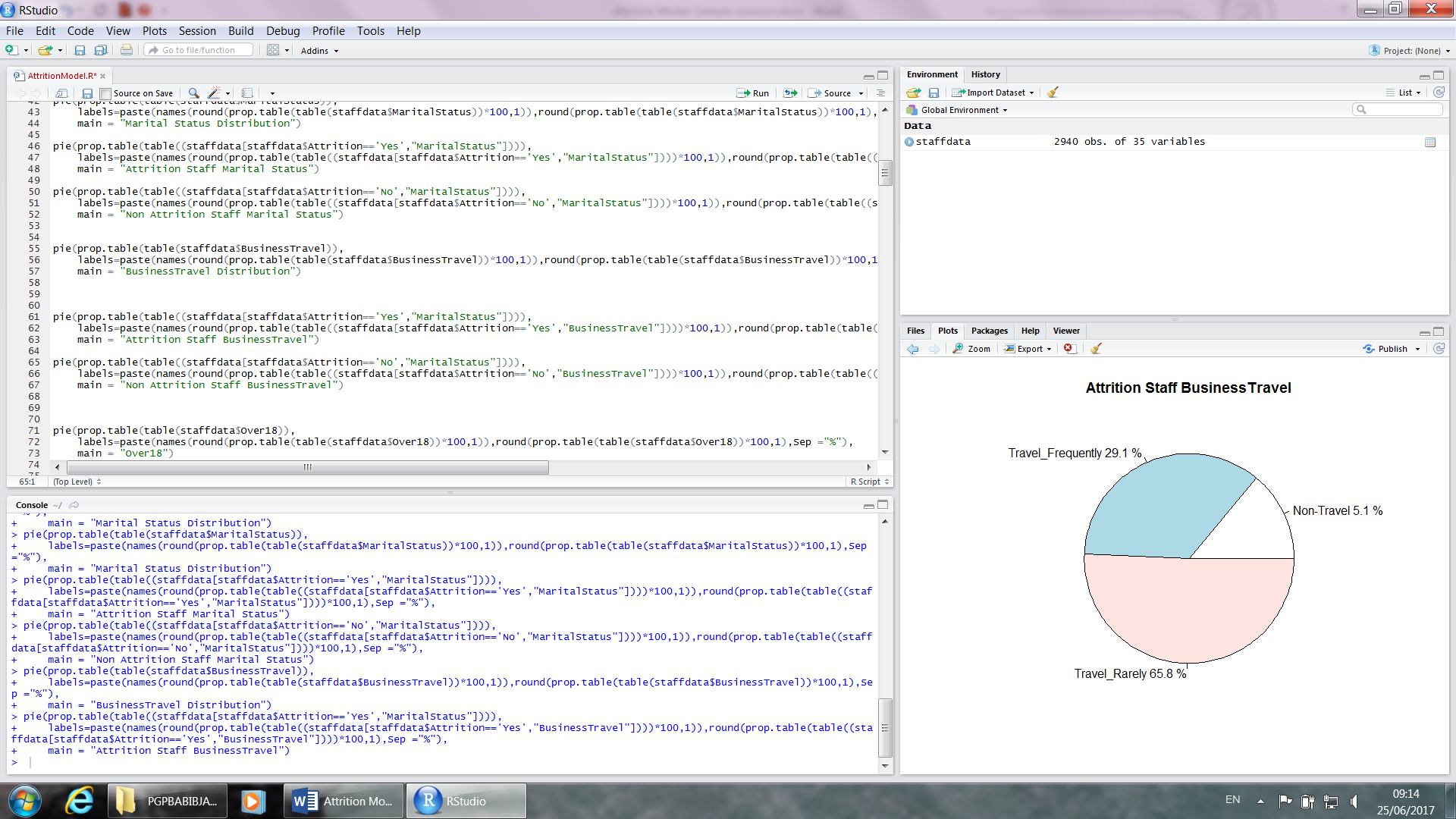
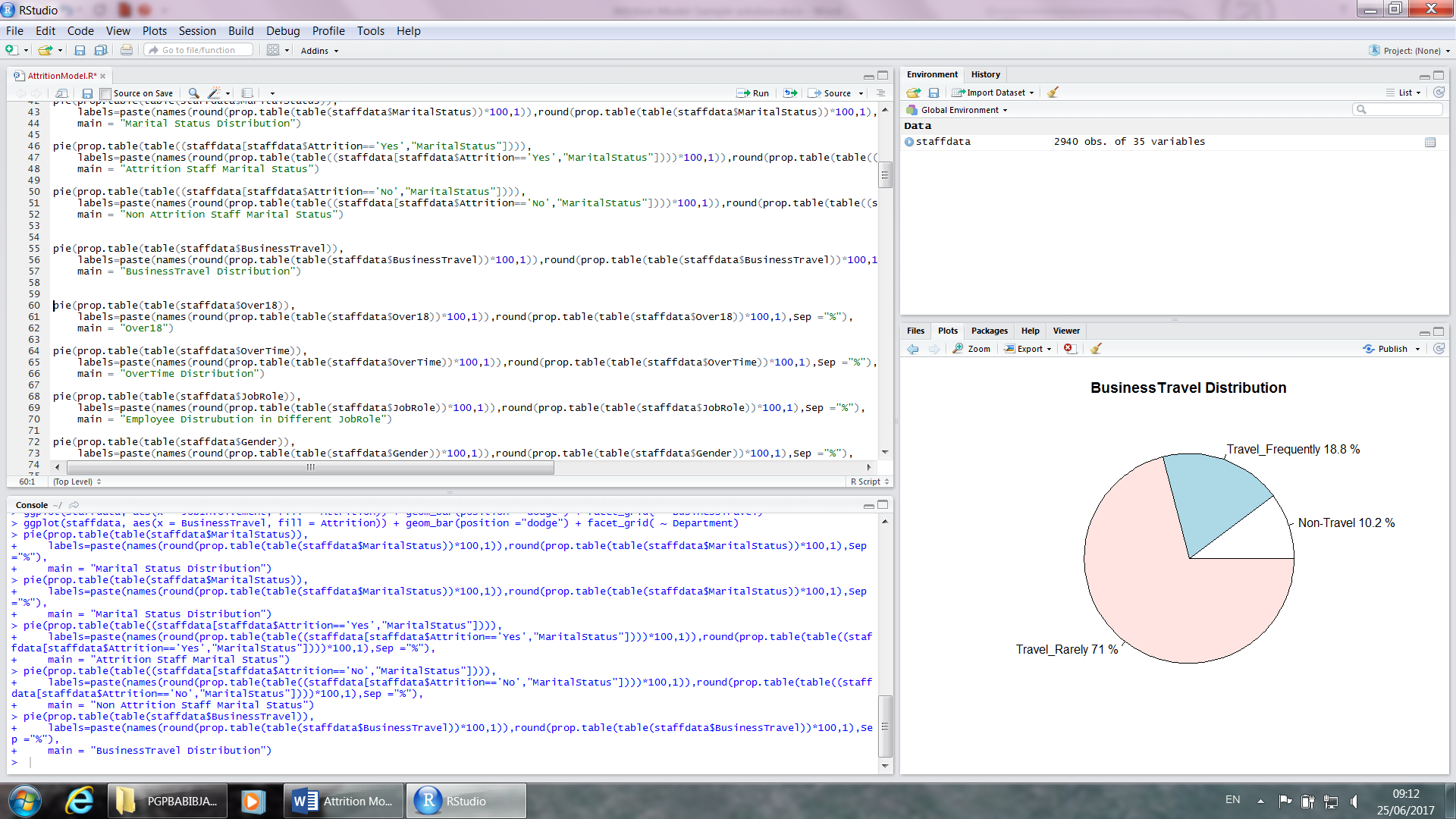
+ main = "Attrition Staff BusinessTravel")

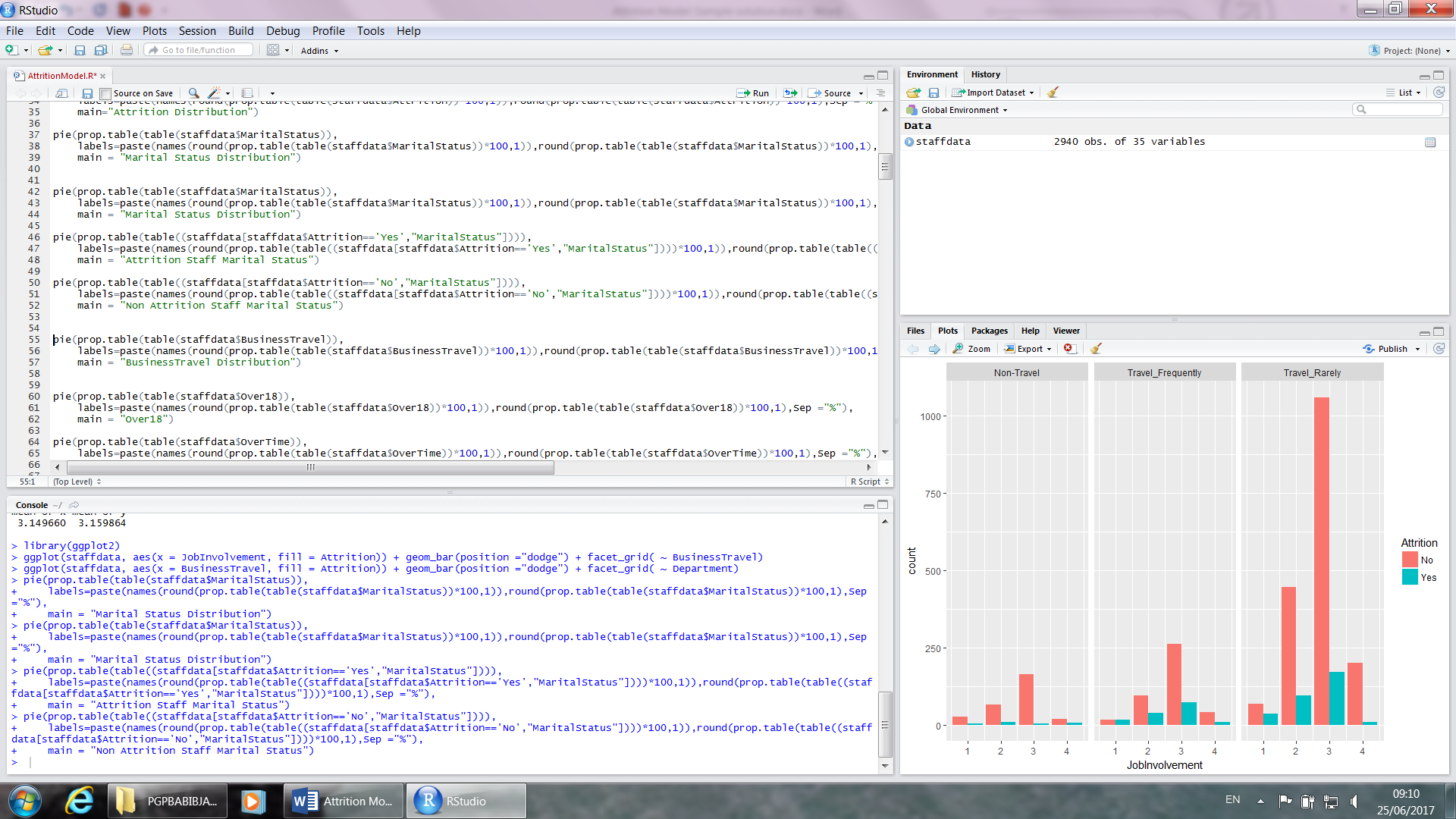
> pie(prop.table(table((staffdata[staffdata$Attrition=='No',"MaritalStatus"]))),

+ labels=paste(names(round(prop.table(table((staffdata[staffdata$Attrition=='No',"BusinessTravel"])))\*100,1)),round(prop.table(table((staffdata[staffdata$Attrition=='No',"BusinessTravel"])))\*100,1),Sep ="%"),

+ main = "Non Attrition Staff BusinessTravel")

let us explore BusinessTravel variable.

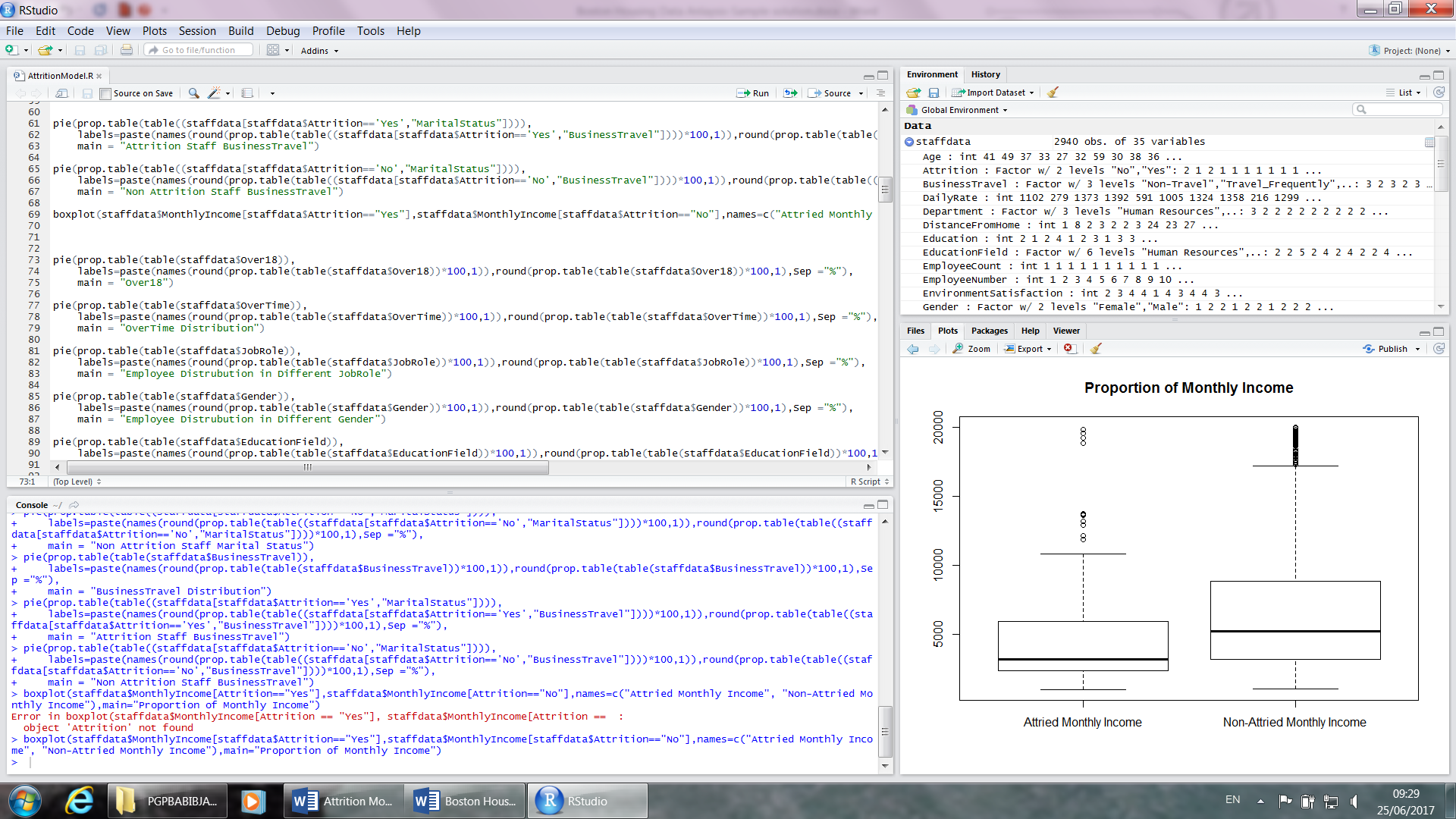




in our dataset 71% travel\_rarely,18.8% travel frequently,10.2% non-travel employee present. In tarried employees 65.8% are travel rarely. Rarely travel employee are more attired than travel frequently and non-travel employee. In hypothesis we assume that rarely travel employee attired less. But in reality it is opposite.

Let us explore Monthlyincome variable.

> boxplot(staffdata$MonthlyIncome[staffdata$Attrition=="Yes"],staffdata$MonthlyIncome[staffdata$Attrition=="No"],names=c("Attried Monthly Income", "Non-Attried Monthly Income"),main="Proportion of Monthly Income")



Above box plot shows mean income of attired staff is comparatively low than mean income of Non-Attired staff. We can conclude Salary play a major role in attrition.

Do EDA for all rest of predictors. Try BOX plot,barcharts ,Pie charts,find out story.

No use columns.

Employee number variable is identification variable. It is specific to employee.it doesn't contain any information values.we will drop it from our model.

> pie(prop.table(table(staffdata$Over18)),

+ labels=paste(names(round(prop.table(table(staffdata$Over18))\*100,1)),round(prop.table(table(staffdata$Over18))\*100,1),Sep ="%"),

+ main = "Over18")

> pie(prop.table(table(staffdata$StandardHours)),

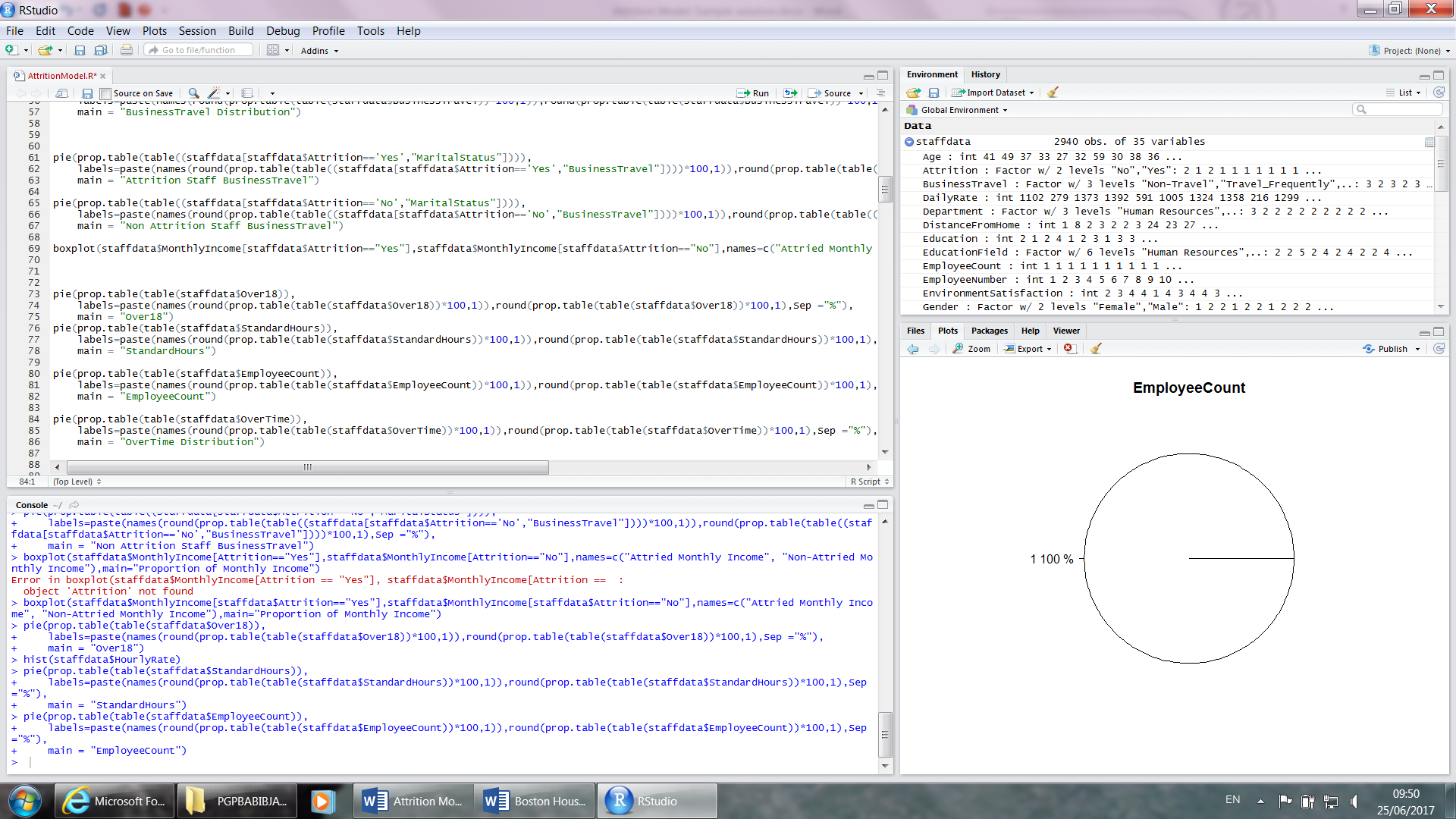
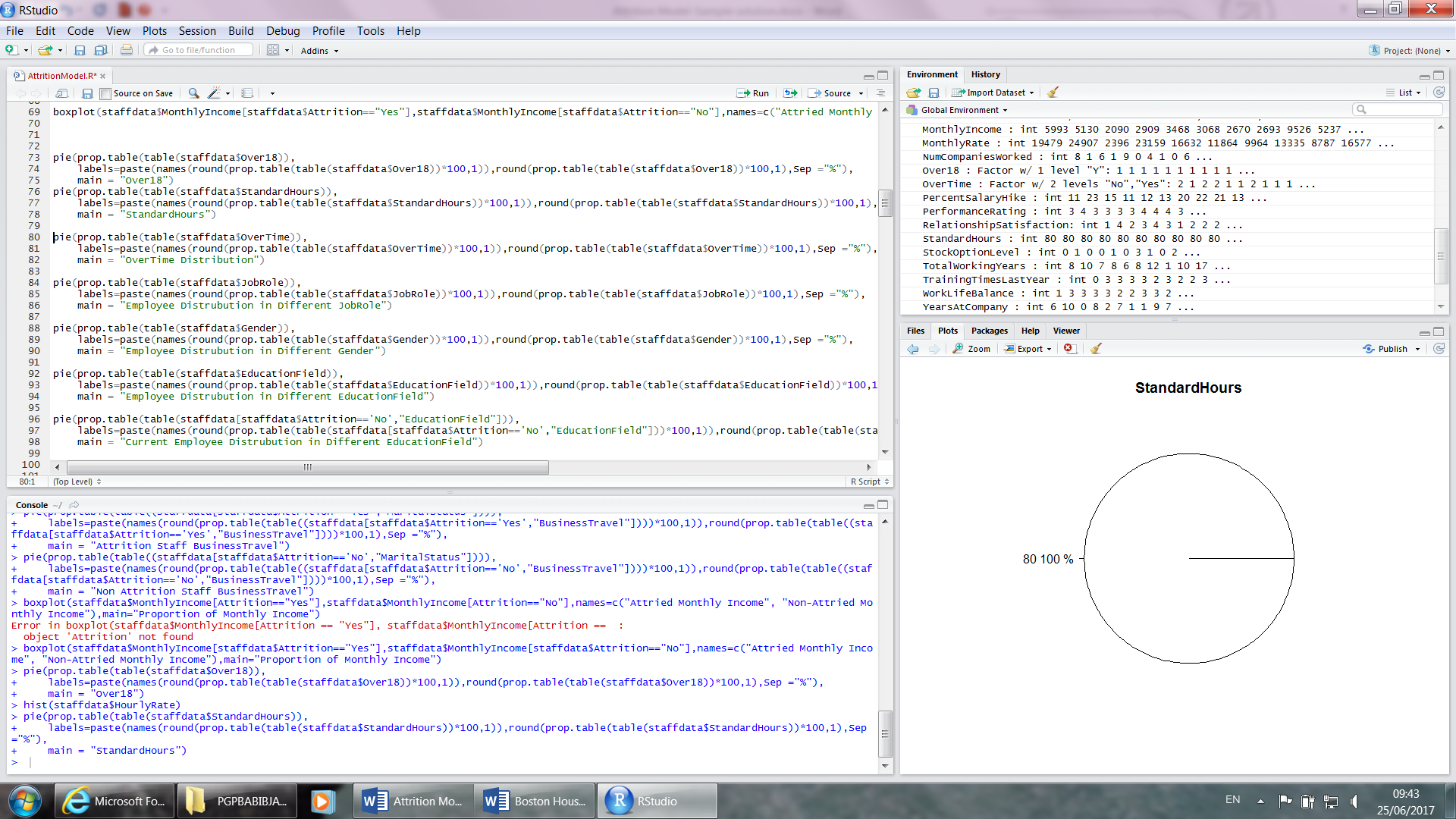
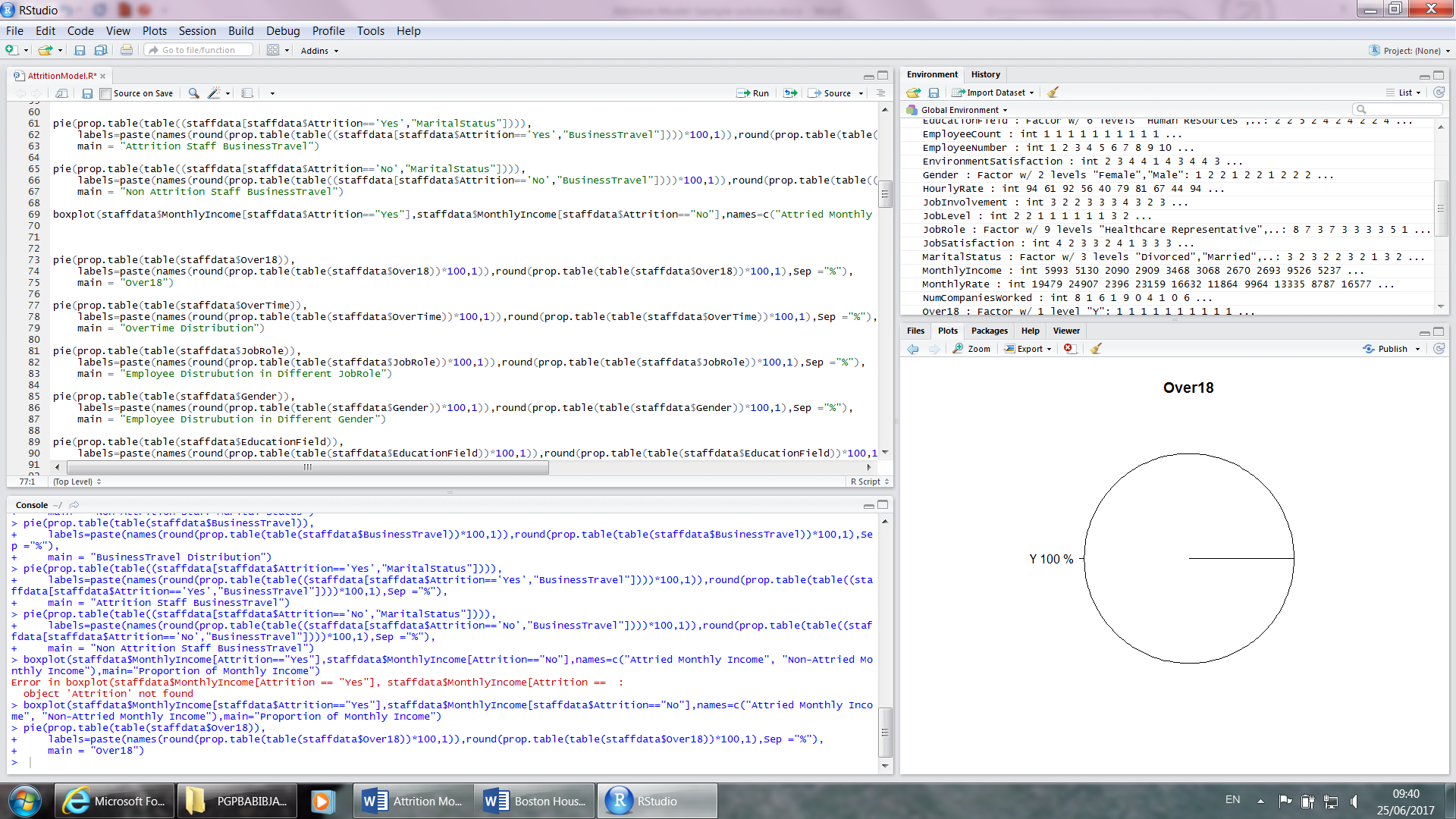
+ labels=paste(names(round(prop.table(table(staffdata$StandardHours))\*100,1)),round(prop.table(table(staffdata$StandardHours))\*100,1),Sep ="%"),

+ main = "StandardHours")

> pie(prop.table(table(staffdata$EmployeeCount)),

+ labels=paste(names(round(prop.table(table(staffdata$EmployeeCount))\*100,1)),round(prop.table(table(staffdata$EmployeeCount))\*100,1),Sep ="%"),

+ main = "EmployeeCount")



Over18 variable contain only one value "Y", Standard hours variable contain only one value 80 and employee count contain count 1 only. All these variables are non-zero variance variable. So we drop those from our model.

Let us check the correlation and scatter plots.

> cortable<-cor(staffdata[Numreicv])

> cortable



|  |
| --- |
| > library(corrplot)  > corrplot(cor(staffdata[Numreicv]), order = "hclust")  > pairs(~Age+DailyRate+DistanceFromHome+HourlyRate+NumCompaniesWorked+MonthlyIncome  +MonthlyRate+PercentSalaryHike+TotalWorkingYears+TrainingTimesLastYear+YearsAtCompany  +YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManager,  data=staffdata[Numreicv],main="Scatterplot Matrix") |
|  |
| |  | | --- | |  | |

From the scatter plot , correlation plot and correlation table, we can see totalworkingyears is highly correlate to age and monthlyincome.In reality it is true.yearswithcurrmanager is highly correlated with yearatcompany and yearatcurrentrole.So there is some multicollinearity present in our dataset.

Let us create some dummy variable for categorical variables. Some are already hot coded.

> staffdatafinal<-subset(staffdata,select=-c(EmployeeCount,EmployeeNumber,Over18,StandardHours))

> View(staffdatafinal)

> staffdatafinal <- dummy.data.frame(staffdatafinal, names = c("BusinessTravel","Department","EducationField","Gender","JobRole","OverTime","MaritalStatus"))

> View(staffdatafinal)

> staffdatafinal$Attrition<-ifelse(staffdatafinal$Attrition=="Yes",1,0)

Let us split dataset into train and test samples. And check the attrition %age in each dataset.

> require(caTools)

Loading required package: caTools

> set.seed(151)

> sample1=sample.split(staffdatafinal$Attrition,SplitRatio = 0.7)

> traindata=subset(staffdatafinal,sample1==TRUE)

> testdata=subset(staffdatafinal,sample1==FALSE)

>

> prop.table(table(staffdatafinal$Attrition))

0 1

0.8387755 0.1612245

> prop.table(table(traindata$Attrition))

0 1

0.8386783 0.1613217

> prop.table(table(testdata$Attrition))

0 1

0.8390023 0.1609977

From Above we can see the event/non-event proportion is maintained in test and train datasets.so our splitting is perfect.

Let us built Neuralnet,randomforest model on Full data. Without scaling any variable.Space in colnames create problem in formula building.Rename the variable after removing the space in variable names.

> colnam <- names(staffdatafinal)

> modelformula <- as.formula(paste("Attrition ~", paste(colnam[!colnam %in% "Attrition"], collapse = " + ")))

> modelformula

> nnetmodel<-neuralnet(modelformula,data = traindata,hidden = 8,threshold = 0.01)

> nnetmodelpredict<-compute(nnetmodel, subset(testdata,select=-c(Attrition)) )

> attrclass<-ifelse(nnetmodelpredict$net.result>0.18,1,0)

> table(attrclass)

attrclass

0 1

681 201

> confusionMatrix(attrclass,testdata$Attrition,positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 603 78

1 137 64

Accuracy : 0.7562358

95% CI : (0.7265055, 0.7842432)

No Information Rate : 0.8390023

P-Value [Acc > NIR] : 1

Kappa : 0.2273938

Mcnemar's Test P-Value : 0.00007635329

Sensitivity : 0.45070423

Specificity : 0.81486486

Pos Pred Value : 0.31840796

Neg Pred Value : 0.88546256

Prevalence : 0.16099773

Detection Rate : 0.07256236

Detection Prevalence : 0.22789116

Balanced Accuracy : 0.63278455

'Positive' Class : 1

> library(randomForest)

> modelformula <- as.formula(paste("as.factor(Attrition) ~", paste(colnam[!colnam %in% "Attrition"], collapse = " + ")))

> modelformula

> rfmodel<-randomForest(modelformula,data=traindata,ntree=100,importance=TRUE)

> print(rfmodel)

Call:

randomForest(formula = modelformula, data = traindata, ntree = 100, importance = TRUE)

Type of random forest: classification

Number of trees: 100

No. of variables tried at each split: 7

OOB estimate of error rate: 4.18%

Confusion matrix:

0 1 class.error

0 1717 9 0.005214368482

1 77 255 0.231927710843

> varImpPlot(rfmodel)

> rfpredict<-predict(rfmodel,subset(testdata,select=-c(Attrition)))

> confusionMatrix(rfpredict,testdata$Attrition,positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 738 30

1 2 112

Accuracy : 0.9637188

95% CI : (0.9491657, 0.9750538)

No Information Rate : 0.8390023

P-Value [Acc > NIR] : < 0.00000000000000022204

Kappa : 0.8540762

Mcnemar's Test P-Value : 0.000001815281

Sensitivity : 0.7887324

Specificity : 0.9972973

Pos Pred Value : 0.9824561

Neg Pred Value : 0.9609375

Prevalence : 0.1609977

Detection Rate : 0.1269841

Detection Prevalence : 0.1292517

Balanced Accuracy : 0.8930148

'Positive' Class : 1

Nuralnet model prediction accuracy is 75% where as randomforest accuracy is 96%.In randomforest model Out of Bag error is 4.18%.This is not the optimal model. These are base models.

Let us preapare a ensamle model by voting process.

> rfpredict1<-data.frame(rfpredict)

> rfpredict1<-cbind(rfpredict1,attrclass)

> rfpredict1$ensamlresult=ifelse(rfpredict1$rfpredict=='0' ,ifelse(rfpredict1$attrclass=='0',"0","1"),"1")

> confusionMatrix(rfpredict1$ensamlresult,testdata$Attrition,positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 601 20

1 139 122

Accuracy : 0.8197279

95% CI : (0.7927371, 0.8445624)

No Information Rate : 0.8390023

P-Value [Acc > NIR] : 0.9438717

Kappa : 0.5015036

Mcnemar's Test P-Value : < 0.00000000000000022

Sensitivity : 0.8591549

Specificity : 0.8121622

Pos Pred Value : 0.4674330

Neg Pred Value : 0.9677939

Prevalence : 0.1609977

Detection Rate : 0.1383220

Detection Prevalence : 0.2959184

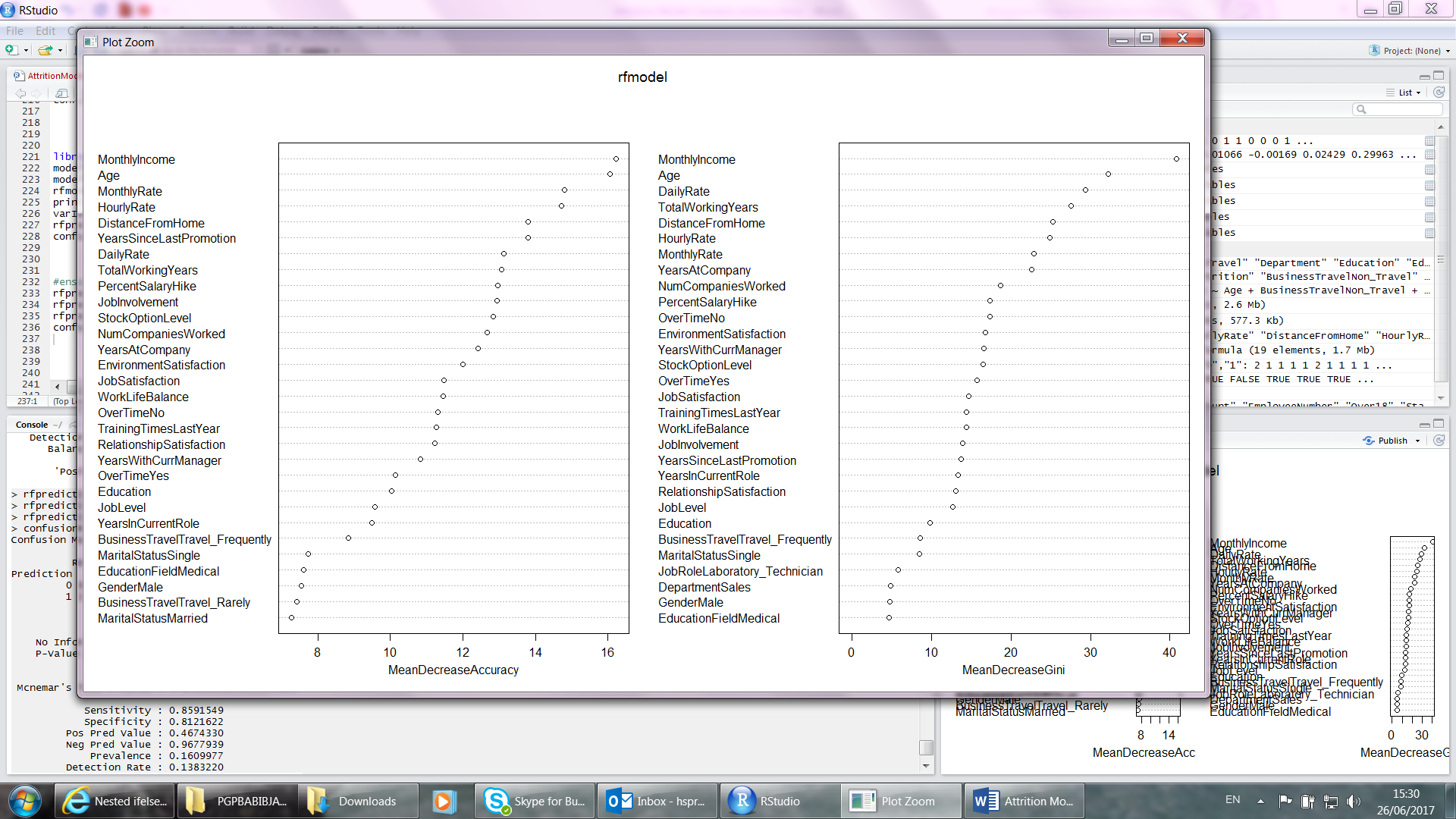
Balanced Accuracy : 0.8356585

'Positive' Class : 1

For model evaluation try Rankorder, ROC,AUC.

Further Model Improvement.

Scale DailyRate, MonthlyIncome, MonthlyRate variables and built Nuralnet model. Don't scale any rating variable. Try with different number hiddle layer and different threshold and find the accuracy model. In randomforest model try with different no of trees ntree option, try option.Find best model among these models.Below is random forest model variable importance plot.Find few top variables ,remove multicollinearity and try the NN and RF models.



Nerualnet model performs verywell in numeric narrow band large dataset. Random forest is ensemble model perform well in low sample dataset.