**DSC 520 Final Project step 1 to 4**

Name:     Soukhna Wade

 Date: 02/28/2020

 Title:

**Final Project Step 1: Getting Started**

**Section 1 – Week 9 – Getting Started**

**Three different datasets :**

1. **Dataset:  IBM HR Analytics Employee Attrition & Performance**

**Introduction:**

This is a fictional data set created by IBM data scientists to uncover the factors that lead to employee attrition. It

explores essential questions: distance from home to the job(using job role and attrition) and comparing the monthly average by education and attrition

**Research questions:**

* What is the different level of higher education in the company?
* How many employees working for the company?
* What is the higher salary?
* How many are satisfying with the salary?
* What is the higher performance rating?
* What is the relationship between employees?
* What is the company's future?
* Is the company profitable?

**Approach:**

* I will import or get the data then visualizing the bivariate relationship, and I will dig to understand the variables. Then I will calculate their correlation and covariance by performing the analysis. Also, I will perform calculations and create a histogram. I will create graphs by using a scatter plot.

**Original source (link):**

<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

* **Source data:**
  + The original purpose of the data is IBM, and this is a fiction dataset created by IBM. The dataset is to predict the attrition of your valuable employees. Pavansubhash collected it three ago and updated the last seven months. It is a CVS file with 1470 observations and 35 variables.

**Identify the Packages:**

* **Required Packages:**
* the following packages: > install.packages("ggm"), >install.packages(Hmisc) & >install.packages(polycor)>install.packages("rcorr") and load rcorr: > library(rcorr)

packages such as: >library(boot); library(ggm); library(ggplot2); library(Hmisc); library(polycor)

* **Plots and tables**:

The scatter plot knows as a scatter graph, scatter chart, scatter gram, or scatter diagram by using the ggplot2 library. And Create histograms.

* **Questions for future steps:**

the dataset by performing statistical modeling and regression analysis to determine the analytical processes for estimating the relationships between a dependent variable.

**How your approach addresses (fully or partially) the problem:**

will organize data for application and perform statistical analysis to process a large amount of data to identify trends, figures, and other relevant information. And I will prepare to identify useful insights and derive analytical data by planning, executing and monitoring, etc.

1. **Dataset:  Africa Economic, Banking and Systemic Crisis Data**

* **Introduction:**

dataset focuses explicitly on the Banking, Debt, Financial, Inflation and Systemic Crises that occurred, from 1860 to 2014, in 13 African countries, including Algeria, Angola, Central African Republic, Ivory Coast, Egypt, Kenya, Mauritius, Morocco, Nigeria, South Africa, Tunisia, Zambia, and Zimbabwe.

**Research questions:**

* Which factors are most associated with Systemic Crises in Africa?" And, "At which the annual rate of inflation does an Inflation Crisis become a practical certainty?"

* **Source (link):**

<https://www.kaggle.com/chirin/africa-economic-banking-and-systemic-crisis-data>

dataset is a derivative of Reinhart et al.'s Global Financial Stability dataset which can be found online ([https://www.hbs.edu/behavioral-finance-and-financial-stability/data/Pages/global.aspx)](https://www.hbs.edu/behavioral-finance-and-financial-stability/data/Pages/global.aspx)

* **Source data:**

The dataset will be valuable to those who seek to understand the dynamics of financial stability within the African context. Chiri collected it, and the United States joined two years ago. It is a CVS file with 1059 observations and 14 variables.

**Identify the Packages:**

* **Required Packages Using R Programming:**

Install the following packages: > install.packages("ggm"), >install.packages(Hmisc) & >install.packages(polycor)>install.packages("rcorr") and load rcorr: > library(rcorr)

load packages such as: >library(boot); library(ggm); library(ggplot2); library(Hmisc); library(polycor)

* **Plots and tables**:

the scatter plot knows as a scatter graph, scatter chart, scatter gram, or scatter diagram by using the ggplot2 library. And I will create histograms.

* **Questions for future steps:**

the dataset by performing statistical modeling and regression analysis to determine the analytical processes for estimating the relationships between a dependent variable.

1. **Dataset: Heart Disease UCI**

* **Introduction:**

the database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.

* **Research questions:**

is to experiment with the Cleveland database that has concentrated on merely attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

* **source (link):**

<https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv>

* **Source data:**
  + The dataset will be valuable to the "goal" field refers to the presence of heart disease in the patient. It is integer-valued from 0 (no presence) to 4.

It was collected By Ronit Feldman (Data Scientist) in Israel BI Developer at Vision.BI Tel Aviv-Yafo, Israel Joined two years ago, last seen a month ago (<https://www.linkedin.com/in/ronit-feldman-a05198118/>). It is a CVS file with 303 observations and 14 variables.

**Identify the Packages:**

* **Required Packages:**

Install the following packages: > install.packages("ggm"), >install.packages(Hmisc) & >install.packages(polycor)>install.packages("rcorr") and load rcorr: > library(rcorr)

load packages such as: >library(boot); library(ggm); library(ggplot2); library(Hmisc); library(polycor) and so on.

**Plots and tables**:

Create the scatter plot, knows as a scatter graph, scatter chart, scatter gram, or scatter diagram by using the ggplot2 library. Also, I will create histograms.

* **Questions for future steps:**

Exploring the dataset by performing statistical modeling and regression analysis to determine the analytical processes for estimating the relationships between a dependent variable.

**Project Step 2: Cleaning Your Data and Exploratory Data Analysis**



**Final Project Step 3**

Soukhna Wade

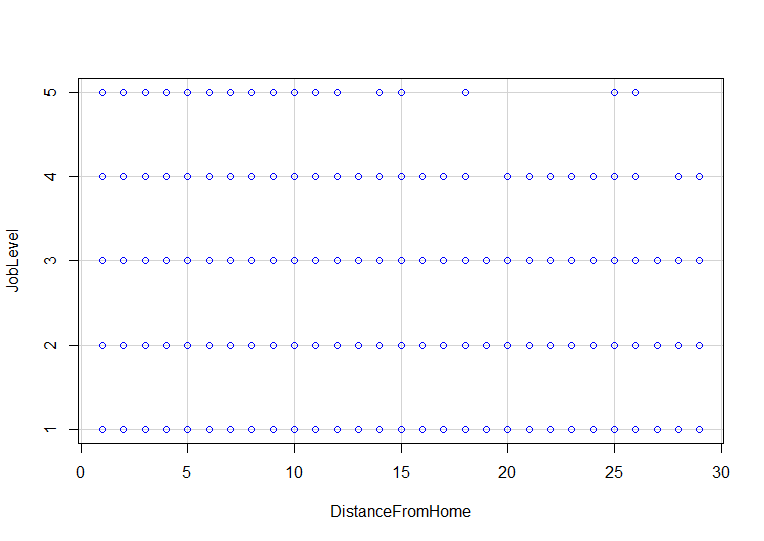
2/28/2020

*Final Project Step 3: Starting Your Writeups*

library(readxl) data <- read.csv(“~/Data/R Book Examples/WA\_Fn-UseC\_-HR-Employee-Attrition.csv”)

View(data) #shows a spreadsheet-like display of the entire data frame summary(data) #provides summary statistics on the columns of the data frame

library(dplyr) glimpse(data) # Explore the data

scatterplot(JobLevel~DistanceFromHome, regLine=FALSE, smooth=FALSE, boxplots=FALSE, data=data)

**SECTION 3**

**INTRODUCTION:**

**This is a fictional data set created by IBM data scientists to uncover the factors that lead to employee attrition. It explores essential questions: distance from home to the job(using job role and attrition) and comparing the monthly average by education and attrition.**

**Question a: Discuss how you plan to uncover new information in the data that is not self-evident.**

**I would use quantitative analysis (QA), which is a technique that seeks to understand behavior by using mathematical and statistical modeling, measurement, and research. I will try to represent a given reality in terms of a numerical value. Therefore, quantitative analysis and a combination of textual will help to derive insights in a holistic manner.**

**Question b: What are different ways you could look at this data to answer the questions you want to answer?**

**I am going to create model and Use R programming to calculate the covariance. Construct a model to check if the current employer likes the job with different combinations of data.**

**Question c: Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.**

**Evidently, I will like to dice, create, join new useful datasets to create the model with the different combinations of the data.**

**Question d: How could you summarize your data to answer key questions?**

**I would totally understand the data. We will construct a model to evaluate the jobs of a person with Income, Job satisfaction, Number of Years, Department etc.**

**I could summarize my data to answer key questions.**

**Question e: What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).**

**I will use the correlation plot, density plot, and scatter plot-possible use ofa dendrogram (or tree diagram) which, is a network structure, and tables to represent results from data analysis.**

**Question f: What do you not know how to do right now that you need to learn to answer your questions?**

**The mean ideas that we have for exploring the data involve an analysis of the overview and tagline fields in the database. To do so, we would require an understanding and relevant knowledge of Data Analysis using R.**

**What I know now that employee salary is not equally shared. I hope that can be solved in the future,**

**Question g: Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.**

**Obviously, we are planning to include many Machine Learning Classification Analysis models in the R for future analysis and prediction**

**I am planning to incorporate machine learning because I want to use regression and modeling.**

**END OF SECTION 3**

**Final Project Step 4**

Soukhna Wade

2/28/2020

SECTION 4

*Forum: 12.1 Discussion: Final Project Step 4: Final Project Submission*

*Final Project Step 4: Final Project Submission*

**LOAD THE LIBRARY needed for calculation, analysis and plotting the data sets are listed below:**

library(boot) library(ggm) library(ggplot2) library(Hmisc) library(polycor) library(rcorr) library(readxl) library(tidyverse) library(tidyr) library(corrplot) library(leaflet) library(lubridate)

library(data.table) library(dplyr) library(VIM) library(DT) library(gridExtra) library(caret) library(Metrics) library(randomForest) library(pROC) library(e1071) library(dtree) library(corrplot) library(DMwR) library(Rcmdr)

*Functions to examine data set are:*

dim(data) #shows the dimensions of the data frame by row and column str(data) # shows the structure of the data frame #summary(data) # provides summary statistics on the columns of the data frame #colnames(data) # shows the name of each column in the data frame #head(data) # shows the first 6 rows of the data frame ## look at the first several rows of the data #tail(data) # shows the last 6 rows of the data frame View(data) # shows a spreadsheet-like display of the entire data frame #rownames(data) #colnames(data) glimpse(data) # Explore the data

attach(data) #attach the data frame to the environment

**To Perform Exploratory Data Analysis(EDA)**

**Import dataset: CSV (WA\_Fn-UseC\_-HR-Employee-Attrition) form The Original source (link):** [**https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset**](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)

library(readxl) data <- read.csv(“~/Data/R Book Examples/WA\_Fn-UseC\_-HR-Employee-Attrition.csv”)

#View(data) # shows a spreadsheet-like display of the entire data frame

colnames(data) # shows the name of each column in the data frame nrow(data) # Shows the columns(variables) number result: [1] 35 variables  
ncol(data) # shows the rows (observations) number result: [1] 1470 Observations dim(data) # shows the dimensions of the data frame by row and column

str(data) # shows the structure of the data frame #rownames(data) #colnames(data) #head(data) # shows the first 6 rows of the data frame ## look at the first several rows of the data #tail(data) # shows the last 6 rows of the data frame summary(data) # provides summary statistics on the columns of the data frame

library(dplyr) glimpse(data) # Explore the data

**To find the missing values(NA)**

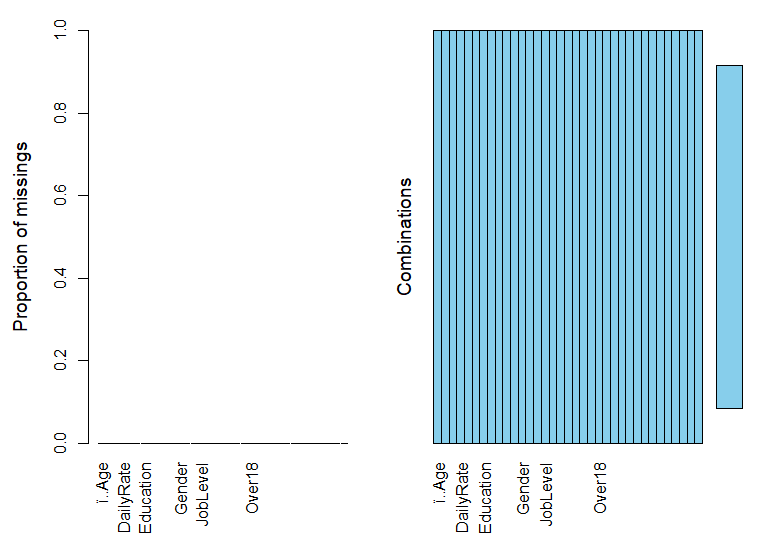
missingValues <- sum(is.na(data)) # to directly find out it there are any missing values in out dataframe / result is [1] 0, I found that my data is cleaned missingValues

apply(is.na(data), 2, sum) # There is no missing value

cat(“DataSet has”,dim(data)[1], " Rows and “, dim(data)[2],” Columns" ) # DataSet has 1470 Rows and 35 Columns ### To check for duplicated Record

sum (is.na(duplicated(data))) # result is [1] 0

VIM::aggr(data)



**To Dropping columns irrelevant columns (step need in EDA to drop unused column)**

#subset( data, select = -c(BusinessTravel,Department,EducationField, Gender, JobRole, MaritalStatus, Over18))

**To Determine correlation**

**data.cor = cor(dataï..Age, method = c(“pearson”, “kendall”, “spearman”)) # result is [1] 0.4978546**

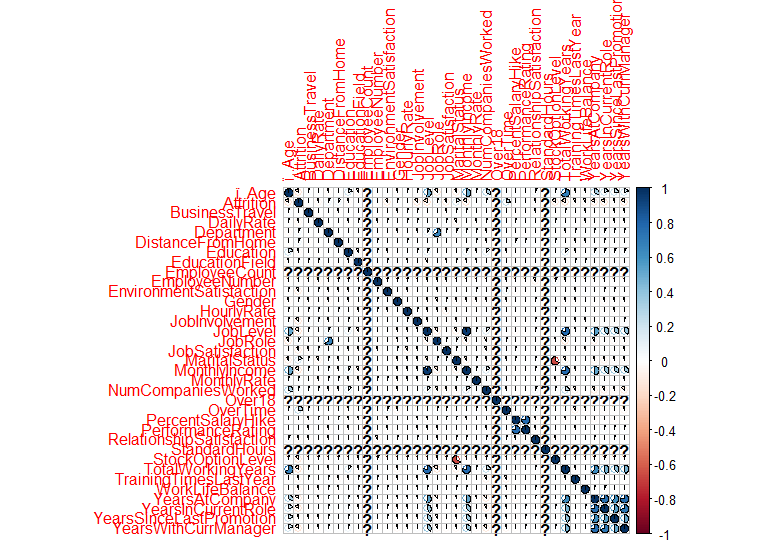
**data.cor = cor(dataï..Age, method = c(“pearson”)) # result is [1] 0.4978546**

**data.cor = cor(dataï..Age, use = “everything”, method = c(“pearson”)) # result is [1] 0.4978546**

data.cor = cor(dataï..Age, use = “complete.obs”) # result is [1] 0.4978546 data.cor = cor(dataï..Age, use = “complete.obs”, method = c(“pearson”, “kendall”, “spearman”)) # [1] 0.4978546 dataï..Age cor(data[,c(“ï..Age”,“MonthlyIncome”)], use=“complete”)

**To find the Correlation Matrix of Data**

library(corrplot) corrplot(cor(sapply(data,as.integer)),method = “pie”)



#Some of the variables are highly correlated, such as: JobLevel and MonthlyIncome / Education and YearsEducation. These variables cause a multicollinearity problem in our dataset. Therefore, we should remove one of them for any group. then we try again our data set with new attributes using Random Forest

**To calculate the covariance of the two variables monthly income and age**

out <- cov(dataï..Age, use = “complete.obs”) #result is [1] 0.4978546

**To Create Elegant Data Visualisations Using the Grammar of Graphics**

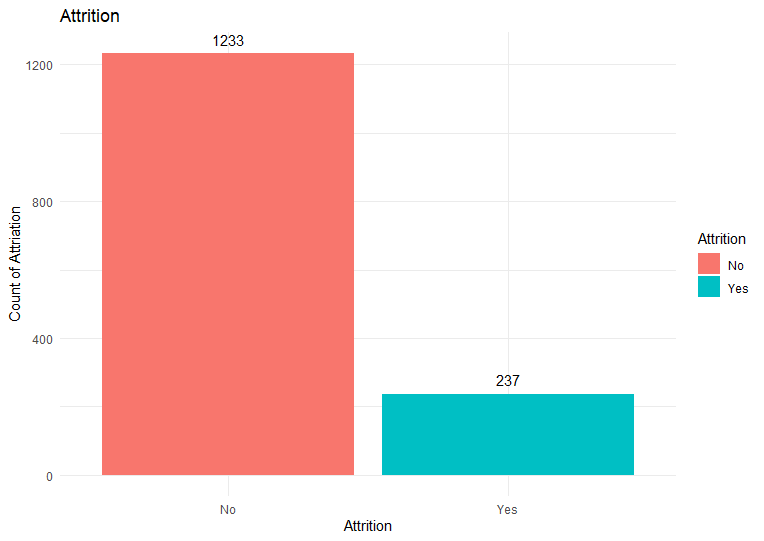
library(ggplot2) ggplot2 <- ggplot(data, aes(x=“ï..Age”, y=“JobSatisfaction”)) + geom\_point(alpha = 0.6) + stat\_smooth(method = “lm”, col = “red”, se = FALSE) ggplot2

**Some attributes are categorical, but in the dataset are integer. We must change them to categorical. Also, we do not need any dummy variable creation, where some machine learning algorithms like RF, Boost, etc. can use categorical variables. For other algorithms like NN, we must change categorical variables more than two-level to dummy variable with two-level (Binary) can be adjusted to number very easy.**

dataPerformanceRating) dataRelationshipSatisfaction) dataStockOptionLevel) dataWorkLifeBalance) dataEducation) dataEnvironmentSatisfaction) dataJobInvolvement) dataJobLevel) dataJobSatisfaction)

**Visualization of Attrition**

data %>% group\_by(Attrition) %>% tally() %>% ggplot(aes(x = Attrition, y = n,fill=Attrition)) + geom\_bar(stat = “identity”) + theme\_minimal()+ labs(x=“Attrition”, y=“Count of Attriation”)+ ggtitle(“Attrition”)+ geom\_text(aes(label = n), vjust = -0.7, position = position\_dodge(0.11))



#The above graph shows 237/1470=0.16 % of the dataset label shows the “Yes” in Attrition. This is a problem that should be handled during the process because an unbalanced dataset will bias the prediction model towards the more common class ( ‘NO’). There are different approaches for dealing with biased data in machine learning. It is ideal to use more data( but unfortunately, here is not possible). Resampling, changing the machine performance metric, using various algorithms, etc.

library(ggplot2) ggplot2 <- ggplot(data=data, aes(x=data$Age, y=null)) + geom\_histogram(breaks=seq(10, 40, by=2), col=“red”, aes(fill=..count..))+ labs(x=“Age”, y=“Count”)+ scale\_fill\_gradient(“Count”, low=“green”, high=“red”) ggplot2

**Building the model**

**Using Raw data by RF: at the first Stage to use RF for getting some information about the prediction split Data to Train and Test**

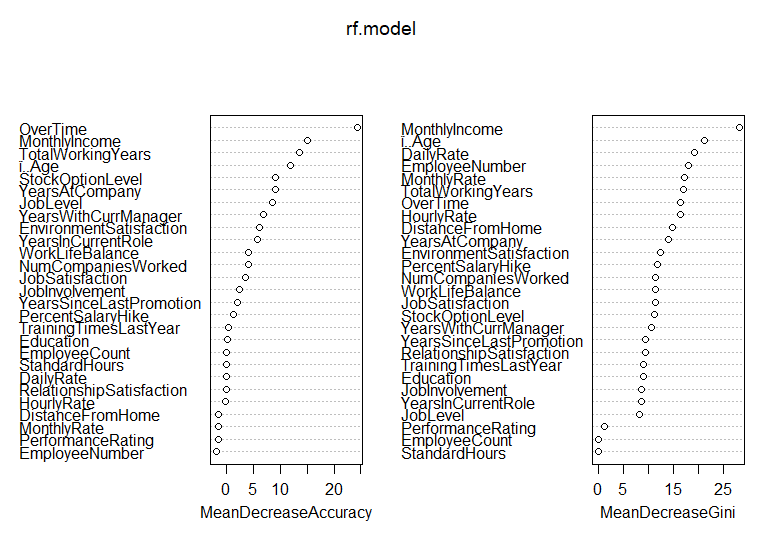
#install.packages(‘pROC’) #install.packages(‘e1071’) #install.packages(“varImpPlot”) library(randomForest) library(pROC) library(e1071) library(varImpPlot)

**To split Data to Train and Test**

rfData <- dataset.seed(123) indexes = sample(1:nrow(rfData), size=0.8\*nrow(rfData)) RFtrain.Data <- rfData[indexes,] RFtest.Data <- rfData[-indexes,]

**To build model**

rf.model <- randomForest(Attrition~.,RFtrain.Data, importance=TRUE,ntree=600) varImpPlot(rf.model)



**To build model**

#rfData <- data #set.seed(123) #indexes = sample(1:nrow(rfData), size=0.8\*nrow(rfData)) #RFRaw.train.data <- rfData[indexes,] #RFRaw.test.data <- rfData[-indexes,] #Raw.rf.model <- randomForest(Attrition~.,RFRaw.train.Data, importance=TRUE,ntree=1000)

#Building the model

rfData <- myData set.seed(123) indexes = sample(1:nrow(rfData), size=0.8\*nrow(rfData)) RFRaw.train.Data <- rfData[indexes,] RFRaw.test.Data <- rfData[-indexes,]

Raw.rf.model <- randomForest(Attrition~.,RFRaw.train.Data, importance=TRUE,ntree=800)

varImpPlot(Raw.rf.model)

Raw.rf.prd <- predict(Raw.rf.model, newdata = RFRaw.test.Data) confusionMatrix(RFRaw.test.Data$Attrition, Raw.rf.prd)

##Confusion Matrix and Statistics ## ##Reference ##Prediction No Yes ##No 244 2 ##Yes 40 8 ## ##Accuracy : 0.8571  
##95% CI : (0.8118, 0.8951) ##No Information Rate : 0.966  
##P-Value [Acc > NIR] : 1  
## ##Kappa : 0.2327  
## ##Mcnemar’s Test P-Value : 0.00000001135  
##  
## Sensitivity : 0.8592  
## Specificity : 0.8000  
## Pos Pred Value : 0.9919  
## Neg Pred Value : 0.1667  
## Prevalence : 0.9660  
## Detection Rate : 0.8299  
## Detection Prevalence : 0.8367  
## Balanced Accuracy : 0.8296  
##  
## ‘Positive’ Class : No  
##

**1. Overall, write a coherent narrative that tells a story with the data as you complete this section.**

**This project is about to build a learning model of a fictional data set created by IBM data scientists to analyze the workforce. The main issue is to deals with the employee at the organization. This project is to uncover the factors that lead to employee attrition and to explore the essential questions such as: showing a breakdown of distance from home by job role and attrition comparing average monthly income by education and attrition. It is a fictional data set created by IBM data scientists. The dataset is supplied by Kaggle and contains HR analytics data of employees that stay and leave. The types of data include metrics such as education level, job satisfaction, and commute distance.**

**2. Summarize the problem statement you addressed.**

**The data mainly deals with employee details. All the data is very clean. Added scatter plots, bar charts for the employee details. Box plots for the bivariate analysis like the number of experiences for the number of projects etc. The final goal of the project is to find if an employee resigns next.**

**3. Summarize how you addressed this problem statement (the data used and the methodology employed).**

**There are many interesting insights. Mainly I have analyzed it in Univariate and Bivariate analysis. So, I found the details of why if an employee is going to leave or not. Many factors could affect an employee to leave a company, like the number of years they worked in the company if they promoted or not, etc.**

**4. Summarize the interesting insights that your analysis provided.**

**The analysis can be used by hour of the company to have a clear idea of how employee behavior is impacting.**

**5. Summarize the implications to the consumer (target audience) of your analysis.**

**I can draw more analysis that can affect employee behavior. Implement the Machine Learning Techniques to find which employee will leave next.**

**6. Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.**

**Employee Number can be accepted as an indicator for the time of joining the company, which can be used for new feature generation. Still, we do not have any metadata about it; then, we will remove it.**

**7. In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.**

END OF SECTION 4

