USE CASE STUDY REPORT

Group No.: Group 21

Student Names: Abhishikth Devabhaktuni and Kirtikumar Waykos

I. Background and Introduction

" Managers tend to blame their turnover problems on everything under the sun, while ignoring the crux of the matter: people don't leave jobs; they leave managers." by Travis BradBerry.

In the current business world, employee attrition is one of the major problems faced by small and large companies. Employee attrition refers to the loss of employees through a natural process, such as retirement, resignation, elimination of a position, personal health, or other similar reasons. With attrition, the employer usually does not fill the position left vacant. A large employee attrition rate results in reduction in size or strength of the workflow. The remaining job duties can also increase the workload for remaining employees.

Using the information in the database can we predict the likelihood of the employee leaving the organization because of attrition? Can we also show which factors are the major causes of employee attrition in the given data?

Employee attrition is an unavoidable problem for organizations. However, it can be addressed at the right time to avoid sudden loss of employees. We start by analyzing the data and finding out the main causes of employee attrition. Then we will compute a Classification model to predict the employees who are most likely to leave. This information can be used to engage them with different strategies and retain them for a longer time.

II. Data Exploration and Visualization

Dataset Structure: 1470 observations (rows), 35 features (variables)

Missing Data: There is no missing data in our dataset.

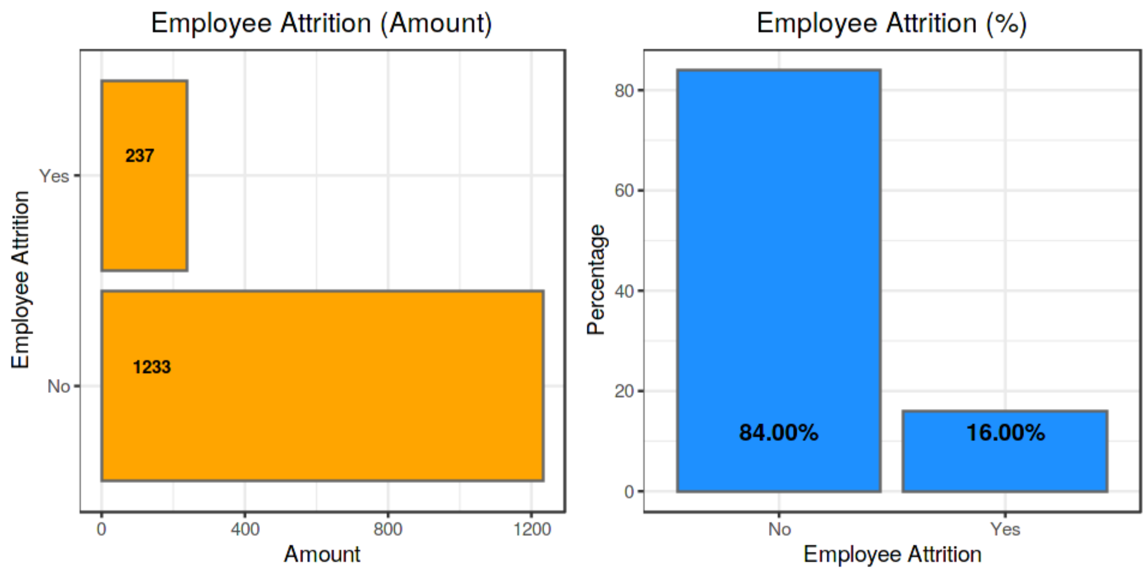
Data Type: We only have two datatypes in this dataset: factors and integers

Label: Attrition is the label in our dataset, and we would like to find out why employees are leaving the organization.

Imbalanced dataset: 1237 (84% of cases) employees did not leave the organization while 237 (16% of cases) did leave the organization making our dataset to be considered **imbalanced** since more people stay in the organization than they leave.

Distribution of our Labels:

This is an important aspect that will be further discussed in this document and that is dealing with **imbalanced dataset. 84%** of employees did not quit the organization while **16%** did leave the organization. Knowing that we are dealing with an imbalanced dataset will help us determine what will be the best approach to implement our predictive model.

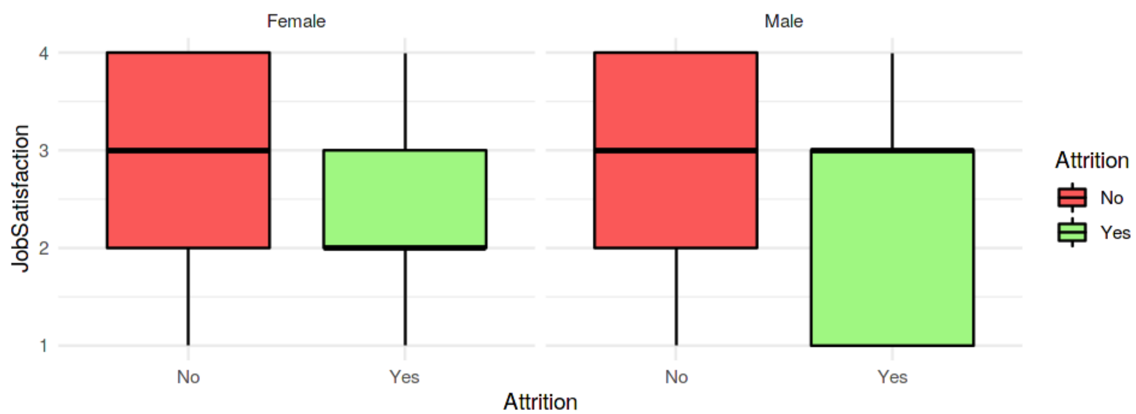


Gender Analysis:

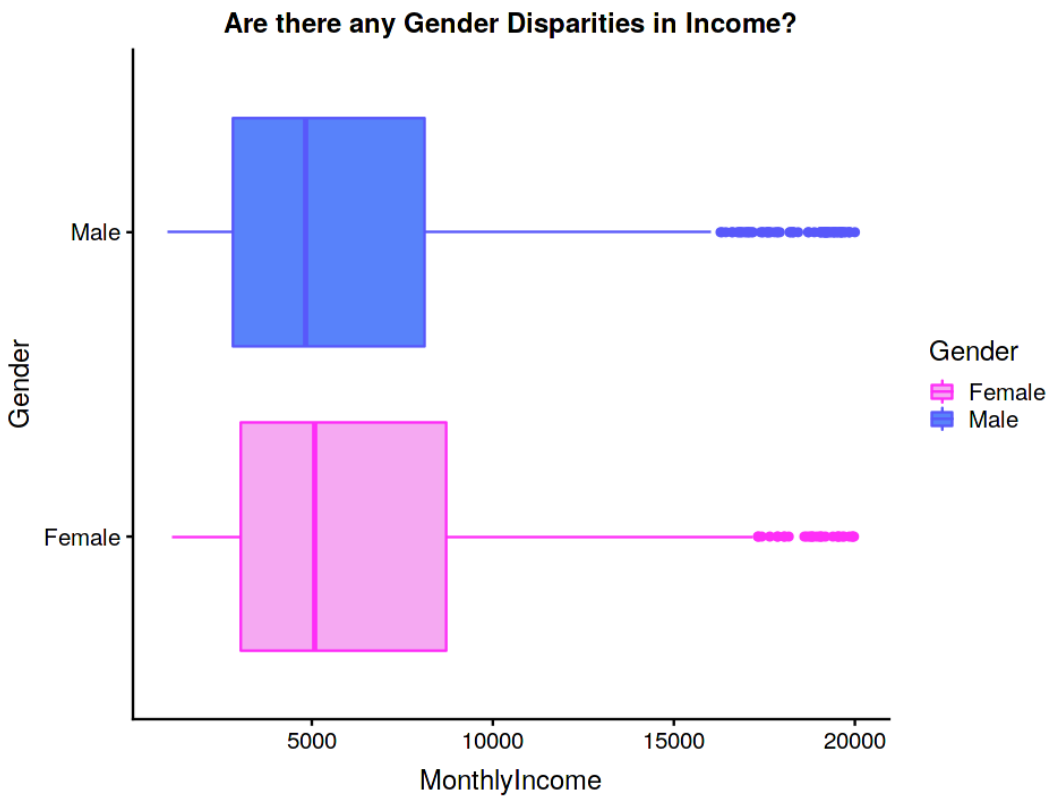
In this section, we will try to see if there are any discrepancies between male and females in the organization. Also, we will look at other basic information such as the age, level of job satisfaction and average salary by gender.

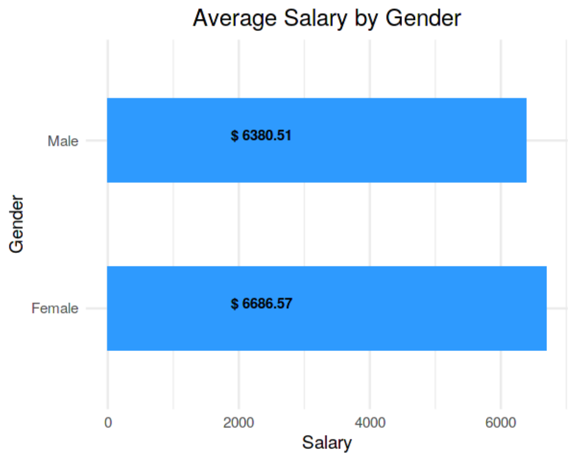
- **Age by Gender:** The average age of females is 37.33 and for males is 36.65 and both distributions are **similar**.
- **Job Satisfaction by Gender:** For individuals who didn't leave the organization, job satisfaction levels are practically the same. However, for people who **left the organization,** females had a lower satisfaction level as opposed to males.
- **Salaries:** The average salaries for both genders are practically the same with **males** having an average of 6380.51 and **females** 6686.57
- **Departments:** There are a higher number of males in the three departments however, females are more predominant in the **Research and Development** department.

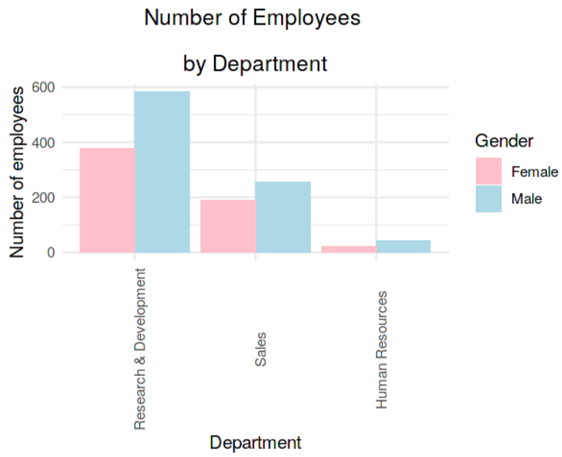

Age Distribution
Female
Male
Mean = 37.33
Years Old
Mean = 36.65
Years Old
density
Age
Mean = 36.92 Years Old
Overall Age

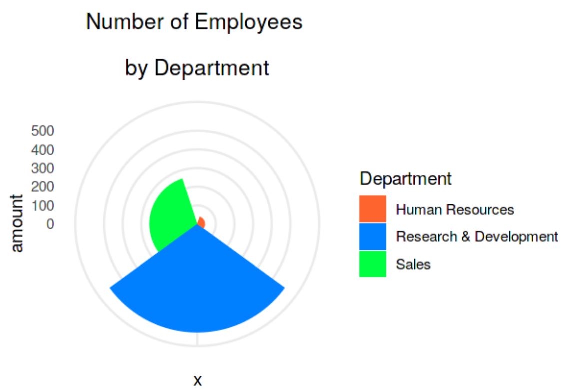

Female
Male
JobSatisfaction
Attrition
No
Yes


density
JobSatisfaction

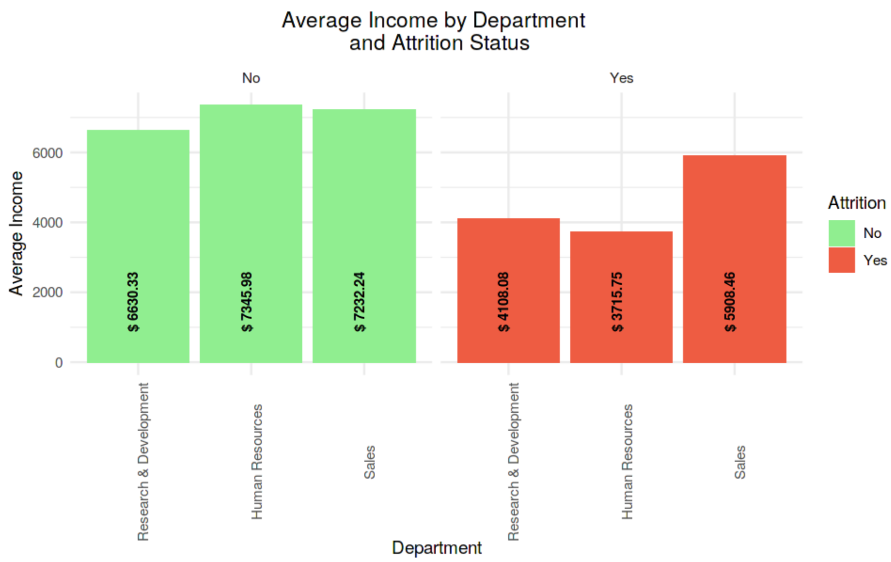

Are there any Gender Disparities in Income?
Male
Female
Gender
5000
10000
15000
20000
MonthlyIncome
Gender
Female
Male







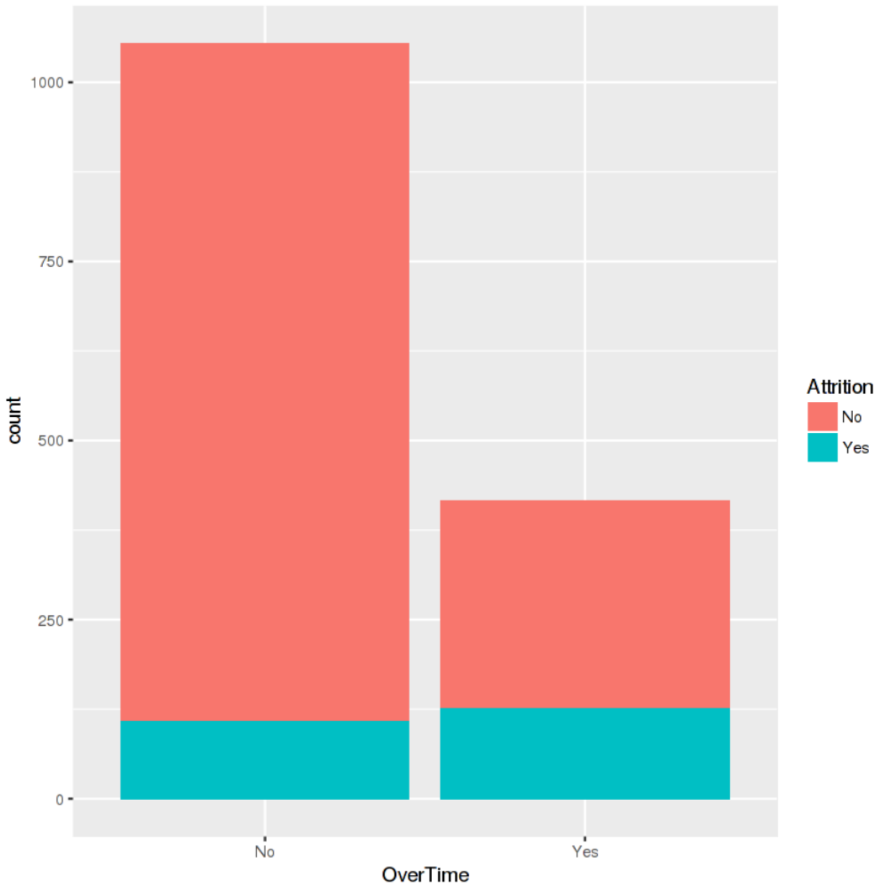
Impact of Income over Attrition:



We can see huge differences in each department by attrition status.

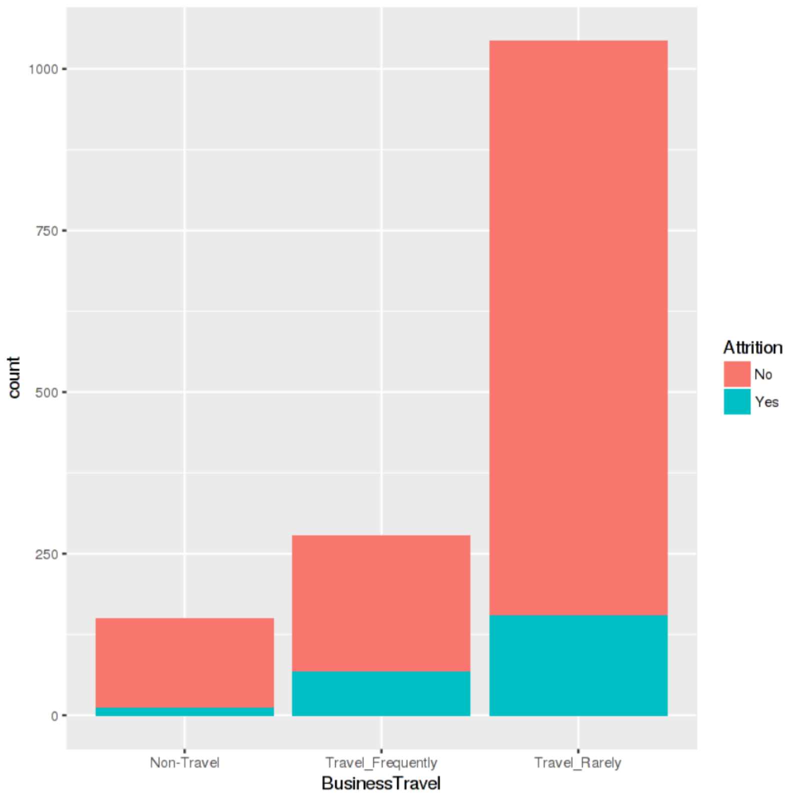
Attrition by Overtime:

Let us look at how Attrition is affected by Overtime.



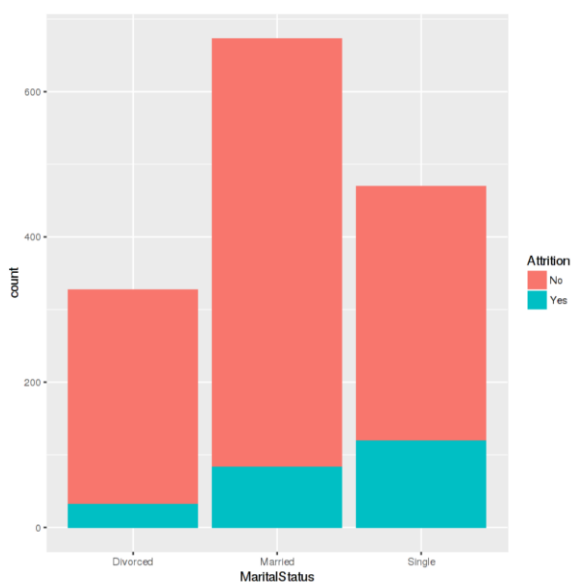
It is readily apparent that the attrition rate is higher for those working overtime.

Attrition and Business Travel:



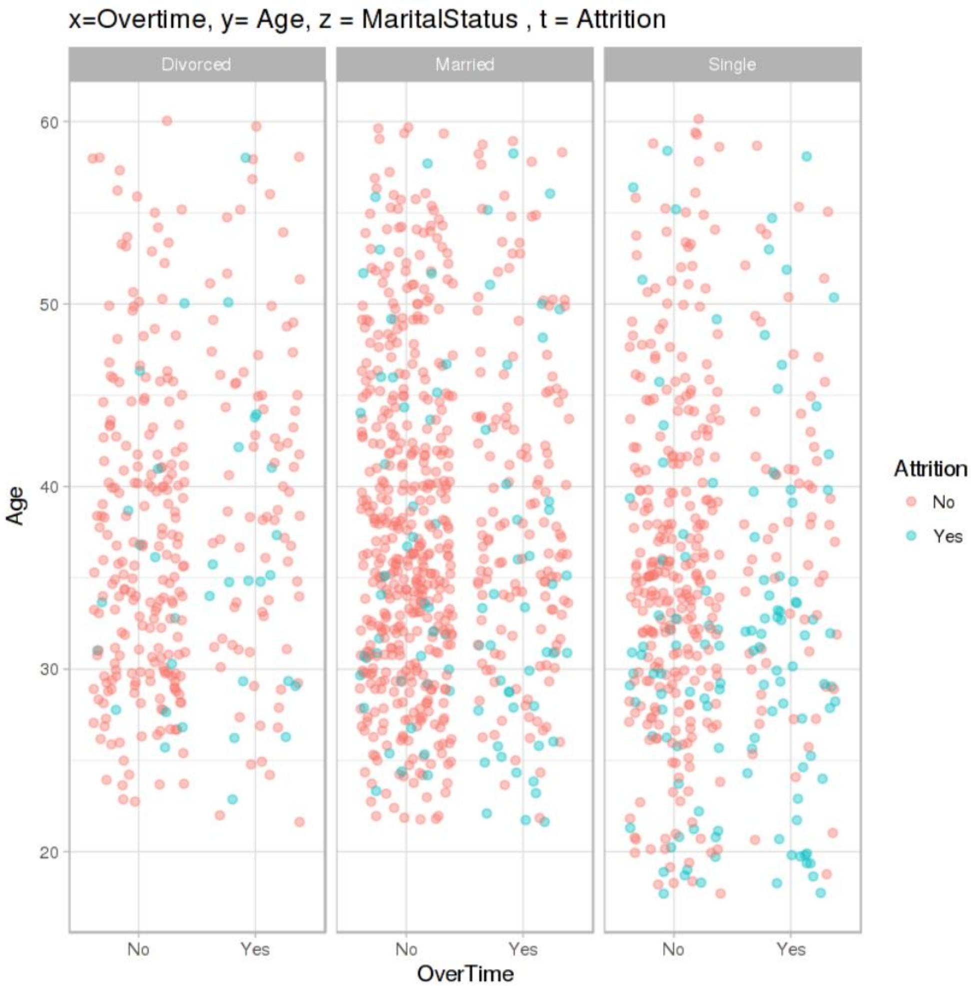
The employees who are required to travel more have a higher rate of attrition compared to those who travel less.

Attrition over Marital status:



Single people are more likely to leave their present company compared to Married and Divorced people.

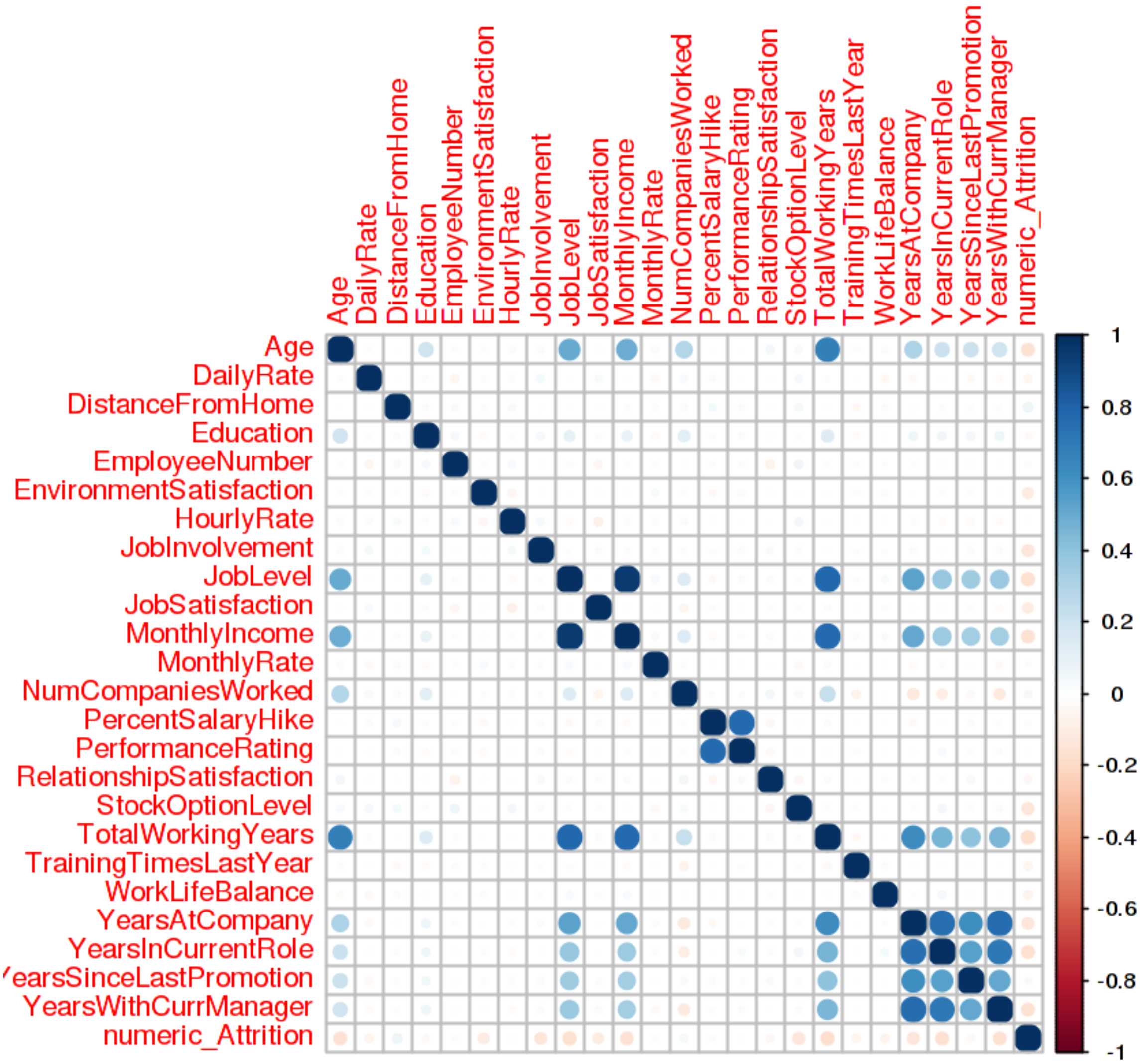
Attrition over Age, Overtime and Marital status:



This graph shows a clear cumulation when Age<35, Single and works Overtime. If these factors can be avoided there will be less attrition.

III. Data Preparation and Preprocessing

Correlation plot:



Our correlation plot shows a clear inverse correlation between our target variable and many attributes like Age, Environment Satisfaction, Job involvement, Job level, Job Satisfaction, e.t.c.

This satisfies our assumptions as those who have a higher salary, higher job level, more involvement and more experience are less likely to leave their job.

We also observe a few columns with redundant data. The column Over18 has the value of "Yes" for all observations. The column StandardHours has the value of 80 for all observations. EmployeeCount is always 1. These columns do not affect the model in anyways. Thus, we drop these columns.

We normalize all the numerical columns to get a more accurate model.

In our dataset, we have columns that are categorical with multiple levels. These columns are preprocessed to allow for implementation of machine learning techniques.

For columns with 2 factors, we turn them into 1 and 0.
For columns with 3 or more factors, dummy variables are created and the original columns are dropped.

IV. Data Mining Techniques and Implementation

As our problem is categorical in nature, we start off with Logistic Regression. Since are data is linear and there is good linear correlation between our variables this model should perform well and serve as a good baseline model.

Logistic regression gives us a good result on our test set with a accuracy of 89.8%.

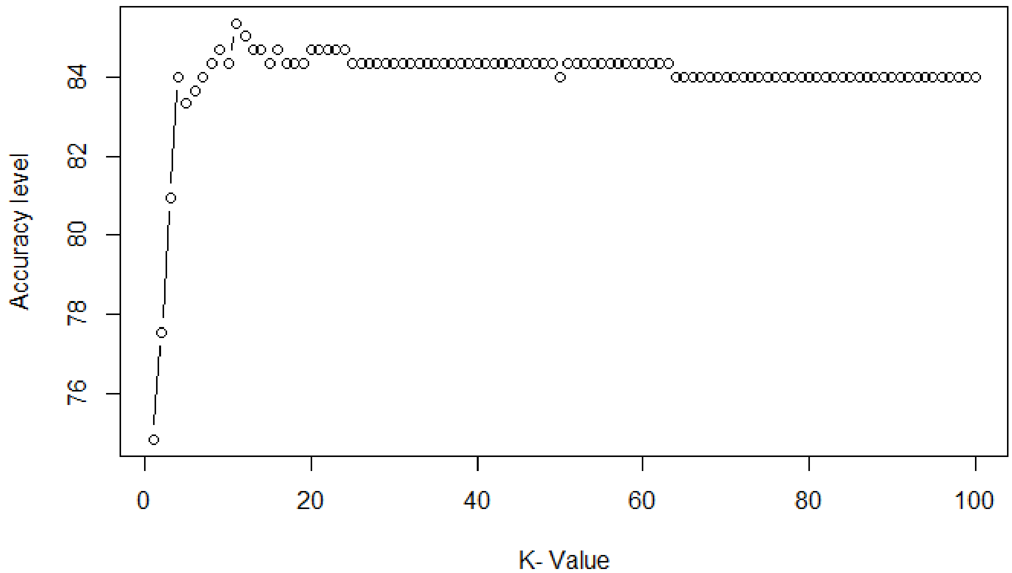
```
          Reference
Prediction   0   1
         0 238  21
         1   9  26

               Accuracy : 0.898
                 95% CI : (0.8575, 0.9301)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 0.002934

                  Kappa : 0.5763

 Mcnemar's Test P-Value : 0.044610

            Sensitivity : 0.9636
            Specificity : 0.5532
         Pos Pred Value : 0.9189
         Neg Pred Value : 0.7429
             Prevalence : 0.8401
         Detection Rate : 0.8095
   Detection Prevalence : 0.8810
      Balanced Accuracy : 0.7584

       'Positive' Class : 0
```

The classification matrix shows good results. The accuracy for “No” is higher when compared to the accuracy for “Yes”. This is not surprising since there is imbalance in the target variable. Logistic Regression results are as expected.

We choose KNN as our second model as it should also perform well over this data. For the initial K value, we choose 38. Our input data ahs 1470 observations. 38 is the square root of this value. This is a good place to start.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  247  46
       Yes   0   1

               Accuracy : 0.8435
                 95% CI : (0.7969, 0.8831)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 0.4755

                  Kappa : 0.0352

 Mcnemar's Test P-Value : 3.247e-11

            Sensitivity : 1.00000
            Specificity : 0.02128
         Pos Pred Value : 0.84300
         Neg Pred Value : 1.00000
             Prevalence : 0.84014
         Detection Rate : 0.84014
   Detection Prevalence : 0.99660
      Balanced Accuracy : 0.51064

       'Positive' Class : No
```

We achieved a accuracy of 84.3%. This is worse than our previous model. To improve this we took a range of k values from 1 to 100 and compared the accuracies. The accuracy plot:



The plot peaks at k=11 with a accuracy of 85.374%.

V. Performance Evaluation

Our data performed better on the Logistic regression model compared to KNN. This is not surprising since the data itself is fairly linear in nature and thus performs better on such models. The KNN model isn’t too bad with an accuracy score of 85%.

VI. Discussion and Recommendation

The data is fairly clean and shows good relation to the target variable. Decision trees or boosting methods might show good performance on this data. Most linear models should show good performance on the dataset.

VII. Summary

We have predicted the attrition rate of employees with a fairly high accuracy. We have also identified a few key factors. The key factors can be used to reduce the rate of attrition. The prediction model can be used to predict the chance of an employee to leave.

Appendix: R Code for use case study

```
mydata <- read.csv("C:/Users/Abhishikth Sagar/Desktop/HR project/WA_Fn-UseC_-
HR-Employee-Attrition.csv", stringsAsFactors = TRUE)
library(dplyr)
library(ggplot2)
library(ggthemes)
colnames(mydata)[1] <- c("Age")
str(mydata)
dim(mydata)
numeric_mydata <- mydata[,c(1,4,6,7,10,11,13,14,15,17,19,20,21,24,25,26,28:35)]
numeric_Attrition = as.numeric(mydata$Attrition)- 1
numeric_mydata = cbind(numeric_mydata, numeric_Attrition)
str(numeric_mydata)
library(corrplot)
M <- cor(numeric_mydata)
corrplot(M, method="circle")
library(caTools)
library(e1071)
library(glmnet)
mydatanew = mydata[,-c(6,9,22)]
str(mydatanew)
l <- ggplot(mydata, aes(OverTime,fill = Attrition))
l <- l + geom_histogram(stat="count")
print(l)
tapply(as.numeric(mydata$Attrition) - 1 ,mydata$OverTime,mean)
### MaritalStatus vs Attiriton
l <- ggplot(mydata, aes(MaritalStatus,fill = Attrition))
l <- l + geom_histogram(stat="count")
```

```
print(l)
tapply(as.numeric(mydata$Attrition) - 1 ,mydata$MaritalStatus,mean)
l <- ggplot(mydata, aes(JobRole,fill = Attrition))
l <- l + geom_histogram(stat="count")
print(l)
tapply(as.numeric(mydata$Attrition) - 1 ,mydata$JobRole,mean)
mean(as.numeric(mydata$Attrition) - 1)
###Gender vs Attrition
l <- ggplot(mydata, aes(Gender,fill = Attrition))
l <- l + geom_histogram(stat="count")
print(l)
tapply(as.numeric(mydata$Attrition) - 1 ,mydata$Gender,mean)
l <- ggplot(mydata, aes(EducationField,fill = Attrition))
l <- l + geom_histogram(stat="count")
print(l)
tapply(as.numeric(mydata$Attrition) - 1 ,mydata$EducationField,mean)
###Department vs Attrition
l <- ggplot(mydata, aes(Department,fill = Attrition))
l <- l + geom_histogram(stat="count")
print(l)
tapply(as.numeric(mydata$Attrition) - 1 ,mydata$Department,mean)
l <- ggplot(mydata, aes(BusinessTravel,fill = Attrition))
l <- l + geom_histogram(stat="count")
print(l)
tapply(as.numeric(mydata$Attrition) - 1 ,mydata$BusinessTravel,mean)
### x=Overtime, y= Age, z = MaritalStatus , t = Attrition
ggplot(mydata, aes(OverTime, Age)) +
  facet_grid(.~MaritalStatus) +
  geom_jitter(aes(color = Attrition),alpha = 0.4) +
  ggtitle("x=Overtime, y= Age, z = MaritalStatus , t = Attrition") +
  theme_light()
split <- sample.split(mydatanew$Attrition, SplitRatio = 0.80)
train <- subset(mydatanew, split == T)
test <- subset(mydatanew, split == F)
library(caret)
model_glm <- glm(Attrition ~ ., data = train, family='binomial')
predicted_glm <- predict(model_glm, test, type='response')
predicted_glm <- ifelse(predicted_glm > 0.5,1,0)
confusionMatrix(factor(predicted_glm),factor(as.numeric(test$Attrition)-1))
train_labels = train$Attrition
test_labels = test$Attrition
knn_train = subset(train,select = -c(Attrition))
knn_test = subset(test,select = -c(Attrition))
knn_data = mydatanew
knn_data[, c("Age","DailyRate",
"Education","EmployeeNumber","EnvironmentSatisfaction","HourlyRate","JobInvolvem
```

```
ent","JobLevel","JobSatisfaction","MonthlyIncome","MonthlyRate","NumCompaniesWo
rked","PercentSalaryHike","PerformanceRating","RelationshipSatisfaction","StandardHo
urs","StockOptionLevel","TotalWorkingYears","TrainingTimesLastYear","WorkLifeBal
ance","YearsAtCompany","YearsInCurrentRole","YearsSinceLastPromotion","YearsWit
hCurrManager")] <- scale(knn_data[, c("Age","DailyRate",
"Education","EmployeeNumber","EnvironmentSatisfaction","HourlyRate","JobInvolvem
ent","JobLevel","JobSatisfaction","MonthlyIncome","MonthlyRate","NumCompaniesWo
rked","PercentSalaryHike","PerformanceRating","RelationshipSatisfaction","StandardHo
urs","StockOptionLevel","TotalWorkingYears","TrainingTimesLastYear","WorkLifeBal
ance","YearsAtCompany","YearsInCurrentRole","YearsSinceLastPromotion","YearsWit
hCurrManager")])
knn_data = subset(knn_data, select = -c(StandardHours))
knn_data$Gender = ifelse(knn_data$Gender == "Male", 1, 0)
knn_data$OverTime = ifelse(knn_data$OverTime == "Yes", 1, 0)
library(psych)
BusinessTravel <- as.data.frame(dummy.code(knn_data$BusinessTravel))
Department <- as.data.frame(dummy.code(knn_data$Department))
EducationField <- as.data.frame(dummy.code(knn_data$EducationField))
JobRole <- as.data.frame(dummy.code(knn_data$JobRole))
MaritalStatus <- as.data.frame(dummy.code(knn_data$MaritalStatus))
JobRole <- rename(JobRole, HR_JobRole = `Human Resources`)
EducationField <- rename(EducationField, HR_Edu = `Human Resources`)
Department <- rename(Department, HR_Dep = `Human Resources`)
library(dplyr)
knn_data = subset(knn_data, select = -
c(BusinessTravel,Department,EducationField,JobRole,MaritalStatus))
knn_data = cbind(knn_data, BusinessTravel, Department, EducationField, JobRole,
MaritalStatus)
knn_train <- subset(knn_data, split == T)
knn_test <- subset(knn_data, split == F)
train_labels = knn_train$Attrition
knn_train = subset(knn_train,select = -c(Attrition))
test_labels = knn_test$Attrition
knn_test = subset(knn_test,select = -c(Attrition))
```