Department of Statistics, SPPU
ST-O13 Statistical Learning and Data Mining

ETE Assignment 2022-23

Ayshik Neogi 2102|*Saurav Jadhav* 2120

23-11-2022

Introduction

A large company named ABC, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and the company needs to replace them with the talent pool available in the job market.

The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons:

1. The former employees' projects get delayed, making it difficult to meet timelines, resulting in a loss of reputation among clients and partners.
2. More often than not, the new employees have to be trained for the job and/or given time to acclimatise themselves to the company.

The purpose of the analysis is to find what factors share a strong relationship with attrition and use them to decide what changes to be made in the workplace to retain employees.

Before we dive into the analysis from the data. It is first important to think about this problem instinctively.

1. How would a firm reduce its attrition rate?
2. And, more importantly, what causes an increase in the firm's attrition rate?
3. Or, what factors make an employee stay in the company?

Heuristic based answer to the problem

From the above three questions, we will be specifically looking at the third question:

What factors make an employee stay in the company?

After looking at the question, the first thing which can immediately come to mind is the good pay scale.

1. Pay scale above the industry standards seems like a good reason for an employee to stay in the company.

As this is a business problem, increasing the pay scale of the employee for the sake of reducing the attrition rate is not an economical solution. Also, this may not be a sole reason due to which an employee might stay into the company.

2. Appreciation to the employees for their good work.
3. Promotion for the deserving candidates.
4. Friendly environment: communication apart from the professional work.
5. Non-toxic work culture. Like asking employees to work on weekends, more than 8 hrs per day, and so on may promote some sort of toxicness in the work.

The above solutions, if imparted might reduce the attrition rate in the company.

Data driven solution to the problem

The data consists of 6 different sheets:

1. Data Dictionary: Consists of metadata.
2. Employee Survey Data: This consists of employee survey data such as job satisfaction, work-life balance, and so on.
3. General data: The data consists of age, employee ID, gender, and other general information about the employee.
4. In time: The data consists of the login time of the employee.
5. Out Time: The data consists of the logout time of the employee.
6. Manager Survey Data: The survey data consists of manager ratings of the employee.

All of the above datais complied in a single sheet and considered for analysis.

Data

```
data<-read.csv("C:/Masters SPPU/3rd Sem/ST 013/ETE Assignment/employee data - general_data.csv")
```

General Information

Number of employees in the company

```
num_employee<-nrow(data);num_employee
```

```
## [1] 4410
```

Number of employees who left the company the previous year

```
num_employees_left<-length(which(data$Attrition=="Yes"));num_employees_left
```

```
## [1] 711
```

Attrition Rate

```
threshold<-num_employees_left/num_employee*100;threshold
```

```
## [1] 16.12245
```

The event under which an employee leaves the company has to be significant. Under the normal conditions, observing an employee leave the company may be unlikely. Observing the extreme cases will be more informative and can help understand the cause of an employee leaving the company. For example, companies paying employees below industry standard will have a higher attrition rate. In our case, the base attrition rate of the company is around 16%. But with the employees who have a low pay scale it is around 18%.

In the data it is verified that, if there is an NA in the employee log-in and log-out time, then an employee was on a leave. And for an NA in log-in time we have an NA in log-out time for that particular date.

Verification that, no one forgot to punch in or punch out

```
Out_Time<-data[,33:293]
In_Time<-data[,294:ncol(data)]
holiday<-c()
for (i in 1:ncol(In_Time))
{
  holiday[i]<-length(which(is.na(In_Time[,i])==TRUE))
}
Out_Time1<-Out_Time[,-which(holiday==num_employee)]
In_Time1<-In_Time[,-which(holiday==num_employee)]

x<-y<-c()
for (i in 1:num_employee)
{
  x[i]<-length(which(is.na(Out_Time1[i,])==TRUE))
  y[i]<-length(which(is.na(In_Time1[i,])==TRUE))
}
length(which(x!=y))
```

```
## [1] 0
```

Three features are constructed:

1. avg_time: This feature indicates the average time an employee spends in the company per day.
2. avg_time_In: This feature indicates the average time at which an employee enters the company daily.
3. employee_leaves: This feature indicates the number leaves taken by an employee apart from holidays throughout the year.

```
employee_leaves<-x

d<-data[,33:ncol(data)]
d[is.na(d)]<-"0"
n<-ncol(d)/2
Out_data<-d[1:n]
In_data<-d[(n+1):522]
time_diff<-intime<-outime<-list()
for(i in 1:n)
{
  outime[[i]]<-intime[[i]]<-time_diff[[i]]<-0*seq(dim(d)[1])
  for (j in 1:dim(d)[1])
  {
    temp<-as.difftime(c(Out_data[[i]][j],In_data[[i]][j]))
    time_diff[[i]][j]<-temp[1]-temp[2]
    if (is.na(time_diff[[i]][j])==TRUE)
      time_diff[[i]][j]<-0
    intime[[i]][j]<-temp[2]
    if (is.na(intime[[i]][j])==TRUE)
      intime[[i]][j]<-0
    outime[[i]][j]<-temp[1]
  }
}

avg_time<-c()
```

```
for(j in 1:dim(d)[1])
{
  a<-c()
  for(i in 1:n)
  {
    a[i]<-time_diff[[i]][j]
  }
  avg_time[j]<-mean(a)
}

avg_time_In<-c()
for(j in 1:dim(d)[1])
{
  b<-c()
  for(i in 1:n)
  {
    b[i]<-intime[[i]][j]
  }
  avg_time_In[j]<-mean(b)
}
Time_data<-data.frame(avg_time,avg_time_In,employee_leaves)
# write.csv(Time_data,file = "Time.csv")
```

Attrition rate of the extreme observations in the ordinal features.

For the below table, 5-percentile and 95-percentile values of the attrition rate are considered. The 5-percentile value indicates the value under which the 5 percent of the data lies. The 95-percentile value indicates the value below which the 95 percent of the data lies. By looking at the below table, we observe the attrition rate for very low and very high values of the features.

```
lower_tail_5perc<-upper_tail_5perc<-lower_tail_10perc<-upper_tail_10perc<-c()
for (i in 1:23)
{
  a<-length(which(is.na(data[,i+8]==TRUE)))
  if(a==0)
  {
    t<-data[,i+8]
  }
  if(a!=0)
  {
    t_NA<-which(is.na(data[,i+8]==TRUE))
    t<-data[,i+8][-t_NA]
  }
  q<-quantile(t,probs = c(0.05,0.95))
  lower_tail_5perc[i]<-length(which(data$Attrition[which(t<=q[1])]=="Yes"))/length(which(t<=q[1]))*100
  upper_tail_5perc[i]<-length(which(data$Attrition[which(t>=q[2])]=="Yes"))/length(which(t>=q[2]))*100
}
Final<-data.frame(lower_tail_5perc,upper_tail_5perc)
rownames(Final)<-colnames(data[,9:31])
colnames(Final)<-c("5 percentile","95 percentile")
Final<-Final[-c(4,6,15),];Final
```

```
##                          5 percentile 95 percentile
```

```
## MonthlyIncome              17.567568     13.513514
## JobLevel                   15.469613     14.857143
## NumCompaniesWorked         14.334471     19.141914
## PercentSalaryHike          14.285714     18.699187
## StockOptionLevel           16.798732     15.294118
## TotalWorkingYears          15.636364     16.450216
## TrainingTimesLastYear      16.000000     11.413043
## YearsAtCompany             34.883721      9.677419
## YearsSinceLastPromotion    18.932874     14.606742
## YearsWithCurrManager       32.319392      6.000000
## Age                        39.175258     12.643678
## DistanceFromHome           15.384615     13.793103
## EnvironmentSatisfaction    15.739645     15.667166
## JobSatisfaction            14.651163     15.727871
## WorkLifeBalance            15.899582     14.977974
## JobInvolvement             21.686747     18.055556
## PerformanceRating          15.755627     18.141593
## avg_time                    9.502262     26.244344
## avg_time_In                 8.144796     21.266968
## Employee_leaves            21.455939      9.703504
```

It is important to observe the higher attrition rate for the particular features. But not only this, we also need to observe lower attrition rates in order to get a picture of why employees stay?

From the above table we observe the following:

1. The employees who worked with more companies are more likely to switch than average.
2. The employees who have not spent much of the time within the company are more likely to make a switch.
3. The employees who stayed for a long time within the company are very less likely to make a switch.
4. The employees who have not spent much time with the current manager are more likely to make a switch.
5. The employees who spent a lot of time with their current manager are less likely to make a switch.
6. Younger employees are more likely to make a switch.
7. The employees who are rated low on a job involvement scale are more likely to make a switch. This is also an indication that their manager may not be satisfied with their performance.
8. The employees who spent lower daily time in the company are less likely to switch the company. This may also be an indication of not having a workload, making an employee satisfied.
9. The employees who spent more daily time in the company are more likely to switch the company. This may indicate higher workload.
10. The employees who tend to arrive early are less likely to leave the job. Maybe this is an indication of interest in the work.
11. The employees who tend to arrive late in the company are more likely to leave the job. It may be an indication of having a casual attitude towards the job.
12. The employees with lower leaves are more likely to leave the job.
13. The employees with more leaves are less likely to make a switch.

Attrition rate within the different levels of the nominal features.

```
D<-data[,c(1:8)]
D<-D[,-c(1,5)]
a<-c()
```

```r
prop<-most_attrition<-dataF<-list()
for (i in 1:ncol(D))
{
  a<-unique(D[,i])
  prop[[i]]<-rep(0,length(a))
  for (j in 1:length(a))
  {
    prop[[i]][j]<-length(which(D[,i]==a[j]&data$Attrition=="Yes"))/length(which(D[,i]==a[j]))*100
  }
  dataF[[i]]<-data.frame(a,prop[[i]])
  colnames(dataF[[i]])<-c("Levels","Attrition rate")
  most_attrition[[i]]<-a[which(prop[[i]]>threshold)]
}
a<-sort(unique(data$Education))
Education_Nominal<-c()
for (j in 1:length(a))
{
  Education_Nominal[j]<-length(which(data$Education==a[j]&data$Attrition=="Yes"))/length(which(data$Edu
}
levels_a<-c("Below College","College","Bachelor","Master","Doctor")
Edu<-data.frame(levels_a,Education_Nominal)
colnames(Edu)<-c("Levels","Attrition rate")
dataF[[7]]<-Edu
dataF
```

```
## [[1]]
##   Levels Attrition rate
## 1 Female       15.30612
## 2   Male       16.66667
##
## [[2]]
##                       Levels Attrition rate
## 1 Healthcare Representative       14.50382
## 2        Research Scientist       18.15068
## 3           Sales Executive       16.87117
## 4           Human Resources       13.46154
## 5         Research Director       23.75000
## 6     Laboratory Technician       16.21622
## 7    Manufacturing Director       11.03448
## 8      Sales Representative       14.45783
## 9                   Manager       13.72549
##
## [[3]]
##     Levels Attrition rate
## 1  Married       12.48143
## 2   Single       25.53191
## 3 Divorced       10.09174
##
## [[4]]
##              Levels Attrition rate
## 1     Travel_Rarely       14.95686
## 2 Travel_Frequently       24.90975
## 3        Non-Travel        8.00000
```

```
##
## [[5]]
##                   Levels Attrition rate
## 1                  Sales       15.02242
## 2 Research & Development       15.71280
## 3        Human Resources       30.15873
##
## [[6]]
##             Levels Attrition rate
## 1    Life Sciences       16.66667
## 2            Other       12.19512
## 3          Medical       16.16379
## 4        Marketing       15.72327
## 5 Technical Degree       11.36364
## 6  Human Resources       40.74074
##
## [[7]]
##          Levels Attrition rate
## 1 Below College       15.29412
## 2       College       18.79433
## 3      Bachelor       15.55944
## 4        Master       15.57789
## 5        Doctor       14.58333
```

From the above tables we observe the following:

1. Research directors are more likely to switch the company.
2. Single employees are more likely to leave the job. The reason might be to explore better opportunities in the initial stages.
3. Human resource employees are more likely to leave the job.
4. The employees who come from a human resource background are more likely to make a switch.
5. The employees who travel frequently are more likely to leave the job.
6. The employees who do not travel are less likely to leave the job.
7. Students with college level of education are more likely to make a switch. Although the attrition rate is not significant.

Analysing the NA values

While studying the data, it is also important to understand the missing values. In employee survey data, suppose there is a missing value in the environment satisfaction, then it can be an indication of an employee not being satisfied with the environment and do not want to say it directly. Hence below we understand the attrition rates for different features where the NAs are present.

Below attrition rates are only calculated for those features where there are missing values. And the attrition rates are calculated only considering the missing values and no other values.

```
d<-data.frame(data$EnvironmentSatisfaction,data$JobSatisfaction,data$WorkLifeBalance,data$TotalWorkingY
a<-b<-c<-c()
for (i in 1:ncol(d))
{
  a[i]<-length(which(is.na(d[,i]==TRUE)))
  b[i]<-length(which(is.na(d[,i]==TRUE)&data$Attrition=="Yes"))
  c[i]<-round(b[i]/a[i]*100,2)
```

```
}
NA_data<-data.frame(a,b,c)
colnames(NA_data)<-c("NA responses","Attritions","Attrition rate")
rownames(NA_data)<-c("EnvioronmentSatisfaction","JobSatisfaction"
                     ,"WorkLifeBalance","TotalWorkingYears","NumCompaniesWorked")
NA_data
```

```
##                           NA responses Attritions Attrition rate
## EnvioronmentSatisfaction           25          5          20.00
## JobSatisfaction                    20          1           5.00
## WorkLifeBalance                    38          4          10.53
## TotalWorkingYears                   9          2          22.22
## NumCompaniesWorked                 19          4          21.05
```

The above table indicates higher attrition rate for NAs for the features, Environment Satisfaction, Total Working Years, Number of Companies Worked. Missing values for Total Working Years and Number of Companies worked, may be an indication of frequent switch in companies for that employee, hence higher attrition rate.

But, the conclusion cannot be drawn due to low NA size.

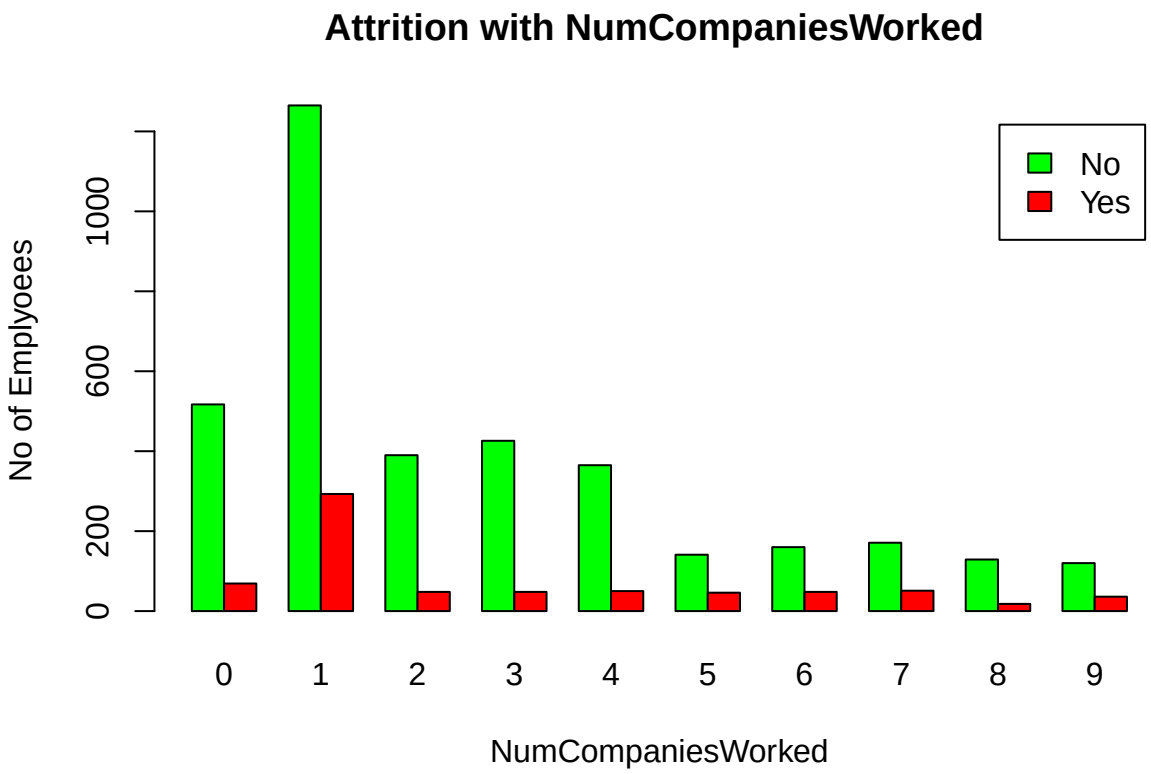
Graphical interpretations

The tables gave us a picture about the attrition for different features. Graphs will help us easily understand the overall picture of attrition for different employee features.

The bars in red indicate the count of employees who left the company with respect to a specific feature. For example, count of employees who left the company with respect to the number of years at the company.

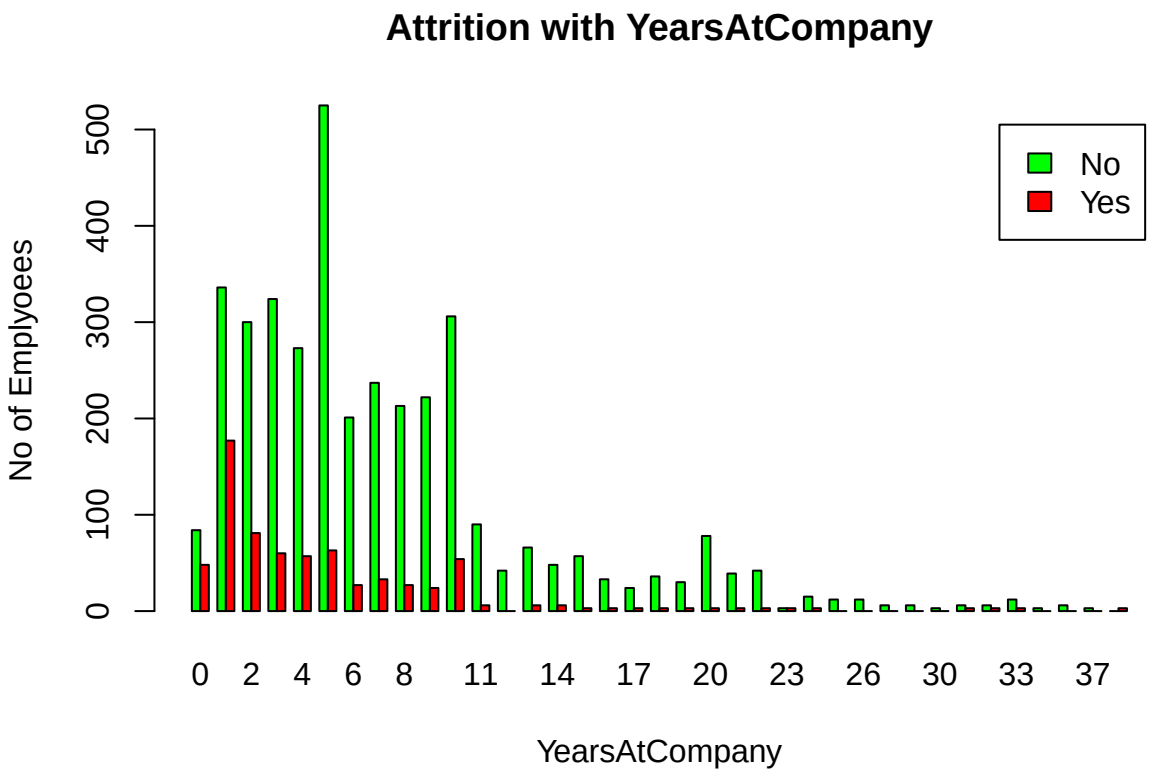
The bars in green indicate the count of employees who are still working in the company for that specific feature.

```
counts <- table(data$Attrition,data$NumCompaniesWorked)
barplot(counts, main="Attrition with NumCompaniesWorked ",
        xlab="NumCompaniesWorked",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



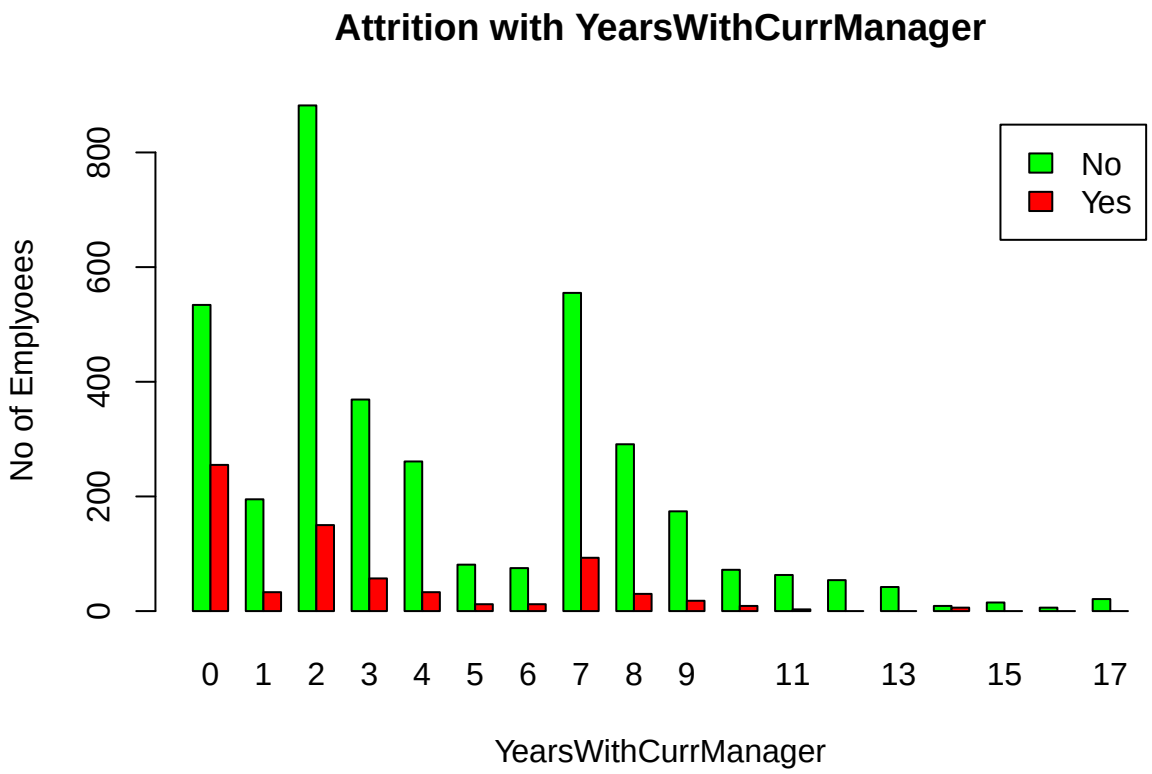
The graph indicates the count for the employees with respect to the number of companies they have worked before.

```
counts <- table(data$Attrition,data$YearsAtCompany)
barplot(counts, main="Attrition with YearsAtCompany",
        xlab="YearsAtCompany",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



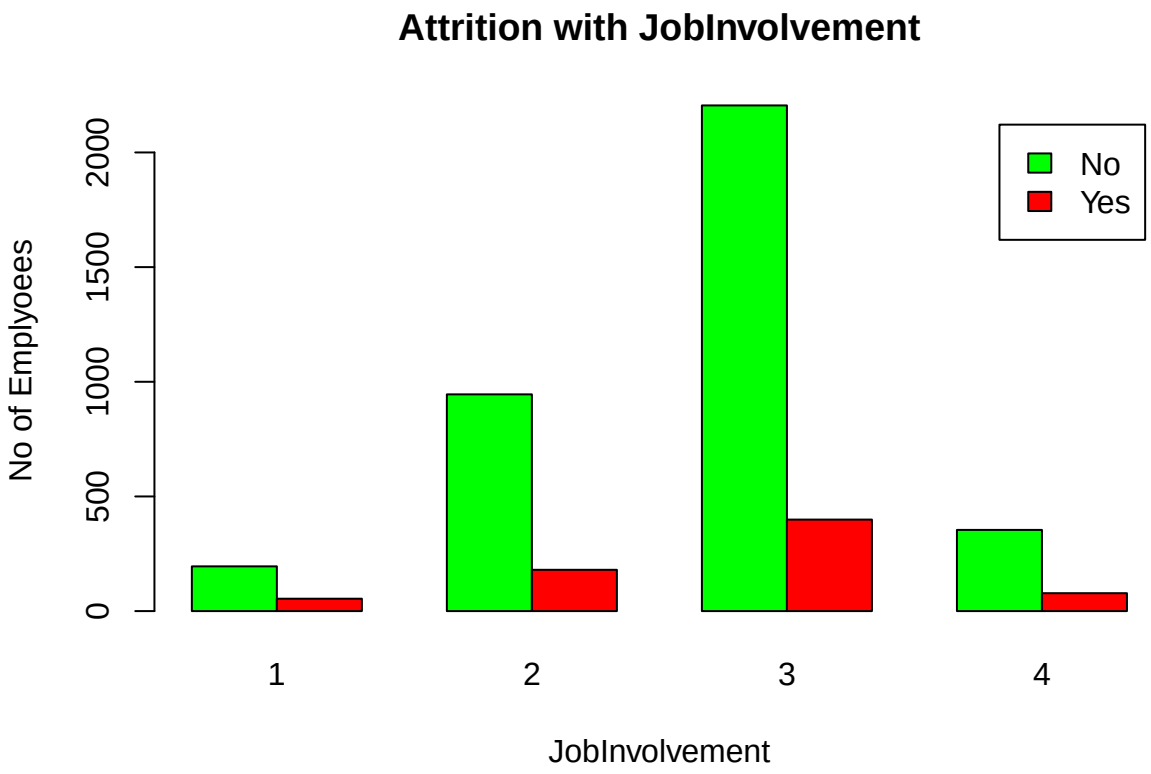
The graph indicates the count for the employees with respect to the years at the company.

```
counts <- table(data$Attrition,data$YearsWithCurrManager)
barplot(counts, main="Attrition with YearsWithCurrManager",
        xlab="YearsWithCurrManager",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



The graph indicates the count for the employees with respect to the years with the current manager.

```
counts <- table(data$Attrition,data$JobInvolvement)
barplot(counts, main="Attrition with JobInvolvement",
        xlab="JobInvolvement",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



The graph indicates the count for the employees with respect to their job involvement.

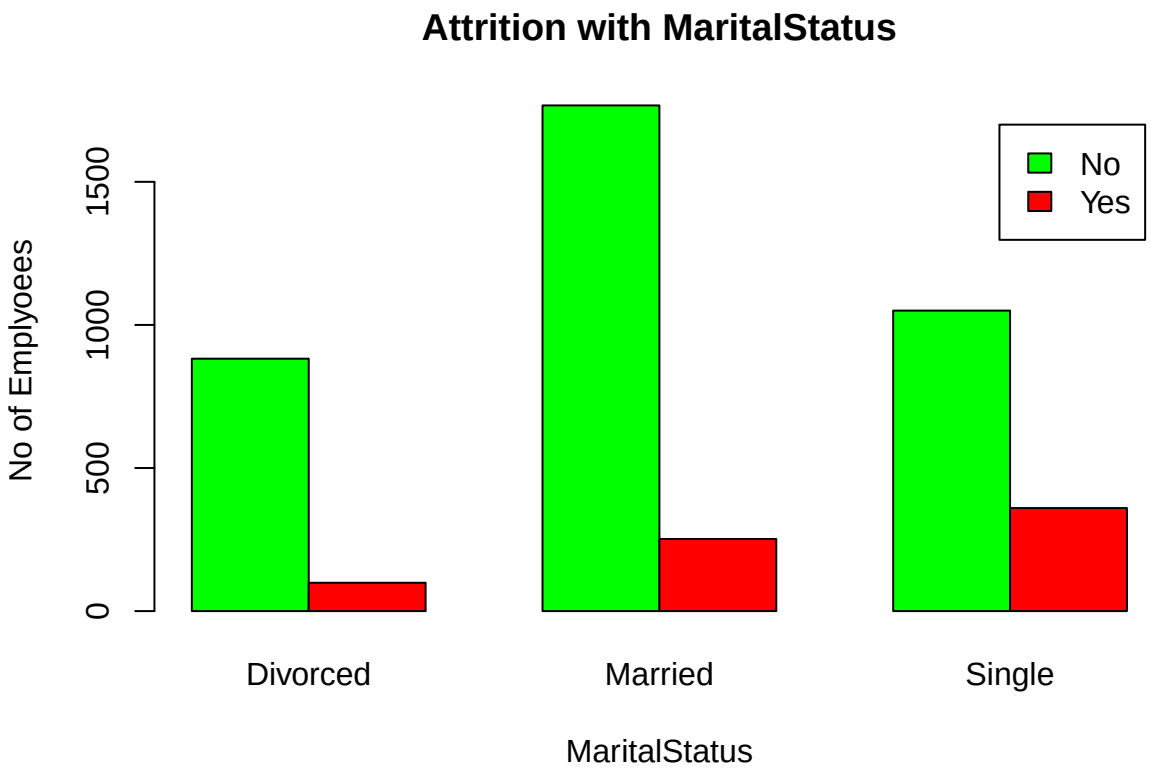
```
counts <- table(data$Attrition,data$Employee_leaves)
barplot(counts, main="Attrition with Employee_leaves",
        xlab="Employee_leaves",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```

The graph indicates the count for the employees with respect to the number of leaves they have taken.

```
counts <- table(data$Attrition,data$JobRole)
barplot(counts, main="Attrition with JobRole",
        xlab="JobRole",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```

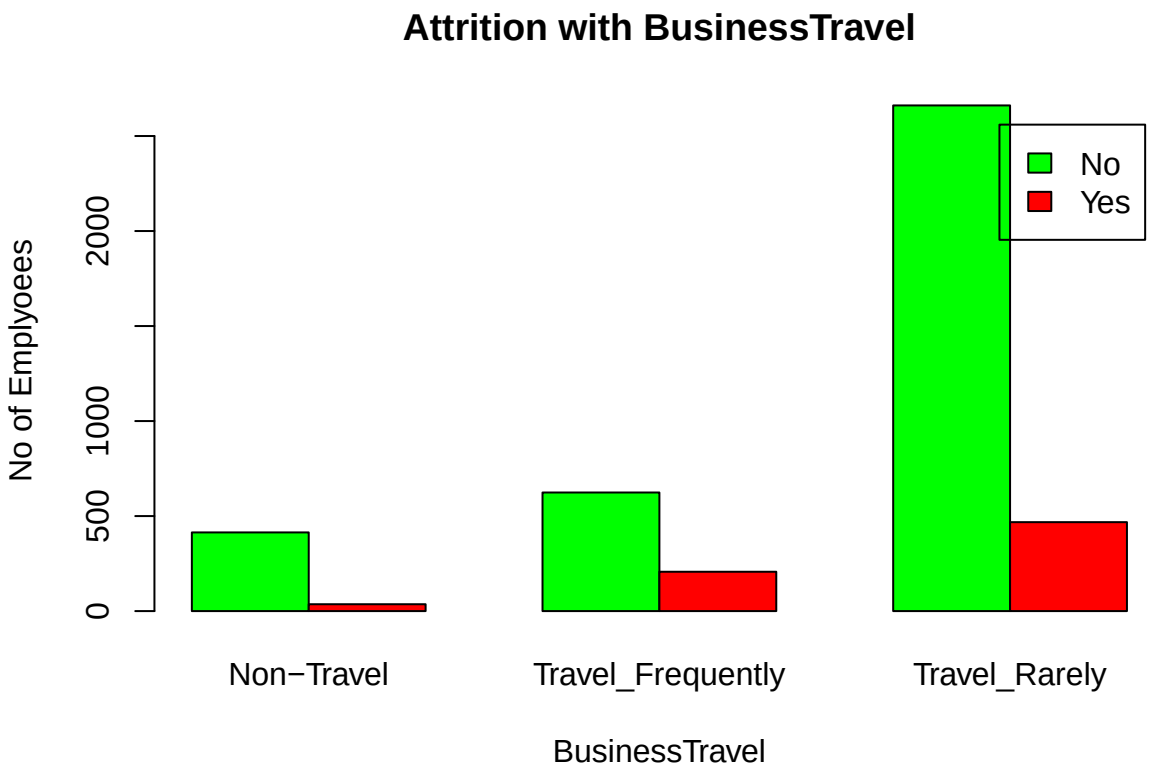
The graph indicates the count for the employees with respect to different job roles.

```
counts <- table(data$Attrition,data$MaritalStatus)
barplot(counts, main="Attrition with MaritalStatus",
        xlab="MaritalStatus",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



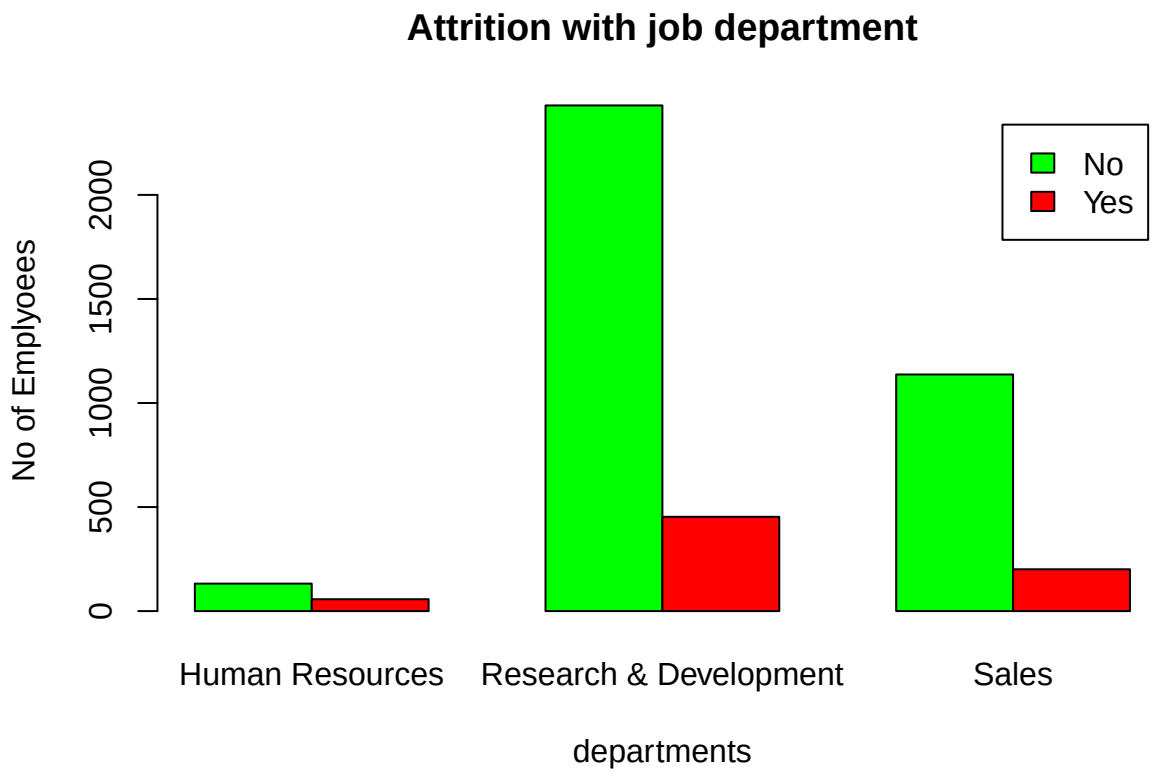
The graph indicates the count for the employees with respect to the employee's marital status.

```
counts <- table(data$Attrition,data$BusinessTravel)
barplot(counts, main="Attrition with BusinessTravel",
        xlab="BusinessTravel",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



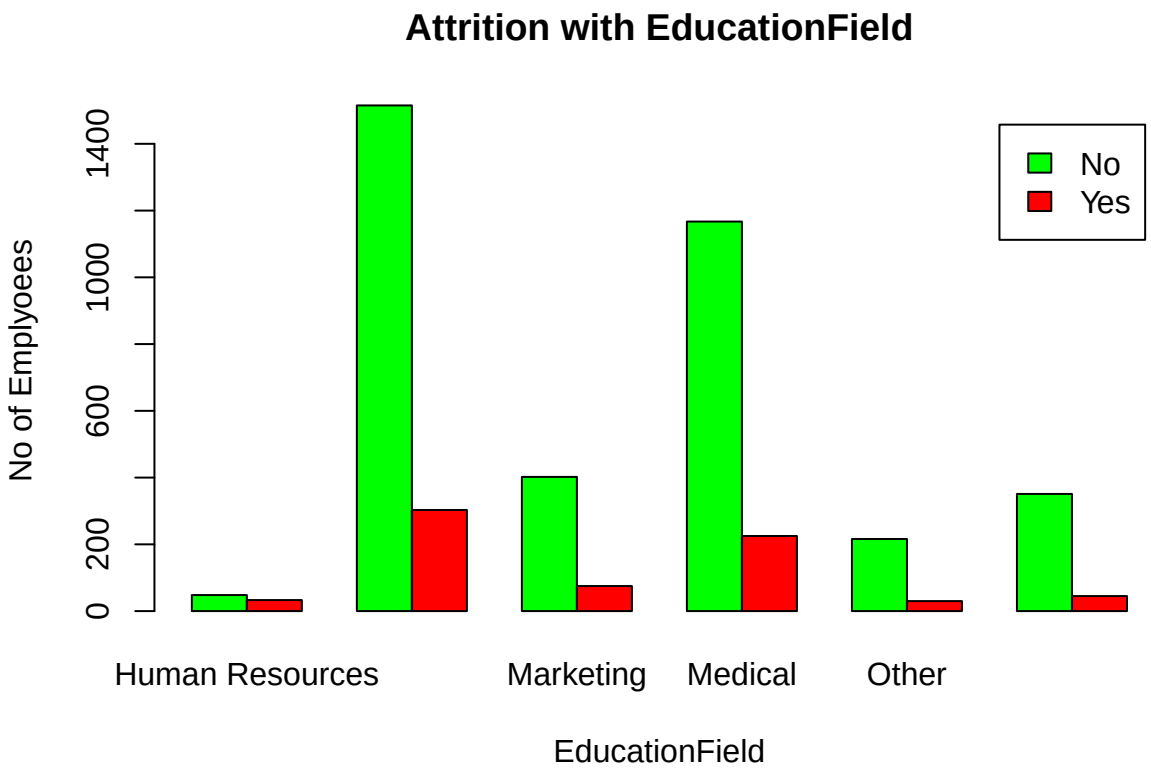
The graph indicates the count for the employees with respect to the employee's frequency of business travel.

```
counts<-table(data$Attrition,data$Department)
barplot(counts, main="Attrition with job department",
        xlab="departments",ylab="No of Emplyoees" ,col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



The graph indicates the count for the employees with respect to the employee's job department.

```
counts <- table(data$Attrition,data$EducationField)
barplot(counts, main="Attrition with EducationField",
        xlab="EducationField",ylab="No of Emplyoees", col=c("green","red"),
        legend = rownames(counts),beside=TRUE)
```



The graph indicates the count for the employees with respect to the employee's education field.

Curve Interpretations

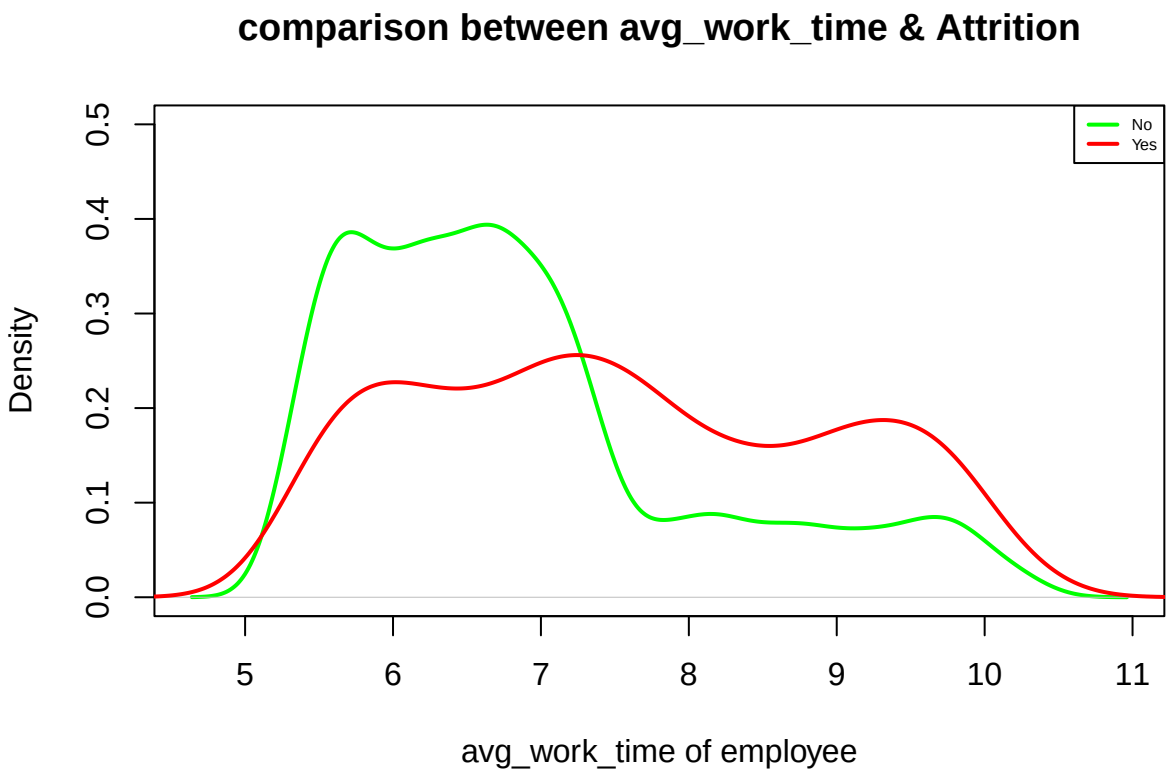
The below curves indicate the values of continuous features of the employees who left the company as well as for those who stayed. The curve for those who left is indicated in red and for those who stayed is indicated in green. The legend indicates the attrition value, that is:

1. 'Yes' for those who left the company.
2. 'No' for those who stayed.

```
Attrition_n<-unclass(as.factor(data$Attrition)) ##Yes<-2,No<-1,converting categorical variable tonumeri
Age<-data$Age
d<-data.frame(Attrition_n,Age)
plot(density(d$Age[which(d$Attrition_n==1)]),main="comparison between age & Attrition",
     xlab="Age of employee",col="green",ylim = c(0,0.06),lwd = 2)
lines(density(d$Age[which(d$Attrition_n==2)]),col = "red",lwd = 2)
legend<-legend(x = "topright",legend = c("No", "Yes")
               ,col= c("green","red"),lwd = 2, cex = 0.5)
```

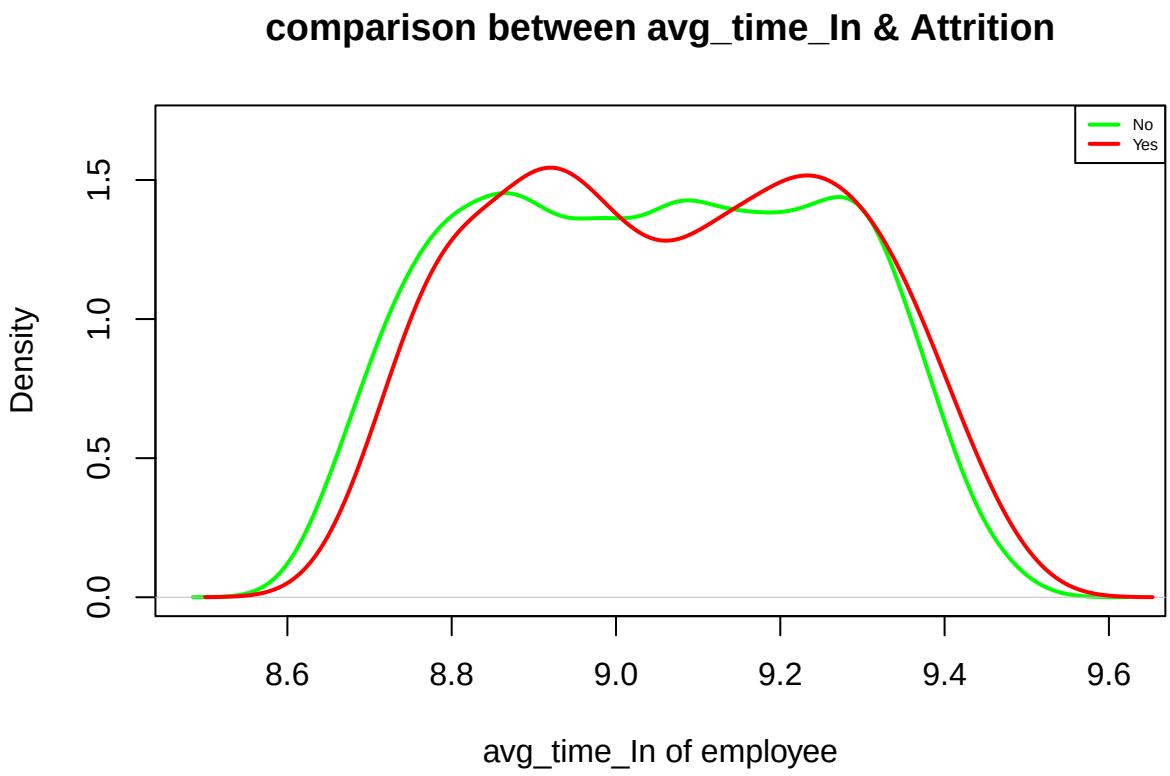
The graph indicate the curves for the employees with respect to the employee's age.

```
avg_work_time<-data$avg_time
d<-data.frame(Attrition_n,avg_work_time)
plot(density(d$avg_work_time[which(d$Attrition_n==1)]),main="comparison between avg_work_time & Attriti
             xlab="avg_work_time of employee",col="green",ylim=c(0,0.5),lwd = 2)
lines(density(d$avg_work_time[which(d$Attrition_n==2)]),col = "red",lwd = 2)
legend<-legend(x = "topright",legend = c("No", "Yes")
               ,col= c("green","red"),lwd = 2,cex = 0.5)
```



The graph indicates the curves for the employees with respect to the employee's average daily work time.

```
avg_time_In<-data$avg_time_In
d<-data.frame(Attrition_n,avg_time_In)
plot(density(d$avg_time_In[which(d$Attrition_n==1)]),main="comparison between avg_time_In & Attrition",
     xlab="avg_time_In of employee",col="green",ylim=c(0,1.7),lwd = 2)
lines(density(d$avg_time_In[which(d$Attrition_n==2)]),col = "red",lwd = 2)
legend<-legend(x = "topright",legend = c("No", "Yes")
               ,col= c("green","red"),lwd = 2,cex = 0.5)
```



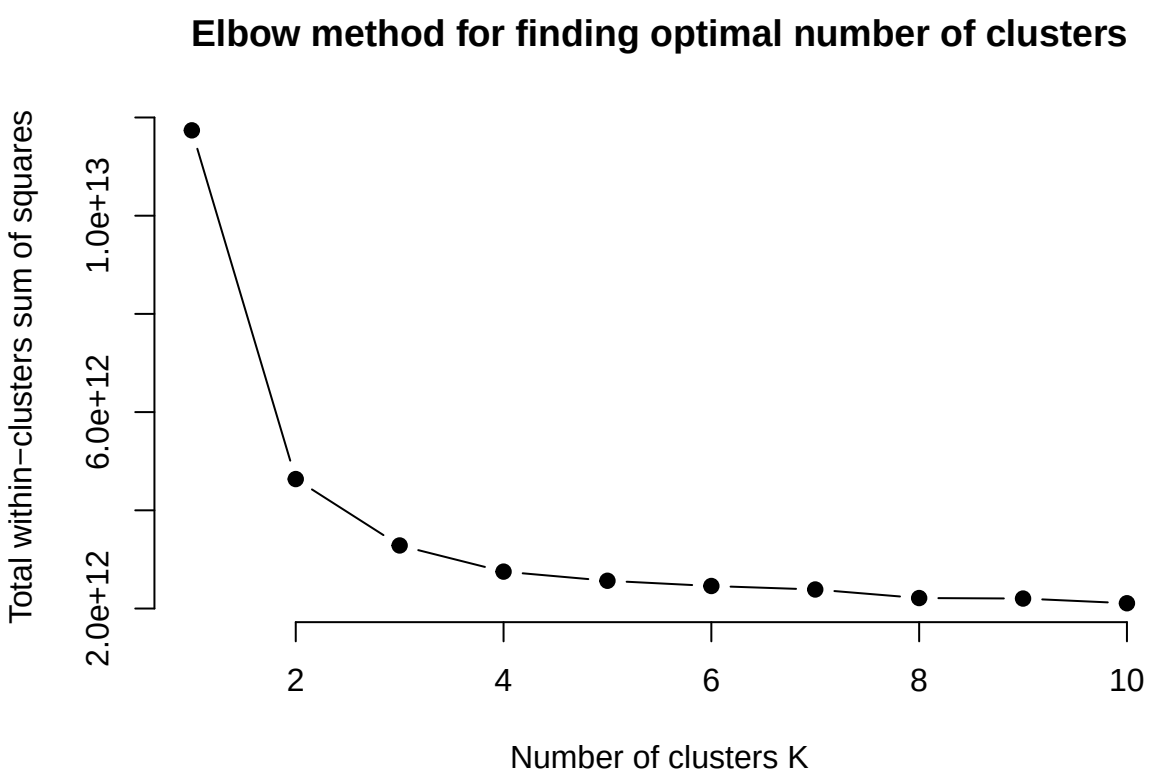
The graph indicates the curves for the employees with respect to the employee's average daily log-in time.

Result from Clustering

```
data1<-data[,1:31]
Non_NA_Data<-na.omit(data1)
data1<-data1[,-c(1,5,23,14)]
data1<-na.omit(data1)

factor<-c(1,2,3,4,5,6)
for(i in factor)
{
  data1[,i]<-as.factor(data1[,i])
}
# install.packages("clustMixType")
library(clustMixType)
k.max<-10
wss<-sapply(1:k.max,
            function(k){kproto(data1, k)$tot.withinss})
```

```
plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares",main = "Elbow method for finding optimal number of clus
```



From the plot, we choose the optimal number of clusters as 3.

Here, as we are dealing with a Mixed Data Type (Categorical and Numerical) we use the K-Prototype Clustering method.

K-Prototype is a clustering method based on partitioning. Its algorithm is an improvement of the K-Means and K-Mode clustering algorithm to handle clustering with the mixed data types.[1]

```
clust<-kproto(data1,3)
```

```
t<-table(clust$cluster,data$Attrition[Non_NA_Data$Employee.ID]);t
```

```
##
##       No  Yes
##   1  363   61
##   2 2511  517
##   3  731  117
```

Attrition rates within the three clusters formed.

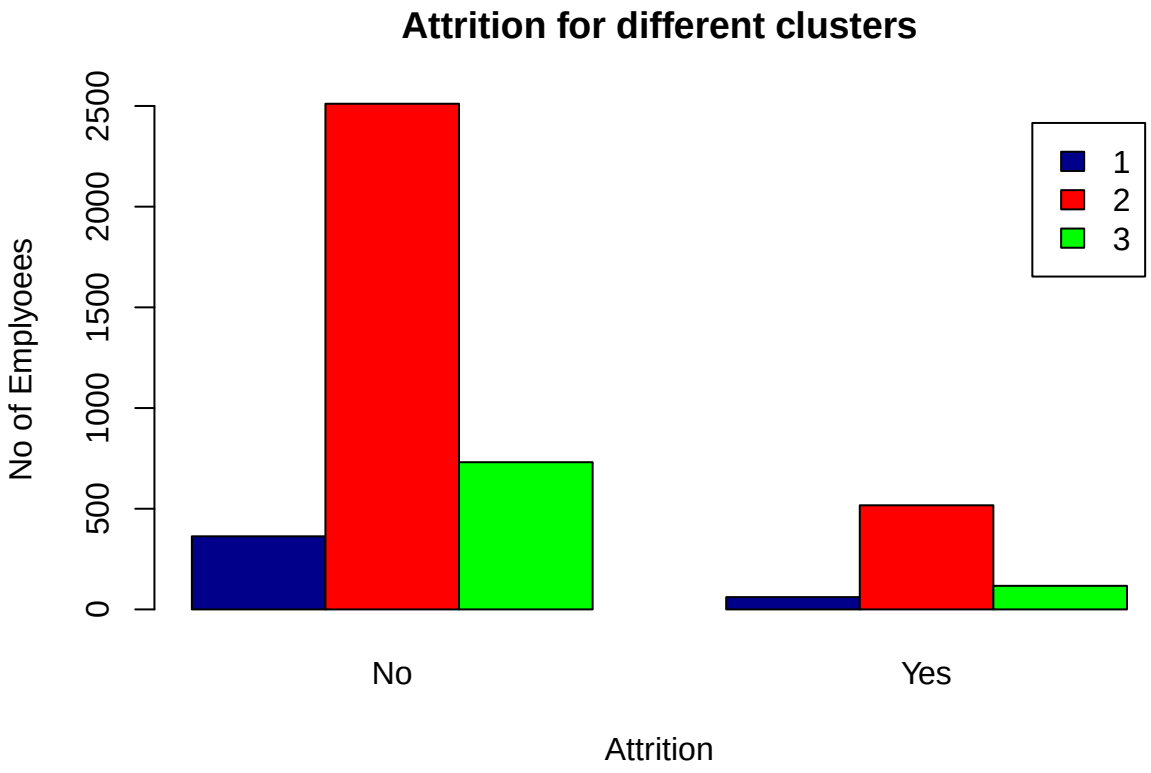
```
attr_rate_clust_1<-t[4]/(t[1]+t[4])*100
attr_rate_clust_2<-t[5]/(t[2]+t[5])*100
attr_rate_clust_3<-t[6]/(t[3]+t[6])*100
Clust_attr_rate<-t(data.frame(attr_rate_clust_1,attr_rate_clust_2,attr_rate_clust_3))
colnames(Clust_attr_rate)<-"Attrition rate"
rownames(Clust_attr_rate)<-c("Cluster 1","Cluster 2","Cluster 3")
Clust_attr_rate
```

```
##           Attrition rate
## Cluster 1       14.38679
## Cluster 2       17.07398
## Cluster 3       13.79717
```

From the table we can observe that the attrition rate for different clusters is not significantly different from the baseline attrition rate of the company. Hence, no conclusion can be drawn.

Diagrammatic representation of the attrition within each of the clusters.

```
barplot(t, main="Attrition for different clusters",xlab="Attrition"
        ,ylab="No of Emplyoees",col=c("darkblue","red","green")
        ,legend = rownames(t),beside=TRUE)
```



Three cohorts are detected from clustering. Interestingly, not much of a difference is observed in the attrition rates for these three different cohorts.

Final recommendations & conclusion

1. Increasing the number of leaves given to an employee might help in reducing the attrition rate for the company.
2. A counter-intuitive suggestion is: delaying promotion can also help in reducing the attrition rate.
3. Another counter-intuitive suggestion is: not to give large percentages of salary hike.
4. Since employees who travel frequently are more likely to leave the job, it is suggested that only those employees should be preferred for business travelling who spent a lot of time in this company. As, the employees who have given a lot of years to this company are less likely to leave the job.

5. Work load should be properly distributed among the employees. The employees who spent more than average daily time in the company are more likely to leave the job.
6. Minimum number of college level educated employees should be hired, since they contribute to a higher attrition rate.
7. Also the company should prefer to have a minimum number of HRs to reduce the attrition rate.

References

1. The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)

2. R version 4.2.1 (2022-06-23 ucrt) – "Funny-Looking Kid" Copyright (C) 2022 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit)