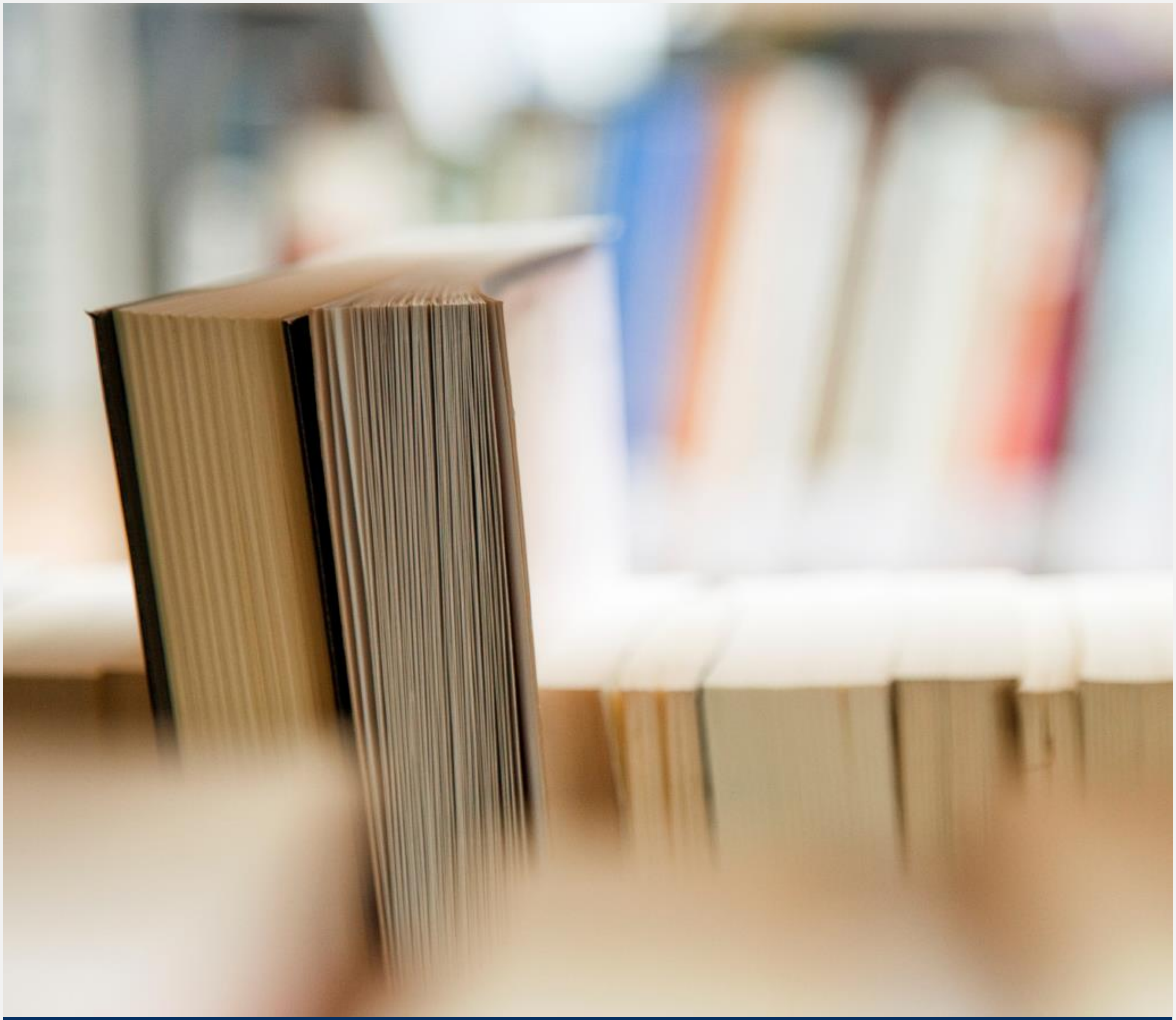


HR EMPLOYEE ATTRITION ANALYTICS

Aishat Abdulgafar

Table of Contents

THE DATA

Description of Data

The IBM HR Analytics Attrition Dataset, which I got on Kaggle, comprises data on employees for the IBM organization. The dataset provides a wealth of information on employees' demographic characteristics, work satisfaction, job environment, roles, and performance indicators, as well as their attrition status. With 35 variables and 1,470 observations, this dataset presents a diverse and extensive range of data for analyzing the drivers of employee attrition. This dataset provides resources for data scientists and researchers who aim to investigate the factors that impact employee retention and engagement in the workplace. Link to the dataset

Loading Packages

To start the data analysis process in my RMarkdown document, I loaded several packages that I needed for my analysis. The following code snippet shows the packages that were installed:

```
```{r}
install.packages("rsdmx")
install.packages("dplyr")
install.packages("ROSE")
install.packages("leaps")
install.packages("glmnet")
library(ggplot2)
library(rsdmx)
library(dplyr)
library(ROSE)
library(leaps)
library(glmnet)
```
```

Loading the Dataset

To load the dataset into R, I used the read.csv() function, which reads data from a CSV file and creates a data frame. After loading the dataset, I used the head() function to display the first five rows of the dataset.

```{r}
df <- read.csv("C:/Users/Aishat/Downloads/IBM-HR-Employee-Attrition.csv")
head(df)
```

Description: df [6 x 35]

	ï..Age <int>	Attrition <chr>	BusinessTravel <chr>	DailyRate <int>	Department <chr>	DistanceFromHome <int>	Education <int>	EducationField <chr>	EmployeeCount <int>
1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1
2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1
3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1
4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1
5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1
6	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1

6 rows | 1-10 of 35 columns

The “is.dataframe” function was then used to ensure that the dataset was properly loaded in to R as a dataframe and the dim() function was used to describe the number of rows and column present in the dataset which displays 1470 observations with 35 variables

```{r}
is.data.frame(df)
dim(df)
```

```
[1] TRUE
[1] 1470   35
```

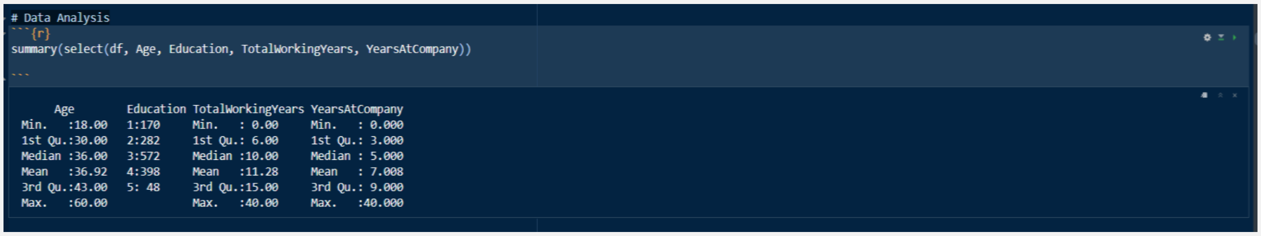
Original Data Structure

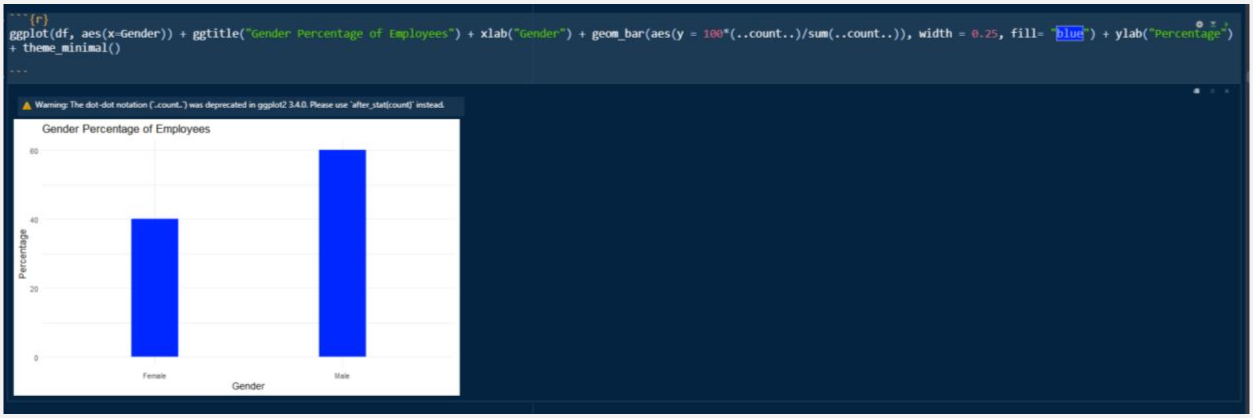
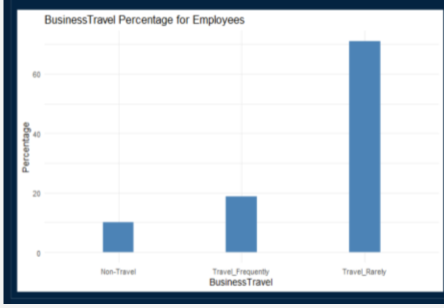
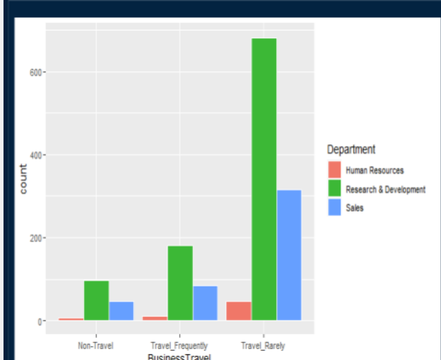
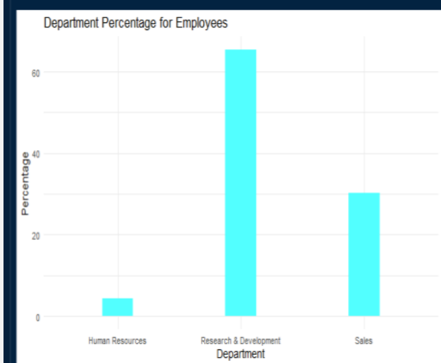
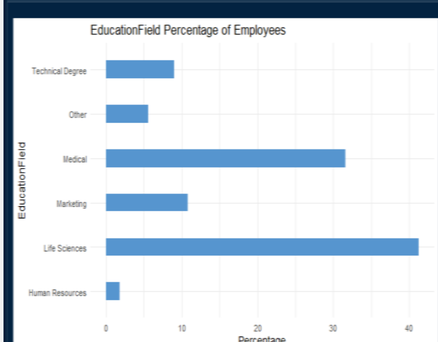
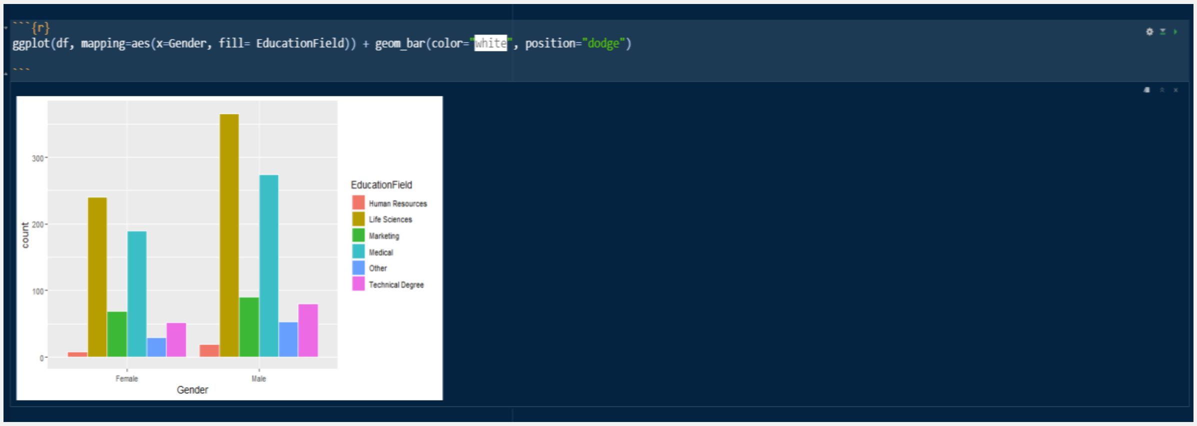
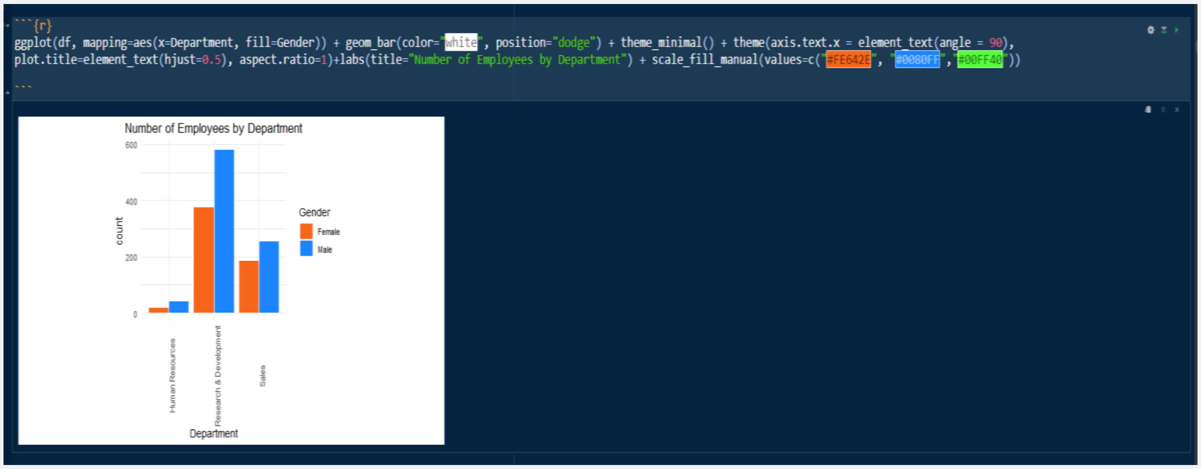
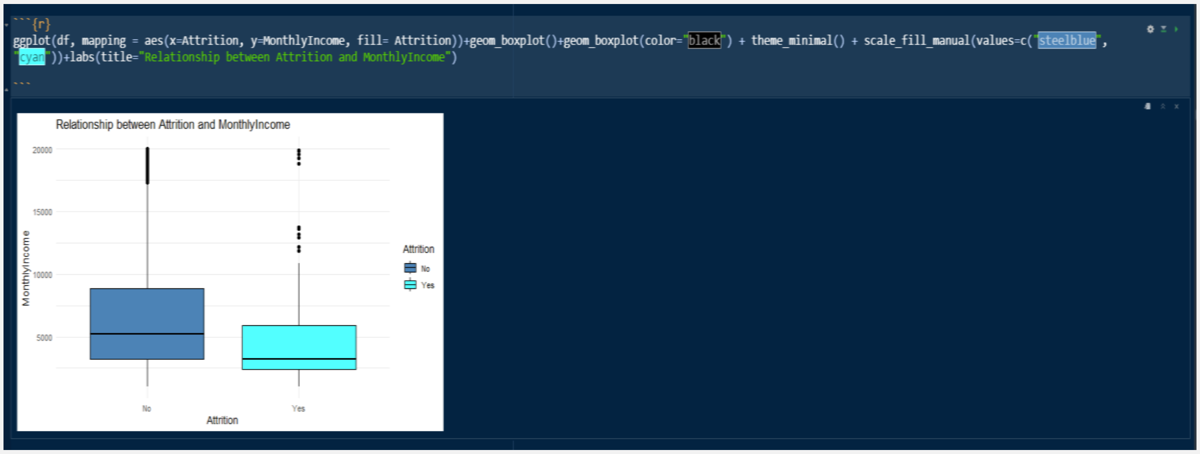
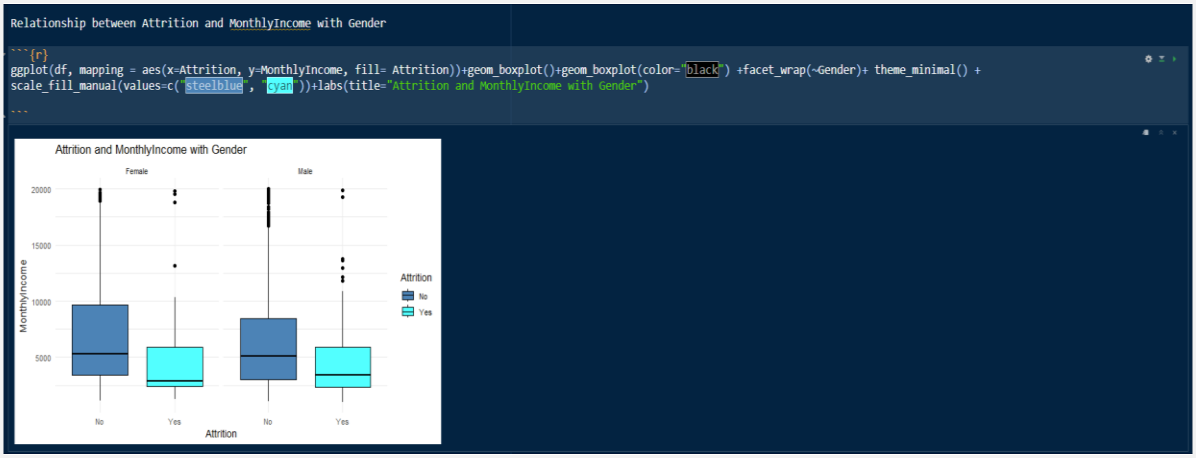
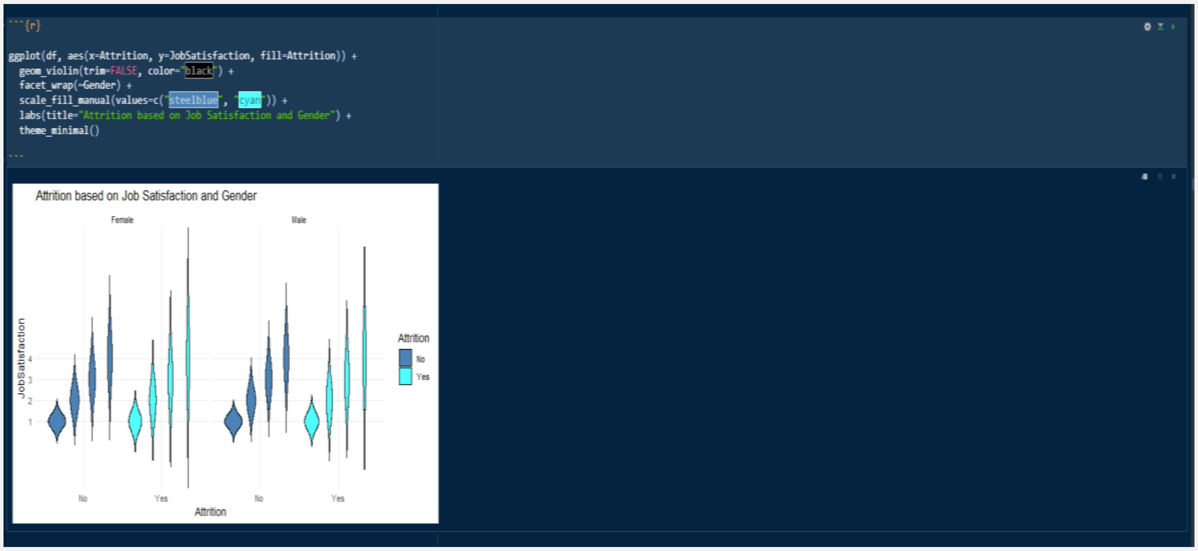
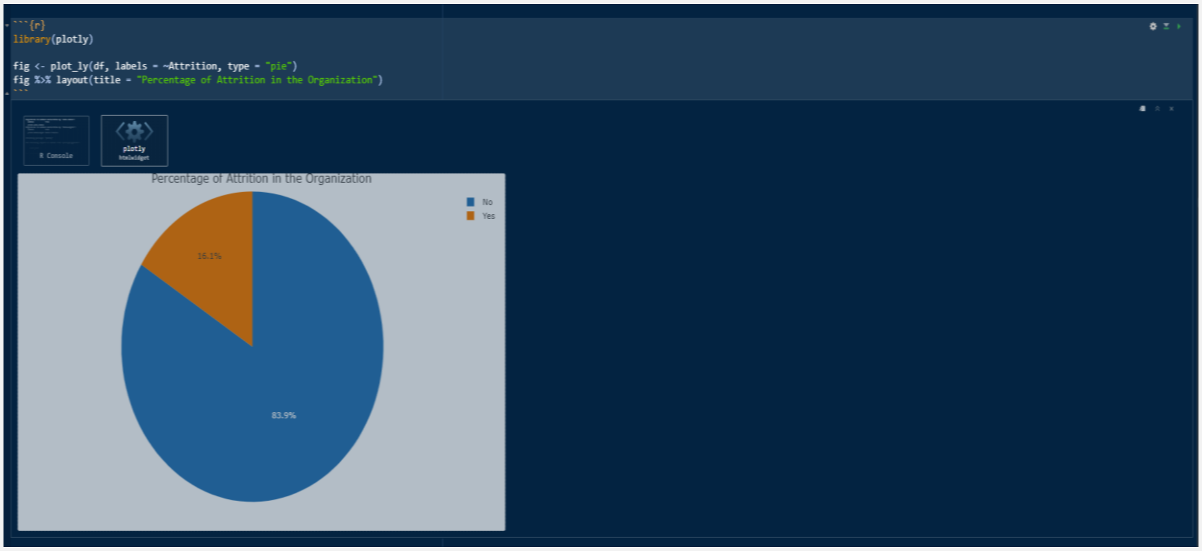
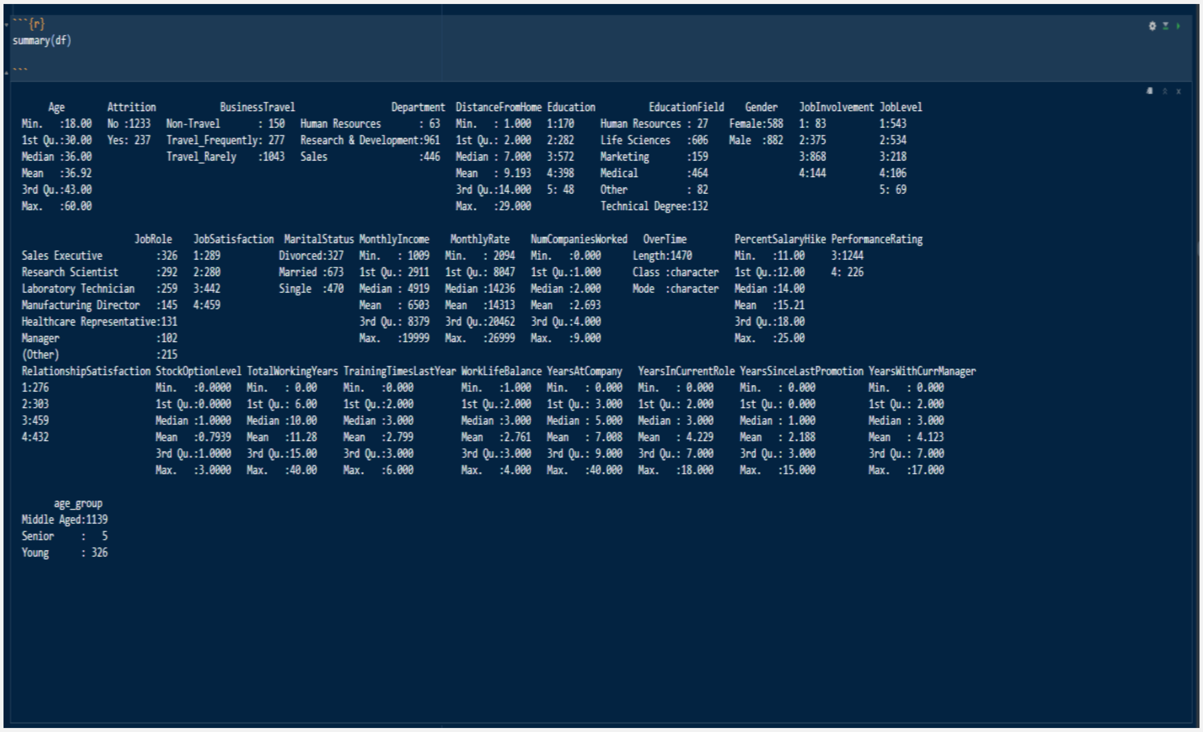
To obtain information about the variables in the dataset, I used the str() function, which displays the structure of the dataset, including its mode and data type. By using the str() function, I was able to determine that the dataset had variables with different modes, including numeric (num), integer (int), and character (chr) modes. This information is useful for performing data cleaning and preparation tasks, such as identifying missing values or correcting data types.

```
```{r}
str(df)

```

 'data.frame':   1470 obs. of 35 variables:
 $ ï..Age   : int 41 49 37 33 27 32 59 30 38 36 ...
 $ Attrition   : chr "Yes" "No" "Yes" "No" ...
 $ BusinessTravel   : chr "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel_Frequently" ...
 $ DailyRate   : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
 $ Department   : chr "Sales" "Research & Development" "Research & Development" "Research & Development" ...
 $ 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   : chr "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
 $ EmployeeCount   : int 1 1 1 1 1 1 1 1 1 1 ...
 $ EmployeeNumber   : int 1 2 4 5 7 8 10 11 12 13 ...
 $ EnvironmentSatisfaction : int 2 3 4 4 1 4 3 4 4 3 ...
 $ Gender   : chr "Female" "Male" "Male" "Female" ...
 $ HourlyRate   : int 94 61 92 56 40 79 81 67 44 94 ...
 $ JobInvolvement   : int 3 2 2 3 3 3 4 3 2 3 ...
 $ JobLevel   : int 2 2 1 1 1 1 1 1 3 2 ...
 $ JobRole   : chr "Sales Executive" "Research Scientist" "Laboratory Technician" "Research Scientist" ...
 $ JobSatisfaction   : int 4 2 3 3 2 4 1 3 3 3 ...
 $ MaritalStatus   : chr "Single" "Married" "Single" "Married" ...
 $ MonthlyIncome   : int 5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
 $ MonthlyRate   : int 19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
 $ NumCompaniesWorked   : int 8 1 6 1 9 0 4 1 0 6 ...
 $ Over18   : chr "Y" "Y" "Y" "Y" ...
 $ OverTime   : chr "Yes" "No" "Yes" "Yes" ...
 $ PercentSalaryHike   : int 11 23 15 11 12 13 20 22 21 13 ...
 $ PerformanceRating   : int 3 4 3 3 3 3 4 4 4 3 ...
 $ RelationshipSatisfaction: int 1 4 2 3 4 3 1 2 2 2 ...
 $ StandardHours   : int 80 80 80 80 80 80 80 80 80 80 ...
 $ StockOptionLevel   : int 0 1 0 0 1 0 3 1 0 2 ...
 $ TotalWorkingYears   : int 8 10 7 8 6 8 12 1 10 17 ...
 $ TrainingTimesLastYear   : int 0 3 3 3 3 2 3 2 2 3 ...
 $ WorkLifeBalance   : int 1 3 3 3 3 2 2 3 3 2 ...
 $ YearsAtCompany   : int 6 10 0 8 2 7 1 1 9 7 ...
```

To display a list of all the variables in the dataset, I used the colnames() function.

```
```{r}
colnames(df)
```

 [1] "ï..Age"   "Attrition"   "BusinessTravel"   "DailyRate"   "Department"   "DistanceFromHome"
 [7] "Education"   "EducationField"   "EmployeeCount"   "EmployeeNumber"   "EnvironmentSatisfaction" "Gender"
[13] "HourlyRate"   "JobInvolvement"   "JobLevel"   "JobRole"   "JobSatisfaction"   "MaritalStatus"
[19] "MonthlyIncome"   "MonthlyRate"   "NumCompaniesWorked"   "Over18"   "OverTime"   "PercentSalaryHike"
[25] "PerformanceRating"   "RelationshipSatisfaction" "StandardHours"   "StockOptionLevel"   "TotalWorkingYears"   "TrainingTimesLastYear"
[31] "WorkLifeBalance"   "YearsAtCompany"   "YearsInCurrentRole"   "YearsSinceLastPromotion" "YearsWithCurrManager"
```

Data Cleaning

The first cleaning on the data was to rename the i..Age column to Age. This was done with the rename function. Then I created a variable and named the variable age_group in order to group the employees into three groups such as young, middle aged and senior using the code below.

```
# Data Cleaning
```{r}
df <- df %>% rename(Age = ï..Age)

```

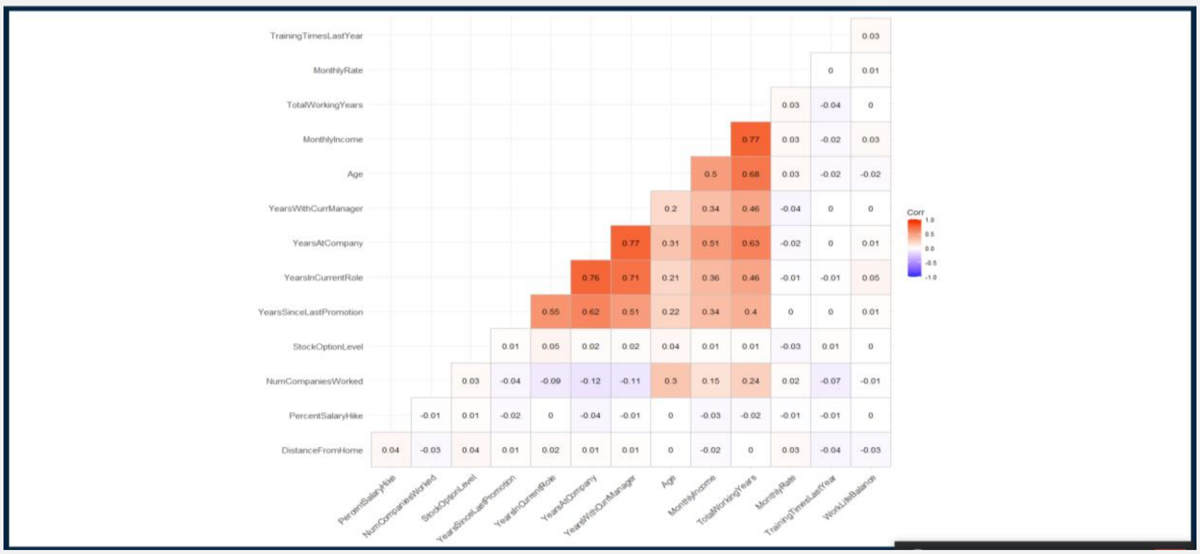
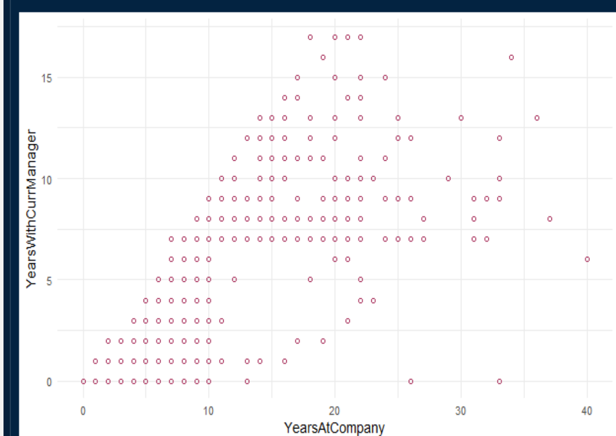
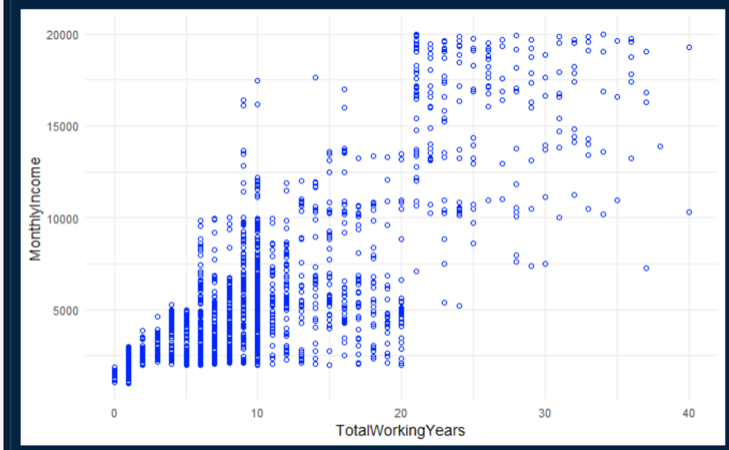
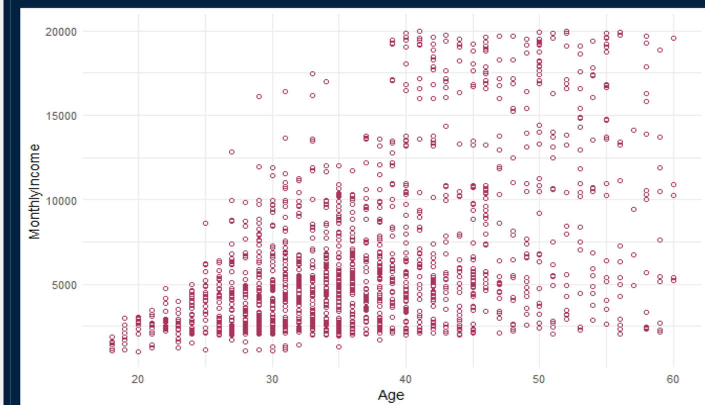
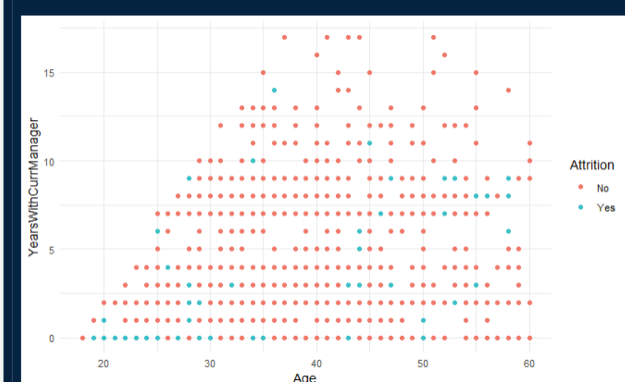
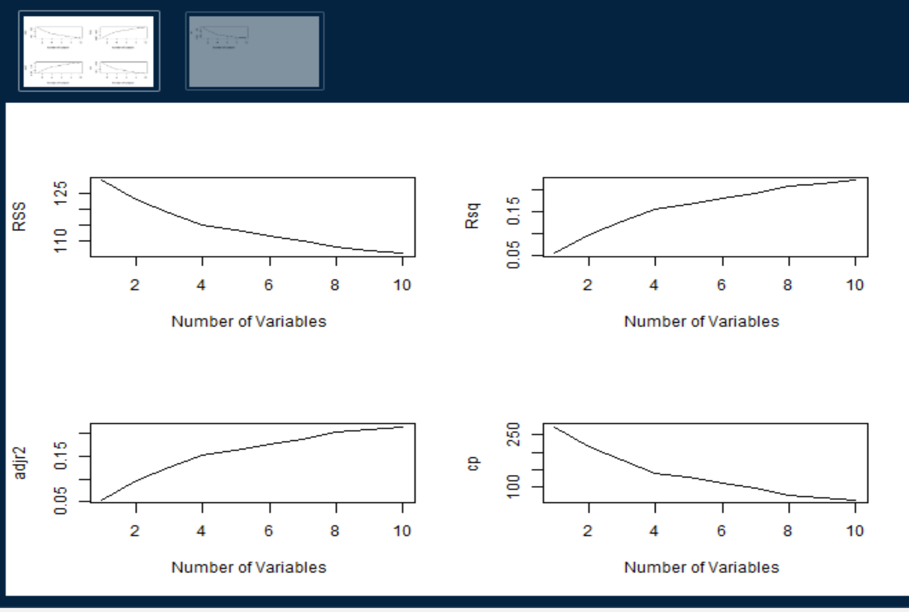
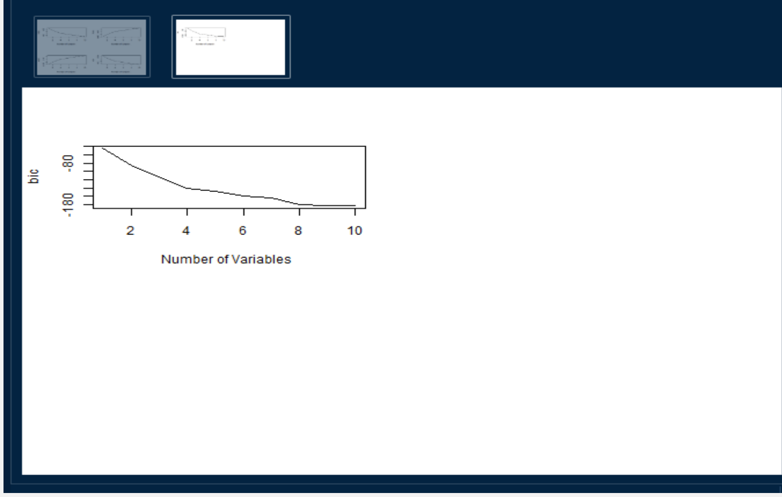
```{r}
# creating a new variable 'age_group' based on the 'Age' variable
df$age_group <- ifelse(df$Age < 30, 'Young',
                       ifelse(df$Age >= 30 & df$Age < 60, 'Middle Aged', 'Senior'))

```
```

As part of the data cleaning and preparation process, I converted some of the variables in the dataset to the appropriate data type. This was necessary to ensure that the variables were correctly represented and could be used in the analysis. I changed some variables to factor variable with levels.

````
```{r}
df$Education <- as.factor(df$Education)
df$Attrition <- as.factor(df$Attrition)
df$BusinessTravel <- as.factor(df$BusinessTravel)
df$Department <- as.factor(df$Department)
df$EducationField <- as.factor(df$EducationField)
df$Gender <- as.factor(df$Gender)
df$JobInvolvement <- as.factor(df$JobInvolvement)
df$JobLevel <- as.factor(df$JobLevel)
df$JobRole <- as.factor(df$JobRole)
df$JobSatisfaction <- as.factor(df$JobSatisfaction)
df$MaritalStatus <- as.factor(df$MaritalStatus)
df$PerformanceRating <- as.factor(df$PerformanceRating)
df$RelationshipSatisfaction <- as.factor(df$RelationshipSatisfaction)
df$age_group <- as.factor(df$age_group)
```
````

I excluded certain variables from the dataset that were deemed unnecessary for the analysis at hand. The removal of these variables aimed to simplify the analysis process and ensure that the remaining variables were relevant and informative for the project.

````
```{r}
df <- df[-c(22, 27, 4, 9, 10, 11, 13)]
```
````

Handling missing values is a crucial step in the data cleaning process, as missing values can affect the accuracy and validity of the analysis results, below is the code that was used.

````
# Handling missing values
```{r}
df <- df[complete.cases(df), ]
df[is.na(df)] <- sapply(df, function(x) ifelse(is.numeric(x), median(x, na.rm = TRUE), x))
```
````

Cleaned Data Frame

To determine if there were any missing values in the dataset after data cleaning, I used the code below.

````
# Cleaned Dataset
```{r}
colSums(sapply(df, is.na))
```
                     Age                Attrition           BusinessTravel               Department         DistanceFromHome                Education           EducationField
                       0                        0                        0                        0                        0                        0                        0
                  Gender           JobInvolvement                 JobLevel                  JobRole          JobSatisfaction            MaritalStatus            MonthlyIncome
                       0                        0                        0                        0                        0                        0                        0
             MonthlyRate       NumCompaniesWorked                 OverTime        PercentSalaryHike        PerformanceRating RelationshipSatisfaction         StockOptionLevel
                       0                        0                        0                        0                        0                        0                        0
       TotalWorkingYears    TrainingTimesLastYear          WorkLifeBalance           YearsAtCompany       YearsInCurrentRole  YearsSinceLastPromotion     YearsWithCurrManager
                       0                        0                        0                        0                        0                        0                        0
               age_group
                       0
````

After the dataset has been cleaned, the str() function was then used again to provide the updated data structure of the dataset. And the head() function was used to display the first few rows of the dataset.

````
```{r}

str(df)

```
````

```
'data.frame':   1470 obs. of  29 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 ...
 $ 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               : Factor w/ 5 levels "1","2","3","4",..: 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 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
 $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3 ...
 $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1 3 2 ...
 $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1 ...
 $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...
 $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
 $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
 $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
 $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
 $ OverTime                : chr  "Yes" "No" "Yes" "Yes" ...
 $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
 $ PerformanceRating       : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 2 2 2 1 ...
 $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3 1 2 2 2 ...
 $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
 $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
 $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
 $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
 $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
 $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
 $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
 $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
 $ age_group               : Factor w/ 3 levels "Middle Aged",..: 1 1 1 1 3 1 1 1 1 1 ...
```

````
```{r}
head(df)
```
````

Description: df [6 x 29]

	Age <int>	Attrition <fctr>	BusinessTravel <fctr>	Department <fctr>	DistanceFromHome <int>	Education <fctr>	EducationField <fctr>	Gender <fctr>	JobInvolvement <fctr>
1	41	Yes	Travel_Rarely	Sales	1	2	Life Sciences	Female	3
2	49	No	Travel_Frequently	Research & Development	8	1	Life Sciences	Male	2
3	37	Yes	Travel_Rarely	Research & Development	2	2	Other	Male	2
4	33	No	Travel_Frequently	Research & Development	3	4	Life Sciences	Female	3
5	27	No	Travel_Rarely	Research & Development	2	1	Medical	Male	3
6	32	No	Travel_Frequently	Research & Development	2	2	Life Sciences	Male	3

6 rows | 1-10 of 29 columns

Description of Data Frame Variables

Below is a table describing each variable in the HR Analytics Attrition Dataset. these descriptions were gotten from the source as described in the data description.

Column Name	Mode	Description
Age	Int	Age of the employee
Attrition	Chr	Level to which employee stays or leave the organization
BusinessTravel	Chr	Levels to how often the employee travels
DailyRate	Int	Daily amount paid to employee
Department	Chr	Department the employee works
DistanceFromHome	Int	The distance from home to work
Education	Int	Educational level of the employee
Education Field	Chr	Employee field of knowledge
EmployeeCount	Int	The count of employee
EmployeeNumber	Int	Employee id number
EnvironmentSatisfaction	Int	Work environment satisfaction of the employee
Gender	Chr	Gender of the employee
HourlyRate	Int	Hourly rate paid to employee
JobInvolvement	Int	How involve the employee is with work
JobLevel	Int	Level of the employee
JobRole	Chr	Role of the employee
JobSatisfaction	Int	Employee satisfaction with the work
MaritalStatus	Chr	Marital status of employee
MonthlyIncome	Int	Monthly income of employee
MonthlyRate	Int	Monthly rate of employee
NumCompaniesWorked	Int	No. of companies employee has worked
Over18	Chr	Determines if employee is over 18years
OverTime	Chr	Determines if the employee works overtime
PercentSalaryHike	Int	Percentage change in employee salary
PerformanceRating	Int	Employee performance rating
RelationshipSatisfaction	Int	Employee and colleagues relationship satisfaction
StandardHours	Int	Hours require from employee to work
StockOptionLevel	Int	Company stock owned by employee
TotalWorkingYears	Int	Working years of employee at the organization

TrainingTimesLastYear	Int	Number of employee training within last year
WorkLifeBalance	Int	determines the work life balance of employees
YearsAtCompany	Int	Years the employee has been with the company
YearsinCurrentRole	Int	Years employee has been the current role
YearsSinceLastPromotion	Int	Years since the last employee promotion
YearsWithCurrManager	Int	Years of employee with the current manager

Expectations

This analysis would focus on exploring the HR analytics attrition dataset to deeply understand the dataset and the variables that contribute to attrition. My analysis would consider various factors that may influence employee turnover, such as age, MonthlyIncome, JobSatisfaction, PerformanceRating and worklife. These variables would be analyzed using appropriate statistical techniques such as different classification analysis to identify the most predictors of attrition. Additionally, data visualization such as scatterplot, pie, boxplots' would be used to identify patterns or trends in the data that may be relevant to the analysis. Ultimately, my analysis would provide actionable insights and recommendations that can be used by the organization to reduce employee turnover and improve employee retention. By effectively predicting attrition.

DATA EXPLORATION AND ANALYSIS

To start my data exploration, I used the code below to show the count, mean, median minimum and maximum values for age, education, TotalWorkingYears and YearsAtCompany for each variable as well as the 1st and 3rd quartiles. This was used to gain a better understanding of the distribution and central tendency of these variables and to identify any potential outliers or data anomalies.



````
# Data Analysis
```{r}
summary(select(df, Age, Education, TotalWorkingYears, YearsAtCompany))
```

      Age        Education TotalWorkingYears YearsAtCompany
 Min.   :18.00   1:170     Min.   : 0.00     Min.   : 0.000
 1st Qu.:30.00   2:282     1st Qu.: 6.00     1st Qu.: 3.000
 Median :36.00   3:572     Median :10.00     Median : 5.000
 Mean   :36.92   4:398     Mean   :11.28     Mean   : 7.008
 3rd Qu.:43.00   5: 48     3rd Qu.:15.00     3rd Qu.: 9.000
 Max.   :60.00             Max.   :40.00     Max.   :40.000
````

To gain insights into the gender distribution of employees in the dataset, I generated a graph that depicts the percentage of gender distribution in the HR analytics attrition dataset. The resulting plot revealed that 60% of the employees are male while 40% are female.



```{r}
summary(df$BusinessTravel)
```

```
      Non-Travel Travel_Frequently     Travel_Rarely
             150               277              1043
```

From the summary statistics above, we can see that the rate of employees that non travel or employees that do not travel at all is the least with a total of 150, and employees that travel rarely has the highest with a total of 1043.

```{r}
ggplot(df, aes(x=BusinessTravel)) + ggtitle("BusinessTravel Percentage for Employees") + xlab("BusinessTravel") + geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.25, fill="steelblue") + ylab("Percentage") + theme_minimal()
```



The bar chart presented above depicts the frequency of work-related travel among employees in the dataset. The analysis reveals that approximately 10% of the employees do not require work-related travel, while roughly 18% travel frequently for business purposes. The majority of employees (approximately 70%) rarely or never travel for work. These findings provide useful insights into the travel requirements of the organization. Below graph shows the distribution of employee business travel in different departments of the organization.

```{r}
ggplot(df, mapping=aes(x=BusinessTravel, fill= Department)) + geom_bar(color="white", position="dodge")
```



The bar chart presented below depicts the distribution of employees across different departments within the organization. The analysis reveals that the human resources department has the highest number of employees, accounting for approximately 70% of the total workforce. The sales department has approximately 30% of the employees, while only about 10% of employees are in the Human Resources department.

```{r}
ggplot(df, aes(x=Department)) + ggtitle("Department Percentage for Employees") + xlab("Department") + geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.25, fill= "cyan") +
ylab("Percentage") + theme_minimal()
```



The graph presented below provides insights into the educational backgrounds of employees within the organization. The analysis reveals that employees with a background in life sciences constitute the highest proportion of the workforce, accounting for approximately 40% of total employees. On the other hand, employees with a background in human resources have the lowest representation within the organization.

```{r}
ggplot(df, aes(x=EducationField)) + ggtitle("EducationField Percentage of Employees") + xlab("EducationField") + geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.4, fill=
"steelblue3") + ylab("Percentage") + coord_flip() + theme_minimal()
```



The graph below also shows the distribution of employees’ educational field across the organization based on their gender. Life sciences and marketing are leading in both. While human resources is the least educational field in both gender.





The two aforementioned plots exhibit a clear trend of male employees outnumbering female employees in both department and educational field, which is consistent with the global trend of males pursuing higher education at a higher rate than females. It is worth noting, however, that this does not necessarily imply any gender bias in the hiring process, as there may be other factors that contribute to this disparity such as societal and cultural norms.

Correlation between Attrition and other Variables

The following code displays the correlation between JobRole, DistanceFromHome, and Attrition in the HR analytics dataset.

```{r}
job.Dist.df<-xtabs(~JobRole+DistanceFromHome+Attrition,data=df)
job.Dist.df
margin.table(job.Dist.df,1)

```

```
, , Attrition = No

                          DistanceFromHome
JobRole                     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
  Healthcare Representative 23 14  4  3  6  4  8  8  5  7  4  3  2  1  2  4  1  4  0  4  1  0  2  3  4  1  2  1  1
  Human Resources            8  9  3  2  1  2  0  5  1  2  1  0  1  0  0  0  1  0  1  0  0  0  0  1  1  1  0  0  0
  Laboratory Technician     27 25 12  9  7 12 10 10 18  6  5  3  2  1  3  5  3  3  6  4  3  1  5  5  2  0  1  3  6
  Manager                   13 22  6  8  6  3  5  3  4  5  3  0  1  0  0  1  2  1  0  0  1  3  0  0  1  6  0  0  3
  Manufacturing Director    23 22  4  6  4  5 10  5  5  6  3  3  1  0  1  3  2  6  3  2  2  1  2  3  4  0  3  2  4
  Research Director         13 10  6  1  3  4  7  5  2  6  1  1  0  3  5  0  1  1  0  0  0  2  2  0  1  1  0  3  0
  Research Scientist        40 32 15  9 14 14 13 10 19 12  2  1  2  5  5  4  3  3  6  4  2  3  7  2  2  6  2  5  3
  Sales Executive           33 37 19 14 11  7 16 21 10 26  5  3  2  6  5  8  2  4  3  5  4  2  3  1  4  7  1  6  4
  Sales Representative       2 12  1  3  3  1  4  3  3  5  1  0  2  1  0  0  0  0  0  2  2  1  1  1  0  0  0  1  1

, , Attrition = Yes

                          DistanceFromHome
JobRole                     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
  Healthcare Representative  0  1  0  0  0  0  0  0  0  0  1  0  0  1  1  0  0  0  0  1  1  0  1  1  0  0  0  0  1
  Human Resources            1  1  0  0  0  1  0  1  1  0  0  0  1  0  0  0  1  1  0  1  0  2  1  0  0  0  0  0  0
  Laboratory Technician      4 11  3  3  2  3  6  4  4  3  0  1  0  2  2  2  2  0  0  0  1  0  0  5  2  0  0  1  1
  Manager                    0  3  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  1
  Manufacturing Director     1  2  1  0  0  0  1  1  0  2  0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  0  0  0  0
  Research Director          0  1  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  Research Scientist         7  4  5  2  3  1  1  2  2  4  1  0  0  1  0  2  2  2  1  0  0  1  1  1  3  0  0  0  1
  Sales Executive            6  2  3  4  2  1  0  2  5  2  2  2  5  0  1  3  0  1  1  1  0  1  1  3  1  3  3  1  1
  Sales Representative       7  3  2  0  3  1  3  0  6  0  0  2  0  0  0  0  0  0  1  1  1  1  0  2  0  0  0  0  0

JobRole
Healthcare Representative           Human Resources     Laboratory Technician                   Manager    Manufacturing Director         Research Director        Research Scientist
                      131                        52                       259                       102                       145                        80                       292
          Sales Executive      Sales Representative
                      326                        83
```

According to the output of the code below, employees in the Sales Executive job role have the highest percentage of individuals who travel long distances to work from home, and they also have the highest rate of attrition.

```{r}
job.Dist.Attr.df<- as.data.frame(margin.table(job.Dist.df, 1))
job.Dist.Attr.df

```

Description: df [9 x 2]

JobRole <fctr>	Freq <int>
Healthcare Representative	131
Human Resources	52
Laboratory Technician	259
Manager	102
Manufacturing Director	145
Research Director	80
Research Scientist	292
Sales Executive	326
Sales Representative	83

9 rows

According to the figure below, there is a relationship between MonthlyIncome and Attrition. This indicates that employees with a lower MonthlyIncome, averaging less than $5000, tend to have a higher attrition rate compared to those with a MonthlyIncome of an average of greater than or equal to $5000.



The plot below reveals that there is no correlation between MonthlyIncome and gender in relation to attrition. This implies that regardless of gender, the likelihood of attrition remains consistent. However, there is a significant association between MonthlyIncome and attrition. Specifically, employees with a lower MonthlyIncome (less than $5000 on average) are more prone to attrition than those with a MonthlyIncome of $5000 or more on average.



The plot below illustrates the distribution of job satisfaction levels among employees who stayed in the organization versus those who left. The plot shows that employees who left the organization had lower job satisfaction levels, with most having a job satisfaction level of 3 or lower. Additionally, the plot indicates that female employees had lower job satisfaction levels compared to their male counterparts.



Checking the distribution of target variable

To gain a better understanding of the attrition variable, I conducted an exploration of the number of employees who left the company and those who remained. This involved utilizing the table function to obtain an aggregate and subsequently plotting the percentage distribution of the frequency.



Summary of the Variables Statistics

The summary () function was employed to provide an overview of the statistical properties and distributions of the variables in the dataset.



````
```{r}
summary(df)

```

      Age        Attrition            BusinessTravel                  Department  DistanceFromHome Education
 Min.   :18.00   No :1233   Non-Travel       : 150   Human Resources       : 63   Min.   : 1.000   1:170
 1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   Research & Development:961   1st Qu.: 2.000   2:282
 Median :36.00              Travel_Rarely    :1043   Sales                 :446   Median : 7.000   3:572
 Mean   :36.92                                                                    Mean   : 9.193   4:398
 3rd Qu.:43.00                                                                    3rd Qu.:14.000   5: 48
 Max.   :60.00                                                                    Max.   :29.000
          EducationField    Gender    JobInvolvement JobLevel
 Human Resources : 27   Female:588   1: 83          1:543
 Life Sciences   :606   Male  :882   2:375          2:534
 Marketing       :159                3:868          3:218
 Medical         :464                4:144          4:106
 Other           : 82                               5: 69
 Technical Degree:132

                      JobRole    JobSatisfaction  MaritalStatus MonthlyIncome    MonthlyRate    NumCompaniesWorked   OverTime
 Sales Executive          :326   1:289           Divorced:327   Min.   : 1009   Min.   : 2094   Min.   :0.000      Length:1470
 Research Scientist       :292   2:280           Married :673   1st Qu.: 2911   1st Qu.: 8047   1st Qu.:1.000      Class :character
 Laboratory Technician    :259   3:442           Single  :470   Median : 4919   Median :14236   Median :2.000      Mode  :character
 Manufacturing Director   :145   4:459                          Mean   : 6503   Mean   :14313   Mean   :2.693
 Healthcare Representative:131                                  3rd Qu.: 8379   3rd Qu.:20462   3rd Qu.:4.000
 Manager                  :102                                  Max.   :19999   Max.   :26999   Max.   :9.000
 (Other)                  :215
 PercentSalaryHike PerformanceRating
 Min.   :11.00     3:1244
 1st Qu.:12.00     4: 226
 Median :14.00
 Mean   :15.21
 3rd Qu.:18.00
 Max.   :25.00

 RelationshipSatisfaction StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
 1:276                    Min.   :0.0000   Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
 2:303                    1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
 3:459                    Median :1.0000   Median :10.00     Median :3.000         Median :3.000   Median : 5.000
 4:432                    Mean   :0.7939   Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008
                          3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
                          Max.   :3.0000   Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000
 YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
 1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
 Median : 3.000     Median : 1.000          Median : 3.000
 Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
 3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :18.000     Max.   :15.000          Max.   :17.000

       age_group
 Middle Aged:1139
 Senior     :   5
 Young      : 326
````

Correlation of Numerical Variables

```
```{r}
numeric1 <- select_if(df, is.numeric)
cor(numeric1)

```
```

```
                                Age DistanceFromHome MonthlyIncome  MonthlyRate NumCompaniesWorked PercentSalaryHike StockOptionLevel TotalWorkingYears TrainingTimesLastYear
Age                      1.000000000     -0.001686120   0.497854567  0.028051167        0.299634758       0.003633585      0.037509712       0.680380536          -0.019620819
DistanceFromHome        -0.001686120      1.000000000  -0.017014445  0.027472864       -0.029250804       0.040235377      0.044871999       0.004628426          -0.036942234
MonthlyIncome            0.497854567     -0.017014445   1.000000000  0.034813626        0.149515216      -0.027268586      0.005407677       0.772893246          -0.021736277
MonthlyRate              0.028051167      0.027472864   0.034813626  1.000000000        0.017521353      -0.006429346     -0.034322830       0.026442471           0.001466881
NumCompaniesWorked       0.299634758     -0.029250804   0.149515216  0.017521353        1.000000000      -0.010238309      0.030075475       0.237638590          -0.066054072
PercentSalaryHike        0.003633585      0.040235377  -0.027268586 -0.006429346       -0.010238309       1.000000000      0.007527748      -0.020608488          -0.005221012
StockOptionLevel         0.037509712      0.044871999   0.005407677 -0.034322830        0.030075475       0.007527748      1.000000000       0.010135969           0.011274070
TotalWorkingYears        0.680380536      0.004628426   0.772893246  0.026442471        0.237638590      -0.020608488      0.010135969       1.000000000          -0.035661571
TrainingTimesLastYear   -0.019620819     -0.036942234  -0.021736277  0.001466881       -0.066054072      -0.005221012      0.011274070      -0.035661571           1.000000000
WorkLifeBalance         -0.021490028     -0.026556004   0.030683082  0.007963158       -0.008365685      -0.003279636      0.004128730       0.001007646           0.028072207
YearsAtCompany           0.311308770      0.009507720   0.514284826 -0.023655107       -0.118421340      -0.035991262      0.015058008       0.628133155           0.003568666
YearsInCurrentRole       0.212901056      0.018844999   0.363817667 -0.012814874       -0.090753934      -0.001520027      0.050817873       0.460364638          -0.005737504
YearsSinceLastPromotion  0.216513368      0.010028836   0.344977638  0.001566800       -0.036813892      -0.022154313      0.014352185       0.404857759          -0.002066536
YearsWithCurrManager     0.202088602      0.014406048   0.344078883 -0.036745905       -0.110319155      -0.011985248      0.024698227       0.459188397          -0.004095526
                        WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
Age                        -0.021490028    0.311308770        0.212901056             0.216513368          0.202088602
DistanceFromHome           -0.026556004    0.009507720        0.018844999             0.010028836          0.014406048
MonthlyIncome               0.030683082    0.514284826        0.363817667             0.344977638          0.344078883
MonthlyRate                 0.007963158   -0.023655107       -0.012814874             0.001566800         -0.036745905
NumCompaniesWorked         -0.008365685   -0.118421340       -0.090753934            -0.036813892         -0.110319155
PercentSalaryHike          -0.003279636   -0.035991262       -0.001520027            -0.022154313         -0.011985248
StockOptionLevel            0.004128730    0.015058008        0.050817873             0.014352185          0.024698227
TotalWorkingYears           0.001007646    0.628133155        0.460364638             0.404857759          0.459188397
TrainingTimesLastYear       0.028072207    0.003568666       -0.005737504            -0.002066536         -0.004095526
WorkLifeBalance             1.000000000    0.012089185        0.049856498             0.008941249          0.002759440
YearsAtCompany              0.012089185    1.000000000        0.758753737             0.618408865          0.769212425
YearsInCurrentRole          0.049856498    0.758753737        1.000000000             0.548056248          0.714364762
YearsSinceLastPromotion     0.008941249    0.618408865        0.548056248             1.000000000          0.510223636
YearsWithCurrManager        0.002759440    0.769212425        0.714364762             0.510223636          1.000000000
```

```
```{r}
library(ggcorrplot)
ab <- cor(numeric1)
ggcorrplot(ab, hc.order =TRUE,
         type ="lower", lab =TRUE)
```
```



The correlation test generated a correlation matrix, which shows that there are a few positive correlations between the numeric variables in the HR analytics attrition dataset. there is a high correlation between MonthlyIncome and TotalWorkingYears, YearsAtCompany and YearsWithCurrManager, YearsInCurrentRole and YearsAtCompany, YearsSinceLastPromotion and YearsInCurrentRole, and medium correlation between Age and MonthlyIncome, and very low correlation between YearsWithCurrManager and Age.

Plots for Correlated Variables

The correlation analysis conducted on MonthlyIncome and TotalWorkingYears revealed a strong positive correlation, indicating that as the age of employees increases, their total working years also increase. This suggests that older employees tend to have higher incomes, which could be due to their accumulated experience and expertise in their respective fields.

````
```{r}
cor(df$YearsAtCompany, df$YearsWithCurrManager)
```
````

```
[1] 0.7692124
```

````
```{r}
ggplot(df, mapping = aes(x=YearsAtCompany, y= YearsWithCurrManager)) + geom_point(color="maroon", shape=1)+theme_minimal()
```
````



Based on the graph below, there is a positive correlation between TotalWorkingYears and MonthlyIncome, meaning that as the TotalWorkingYears of the employees increase, their MonthlyIncome tends to increase as well.

```
```{r}
cor(df$MonthlyIncome, df$TotalWorkingYears)
```
```

```
[1] 0.7728932
```

```
```{r}
ggplot(df, mapping = aes(x=TotalWorkingYears, y= MonthlyIncome)) + geom_point(color="blue", shape=1)+theme_minimal()
```
```



The graph below shows the correlation between YearsInCurrentRole and YearsAtCompany of employees. This depicts that as the number of years at the company increases the years in current role also increases. Although attrition tends to be higher between 0-5 and 20-25 years at the company and 5-10years of employees in current role.

```{r}
cor(df$YearsInCurrentRole, df$YearsAtCompany)
```

```
[1] 0.7587537
```

```{r}
ggplot(df, mapping = aes(x=YearsAtCompany, y= YearsInCurrentRole, color=Attrition)) + geom_point()+theme_minimal()
```

There is a positive correlation of 0.548 between YearsSinceLastPromotion and YearsInCurrentRole. This shows that have had between 5-7years since their last promotion had high attrition rate and employees that have spent 10 years and above in their current role had the least attrition.

```{r}
cor(df$YearsSinceLastPromotion, df$YearsInCurrentRole)
```

```
[1] 0.5480562
```

```{r}
ggplot(df, mapping = aes(x=YearsSinceLastPromotion, y= YearsInCurrentRole, color=Attrition)) + geom_point()+theme_minimal()
```

There is a positive correlation between Age and MonthlyIncome of employees. This shows that as the Age of the employees increases, their MonthlyIncome also increases.

```{r}
cor(df$Age, df$MonthlyIncome)
```

```
[1] 0.4978546
```

```{r}
ggplot(df, mapping = aes(x=Age, y= MonthlyIncome)) + geom_point(color="maroon", shape=1)+theme_minimal()
```



There's a slight correlation between YearsWithCurrManager and Age. This means that as the age of the employees increases, a medium number of employees remain with their current managers and that employees that are between 20-30years of age and have spent less than 10years with their current managers had high attrition.

```{r}
cor(df$YearsWithCurrManager, df$Age)
```

```
[1] 0.2020886
```

```{r}
ggplot(df, mapping = aes(x=Age, y=YearsWithCurrManager, color=Attrition)) + geom_point()+theme_minimal()
```



Predictive Modeling

I began by loading the ROSE package, which was previously installed to facilitate the analysis. The package was utilized to oversample the minority class (Attrition) in the HR employee attrition dataset using the ovun.sample() function. I then used the table() function to count the number of instances in each class of the Attrition target variable.

```
```{r}
new.df <- ovun.sample(Attrition ~ ., data = df, method = "over", N = nrow(df), seed = 1234)$data
View(new.df)
class_count <- table(new.df$Attrition)
class_count
```

  No Yes
1233 237
```

The following code was used to split the dataset into training and testing sets. The purpose of this split was to use the training set to build a model to predict the response variable "Attrition" and then test the model's accuracy using the testing set.

```
```{r}

set.seed(1234)
new.df_split <- sample(x=nrow(new.df), size=.70*nrow(new.df))
train <-new.df[new.df_split,]
test<-new.df[-new.df_split,]
```
```

Logistics Regression

To find the best Logistic regression model for the HR employee analytics attrition dataset, I created several models using different variables in the dataset. The goal was to use MSE to determine the best model. The first model utilizes feature selection variables in the dataset, with Attrition as the response variable. The hypothesis was that the other variables in the "new.df" dataset would have an impact on predicting Attrition for the organization. I used the validation set approach to validate the model, by calculating the MSE for the different models.

Hypothesis testing

This is to predict the Attrition based with the variables present in the dataset. The response variable for this prediction is categorical.

Model 1

```{r}
glm.fit <-glm(Attrition~., data =train, family="binomial")
summary(glm.fit)

```

```
Call:
glm(formula = Attrition ~ ., family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5522  -0.4130  -0.1812  -0.0347   3.5966  

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                       -1.745e+01  6.315e+02  -0.028 0.977953    
Age                               -1.623e-02  2.214e-02  -0.733 0.463466    
BusinessTravelTravel_Frequently    3.109e+00  6.356e-01   4.892 9.99e-07 ***
BusinessTravelTravel_Rarely        1.694e+00  6.026e-01   2.811 0.004944 ** 
DepartmentResearch & Development   1.602e+01  6.315e+02   0.025 0.979763    
DepartmentSales                    1.274e+01  6.315e+02   0.020 0.983909    
DistanceFromHome                   4.146e-02  1.434e-02   2.890 0.003851 ** 
Education2                        -1.515e-01  4.598e-01  -0.330 0.741763    
Education3                         1.744e-01  3.698e-01   0.472 0.637245    
Education4                         9.468e-01  4.132e-01   2.291 0.021947 *  
Education5                         1.524e-02  8.426e-01   0.018 0.985571    
EducationFieldLife Sciences       -6.319e-01  1.212e+00  -0.521 0.602115    
EducationFieldMarketing           -3.841e-01  1.275e+00  -0.301 0.763257    
EducationFieldMedical             -4.180e-01  1.197e+00  -0.349 0.726861    
EducationFieldOther               -4.559e-01  1.268e+00  -0.360 0.719214    
EducationFieldTechnical Degree     5.176e-01  1.266e+00   0.409 0.682536    
GenderMale                         3.276e-01  2.422e-01   1.353 0.176215    
JobInvolvement2                   -1.150e+00  4.821e-01  -2.385 0.017084 *  
JobInvolvement3                   -1.819e+00  4.729e-01  -3.847 0.000120 ***
JobInvolvement4                   -1.474e+00  5.785e-01  -2.549 0.010817 *  
JobLevel2                         -8.252e-01  5.348e-01  -1.543 0.122831    
JobLevel3                          6.973e-01  9.638e-01   0.723 0.469381    
JobLevel4                         -1.461e+00  1.707e+00  -0.856 0.392040    
JobLevel5                          4.197e+00  2.496e+00   1.682 0.092595 .  
JobRoleHuman Resources             1.787e+01  6.315e+02   0.028 0.977429    
JobRoleLaboratory Technician       1.491e+00  8.111e-01   1.838 0.066091 .  
JobRoleManager                     1.474e+00  1.836e+00   0.803 0.422216    
JobRoleManufacturing Director      1.991e+00  7.462e-01   2.669 0.007611 ** 
JobRoleResearch Director          -9.653e-01  1.836e+00  -0.526 0.599023    
JobRoleResearch Scientist          7.423e-01  8.001e-01   0.928 0.353548    
JobRoleSales Executive             5.337e+00  1.987e+00   2.686 0.007223 ** 
JobRoleSales Representative        5.077e+00  2.048e+00   2.479 0.013167 *  
JobSatisfaction2                  -4.264e-01  3.992e-01  -1.068 0.285434    
JobSatisfaction3                  -2.691e-01  3.439e-01  -0.783 0.433904    
JobSatisfaction4                  -1.010e+00  3.608e-01  -2.799 0.005123 ** 
MaritalStatusMarried               3.056e-01  3.776e-01   0.809 0.418382    
MaritalStatusSingle                1.707e+00  4.749e-01   3.595 0.000324 ***
MonthlyIncome                     -2.793e-04  1.319e-04  -2.118 0.034140 *  
MonthlyRate                       -8.427e-06  1.698e-05  -0.496 0.619752    
NumCompaniesWorked                 1.612e-01  5.141e-02   3.136 0.001715 ** 
OverTimeYes                        1.938e+00  2.585e-01   7.497 6.54e-14 ***
PercentSalaryHike                  3.484e-02  5.289e-02   0.659 0.510013    
PerformanceRating4                -3.109e-03  5.305e-01  -0.006 0.995325    
RelationshipSatisfaction2         -4.694e-01  3.657e-01  -1.284 0.199302    
RelationshipSatisfaction3         -8.514e-01  3.456e-01  -2.463 0.013764 *  
RelationshipSatisfaction4         -9.909e-01  3.463e-01  -2.862 0.004216 ** 
StockOptionLevel                  -1.799e-01  2.340e-01  -0.769 0.442033    
TotalWorkingYears                  7.346e-04  4.007e-02   0.018 0.985372    
TrainingTimesLastYear             -1.636e-01  9.904e-02  -1.652 0.098496 .  
WorkLifeBalance                   -5.607e-01  1.664e-01  -3.369 0.000753 ***
YearsAtCompany                     1.636e-01  4.698e-02   3.482 0.000497 ***
YearsInCurrentRole                -2.751e-01  6.284e-02  -4.378 1.20e-05 ***
YearsSinceLastPromotion            3.020e-01  5.654e-02   5.341 9.22e-08 ***
YearsWithCurrManager              -2.293e-01  5.972e-02  -3.839 0.000124 ***
age_groupSenior                   -1.436e+01  1.158e+03  -0.012 0.990105    
age_groupYoung                     9.778e-01  3.885e-01   2.516 0.011854 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 896.03  on 1028  degrees of freedom
Residual deviance: 526.75  on  973  degrees of freedom
AIC: 638.75

Number of Fisher Scoring iterations: 15
```

```{r}
probs <- predict(glm.fit, test, type = "response")
predict <- rep("No", length(probs))
predict[probs > 0.5] <- "Yes"
table(predict, test$Attrition)
```

```
predict  No Yes
    No  345  34
    Yes  21  41
```

```{r}
mean(predict != test$Attrition)
```

```
[1] 0.1247166
```

Based on the regression analysis results, many of the variables in the dataset do not have a significant association with Attrition, as indicated by their p-values being greater than 0.05. However, certain variables, such as BusinessTravel, DistanceFromHome, JobInvolvement, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, NumCompaniesWorked, OverTime, RelationshipSatisfaction, WorkLifeBalance, and YearsAtCompany, YearsWithCurrManager, YearsSinceLastPromotion, YearsInCurrentRole and age_group do have a significant association with Attrition, as their p-values are less than 0.05. Further model validation revealed a mean error rate of 12.5%. To identify the most important variables for the dataset, a subset selection method can be used.

Forward Selection

```{r}

fwd.set = regsubsets(Attrition~. , data=new.df,nvmax=10, method ="forward")
```

```{r}
summary(fwd.set)

```

```
Subset selection object
Call: regsubsets.formula(Attrition ~ ., data = new.df, nvmax = 10,
    method = "forward")
55 Variables  (and intercept)
```

Backward Selection

```
```{r}
bkd.set <- regsubsets(Attrition ~ ., data = new.df, nvmax = 10, method = "backward")
summary(bkd.set)

```

Subset selection object
Call: regsubsets.formula(Attrition ~ ., data = new.df, nvmax = 10,
    method = "backward")
55 Variables (and intercept)
```

```
```{r}
coef(fwd.set ,7)
```

        (Intercept) BusinessTravelTravel_Frequently   JobRoleLaboratory Technician   JobRoleSales Representative   MaritalStatusSingle   OverTimeYes
         1.13660322   0.13707323   0.11466193   0.19359260   0.15479514   0.19623690
    WorkLifeBalance   age_groupYoung
        -0.06062613   0.10955639

```{r}
coef(bkd.set ,7)
```

        (Intercept) BusinessTravelTravel_Frequently   JobLevel2   MaritalStatusSingle   MonthlyIncome   OverTimeYes
       1.263327e+00   1.381813e-01   -9.738518e-02   1.541716e-01   -1.069138e-05   1.902063e-01
    WorkLifeBalance   age_groupYoung
      -5.500275e-02   8.928423e-02
```

The results of forward selection and backward selection for the best models using one to six variables are very similar. However, there is a slight difference in the best models using seven variables when comparing forward and backward stepwise selection. As the difference in the number of variables in the output of each subset selection is not significant, I have chosen to use the training data for the selection process.

```
```{r}
fwd.set <- regsubsets(Attrition~. , data=train, nvmax=10, method ="forward")
bkd.set <- regsubsets(Attrition ~. , data=train, nvmax=10, method ="backward")

```

```{r}
coef(fwd.set ,7)
```

              (Intercept) BusinessTravelTravel_Frequently   JobLevel2   MaritalStatusSingle   MonthlyIncome   OverTimeYes
             1.095172e+00   1.538561e-01   -9.142839e-02   1.495537e-01   -1.291149e-05   1.611443e-01
    YearsSinceLastPromotion   age_groupYoung
             1.215869e-02   1.117427e-01

```{r}
coef(bkd.set ,7)
```

              (Intercept) BusinessTravelTravel_Frequently   MaritalStatusSingle   MonthlyIncome   OverTimeYes   YearsInCurrentRole
             1.086661e+00   1.508916e-01   1.509363e-01   -8.566287e-06   1.604221e-01   -1.782314e-02
    YearsSinceLastPromotion   age_groupYoung
             2.238839e-02   1.162627e-01
```

```{r}
par(mfrow=c(2,2))
plot(summary(fwd.set)$rss ,xlab="Number of Variables ",ylab="RSS", type="l")
plot(summary(fwd.set)$rsq ,xlab="Number of Variables ",ylab="Rsq", type="l")
plot(summary(fwd.set)$adjr2 ,xlab="Number of Variables ",ylab="adjr2", type="l")
plot(summary(fwd.set)$cp ,xlab="Number of Variables ",ylab="cp", type="l")
plot(summary(fwd.set)$bic ,xlab="Number of Variables ",ylab="bic", type="l")

```




```{r}
par(mfrow=c(2,2))
plot(summary(fwd.set)$rss ,xlab="Number of Variables ",ylab="RSS", type="l")
plot(summary(fwd.set)$rsq ,xlab="Number of Variables ",ylab="Rsq", type="l")
plot(summary(fwd.set)$adjr2 ,xlab="Number of Variables ",ylab="adjr2", type="l")
plot(summary(fwd.set)$cp ,xlab="Number of Variables ",ylab="cp", type="l")
plot(summary(fwd.set)$bic ,xlab="Number of Variables ",ylab="bic", type="l")
```



I will use the variables selected by both forward selection and backward selection to predict Model 2.

Model 2

````
```{r}
glm.fit2<-glm(Attrition~BusinessTravel + JobLevel+ MaritalStatus + MonthlyIncome + OverTime + YearsSinceLastPromotion + age_group, data =train, family="binomial")
summary(glm.fit2)

```


Call:
glm(formula = Attrition ~ BusinessTravel + JobLevel + MaritalStatus +
    MonthlyIncome + OverTime + YearsSinceLastPromotion + age_group,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9403  -0.5460  -0.3463  -0.1839   2.7712

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -3.161e+00  5.994e-01  -5.273 1.34e-07 ***
BusinessTravelTravel_Frequently  2.023e+00  4.856e-01   4.166 3.09e-05 ***
BusinessTravelTravel_Rarely   9.750e-01  4.654e-01   2.095 0.036178 *
JobLevel2                    -4.837e-01  3.224e-01  -1.500 0.133525
JobLevel3                     6.214e-01  6.843e-01   0.908 0.363846
JobLevel4                     7.545e-01  1.273e+00   0.593 0.553508
JobLevel5                     2.944e+00  1.641e+00   1.794 0.072819 .
MaritalStatusMarried          2.394e-01  3.107e-01   0.770 0.441102
MaritalStatusSingle           1.368e+00  3.010e-01   4.544 5.53e-06 ***
MonthlyIncome                -2.578e-04  9.383e-05  -2.748 0.005998 **
OverTimeYes                   1.292e+00  2.031e-01   6.361 2.01e-10 ***
YearsSinceLastPromotion       1.286e-01  3.162e-02   4.068 4.74e-05 ***
age_groupSenior              -1.163e+01  4.873e+02  -0.024 0.980953
age_groupYoung                7.502e-01  2.239e-01   3.350 0.000808 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 896.03  on 1028  degrees of freedom
Residual deviance: 685.51  on 1015  degrees of freedom
AIC: 713.51

Number of Fisher Scoring iterations: 13
````

The output suggests that employees who frequently travel and those with employees that their MaritalStatusis single are more likely to experience attrition, as there is a significant relationship between attrition and these variables, indicated by p-values less than 0.05. Additionally, there is a relationship between attrition and variables such as MonthlyIncome, OvertimeYes, YearsSinceLastPromotion and age_groupYoung.

````
```{r}
probs2 <- predict(glm.fit2, test, type = "response")
predict2 <- rep("No", length(probs2))
predict2[probs2 > 0.5] <- "Yes"
table(predict2, test$Attrition)

```


predict2  No Yes
     No  357  56
     Yes   9  19
````

````
```{r}
mean(predict2 != test$Attrition)
```

[1] 0.1473923
````

For model 2, the mean test error is 14.7%.

Model 3

````
```{r}
glm.fit3 <-glm(Attrition~BusinessTravel + MaritalStatus + MonthlyIncome + OverTime + YearsSinceLastPromotion + age_group + NumCompaniesWorked + OverTime, data =train, family="binomial")
summary(glm.fit3)
```
````

```
Call:
glm(formula = Attrition ~ BusinessTravel + MaritalStatus + MonthlyIncome +
    OverTime + YearsSinceLastPromotion + age_group + NumCompaniesWorked +
    OverTime, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9258  -0.5413  -0.3371  -0.1785   3.2316

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -4.177e+00  5.801e-01  -7.200 6.04e-13 ***
BusinessTravelTravel_Frequently  2.027e+00  4.804e-01   4.220 2.44e-05 ***
BusinessTravelTravel_Rarely      9.847e-01  4.596e-01   2.143 0.032131 *
MaritalStatusMarried        2.842e-01  3.117e-01   0.912 0.361852
MaritalStatusSingle         1.487e+00  3.033e-01   4.903 9.46e-07 ***
MonthlyIncome              -1.680e-04  3.441e-05  -4.882 1.05e-06 ***
OverTimeYes                 1.302e+00  2.015e-01   6.461 1.04e-10 ***
YearsSinceLastPromotion     1.492e-01  3.213e-02   4.643 3.43e-06 ***
age_groupSenior            -1.225e+01  4.595e+02  -0.027 0.978738
age_groupYoung              1.061e+00  2.265e-01   4.684 2.81e-06 ***
NumCompaniesWorked          1.300e-01  3.856e-02   3.372 0.000745 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 896.03  on 1028  degrees of freedom
Residual deviance: 691.50  on 1018  degrees of freedom
AIC: 713.5

Number of Fisher Scoring iterations: 13
```

````
```{r}
probs3 <- predict(glm.fit3, test, type = "response")
predict3 <- rep("No", length(probs3))
predict3[probs3 > 0.5] <- "Yes"
table(predict3, test$Attrition)
```
````

```
predict3  No Yes
     No  359  54
     Yes   7  21
```

````
```{r}
mean(predict3 != test$Attrition)
```
````

```
[1] 0.138322
```

In this analysis, almost all the predicting variables had a p-value less than 0.05, indicating a significant association with the response variable "Attrition". The mean test error rate was found to be 13.8%, and there was a decrease in the mean test error rate between Model 2 and Model 3. For the logistic regression model, I plan to use the variables from Model 2 for further classification, as this model had the highest mean test error and includes the variables suggested by both forward and backward subset selection methods. I will now proceed with the analysis using Linear Discriminant Analysis and Quadratic Discriminant Analysis.

Linear Discriminant Analysis

````
```{r}
lda.fit<-lda(Attrition~BusinessTravel + JobLevel+ MaritalStatus + MonthlyIncome + OverTime + YearsSinceLastPromotion + age_group, data =train, family="binomial")
lda.fit

```

Call:
lda(Attrition ~ BusinessTravel + JobLevel + MaritalStatus + MonthlyIncome +
    OverTime + YearsSinceLastPromotion + age_group, data = train,
    family = "binomial")

Prior probabilities of groups:
       No       Yes
0.8425656 0.1574344

Group means:
    BusinessTravelTravel_Frequently BusinessTravelTravel_Rarely JobLevel2 JobLevel3  JobLevel4  JobLevel5 MaritalStatusMarried MaritalStatusSingle MonthlyIncome OverTimeYes YearsSinceLastPromotion
No                        0.1614764                   0.7220300 0.3886967 0.1568627 0.08073818 0.04036909            0.4959631           0.2791234      6704.797   0.2399077                2.159170
Yes                       0.3765432                   0.5864198 0.2160494 0.1049383 0.01851852 0.02469136            0.3086420           0.5802469      4496.951   0.5000000                2.364198
    age_groupSenior age_groupYoung
No      0.003460208      0.2018454
Yes     0.000000000      0.4320988

Coefficients of linear discriminants:
                                          LD1
BusinessTravelTravel_Frequently  1.4198851741
BusinessTravelTravel_Rarely      0.4155137360
JobLevel2                       -0.5701426337
JobLevel3                        0.0021558010
JobLevel4                        0.2290991722
JobLevel5                        1.1397945165
MaritalStatusMarried             0.1338035090
MaritalStatusSingle              1.1174305663
MonthlyIncome                   -0.0001236882
OverTimeYes                      1.1004774568
YearsSinceLastPromotion          0.0871251239
age_groupSenior                 -0.3233959652
age_groupYoung                   0.7164320084
````

````
```{r}
pred.lda<-predict(lda.fit, test)
table(pred.lda$class, test$Attrition)

```

       No Yes
  No  353  55
  Yes  13  20

```{r}
mean(pred.lda$class != test$Attrition)

```

 [1] 0.154195
````

The mean test error for linear Discriminant Analysis is 15.4%.

Regression Analysis

For the Regression Analysis, I have selected the "MonthlyIncome" variable as the response variable. To carry out this analysis, I will be using Linear Regression, Ridge Regression, and Lasso Regression. I will also use both the validation set approach and the cross-validation approach to validate the models. To perform this analysis, I will be using the standardized dataset.

Before performing the regression analysis, I used forward and backward subset selection methods to identify the most important variables for predicting MonthlyIncome. This approach helps to reduce the number of variables used in the analysis, ensuring that the selected variables are the most relevant for predicting MonthlyIncome. By doing so, I can improve the accuracy and efficiency of the regression models.

````
```{r}
fwd.set2 = regsubsets(MonthlyIncome~. , data=new.df,nvmax=10, method ="forward")
bkd.set2 = regsubsets(MonthlyIncome~. , data=new.df,nvmax=10, method ="backward")
coef(fwd.set2 ,7)

```
             (Intercept)                JobLevel2                JobLevel3                JobLevel4                JobLevel5           JobRoleManager
             -0.74261014               0.53838411               1.29769623               2.07820130               2.63046781               0.72035919
JobRoleResearch Director        TotalWorkingYears
              0.72988369               0.05307624

```{r}
coef(bkd.set2 ,7)
````

````
              (Intercept)                 JobLevel2                 JobLevel3                 JobLevel4                 JobLevel5
               -0.7642192                 0.5560108                 1.3497695                 2.2023160                 2.7632060
           JobRoleManager  JobRoleResearch Director JobRoleSales Representative
                0.7120786                 0.7176898                -0.1085601
````

````
```{r}
split<- sample(x=nrow(new.df), size=.70*nrow(new.df))
train2 <-new.df[split,]
test2 <-new.df[-split,]
fwd.set2 = regsubsets(MonthlyIncome~. , data=train2,nvmax=10, method ="forward")
bkd.set2 = regsubsets(MonthlyIncome~. , data=train2,nvmax=10, method ="backward")
coef(fwd.set2 ,7)

```
             (Intercept)                JobLevel3                JobLevel4                JobLevel5           JobRoleManager JobRoleResearch Director
              -0.4843534                0.8100471                1.4184060                1.9374151                0.8788565                0.9346958
   JobRoleSales Executive        TotalWorkingYears
               0.3266593                0.1944928

```{r}
coef(bkd.set2 ,7)

```
              (Intercept)                 JobLevel2                 JobLevel3                 JobLevel4                 JobLevel5
               -0.7691242                 0.5586217                 1.3418100                 2.2029724                 2.7592819
           JobRoleManager  JobRoleResearch Director JobRoleSales Representative
                0.7131939                 0.7491490                -0.1198454
````

Multiple Linear Regression

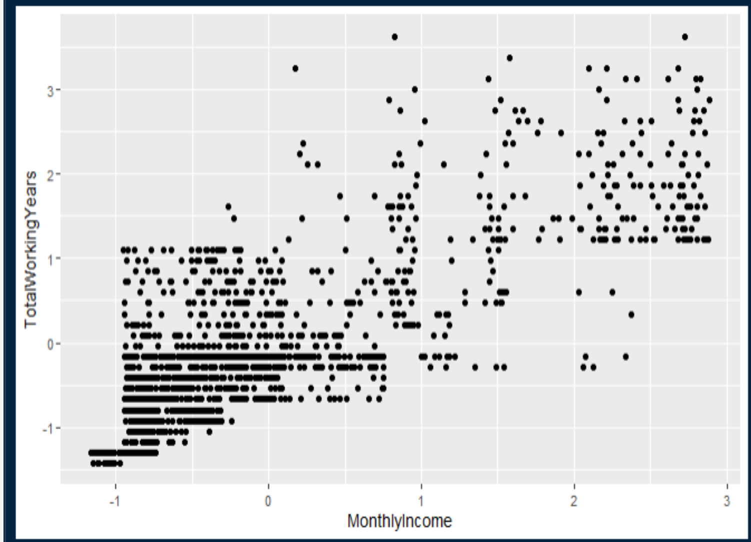
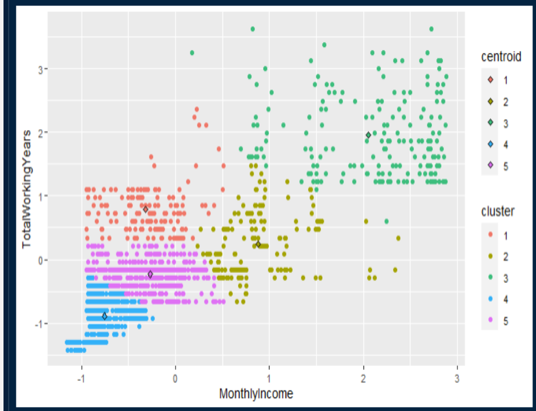
To test my prediction that Monthly Income of employees in the organization can be predicted by a mix of different variables, I used simple linear regression and hypothesis testing. The null hypothesis was that all the regression coefficients were equal to zero, while the alternative hypothesis stated that at least one regression coefficient was not equal to zero. The model equation was $Y= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots\dots\dots\dots \beta_n X_n + \varepsilon$.
````

```{r}
lm.fit<-lm(MonthlyIncome~., data=train2)
summary(lm.fit)

```

```
Call:
lm(formula = MonthlyIncome ~ ., data = train2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.66895 -0.14378 -0.01322  0.12790  0.99522 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -0.5088007  0.1289927  -3.944 8.57e-05 ***
Age                             -0.0185356  0.0126879  -1.461  0.14437    
AttritionYes                    -0.0455440  0.0239730  -1.900  0.05775 .  
BusinessTravelTravel_Frequently  0.0204738  0.0281630   0.727  0.46742    
BusinessTravelTravel_Rarely      0.0056640  0.0242347   0.234  0.81525    
DepartmentResearch & Development 0.2094121  0.1278874   1.637  0.10186    
DepartmentSales                  0.1457708  0.1308311   1.114  0.26547    
DistanceFromHome                -0.0035780  0.0073281  -0.488  0.62548    
Education2                      -0.0204961  0.0275447  -0.744  0.45699    
Education3                       0.0062596  0.0243718   0.257  0.79736    
Education4                       0.0106736  0.0264320   0.404  0.68644    
Education5                      -0.0132894  0.0449468  -0.296  0.76755    
EducationFieldLife Sciences     -0.1134235  0.0794634  -1.427  0.15379    
EducationFieldMarketing         -0.1124226  0.0843632  -1.333  0.18297    
EducationFieldMedical           -0.1207068  0.0796083  -1.516  0.12978    
EducationFieldOther             -0.1671349  0.0842402  -1.984  0.04753 *  
EducationFieldTechnical Degree  -0.0829342  0.0831837  -0.997  0.31901    
GenderMale                       0.0216974  0.0149548   1.451  0.14714    
JobInvolvement2                 -0.0528984  0.0351033  -1.507  0.13215    
JobInvolvement3                 -0.0799431  0.0335624  -2.382  0.01741 *  
JobInvolvement4                 -0.0673179  0.0394445  -1.707  0.08821 .  
JobLevel2                        0.3207410  0.0288717  11.109  < 2e-16 ***
JobLevel3                        1.0070561  0.0397098  25.360  < 2e-16 ***
JobLevel4                        1.7784341  0.0603563  29.466  < 2e-16 ***
JobLevel5                        2.3209563  0.0715079  32.457  < 2e-16 ***
JobRoleHuman Resources          -0.0101089  0.1258332  -0.080  0.93599    
JobRoleLaboratory Technician    -0.2611914  0.0363608  -7.183 1.35e-12 ***
JobRoleManager                   0.7469152  0.0563512  13.255  < 2e-16 ***
JobRoleManufacturing Director   -0.0143636  0.0326730  -0.440  0.66031    
JobRoleResearch Director         0.7522769  0.0469257  16.031  < 2e-16 ***
JobRoleResearch Scientist       -0.2561059  0.0369466  -6.932 7.56e-12 ***
JobRoleSales Executive           0.0717982  0.0690117   1.040  0.29842    
JobRoleSales Representative     -0.2501388  0.0786455  -3.181  0.00152 ** 
JobSatisfaction2                -0.0006363  0.0241429  -0.026  0.97898    
JobSatisfaction3                -0.0098921  0.0215441  -0.459  0.64623    
JobSatisfaction4                 0.0046487  0.0211560   0.220  0.82612    
MaritalStatusMarried             0.0099824  0.0196628   0.508  0.61179    
MaritalStatusSingle             -0.0208384  0.0272116  -0.766  0.44398    
MonthlyRate                      0.0055740  0.0072664   0.767  0.44321    
NumCompaniesWorked               0.0210314  0.0083794   2.510  0.01224 *  
OverTimeYes                      0.0249545  0.0170415   1.464  0.14342    
PercentSalaryHike                0.0086448  0.0116185   0.744  0.45702    
PerformanceRating4              -0.0276322  0.0317577  -0.870  0.38446    
RelationshipSatisfaction2        0.0219241  0.0230831   0.950  0.34246    
RelationshipSatisfaction3        0.0106970  0.0213497   0.501  0.61646    
RelationshipSatisfaction4        0.0126715  0.0215003   0.589  0.55575    
StockOptionLevel                -0.0122453  0.0099922  -1.225  0.22069    
TotalWorkingYears                0.0667267  0.0168189   3.967 7.80e-05 ***
TrainingTimesLastYear           -0.0048177  0.0073040  -0.660  0.50966    
WorkLifeBalance                 -0.0051221  0.0074932  -0.684  0.49441    
YearsAtCompany                   0.0112501  0.0162505   0.692  0.48892    
YearsInCurrentRole              -0.0055822  0.0118882  -0.470  0.63877    
YearsSinceLastPromotion         -0.0019559  0.0095050  -0.206  0.83701    
YearsWithCurrManager            -0.0017079  0.0123886  -0.138  0.89038    
age_groupSenior                  0.1581342  0.1377812   1.148  0.25137    
age_groupYoung                  -0.0448914  0.0252125  -1.781  0.07530 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2287 on 973 degrees of freedom
Multiple R-squared:  0.9474,	Adjusted R-squared:  0.9444 
F-statistic: 318.7 on 55 and 973 DF,  p-value: < 2.2e-16
```

The multiple linear regression output shows that variables such as Joblevel, Job Role, TotalWorking, and NumCompaniesWorked have a p-value less than 0.05, indicating their significant relationship with predicting MonthlyIncome. Thus, the null hypothesis is rejected. The F-statistic greater than 1 suggests at least one predictor is related to the response variable. The Adjusted R2 value of 0.9444 implies a good model fit, but overfitting is a concern. Therefore, significant variables will be selected and used for the next model to avoid overfitting.

````
```{r}
pred.lm <- predict(lm.fit, test2)
mean((pred.lm - test2$MonthlyIncome)^2)

```

 [1] 0.04454692
````

After validating the model, I found that the least square error for the model is 0.0445, indicating a good fit.

````
```{r}
lm.fit3<-lm(MonthlyIncome~TotalWorkingYears +YearsSinceLastPromotion+ JobRole +JobLevel , data=train2)
summary(lm.fit3)

```

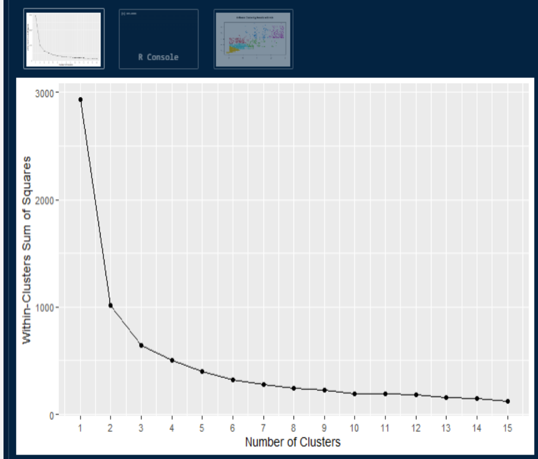

Call:
lm(formula = MonthlyIncome ~ TotalWorkingYears + YearsSinceLastPromotion +
    JobRole + JobLevel, data = train2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66371 -0.14170 -0.01339  0.13614  0.92593

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.469001   0.036720 -12.772  < 2e-16 ***
TotalWorkingYears            0.076720   0.012674   6.054 1.99e-09 ***
YearsSinceLastPromotion     -0.002492   0.007884  -0.316 0.751997
JobRoleHuman Resources      -0.168298   0.048590  -3.464 0.000555 ***
JobRoleLaboratory Technician -0.264976   0.035401  -7.485 1.55e-13 ***
JobRoleManager               0.715806   0.049328  14.511  < 2e-16 ***
JobRoleManufacturing Director -0.017750   0.032167  -0.552 0.581211
JobRoleResearch Director     0.756097   0.045920  16.465  < 2e-16 ***
JobRoleResearch Scientist   -0.252582   0.036207  -6.976 5.47e-12 ***
JobRoleSales Executive       0.005158   0.028404   0.182 0.855947
JobRoleSales Representative -0.326474   0.046676  -6.994 4.83e-12 ***
JobLevel2                    0.328970   0.027536  11.947  < 2e-16 ***
JobLevel3                    1.014835   0.038402  26.427  < 2e-16 ***
JobLevel4                    1.767829   0.059178  29.873  < 2e-16 ***
JobLevel5                    2.313127   0.069745  33.165  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.229 on 1014 degrees of freedom
Multiple R-squared:  0.9451,	Adjusted R-squared:  0.9443
F-statistic:  1246 on 14 and 1014 DF,  p-value: < 2.2e-16
````

In the second multiple linear regression, the p-values of Joblevel, Job Role and Total Working Years are less than 0.05, indicating a significant relationship with MonthlyIncome. We reject the null hypothesis and conclude that these variables are significant predictors of MonthlyIncome, except for the YearsSinceLastPromotion, Job Role Manufacturing Director and Sales Executives, which show no significant relationship with MonthlyIncome. The F-statistic for the second model is greater than the first, indicating at least one predictor related to the response variable. The adjusted R2 value is 0.9443, indicating a good model fit. However, to avoid overfitting, we will remove the variables showing significant relationship with the response variable for the next model. The second linear regression model is similar to the first one, with a larger F-statistic. The model has a 94.43% less variance of errors compared to the variance of the response variable and standard deviation.

```{r}
pred.lm3 <- predict(lm.fit3, test2)
mean((pred.lm3 - test2$MonthlyIncome)^2)
```

```
[1] 0.04251559
```

Validating the second linear regression model showed a 0.129 least square error, which is lower than the first linear model. Therefore, the second linear model is preferred for predicting Monthly Income using multiple regression analysis.

Ridge Regression

To further the analysis, I applied ridge and lasso regression techniques using the same variables from the second linear model for regularization.

```{r}
train.mat<- model.matrix(MonthlyIncome~ TotalWorkingYears +YearsSinceLastPromotion+ JobRole +JobLevel , data = train2)
test.mat <- model.matrix(MonthlyIncome~ TotalWorkingYears +YearsSinceLastPromotion+ JobRole +JobLevel, data = test2)
grid <- 10 ^ seq(4, -2, length = 100)
fit.ridge <- glmnet(train.mat, train2$MonthlyIncome, alpha = 0, lambda = grid, thresh = 1e-12)
cv.ridge <- cv.glmnet(train.mat, train2$MonthlyIncome, alpha = 0, lambda = grid, thresh = 1e-12)
bestlam.ridge <- cv.ridge$lambda.min
bestlam.ridge
pred.ridge <- predict(fit.ridge, s = bestlam.ridge, newx = test.mat)
mean((pred.ridge - test2$MonthlyIncome)^2)
```

```
[1] 0.01
[1] 0.04326964
```

The ridge regression analysis using the variables from the second linear model resulted in an MSE of 0.043, which is comparable to the least squares 2nd linear model. Therefore, there is little or no significant difference between the two models.

Lasso Regression

```{r}
fit.lasso <- glmnet(train.mat, train2$MonthlyIncome, alpha = 1, lambda = grid, thresh = 1e-12)
cv.lasso <- cv.glmnet(train.mat, train2$MonthlyIncome, alpha = 1, lambda = grid, thresh = 1e-12)
bestlam.lasso <- cv.lasso$lambda.min
bestlam.lasso
pred.lasso <- predict(fit.lasso, s = bestlam.lasso, newx = test.mat)
mean((pred.lasso - test2$MonthlyIncome)^2)
```

```
[1] 0.01
[1] 0.04388959
```

The mean squared error (MSE) for the lasso regression was found to be 0.043, which is similar to the MSE values of the least squares 2nd linear model and the ridge regression. To evaluate the accuracy of the models in predicting Monthly Income, we calculated the R2 values for all the regression techniques used.

```{r}
test.avg <- mean(test2$MonthlyIncome)
lm.r2 <- 1 - mean((pred.lm - test2$MonthlyIncome)^2) / mean((test.avg - test2$MonthlyIncome)^2)
ridge.r2 <- 1 - mean((pred.ridge - test2$MonthlyIncome)^2) / mean((test.avg - test2$MonthlyIncome)^2)
lasso.r2 <- 1 - mean((pred.lasso - test2$MonthlyIncome)^2) / mean((test.avg - test2$MonthlyIncome)^2)
lm.r2
ridge.r2
lasso.r2
```

```
[1] 0.9607836
[1] 0.9619081
[1] 0.9613623
```

The test R2 for the three regression models were calculated and compared. The results show that the test R2 for least squares is 0.9607836, for ridge regression is 0.9619081, and for lasso regression is 0.9613623. The similarity in their values suggests that the three models have almost the same accuracy in predicting Monthly Income. Therefore, all three models can be considered to have a high accuracy in predicting Monthly Income.

K-Means Clustering

```{r}
my_data<-new.df[,c(14,22)]
my_data <- as.data.frame(my_data)
ggplot(data=my_data, mapping=aes(x=MonthlyIncome, y=TotalWorkingYears)) + geom_point()
```



````
```{r}
set.seed(1)
fit <- kmeans(x=my_data, centers=5)
my_data$cluster <- factor(fit$cluster)
ggplot() + geom_point(data=my_data, mapping=aes(x=MonthlyIncome, y=TotalWorkingYears, color=cluster)) + geom_point(data=data.frame(fit$centers, centroid=as.factor(1:nrow(fit$centers))),
mapping=aes(x=MonthlyIncome, y=TotalWorkingYears, fill=centroid), shape=23, color="black")
my_data1 <- as.data.frame(scale(my_data[,1:2]))
wss <- (nrow(my_data1)-1) * sum(apply(X=my_data1, MARGIN=2, FUN=var))

```
````




````
```{r}
set.seed(1)
for (i in 2:15) wss[i] <- sum(kmeans(my_data1, centers=i)$withinss)
wss.df <- data.frame(cluster = 1:15, wss = wss)
ggplot(data=wss.df, mapping=aes(x=cluster, y=wss)) + geom_point() + geom_line() + scale_x_continuous(breaks = seq(from=1, to=15, by=1)) + labs(x="Number of Clusters", y="Within-Clusters Sum
of Squares")
km.out=kmeans (my_data1,6, nstart =20)
km.out$tot.withinss
plot(my_data1, col=(km.out$cluster +1), main="K-Means Clustering Results with K=6", xlab="", ylab="", pch=20, cex=1)

```
````



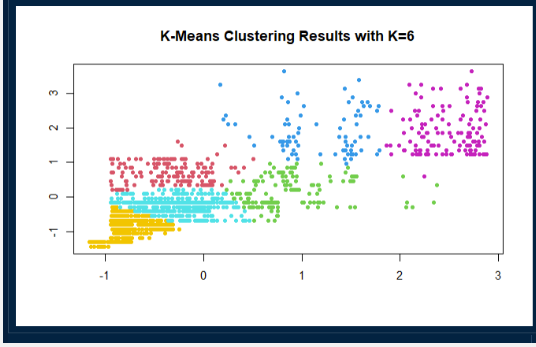
```{r}
set.seed(1)
for (i in 2:15) wss[i] <- sum(kmeans(my_data1, centers=i)$withinss)
wss.df <- data.frame(cluster = 1:15, wss = wss)
ggplot(data=wss.df, mapping=aes(x=cluster, y=wss)) + geom_point() + geom_line() + scale_x_continuous(breaks = seq(from=1, to=15, by=1)) + labs(x="Number of Clusters", y="Within-Clusters Sum of Squares")
km.out=kmeans (my_data1,6, nstart =20)
km.out$tot.withinss
plot(my_data1, col=(km.out$cluster +1), main="K-Means Clustering Results with K=6", xlab="", ylab="", pch=20, cex=1)
```

R Console

```
[1] 325.8749
```

```{r}
set.seed(1)
for (i in 2:15) wss[i] <- sum(kmeans(my_data1, centers=i)$withinss)
wss.df <- data.frame(cluster = 1:15, wss = wss)
ggplot(data=wss.df, mapping=aes(x=cluster, y=wss)) + geom_point() + geom_line() + scale_x_continuous(breaks = seq(from=1, to=15, by=1)) + labs(x="Number of Clusters", y="Within-Clusters Sum of Squares")
km.out=kmeans (my_data1,6, nstart =20)
km.out$tot.withinss
plot(my_data1, col=(km.out$cluster +1), main="K-Means Clustering Results with K=6", xlab="", ylab="", pch=20, cex=1)
```

R Console



SUMMARY

The HR Analytics Employee Attrition dataset is composed of 35 variables that detail the reasons why employees leave an organization. To prepare the dataset for analysis, redundant variables were removed and some variable data types were changed, resulting in a new dataset with 29 variables. Exploratory analysis revealed an imbalanced distribution, which could create classification problems during analysis. To address this, the ovun.sample () function was used to balance the data.

Prior to classification analysis, the dataset was split into training and test sets for validation purposes. Three logistic regression models were created with varying mean test error rates, with Model 2 having the lowest error rate of 14.7%. Model 2 was chosen as the best model due to potential overfitting with Model 1. Forward subset selection and backward subset selection were used to determine statistically significant variables in predicting attrition in the organization, with BusinessTravel, JobLevel, JobRole, JobSatisfaction, NumCompaniesWorked, OverTime, and TotalWorkingYears being identified as significant variables.

Subsequent classification analysis was performed using LDA and QDA with the variables used in Model 2, resulting in higher mean test error rates of 26.1% and 31.1%, respectively. Comparison of the three models showed that Model 2 remained the best choice, with mean test error rate computed using the validation set approach.

To make useful future predictions, Monthly Income was set as the response variable and the dataset was standardized prior to regression analysis. Multiple regression, Ridge regression, and Lasso Regression were used to predict Monthly Income, with TotalWorkingYears, YearsSinceLastPromotion, JobRole, and JobLevel identified as statistically significant variables. The 2nd Linear Model had the least squares error of 0.129, which was slightly lower than Ridge Regression and Lasso Regression. The test R2 for the 2nd Linear model was 0.8720748, while the test R2 for Ridge Regression and Lasso Regression were 0.875482 and 0.873913, respectively, indicating little difference in accuracy among the models.

Finally, KNN nearest neighbor and K-means clustering were computed using the standardized dataset, with k=6 identified as the best value. A graph was plotted to show the number of clusters based on Monthly Income and TotalWorkingYears of Employees.