Monash University: Assessment Cover Sheet

Student name	Nguyen	Trung Kien	
School/Campus		**Student's I.D. number**	29057957
Unit name	FIT5141 Advanced topics in information technology S2 2019		
Lecturer's name		**Tutor's name**	Seyedali Meghdadi
Assignment name	FIT5141 Assignment 3 - Report	**Group Assignment: No** **Note, each student must attach a coversheet**	

Lab/Tute Class:	**Lab/Tute Time: 8 am Tuesday**	**Word Count:**
Due date: **27-10-2019**	**Submit Date: 27-10-2019**	**Extension granted □**

If an extension of work is granted, specify date and provide the signature of the lecturer/tutor. Alternatively, attach an email printout or handwritten and signed notice from your lecturer/tutor verifying an extension has been granted.

Extension granted until (date): ..../......./............ Signature of lecturer/tutor: ....

Late submissions policy	**Days late**	**Penalty applied**
Penalties apply to late submissions and may vary between faculties. Please refer to your faculty's late assessment policy for details.		

Patient/client confidentiality: Where a patient/client case study is undertaken a signed Consent Form must be obtained.

Intentional plagiarism or collusion amounts to cheating under Part 7 of the Monash University (Council) Regulations

Plagiarism: Plagiarism means to take and use another person's ideas and or manner of expressing them and to pass these off as one's own by failing to give appropriate acknowledgement. This includes material from any source, staff, students or the Internet - published and unpublished works.

Collusion: Collusion means unauthorised collaboration on assessable written, oral or practical work with another person. Where there are reasonable grounds for believing that intentional plagiarism or collusion has occurred, this will be reported to the Associate Dean (Education) or nominee, who may disallow the work concerned by prohibiting assessment or refer the matter to the Faculty Discipline Panel for a hearing.

Student Statement:

- I have read the university's Student Academic Integrity Policy and Procedures
- I understand the consequences of engaging in plagiarism and collusion as described in Part 7 of the Monash University (Council) Regulations (academic misconduct).
- I have taken proper care to safeguard this work and made all reasonable efforts to ensure it could not be copied. No part of this assignment has been previously submitted as part of another unit/course.
- I acknowledge and agree that the assessor of this assignment may, for the purposes of assessment, reproduce the assignment and:

  i. provide it to another member of faculty and any external marker; and/or

  ii. submit to a text matching/originality checking software; and/or

  iii. submit it to a text matching/originality checking software which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking.

  I certify that I have not plagiarised the work of others or participated in unauthorised collaboration or otherwise breached the academic integrity requirements in the Student Academic Integrity Policy.

Date: ....27.../....10.../.....2019........ Signature: Trung Kien Nguyen *

Privacy Statement:

For information about how the University deals with your personal information go to http://privacy.monash.edu.au/guidelines/collection-personal-information.html#enrol

FIT5141 – Advanced Topics in Information Technology

Assignment 3: Attrition in Organization

The Story behind the Employees Quit the Company

Why are Employees Leaving their Jobs?

MONASH University

Student name: Trung Kien Nguyen

Student ID: 29057957

Content

List of figures

List of tables

1. Introduction

Employees are the backbone of any organization and they are one of the many keys to success in business. Hiring a talent worker is always challenging but keep them stay in the company is a difficult task. It is always a significant loss for a company when they invest time and resources in an employee who then leaves prematurely (Katsikea, 2015). The company does not lose good employees but they also lose the working relationship, contacts with customers, pieces of knowledge and experiences that employees have developed or even the energy dedication that the employees have brought to the jobs (Pestronk, 2014), as the consequence, it puts the company in challenging situations and have negative impact to the operational efficiency.

There are so many reasons that cause the turnover, in many cases, it is the working environment, relationship satisfaction or even job satisfaction rather than low pay that prompts an employee to leave. Thus, this report is conducted to analyze the number of different factors which have an impact on the attrition of employee from basic information such as gender of employees, their average salary, department, the job role to the qualitative factors such as working environment, relationship satisfaction or job satisfaction, etc. The dataset is taken from Human Resource Department of IBM. The main purpose of the report is to find the answer to the question of what is the main factors that have the most impact on the employees' turnover.

2. General Information of Dataset

2.1. Data Wrangling

The dataset contains the answers from 1470 respondents working in different departments with different job roles. Datasets will be checked and validated in terms of its format and values by using R and a variety of libraries. Following is the summary of the datasets and several steps of data wrangling.

- Transform the **chr** data format to **factor** format.
- Change the following attribute: Education, JobInvolvement, StockOptionLevel, WorkLifeBalance and JobLevel from **integer** format to **factor** format.
- Remove following unnecessary attributes: EmployeeCount, Over18, StandardHours.

For further analysis, the following variables will be added into the dataset. Those new attributes are made based on the current attributes

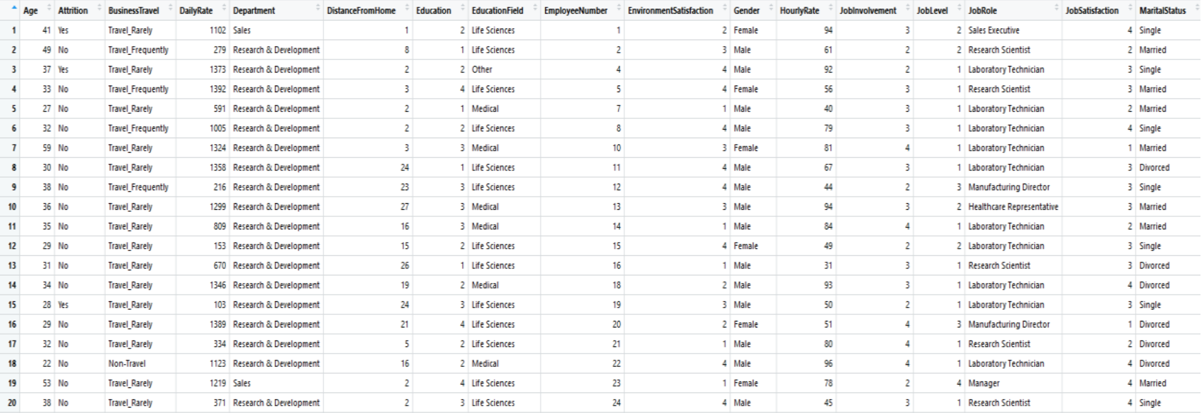
Attribute	**Current Value**	**Format**	**New Attribute**	**Format**	**New Value**
Age	18, 19, 20, 21….	Int	Generation	Factor	Millennial (18-30) Experience (30-40) Settle (40-50) Old (50-60)

Education	1, 2, 3, 4, 5	Int	Educational_Levels	Factor	Without College D. College D. Bachelors D. Master D. Ph.D. D.
YearsWith CurrManaer	0, 2, 3, 5, 6, 8…	Int	CatYearManager	Factor	Recently Hired 2-4 Years Hired Long Established Manager

Table 1: Feature transformation

2.2. Data Description

The overview of the final dataset.



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus
1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	2	Female	94	3	2	Sales Executive	4	Single
2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	2	3	Male	61	2	2	Research Scientist	2	Married
3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	4	4	Male	92	2	1	Laboratory Technician	3	Single
4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	5	4	Female	56	3	1	Research Scientist	3	Married
5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	7	1	Male	40	3	1	Laboratory Technician	2	Married
6	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	8	4	Male	79	3	1	Laboratory Technician	4	Single
7	59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	10	3	Female	81	4	1	Laboratory Technician	1	Married
8	30	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	11	4	Male	67	3	1	Laboratory Technician	3	Divorced
9	38	No	Travel_Frequently	216	Research & Development	23	3	Life Sciences	12	4	Male	44	2	3	Manufacturing Director	3	Single
10	36	No	Travel_Rarely	1299	Research & Development	27	3	Medical	13	3	Male	94	3	2	Healthcare Representative	3	Married
11	35	No	Travel_Rarely	809	Research & Development	16	3	Medical	14	1	Male	84	4	1	Laboratory Technician	2	Married
12	29	No	Travel_Rarely	153	Research & Development	15	2	Life Sciences	15	4	Female	49	2	2	Laboratory Technician	3	Single
13	31	No	Travel_Rarely	670	Research & Development	26	1	Life Sciences	16	1	Male	31	3	1	Research Scientist	3	Divorced
14	34	No	Travel_Rarely	1346	Research & Development	19	2	Medical	18	2	Male	93	3	1	Laboratory Technician	4	Divorced
15	28	Yes	Travel_Rarely	103	Research & Development	24	3	Life Sciences	19	3	Male	50	2	1	Laboratory Technician	3	Single
16	29	No	Travel_Rarely	1389	Research & Development	21	4	Life Sciences	20	2	Female	51	4	3	Manufacturing Director	1	Divorced
17	32	No	Travel_Rarely	334	Research & Development	5	2	Life Sciences	21	1	Male	80	4	1	Research Scientist	2	Divorced
18	22	No	Non-Travel	1123	Research & Development	16	2	Medical	22	4	Male	96	4	1	Laboratory Technician	4	Divorced
19	53	No	Travel_Rarely	1219	Sales	2	4	Life Sciences	23	1	Female	78	2	4	Manager	4	Married
20	38	No	Travel_Rarely	371	Research & Development	2	3	Life Sciences	24	4	Male	45	3	1	Research Scientist	4	Single

Figure 1: Table overview of the dataset with some first features produced by R.

Data Description

Feature	Format	Example	Description
Age	int	48, 49, 50, 18, 20, 22…	Self-description
Attrition	Factor	Yes, No, Yes, No	Attrition status
BusinessTravel	Factor	Travel_Rarely, Non-Travel Travel_Frequently,	Self-description
DailyRate	int	1102, 279, 1373….	Self-description
Department	Factor	Sales, Research …	Self-description
DistanceFromHome	int	1, 8, 2, 3 ….	Self-description
Education	Factor	1, 2, 3, 4, 5	Without College D College D, Bachelor D Master D, Doctor D
EducationField	Factor	Life Sciences, Medical…	Self-description
EmployeeNumber	int	1,2,4,7,6,8	Self-description
EnvironmentSatisfaction	int	1,2,3,4	Low, Medium High, Very high
Gender	Factor	Female, Male	Self-description
HourlyRate	int	94, 61, 92…	Self-description

JobInvolvement	Factor	4, 2, 1, 3	4- Very High, 2-Medium 3- High, 1- Low
JobLevel	Factor	1, 2	Self-description
JobRole	Factor	Sales Executive, Research…	Self-description
JobSatisfaction	int	3, 4, 2, 1	3- High, 4- Very High, 2- Medium, 1- Low
MaritalStatus	int	Single, Married, Divorced	Self-description
MonthlyIncome	int	8000, 10000	Self-description
MonthlyRate	int	19479, 24970….	Self-description
NumCompaniesWorked	int	1, 2, 8, 10, …	Self-description
PercentSalaryHike	int	11, 23, 15…	Self-description
PerformanceRating	int	2, 4, 3, 1	2- Good 4- Outstanding 3- Excellent 4- Bad
RelationshipSatisfaction	int	3, 2, 1, 4	3- High, 2- Medium 1- Low, 4- Very High
StockOptionLevel	Factor	3, 2, 1, 0	Self-description
TotalWorkingYears	int	10, 8 ,9…	Self-description
TrainingTimesLastYear	int	0, 2, 3, 6…	Self-description
WorkLifeBalance	Factor	1, 4, 3, 2	1- Bad, 4- Best 3- Better, 2- Good
YearsAtCompany	int	6, 10, 0 ,8 ,2 …	Self-description
YearsInCurrentRole	int	1,2,3,4…	Self-description
YearsSinceLastPromotion	int	0, 1, 2, 3, 4 ….	Self-description
YearsWithCurrManager	int	5, 7, 10…	Self-description
Generation	Factor	Millennial (18-30) Experience (30-40)	Self-description
Educational_Levels	Factor	Bachelors D, Master D..	Self-description
CatYearManager	Factor	Recently Hired, 2-4 Years Hired	Self-description

Table 2: Dataset Overview

The overview the general descriptions of the dataset after wrangling which is used for this report.

Datasets Name	**IBM_HR_Analytics.**
Number of rows	**1470**
Number of attributes	**35**
Format of values	**int, factor**
Number of **numeric** attributes	**15**
Number of **factor** attributes	**20**
Missing data	**No**
Label	**Attrition (Yes, No)**
Imbalanced datasets	**1237 (84% of cases) say No** **237 (16% of cases) say Yes**

Table 3: Data Summary

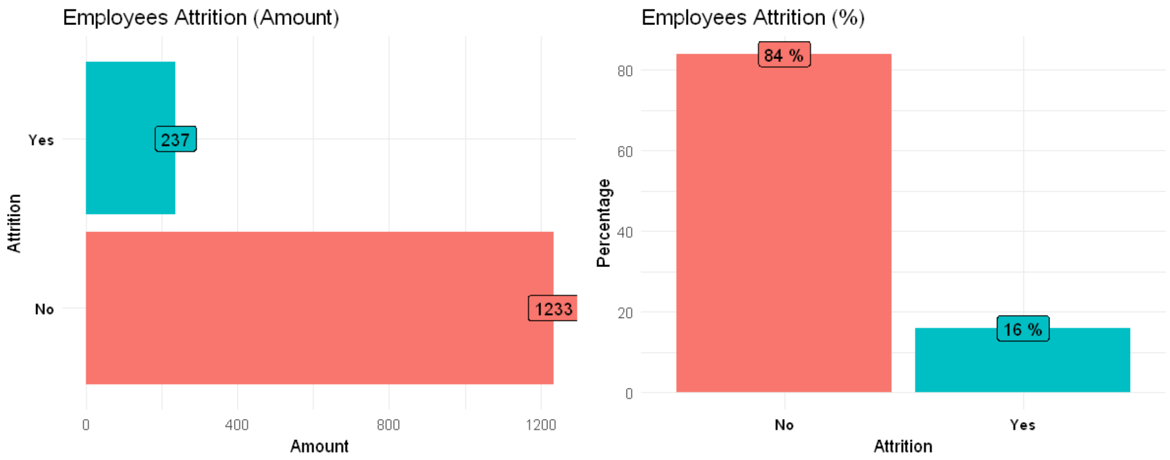
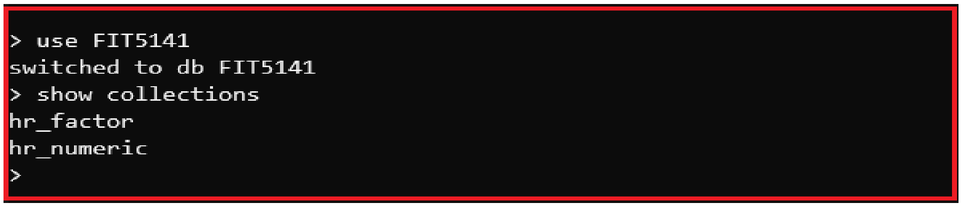


Figure 2: Employees Attrition

2.3. MongoDB Preparation

In this section, the datasets will be broken down into collections in order to import it to the MongoDB database. The first collection includes all the numeric attributes of the datasets while the second collection contains the factor attributes of the datasets. The datasets, namely "**FIT5141**" and the two collections for numeric and factor attributes named **hr_numeric** and **hr_factor**, respectively.

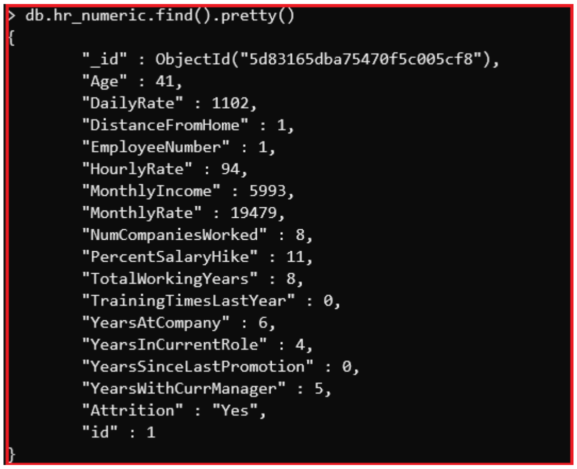
Before inserting data into collections of the database, several steps need to take in order for convenience purposes. Firstly, two data frames which contain the numeric attributes and factor attributes need to have the same id for each row. The benefit is for joining two collections to analyze data later. Secondly, the "Attrition" column which lives in the factor table will be replicated to the numeric data frame. The purpose is to reduce the joining step so improving the performance.



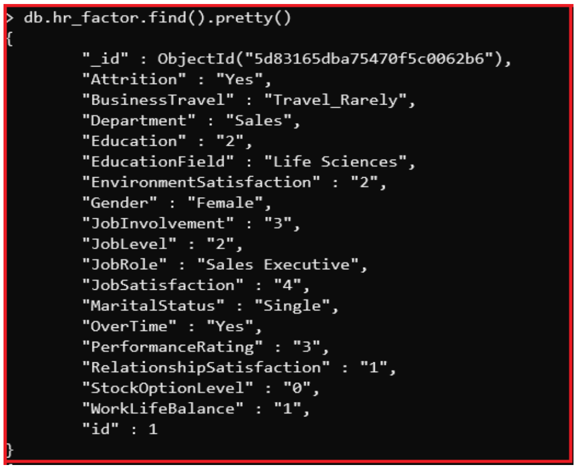
```
> use FIT5141
switched to db FIT5141
> show collections
hr_factor
hr_numeric
>
```

Figure 3: MongoDB database collections

The analysis that has done in this report is also produced by Tableau. The detail of importing data into Tableau is provided under Appendix section.



```
> db.hr_numeric.find().pretty()
{
        "_id" : ObjectId("5d83165dba75470f5c005cf8"),
        "Age" : 41,
        "DailyRate" : 1102,
        "DistanceFromHome" : 1,
        "EmployeeNumber" : 1,
        "HourlyRate" : 94,
        "MonthlyIncome" : 5993,
        "MonthlyRate" : 19479,
        "NumCompaniesWorked" : 8,
        "PercentSalaryHike" : 11,
        "TotalWorkingYears" : 8,
        "TrainingTimesLastYear" : 0,
        "YearsAtCompany" : 6,
        "YearsInCurrentRole" : 4,
        "YearsSinceLastPromotion" : 0,
        "YearsWithCurrManager" : 5,
        "Attrition" : "Yes",
        "id" : 1
}
```



```
> db.hr_factor.find().pretty()
{
        "_id" : ObjectId("5d83165dba75470f5c0062b6"),
        "Attrition" : "Yes",
        "BusinessTravel" : "Travel_Rarely",
        "Department" : "Sales",
        "Education" : "2",
        "EducationField" : "Life Sciences",
        "EnvironmentSatisfaction" : "2",
        "Gender" : "Female",
        "JobInvolvement" : "3",
        "JobLevel" : "2",
        "JobRole" : "Sales Executive",
        "JobSatisfaction" : "4",
        "MaritalStatus" : "Single",
        "OverTime" : "Yes",
        "PerformanceRating" : "3",
        "RelationshipSatisfaction" : "1",
        "StockOptionLevel" : "0",
        "WorkLifeBalance" : "1",
        "id" : 1
}
```

Figure 4: First and second collection of the database

3. Gender Analysis

This section will look into the gender perspective to see if there are any discrepancies between make and females in the organization as well as other basic information include the average monthly salary, age, and job satisfaction by gender.

The following are the questions need to find answers in this part.

- What is the distribution of gender in the dataset and the age distribution between gender among respondents? Is there any substantial disparity?
- What is the distribution of job satisfaction between employees who stay and leave? Is any type of gender more unsatisfied or dissatisfied?
- What is income distribution by gender? Is there any difference distribution between gender in each department?

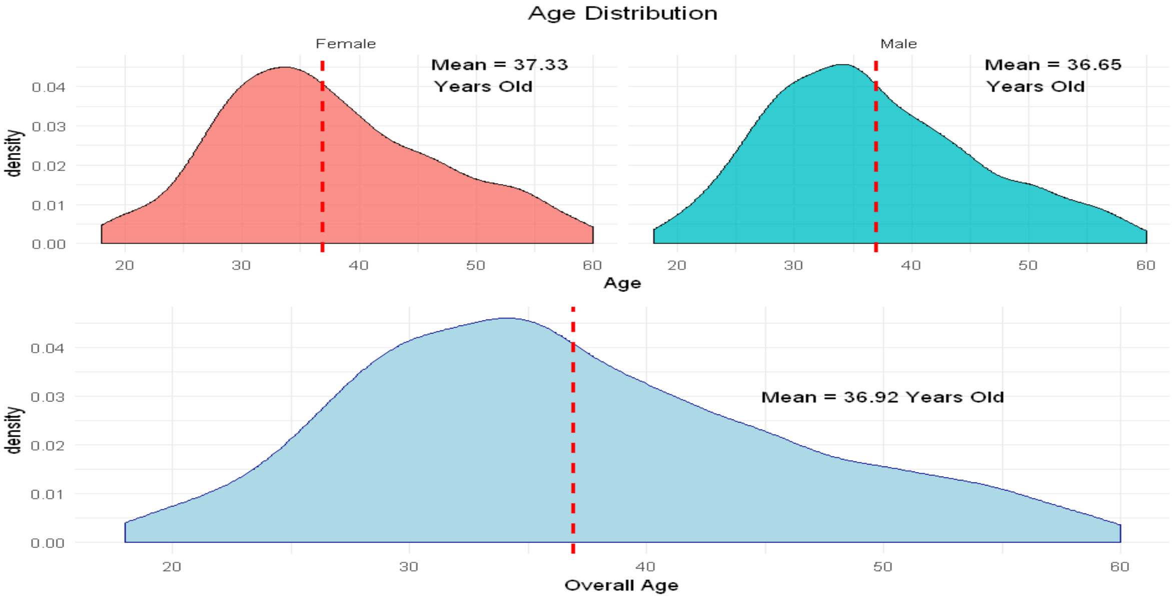


Figure 5: Age Distribution

There were 588 female and 882 male respondents in the survey with similar average age which is 37.33 and 36.65 for females and males respectively, so there are no significant discrepancies in terms of age in the dataset.

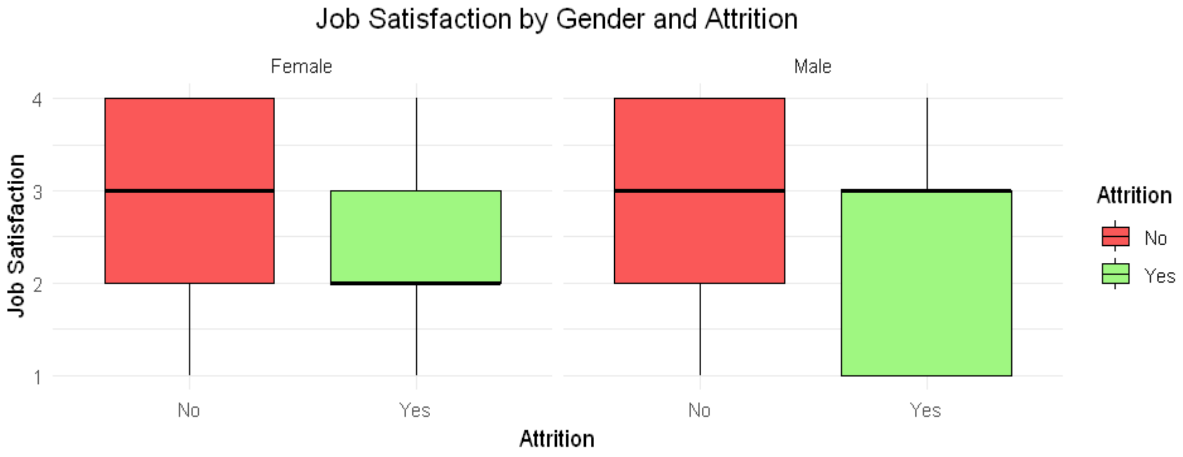


Figure 6: Job Satisfaction by Gender and Attrition

It seems like the male employees who want to quit the organization have lower job satisfaction, also, it is obvious that people no matter male or female will stay in the company if they feel satisfied with their job.

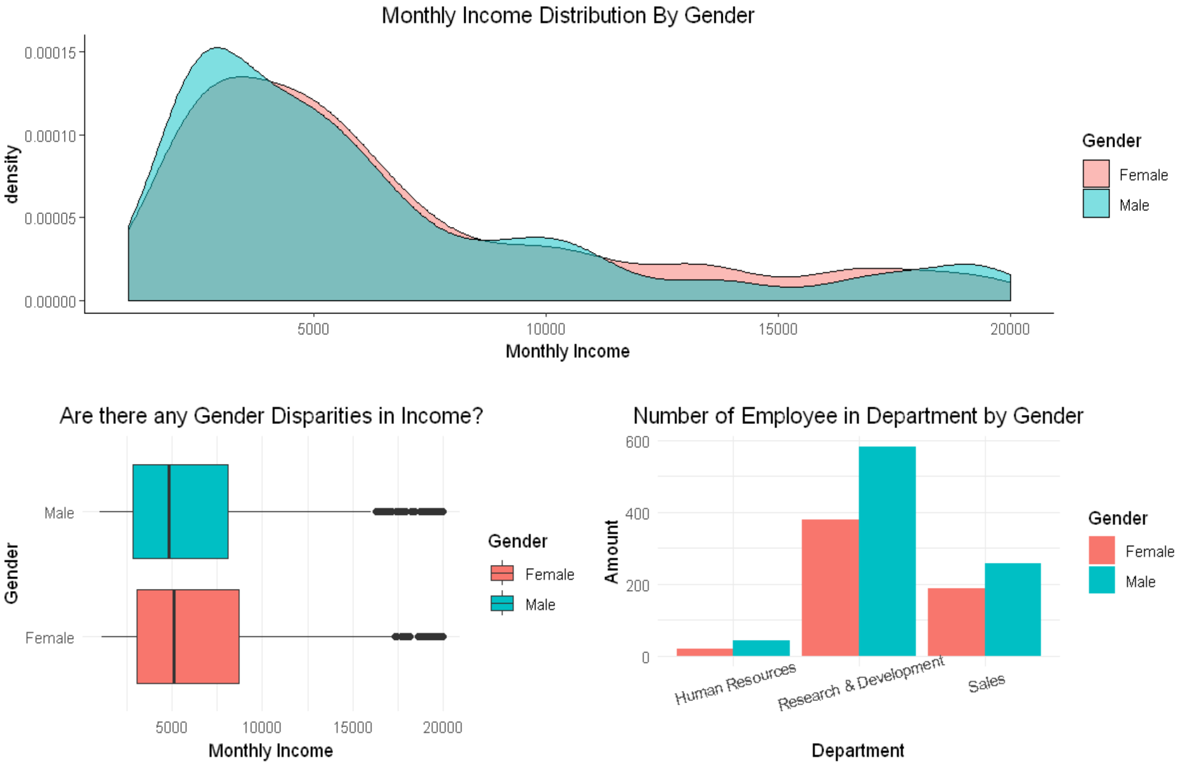


Figure 7: Monthly Income by Gender and the Disparities by Department

Those the number of female respondents in the survey is less than number of male employees, but the average salary of female is higher. This is interesting as the median salary for men is roughly 21 percent higher than for women regardless of job type (Payscale, 2019). Perhaps, there is a bias in the survey as it concentrates on 3 major departments only. In addition, though the number of male employees in each department is dominant, but male employees prefer to work in research development than women, on the other hand, females prefer to work in sales positions more than males.

4. Generation and Education Analysis

This section examines the relationship between the generation and education of the employee towards the attrition. Following is the question to find the answer for the analysis.

- The attrition status and income by the generation
- What is the average number of companies previously worked for each generation? Is that true that the past generation used to stay longer in the company while the millennials tend to switch companies more often?
- Does education affect to the attrition status?

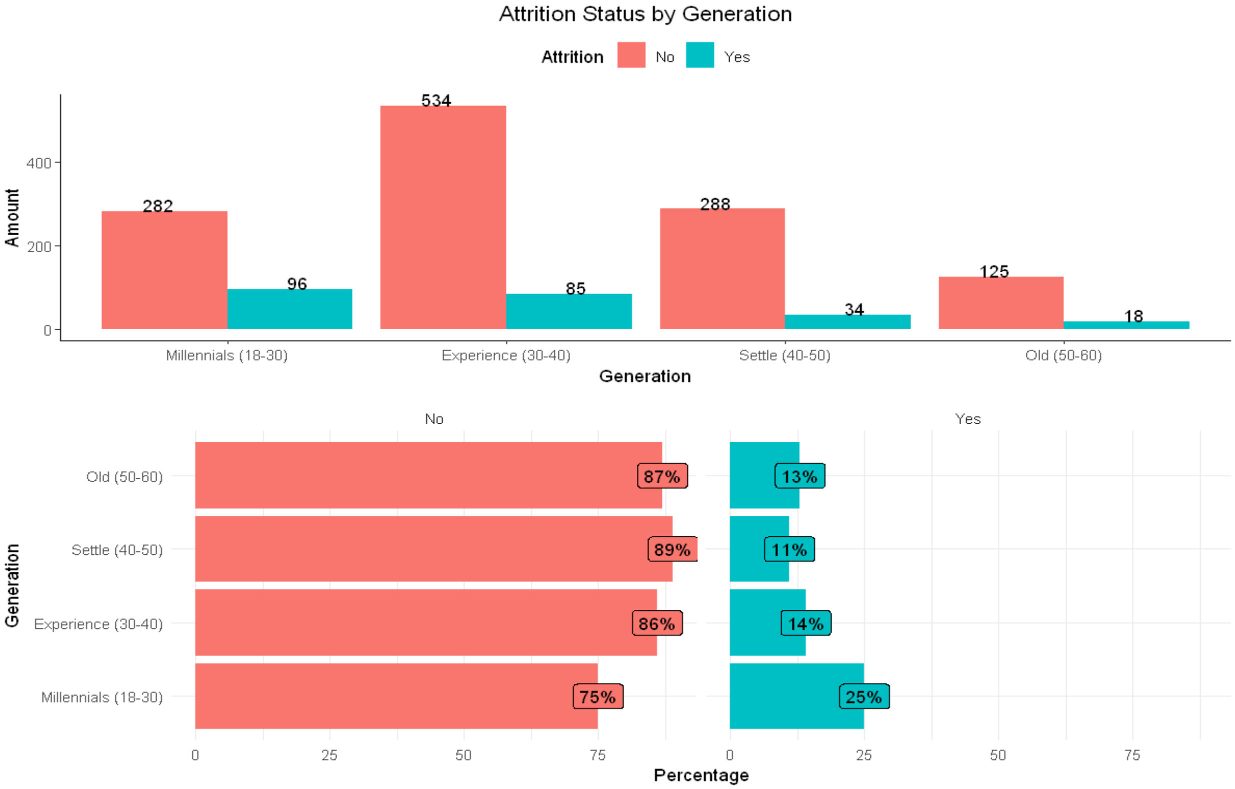


Figure 8: Attrition Status by Generation

It is understandable when the employee between 40 to 60 years old have the lowest rate of attrition while the millennials have the highest rate. The new generation like millennials opts for more opportunities and other jobs that satisfy their needs. The old generation has a turnover rate higher than the settle (40-50) generation, the reason could be they are approximating retirement.

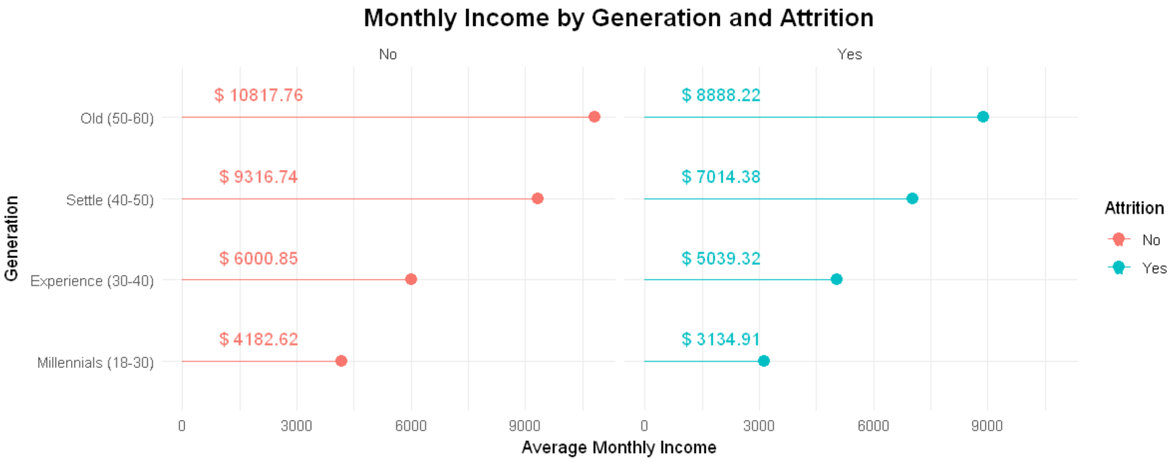


Figure 9:Monthly Income by Generation and Attrition

Whatever the generation, it is clear that people who want to quit an organization have a lower average salary than people who want to stay. The depth-look in income toward to the attrition will be conducted more in section 5.

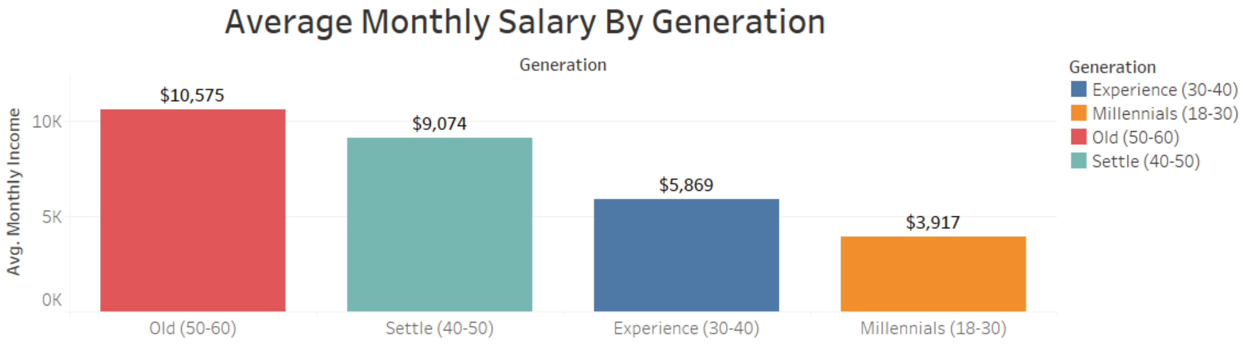


Figure 10: Average Monthly Salary by Generation

It is true that the older employees get, the more salary they receive. This is because the organization mainly paid the employee based on their experience.

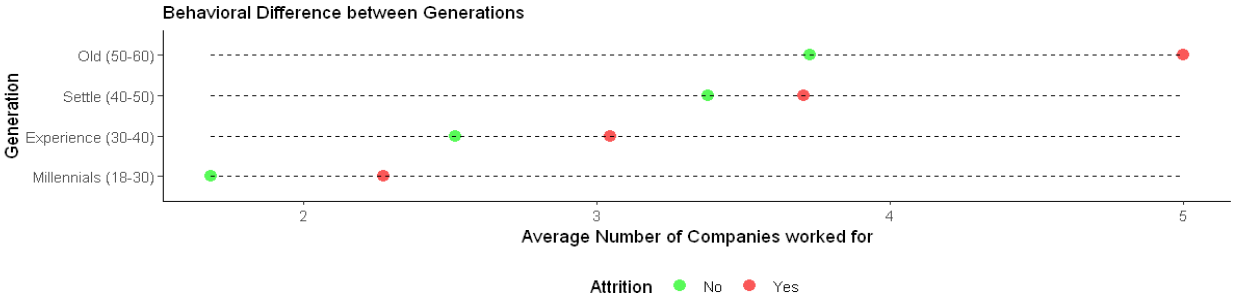


Figure 11: Behavioral Difference between Generations

Also, no wonder that the older generation works for more companies than the other generation while the millennials are still relatively young. There is a trend that whatever of the generations, the people who quite the companies always work for more companies than people who stay.

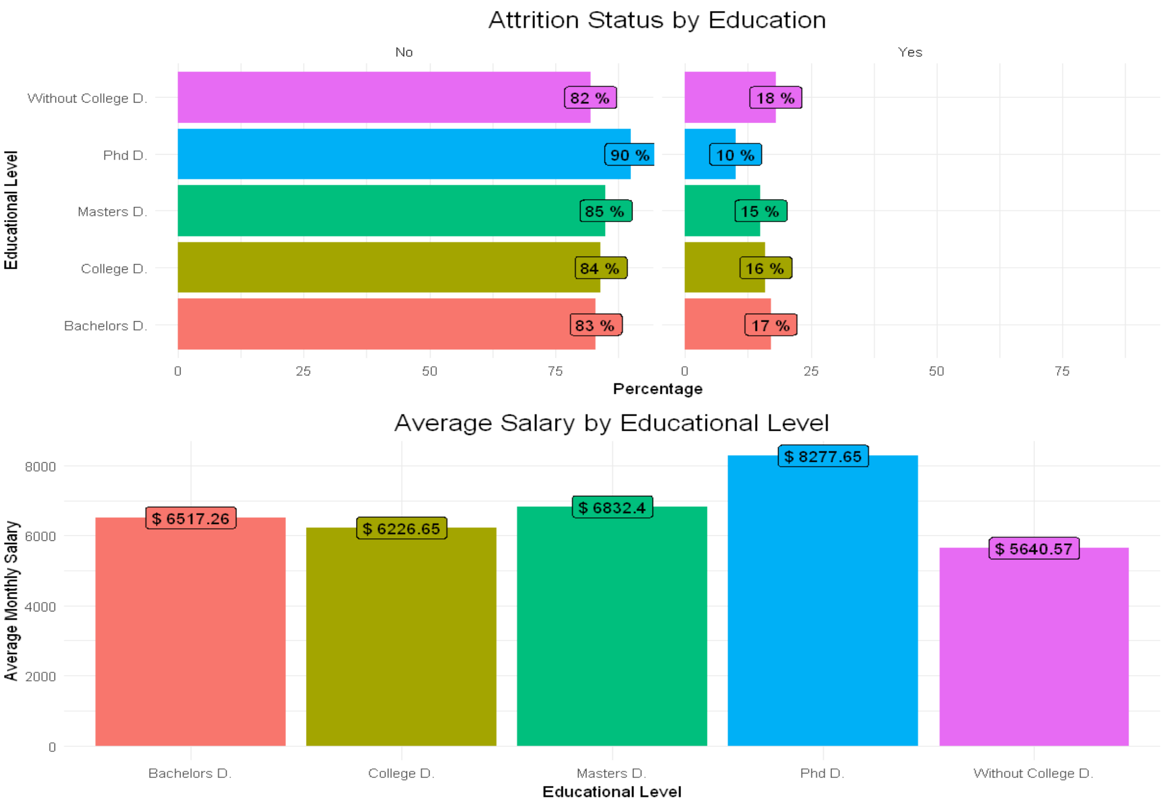


Figure 12: Educational Level, Average Salary and the Attrition

Based on the survey, it seems like the more degree employees have, the less turnover rate they are. Ph.D. employees have a turnover rate of only around 10% while without college D group has the highest turnover rate with 18%. The education level also reflects the average monthly salary as the Ph.D. D group gets the highest pay while the without college D group receive the lowest salary. In conclusion, the educational levels affect the salary and salary affect attrition status.

5. The Impacts of Income Towards to Attrition

This section will look in-depth to one of the most important factors which may have a significant impact on attrition. In order to answer the question that how much importance does each worker gives to the salary they earn, there are some questions need to be asked to reach the conclusion.

- How much does an employee earn in each department? Is there any noteworthy difference between individuals who leave and stay?
- What is the average income of employees by job role? Do people who do not quit earn more people who quit the same job role?

- Is there any significant difference in salary by job satisfaction? Is it true that an individual who has lower job satisfaction because they get lower paid?
- Do employees quit the organization because they have a lower income?
- Do employees with higher performance ratings earn more than with lower performance ratings? Does it affect attrition status?

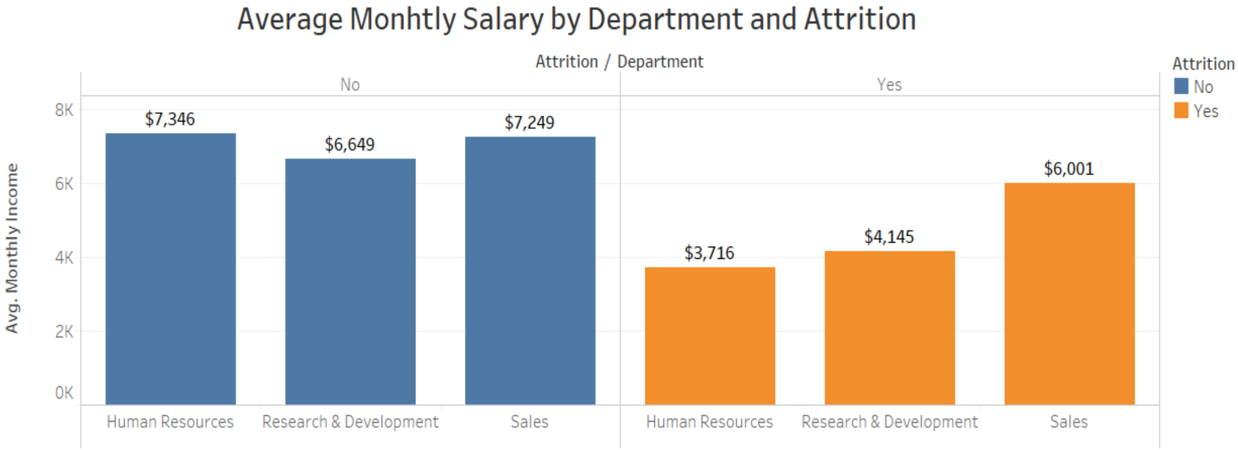


Figure 13: Average Monthly Salary by Department and Attrition

There is a huge different attrition status in each department. It is clear employees who quit get lower salaries than people who stay. Also, people who work in the Human Resource Department and stay in organization has average monthly salary more than double average income of people who work in the same department but want to quit.

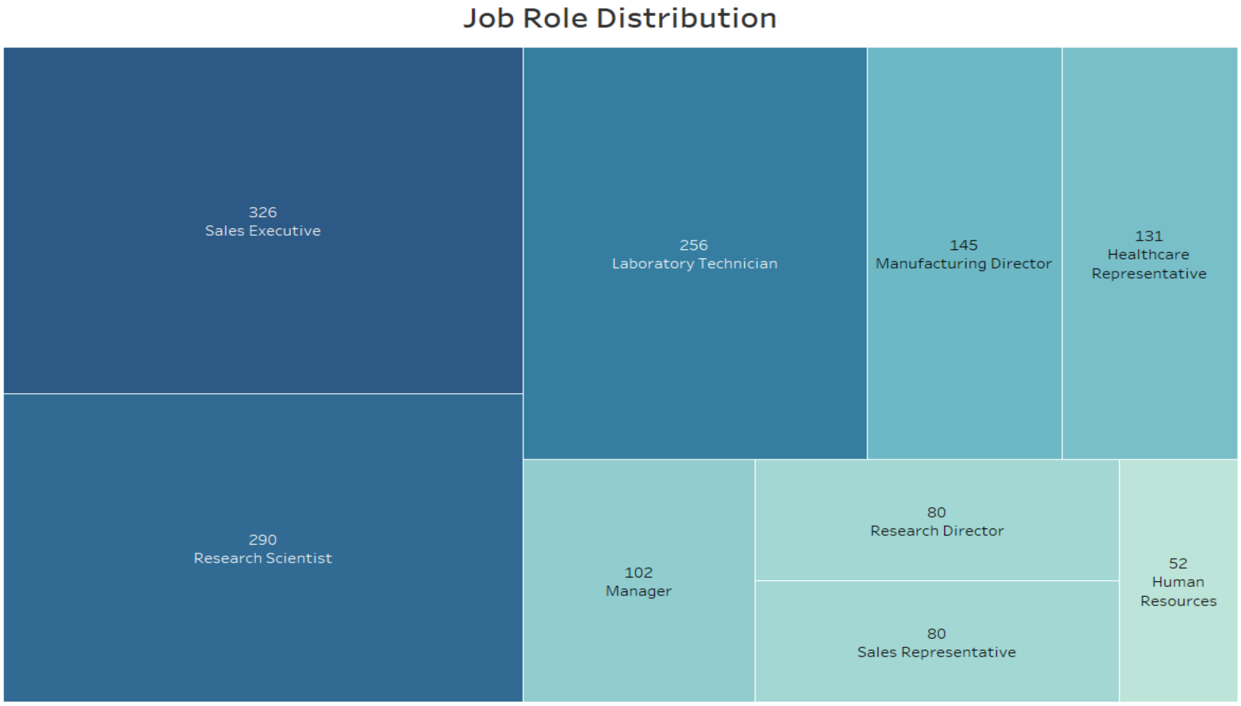


Figure 14: Number of Employees by Job Role

In terms of job roles, the number of employees works as the Sale Executive accounts the most, the following is Research Scientist and Laboratory Technician with 326, 292, and 259 respondents, respectively, while only 52 respondents with the title are Human Resources.

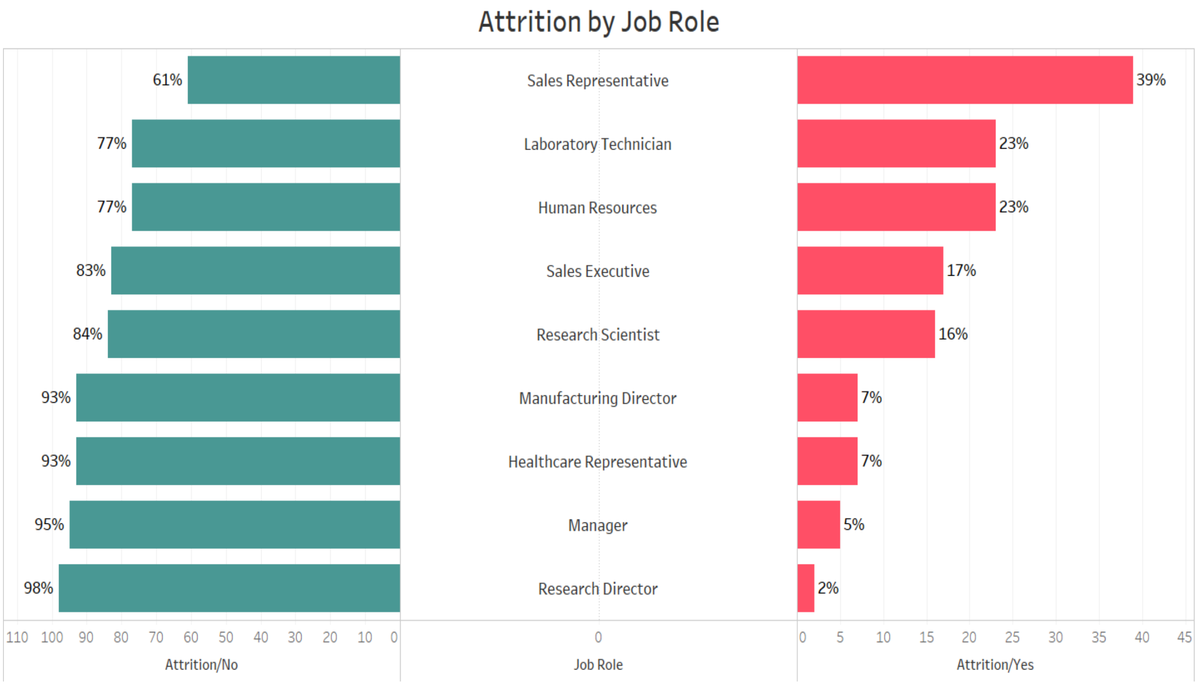


Figure 15:Attrition by Job Roles

Regarding the turnover rate, there is around 40 percent of respondents work as a Sales Representative want to quit their job, the following is Laboratory Technician and Human Resources.

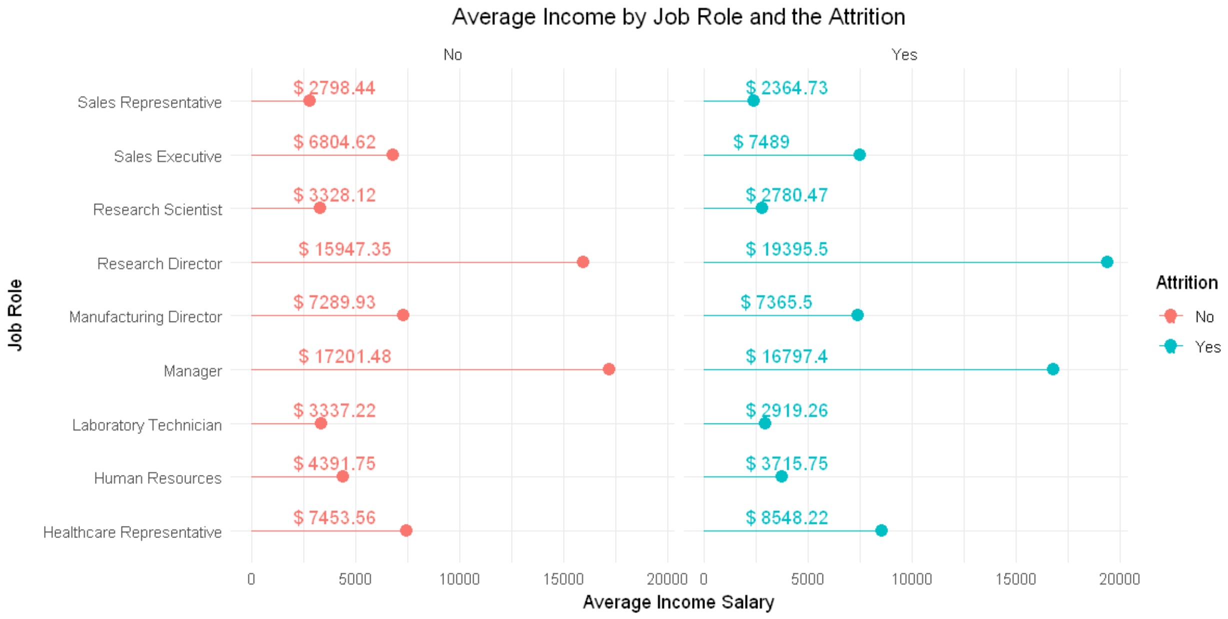


Figure 16: Average Income by Job Role and the Attrition

Look into the average salary by job role, it can be seen that almost the people who want to leave have a lower salary than people who stay but the Manager, Healthcare Representative, and Research Director, Sale Executive. Thus, the hypothesis that employees who have higher average monthly salary will stay longer than employees who have lower average monthly income seems be true when considered by department, however, when coming to the Job Role, this hypothesis has some exceptions. Hence, monthly income is a vital factor but there are also other important factors that affect the employees' decision rather than just money. Consider the job satisfaction factor in the graph below.

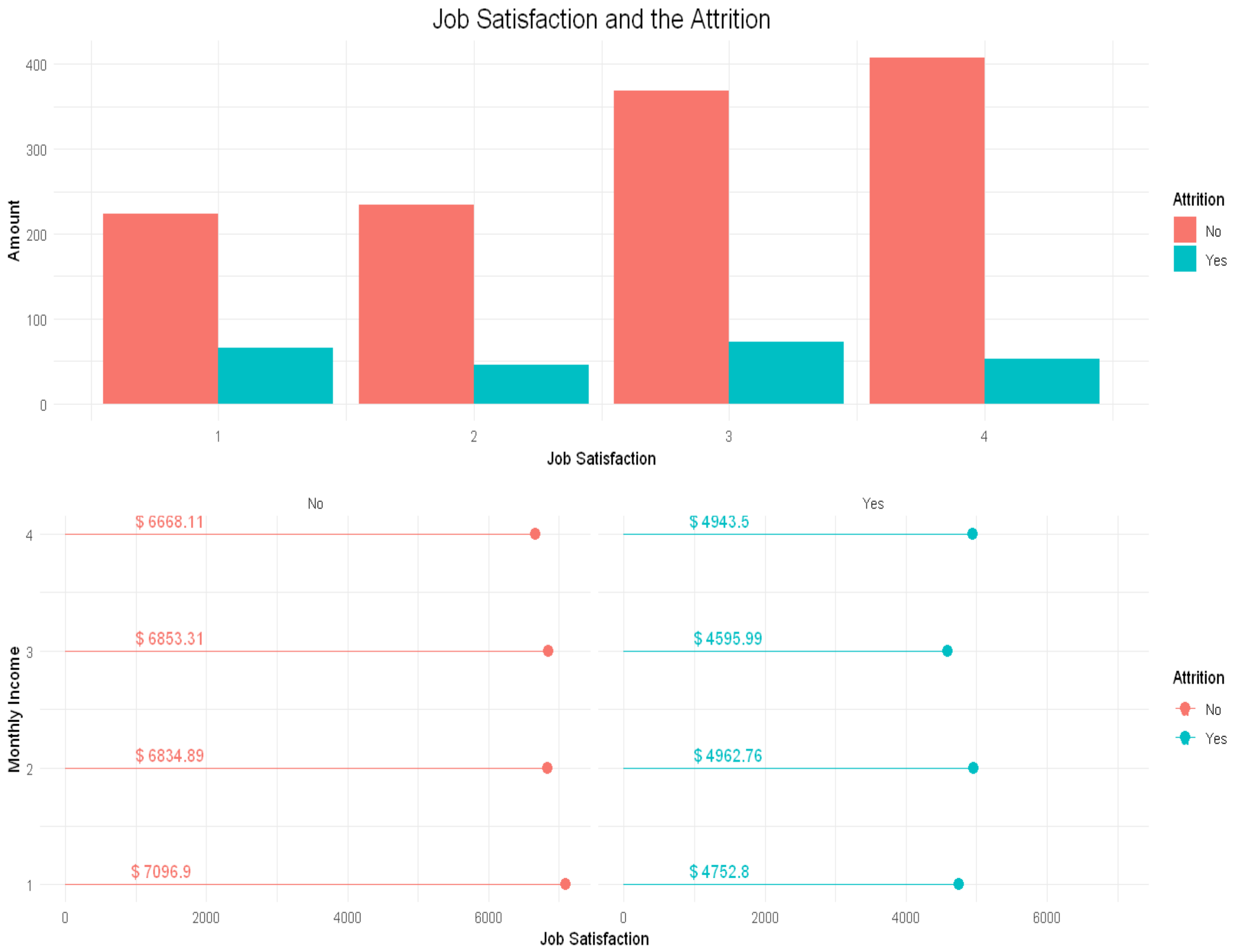


Figure 17: Job Satisfaction with Monthly Income and Attrition Status

The higher the rate of job satisfaction employee gives for the organization, the longer employees will stay. However, the interesting here is even when employee gives the organization only 1 star for job level satisfaction, but they willing to stay if they get high monthly income. In addition, there is relationship between job satisfaction and monthly income and that relationship affects attrition.

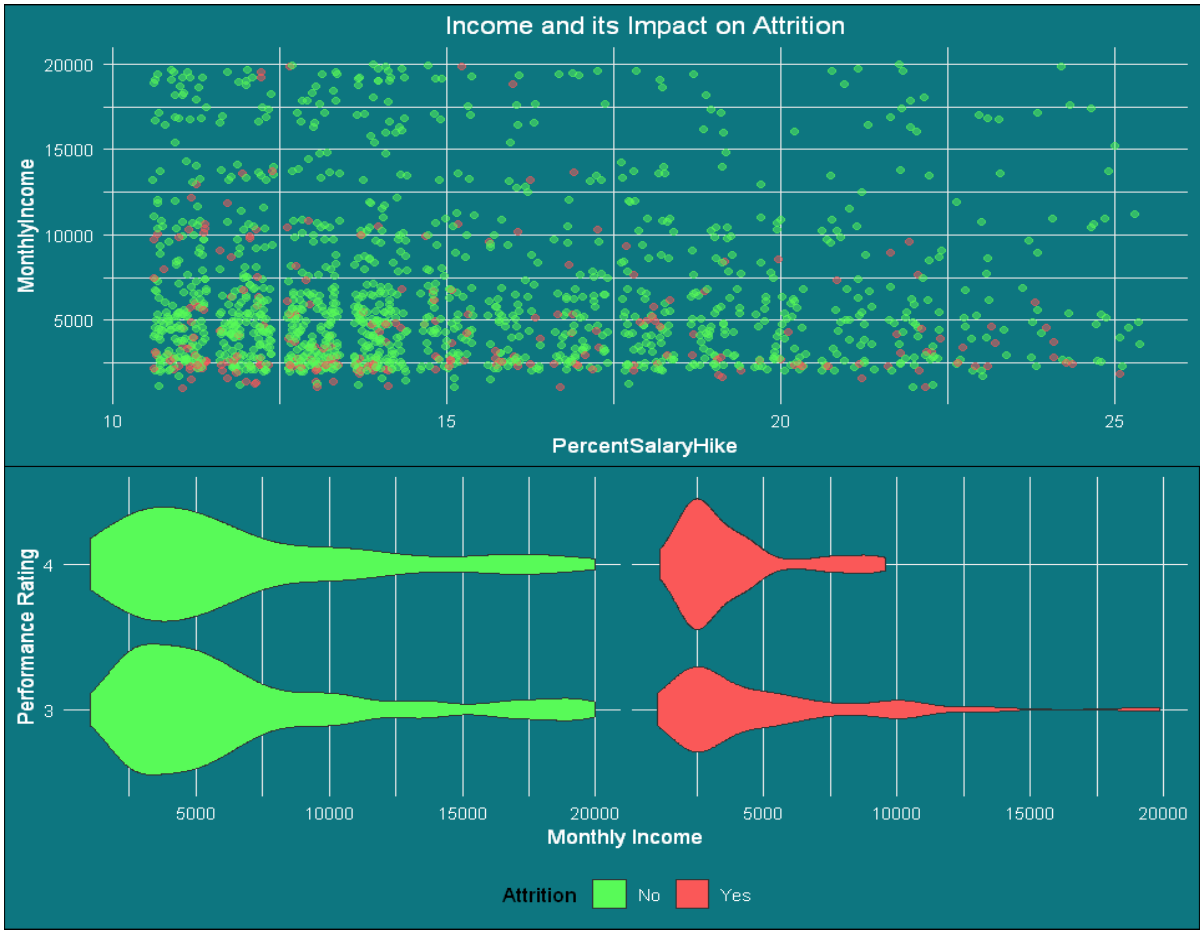


Figure 18:Income and its Impact on the Attrition

There is no employee say yes to the attrition if they receive more than $10000 monthly salary and have a percent salary hike greater than 17.5%. It is witnessed that employees still quit the organization even when they get paid more than $15000 per month, but the percent salary hike is low. In term of performance rating, if the organization evaluate the employees' performance lower than their work by paying a lower salary or lower percent salary hike, they might want to consider leaving the companies.

6. Working Environment Analysis

From the previous analysis, a monthly salary is a vital factor but not the only reason to cause attrition. This section will look in-depth the other factors such as working overtime, working environments, etc, to identify how important those factors that affect the turnover. Manager, Research Director, HealthCare Representative and Sale Representative have higher income but also have a higher turnover rate if income does not affect their decision to quit the company, whether the working environment is the main reason?

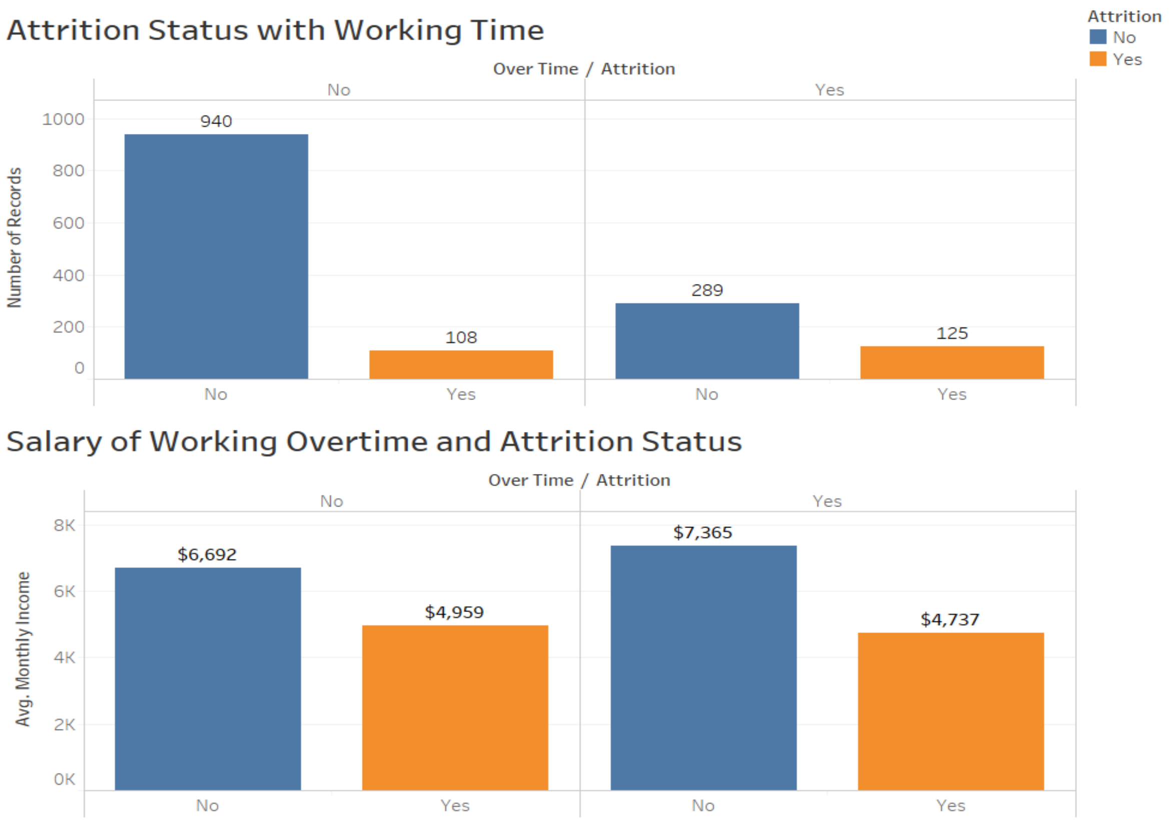


Figure 19:Working Overtime and Income

Look in the graph above, in both group which is working overtime and no working overtime, approximately 30% employees want to quit their job if they have work overtime in comparison to only 10% of the no working time group. The interesting about employees who do not want to quit even they report to have work overtime condition is their income higher than average. In addition, employees in both groups who want to quit have a low salary.

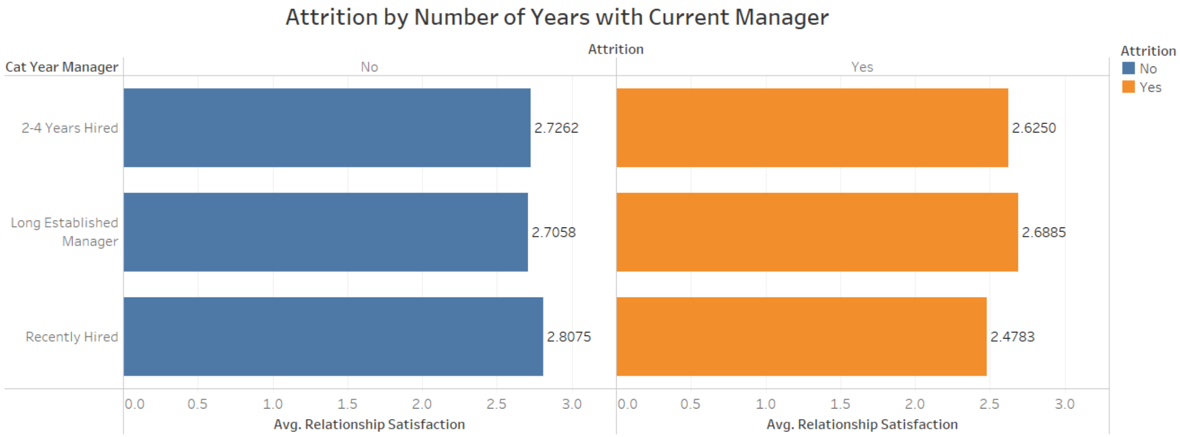


Figure 20: Attrition by Number of Years with current Manager

The average relationship of employees that work with a current manager who left is lower than employees who want to stay.

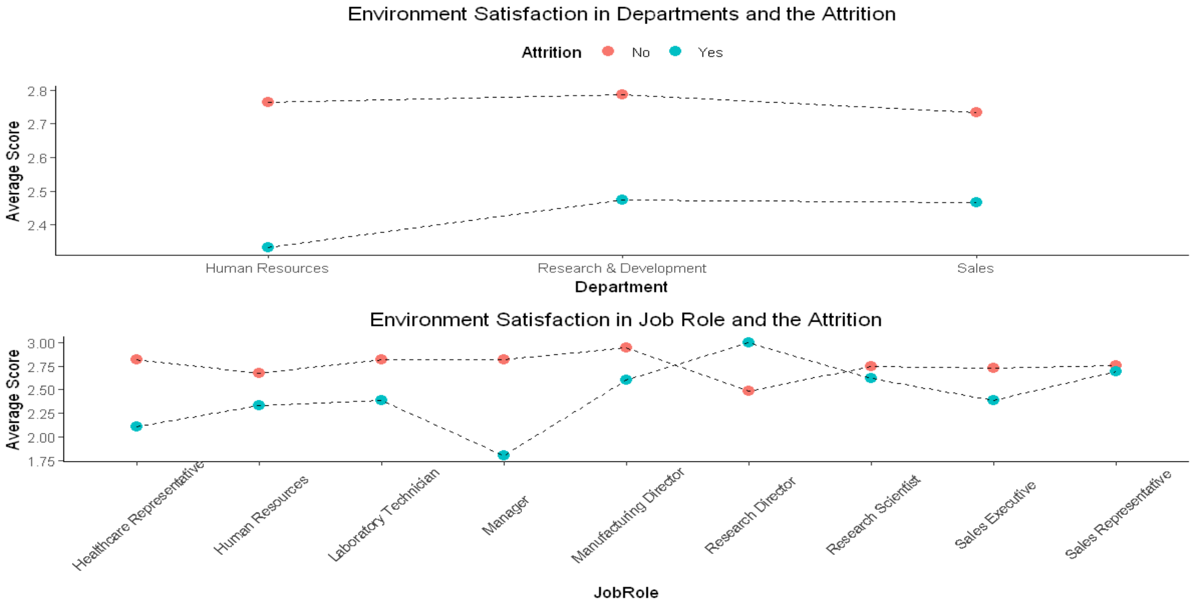


Figure 21:Environment Satisfaction in Department, Job Role and the Attrition

From the previous section, the Manager, Healthcare Representative left the organization even when their income is high, mainly because those employees need to deal with the low working environment. Sale Representative also has a higher turnover rate than the average but no witness the low score for working environment, that could be because most sale representative work outside the organization.

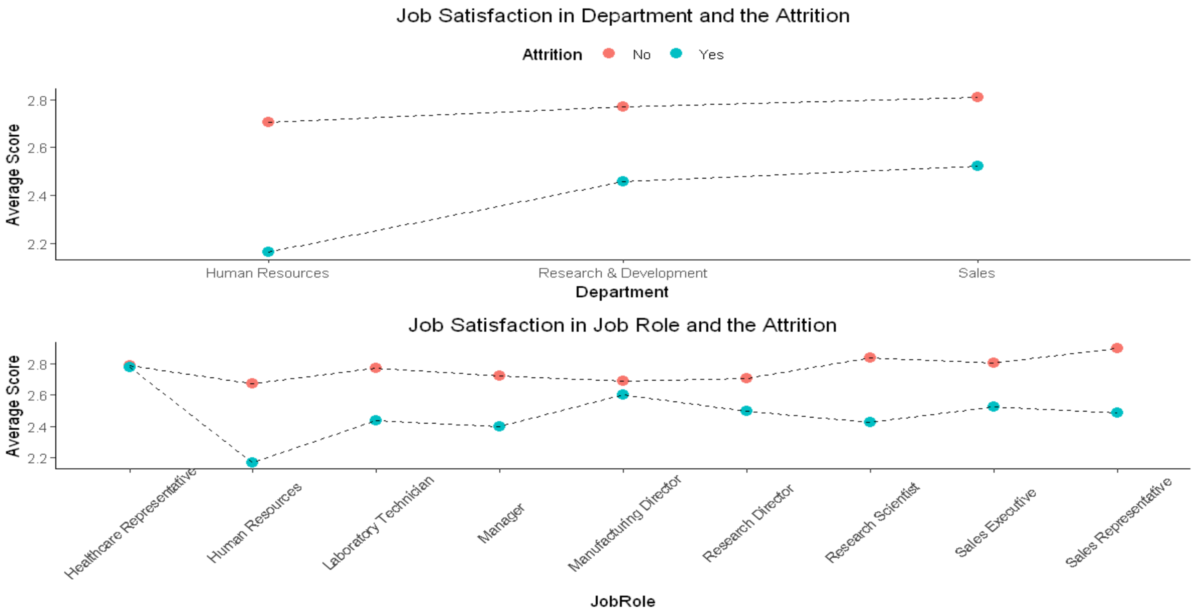


Figure 22: Job Satisfaction in Department, Job Role and the Attrition

When coming to job satisfaction, it is an obvious trend that employees no matter the departments they are working or their job role, will leave the organization if average score for job satisfaction is low.

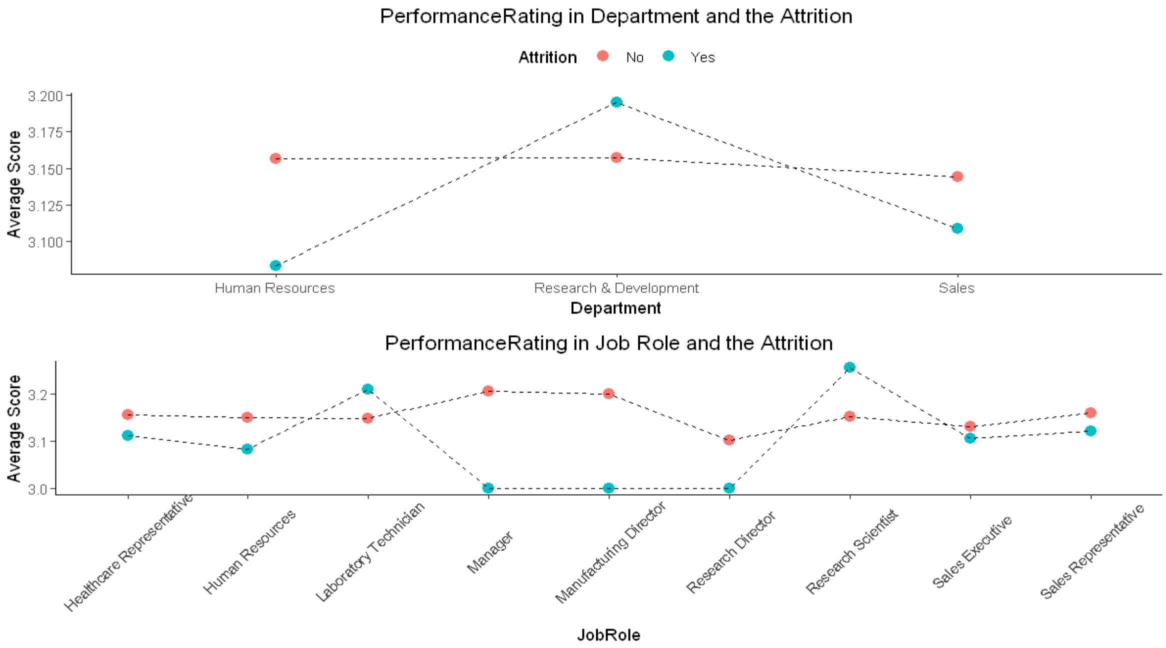


Figure 23: Performance Rating in Department, Job Role and the Attrition

In terms of performance rating, the Manager, Manufacturing Director and Research Director have the lowest rate. These factors and the working environment factors are the main reason to cause the turnover in Manager, Director rather than income.

7. An In-depth Look at Attrition

Among all the employees who want to quit the organization in this survey, more than 54% of respondents said that they work overtime.

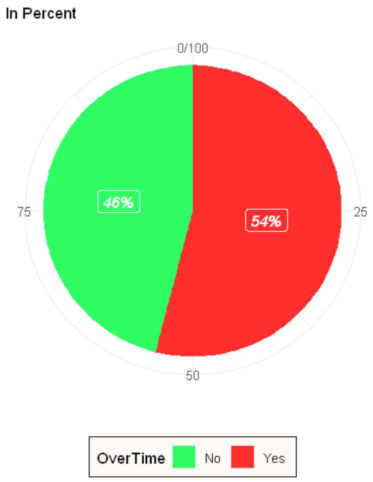


Figure 24: Working Overtime Attrition

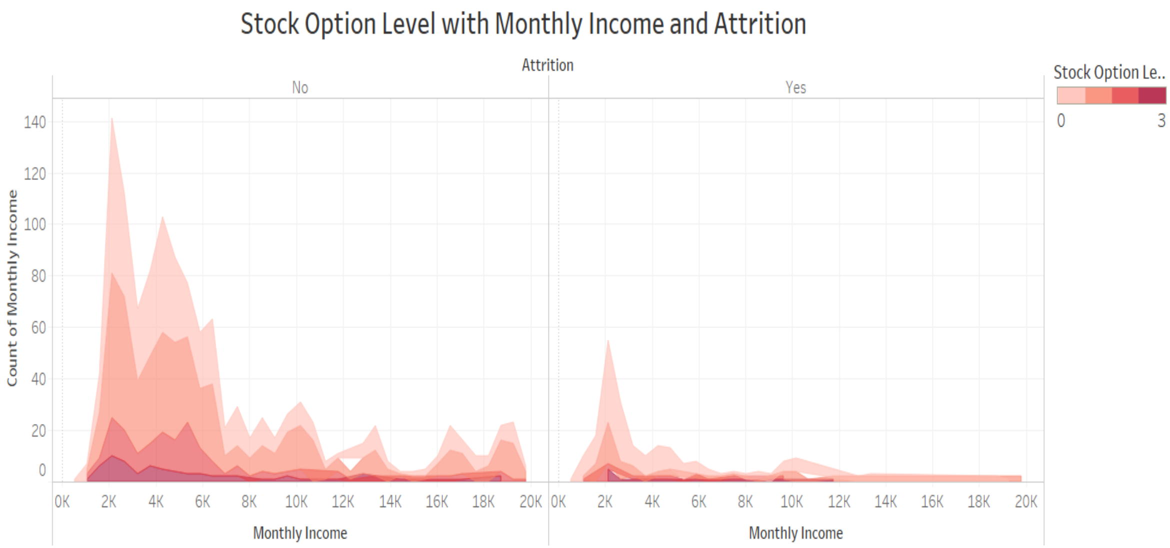


Figure 25: Stock Options Level with Monthly Income and Attrition

Stock options with income appear dominant to the turnover.

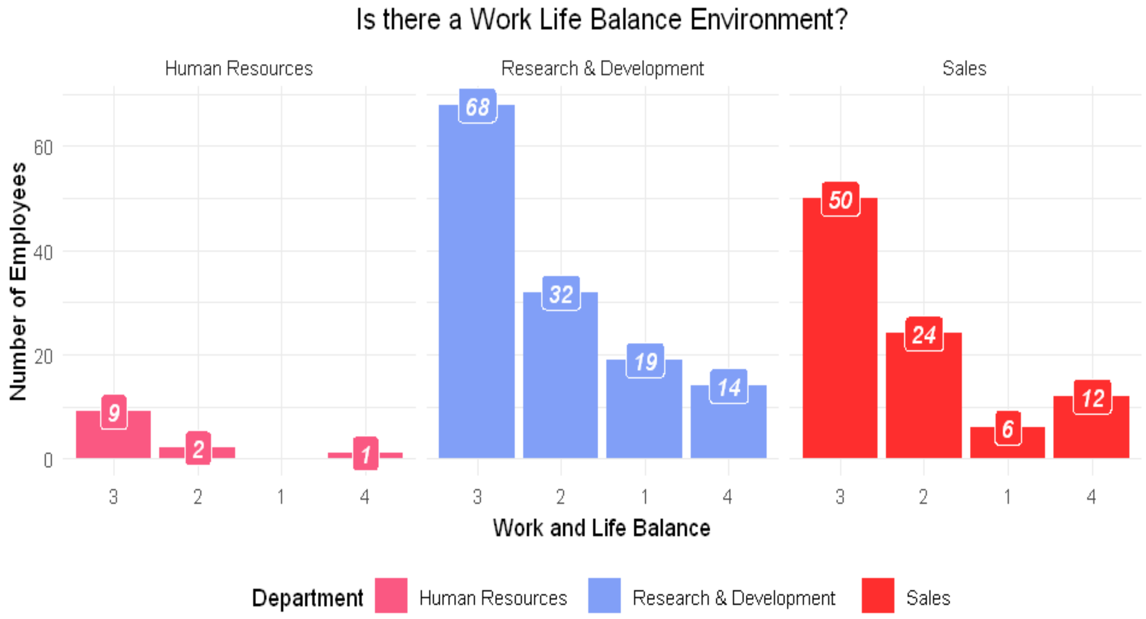


Figure 26: Work-Life Balanced Environment

It is interesting when people who quit actually have quite a good work-life balance. Perhaps, the main reason to quit is about the working environment, job satisfaction and monthly salary rather than the ability to balance between work-life.

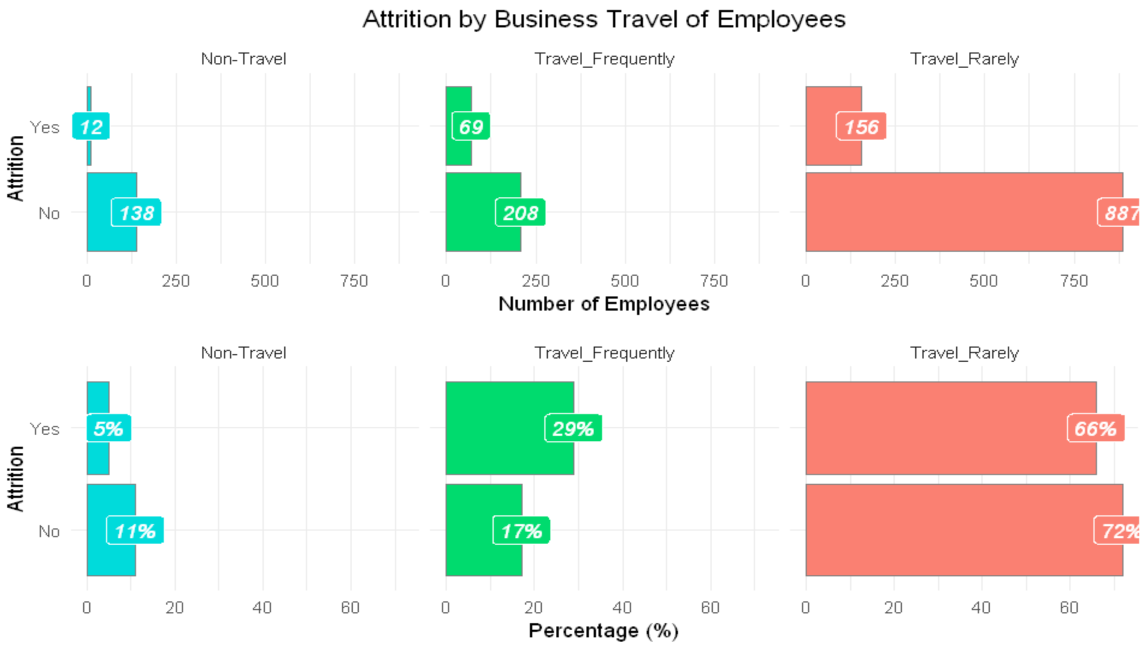


Figure 27: Attrition by Business Travel of Employees

It seems unfair when employees who travel at work frequently get average income lower than those who travel rarely and non-travel. Among all the employees who quit, only 5% from non-travel while travel rarely and travel frequently account 66% and 29%, respectively. It is understandable because the average income non-travel employee receives more than travel frequently and travel rarely.

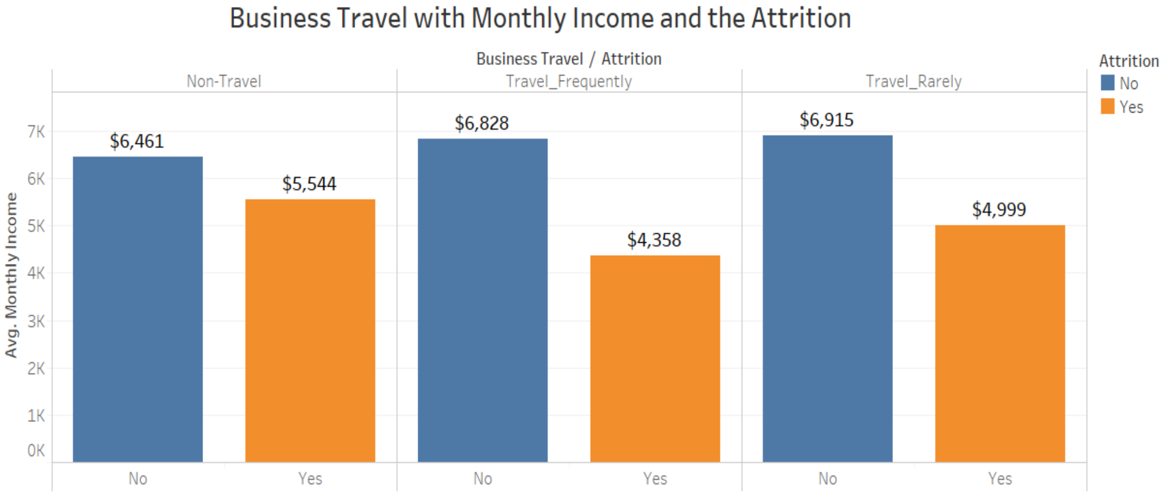


Figure 28: Income by Business Travel and Attrition

8. Model Analysis and Prediction

8.1. Correlations and Bi-variate Analysis

This section will give an understanding of what features have a positive correlation with each other to see whether there is an association between two features.

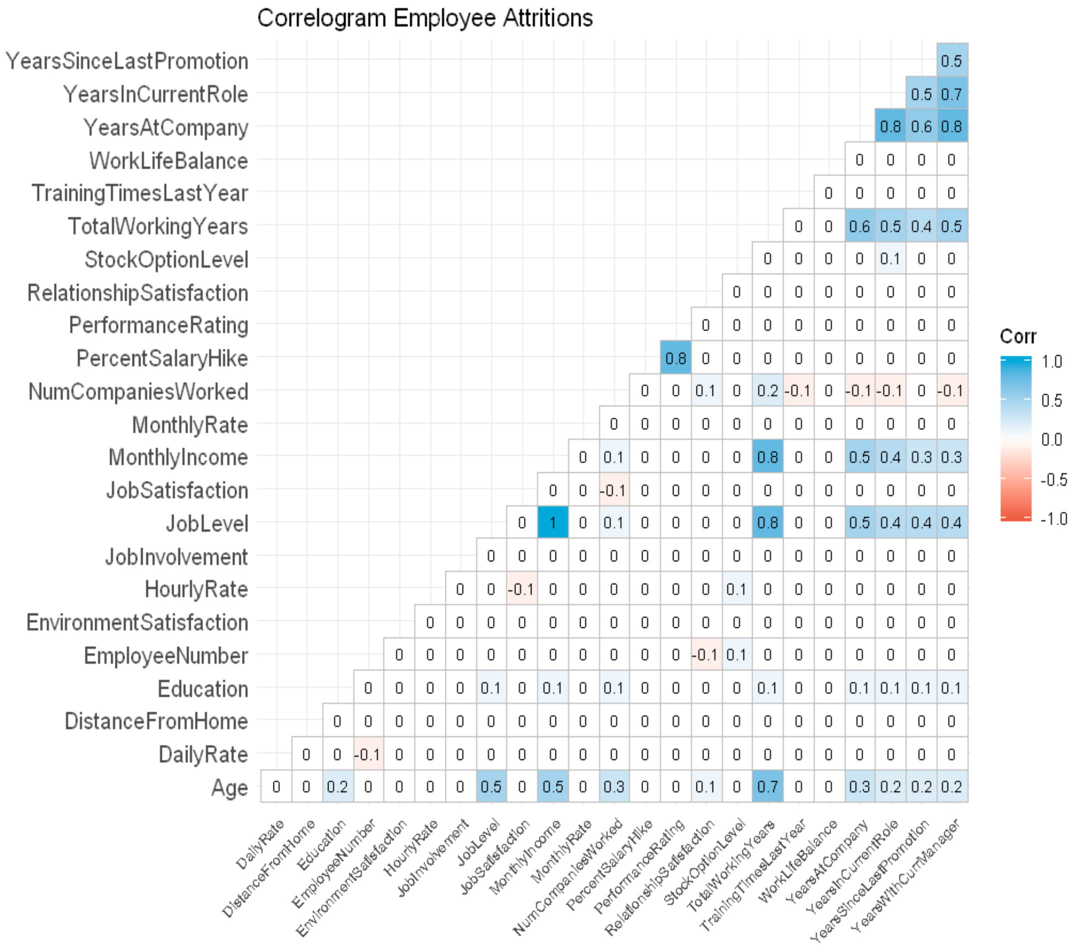


Figure 29: Correlogram Employee Attritions

It seems like the total working years the higher the monthly income of an employee, also, the higher percent salary hike, the higher performance rating. In addition, the higher the years with current managers, the higher the years since the last promotion and higher the age, the higher the monthly income.

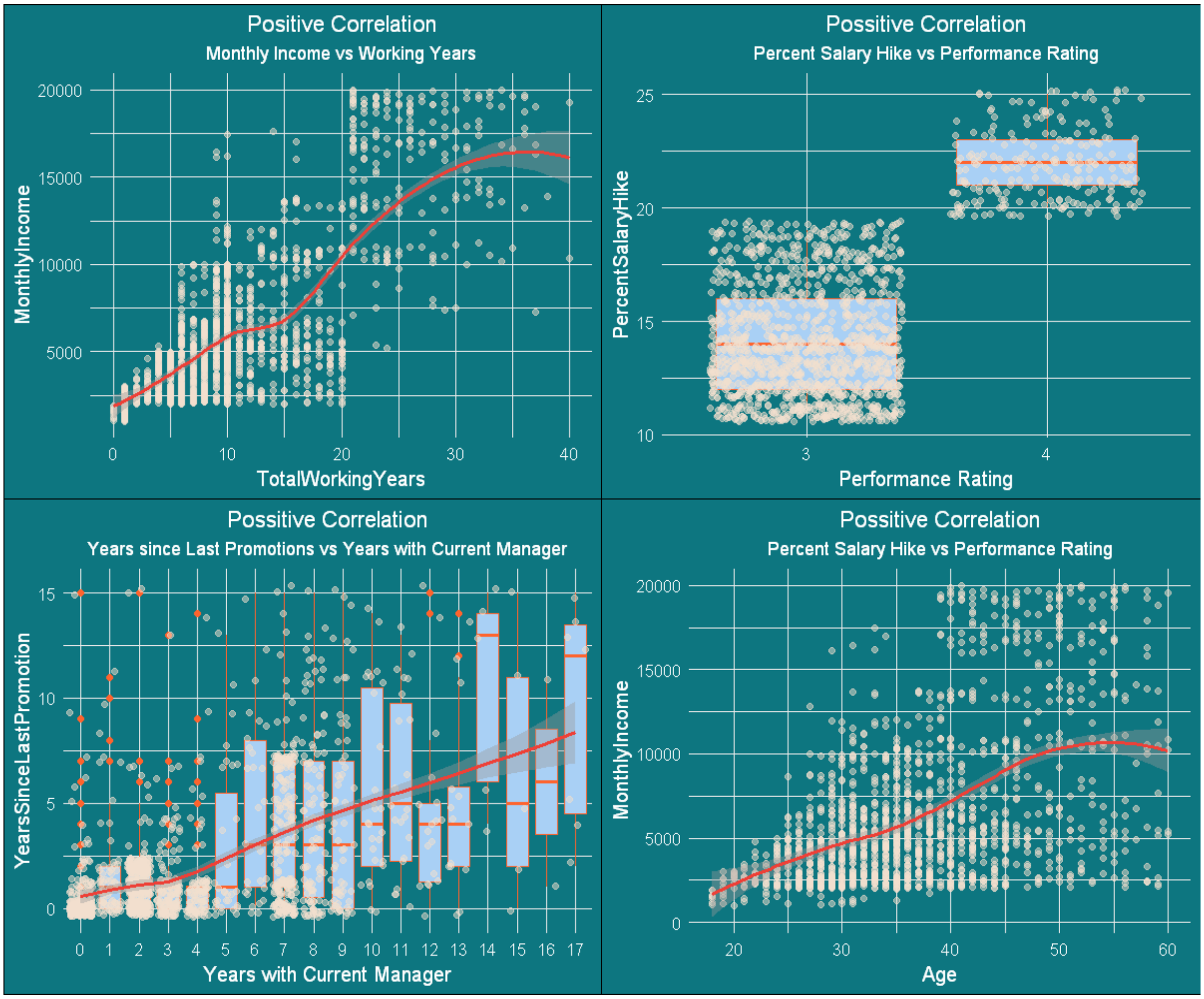


Figure 30: Bivariate Analysis between Features

8.2. Attrition Prediction Model

8.2.1. Decision Tree Theory

- **Gene Impurity**: If all the training instances belong to the same class the impurity of a specific node will be equivalent to zero. For instance, let's assume that all employees who had a WorkLifeBalance $<$ 2, decided to leave the organization. In this case, the impurity will also be zero.
- **Calculating Impurity:**

$$\boldsymbol{G_i = 1 - \sum_{k=1}^{n} p_1, k^2}$$

Where p, i, k is the ratio of class k instances in the i^{th} node

- **Classification and Regression Tree (CART) algorithm:** the main idea of this concept is to split the training set into two smaller subsets and create a conditional statement using one feature. (For instance, MonthlyIncome $<$ 25k). How does the algorithm determine the conditional statement? It looks for the conditional statement that creates the "purest" subset (remember impurity = 0 or in this case the lowest impurity)
- **The formula of the CART Training Algorithm (Finding our threshold for each feature):**

$$\boldsymbol{J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}}$$

where:

$G_{left/right}$ = determines the impurity of both subsets

$m_{left/right}$ = total number of instances on each subset.

References: (Géron, 2019) Hands-on Machine Learning with Scikit-Learn and TensorFlow (Chapter 6: Decision Trees)

8.2.2. Decision Tree Model

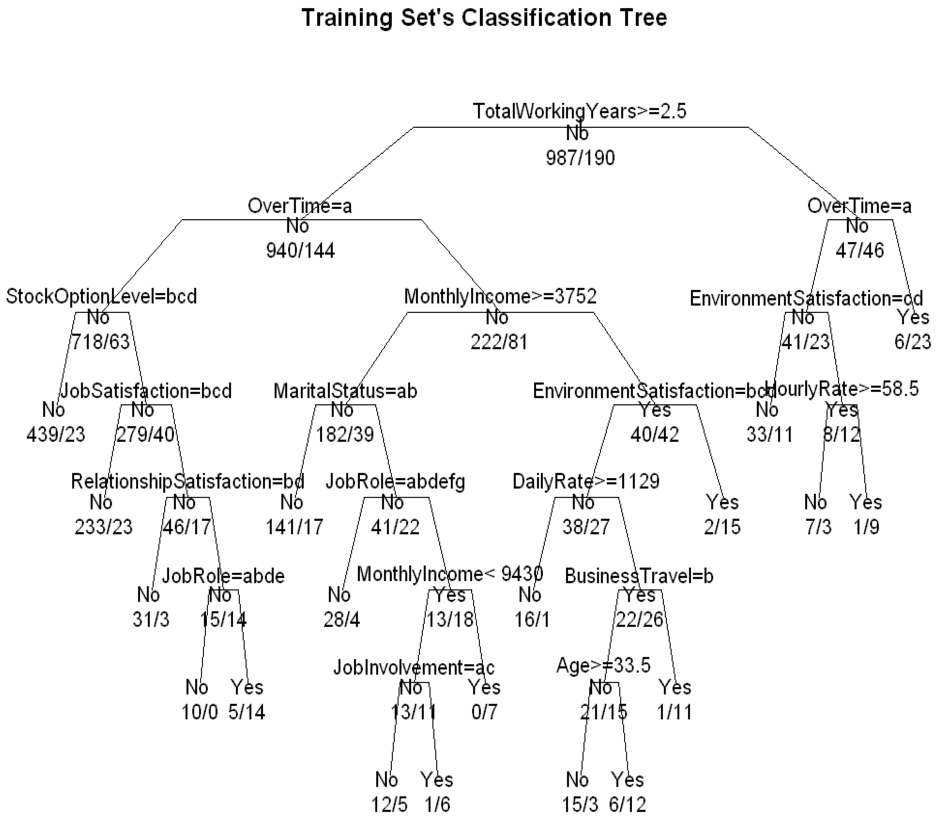


Figure 31: Decision Tree of the Dataset

According to the decision tree model, it considers these six attributes as the most important:

- **Monthly Income**: income might be a determinant factor for employees to leave
- **Job Role**: employees do not like their current position
- **Job Level**: employees believe they deserve a higher job level are prone to leaving the organization
- **Overtime**: working overtime affect the quality of life of employees
- **Total Working Years**: employees might be retiring or looking for more challenge position
- **Age**

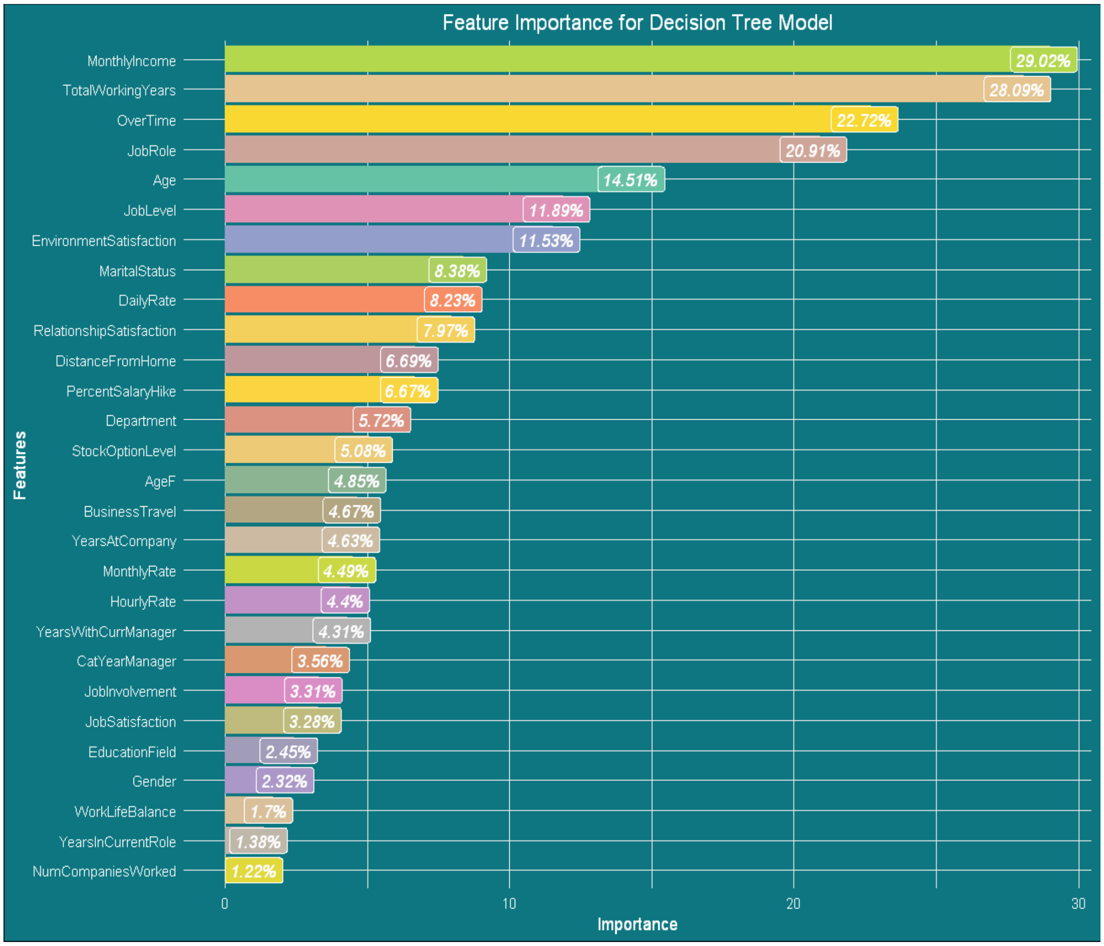


Figure 32: Features importance by Decision Tree

Those important factors that spot by the decision tree confirm the hypothesis and data analysis from previous sections as it concerns monthly income, working overtime, etc are the most critical factors that affect attrition.

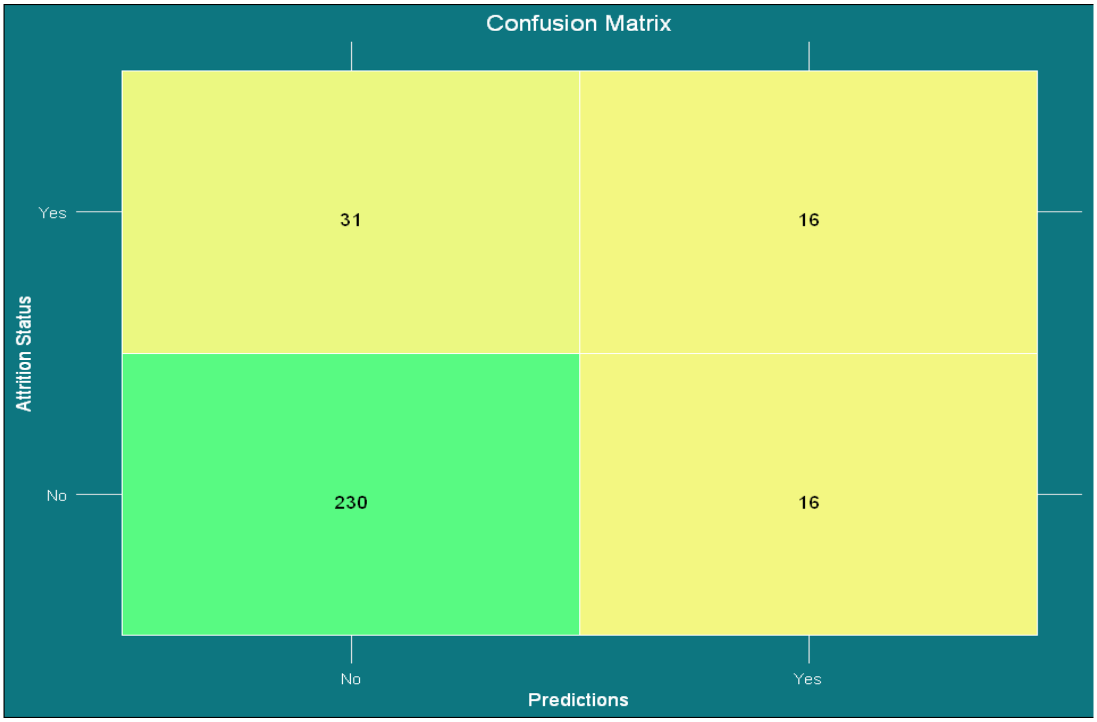


Figure 33: Confusion Matrix

In terms of accuracy, this model gives 82% accuracy. The main reason the accuracy is not high because the dataset is small as it only has 1470 rows, also, as the dataset is imbalanced data, it could affect the accuracy of the algorithms.

In order to overcome this challenge, there are several ways to consider to improve accuracy.

- Collect more data
- Handle the imbalanced datasets by using the oversampling technique
- Perform more feature engineering
- Use cross-validation
- Implement different algorithms and tune its parameters.

In this case, it is impossible to get more data, therefore, to improve the accuracy, the following steps have taken before feed data into the algorithm.

- Use the oversampling technique to addresses the imbalanced datasets
- Use cross-validation with 10 folds and repeat 3 times
- Use random forest instead of a decision tree because the decision tree cannot handle the sensitive data.

The accuracy increased from **82%** to **85.86%.**

8.2.3. Examine the Impact of the Importance Features to the Attrition

To test the hypothesis that those factors: MonthlyIncome, OverTime, JobRole, Age, TotalWorkingYears, EnvironmentSatisfaction are the most important attributes that have a high effect on the accuracy of the model. This section will perform several examinations to evaluate precision of different algorithms with data that includes all the features and the data that include only important features.

Dataset Type	Algorithms	Cross-Validation	Accuracy
All feature	Decision Tree	No	82.76%
All feature	Decision Tree	Yes	83.1%
Only Important Feature	Decision Tree	Yes	83.28%
All feature	Random Forest	Yes	85.86%
Only Important Feature	Random Forest	Yes	84.23%

Table 4: Accuracy of different algorithms

When excluding the less important feature from the dataset, the accuracy slightly increased if using the decision tree algorithm with cross-validation, however, the accuracy also slightly decreased if using the random forest algorithm. The reason mainly because the decision tree considers all the features of the data and it calculates the impurity of each feature, while random forest considers small subsets of feature and construct a multiple of decision tree at training time and outputting the class that is the mode of the class or mean predictions of the individual trees. Though the five important factors are the most important but still there are some other factors that can contribute even a small change to the output which is the attrition.

This experiment also confirms that those factors are the most vital features as with only those six features can achieve very close to the accuracy of the random forest with all the feature. When coming to the decision tree, using only six important features can outperform the accuracy of the decision tree with all features.

9. Conclusion and Recommendation

In the flat economy, the employee is the backbone of any organization and one the main key to successful business. If the company takes care employees well, employees will take care company's customers well. While most employees stay in the organization if they receive high paying salaries, but it is important to note that the income is not the only main factor contributes to the turnover rate. If the company cannot afford high paying for their employees, make sure they satisfied with the work environment, pleased with job satisfaction as well as working relationship and reduce working overtime.

In addition, prediction of the turnover of an employee can be tricky as it depends on many conditions. The most obvious way to increase accuracy is by collecting more datasets. As the data tell stories its sell, it is worth to spend time and resources to get more data. Enhancing a model performance can be challenging at times, for this dataset, it would be better to spend time to do more feature engineering by extracting more information from existing data. Some methods such as feature transformation, feature selection, feature creation could be helpful to improve model accuracy. Finally, as machine learning algorithms are driven by parameters and each algorithm has different parameters to tune, so, intuitive optimization of all these parameters will result in better and more accurate models.

REFERECES

Géron, A. &. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition (2nd ed.).* O'Reilly Media Company.

Katsikea, E. T. (2015). Why people quit: Explaining employee turnover intentions among export sales managers. *International Business Review*, 24(3), 367-379.

Payscale. (2019). *THE STATE OF THE GENDER PAY GAP.* Payscale.

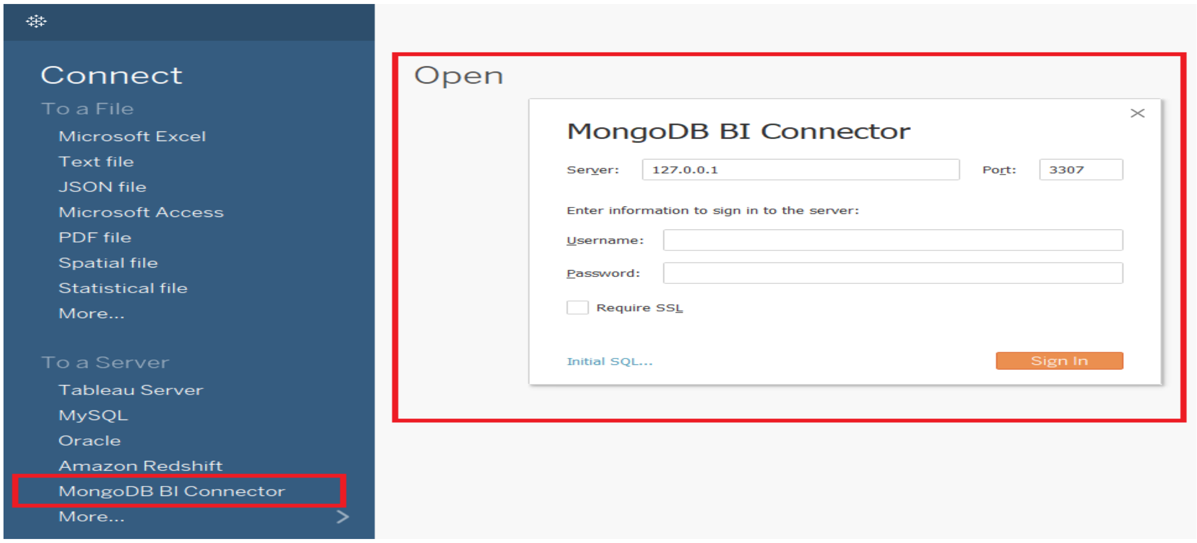
Pestronk, M. (2014). Employee might quit, but commission stays with ex-employer. *Travel Weekly,*, 73(51), 9.

APPENDIX

This section contains supplementary the R code for all the visualizations and wrangling above. For more information, refer to the R file and Tableau files.

For Tableau and MongoDB connection

Open Tableau Desktop and select MongoDB Connector from **To a Server** section. In the popup, the server: **127.0.0.1** and the port is **3307**, the server URL and port come from the red box in the figure of running **mongosqld.exe** above.



MongoDB BI Connector

Server: 127.0.0.1 Port: 3307

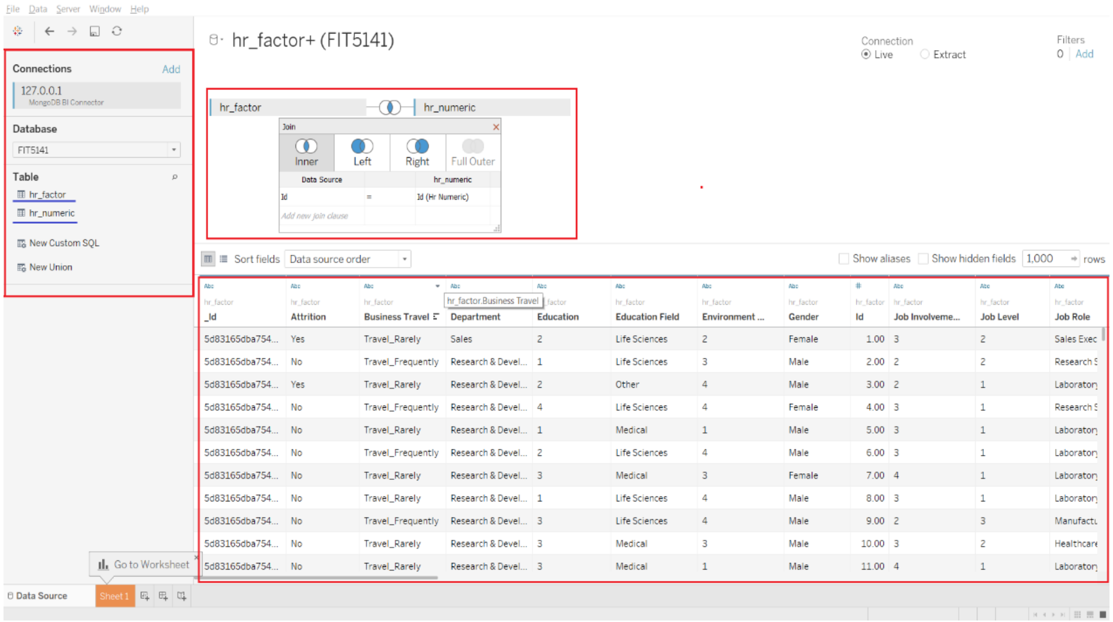
Enter information to sign in to the server:

Username:

Password:

□ Require SSL

After Tableau connects successful to MongoDB, select correct database which in this case is FIT5147, two tables or in other word, two collections of the database will be shown below the database name. Select two collections and join two collections by the **id** attribute.



hr_factor+ (FIT5141)

Connection: ● Live ○ Extract

Connections: 127.0.0.1 MongoDB BI Connector

Database: FIT5141

Table: hr_factor, hr_numeric

hr_factor — hr_numeric

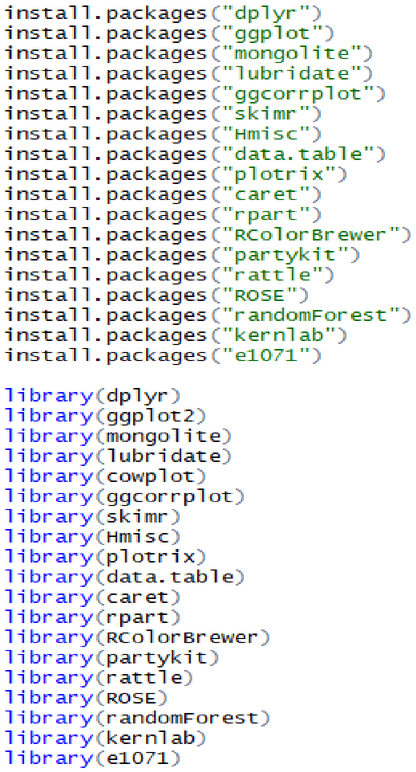
Data Source		hr_numeric
Id	=	Id (Hr Numeric)

Sort fields: Data source order

□ Show aliases □ Show hidden fields 1,000 rows

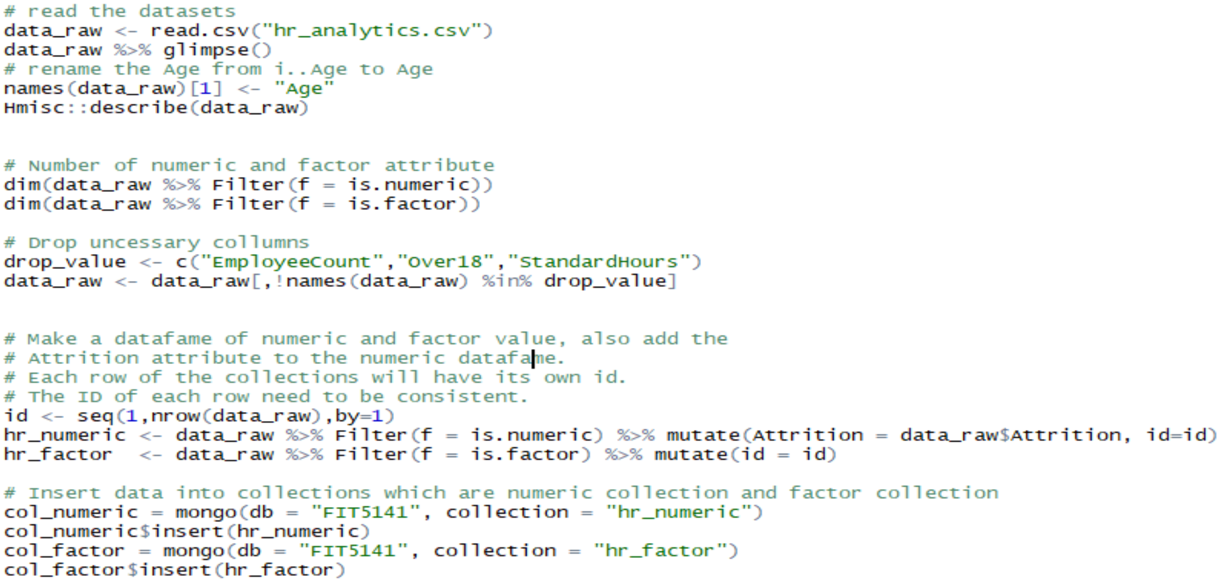
hr_factor _Id	hr_factor Attrition	hr_factor Business Travel	hr_factor Department	hr_factor Education	hr_factor Education Field	hr_factor Environment ...	hr_factor Gender	hr_factor Id	hr_factor Job Involveme...	hr_factor Job Level	hr_factor Job Role
5d83165dba754...	Yes	Travel_Rarely	Sales	2	Life Sciences	2	Female	1.00	3	2	Sales Exec
5d83165dba754...	No	Travel_Frequently	Research & Devel...	1	Life Sciences	3	Male	2.00	2	2	Research S
5d83165dba754...	Yes	Travel_Rarely	Research & Devel...	2	Other	4	Male	3.00	2	1	Laboratory
5d83165dba754...	No	Travel_Frequently	Research & Devel...	4	Life Sciences	4	Female	4.00	3	1	Research S
5d83165dba754...	No	Travel_Rarely	Research & Devel...	1	Medical	1	Male	5.00	3	1	Laboratory
5d83165dba754...	No	Travel_Frequently	Research & Devel...	2	Life Sciences	4	Male	6.00	3	1	Laboratory
5d83165dba754...	No	Travel_Rarely	Research & Devel...	3	Medical	3	Female	7.00	4	1	Laboratory
5d83165dba754...	No	Travel_Rarely	Research & Devel...	1	Life Sciences	4	Male	8.00	3	1	Laboratory
5d83165dba754...	No	Travel_Frequently	Research & Devel...	3	Life Sciences	4	Male	9.00	2	3	Manufactu
5d83165dba754...	No	Travel_Rarely	Research & Devel...	3	Medical	3	Male	10.00	3	2	Healthcare
5d83165dba754...	No	Travel_Rarely	Research & Devel...	3	Medical	1	Male	11.00	4	1	Laboratory

Install and import library



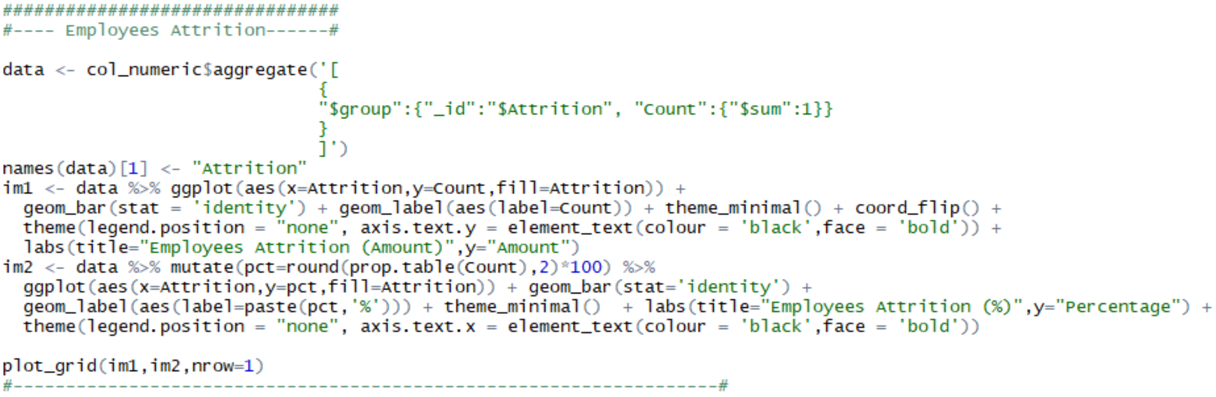
```
install.packages("dplyr")
install.packages("ggplot")
install.packages("mongolite")
install.packages("lubridate")
install.packages("ggcorrplot")
install.packages("skimr")
install.packages("Hmisc")
install.packages("data.table")
install.packages("plotrix")
install.packages("caret")
install.packages("rpart")
install.packages("RColorBrewer")
install.packages("partykit")
install.packages("rattle")
install.packages("ROSE")
install.packages("randomForest")
install.packages("kernlab")
install.packages("e1071")

library(dplyr)
library(ggplot2)
library(mongolite)
library(lubridate)
library(cowplot)
library(ggcorrplot)
library(skimr)
library(Hmisc)
library(plotrix)
library(data.table)
library(caret)
library(rpart)
library(RColorBrewer)
library(partykit)
library(rattle)
library(ROSE)
library(randomForest)
library(kernlab)
library(e1071)
```

Read the data, break into 2 collections and insert to MongoDB database



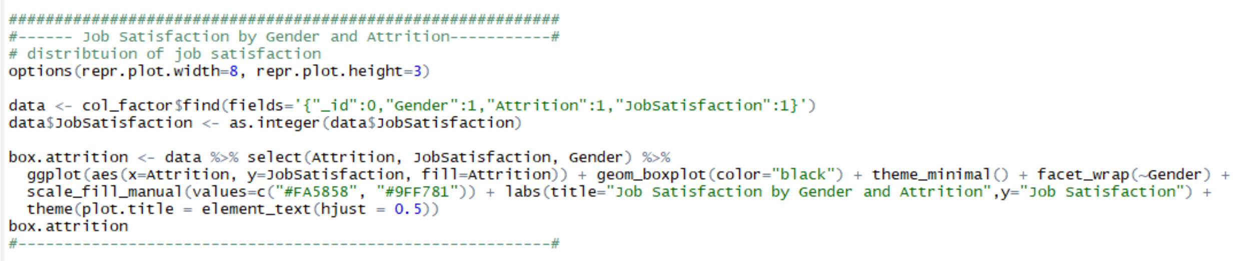
```
# read the datasets
data_raw <- read.csv("hr_analytics.csv")
data_raw %>% glimpse()
# rename the Age from i..Age to Age
names(data_raw)[1] <- "Age"
Hmisc::describe(data_raw)


# Number of numeric and factor attribute
dim(data_raw %>% Filter(f = is.numeric))
dim(data_raw %>% Filter(f = is.factor))

# Drop uncessary collumns
drop_value <- c("EmployeeCount","Over18","StandardHours")
data_raw <- data_raw[,!names(data_raw) %in% drop_value]


# Make a datafame of numeric and factor value, also add the
# Attrition attribute to the numeric datafame.
# Each row of the collections will have its own id.
# The ID of each row need to be consistent.
id <- seq(1,nrow(data_raw),by=1)
hr_numeric <- data_raw %>% Filter(f = is.numeric) %>% mutate(Attrition = data_raw$Attrition, id=id)
hr_factor  <- data_raw %>% Filter(f = is.factor) %>% mutate(id = id)

# Insert data into collections which are numeric collection and factor collection
col_numeric = mongo(db = "FIT5141", collection = "hr_numeric")
col_numeric$insert(hr_numeric)
col_factor = mongo(db = "FIT5141", collection = "hr_factor")
col_factor$insert(hr_factor)
```

Employee Attrition - Graph Number 2



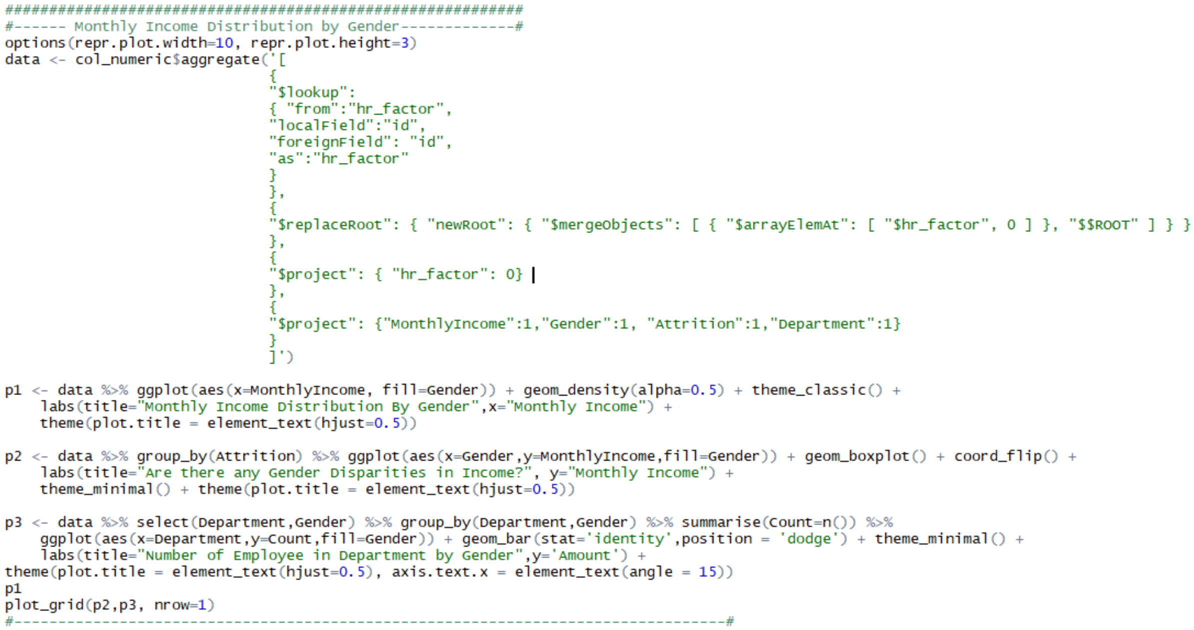
```
################################
#---- Employees Attrition------#

data <- col_numeric$aggregate('[
                              {
                              "$group":{"_id":"$Attrition", "Count":{"$sum":1}}
                              }
                              ]')
names(data)[1] <- "Attrition"
im1 <- data %>% ggplot(aes(x=Attrition,y=Count,fill=Attrition)) +
  geom_bar(stat = 'identity') + geom_label(aes(label=Count)) + theme_minimal() + coord_flip() +
  theme(legend.position = "none", axis.text.y = element_text(colour = 'black',face = 'bold')) +
  labs(title="Employees Attrition (Amount)",y="Amount")
im2 <- data %>% mutate(pct=round(prop.table(Count),2)*100) %>%
  ggplot(aes(x=Attrition,y=pct,fill=Attrition)) + geom_bar(stat='identity') +
  geom_label(aes(label=paste(pct,'%'))) + theme_minimal()  + labs(title="Employees Attrition (%)",y="Percentage") +
  theme(legend.position = "none", axis.text.x = element_text(colour = 'black',face = 'bold'))

plot_grid(im1,im2,nrow=1)
#---------------------------------------------------------------------#
```

Job Satisfaction and Attrition – Graph Number 6



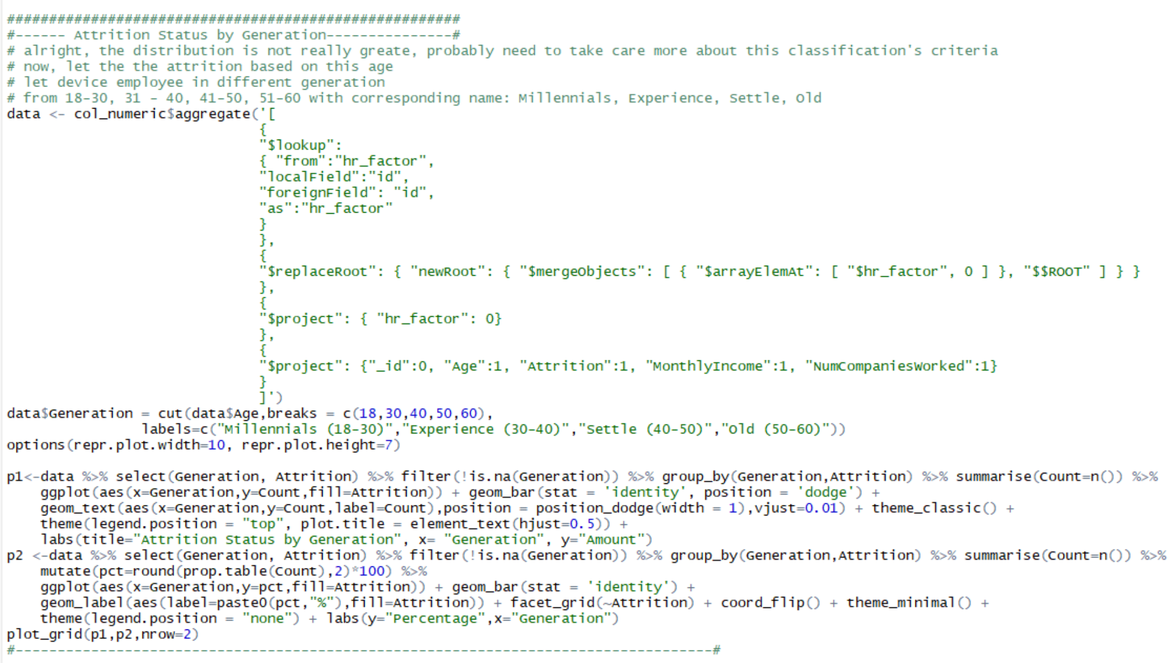
```
############################################################
#------ Job Satisfaction by Gender and Attrition-----------#
# distribtuion of job satisfaction
options(repr.plot.width=8, repr.plot.height=3)

data <- col_factor$find(fields='{"_id":0,"Gender":1,"Attrition":1,"JobSatisfaction":1}')
data$JobSatisfaction <- as.integer(data$JobSatisfaction)

box.attrition <- data %>% select(Attrition, JobSatisfaction, Gender) %>%
  ggplot(aes(x=Attrition, y=JobSatisfaction, fill=Attrition)) + geom_boxplot(color="black") + theme_minimal() + facet_wrap(~Gender) +
  scale_fill_manual(values=c("#FA5858", "#9FF781")) + labs(title="Job Satisfaction by Gender and Attrition",y="Job Satisfaction") +
  theme(plot.title = element_text(hjust = 0.5))
box.attrition
#---------------------------------------------------------#
```

Monthly Income Distribution by Gender – Graph Number 7



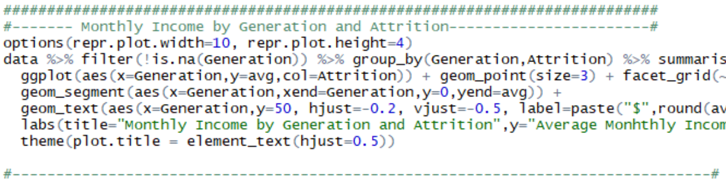
```
##########################################################
#------ Monthly Income Distribution by Gender-------------#
options(repr.plot.width=10, repr.plot.height=3)
data <- col_numeric$aggregate('[
                              {
                              "$lookup":
                              { "from":"hr_factor",
                              "localField":"id",
                              "foreignField": "id",
                              "as":"hr_factor"
                              }
                              },
                              {
                              "$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }
                              },
                              {
                              "$project": { "hr_factor": 0} |
                              },
                              {
                              "$project": {"MonthlyIncome":1,"Gender":1, "Attrition":1,"Department":1}
                              }
                              ]')

p1 <- data %>% ggplot(aes(x=MonthlyIncome, fill=Gender)) + geom_density(alpha=0.5) + theme_classic() +
    labs(title="Monthly Income Distribution By Gender",x="Monthly Income") +
    theme(plot.title = element_text(hjust=0.5))

p2 <- data %>% group_by(Attrition) %>% ggplot(aes(x=Gender,y=MonthlyIncome,fill=Gender)) + geom_boxplot() + coord_flip() +
    labs(title="Are there any Gender Disparities in Income?", y="Monthly Income") +
    theme_minimal() + theme(plot.title = element_text(hjust=0.5))

p3 <- data %>% select(Department,Gender) %>% group_by(Department,Gender) %>% summarise(Count=n()) %>%
    ggplot(aes(x=Department,y=Count,fill=Gender)) + geom_bar(stat='identity',position = 'dodge') + theme_minimal() +
    labs(title="Number of Employee in Department by Gender",y='Amount') +
theme(plot.title = element_text(hjust=0.5), axis.text.x = element_text(angle = 15))
p1
plot_grid(p2,p3, nrow=1)
#------------------------------------------------------------------------------#
```

Attrition Status by Gender – Graph number 8

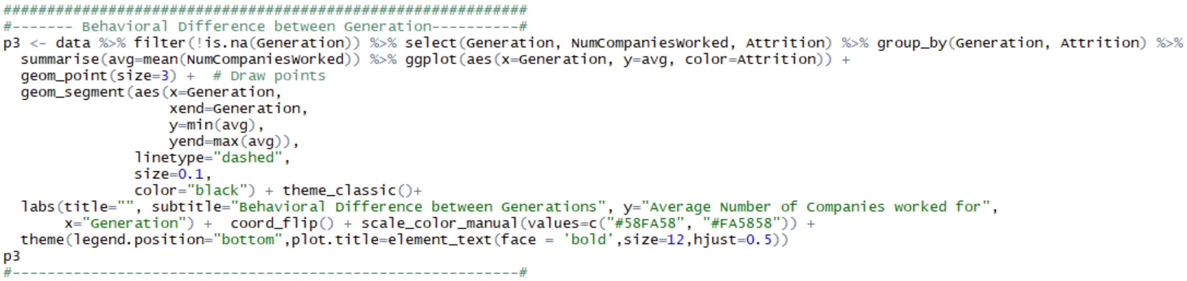


```
######################################################
#------ Attrition Status by Generation---------------#
# alright, the distribution is not really greate, probably need to take care more about this classification's criteria
# now, let the the attrition based on this age
# let device employee in different generation
# from 18-30, 31 - 40, 41-50, 51-60 with corresponding name: Millennials, Experience, Settle, Old
data <- col_numeric$aggregate('[
                              {
                              "$lookup":
                              { "from":"hr_factor",
                              "localField":"id",
                              "foreignField": "id",
                              "as":"hr_factor"
                              }
                              },
                              {
                              "$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }
                              },
                              {
                              "$project": { "hr_factor": 0}
                              },
                              {
                              "$project": {"_id":0, "Age":1, "Attrition":1, "MonthlyIncome":1, "NumCompaniesWorked":1}
                              }
                              ]')
data$Generation = cut(data$Age,breaks = c(18,30,40,50,60),
                labels=c("Millennials (18-30)","Experience (30-40)","Settle (40-50)","Old (50-60)"))
options(repr.plot.width=10, repr.plot.height=7)

p1<-data %>% select(Generation, Attrition) %>% filter(!is.na(Generation)) %>% group_by(Generation,Attrition) %>% summarise(Count=n()) %>%
    ggplot(aes(x=Generation,y=Count,fill=Attrition)) + geom_bar(stat = 'identity', position = 'dodge') +
    geom_text(aes(x=Generation,y=Count,label=Count),position = position_dodge(width = 1),vjust=0.01) + theme_classic() +
    theme(legend.position = "top", plot.title = element_text(hjust=0.5)) +
    labs(title="Attrition Status by Generation", x= "Generation", y="Amount")
p2 <-data %>% select(Generation, Attrition) %>% filter(!is.na(Generation)) %>% group_by(Generation,Attrition) %>% summarise(Count=n()) %>%
    mutate(pct=round(prop.table(Count),2)*100) %>%
    ggplot(aes(x=Generation,y=pct,fill=Attrition)) + geom_bar(stat = 'identity') +
    geom_label(aes(label=paste0(pct,"%"),fill=Attrition)) + facet_grid(~Attrition) + coord_flip() + theme_minimal() +
    theme(legend.position = "none") + labs(y="Percentage",x="Generation")
plot_grid(p1,p2,nrow=2)
#-----------------------------------------------------------------------------------#
```

Monthly Income by Generation and behavioral difference between generation – Graph 9 and 10

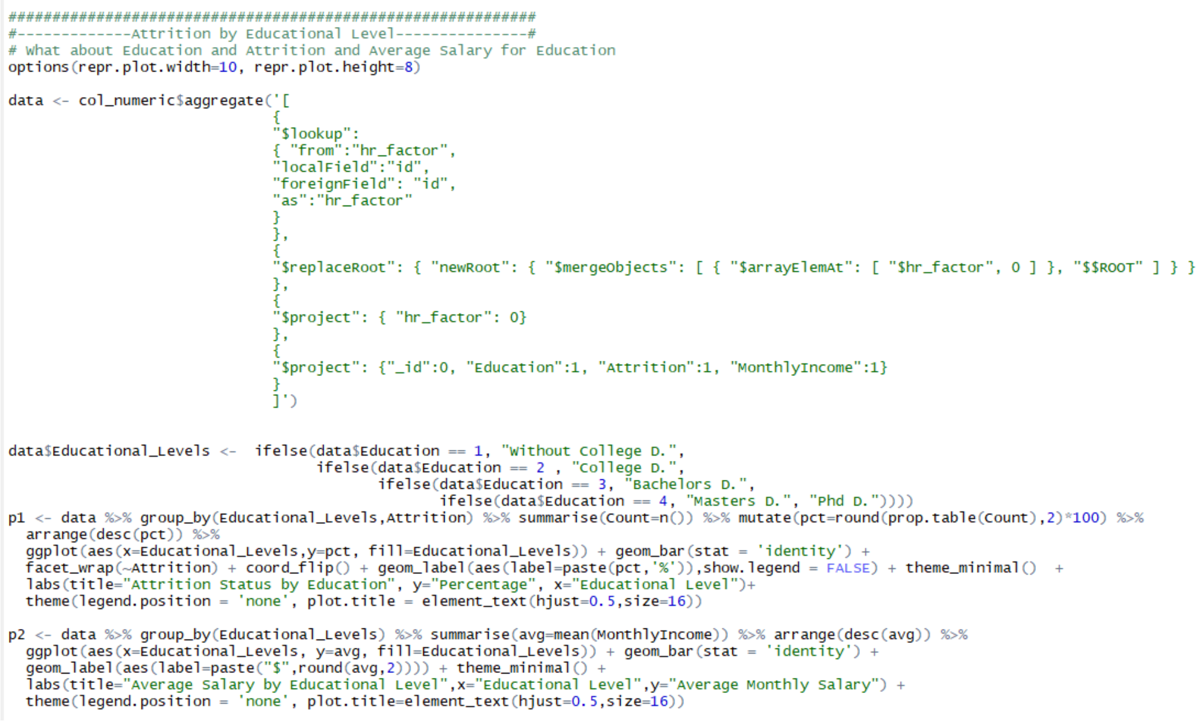


```
##########################################################################
#------- Monthly Income by Generation and Attrition----------------------#
options(repr.plot.width=10, repr.plot.height=4)
data %>% filter(!is.na(Generation)) %>% group_by(Generation,Attrition) %>% summarise(avg=mean(MonthlyIncome)) %>%
  ggplot(aes(x=Generation,y=avg,col=Attrition)) + geom_point(size=3) + facet_grid(~Attrition) + coord_flip() + theme_minimal() +
  geom_segment(aes(x=Generation,xend=Generation,y=0,yend=avg)) +
  geom_text(aes(x=Generation,y=50, hjust=-0.2, vjust=-0.5, label=paste("$",round(avg,2)))) +
  labs(title="Monthly Income by Generation and Attrition",y="Average Monhthly Income") +
  theme(plot.title = element_text(hjust=0.5))

#-------------------------------------------------------------------------------#
```



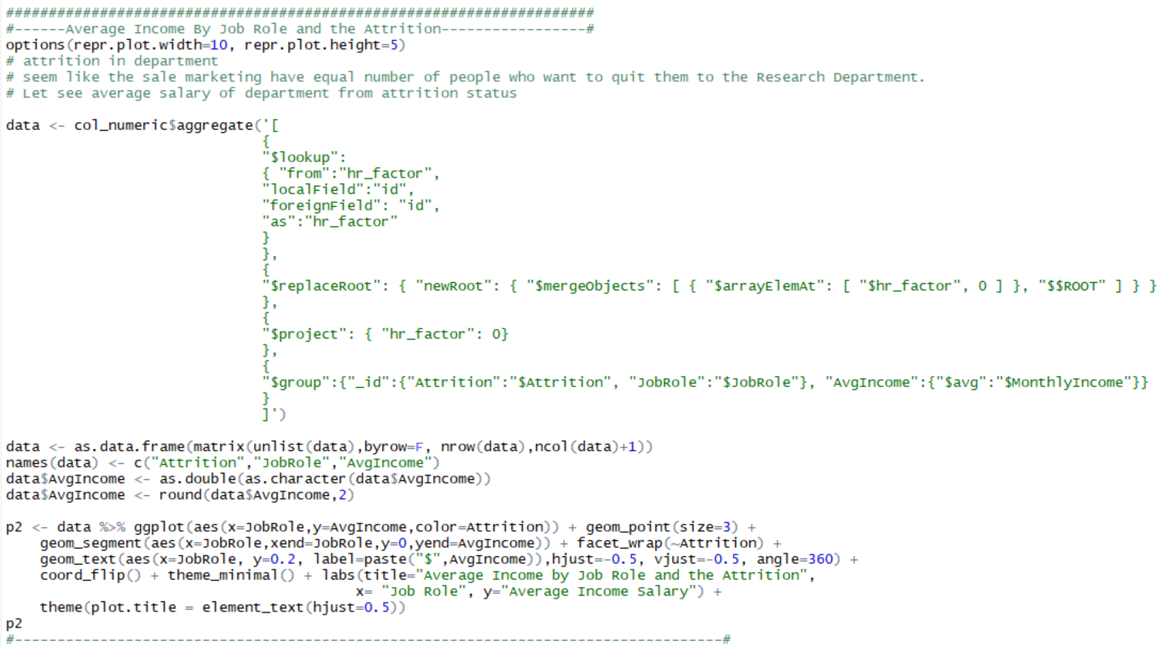
```
############################################################
#------- Behavioral Difference between Generation----------#
p3 <- data %>% filter(!is.na(Generation)) %>% select(Generation, NumCompaniesWorked, Attrition) %>% group_by(Generation, Attrition) %>%
  summarise(avg=mean(NumCompaniesWorked)) %>% ggplot(aes(x=Generation, y=avg, color=Attrition)) +
  geom_point(size=3) +  # Draw points
  geom_segment(aes(x=Generation,
                   xend=Generation,
                   y=min(avg),
                   yend=max(avg)),
               linetype="dashed",
               size=0.1,
               color="black") + theme_classic()+
  labs(title="", subtitle="Behavioral Difference between Generations", y="Average Number of Companies worked for",
       x="Generation") +  coord_flip() + scale_color_manual(values=c("#58FA58", "#FA5858")) +
  theme(legend.position="bottom",plot.title=element_text(face = 'bold',size=12,hjust=0.5))
p3
#----------------------------------------------------------#
```

Attrition by Educational Levels – Graph Number 12



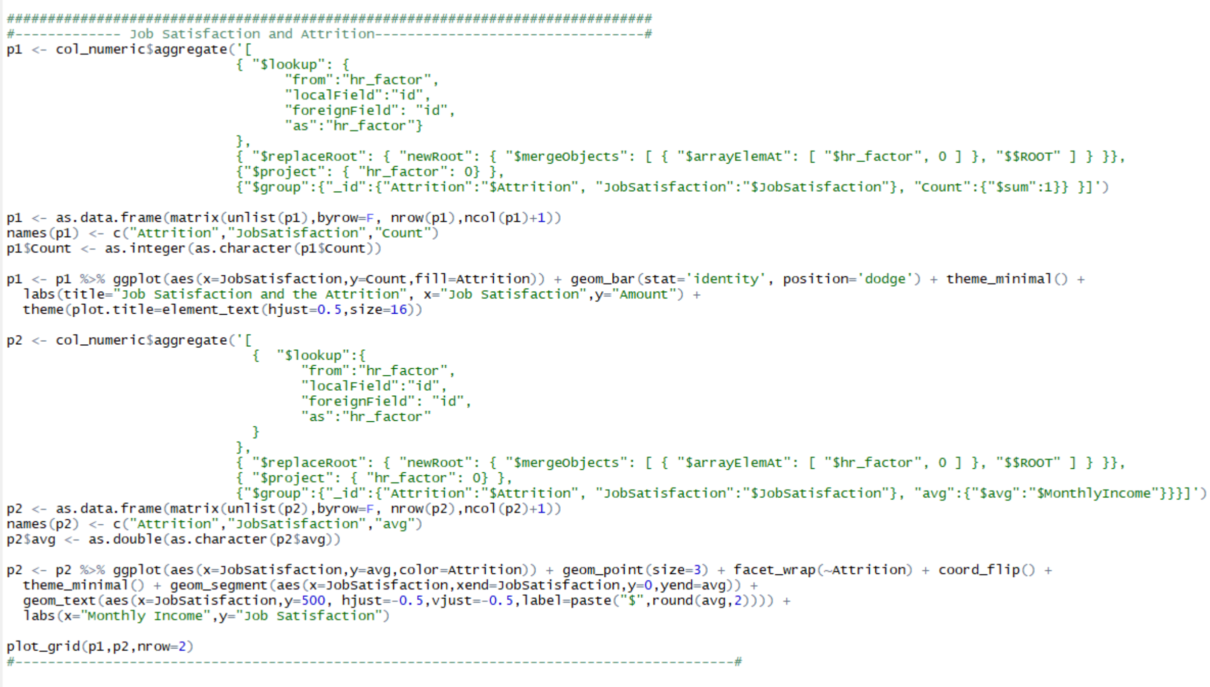
```
############################################################
#-------------Attrition by Educational Level---------------#
# What about Education and Attrition and Average Salary for Education
options(repr.plot.width=10, repr.plot.height=8)

data <- col_numeric$aggregate('[
                               {
                               "$lookup":
                               { "from":"hr_factor",
                               "localField":"id",
                               "foreignField": "id",
                               "as":"hr_factor"
                               }
                               },
                               {
                               "$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }
                               },
                               {
                               "$project": { "hr_factor": 0}
                               },
                               {
                               "$project": {"_id":0, "Education":1, "Attrition":1, "MonthlyIncome":1}
                               }
                               ]')


data$Educational_Levels <-  ifelse(data$Education == 1, "Without College D.",
                                   ifelse(data$Education == 2 , "College D.",
                                          ifelse(data$Education == 3, "Bachelors D.",
                                                 ifelse(data$Education == 4, "Masters D.", "Phd D."))))
p1 <- data %>% group_by(Educational_Levels,Attrition) %>% summarise(Count=n()) %>% mutate(pct=round(prop.table(Count),2)*100) %>%
  arrange(desc(pct)) %>%
  ggplot(aes(x=Educational_Levels,y=pct, fill=Educational_Levels)) + geom_bar(stat = 'identity') +
  facet_wrap(~Attrition) + coord_flip() + geom_label(aes(label=paste(pct,'%')),show.legend = FALSE) + theme_minimal()  +
  labs(title="Attrition Status by Education", y="Percentage", x="Educational Level")+
  theme(legend.position = 'none', plot.title = element_text(hjust=0.5,size=16))

p2 <- data %>% group_by(Educational_Levels) %>% summarise(avg=mean(MonthlyIncome)) %>% arrange(desc(avg)) %>%
  ggplot(aes(x=Educational_Levels, y=avg, fill=Educational_Levels)) + geom_bar(stat = 'identity') +
  geom_label(aes(label=paste("$",round(avg,2)))) + theme_minimal() +
  labs(title="Average Salary by Educational Level",x="Educational Level",y="Average Monthly Salary") +
  theme(legend.position = 'none', plot.title=element_text(hjust=0.5,size=16))
```

Average Income by Job Role and Attrition – Graph Number 16



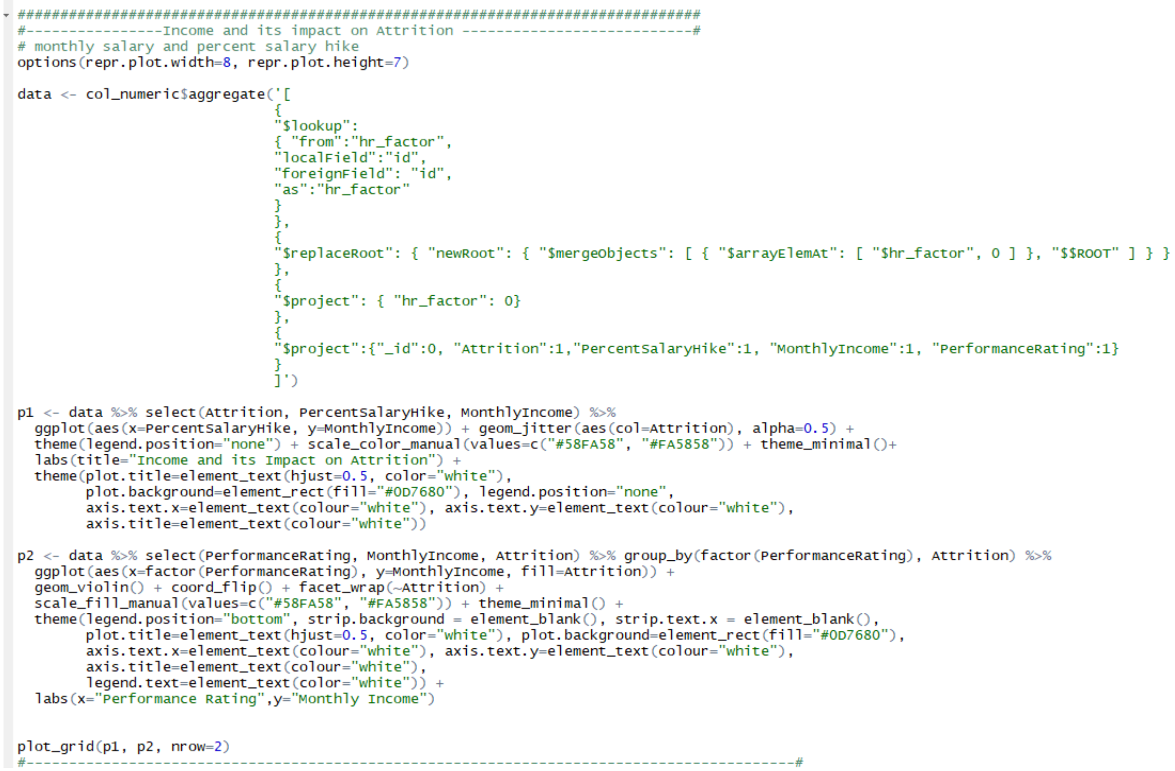
```
#####################################################################
#------Average Income By Job Role and the Attrition-----------------#
options(repr.plot.width=10, repr.plot.height=5)
# attrition in department
# seem like the sale marketing have equal number of people who want to quit them to the Research Department.
# Let see average salary of department from attrition status

data <- col_numeric$aggregate('[
                              {
                              "$lookup":
                              { "from":"hr_factor",
                              "localField":"id",
                              "foreignField": "id",
                              "as":"hr_factor"
                              }
                              },
                              {
                              "$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }
                              },
                              {
                              "$project": { "hr_factor": 0}
                              },
                              {
                              "$group":{"_id":{"Attrition":"$Attrition", "JobRole":"$JobRole"}, "AvgIncome":{"$avg":"$MonthlyIncome"}}
                              }
                              ]')

data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","JobRole","AvgIncome")
data$AvgIncome <- as.double(as.character(data$AvgIncome))
data$AvgIncome <- round(data$AvgIncome,2)

p2 <- data %>% ggplot(aes(x=JobRole,y=AvgIncome,color=Attrition)) + geom_point(size=3) +
    geom_segment(aes(x=JobRole,xend=JobRole,y=0,yend=AvgIncome)) + facet_wrap(~Attrition) +
    geom_text(aes(x=JobRole, y=0.2, label=paste("$",AvgIncome)),hjust=-0.5, vjust=-0.5, angle=360) +
    coord_flip() + theme_minimal() + labs(title="Average Income by Job Role and the Attrition",
                                          x= "Job Role", y="Average Income Salary") +
    theme(plot.title = element_text(hjust=0.5))
p2
#-----------------------------------------------------------------------------------#
```

Job Satisfaction and Attrition – Graph Number 17



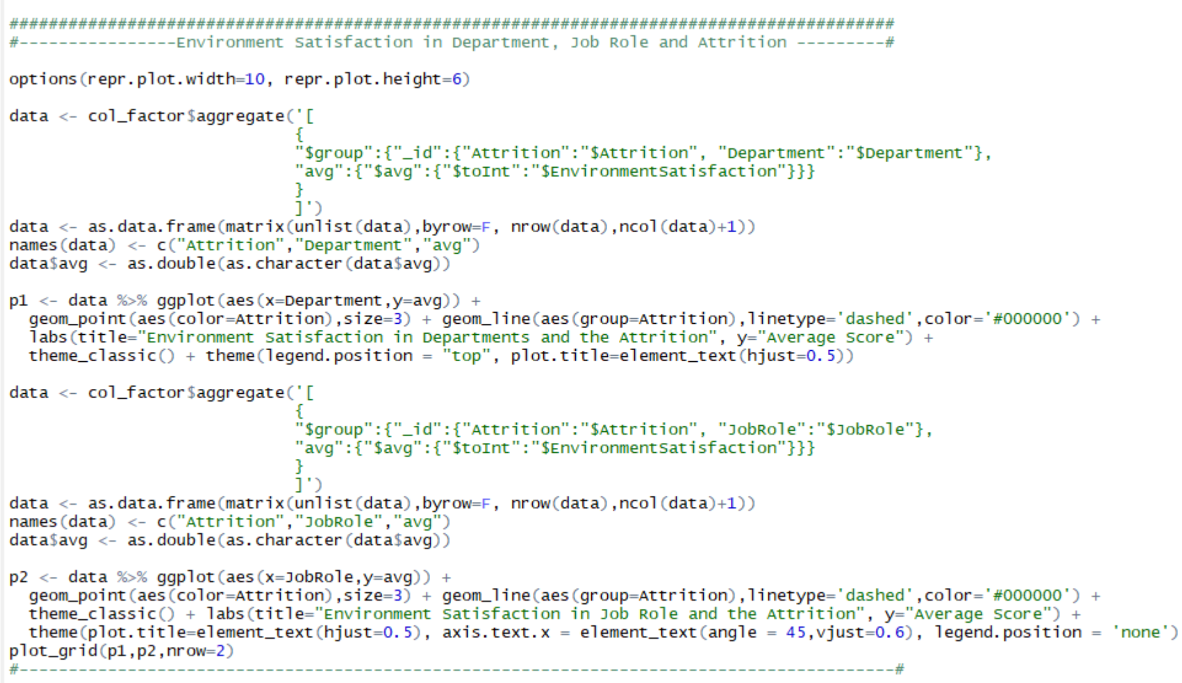
```
################################################################################
#------------- Job Satisfaction and Attrition---------------------------------#
p1 <- col_numeric$aggregate('[
                             { "$lookup": {
                                   "from":"hr_factor",
                                   "localField":"id",
                                   "foreignField": "id",
                                   "as":"hr_factor"}
                             },
                             { "$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }},
                             {"$project": { "hr_factor": 0} },
                             {"$group":{"_id":{"Attrition":"$Attrition", "JobSatisfaction":"$JobSatisfaction"}, "Count":{"$sum":1}} }]')

p1 <- as.data.frame(matrix(unlist(p1),byrow=F, nrow(p1),ncol(p1)+1))
names(p1) <- c("Attrition","JobSatisfaction","Count")
p1$Count <- as.integer(as.character(p1$Count))

p1 <- p1 %>% ggplot(aes(x=JobSatisfaction,y=Count,fill=Attrition)) + geom_bar(stat='identity', position='dodge') + theme_minimal() +
  labs(title="Job Satisfaction and the Attrition", x="Job Satisfaction",y="Amount") +
  theme(plot.title=element_text(hjust=0.5,size=16))

p2 <- col_numeric$aggregate('[
                               {  "$lookup":{
                                     "from":"hr_factor",
                                     "localField":"id",
                                     "foreignField": "id",
                                     "as":"hr_factor"
                               }
                             },
                             { "$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }},
                             { "$project": { "hr_factor": 0} },
                             {"$group":{"_id":{"Attrition":"$Attrition", "JobSatisfaction":"$JobSatisfaction"}, "avg":{"$avg":"$MonthlyIncome"}}}]')
p2 <- as.data.frame(matrix(unlist(p2),byrow=F, nrow(p2),ncol(p2)+1))
names(p2) <- c("Attrition","JobSatisfaction","avg")
p2$avg <- as.double(as.character(p2$avg))

p2 <- p2 %>% ggplot(aes(x=JobSatisfaction,y=avg,color=Attrition)) + geom_point(size=3) + facet_wrap(~Attrition) + coord_flip() +
  theme_minimal() + geom_segment(aes(x=JobSatisfaction,xend=JobSatisfaction,y=0,yend=avg)) +
  geom_text(aes(x=JobSatisfaction,y=500, hjust=-0.5,vjust=-0.5,label=paste("$",round(avg,2)))) +
  labs(x="Monthly Income",y="Job Satisfaction")

plot_grid(p1,p2,nrow=2)
#-------------------------------------------------------------------------------------------#
```

Income and its impact on Attrition – Graph Number 18



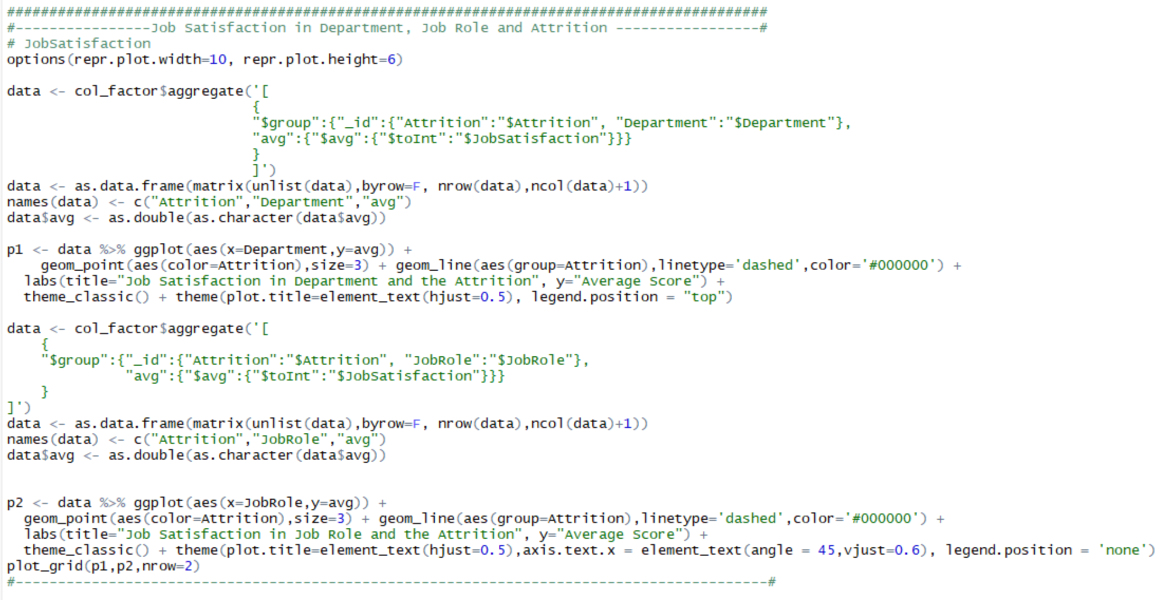
```
################################################################################
#----------------Income and its impact on Attrition ---------------------------#
# monthly salary and percent salary hike
options(repr.plot.width=8, repr.plot.height=7)

data <- col_numeric$aggregate('[
                               {
                               "$lookup":
                               { "from":"hr_factor",
                               "localField":"id",
                               "foreignField": "id",
                               "as":"hr_factor"
                               }
                               },
                               {
                               "$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }
                               },
                               {
                               "$project": { "hr_factor": 0}
                               },
                               {
                               "$project":{"_id":0, "Attrition":1,"PercentSalaryHike":1, "MonthlyIncome":1, "PerformanceRating":1}
                               }
                               ]')

p1 <- data %>% select(Attrition, PercentSalaryHike, MonthlyIncome) %>%
  ggplot(aes(x=PercentSalaryHike, y=MonthlyIncome)) + geom_jitter(aes(col=Attrition), alpha=0.5) +
  theme(legend.position="none") + scale_color_manual(values=c("#58FA58", "#FA5858")) + theme_minimal()+
  labs(title="Income and its Impact on Attrition") +
  theme(plot.title=element_text(hjust=0.5, color="white"),
        plot.background=element_rect(fill="#0D7680"), legend.position="none",
        axis.text.x=element_text(colour="white"), axis.text.y=element_text(colour="white"),
        axis.title=element_text(colour="white"))

p2 <- data %>% select(PerformanceRating, MonthlyIncome, Attrition) %>% group_by(factor(PerformanceRating), Attrition) %>%
  ggplot(aes(x=factor(PerformanceRating), y=MonthlyIncome, fill=Attrition)) +
  geom_violin() + coord_flip() + facet_wrap(~Attrition) +
  scale_fill_manual(values=c("#58FA58", "#FA5858")) + theme_minimal() +
  theme(legend.position="bottom", strip.background = element_blank(), strip.text.x = element_blank(),
        plot.title=element_text(hjust=0.5, color="white"), plot.background=element_rect(fill="#0D7680"),
        axis.text.x=element_text(colour="white"), axis.text.y=element_text(colour="white"),
        axis.title=element_text(colour="white"),
        legend.text=element_text(color="white")) +
  labs(x="Performance Rating",y="Monthly Income")


plot_grid(p1, p2, nrow=2)
#-----------------------------------------------------------------------------------------#
```

Environment Satisfaction in Department, Job Role and Attrition – Graph Number 21



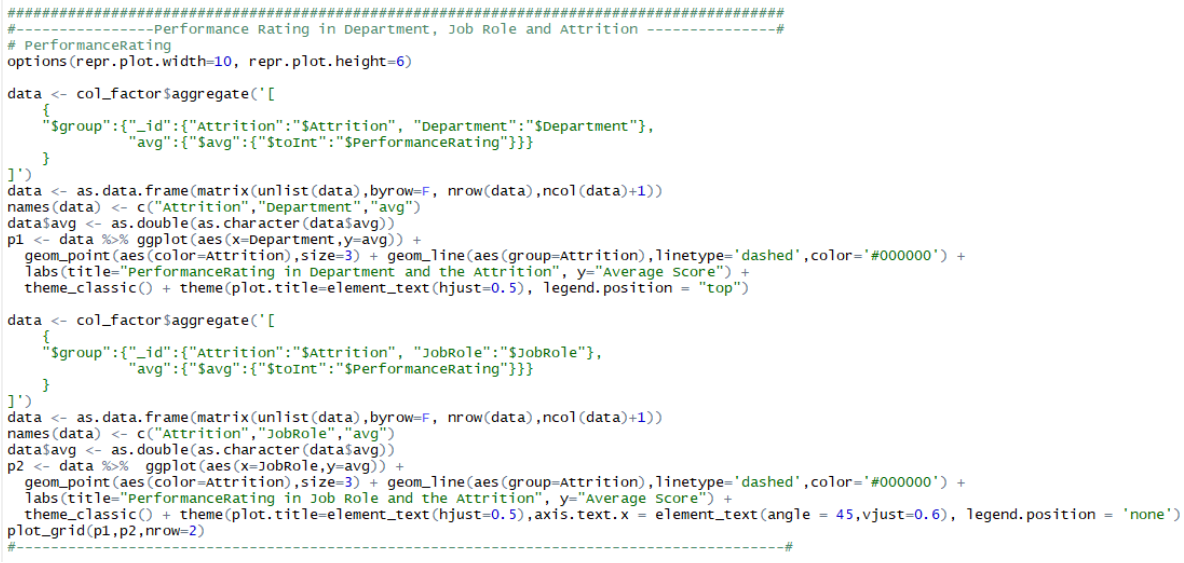
```
##########################################################################################
#----------------Environment Satisfaction in Department, Job Role and Attrition ---------#

options(repr.plot.width=10, repr.plot.height=6)

data <- col_factor$aggregate('[
                             {
                             "$group":{"_id":{"Attrition":"$Attrition", "Department":"$Department"},
                             "avg":{"$avg":{"$toInt":"$EnvironmentSatisfaction"}}}
                             }
                             ]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","Department","avg")
data$avg <- as.double(as.character(data$avg))

p1 <- data %>% ggplot(aes(x=Department,y=avg)) +
  geom_point(aes(color=Attrition),size=3) + geom_line(aes(group=Attrition),linetype='dashed',color='#000000') +
  labs(title="Environment Satisfaction in Departments and the Attrition", y="Average Score") +
  theme_classic() + theme(legend.position = "top", plot.title=element_text(hjust=0.5))

data <- col_factor$aggregate('[
                             {
                             "$group":{"_id":{"Attrition":"$Attrition", "JobRole":"$JobRole"},
                             "avg":{"$avg":{"$toInt":"$EnvironmentSatisfaction"}}}
                             }
                             ]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","JobRole","avg")
data$avg <- as.double(as.character(data$avg))

p2 <- data %>% ggplot(aes(x=JobRole,y=avg)) +
  geom_point(aes(color=Attrition),size=3) + geom_line(aes(group=Attrition),linetype='dashed',color='#000000') +
  theme_classic() + labs(title="Environment Satisfaction in Job Role and the Attrition", y="Average Score") +
  theme(plot.title=element_text(hjust=0.5), axis.text.x = element_text(angle = 45,vjust=0.6), legend.position = 'none')
plot_grid(p1,p2,nrow=2)
#----------------------------------------------------------------------------------------#
```

Job Satisfaction in Department, Job Role and Attrition – Graph Number 22



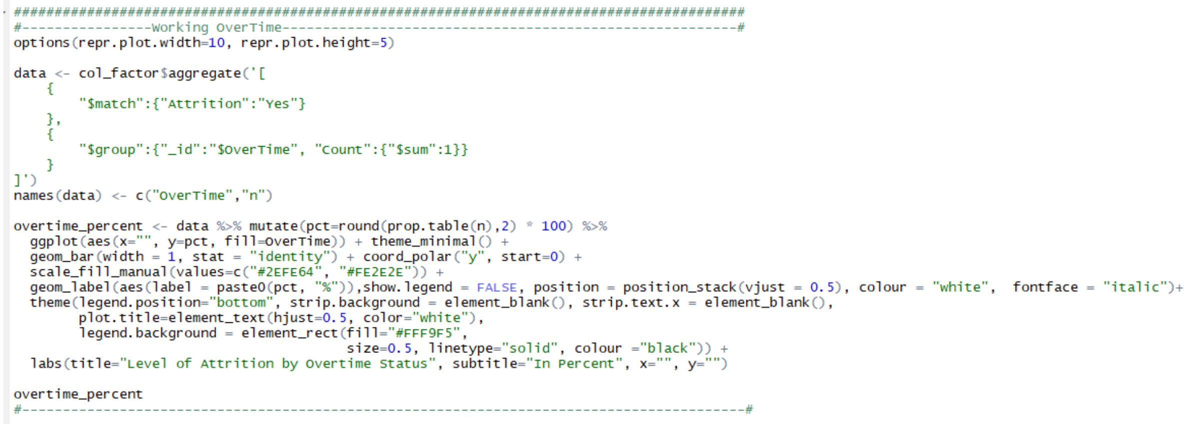
```
#########################################################################################
#----------------Job Satisfaction in Department, Job Role and Attrition ----------------#
# JobSatisfaction
options(repr.plot.width=10, repr.plot.height=6)

data <- col_factor$aggregate('[
                             {
                             "$group":{"_id":{"Attrition":"$Attrition", "Department":"$Department"},
                             "avg":{"$avg":{"$toInt":"$JobSatisfaction"}}}
                             }
                             ]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","Department","avg")
data$avg <- as.double(as.character(data$avg))

p1 <- data %>% ggplot(aes(x=Department,y=avg)) +
    geom_point(aes(color=Attrition),size=3) + geom_line(aes(group=Attrition),linetype='dashed',color='#000000') +
  labs(title="Job Satisfaction in Department and the Attrition", y="Average Score") +
  theme_classic() + theme(plot.title=element_text(hjust=0.5), legend.position = "top")

data <- col_factor$aggregate('[
    {
    "$group":{"_id":{"Attrition":"$Attrition", "JobRole":"$JobRole"},
              "avg":{"$avg":{"$toInt":"$JobSatisfaction"}}}
    }
]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","JobRole","avg")
data$avg <- as.double(as.character(data$avg))


p2 <- data %>% ggplot(aes(x=JobRole,y=avg)) +
  geom_point(aes(color=Attrition),size=3) + geom_line(aes(group=Attrition),linetype='dashed',color='#000000') +
  labs(title="Job Satisfaction in Job Role and the Attrition", y="Average Score") +
  theme_classic() + theme(plot.title=element_text(hjust=0.5),axis.text.x = element_text(angle = 45,vjust=0.6), legend.position = 'none')
plot_grid(p1,p2,nrow=2)
#---------------------------------------------------------------------------------------#
```

Performance Rating in Department, Job Role and Attrition - Graph Number 23



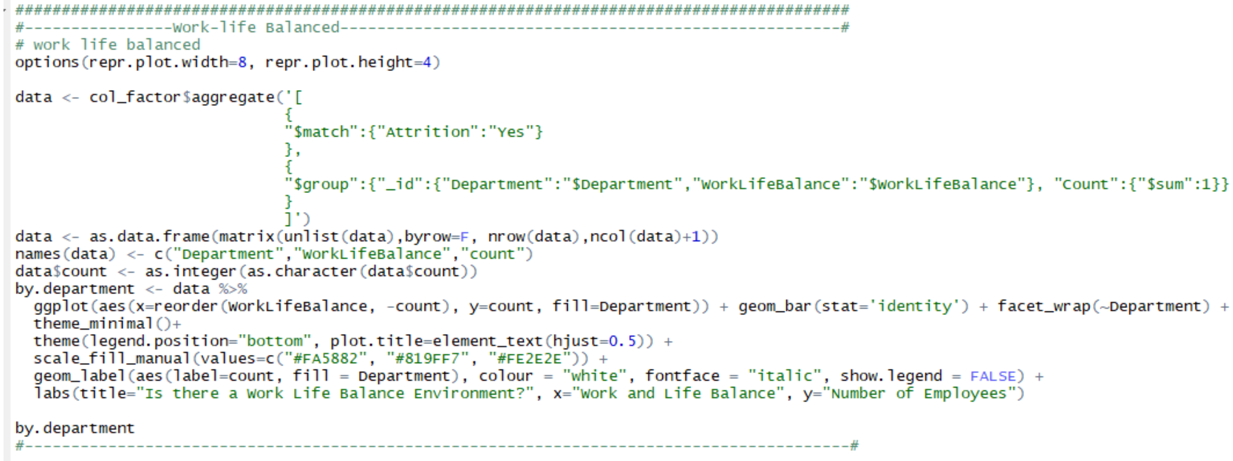
```
##########################################################################################
#----------------Performance Rating in Department, Job Role and Attrition ---------------#
# PerformanceRating
options(repr.plot.width=10, repr.plot.height=6)

data <- col_factor$aggregate('[
    {
    "$group":{"_id":{"Attrition":"$Attrition", "Department":"$Department"},
              "avg":{"$avg":{"$toInt":"$PerformanceRating"}}}
    }
]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","Department","avg")
data$avg <- as.double(as.character(data$avg))
p1 <- data %>% ggplot(aes(x=Department,y=avg)) +
  geom_point(aes(color=Attrition),size=3) + geom_line(aes(group=Attrition),linetype='dashed',color='#000000') +
  labs(title="PerformanceRating in Department and the Attrition", y="Average Score") +
  theme_classic() + theme(plot.title=element_text(hjust=0.5), legend.position = "top")

data <- col_factor$aggregate('[
    {
    "$group":{"_id":{"Attrition":"$Attrition", "JobRole":"$JobRole"},
              "avg":{"$avg":{"$toInt":"$PerformanceRating"}}}
    }
]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","JobRole","avg")
data$avg <- as.double(as.character(data$avg))
p2 <- data %>%  ggplot(aes(x=JobRole,y=avg)) +
  geom_point(aes(color=Attrition),size=3) + geom_line(aes(group=Attrition),linetype='dashed',color='#000000') +
  labs(title="PerformanceRating in Job Role and the Attrition", y="Average Score") +
  theme_classic() + theme(plot.title=element_text(hjust=0.5),axis.text.x = element_text(angle = 45,vjust=0.6), legend.position = 'none')
plot_grid(p1,p2,nrow=2)
#----------------------------------------------------------------------------------------#
```

Working Overtime – Graph Number 24



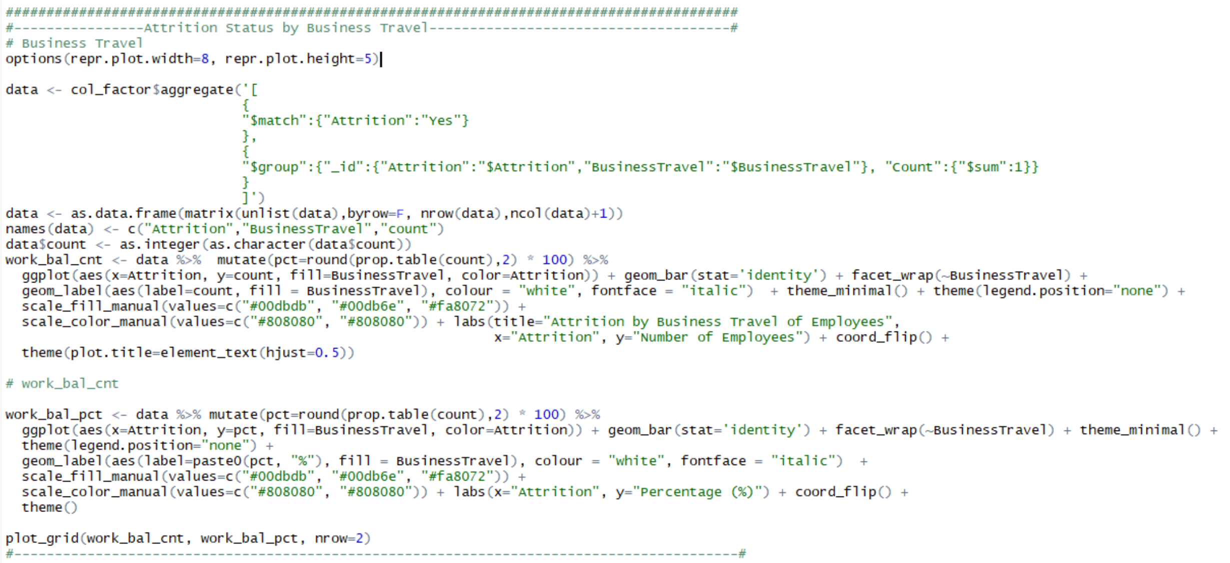
```
#########################################################################################
#----------------Working OverTime--------------------------------------------------------#
options(repr.plot.width=10, repr.plot.height=5)

data <- col_factor$aggregate('[
    {
        "$match":{"Attrition":"Yes"}
    },
    {
        "$group":{"_id":"$OverTime", "Count":{"$sum":1}}
    }
]')
names(data) <- c("OverTime","n")

overtime_percent <- data %>% mutate(pct=round(prop.table(n),2) * 100) %>%
  ggplot(aes(x="", y=pct, fill=OverTime)) + theme_minimal() +
  geom_bar(width = 1, stat = "identity") + coord_polar("y", start=0) +
  scale_fill_manual(values=c("#2EFE64", "#FE2E2E")) +
  geom_label(aes(label = paste0(pct, "%")),show.legend = FALSE, position = position_stack(vjust = 0.5), colour = "white",  fontface = "italic")+
  theme(legend.position="bottom", strip.background = element_blank(), strip.text.x = element_blank(),
        plot.title=element_text(hjust=0.5, color="white"),
        legend.background = element_rect(fill="#FFF9F5",
                                         size=0.5, linetype="solid", colour ="black")) +
  labs(title="Level of Attrition by Overtime Status", subtitle="In Percent", x="", y="")

overtime_percent
#-----------------------------------------------------------------------------------------#
```

Work-life balance – Graph Number 26



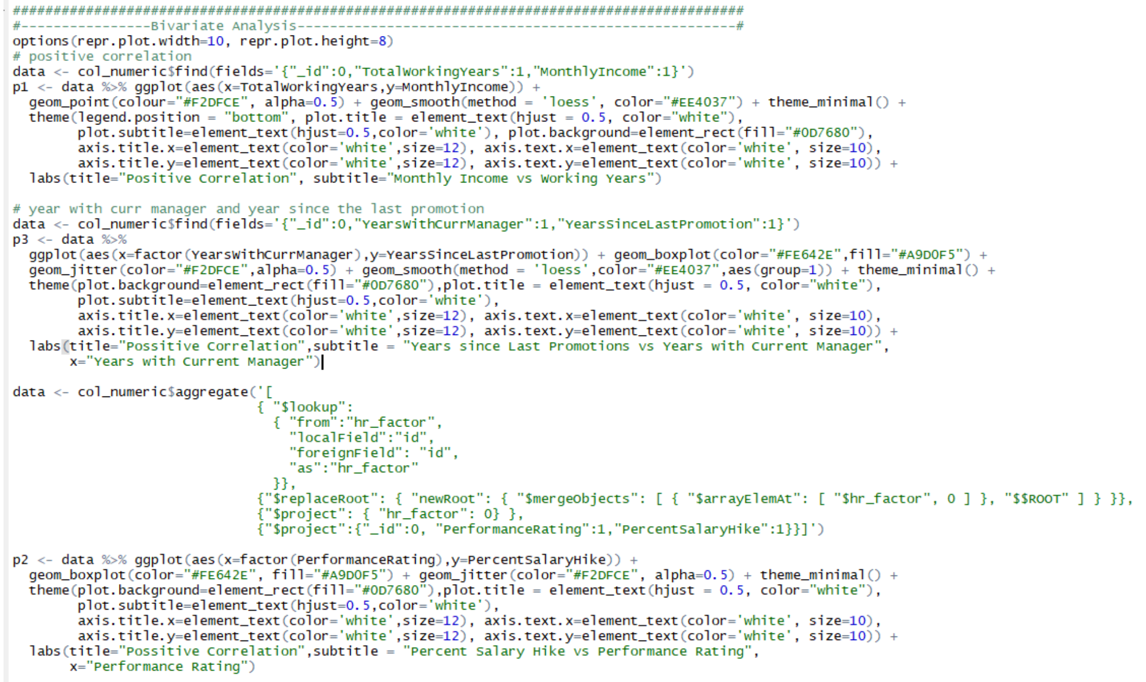
```
##########################################################################################
#----------------Work-life Balanced------------------------------------------------------#
# work life balanced
options(repr.plot.width=8, repr.plot.height=4)

data <- col_factor$aggregate('[
                              {
                              "$match":{"Attrition":"Yes"}
                              },
                              {
                              "$group":{"_id":{"Department":"$Department","WorkLifeBalance":"$WorkLifeBalance"}, "Count":{"$sum":1}}
                              }
                              ]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Department","WorkLifeBalance","count")
data$count <- as.integer(as.character(data$count))
by.department <- data %>%
  ggplot(aes(x=reorder(WorkLifeBalance, -count), y=count, fill=Department)) + geom_bar(stat='identity') + facet_wrap(~Department) +
  theme_minimal()+
  theme(legend.position="bottom", plot.title=element_text(hjust=0.5)) +
  scale_fill_manual(values=c("#FA5882", "#819FF7", "#FE2E2E")) +
  geom_label(aes(label=count, fill = Department), colour = "white", fontface = "italic", show.legend = FALSE) +
  labs(title="Is there a Work Life Balance Environment?", x="Work and Life Balance", y="Number of Employees")

by.department
#-----------------------------------------------------------------------------------------#
```

Attrition and Business Travel – Graph Number 27



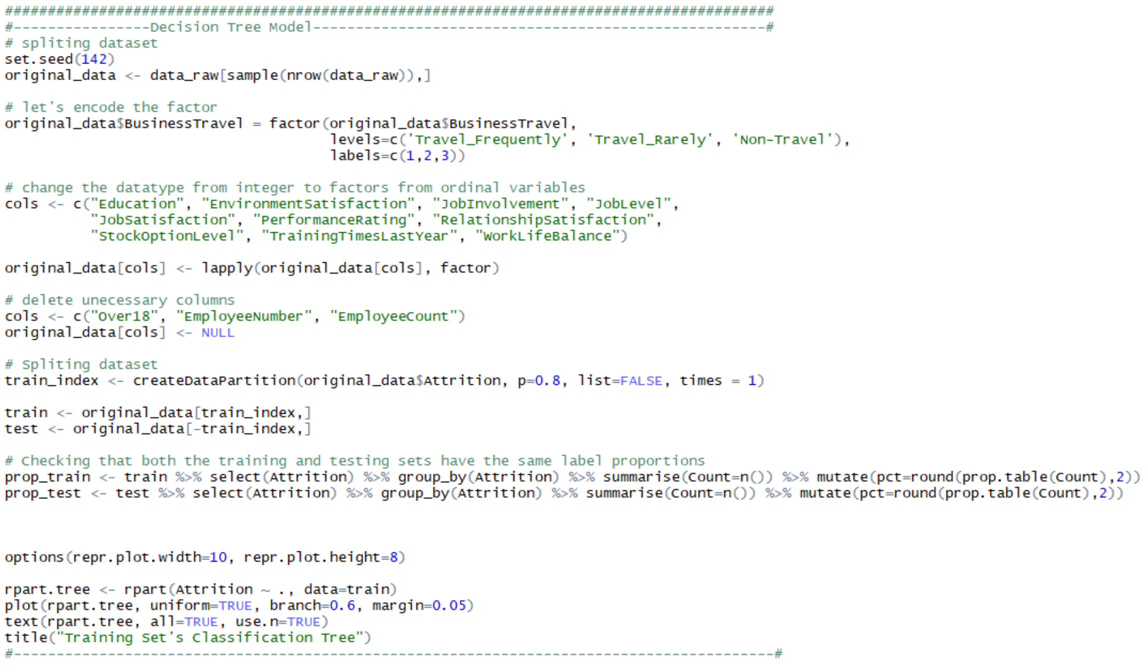
```
##########################################################################################
#----------------Attrition Status by Business Travel-------------------------------------#
# Business Travel
options(repr.plot.width=8, repr.plot.height=5)

data <- col_factor$aggregate('[
                             {
                             "$match":{"Attrition":"Yes"}
                             },
                             {
                             "$group":{"_id":{"Attrition":"$Attrition","BusinessTravel":"$BusinessTravel"}, "Count":{"$sum":1}}
                             }
                             ]')
data <- as.data.frame(matrix(unlist(data),byrow=F, nrow(data),ncol(data)+1))
names(data) <- c("Attrition","BusinessTravel","count")
data$count <- as.integer(as.character(data$count))
work_bal_cnt <- data %>%  mutate(pct=round(prop.table(count),2) * 100) %>%
  ggplot(aes(x=Attrition, y=count, fill=BusinessTravel, color=Attrition)) + geom_bar(stat='identity') + facet_wrap(~BusinessTravel) +
  geom_label(aes(label=count, fill = BusinessTravel), colour = "white", fontface = "italic")  + theme_minimal() + theme(legend.position="none") +
  scale_fill_manual(values=c("#00dbdb", "#00db6e", "#fa8072")) +
  scale_color_manual(values=c("#808080", "#808080")) + labs(title="Attrition by Business Travel of Employees",
                                                            x="Attrition", y="Number of Employees") + coord_flip() +
  theme(plot.title=element_text(hjust=0.5))

# work_bal_cnt

work_bal_pct <- data %>% mutate(pct=round(prop.table(count),2) * 100) %>%
  ggplot(aes(x=Attrition, y=pct, fill=BusinessTravel, color=Attrition)) + geom_bar(stat='identity') + facet_wrap(~BusinessTravel) + theme_minimal() +
  theme(legend.position="none") +
  geom_label(aes(label=paste0(pct, "%"), fill = BusinessTravel), colour = "white", fontface = "italic")  +
  scale_fill_manual(values=c("#00dbdb", "#00db6e", "#fa8072")) +
  scale_color_manual(values=c("#808080", "#808080")) + labs(x="Attrition", y="Percentage (%)") + coord_flip() +
  theme()

plot_grid(work_bal_cnt, work_bal_pct, nrow=2)
#-----------------------------------------------------------------------------------------#
```

Bi-variate analysis – Graph Number 30



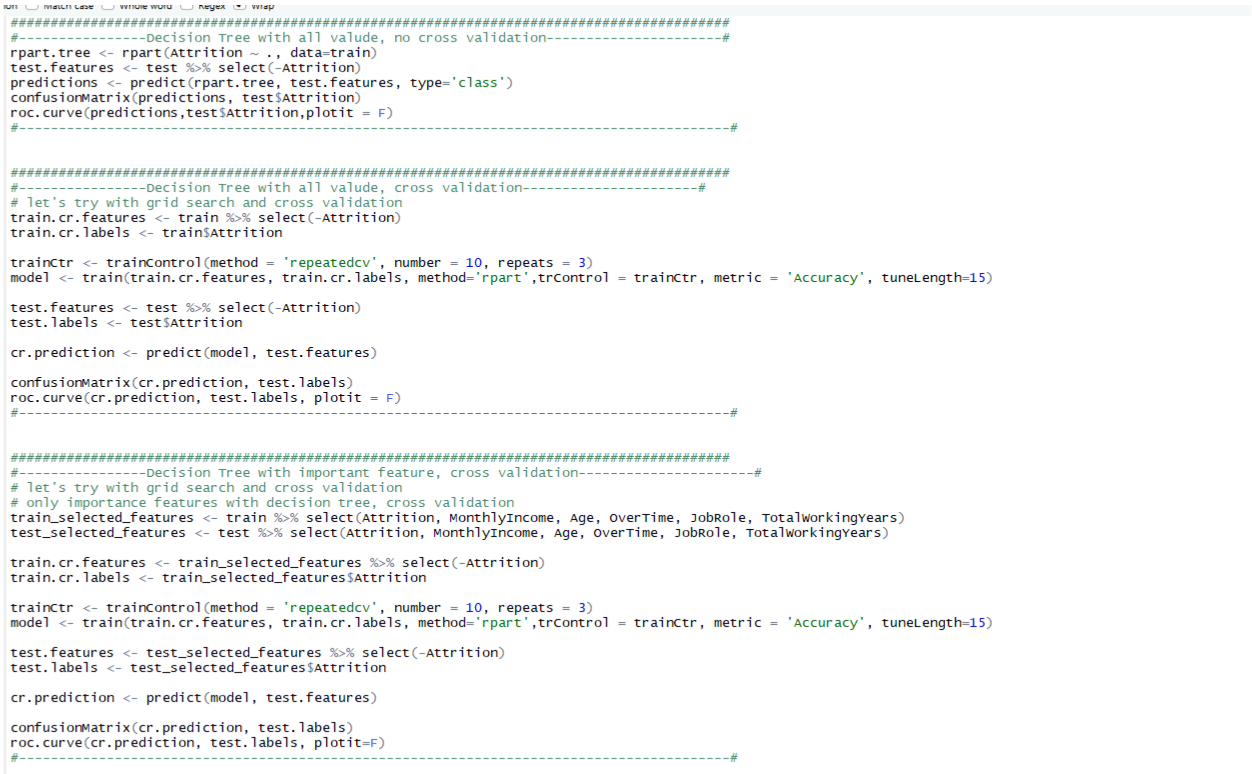
```
##########################################################################################
#----------------Bivariate Analysis------------------------------------------------------#
options(repr.plot.width=10, repr.plot.height=8)
# positive correlation
data <- col_numeric$find(fields='{"_id":0,"TotalWorkingYears":1,"MonthlyIncome":1}')
p1 <- data %>% ggplot(aes(x=TotalWorkingYears,y=MonthlyIncome)) +
  geom_point(colour="#F2DFCE", alpha=0.5) + geom_smooth(method = 'loess', color="#EE4037") + theme_minimal() +
  theme(legend.position = "bottom", plot.title = element_text(hjust = 0.5, color="white"),
        plot.subtitle=element_text(hjust=0.5,color='white'), plot.background=element_rect(fill="#0D7680"),
        axis.title.x=element_text(color='white',size=12), axis.text.x=element_text(color='white', size=10),
        axis.title.y=element_text(color='white',size=12), axis.text.y=element_text(color='white', size=10)) +
  labs(title="Positive Correlation", subtitle="Monthly Income vs Working Years")

# year with curr manager and year since the last promotion
data <- col_numeric$find(fields='{"_id":0,"YearsWithCurrManager":1,"YearsSinceLastPromotion":1}')
p3 <- data %>%
  ggplot(aes(x=factor(YearsWithCurrManager),y=YearsSinceLastPromotion)) + geom_boxplot(color="#FE642E",fill="#A9D0F5") +
  geom_jitter(color="#F2DFCE",alpha=0.5) + geom_smooth(method = 'loess',color="#EE4037",aes(group=1)) + theme_minimal() +
  theme(plot.background=element_rect(fill="#0D7680"),plot.title = element_text(hjust = 0.5, color="white"),
        plot.subtitle=element_text(hjust=0.5,color='white'),
        axis.title.x=element_text(color='white',size=12), axis.text.x=element_text(color='white', size=10),
        axis.title.y=element_text(color='white',size=12), axis.text.y=element_text(color='white', size=10)) +
  labs(title="Possitive Correlation",subtitle = "Years since Last Promotions vs Years with Current Manager",
       x="Years with Current Manager")

data <- col_numeric$aggregate('[
                              { "$lookup":
                                { "from":"hr_factor",
                                  "localField":"id",
                                  "foreignField": "id",
                                  "as":"hr_factor"
                                }},
                              {"$replaceRoot": { "newRoot": { "$mergeObjects": [ { "$arrayElemAt": [ "$hr_factor", 0 ] }, "$$ROOT" ] } }},
                              {"$project": { "hr_factor": 0} },
                              {"$project":{"_id":0, "PerformanceRating":1,"PercentSalaryHike":1}}]')

p2 <- data %>% ggplot(aes(x=factor(PerformanceRating),y=PercentSalaryHike)) +
  geom_boxplot(color="#FE642E", fill="#A9D0F5") + geom_jitter(color="#F2DFCE", alpha=0.5) + theme_minimal() +
  theme(plot.background=element_rect(fill="#0D7680"),plot.title = element_text(hjust = 0.5, color="white"),
        plot.subtitle=element_text(hjust=0.5,color='white'),
        axis.title.x=element_text(color='white',size=12), axis.text.x=element_text(color='white', size=10),
        axis.title.y=element_text(color='white',size=12), axis.text.y=element_text(color='white', size=10)) +
  labs(title="Possitive Correlation",subtitle = "Percent Salary Hike vs Performance Rating",
       x="Performance Rating")
```

Decision Tree Model



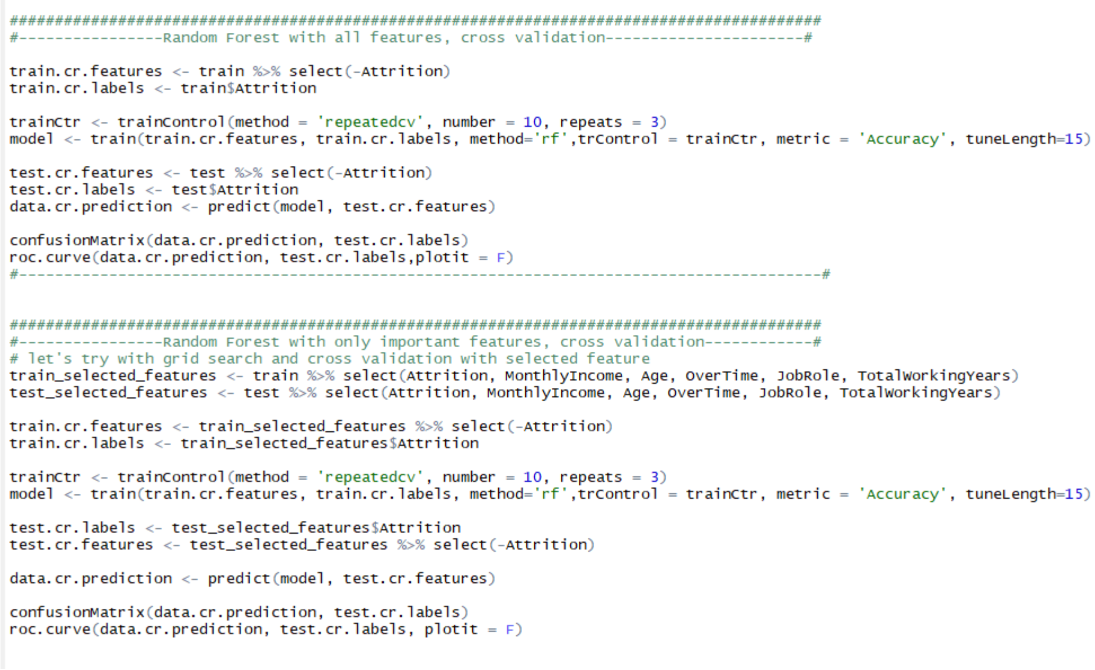
```
##########################################################################################
#----------------Decision Tree Model-----------------------------------------------------#
# spliting dataset
set.seed(142)
original_data <- data_raw[sample(nrow(data_raw)),]

# let's encode the factor
original_data$BusinessTravel = factor(original_data$BusinessTravel,
                                      levels=c('Travel_Frequently', 'Travel_Rarely', 'Non-Travel'),
                                      labels=c(1,2,3))

# change the datatype from integer to factors from ordinal variables
cols <- c("Education", "EnvironmentSatisfaction", "JobInvolvement", "JobLevel",
          "JobSatisfaction", "PerformanceRating", "RelationshipSatisfaction",
          "StockOptionLevel", "TrainingTimesLastYear", "WorkLifeBalance")

original_data[cols] <- lapply(original_data[cols], factor)

# delete unecessary columns
cols <- c("Over18", "EmployeeNumber", "EmployeeCount")
original_data[cols] <- NULL

# Spliting dataset
train_index <- createDataPartition(original_data$Attrition, p=0.8, list=FALSE, times = 1)

train <- original_data[train_index,]
test <- original_data[-train_index,]

# Checking that both the training and testing sets have the same label proportions
prop_train <- train %>% select(Attrition) %>% group_by(Attrition) %>% summarise(Count=n()) %>% mutate(pct=round(prop.table(Count),2))
prop_test <- test %>% select(Attrition) %>% group_by(Attrition) %>% summarise(Count=n()) %>% mutate(pct=round(prop.table(Count),2))


options(repr.plot.width=10, repr.plot.height=8)

rpart.tree <- rpart(Attrition ~ ., data=train)
plot(rpart.tree, uniform=TRUE, branch=0.6, margin=0.05)
text(rpart.tree, all=TRUE, use.n=TRUE)
title("Training Set's Classification Tree")
#-----------------------------------------------------------------------------------------#
```

Decision Model with Cross Validation



```
##########################################################################################
#---------------Decision Tree with all valude, no cross validation------------------------#
rpart.tree <- rpart(Attrition ~ ., data=train)
test.features <- test %>% select(-Attrition)
predictions <- predict(rpart.tree, test.features, type='class')
confusionMatrix(predictions, test$Attrition)
roc.curve(predictions,test$Attrition,plotit = F)
#------------------------------------------------------------------------------------------#


##########################################################################################
#---------------Decision Tree with all valude, cross validation------------------------#
# let's try with grid search and cross validation
train.cr.features <- train %>% select(-Attrition)
train.cr.labels <- train$Attrition

trainCtr <- trainControl(method = 'repeatedcv', number = 10, repeats = 3)
model <- train(train.cr.features, train.cr.labels, method='rpart',trControl = trainCtr, metric = 'Accuracy', tuneLength=15)

test.features <- test %>% select(-Attrition)
test.labels <- test$Attrition

cr.prediction <- predict(model, test.features)

confusionMatrix(cr.prediction, test.labels)
roc.curve(cr.prediction, test.labels, plotit = F)
#------------------------------------------------------------------------------------------#


##########################################################################################
#---------------Decision Tree with important feature, cross validation------------------------#
# let's try with grid search and cross validation
# only importance features with decision tree, cross validation
train_selected_features <- train %>% select(Attrition, MonthlyIncome, Age, OverTime, JobRole, TotalWorkingYears)
test_selected_features <- test %>% select(Attrition, MonthlyIncome, Age, OverTime, JobRole, TotalWorkingYears)

train.cr.features <- train_selected_features %>% select(-Attrition)
train.cr.labels <- train_selected_features$Attrition

trainCtr <- trainControl(method = 'repeatedcv', number = 10, repeats = 3)
model <- train(train.cr.features, train.cr.labels, method='rpart',trControl = trainCtr, metric = 'Accuracy', tuneLength=15)

test.features <- test_selected_features %>% select(-Attrition)
test.labels <- test_selected_features$Attrition

cr.prediction <- predict(model, test.features)

confusionMatrix(cr.prediction, test.labels)
roc.curve(cr.prediction, test.labels, plotit=F)
#------------------------------------------------------------------------------------------#
```

Random Forest



```
##########################################################################################
#----------------Random Forest with all features, cross validation-----------------------#

train.cr.features <- train %>% select(-Attrition)
train.cr.labels <- train$Attrition

trainCtr <- trainControl(method = 'repeatedcv', number = 10, repeats = 3)
model <- train(train.cr.features, train.cr.labels, method='rf',trControl = trainCtr, metric = 'Accuracy', tuneLength=15)

test.cr.features <- test %>% select(-Attrition)
test.cr.labels <- test$Attrition
data.cr.prediction <- predict(model, test.cr.features)

confusionMatrix(data.cr.prediction, test.cr.labels)
roc.curve(data.cr.prediction, test.cr.labels,plotit = F)
#-----------------------------------------------------------------------------------------#


##########################################################################################
#----------------Random Forest with only important features, cross validation------------#
# let's try with grid search and cross validation with selected feature
train_selected_features <- train %>% select(Attrition, MonthlyIncome, Age, OverTime, JobRole, TotalWorkingYears)
test_selected_features <- test %>% select(Attrition, MonthlyIncome, Age, OverTime, JobRole, TotalWorkingYears)

train.cr.features <- train_selected_features %>% select(-Attrition)
train.cr.labels <- train_selected_features$Attrition

trainCtr <- trainControl(method = 'repeatedcv', number = 10, repeats = 3)
model <- train(train.cr.features, train.cr.labels, method='rf',trControl = trainCtr, metric = 'Accuracy', tuneLength=15)

test.cr.labels <- test_selected_features$Attrition
test.cr.features <- test_selected_features %>% select(-Attrition)

data.cr.prediction <- predict(model, test.cr.features)

confusionMatrix(data.cr.prediction, test.cr.labels)
roc.curve(data.cr.prediction, test.cr.labels, plotit = F)
```