­­

Higher Diploma in Data Analytics (HDSDA) Programming for Big Data

Student Name: Siobhan PurcellStudent Number: 18195342  
Continuous Assignment 2

April 2019

Lecturer: Symeon Charalabides

Table of Contents

[Table of Contents 2](#_Toc7382302)

[Table of Figures 4](#_Toc7382303)

[Table of Tables 5](#_Toc7382304)

[Project Overview Introduction 6](#_Toc7382305)

[1. Report 1. Analysis of the Titanic disaster survival dataset for predicting passenger survival 7](#_Toc7382306)

[1.1. Dataset Description 7](#_Toc7382307)

[1.2. Literature Review 8](#_Toc7382308)

[1.3. Analysis Objectives 8](#_Toc7382309)

[1.4. Cleaning and pre-processing the dataset 9](#_Toc7382310)

[1.4.1. Missing Value Imputation 9](#_Toc7382311)

[1.4.2. Data transformation and Preparation 11](#_Toc7382312)

[1.5. Results 12](#_Toc7382313)

[1.5.1. Exploring Passengers’ Demographic profiles 12](#_Toc7382314)

[1.5.2. Sex and Survived 14](#_Toc7382315)

[1.5.3. Age and Survived 15](#_Toc7382316)

[1.5.4. Pclass and Survived 16](#_Toc7382317)

[1.6. Conclusions and Summary of Results 19](#_Toc7382318)

[2. Report 2. Analysis of Human Resources dataset: Improving Employee Retention by Predicting Employee Attrition 20](#_Toc7382319)

[2.1. Dataset Description 20](#_Toc7382320)

[2.2. Literature Review 21](#_Toc7382321)

[2.3. Analysis Objectives 22](#_Toc7382322)

[2.4. Preparing and pre-processing the dataset 22](#_Toc7382323)

[2.5. Results 24](#_Toc7382324)

[2.5.1. Statistical Overview 24](#_Toc7382325)

[2.5.2. Exploring Company’s Business and Department Infrastructure 25](#_Toc7382326)

[2.5.3. Correlation matrix and heatmap 26](#_Toc7382327)

[2.5.4. Distributions of Satisfaction, Evaluation & Average Monthly Hours 28](#_Toc7382328)

[2.5.5. Characterising employees and predicting who may leave company next 34](#_Toc7382329)

[2.6. Conclusions and Summary of Results 36](#_Toc7382330)

[3. References cited 37](#_Toc7382331)

[3.1. Titanic dataset 37](#_Toc7382332)

[3.1.1. Material for literature review 37](#_Toc7382333)

[3.1.2. Supporting material For R studio data analysis 37](#_Toc7382334)

[3.2. Human Resources dataset 39](#_Toc7382335)

[3.2.1. Material for literature review 39](#_Toc7382336)

[3.2.2. Material For R data analysis 39](#_Toc7382337)

Table of Figures

[Figure 1: Distribution of missing data in the Titanic dataset. 9](#_Toc7382338)

[Figure 2: The distribution passenger ages after exclusion of missing values (A) boxplot and representing the distribution shape of passenger ages. (B) A boxplot showing distribution of age data by sex and their degree of variability. 10](#_Toc7382339)

[Figure 3: The final cleaned, filtered and transformed data that will be used for analysis. 11](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382340)

[Figure 4: A statistical summary on passenger survivability, class, gender, age and title. 12](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382341)

[Figure 5: Relative frequency bar charts for the variables (A) Sex (B) Age (C) and Class. 13](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382342)

[Figure 6: Frequency distribution of the Titanic survivors and victims 14](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382343)

[Figure 7: Proportion of Survivors by Sex 14](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382344)

[Figure 8: Proportion of survivors by age 15](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382345)

[Figure 9: Proportion of survivors by age group 15](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382346)

[Figure 10: Count of total survivors by class 16](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382347)

[Figure 11: Distribution of Survivors by Title 17](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382348)

[Figure 12: Frequency distribution of survivors by class and title 18](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382349)

[Figure 13: Statistical summary on the HR dataset. 22](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382350)

[Figure 14: Frequency distribution of employees who left and stayed with the company 24](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382351)

[Figure 15: (A) Distibution of employees across different departments. (B) Distribution of employees who left across different departments. 25](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382352)

[Figure 16: Heatmap of Correlation Matrices A) Correlation heat map of all company employees (B) Correlation heatmap of employees who left the company 27](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382353)

[Figure 17: Distribution of satisfaction levels, evaluation score, and average monthly hours across employees (A) Employees who stayed with company (B )Employees who left from company. 28](file:///C:\Users\Siobhan\Documents\Data%20Analytics%20Course%20(Siobhan)\Semester%201\Programming%20for%20Big%20Data\Project\Final%20Submission%20Folder%20CA2\Programming%20for%20Big%20Data%20CA2_Siobhan%20Purcell_Final_Report.docx#_Toc7382354)

[Figure 18: Frequency distribution of employees who stayed (0) and left (1) by project count. 29](#_Toc7382355)

[Figure 19: Boxplots of employee turnover by (A) project counts vs average monthly hours (B) project counts vs average monthly hours. 31](#_Toc7382356)

[Figure 20: Population clusters of employee turnover by (A) satisfaction\_levels vs evaluation\_score (B) satisfaction\_level vs average\_monthly\_hours 33](#_Toc7382357)

[Figure 21: A multivariate comparison between Evaluation\_Score, Satisfaction\_level and average\_monthly\_hours. 34](#_Toc7382358)

[Figure 22: Percentage of current employees that may leave according to the clustering features of the valuable employees that departed. 35](#_Toc7382359)

Table of Tables

[Table 1:Summary description of the Titanic dataset and variables 7](#_Toc7382420)

[Table 2: Absolute rate of survival by class 16](#_Toc7382421)

[Table 3: Relative frequency of survivors by class 16](#_Toc7382422)

[Table 4: Relative frequency of passenger survivors by titles 17](#_Toc7382423)

[Table 5: Summary description of the Human Resources dataset 20](#_Toc7382424)

[Table 6: Summary table of the changes that were made to the dataset’s variable names 23](#_Toc7382425)

[Table 7: Frequency of employees that stayed and departed from the company 24](#_Toc7382426)

[Table 8: Relative Frequency of resigned employees by department 26](#_Toc7382427)

[Table 9: Summary of variable values of each employee subgroup that left and their assigned employee type behaviour 34](#_Toc7382428)

Project Overview Introduction

For this project, two large datasets were analysed using statistical programming software R. The first selected dataset presents a wealth of historical information on a sample number of passengers who embarked upon the RMS Titanic’s fateful journey. The second is a much larger and modern dataset containing simulated Human Resources information on the employment status and attributes of employees for a company. Both datasets are sufficiently rich and complex for conducting deep exploratory analyses. Each present their own unique challenges and shall be discussed in further detail in their respective sections.

# Report 1. Analysis of the Titanic disaster survival dataset for predicting passenger survival

## Dataset Description

This dataset provides demographic and survival information from 891 of the 2224 passengers who embarked upon one of the most **i**nfamous and tragic shipwrecks in history; the sinking of the RMS Titanic.

All historical records were originally collected by the British Board of Trade during investigative proceedings (British Board of Trade; 1990) and the raw file was derived from the online community sharing platform, Kaggle. The dataset contains 891 observations with 12 variables; a summary description for each passenger attribute can be found in in Table 1.

Table 1:Summary description of the Titanic dataset and variables

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Key** |
| Survived | Indication of Survival | 0 = No, 1 = Yes |
| Pclass | Ticket class | Integer: 1 = 1st, 2 = 2nd, 3 = 3rd |
| Name | Full name of passenger | Characters |
| Sex | Sex | Characters: Male; Female |
| Age | Age in years | Numeric |
| SibSp | No. of siblings / spouses | Integer |
| Parch | No. of parents / children aboard | Integer |
| Ticket | Issued Ticket No. | Factor |
| Fare | Passenger fare for ticket | Numeric |
| Cabin | Cabin No. occupied | Factor |
| Embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

## Literature Review

On April 15th, 1912, the Titanic collided with an iceberg which resulted in the loss of many lives due to the insufficient number of lifeboats to accommodate all passengers on board.

Preliminary research from historical sources would imply a higher chance of survival to be observed within certain groups of people, particularly, amongst adult females and accompanying children over adult males. Due to the traditional documented policy of “women and children first” this population would have been afforded more priority for securing a lifeboat space. (Everett, M., & Neill, 1912 ; J.B. Geller, 1998).

Consistent documentation from investigative reports have also supported that socioeconomic status is additional determining factor for survivability. Studies have shown that people from different education backgrounds and income, display a significant difference in mortality rates (Kitagawa E.M., Hauser P.M., 1973). Therefore, those travelling within the rich and prominent upper class may have had better access to safety information and privileges from the crew, thus increasing their likelihood of survival during the disaster over third class passengers (Bijan, A., 2014; J.R. Henderson, 1998).

## Analysis Objectives

Following a comprehensive review of the literature, the insights gathered were used to identify the most intuitive and relevant variables to address them. Hence, the main aims of this investigation were:

* To assess the demographic structure of the population cohort within the dataset according to their gender, age and social status, as represented by the variables “Sex”, “Age”, “Pclass”, respectively.
* To examine the relationship between these listed variables a passenger’s likelihood of survival.
* To extract new features from the existing dataset that could provide greater insight into survivability.

## Cleaning and pre-processing the dataset

Given the historical nature of these records, it was important initially inspect and cleanse the data for any potential quality issues and to transform it to datatype more amenable for exploratory data analysis in the next section.

### Missing Value Imputation

Immediately upon first review of the tabular dataset, a significant number (NA=177) of missing values were detected within the “Age” variable. This amount was even further increased following the conversion of all empty strings as NA values, revealing (after converting all empty strings into NA values using the na.strings argument, this total amount was even further substantially increased), revealing additional blank data within the columns “Cabin” (NA=687) and “Embarked” (NA=2). The distribution pattern of all total missing values was then visualised using the VIM package (Figure 1).

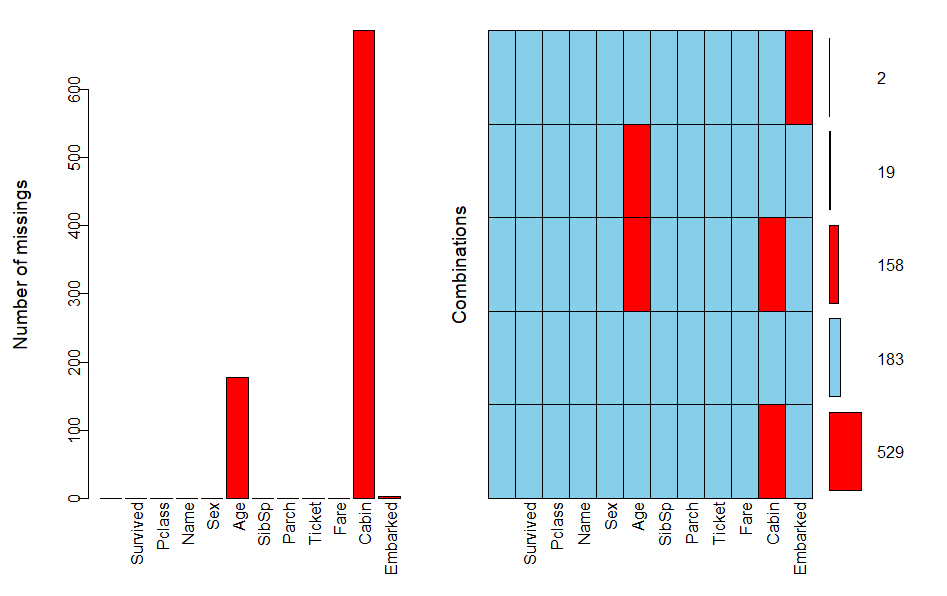


Figure 1: Distribution of missing data in the Titanic dataset.

Careful consideration was undertaken to address the incomplete age data, given that the variable “Age” would be used for further analysis in this study. The final decided technique was to impute these missing values with either the calculated mean or the median of all the passengers, depending on distribution of the data. Following removal of all NA values with the na.rm() function to accommodate descriptive statistics, a box and histogram was applied to determine which measure of central tendency would provide a more accurate prediction of a passenger’s age. (Figure 1A, 1B). While the age distribution of the entire population appears well-distributed, there are a several extreme outliers within the dataset, particularly among males, which could skew the mean. Thus, the median age chosen as the final most appropriate measure for populating the empty values.

As both “Cabin” and “Fare” would be omitted later from the final feature array, their incomplete values did not generate any raising concerns.

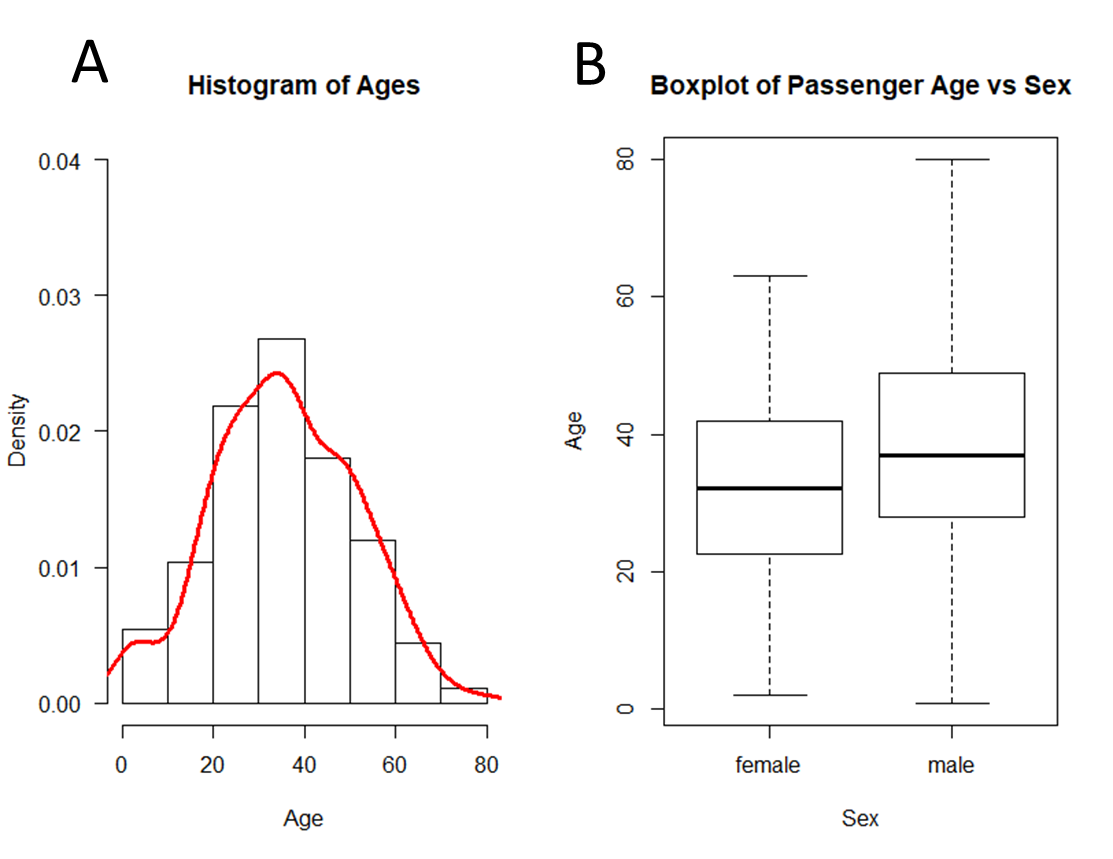


Figure 2: The distribution passenger ages after exclusion of missing values (A) boxplot and representing the distribution shape of passenger ages. (B) A boxplot showing distribution of age data by sex and their degree of variability.

### Data transformation and Preparation

The cleaned data was then subsetted into a new data frame to only include variables that had been selected for further exploratory analysis which include “Age”, “Pclass”, “Sex”, “Survived” and “Name”.

Additionally, a new feature called “Title” was created by extracting the embedded titles of “Miss”, “Mr”, “Mrs”, “Master” and “Other” from the strings of each passengers’ “Name” value. The information gathered may provide a deeper insight on a passenger’s chance of survival.

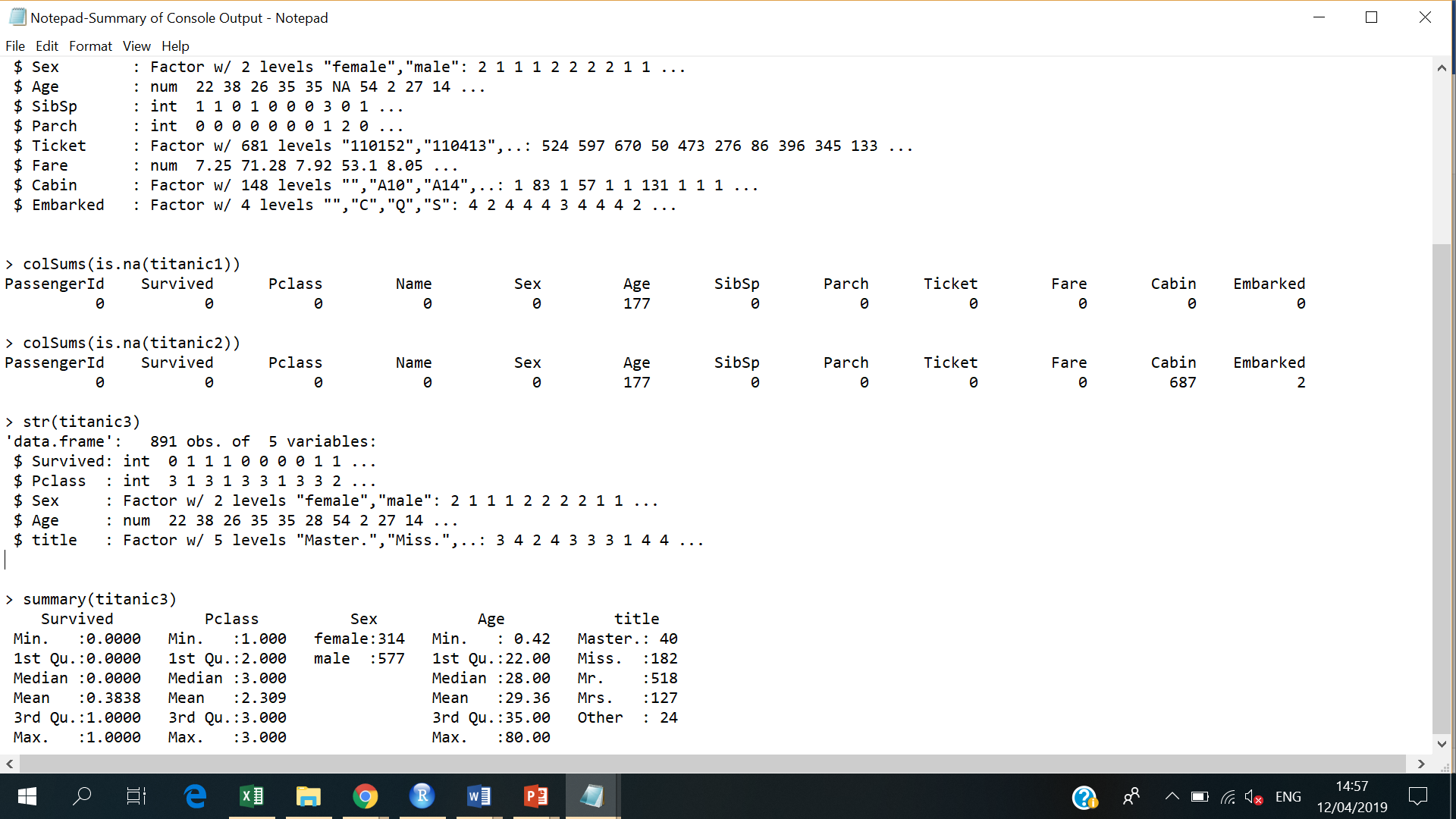
After omitting the “Name” column, the final selected features were converted into either a numeric or categorical date type which would be more amenable for data analysis in the next section. The final filtered and cleansed dataset is summarised in the table below.

Figure 3: The final cleaned, filtered and transformed data that will be used for analysis.

## 

## Results

This section will discuss the main findings and insights that were gathered from the analysis. For a comprehensive view of how the dataset was explored and supplementary results, please refer to the original R script.

### Exploring Passengers’ Demographic profiles

Initially, preliminary descriptive statistics was performed to provide a summary about the absolute frequency distributions of age, sex, travelling class, title and survivors/victims amongst passengers.

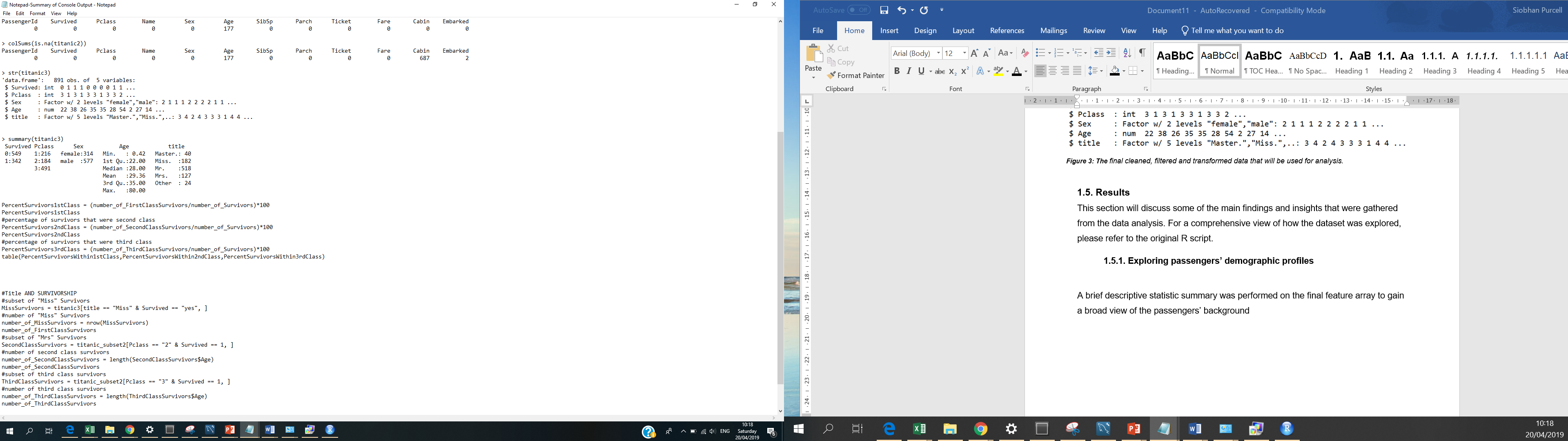


Figure 4: A statistical summary on passenger survivability, class, gender, age and title.

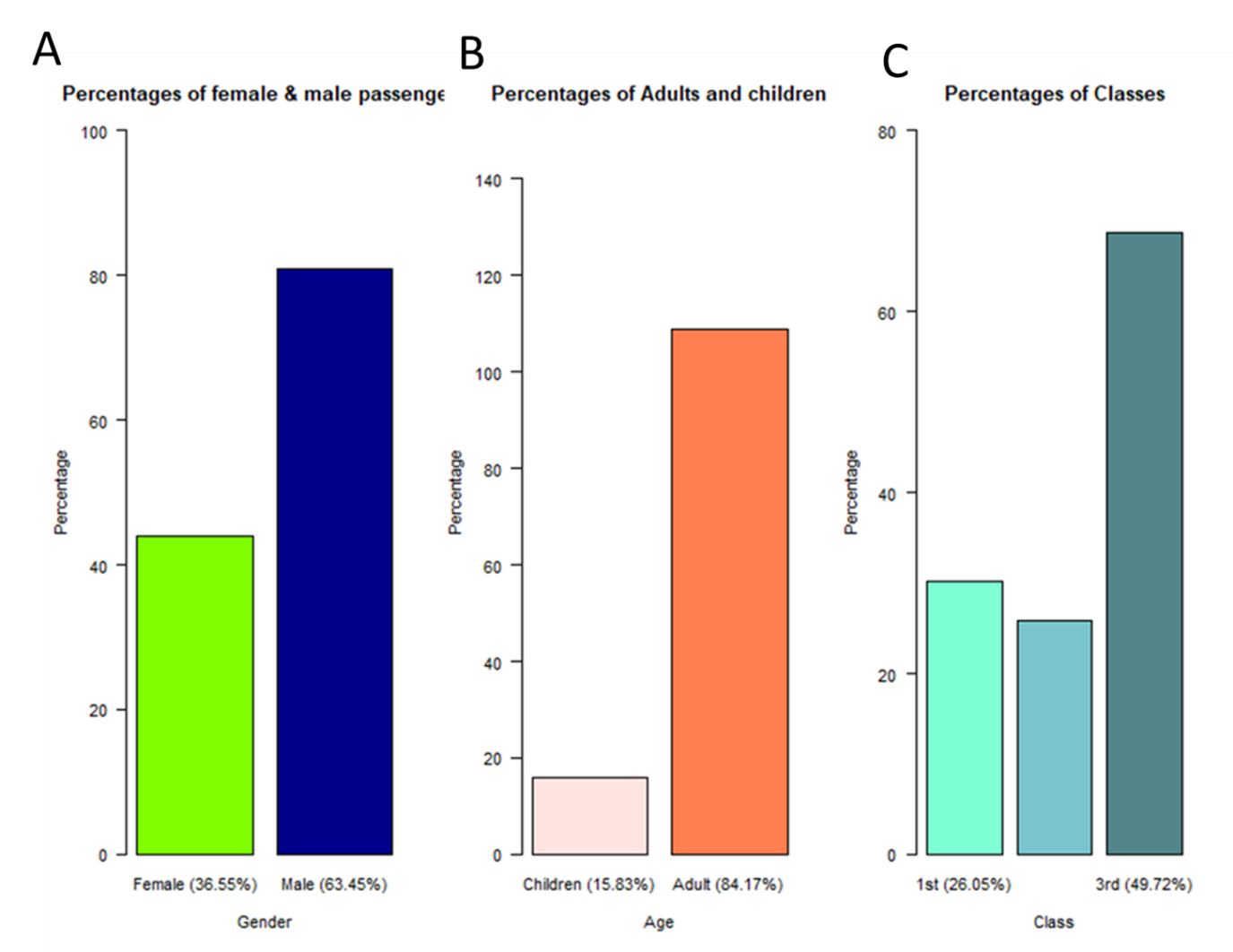
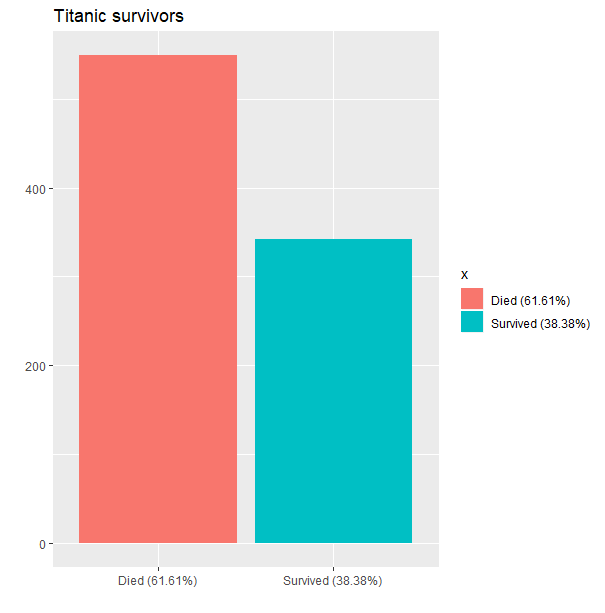
 For categorical univariate analysis, several bar plots were created to display the relative frequency distributions of passengers according to their gender, age and class categories (Figure 5). To accommodate these illustrations, the ages of passengers were stratified into two age groups; children (age ≤18) and adults (age ≥18).

Figure 5: Relative frequency bar charts for the variables (A) Sex (B) Age (C) and Class.

These findings show there is an imbalance amongst the distribution of each variable amongst the sample population. Clearly, there were proportionally almost twice as many men (64%) on board the Titanic than women than women (37%) on board (Figure 1A). Similarly, the majority of passengers were adults with children only comprising of only 16% (Figure 1B). Again, the 3 categories of Pclass are not evenly represented in their distribution. It appears that the majority of passengers were comprised of 3rd class, followed by 1st and 2st (Figure 1C).

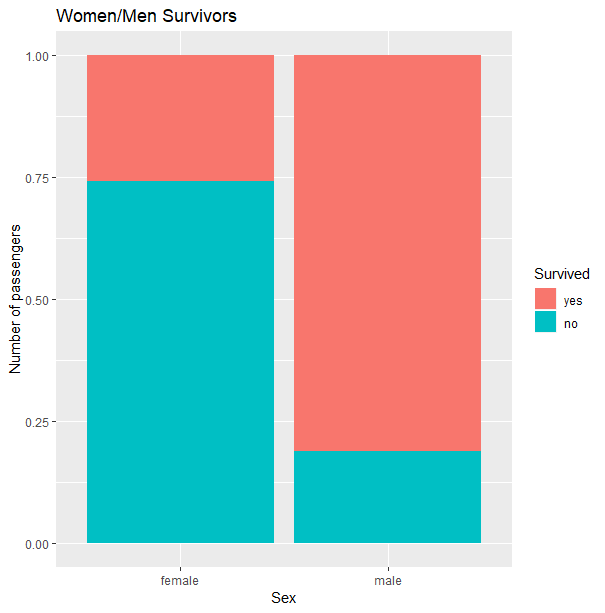
A barchart was also used to describe the relative frequencies of survivors and victims for the whole sample.

With regards to the survival rate within the dataset, only 38% survived overall and, sadly, the majority had perished.

The next step of this analysis will explore if sex, age and economic status could somehow be associated with a passenger’s chance for survival.

Figure 6: Frequency distribution of the Titanic survivors and victims

### Sex and Survived



As demonstrated in Figure 7, when compared as a percentage of survivors by gender, females had a larger proportion of survivors for survival (74%) relative to males (19%). This indicates that women were more likely to survive than men.

Figure 7: Proportion of Survivors by Sex

### Age and Survived

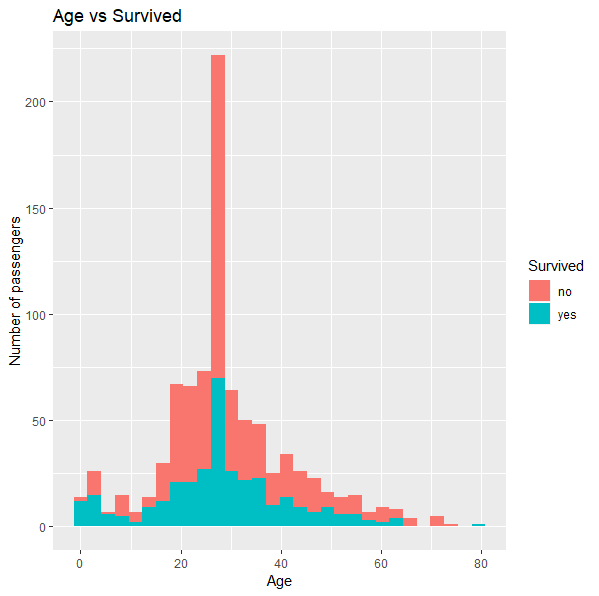
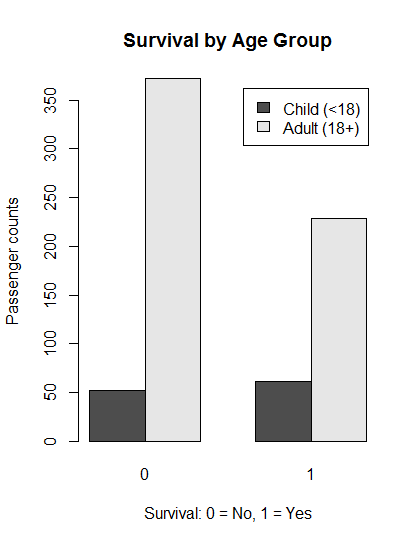


Figure 8: Proportion of survivors by age

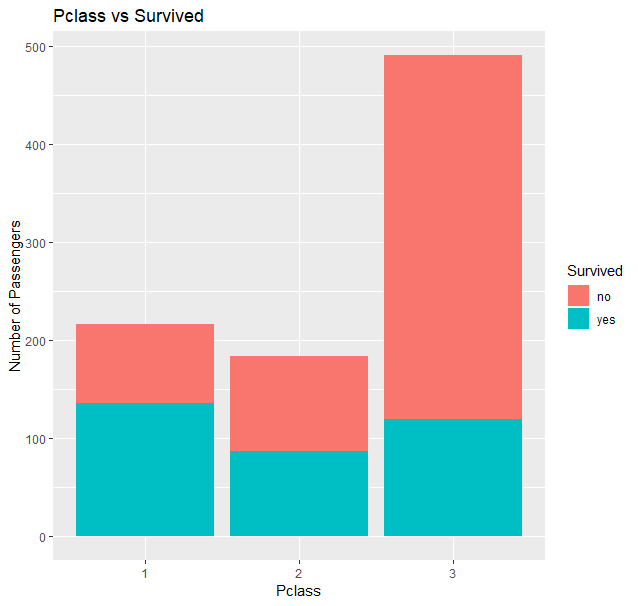
A histogram was generated to understand the distribution of survivors by age (Figure 8). While the relationship here is not simple, there are observable differences between survivors and victims across different age ranges groups which was further inspected.

****

As depicted in the bar chart for age groups (Figure 9), the distribution of children who survived and died is very similar. When compared alongside adults, it is evident that a larger proportion of children lived. Thus, children had a much higher survival advantage than older passengers.

Figure 9: Proportion of survivors by age group

### Pclass and Survived

Another research goal was to explore if a passenger’s socioeconomic status could have affected their likelihood for survival. As 1st class passengers had better access to facilities and benefits on board, one would expect to see a higher proportion of survivors within this level, relative to the other classes.

Notably, the share of survivors by class does not reflect the distribution of passengers overall (Figure 10). Whilst the survival percentage differences between each class appear minor in Table 2, it is important to note that the 3rd class comprised the majority of passengers

Figure 10: Count of total survivors by class

Table 2: Absolute rate of survival by class

|  |  |  |
| --- | --- | --- |
| 1st class | 2nd class | 3rd class |
| 40 % | 25 % | 35 % |

> table(PercentSurvivors1stClass, PercentSurvivors2ndClass, PercentSurvivors3rdClass)

This discrepancy was amplified when the percentage share of survivors within each group was calculated and compared (Table 3).

Table 3: Relative frequency of survivors by class

|  |  |  |
| --- | --- | --- |
| 1st class | 2nd class | 3rd class |
| 63% % | 47 % | 24 % |

> table(PercentSurvivorsWithin1stClass, PercentSurvivorsWithin1stClass, Percent3rdClassSurvivors)

From this perspective, a much more striking difference was observed between the levels. As expected, 3rd class passengers had suffered the poorest outcome, with a survival rate of just under 25%. This may be explained by the location of their living quarters on the lower decks of the ship or else by the potential discrimination of their social status. Whatever the case may be, the results support the assumption that, with those travelling in 1st class had a much higher probability in securing a lifeboat space.

To further expand upon the relationship between passenger class and survival, each person’s rank or title were explored in further detail.

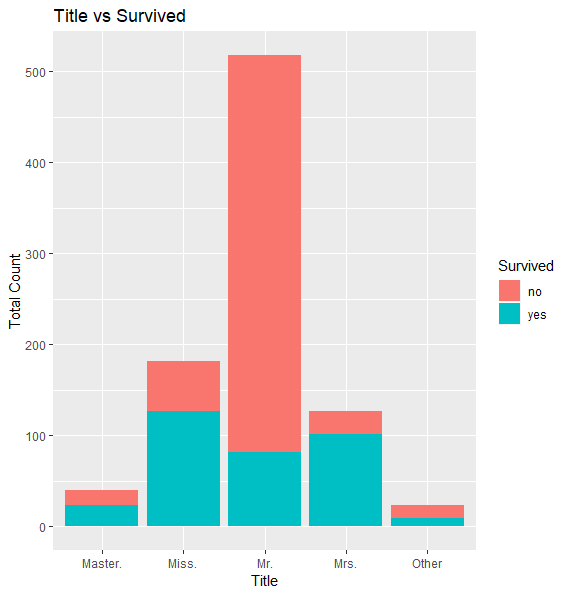


Figure 11: Distribution of Survivors by Title

Table 4: Relative frequency of passenger survivors by titles

|  |  |
| --- | --- |
| **Master** | 57.5 % |
| **Mr.** | 15.83 % |
| **Miss.** | 69.78 % |
| **Mrs.** | 79.53 % |
| **Other** | 37.5 % |

> table(PercentMasterSurvivors, PercentMissSurvivors, PercentMrSurvivors, PercentMrsSurvivors, PercentOtherSurvivors)

To further expand upon the relationship between passenger class and survival, each person’s rank or title were explored in further detail. As shown by the bar chart (Figure 11), and relative frequency table above (Table 4), the survival rates of females owning the titles “Miss” and “Mrs” in their names are very similar and have highest proportion of survivors. Amongst the subgroup of males by title, those bearing the honorific address of “Master” have a much a greater chance of survival. This may be confounded by age.

Furthermore, some additional interesting insights were gathered when survivorship was plotted against both class and title. Figure 12 shows that title is a meaningful predictor for survivability, particularly among the upper-class passengers. Indeed, the “women and children first” policy is much more striking as it appears that 1st and 2nd class females with the were far more likely to survive than other travellers, revealing a 90% survival rate. Moreover, all of the “Master” passengers in both upper classes have a 100% probability of survival.

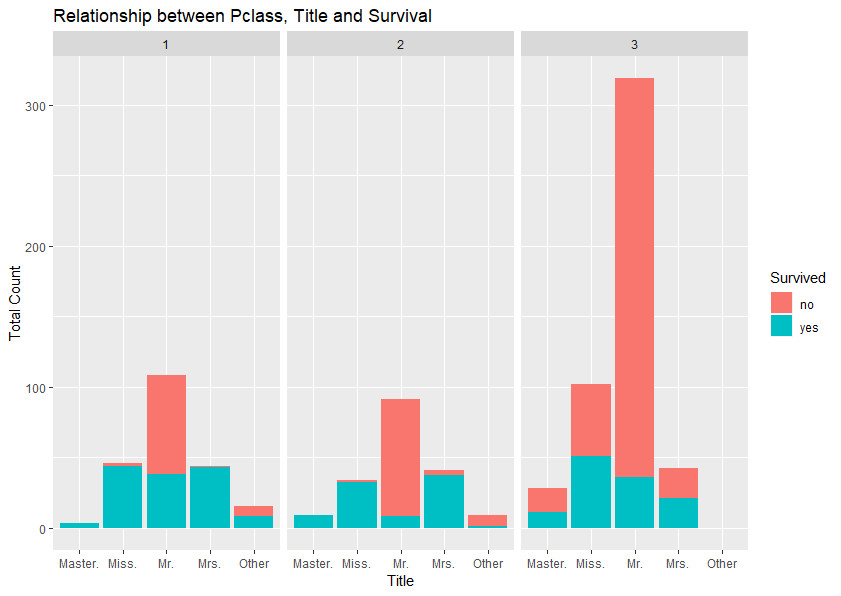
****

Figure 12: Frequency distribution of survivors by class and title

## Conclusions and Summary of Results

In conclusion, the results from this analysis demonstrate the relative importance of gender and socioeconomic class in determining passenger survival and revealed a number of interesting insights:

* The policy of “women and children” appears to have be rigidly enforced during the event of the disaster as the chances for survival were far greater for this cohort.
* There is a particularly strong correlation between being in the upper 1st class and survival.
* These odds were highest for females travelling within 1st or 2nd class.
* In contrast, the odd of survival was worst for males in 3rd class.
* However, males that bore the title of “Master” had a far better outcome amongst overall male cohort.

# Report 2. Analysis of Human Resources dataset: Improving Employee Retention by Predicting Employee Attrition

## Dataset Description

The second analysis concerns a dataset acquired from the *“*Human Resources Analytics” Kaggle challenge. It provides summary information on 14999 employees, with respect to ten work-related variables, including whether or not they left (1=Yes) or stayed (0=No). Further details and description for each field in the dataset can be found in Table 5. Due to the sensitive nature of employee data, all material within this dataset has been simulated*.*

Table 5: Summary description of the Human Resources dataset

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Key** |
| **Satisfication\_level** | Whether an employee is satisfied or not with their work | Numeric (ranges from 0 to 1) |
| **Last\_evaluation** | Last performance rating of an employee by the company | Numeric (ranges from 0 to 1) |
| **number\_project** | No. of projects worked on (yearly basis) | Integer |
| **Average\_montly\_hours** | Average monthly working hours | Integer |
| **time\_spend\_company** | No of years spent at the company | Integer |
| **Work\_accident** | Whether they have had a work accident in the last 2 years | Integer  0 = No, 1 = Yes |
| **left** | Whether the employee has left the company | Integer  0 = No, 1 = Yes |
| **Promotion\_last\_5years** | Whether they have had a promotion in the last 5 years | Integer  0 = No, 1 = Yes |
| **salary** | Salary of the employee | Factor (3 levels) |
| **Sales** | employee working in which departments | Factor (10 levels) |

## Literature Review

The consequences of employee attrition can have a detrimental impact for the costs of any industry or organisation. According to case studies drawn from several research papers, the average costs of turnover for a business is approximately 20-25% of a worker’s annual salary. These expenses reflect the loss of productivity after their departure as well as the associated additional investments of recruiting and training a replacement (Tracey, J.B. & Hinkin, T.R., 2008; Crowe, E., & Schaefer B., 2007; Jones C.B., 2005; Tracey, J.B. & Hinkin, T.R, 2000). Furthermore, high employee turnover could damage a company’s performance by prohibiting them from retaining this collective knowledge base and experience in the long term.

There have been considerable number of studies on factors responsible for employee attrition, with the most notable determinants including overall job satisfaction, job performance, salary and promotional opportunities (Cotton, J.L. and Tuttle, J.M., 1986; E. Ribes, K. Touahri and B., Perhame, 2017). Another recent study revealed unsatisfactory work-life balance and work stress can cause low employee satisfaction and result in early employee reassignment (Jessica Sze-Yin Ho, 2010)

Therefore, it is important for organisations to predict which of their employees are likely to prematurely leave and to develop strategies for retaining their most valuable employees.

## Analysis Objectives

The primary focus of the project is to:

* Identify which variables are the main drivers of employee churn
* Use gathered insights to identify the company’s current high value employees that are at high risk for turnover

## Preparing and pre-processing the dataset

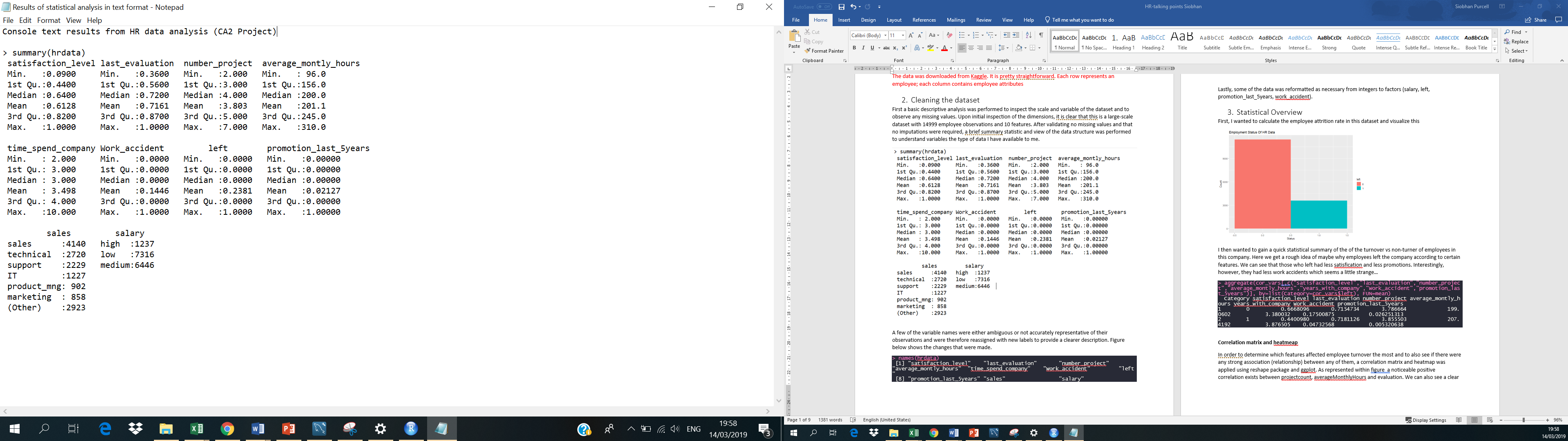
First, a basic summary statistic was performed on the to understand the type of variables within and to determine if there may be any quality issues that needed to be addressed (Figure 13).

Figure 13: Statistical summary on the HR dataset.

Unlike the previous dataset, all of the records were of excellent quality, with no missing values and or extreme outliers to suggest false or incomplete data. However, upon further careful inspection it appeared that some of the feature names were ambiguous or else they did not match the observation values appropriately. These variables were therefore reassigned with new labels to provide a clearer description of the data. A summary of these variable name changes can be viewed in Table 6 below.

Table 6: Summary table of the changes that were made to the dataset’s variable names

|  |  |
| --- | --- |
| **Original dataset** | **New dataset** |
| “last\_evaluation” | evaluation\_score |
| “average\_montly\_hours” | “average\_monthly\_hours” |
| “time\_spend\_company” | “years\_with\_company” |
| “sales” | “department” |

## Results

This section will discuss the main findings and insights that were gathered from the analysis. For a comprehensive view of how the dataset was explored and supplementary results, please refer to the original R script.

### Statistical Overview

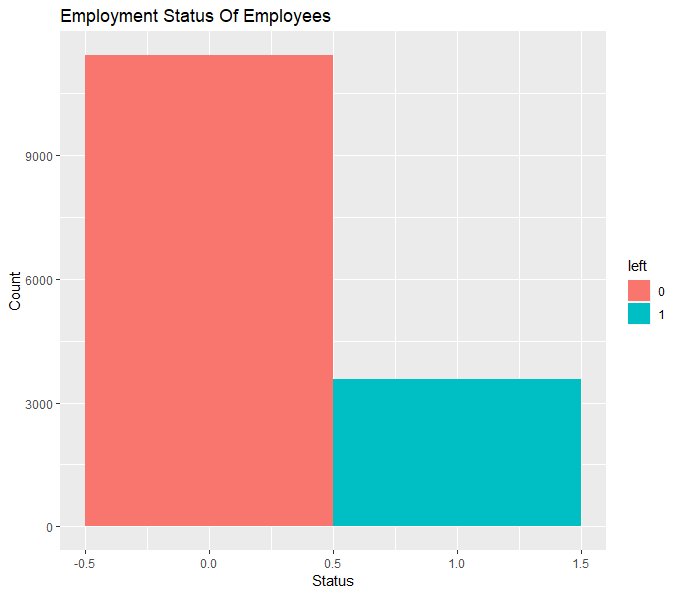
First, to understand the company’s attrition rate, bar chart was also applied to plot the frequency distribution of employees who left (1) and stayed (0) with the company. In all employee records, 11428 people remained with the company, which accounts for 76% of the total population. Thus, the company had a calculated turnover rate of 24%(Table 7).

Figure 14: Frequency distribution of employees who left and stayed with the company

Table 7: Frequency of employees that stayed and departed from the company

|  |  |
| --- | --- |
| Status | % Freq |
| Stayed | 76% |
| Departed | 24% |

>table(percent\_stayed,percent\_left)

### Exploring Company’s Business and Department Infrastructure

The business nature of the company was then explored by examining the distribtion of employees was evenly across different departments (Figure 15). As depicted in Figure 1A, the departments are imbalanced in terms of employee counts. It appears that sales, technical and support account for the majority of the employee dataset, suggesting that this could be a technical company that has a technical product which could require sales, support and ongoing development and maintenance.

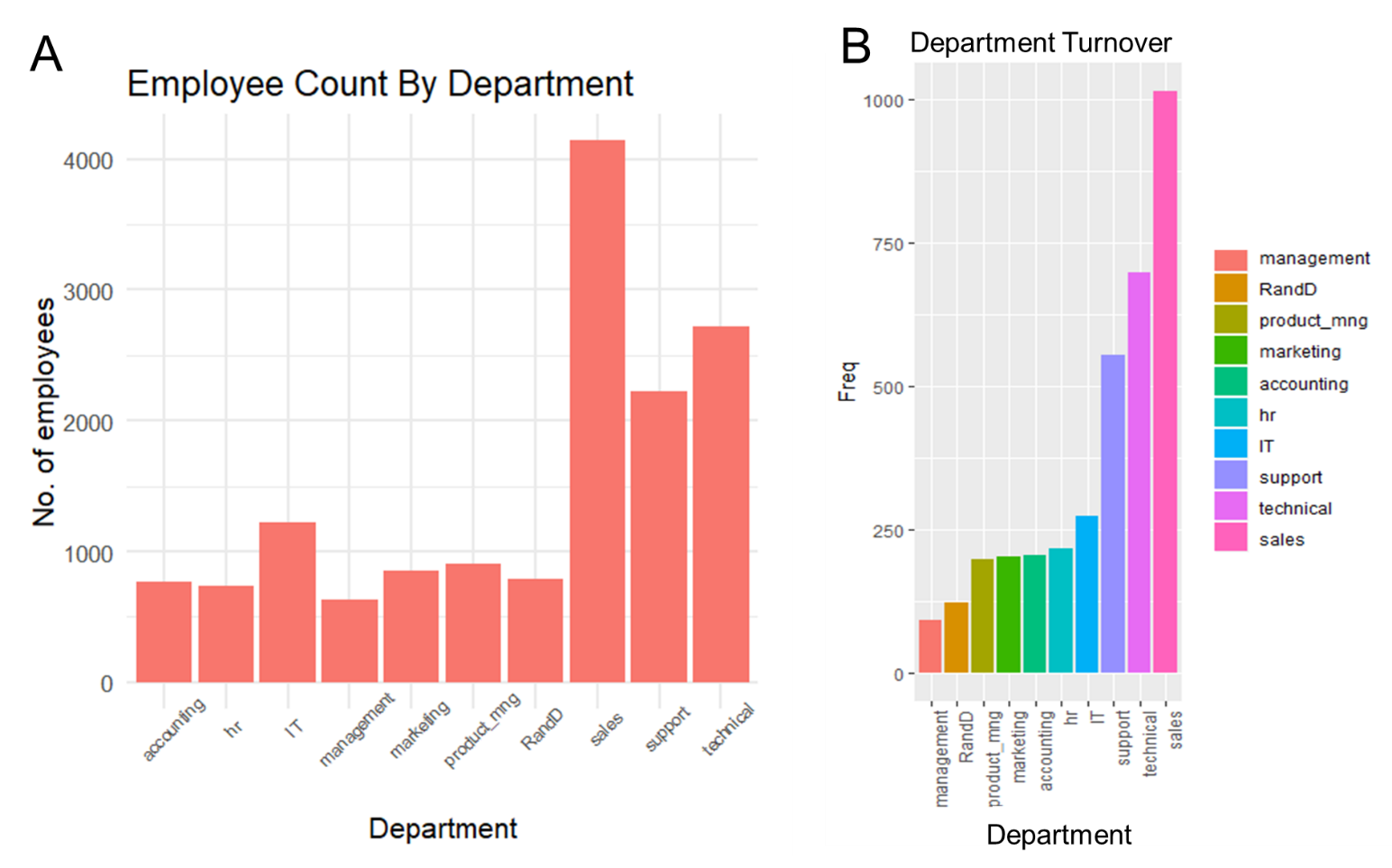
****We also observe that these departments also have the highest employee turnover rate (Figure 1B). With a This would suggest that there is a fast hiring and high recruitment process within these departments. Perhaps the working environment for these new recruits are not optimal and it would be interesting to further explore a more direct underlying cause for this. By contrast, we observe the lowest attrition rate within the management department.

Figure 15: (A) Distibution of employees across different departments. (B) Distribution of employees who left across different departments.

However, when compared as a percentage of employee who left by department, HR is the one with highest attrition rate (Table 7).

Table 8: Relative Frequency of resigned employees by department

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| hr | accounting | technical | support | sales | marketing | IT | Product\_mng | RandD | management |
| 27% | 29%% | 22% | 25% | 24% | 24% | 22% | 22% | 15% | 14% |

> percent\_left\_within\_dept<- round((dept\_turnover\_left/dept.table)\*100)

### Correlation matrix and heatmap

As part of the preliminary analysis, one of the main objectives was to determine which key attributes affected employee turnover the most for this company. To achieve this, an initial correlation analysis for numerical variables was performed of all the employees (Figure 16 A) and employees who resigned (Figure 16B) and was visualised using a heatmap matrix. In Figure 16A, a strong negative correlation exists between employee turnover and satisfaction levels. As represented by both heatmaps, several positive relationships were observed between “number\_project”, “average\_monthly\_hours” and “evaluation\_score” suggesting that the company may highly evaluate those who worked longer hours and completed more projects. However, overall, only significant correlation was found between these attributes across the subgroup of departed employees. As these appear the most intuitive for predicting employee attrition, they were selected for this study for the further exploratory analysis in the next section.

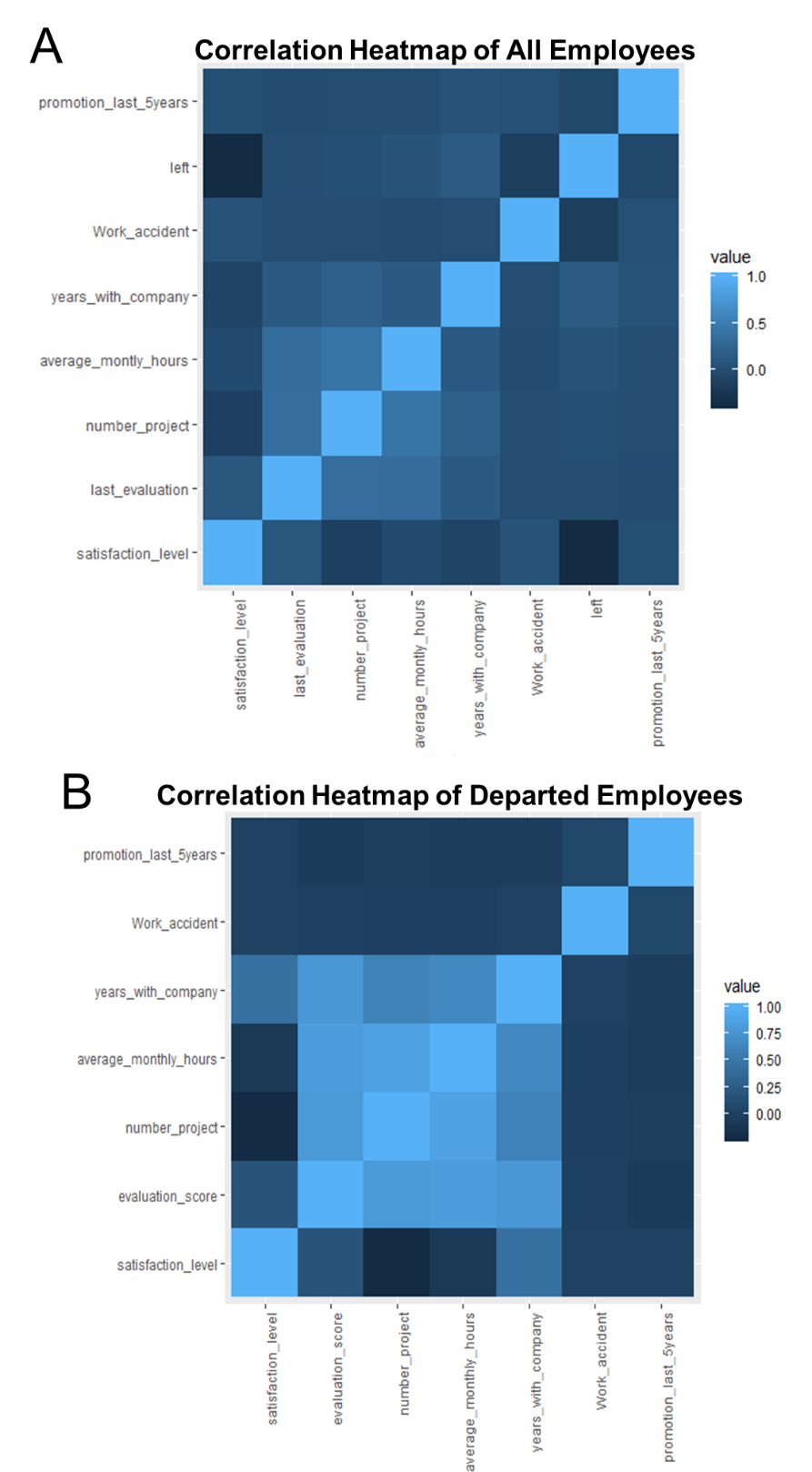


Figure 16: Heatmap of Correlation Matrices A) Correlation heat map of all company employees (B) Correlation heatmap of employees who left the company

### Distributions of Satisfaction, Evaluation & Average Monthly Hours

Based on previous findings, the distributions of “satisfaction level”, “evaluation\_score” and “average\_monthly\_hours” were first plotted using histograms and compared between employees who stayed with (Figure 4A.) and left from the company (Figure 4B). Using an unpaired student t-test, the mean evaluation scores and mean average number of hours worked between employees who left and stayed were significantly different.

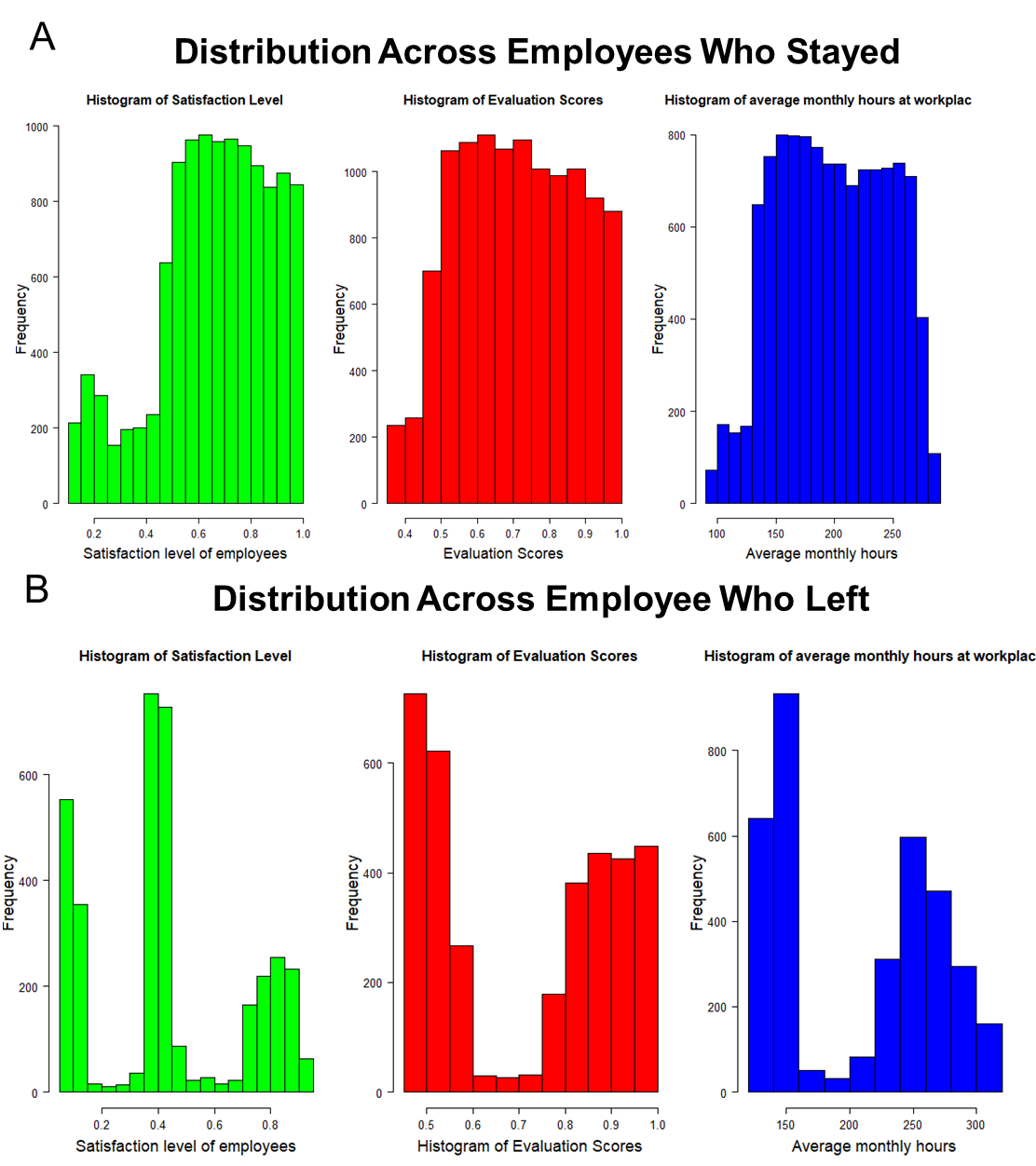


Figure 17: Distribution of satisfaction levels, evaluation score, and average monthly hours across employees (A) Employees who stayed with company (B )Employees who left from company.

The shape of these distributions across employees that left are very interesting compared with those that stayed. Most of the distributions are normal for those that stayed except the distribution of satisfaction levels which appears slightly skewed to the right (Figure 17A).

Regarding employees that left, none of the distributions appear normal but are all instead quite polarized (Figure 17B). A bimodal distribution is observed for employees that had low (<0.5) and high evaluation scores (>0.75). In parallel to this, two similar sharp peaks are observed at the ends of the “average monthly hours” histogram. This reveals the existence of two subgroups that had either worked a lot (>250 hours) or below average (<150 hours.)The similar distributions patterns of these attributes are consistent with observations made previously from the heatmap. Regarding satisfaction level a tri-modal distribution was observed among departed employees. Overall, the results from this analysis suggest that certain employees who left company could be grouped by certain features.

#### Turnover vs. Number\_Project

Next, the frequency distribution patterns of employees by the number of project was explored (Figure 18).

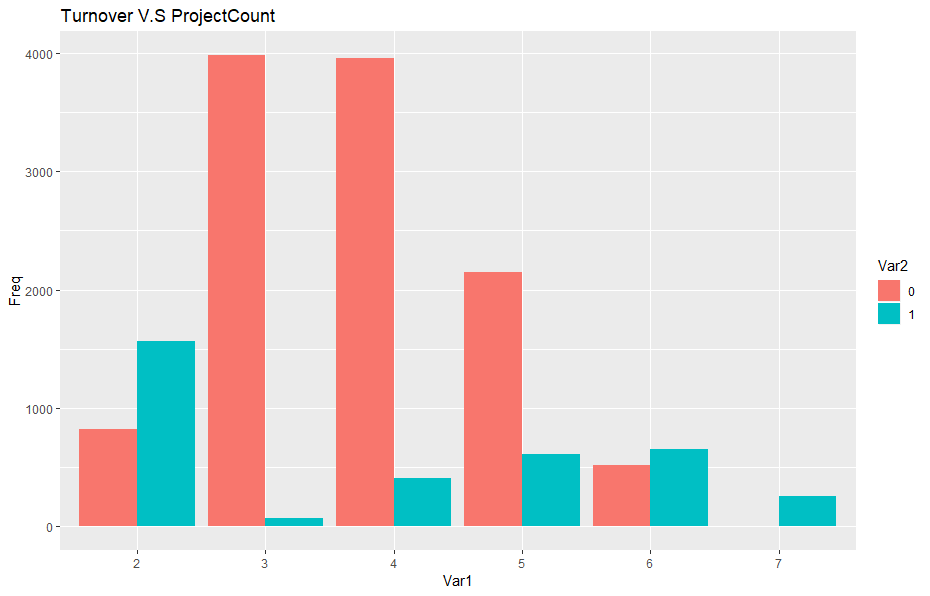


Figure 18: Frequency distribution of employees who stayed (0) and left (1) by project count.

Again, a similar distribution of employees leaving at the high and low spectrum of the curve is observed. It is also worth noting that none of the employees with 7 projects remained with the company, suggesting that a higher workload could be associated with employee turnover. By contrast, those tasked with less assignments may have felt underchallenged or lowly valued by the company, thus leaving as a result.

#### Number\_Project vs Average\_Monthly\_Hours and Evaluation\_Score

Next, further examination was carried out to compare the relationships between employee turnover against “number project” and “average\_monthly\_hours” and employee turnover against “number\_project” vs “evaluation\_score” (Figure 19). In figure 19 A., the average monthly hours of employees increased as they took on more projects, particularly those who left the company. This may suggest that their work life balance may have been compromised with extra hours as they worked to meet multiple project deadlines. Interestingly, in contrast, those who remained with the company maintained consistent working hours, despite the increased number of projects. This suggests that employees with more project would tend to stay with company longer if they didn’t work overtime and had a better work-life balance.

Interestingly, employee turnover by “number\_project” vs “evaluation\_score” in figure 19 B. is almost parallel to the previous results. In a similar pattern, there is a drastic change in the evaluation scores of employees when they have more than 3 projects. Employees who undertook 2 projects scored were poorly esteemed by the company and left. Those who worked on more projects and had high evaluation scores had also departed. A consistent performance score was observed amongst employees who stayed with increasing project number. On average, employees who left were deemed as better performers by the company than those who stayed suggested a loss of valuable workers.

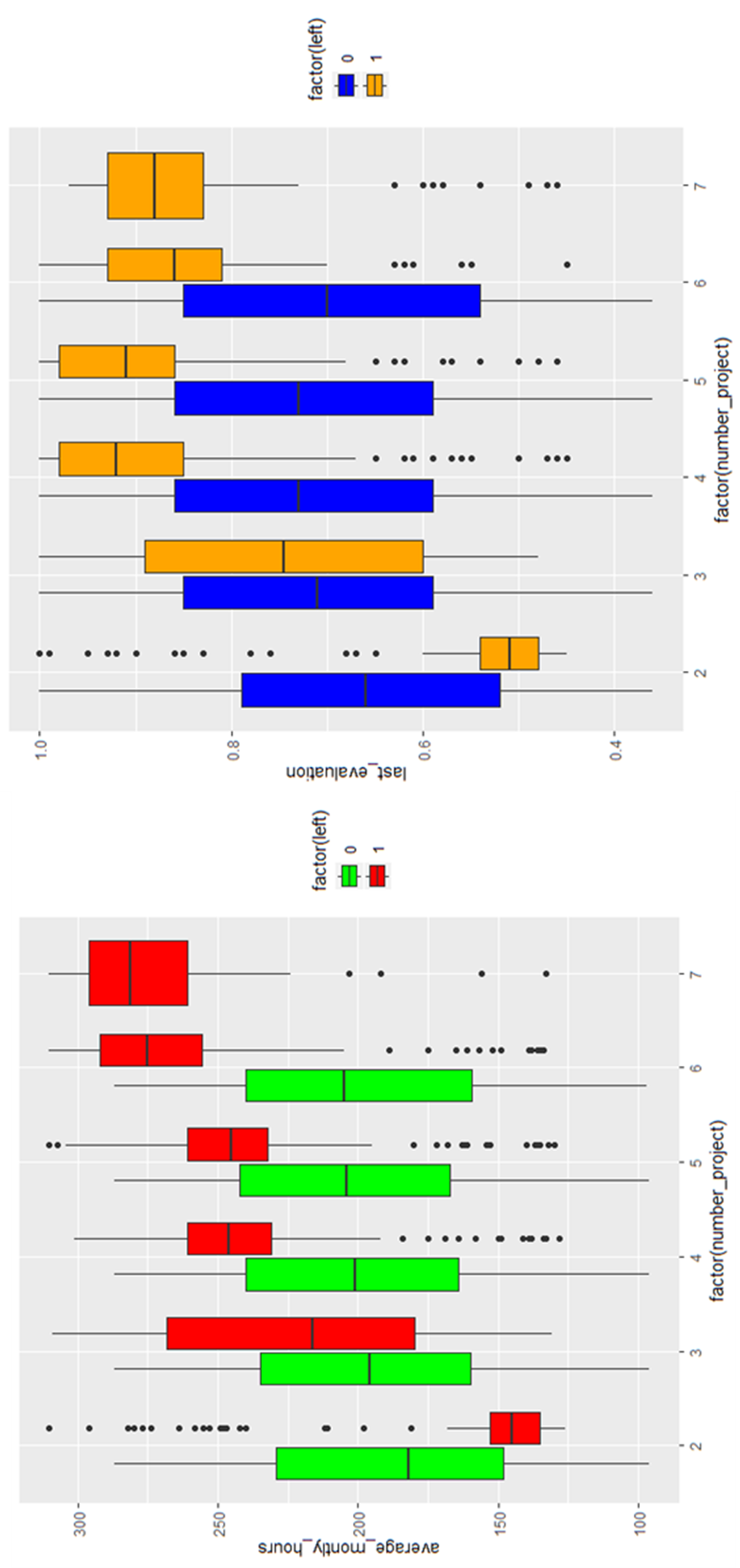


Figure 19: Boxplots of employee turnover by (A) project counts vs average monthly hours (B) project counts vs average monthly hours.

#### Satisfaction vs Evaluation score and Average monthly working hours

The satisfaction levels of departed employees were plotted against their evaluation scores and their average hours worked per month, revealing compelling insights as to why they may have left the company (Figure 20 A and B).

From this, three distinct clusters were observed. The first cluster (Cluster 1) may represent employees that were hard-working, as indicated by their high evaluation scores (>0.75,) and high number of working hours (>237.5 hrs/month). However, with a low satisfaction score of <0.2 they may have felt unhappy with their situation, possibly because they could have felt overworked. The second cluster (Cluster 2) describes a group of employees that were deemed as poor performers, lazy and unhappy, as reflected by their low evaluation scores (<0.5), below average number of hours worked (<170) and low satisfaction levels (<0.45), respectively. This group may have left because they were bad workers or else may have possibly felt under challenged and unengaged with less work. The third upper right cluster (Cluster 3) reveals a group of highly satisfied employees (~ >0.7) who were evaluated highly for their performance (>0.8) and were hard-working, as reflected by their high average working hours every month (>212.5 hrs/month and <275 hrs/month). This group may have left because they found another job opportunity elsewhere.

Furthermore, a multivariate comparison of “satisfaction levels”, “evaluation\_score” and “average\_monthly\_hours” was performed in order to better visualise the relationships between these variables amongst this cohort of employees (Figure 20). From this figure, we observe a clearer image of these three distinct populations. An interactive web-based version of this 3D model plot can be view from the original R script or else directly accessed via the link <https://plot.ly/~purcelsi/1/#/>.

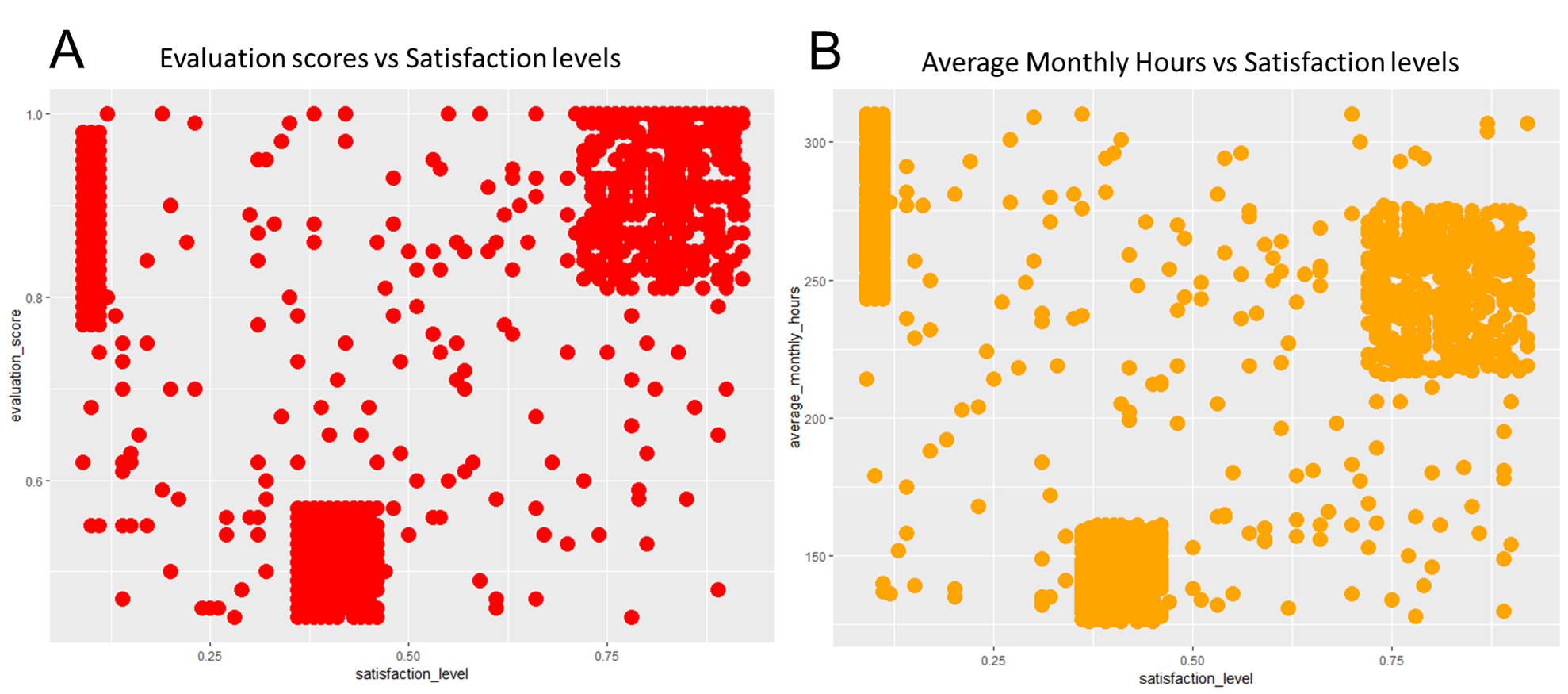


Figure 20: Population clusters of employee turnover by (A) satisfaction­\_levels vs evaluation\_score (B) satisfaction\_level vs average\_monthly\_hours

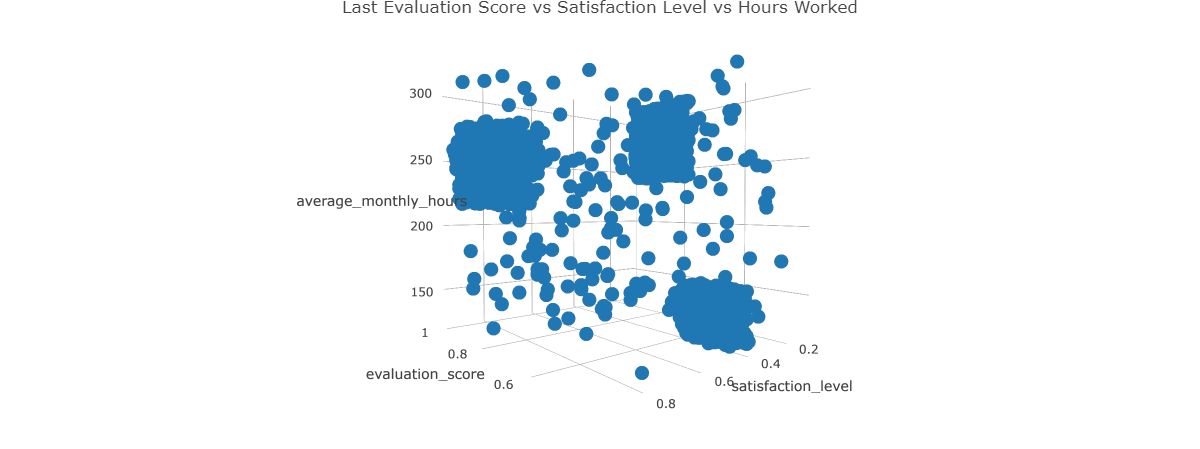


Figure 21: A multivariate comparison between Evaluation\_Score, Satisfaction\_level and average\_monthly\_hours.

### Characterising employees and predicting who may leave company next

Findings from the previous analysis results indicate that certain employees who left the company can be categorized into specific subgroups, according to their shared range of values for the variables “Evaluation\_Score”, “Satisfaction\_Level”, and Average\_monthly\_hours”. The unique features of each cluster was then used to define the potential work behaviour of the company’s current employees and predict their future with the business. Using a for loop and an if/else statement, employees that shared the same combination of values for either of these clusters were categorised as either “Ideal and Ambitious”, “Ideal and Sad/Stressed” and “Poor employee”, with values being stored inside a new feature column called “employee\_type”. A summary table for each of these cluster values and their assigned work category type is summarized in Table 8.

Table 9: Summary of variable values of each employee subgroup that left and their assigned employee type behaviour

|  |  |  |  |
| --- | --- | --- | --- |
| satisfaction\_level | Evaluation\_Score | Average\_Monthly\_Hours | Employee\_type |
| >0.7 | >0.8 | >212.5 & <275 | Ideal and Ambitious |
| >0.35 | >0.45 | >0.45 | Ideal and Sad/Stressed |
| >0.2 | >0.75 | >237.5 | Poor employee |

Employees that were classified as “Ideal and Ambitious” and “Ideal but Sad/Stressed” would be deemed as valuable to the company and were further isolated to determine their frequency. The results from Figure 22 would indicate that 7.6% of the company’s current most ideal employees may leave soon.

Thus, analysing employee data can be useful for the company to identify valuable employees who may leave the company soon. The findings from this analysis could provide actionable insights about employee attrition and assist Human Resources to generate better policies for improving employee retention by creating a better work environment. The company could perhaps mitigate the potential departure of some of their most valuable employees by exploring their other work-associated attributes such as salary levels and promotions.

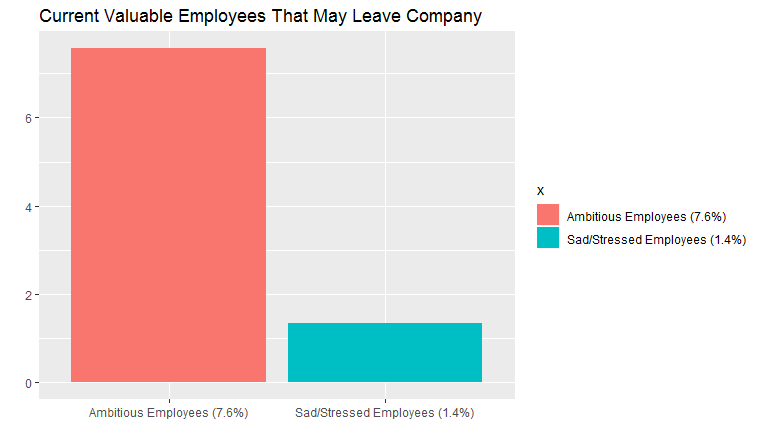


Figure 22: Percentage of current employees that may leave according to the clustering features of the valuable employees that departed.

## Conclusions and Summary of Results

Overall, the following conclusions can be derived from this exploratory analysis:

* Satisfaction plays is the highest driver for employee turnover.
* Employees carrying our 2, 6 or 7 projects are likely to quit from the company.
* Resigned employees can fall into three distinct subgroups based on their satisfaction levels, evaluation score and average number of working hours.
* Employees typically left when they were either underworked (> 150hr/month) or when they are overworked (> 250 hr/month).
* Employees that are either very poor or high performers may contribute to a high attrition rate.
* 7.6% of the company’s current most ideal employees may leave soon.

# References cited

## Titanic dataset

### Material for literature review

British Board of Trade. 1990. Report on the Loss of the ‘Titanic’ (S.S.). British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing.

Everett, M., & Neill, H. 1912. (eds). The Story of the Wreck of the Titanic (memorial edition), Chicago: Homewood Press.

*Titanic: Women and Children First*, by Judith B. Geller (Norton, 1998)

Kitagawa, E.M. and Hauser, P.M., 1973. Differential mortality in the United States: A study in socioeconomic epidemiology.

Bijan, A., 2014. Titanic and the People on Board: A Look at the First, Second, and Third Classes.

John R. Henderson, (1998) *Demographics of the TITANIC Passengers: Deaths, Survivals, Lifeboat Occupancy*, Ithaca College Library. <http://www.icyousee.org/titanic.html> Last Accessed [30/05/2016]

### Supporting material For R studio data analysis

Kaggle. 2015. *Data –Titanic: Machine Learning from Disaster.* Available at <https://www.kaggle.com/c/titanic/data> Last Accessed [30/02/2019]

*I would like to acknowledge the learning I received from several sources including kernals from the following websites and their authors*:

**David Langer**

<https://github.com/EasyD/IntroToDataScience> Last Accessed [30/02/2019]

**Erik Bruin**

<https://www.kaggle.com/erikbruin/titanic-2nd-degree-families-and-majority-voting> Last Accessed [30/02/2019]

**Fabrice Tereszkiewicz**

<http://www.rpubs.com/fatz/titanic-exploration> Last Accessed [12/02/2019]

## Human Resources dataset

### Material for literature review

Cotton, J.L. and Tuttle, J.M., 1986. Employee turnover: A meta-analysis and review with implications for research. *Academy of management Review*, *11*(1), pp.55-70.

Ribes, E., Touahri, K. and Perthame, B., 2017. Employee turnover prediction and retention policies design: a case study. *arXiv preprint arXiv:1707.01377*.

Tracey, J.B. and Hinkin, T.R., 2008. Contextual factors and cost profiles associated with employee turnover. *Cornell Hospitality Quarterly*, *49*(1), pp.12-27.

Hinkin, T.R. and Tracey, J.B., 2000. The cost of turnover: Putting a price on the learning curve. *Cornell hotel and restaurant administration quarterly*, *41*(3), pp.14-21.

Simons, T. and Hinkin, T., 2001. The effect of employee turnover on hotel profits: A test across multiple hotels. *Cornell Hotel and Restaurant Administration Quarterly*, *42*(4), pp.65-69.

Barnes, G., Crowe, E. and Schaefer, B., 2007. The Cost of Teacher Turnover in Five School Districts: A Pilot Study. *National Commission on Teaching and America's Future*.

Jones, C.B., 2005. The costs of nurse turnover, part 2: application of the nursing turnover cost calculation methodology. *Journal of Nursing Administration*, *35*(1), pp.41-49.

### Material For R data analysis

The original dataset file from Kaggle competition Human Resources Analytics. (However this dataset is no longer publicly available since last access date) <https://www.kaggle.com/deleted-dataset/358> Last Accessed and Downloaded [12/03/2019]

*I would like to acknowledge the learning I received from several sources including kernals from the following websites and their authors*:

**Randy Lao**

<https://www.linkedin.com/pulse/learning-data-science-kaggles-human-resources-analytics-randy-lao-1/> Last Accessed [10/04/2019]

**Giusti Elena, Sarah Wong, Zohaib Gulzar, Tatsuya Nagata** [http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2017/AnalysisofHR[1].html](http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2017/AnalysisofHR%5b1%5d.html) Last Accessed [02/04/2019]

**Lee Gang**

<http://www.rpubs.com/fechard/HR_Analytics> Last Accessed [10/04/2019]