INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

TYPE INTAKE CODE

HAND OUT DATE: 25 JULY 2022

HAND IN DATE: 8 AUGUST 2022

WEIGHTAGE: 50%

INSTRUCTIONS TO CANDIDATES:

1 Submit your assignment at the administrative counter

2 Students are advised to underpin their answers with the use of APA (American Psychological Association) citation

3 Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld

4 Cases of plagiarism will be penalized

5 The assignment should be bound in an appropriate style (comb bound or stapled).

6 Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where

appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.

7 You must obtain 50% overall to pass this module.

Table of Contents

1. Introduction and Assumptions

This assignment required us to do data analytics and investigate a dataset which consist of staff data within an organization, in order to identify potential unclosed issues inside the human resources management and provide meaningful insights available for decision making using RStudio. In this dataset, there's a total of 18 columns and 49654 rows which includes personal detail of the staff. The assumption in this case is the dataset consist of clean data, so it is unnecessary to do data cleaning, however, to get proper insights, solid data exploration is required along with using data visualization and manipulation to prove certain points.

2. Data Import/ Pre-processing / Exploration

2.1 Data Import

```
#Import Data set, pre-processing
employeeAttrition = read.csv("C:\\Users\\yhcominthru\\Desktop\\Year 2 Sem 1\\
                             PFDA\\ASsignment\\employee_attrition.csv", header = TRUE)
```

Figure 1: Source code of importing csv file into the R console

Firstly, data import of csv format of the excel file has been completed by using the read.csv function that reads the csv file. The header is set to true to ensure RStudio accepts the headers as parameters.

2.2 Data Pre-processing

```
library(ggplot2)
library(crayon)
library(plotrix)
library(dplyr)
library(tidyr)
library(stringr)
```

Figure 2: Source code of the libraries used (pre-processing)

```
names(employeeAttrition) = c("ID", "Record Date", "Birth Date", "Hired Date", "Termination Date", "Age",
                             "Length of Service(Year)", "City", "Department", "Job Title",
                             "Store Name", "Gender", "Gender(Full)", "Reason of Termination",
                             "Type of Termination", "Year of Status", "Employee Status", "Business Unit")
```

Figure 3: Source code of re-naming the columns (pre-processing)

Prior of doing proper data exploration to obtain useful information, data processing, which is preformed of raw data is crucial as the data is transformed into a format that can be processed easily and effectively. (Lawton, 2022) This has been completed by ensuring that the proper libraries required in this session is prepared and renaming the column names within the dataset.

2.3 Data Exploration

Data exploration is the following step as it provides benefits in terms of allowing better decisions, having a broad understanding of the business or company, and familiarizing oneself with the data. (“What Is Data Exploration?”, 2020)

```
#structure
str(employeeAttrition)

#number of rows and columns
dim(employeeAttrition)

#Column Names
names(employeeAttrition)

#Summary of the data
summary(employeeAttrition)
```

```
unique(employeeAttrition$City)
unique(employeeAttrition$`Store Name`)
unique(employeeAttrition$`Reason of Termination`)
unique(employeeAttrition$`Type of Termination`)
unique(employeeAttrition$`Business Unit`)
unique(employeeAttrition$Department)
unique(employeeAttrition$`Job Title`)
unique(employeeAttrition$`Employee Status`)
```

```
terminationList = data.frame(filter(employeeAttrition, employeeAttrition$`Reason of Termination` == "Layoff"))
View(terminationList)

resignationList = data.frame(filter(employeeAttrition, employeeAttrition$`Reason of Termination` == "Resignaton"))
View(resignationList)

retirementList = data.frame(filter(employeeAttrition, employeeAttrition$`Reason of Termination` == "Retirement"))
View(retirementList)

naList = data.frame(filter(employeeAttrition, employeeAttrition$`Reason of Termination` == "Not Applicable"
                           & employeeAttrition$`Employee Status` == "ACTIVE"))
View(naList)


orderedByID_Year =  data.frame(employeeAttrition[order(employeeAttrition[,1], -employeeAttrition[,7]),  ])
View(orderedByID_Year)

removeDuplicates = orderedByID_Year[!duplicated(employeeAttrition$ID), ]
View(removeDuplicates)

uniqueList = unique(rbind.data.frame(removeDuplicates, terminationList, resignationList, retirementList))
View(uniqueList)
```

Figure 4: Source codes of data exploration

The application of data exploration on this dataset was started by getting to know the basic structure of the dataset, which are the structure, dimension, column names and summary of the data recorded. Moving on, I also used unique() function on certain columns that have the character data type to get a clearer overview on what are the options recorded in each column to seek on potential perspectives to dive deeper when it comes to analyzing this dataset.

With that mentioned, I noticed that employees' record is input every year, hence there's multiple entries with the same employee ID, until they resigned, being laid off or retired. In order to get the final list where the latest employees' records are the ones remain, the employeeAttrition data frame is ordered ascending by ID, then descending by the length of service(year), where the latest record of a particular employee is available. After that, every entry that consist of a duplicate employee ID is being removed using !duplicated() function, and it is stored in removedDuplicates. Lastly, the unique function is reused after binding the removeDuplicates along with data frames that each consist of different type of reason of termination. In order to confirm the final data frame, which is uniqueList, is complete and filtered out the unwanted duplicate ID entries, View() function is used along with filtering the data frame using different reasons of termination, the numbers matched when the same is applied to the first data frame (employeeAttrition), hence it can be confirmed to contain the up-to-date data of every employee.

3. Data Analysis

Data analysis is the next step to be completed to gather information from this dataset. 5 questions have been prepared and data visualization is required to understand the given data, which can be completed using packages like ggplot2, dplyr to create charts and graphs for better comprehension.

3.1 Question 1: Why do the employees want to resign?

3.1.1 Analysis 1: The amount of people resigned by age

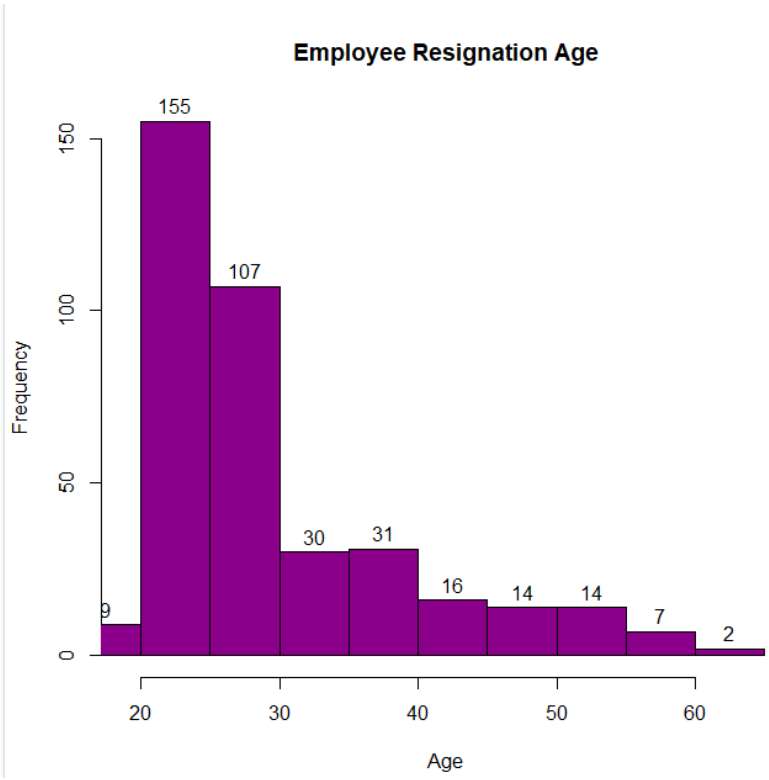
```
#Question 1: Why do the employees want to resign?
#Analysis 1.1: The amount of people resigned by age

employeeResignationAge = c(resignationList$Age)
employeeAgeHist = hist(employeeResignationAge, main = "Employee Resignation Age",xlab = "Age", xlim = c(19,65), col = "darkmagenta")
text(employeeAgeHist$mids, employeeAgeHist$counts, labels = employeeAgeHist$counts, adj=c(0.5,-0.5))


resignationListMale = data.frame(filter(resignationList, resignationList$Gender == "M"))
resignationListMaleAge = c(resignationListMale$Age)
resignationListMaleHist = hist(resignationListMaleAge, main = "Male Employees Resignation Age",xlab = "Age", xlim = c(19,65), col = "blue")
text(resignationListMaleHist$mids,resignationListMaleHist$counts, labels = resignationListMaleHist$counts, adj=c(0.5,-0.5))
```

Figure 5: Source code of Analysis 1.1

In Analysis 1.1, the quantity of resignations for each age group is being searched for. In order to obtain a dataset with only people that resigned, data exploration and filtering of the dataset has been done using functions such as View(), filter(), which was mentioned in section 2.3, then vectorizing one of the results and the dataset required for this analysis, which is resignationList, and created a histogram using hist() which is followed by the application of text() function to add labels regarding the frequency of each bar. The structure is applied to find out the number of males and females resigned in each age group.


Employee Resignation Age
Frequency
0
50
100
150
155
107
30
31
16
14
14
7
2
9
20
30
40
50
60
Age


Male Employees Resignation Age
0
20
40
60
80
89
21
16
17
7
8
10
3
3
20
30
40
50
60
Age

Figure 6: Results of Analysis 1.1

The result shows that most of the employees resigned within the age group of 20 to 30 (262/385, approximately 68%), with 110 of them being male (110/262, approximately 41%). It is assumed that employees that resign in this age group are mainly resigning after internship, part time job or looking for another job opportunity. One fact that is worth investigating more is the fact that the first half of this age group has more males that resigned (89/155, approximately 57%); however, the latter half of this age group has more females that resigned (86/107, approximately 80%). With this assumption and fact in mind, the ex-employees' jobs for each gender should be investigated.

3.1.2 Analysis 2: The amount of people resigned based on their jobs

```
#Analysis 1.2: The amount of people resigned based on their jobs
jobCountAndGender = data.frame(table(resignationList$Job.Title, resignationList$Gender))
names(jobCountAndGender) = c("Job", "Gender", "Frequency")
jobCountAndGender$Gender = factor(jobCountAndGender$Gender, levels = c("M", "F"))
jobCountAndGenderOrdered = data.frame(jobCountAndGender[order(jobCountAndGender$Job),])
View(jobCountAndGenderOrdered)

jobs = c(jobCountAndGenderOrdered$Job)
frequency = c(jobCountAndGenderOrdered$Frequency)
gender = c("F", "M")

barplot(frequency, names.arg = jobs, xlab = "Jobs", ylab = "Frequency", col = c("red", "blue"),
        main = "Employees Resigned according to jobs", border = "black")
legend("topright", legend = gender, cex = 1.3, fill = c("red", "blue"))
```

Figure 7: Source code of Analysis 1.2

In analysis 1.2, I'm looking for the total amount of people resigned according to their jobs. I started things of by creating a table from the job title column and gender column and then convert it to a data frame, along with renaming the column names using names() function. With this, the frequency of each repeating occurrence is accumulated. Next up, I re-order the factors in the gender column then I arrange the data frame using order() function, to ensure that the same jobs are grouped along with proper gender order. From there onwards, I vectorized each column and came up with a bar plot using the barplot() function, along with legend() function to demonstrate the legend.

	Job	Gender	Frequency
1	Baker	F	14
10	Baker	M	16
2	Cashier	F	84
11	Cashier	M	95
3	Dairy Person	F	43
12	Dairy Person	M	21
4	HRIS Analyst	F	0
13	HRIS Analyst	M	1
5	Meat Cutter	F	13
14	Meat Cutter	M	16
6	Processed Foods Manager	F	1
15	Processed Foods Manager	M	0
7	Produce Clerk	F	14
16	Produce Clerk	M	13
8	Shelf Stocker	F	41
17	Shelf Stocker	M	12
9	Store Manager	F	1
18	Store Manager	M	0

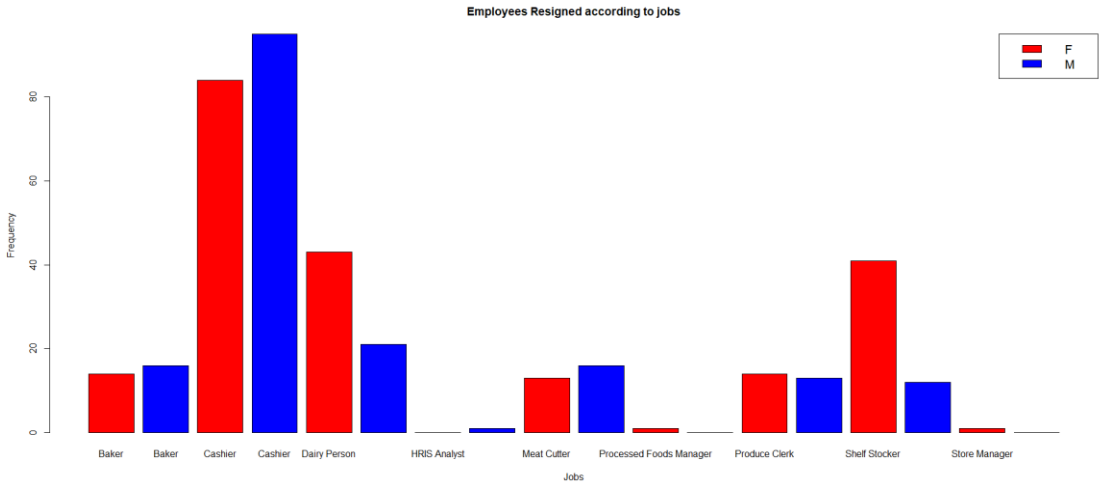


Figure 8: Result of Analysis 1.2

From Figure 8, the people that resigned were mainly having lower skilled jobs (cashier, dairy person, shelf stocker, occupying 296 out of 385 resignations), whereas higher skilled jobs such as HRIS Analyst and managers have drastically lesser resignations, with a total of 3 resignations, out of 385 of them. According to Banks (2018), a job is defined as low skilled job when it doesn't require college degree nor

specialized training, only might require on-the-job training, which usually takes up less than one month. In addition, these type of jobs does not provide enough income to cover expenses due to their low skill requirement. With this fact discovered and integrating it with the discoveries in analysis 1.1, it's safe to say that the majority of the employees resigned from low-skilled, hard labor jobs, whether it's part-time or full time, to find other jobs with better salary, and more room to improve themselves in terms of skill set. This scenario stands out more in the resignations of females who were working as shelf stockers as that job takes up a lot of energy and it takes a toll on females' bodies who are genetically more delicate than males.

3.1.3 Analysis 3: The average resignation age according to jobs

```
resignedBakers = resignationList[which(resignationList$Job.Title == 'Baker'),]
resignedCashiers = resignationList[which(resignationList$Job.Title == 'Cashier'),]
resignedDairyPersons = resignationList[which(resignationList$Job.Title == 'Dairy Person'),]
resignedHRISAnalysts = resignationList[which(resignationList$Job.Title == 'HRIS Analyst'),]
resignedMeatCutters = resignationList[which(resignationList$Job.Title == 'Meat Cutter'),]
resignedFoodManagers = resignationList[which(resignationList$Job.Title == 'Processed Foods Manager'),]
resignedClerks = resignationList[which(resignationList$Job.Title == 'Produce Clerk'),]
resignedShelfStockers = resignationList[which(resignationList$Job.Title == 'Shelf Stocker'),]
resignedStoreManagers = resignationList[which(resignationList$Job.Title == 'Store Manager'),]

averageJobResignationAge = c(mean(resignedBakers$Age), mean(resignedCashiers$Age),mean(resignedDairyPersons$Age), mean(resignedHRISAnalysts$Age),
                             mean(resignedMeatCutters$Age), mean(resignedFoodManagers$Age), mean(resignedClerks$Age), mean(resignedShelfStockers$Age), mean(resignedStoreManagers$Age))
ResignationAgeRounded = round(averageJobResignationAge, digits = 2)
resignedJobsList = c('Baker', 'Cashier',  'Dairy Person', 'HRIS Analyst', 'Meat Cutter', 'Processed Foods Manager','Produce Clerk', 'Shelf Stocker','Store Manager' )
averageAgeAndJob = data.frame(ResignationAgeRounded, resignedJobsList)
names(averageAgeAndJob) = c("Average Age", "Job")
View(averageAgeAndJob)
ggplot(averageAgeAndJob, aes(x=averageAgeAndJob$Job, y=averageAgeAndJob$`Average Age`)) +
  labs(title="Average Resignation Age According to Jobs", x="Job Titles", y = "Average Age")+
  geom_bar(stat="identity", color="black", fill="blue", width=0.5) +
  geom_text(aes(label=averageAgeAndJob$`Average Age`), vjust=1.6, color="white", size=3.5) +
  theme_minimal()
```

Figure 9: Source Code of Analysis 1.3

In analysis 1.3, I came up with new data frames according to their jobs, calculated the mean age for each of those tables and made a new table using that data. I used ggplot to come up with the bar plot that shows this data.

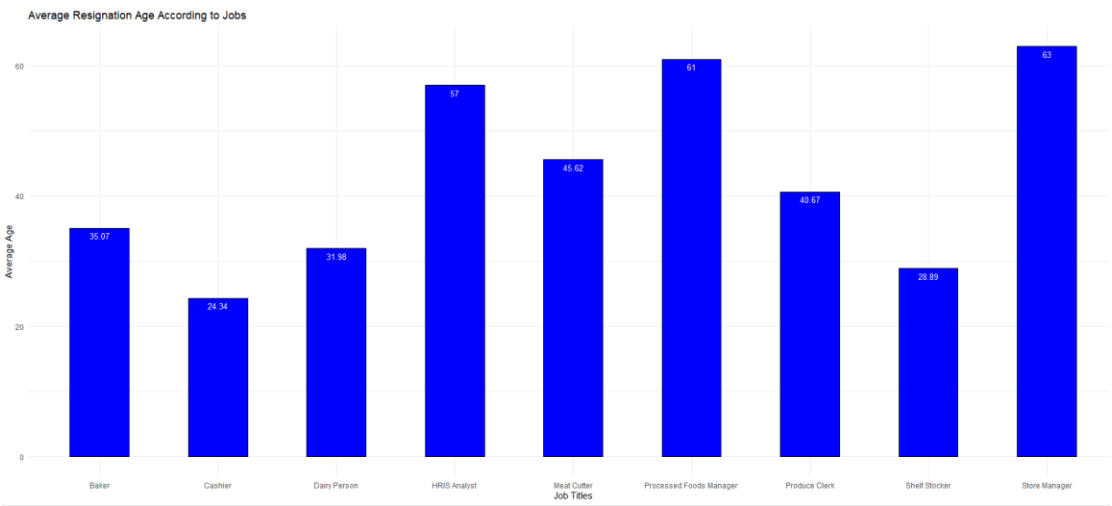


Figure 10: Result of Analysis 1.3

From figure 10, we can see that lower skilled jobs, which are cashiers, dairy persons and shelf stockers have a lower average resignation age, which aligns to the analysis in 1.2, where they need more income, a job that provides them with more lessons and opportunities or not having enough stamina to take on these jobs, whereas jobs that requires expertise, from baker to analysts and managers, all have higher average resignation age.

3.1.4 Analysis 4: The amount of people resigned for each store & unit

```
#Analysis 1.3: The amount of people resigned based on store and unit
orderedByStoreDeparmentUnit = data.frame(resignationList[order(resignationList$Store.Name, resignationList$Business.Unit), ] )
View(orderedByStoreDeparmentUnit)

storeNameAndUnit = data.frame(table(resignationList$Store.Name, resignationList$Business.Unit))
names(storeNameAndUnit) = c("Store Name", "Unit", "Frequency")
storeNameAndUnitOrdered = data.frame(storeNameAndUnit[order(storeNameAndUnit[,1], storeNameAndUnit[,2]), ])
View(filter(storeNameAndUnitOrdered, storeNameAndUnitOrdered$Frequency != 0))

ggplot(storeNameAndUnitOrdered, aes(fill=`Unit`, y=`Frequency`, x=`Store.Name`)) +
  geom_bar(position="stack", stat="identity")  +
  labs(title="Job Resignation for Each Store", x="Stores", y = "Frequency") +
  guides(fill = guide_legend((title = "Unit")))
```

Figure 11: Source Code of Analysis 1.4

In my last analysis for question 1, I'm looking at the amount of people resigned according to store name and unit. In order to achieve that, I made a table using store name and business unit to calculate the amount for each store and classify them by business unit.

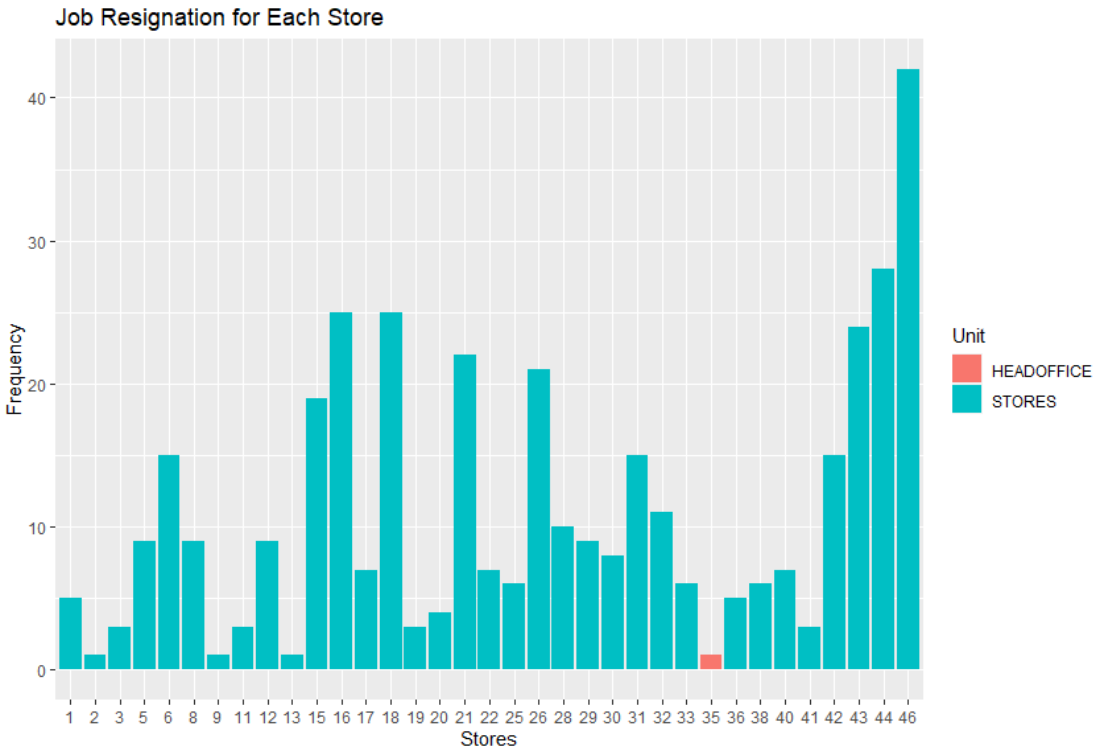


Figure 12: Result of Analysis 1.4

According to figure 12, only store 35 from head office unit has people resigned whereas the rest are resignations from different stores with store 46, 44, 43, 15, 18 21 and 26 having more resignations.

Throughout the 4 analyses in question 1, I found out that most resignations came from the age group of 20 to 30, and majority of them are having lower skilled jobs, which then lead to lower skilled jobs having a lower number of average resignation age, whereas higher skilled jobs have higher resignation age. Lastly, those that are working lower skilled jobs are mostly scattered in the stores

3.2 Question 2: Why do the employees get laid off?

3.2.1 Analysis 1: The number of employees that were laid off every year

```
#Analysis 2.1: The number of employees that were laid off every year
terminationList = data.frame(filter(employeeAttrition, employeeAttrition$`Reason of Termination` == "Layoff"))
view(terminationList)
terminationByYear = data.frame(terminationList[order(terminationList$Year.of.Status), ])
view(terminationByYear)

ggplot(terminationByYear, aes(x=as.factor(`Year.of.Status`))) +
  geom_bar(stat ="count") + labs(title = "Number of termination by year", x = "Year", y = "Quantity" )
```

Figure 13: Source Code of Analysis 2.1

In our first analysis for question 2, I wanted to look for the quantity for the laid off employees every year. I did this by filtering out other rows of data that doesn't have layoff as their reason of termination and I use ggplot on the year of status to get the desired results.

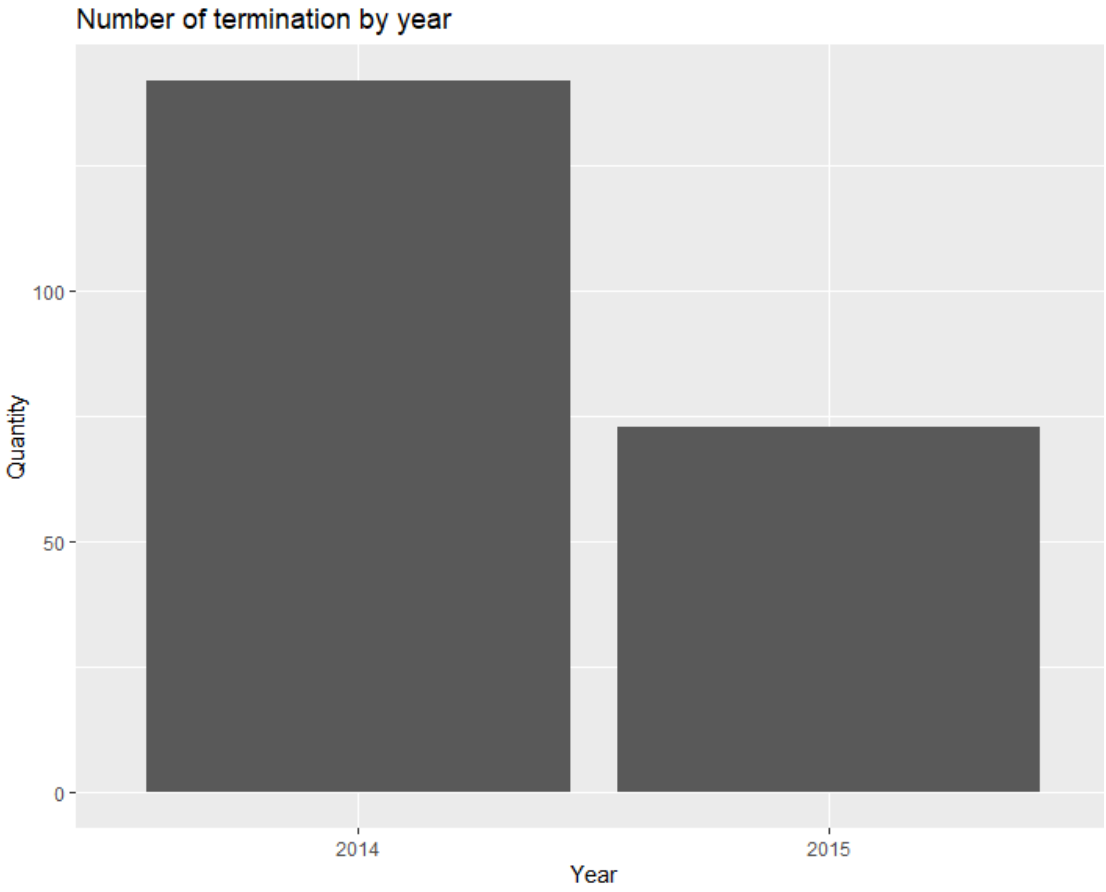


Figure 14: Result of Analysis 2.1

From figure 14, there's only 2 years with resignations, which are 2014 and 2015, with 2014 having more than 100 whereas year 2015 have only near 75.

3.2.2 Analysis 2: The number of employees that were laid off based on age

```
#Analysis 2.2: The number of employees that were laid off based on age
ageArranged = data.frame(table(terminationList$Age))
View(ageArranged)
ggplot(ageArranged, aes(y=Freq, x=Var1)) + geom_bar(stat = "identity", color = "black") +
  labs(title = "Number of employees being laid off based on age", x = "Age", y = "Quantity") +
  geom_text(aes(label=Freq), vjust=1.6, color="white", size=3.5) +
  theme_minimal()
```

Figure 15: Source Code of Analysis 2.2

In analysis 2.2, I'm looking for the number of laid off employees according to age. To do so, I made a table using age and converted it into a data frame before I used ggplot to make it a barplot to view it and obtain insights from it.

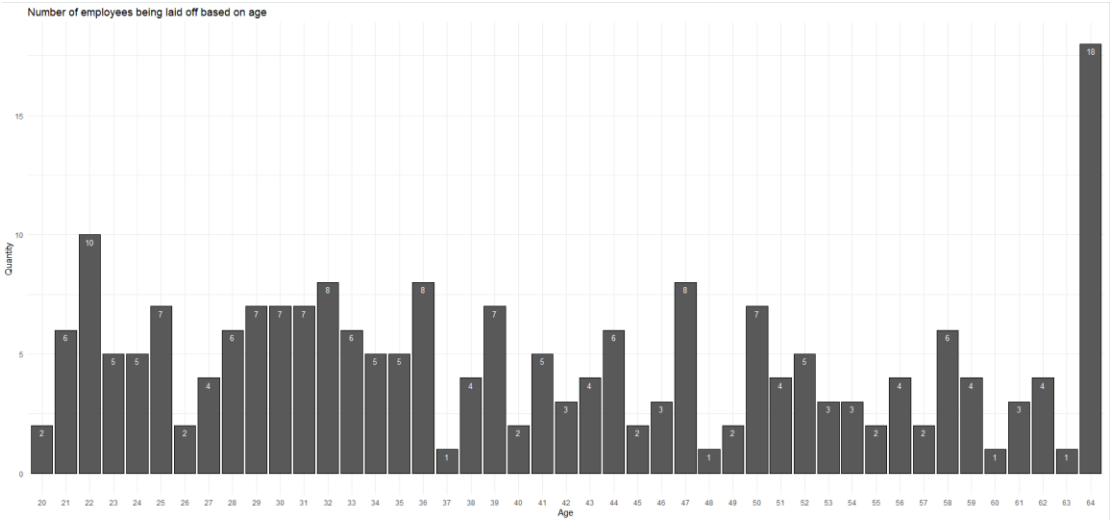


Figure 16: Result of Analysis 2.2

According to figure 16, the majority of laid of employees are clustered in the left side of the barplot, which is within the age frame of 20 to 36 and also with age 64 having the biggest number of employees resigned. This makes sense as age 64 is around or is the retirement age. With the analysis from question 1, I have an guess that the majority of those that are laid of in the age frame of 20 to 36 are mainly

working on lower skilled jobs, which will be investigated in analysis 2 of this question.

3.2.3 Analysis 3: The number of employees that were laid off based on job title; store & unit (according to job title, according to store & unit)

```
#Analysis 2.3: The number of employees that were laid off based on job title ; store & unit (according to job title, according to store & unit)
sample_frac(terminationList, 1) %>% select(`ID`, `Store.Name`, `Business.Unit`)
jobArranged = data.frame(table(terminationList$Job.Title))
jobArrangedOrdered = data.frame(jobArranged[order(-jobArranged$Freq), ])
View(jobArrangedOrdered)
storeArranged = data.frame(table(terminationList$Store.Name))
view(storeArranged)
unitArranged = data.frame(table(terminationList$Business.Unit))
View(unitArranged)
```

Figure 17: Source Code of Analysis 2.3

By tabulating the list of terminated employees by their job title, store and unit I am able to find out on the number of laid off employees for each job and store separately.

jobArrangedOrdered | Visualization2.R | Filter

	Var1	Freq
3	Cashier	67
5	Dairy Person	46
6	Meat Cutter	27
11	Shelf Stocker	18
1	Baker	17
9	Produce Clerk	16
12	Store Manager	6
2	Bakery Manager	4
7	Meats Manager	4
10	Produce Manager	4
4	Customer Service Manager	3
8	Processed Foods Manager	3

	Var1	Freq
1	4	1
2	7	4
3	9	14
4	10	2
5	11	39
6	13	30
7	14	21
8	20	31
9	23	6
10	24	6
11	27	17
12	34	5
13	37	9
14	39	24
15	45	6

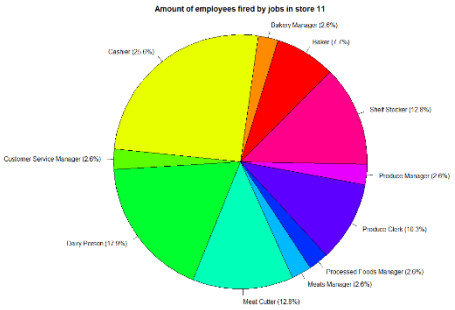
	Var1	Freq
1	STORES	215

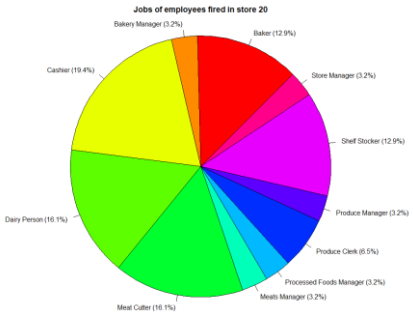
Figure 18: Result of Analysis 2.3

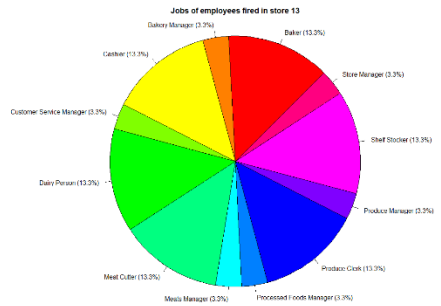
From the results in figure 18, we can tell that it's the lower skilled jobs that have the highest chance of being laid off, with cashier, dairy person, meat cutter, shelf stocker having 67, 46, 27, and 18 employees being laid off out of 215 employees that were laid off. As for stores, we have store 11, 20, 13, 39, 14, 27 and 9 having the biggest number of laid off employees. In order to figure out the correlation between

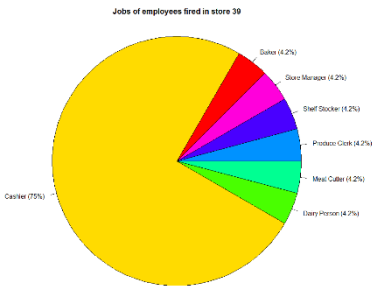
the jobs and the stores, the following code in figure was ran for the stores mentioned to visualize the number of employees being laid off for each jobs in the designated stores.

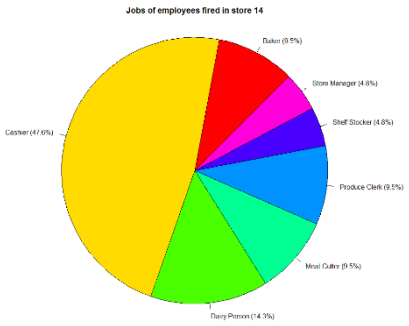
```
store11 = sample_frac(terminationList, 1) %>% subset(`Store.Name` == 11) %>% select(`ID`, `Job.Title`, `Store.Name`) %>% arrange(`Job.Title`)
store11Frame = as.data.frame(table(store11$Job.Title))
store11Frame$percent = round(100*store11Frame$Freq/sum(store11Frame$Freq), digits = 1)
store11Frame$label = paste(store11Frame$Var1," (", store11Frame$percent,"%)", sep = "")
pie(store11Frame$Freq, labels = store11Frame$label, main = "Jobs of employees fired in store 11", radius = 1,init.angle = 45, col = rainbow(nrow(store11Frame)))
```

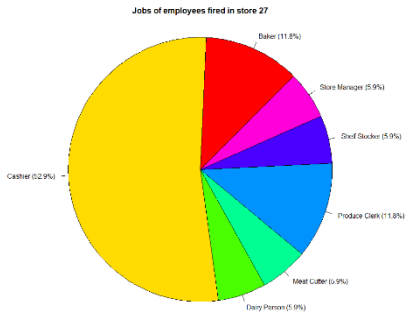












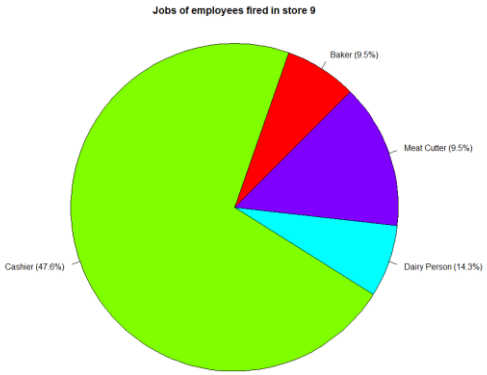


Figure 19: Source Code & Result of Extra Analysis on 2.3

According to the pie charts visualized in figure, it can be proven that these stores that have the bigger number of employees being laid off are employees with lower skilled jobs (cashiers, meat cutters, dairy persons, shelf stockers), with store 39, 14, 27 and 9 being more obvious in this situation with near to 50% or over 50% of laid off employees being cashiers.

To wrap things up for question 2, the information discovered are only years 2014 and 2015 have employees being cut off, in which majority of them are from the younger age group, which is 20 to 30 or at retirement age. 158 out of 215 (73%) laid off employees were working on lower skilled jobs. The stores that have the biggest number of laid off employees are store 11, 20, 13, 39, 14, 27 and 9 with each store having near 50% or more than 50% of laid off employees were working on lower skilled jobs.

3.3 Question 3: Does the organization have enough human resources to fill in the vacant spots from termination?

3.3.1 Analysis 1: The number of new employees and terminated employees per year

```
#Analysis 3.1: The number of new employees and terminated employees per year
fullTerminationList = uniqueList %>% filter(uniqueList$`Reason.of.Termination` == "Layoff"|uniqueList$`Reason.of.Termination` == "Resignaton" |uniqueList$`Reason.of.Termination` == "Retirement") %>%
  select(`ID`, `Reason.of.Termination`, `Store.Name`,`Year.of.Status`) %>% arrange(`Year.of.Status`)
view(fullTerminationList)
terminationsByYear = table(fullTerminationList$Year.of.Status)  # line chart later

fullNewEmployeeList = uniqueList %>% filter(uniqueList$Length.of.Service.Year. == 0) %>% select(`ID`, `Year.of.Status`, `Job.Title`, `Store.Name`) %>% arrange(`Year.of.Status`)
newEmployeesByYear = table(fullNewEmployeeList$Year.of.Status)

plot(terminationsByYear,type = "o", col = "red", xlab = "Year", ylab = "Amount",
     main = "Number of terminated employees every year")

plot(newEmployeesByYear,type = "o", col = "red", xlab = "Year", ylab = "Amount",
     main = "Number of new employees every year")
```

Figure 20: Source Code of Analysis 3.1

In analysis 3.1, I curate the full list of terminated employees and list of all new employees by piping the data from uniqueList and filter out the unwanted data with the certain condition(s), from there I select the columns I needed and arrange them by year for a better view. Lastly, I tabulated the filtered data by using table() function and made a line plot for each dataset to compare between the two.

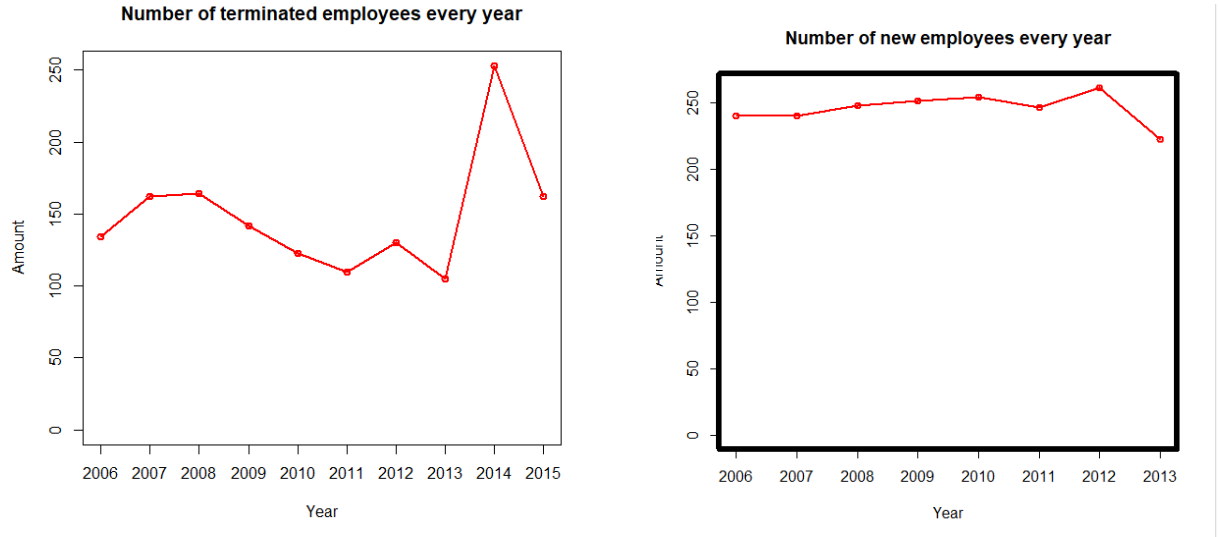


Figure 21: Results of Analysis 3.1

By comparing the 2 line charts in figure 21, we can see that the number of terminated employees was able to be covered by newly employed employees throughout the years of 2006 to 2013. However, in year 2014 and 2015. There's 0 new employees, only a huge spike in terminated employees.

3.3.2 Analysis 2: The number of working employees for each year

```
#Analysis 3.2: The number of working employees for each year
workingEmployeeList = uniqueList %>% filter(uniqueList$Reason.of.Termination == "Not Applicable") %>% select(`ID`, `Year.of.Status`, `Store.Name`,`Job.Title`) %>% arrange(`Year.of.Status`)
View(data.frame(workingEmployeeList))
workingEmployeeListByYear = table(workingEmployeeList$Year.of.Status)
View(workingEmployeeListByYear)
plot(fullEmployeeListByYear,type = "o", col = "red", xlab = "Year", ylab = "Amount",
     main = "Number of total working employees every year")
```

Figure 22: Source Code of Analysis 3.2

Due to the analysis from question 3.1, I was wondering whether the number of working employees for each year can further prove and back it up. Hence, I filtered out unwanted data and only left myself with working employees of each year, tabulate that data and came up with a line plot.

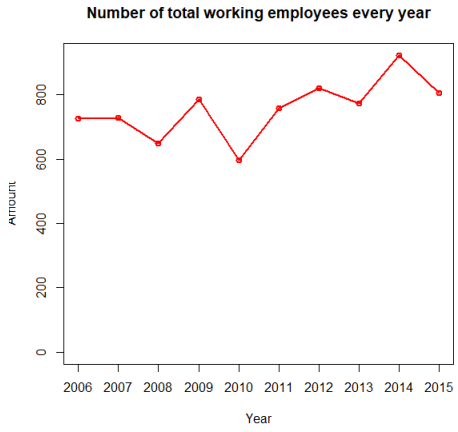


Figure 22: Result of Analysis 3.2

From figure 22, it has proven that analysis 3.2 is able to back analysis 3.1. Throughout the years where newly employed employees were consistent, the number of working employees per year was in the consolidation area, around 600 to 800 employees per year. Until the early 2010s, in around year 2012 ~ 2014, where there's a slight uptrend to break out of the consolidation area which then lead to a sharp fall off of employees per year from year 2014 to 2015.

3.3.3 Analysis 3: The number of new employees and terminated employees for the stores that terminated the most amount of people

```
#Store 11
store11Termination = as.data.frame(fullTerminationList[fullTerminationList$Store.Name == 11, ])
store11TerminationPerYear = as.data.frame(table(store11Termination$Year.of.Status))
View(store11TerminationPerYear)

store11New = as.data.frame(fullNewEmployeeList[fullNewEmployeeList$Store.Name == 11, ])
store11NewPerYear = as.data.frame(table(store11New$Year.of.Status))
View(store11NewPerYear)

store11Final = full_join(store11TerminationPerYear, store11NewPerYear, by = "Var1", all.x = TRUE)
store11Final[is.na(store11Final)] = 0
names(store11Final)[2:3] = c("Terminated", "New")
View(store11Final)

cols <- c('red','blue');
ylim <- c(0,max(store11Final[c('Terminated','New')])*1.8);
par(lwd=6);
store11Bar = barplot(t(store11Final[c('Terminated','New')]),main = "New and Terminated Employees in Store 11 by year", beside=T,ylim=ylim,border=cols,col='white',
  names.arg=store11Final$`Var1`,xlab='Year',ylab='Amount',legend = TRUE,args.legend=list(text.col=cols,col=cols,border=cols,bty='n'));
box();
```

Figure 23: Source Code of Analysis 3.3

On analysis 3.3, I was trying to find the answer for my question for the stores that terminated the greatest number of employees for each year. To do so, I first filtered out the dataset using store name and tabulate them using year of status. The mentioned step was conducted to create 2 tables, a table for terminations every year, another table for new employees every year. However, there might be years where there's no resignations or employments, hence making the merged data leading to potential error during visualization of the data. This was solved by setting all.x = TRUE where there will be N/A for every vacant spot, to ensure that the merged tables have the same number of rows, from there, the ones with N/A are being converted to 0, and the comparison bar plot is visualized. The mentioned steps are conducted on store 11, 20, 13, 39, 14, 27 and 9.

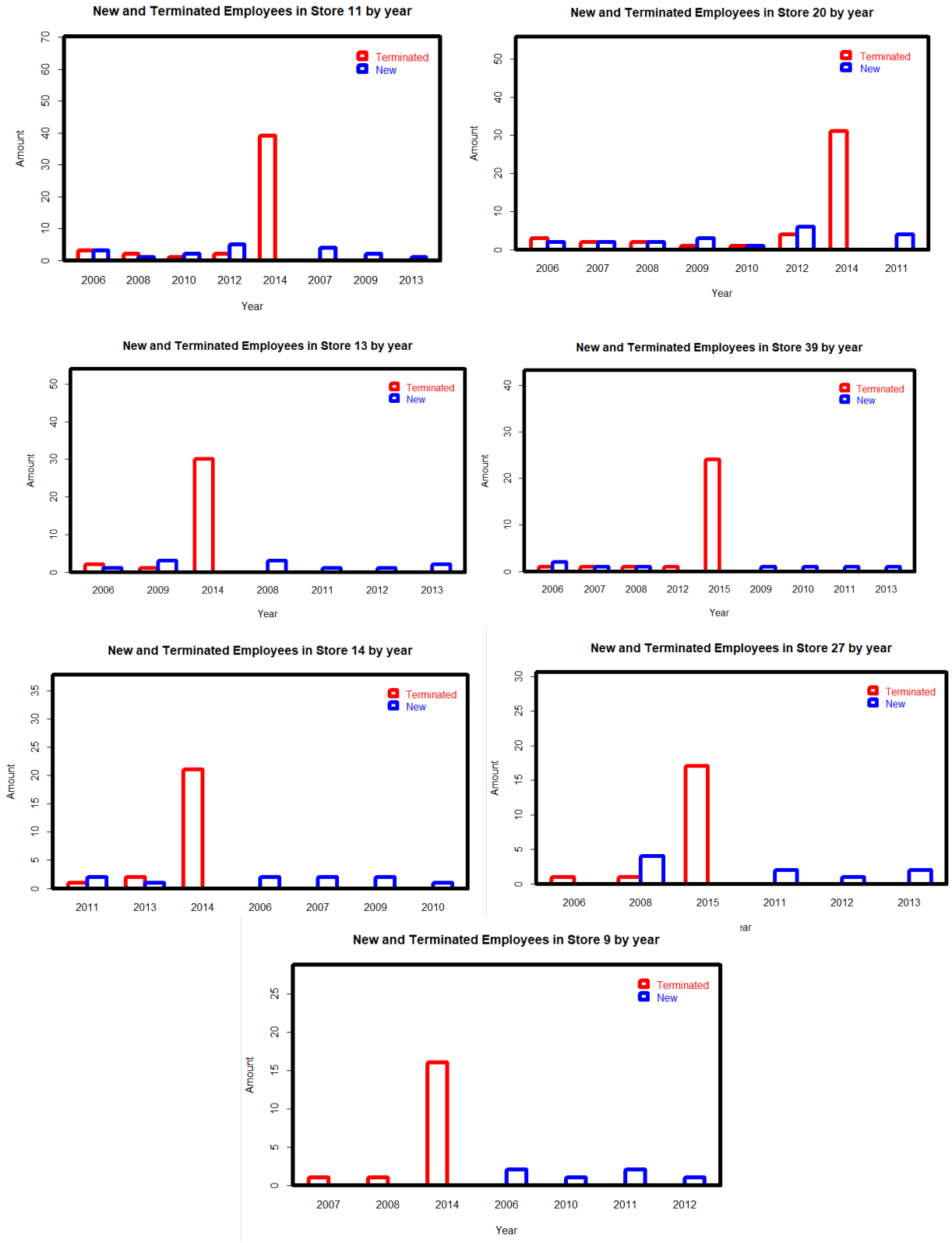


Figure 24: Results of Analysis 3.3

By viewing the comparing bar plots in figure 24, these stores have a similar pattern, where the comparison between new employees and terminated employees is almost non-existent until year 2014 and 2015 where it's only terminations. Throughout the several analyses from question 3 and analysis 2.1, it has shown that the company had to cut down the employees during 2014 and 2015.

Prior of proceeding to question 4, the insights gained from the analyses in question 3 are this company was able to cover the cut-off employees throughout years 2006 to 2013 with their newly employed employees, hence the steady increase of working employees for the year within the region shown in the chart. But, the exact opposite case happened in years 2014 and 2015, where they hired none and cut off a lot, which lead to a steep reduction in terms of number of working employees on the upcoming year. This same case applies for the stores that terminated the bigger amount of employees where only in 2014 and 2015, they couldn't cover the terminated employees with fresh employed employees.

3.4 Question 4: Is there any store being affected by the terminations? If yes, how's the condition?

3.4.1 Analysis 1: How many stores are active?

```
#Analysis 4.1: How many stores are active?
employeeAttrition %>%
  filter(employeeAttrition$Employee.Status == "ACTIVE") %>%
  ggplot(aes(x = as.factor(Store.Name), y = as.factor(Year.of.Status))) +
  geom_count(aes(color = ..n..))+
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  labs(titles = "No. of active employees per store in 2006 ~ 2015", x = "Store Name", y = "Year")
```

Figure 25: Source Code of Analysis 4.1

Due to the findings of the previous questions, I want to dig further into the effects of the terminations. Hence, I filtered out the employees that're terminated and from there I used the ..x.. syntax in geom_count function to view points by their occurrence.

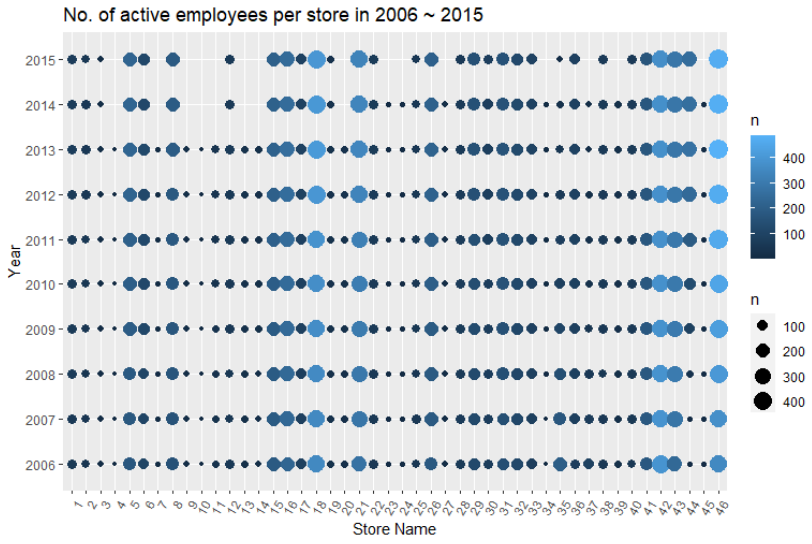


Figure 26: Results of Analysis 4.1

The result of this visualization is shown in figure, where the stores that were closed, which means not active anymore all happened in years 2014 and 2015. As for the rest of the stores, they aren't affected that much as the size of the dots for each store throughout the years don't show that much of a difference.

3.4.2 Analysis 2: How many stores are facing potential employee shortage in year 2014 and 2015?

```
#Analysis 4.2: How many stores are facing potential employee shortage in year 2014 and 2015?

activeEmployees = filter(employeeAttrition, employeeAttrition$Employee.Status == "ACTIVE")

for(x in 1:46){
  store = filter(activeEmployees, activeEmployees$Store.Name == x)
  sumActiveEmployees = 0
  for(y in 2006:2015){
    storeYear = filter(store, store$Year.of.Status == y)
    sumActiveEmployees = sumActiveEmployees + nrow(storeYear)
  }
  storeAverage = sumActiveEmployees / 10
  round(storeAverage, 2)
  for(z in 2014:2015){
    storeYear = filter(store, store$Year.of.Status == z)
    if(sum(storeYear$Employee.Status == 'ACTIVE') < storeAverage){
      difference = storeAverage - sum(storeYear$Employee.Status == 'ACTIVE')
      print(paste("Store ", x, " is facing employee shortage on year ", z, " with ", sum(storeYear$Employee.Status == 'ACTIVE'), " amount of employees, that's ", difference, " lesser than the average"  ))

    }
  }
}
```

Figure 27: Source Code of Analysis 4.2

For this analysis, a standard of stores being considered as having this concern is required, and I define a year to be considered so when it's below the average employees per year for that store. I looped through the years using for loop and calculated the total then the average employees per year for each store. To add up to that, whenever a store has a particular year that's below average employee per year of that store, the result will be shown.

```
"Store  4  is facing employee shortage on year  2014  with  0  amount of employees, that's  0.8  lesser than the average"
"Store  4  is facing employee shortage on year  2015  with  0  amount of employees, that's  0.8  lesser than the average"
"Store  7  is facing employee shortage on year  2014  with  0  amount of employees, that's  3.7  lesser than the average"
"Store  7  is facing employee shortage on year  2015  with  0  amount of employees, that's  3.7  lesser than the average"
"Store  9  is facing employee shortage on year  2014  with  0  amount of employees, that's  11.1  lesser than the average"
"Store  9  is facing employee shortage on year  2015  with  0  amount of employees, that's  11.1  lesser than the average"
"Store  10  is facing employee shortage on year  2014  with  0  amount of employees, that's  1.6  lesser than the average"
"Store  10  is facing employee shortage on year  2015  with  0  amount of employees, that's  1.6  lesser than the average"
"Store  11  is facing employee shortage on year  2014  with  0  amount of employees, that's  27.5  lesser than the average"
"Store  11  is facing employee shortage on year  2015  with  0  amount of employees, that's  27.5  lesser than the average"
"Store  13  is facing employee shortage on year  2014  with  0  amount of employees, that's  20.3  lesser than the average"
"Store  13  is facing employee shortage on year  2015  with  0  amount of employees, that's  20.3  lesser than the average"
"Store  14  is facing employee shortage on year  2014  with  0  amount of employees, that's  15.8  lesser than the average"
"Store  14  is facing employee shortage on year  2015  with  0  amount of employees, that's  15.8  lesser than the average"
"Store  20  is facing employee shortage on year  2014  with  0  amount of employees, that's  21  lesser than the average"
"Store  20  is facing employee shortage on year  2015  with  0  amount of employees, that's  21  lesser than the average"
"Store  23  is facing employee shortage on year  2015  with  0  amount of employees, that's  5.8  lesser than the average"
"Store  24  is facing employee shortage on year  2015  with  0  amount of employees, that's  4.8  lesser than the average"
"Store  25  is facing employee shortage on year  2014  with  52  amount of employees, that's  1  lesser than the average"
"Store  25  is facing employee shortage on year  2015  with  52  amount of employees, that's  1  lesser than the average"
"Store  27  is facing employee shortage on year  2015  with  0  amount of employees, that's  11.7  lesser than the average"
"Store  28  is facing employee shortage on year  2014  with  65  amount of employees, that's  2.2  lesser than the average"
"Store  28  is facing employee shortage on year  2015  with  63  amount of employees, that's  4.2  lesser than the average"
"Store  34  is facing employee shortage on year  2015  with  0  amount of employees, that's  3.2  lesser than the average"
"Store  35  is facing employee shortage on year  2014  with  26  amount of employees, that's  67.6  lesser than the average"
"Store  35  is facing employee shortage on year  2015  with  14  amount of employees, that's  79.6  lesser than the average"
"Store  36  is facing employee shortage on year  2015  with  87  amount of employees, that's  0.599999999999994  lesser than the average"
"Store  37  is facing employee shortage on year  2014  with  12  amount of employees, that's  24.1  lesser than the average"
"Store  37  is facing employee shortage on year  2015  with  0  amount of employees, that's  36.1  lesser than the average"
"Store  39  is facing employee shortage on year  2015  with  0  amount of employees, that's  20.3  lesser than the average"
"Store  41  is facing employee shortage on year  2014  with  174  amount of employees, that's  1.59999999999999  lesser than the average"
"Store  41  is facing employee shortage on year  2015  with  169  amount of employees, that's  6.59999999999999  lesser than the average"
"Store  42  is facing employee shortage on year  2014  with  376  amount of employees, that's  4.5  lesser than the average"
"Store  42  is facing employee shortage on year  2015  with  370  amount of employees, that's  10.5  lesser than the average"
"Store  43  is facing employee shortage on year  2014  with  287  amount of employees, that's  0.199999999999989  lesser than the average"
"Store  43  is facing employee shortage on year  2015  with  287  amount of employees, that's  0.199999999999989  lesser than the average"
"Store  45  is facing employee shortage on year  2015  with  0  amount of employees, that's  5.4  lesser than the average"
```

Figure 28: Result of Analysis 4.2

From the results above, it is shown that majority of the stores that have this concern or are already affected by this concern (0 employees, store closed) came from year 2014 and 2015. Aside from the stores that're closed, stores 25, 28, 35, 36, 41, 42 and 43 have potential employee shortage issues.

The information obtained from analysis 4.1 and analysis 4.2 are not bright signs for this company as having stores to be closed signifies the reduce of revenue earned.

3.5 Question 5: Are any jobs affected by the terminations?

3.5.1 Analysis 1: How many jobs still consist of employees?

```
#Analysis 5.1: How many jobs still consist of employees?
employeeAttrition %>%
  filter(employeeAttrition$Employee.Status == "ACTIVE") %>%
  ggplot(aes(x = Job.Title, y = as.factor(Year.of.Status))) +
  geom_count(aes(color = ..n..))+
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  labs(titles = "No. of active employees per job title in 2006 ~ 2015", x = "Job Title", y = "Year")
```

Figure 29: Source Code of Analysis 5.1

In order to answer this question, I first filtered out the terminated entries, used the ..n.. syntax in aes() command of ggplot() function in order to know the frequency for each job on each year.

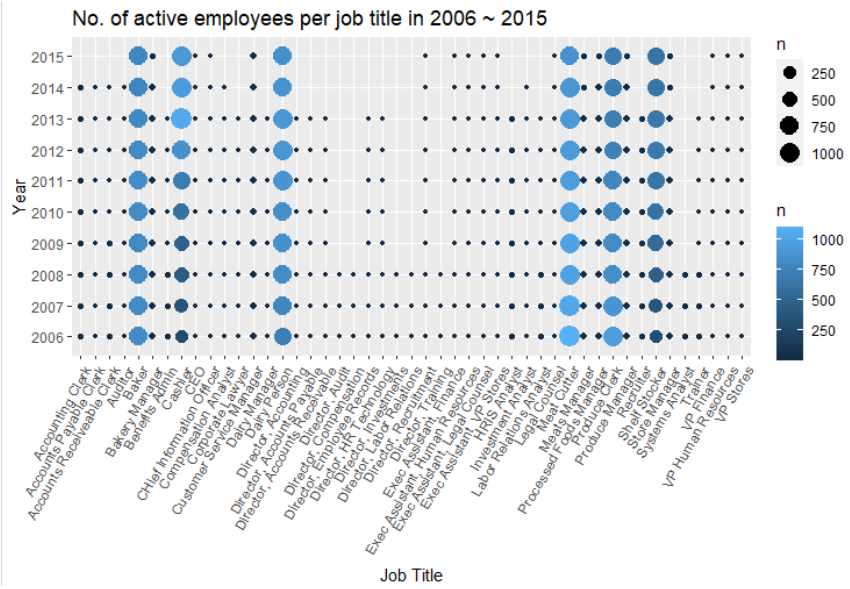


Figure 30: Result of Analysis 5.1

As shown in figure 30, certain jobs went missing, with majority of them being around years 2014 and 2015, which are the clerks, admins, trainers, recruiters, and some of the analysts. However, some of the directors are not being replaced since earlier years. This means that the reason of massive termination is because there were a number of vacant job applications, making the organizing of the employees bad.

3.5.2 Analysis 2: Why are certain director jobs currently vacant?

```
#Analysis 5.2:  Why are certain director jobs currently vacant?

directors = employeeAttrition %>%
  filter(str_detect(Job.Title, "Director"))
directors = as.data.frame(directors)
View(directors)

directors %>%
  filter(Type.of.Termination == "Voluntary") %>%
  ggplot(aes(x = Job.Title, y = Reason.of.Termination)) +
  geom_tile(size = 1, fill = "plum") +
  coord_flip() +
  facet_wrap(~Year.of.Status) +
  labs(x = "Job Title", y = "Termination reason", title = "Director Status", fill = "Type of Termination")
```

Figure 31: Source Code of Analysis 5.2

In this analysis, I wanted to know the reason behind the complete vacancy of these certain director jobs. First up, I filtered the data by using str_detect. Moving on, I used ggplot to visualize it in a tile plot.

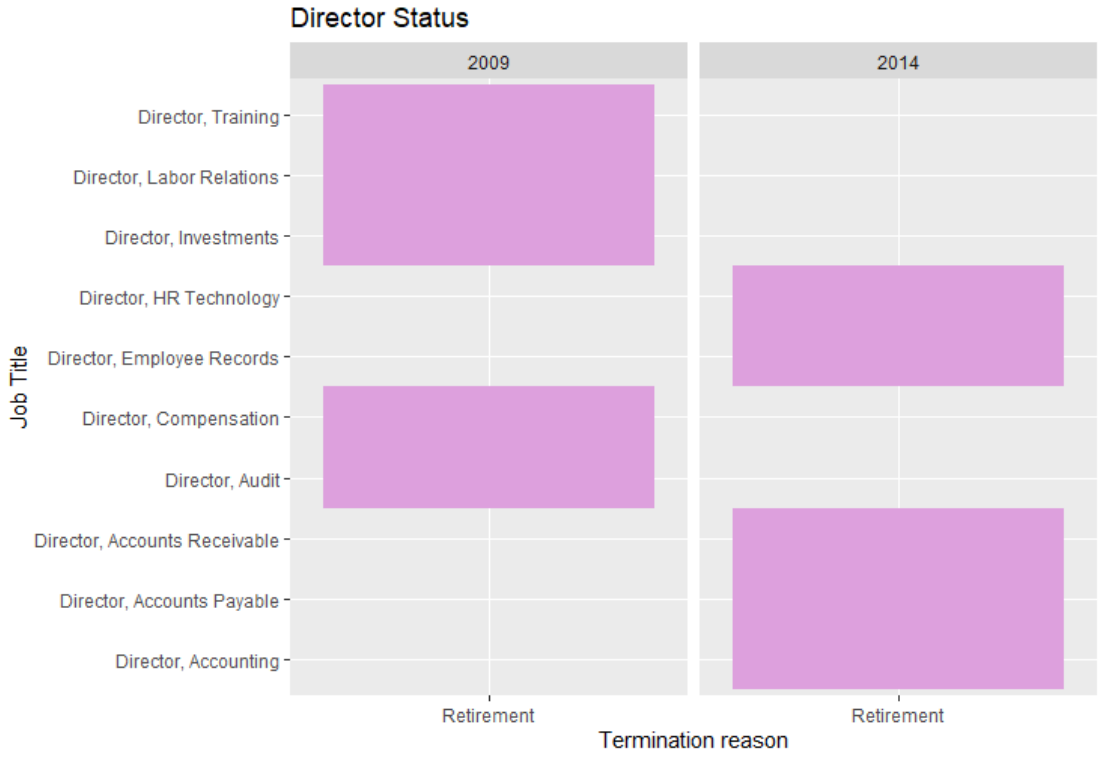


Figure 32: Result of Analysis 5.2

According to the visualization in figure 32, it's shown that all complete vacancies of directors are due to retirement. However, the company didn't recruit any replacement for these directors, which is a warning sign for the company as according

to Poole (2017), directors are responsible for promoting the success of the company, for avoiding conflicts of interest, for having proper independent judgement and exercise them. Without directors, a company's department can fall apart with ease.

3.5.3 Analysis 3: Why are certain analyst jobs currently vacant?

```
#Analysis 5.3: Why are certain analyst jobs currently vacant?

analysts = employeeAttrition %>%
  filter(str_detect(Job.Title, "Analyst")) %>%
  filter(Employee.Status == "TERMINATED")
analysts = as.data.frame(analysts)
View(analysts)

analysts %>%
  ggplot(aes(x = Job.Title, y = Reason.of.Termination)) +
  geom_tile(size = 1, fill = "plum") +
  coord_flip() +
  facet_wrap(~Year.of.Status) +
  labs(x = "Job Title", y = "Termination reason", title = "Analyst Status", fill = "Type of Termination")

employeeAttrition %>%
  filter(str_detect(Job.Title, "Analyst")) %>%
  ggplot(aes(x = Job.Title, y = Employee.Status)) +
  geom_tile(size = 1, fill = "plum") +
  coord_flip() +
  facet_wrap(~Year.of.Status) +
  labs(x = "Job Title", y = "Employee Status", title = "Analyst Status")
```

Figure 33: Source Code of Analysis 5.3

In this analysis, I filtered out rows of data that don't consist of "TERMINATED" and "Analyst". Afterwards, I came u with 2 tile plots, 1 to view the reason of termination, and the other one to view employee status, each are segregated according to year.

Figure 34: Results of Analysis 5.3

According to the results in figure 34, the majority of analysts are terminated due to retirement and the number of analysts has been reducing since year 2010 with systems analyst being vacant, where the other four analysts followed this route in years 2014 and 2015. The unavailability to employ new data analysts makes the company not available to comprehend the demographic, not having access to evidence to make decisions, unavailable to test and retest and the company will be in a passive stance instead of having a proactive approach towards business growth. (Crook, 2017)

4. Conclusion

In conclusion, the company is facing a big termination session of employees due to the lack of replacements for retired analysts and directors, which are both roles with ability to pivot the company towards success or plummet. And with that being the leading factor, the company couldn’t sustain bigger number of employees and had to initiate the mentioned termination session, which leads to lower skilled people losing their jobs and several shops being closed, which then leads to lesser yearly revenue for the company.

5. References

Banks, C. (2018). Types of Low Skill Jobs. Career Trend. https://careertrend.com/list-7404193-types-low-skill-jobs.html

Crook, C. (2017, April 7). 4 Reasons You Need to Hire a Data Analyst. Cogs Agency. https://cogsagency.com/analytics-and-business-intelligence/4-reasons-need-hire-data-analyst/

Lawton, G. (2022). data preprocessing. TechTarget. https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing#:~:text=Data%20preprocessing%2C%20a%20component%20of,for%20the%20data%20mining%20process.

Poole, G. (2017, October 6). Directors' duties and why they are so important |Stephens Scown. Stephens Scown. https://www.stephens-scown.co.uk/corporate-commercial/directors-duties-important/

What is Data Exploration? (2020). TIBCO Software. https://www.tibco.com/reference-center/what-is-data-exploration#:~:text=Data%20exploration%20is%20the%20first,and%20get%20to%20insights%20faster.