



ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APU2F2109CS(DA)

HAND OUT DATE: 4 OCTOBER 2021

HAND IN DATE: 22 NOVEMBER 2021

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 references (cited using the American Psychological Association (APA) Referencing).**
- 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

1.0 Introduction and Assumption	4
2.0 Pre-Analysis	5
2.1 Data Import	5
2.2 Library	6
2.2.1 dplyr	6
2.2.2 ggplots2	6
2.3 Data Cleaning	7
2.4 Data Pre-Processing	9
2.5 Data Transformation	11
3.0 Questions and Analysis	12
3.1 Question 1 – Why people leaving their jobs?	12
3.1.1 Analysis 1 – Termination Reason and Age	12
3.1.2 Analysis 2 – Termination Reason and Job Title	14
3.1.3 Analysis 3 – Termination Reason and City	15
3.2 Question 2 – How is the Company current condition?	16
3.2.1 Analysis 4 – Comparing the workforce between 2006 and 2015	16
3.2.2 Analysis 5 – Comparing store condition between 2006 and 2015	19
3.2.3 Analysis 6 – Comparing city presence between 2006 and 2015	21
3.2.3 Analysis 7 – Finding Layoff rate per year	24
3.3 Question 3 – Which Store has the highest turnover rate and why?	26
3.3.1 Analysis 8 – Finding the store with the highest turnover rate	26
3.3.2 Analysis 9 – Analysing store with most layoff's job distribution	27
3.3.3 Analysis 10 – Analysing store with most resignation's job distribution	29
3.4 Question 4 – When is the best time for hiring spree?	31
3.4.1 Analysis 11 – Amount of Resignation per Month (Per Year)	31
3.4.2 Analysis 12 – Amount of Resignation per Month (Average)	32

4.0 Extra Feature	33
4.1 Extra Feature 1 - Comparison Bar Chart	33
4.2 Boxplot with Jitter overlay and mean position	35
Conclusion	36
References	37

1.0 Introduction and Assumption

Running a company is never as easy as a home scale operation because due to the amount of people that are needed, causing a lot of complex problems arise, from scheduling conflict, inter-personnel conflict, managerial-employee conflict, and each personal conflict that everyone have. The bigger the scale of operation the company run, these problems will grow bigger in size and scale.

To counteract this problem, most company has a specialized department with the purpose to deal and resolve the issues within the employees, fittingly named Human Resources Department. Even though they rarely go actively running an operation, they exist to make sure all other departments run their objectives well and should any problems arise between departments, or even inter-departments, they should find out the roots and fix it quickly. Human Resources Team also routinely keep check towards all employees, making sure they are in good condition in all aspects, such as mentally, physically, emotionally, and financially to work in their maximum capabilities.

However, no matter how good the Human Resources inside a company, there will always be some missed evaluation and judgements that may cost a company its manpower. In this case, a company Human Resources department's dataset about their employee attritions record between 2006-2015. The assumption is over the span of 10 years, there can be some missed evaluations that can be discovered to help the company in the future to prepare beforehand by doing observation towards the dataset. The observations will be represented with R language, with explanation for each analysis will be provided.

2.0 Pre-Analysis

2.1 Data Import

```
main_data<-read.csv(file="C:/Users/User/Documents/APU/Year 2 Semester 3/Programming for Data Analysis/Assignment/employee_attrition.csv",header=TRUE,sep=",")
```

Figure 2.1.1 Snippet Code of Data Import

Before any analysis can be the dataset should be imported inside RStudio, the tools that will be used to generate all observations using R language. To import it, function `read.csv()` is used to import the .csv file into R. After the code has been executed, the data will be stored within `main_data` dataframe.

2.2 Library

2.2.1 dplyr

dplyr library is a grammar of data manipulation library in R. It has many useful features such as `arrange()`, `select()`, `summarise()`, `filter()`, `group_by()`, `mutate()` and many more.

2.2.2 ggplots2

ggplots2 library is used to create data visualizations in R. The library is easy to use as we only need to input the variables we want to use to visualize and choose which graph will be used to represent it. The visualization is also easy to modify. The graph can be a bar chart, histogram, line chart, scatter plot, box plot, and many more.

2.3 Data Cleaning

Before any analysis started, we must do some cleaning towards the data. The process called Data Cleaning means to identify, correct, or removing inaccurate raw data for down streaming purposes (Burns, 2021). This process is necessary so during the analysis process, there is no more problem that must be solved first, that may cause any previous work to be invalid.

```
unique_main_data <- unique(main_data)
```

Figure 2.3.1 Snippet Code to check duplicated data

To check for any duplicate, we use `unique()` function. As the snippet code above indicates, we took the data frame, use `unique()` and move the result to the `unique_main_data` data frame. `Unique()` function produces only one for each unique data, so any duplicate will be removed.



Data		
▶ <code>main_data</code>	49653 obs. of 18 variables	
▶ <code>unique_main_d...</code>	49653 obs. of 18 variables	

Figure 2.3.2 result from `unique()` function

After the `unique()` function has been executed, the resulting data frame and the original data frame has the same amount of observation. This conclude that the dataset given does not have any duplicated data inside it.

```
main_data$recorddate_key <- sub(" 0:00","",main_data$recorddate_key)
```

Figure 2.3.3 Snippet Code to remove timestamp from `recorddate_key`

Data cleaning also means removing any piece of data that can be deemed unnecessary. Inside the `recorddate_key` column, there is a timestamp of when the data is taken. But since all timestamp is set at “0:00”, the data cannot provide any analysis benefit, and can be removed so the column can be formatted.



recorddate_key
12/31/2006
12/31/2007
12/31/2008
12/31/2009
12/31/2010

Figure 2.3.4 Result after removing the timestamp

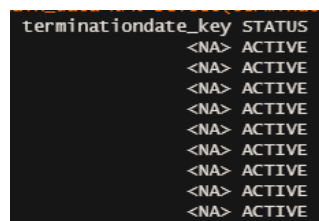
After the changes, the recorddate_key will only contain a date.

The last changes that can be taken to make the analysis process easier is to remove the termination date from the data for anyone who status is still active. From the data provided, if someone is still active after the 2015 record, the termination date will be “01/01/1900”. The changes will delete the termination date for all active personnel for easier understanding.

```
#Data Cleaning: removing terminated date if terminationdate_key == '1900-01-01'  
main_data$terminationdate_key[main_data$terminationdate_key == '1/1/1900'] <- NA
```

Figure 1.3.5 Snippet code to replace all active personnel termination date

Using the code above, we delete all termination date that valued at “1/1/1900” into NA (not available) as the shown result below.



terminationdate_key	STATUS
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE
<NA>	ACTIVE

Figure 2.3.6 Result from replacing all active personnel termination date

For the data with valid termination date, but the status of the personnel is still active, due to the Status update followed the record date, the data will remain.

2.4 Data Pre-Processing

After all cleaning has been done, data pre-processing is next. Data Pre-Processing contain steps to format the data into a more usable format (GeeksforGeeks, 2021). In this dataset, first pre-processing that can be done are all the date's column, can be transformed into R date values. These changes can prove beneficial during analysis when we only need the year or month of the data, which can be easily extracted from the date value if the date has been transformed into the correct data type.

```
#Data Transformation #2: formatting date data into date
main_data$recorddate_key <- as.Date(as.character(main_data$recorddate_key), format = "%m/%d/%Y")
main_data$birthdate_key <- as.Date(as.character(main_data$birthdate_key), format = "%m/%d/%Y")
main_data$orighiredate_key <- as.Date(as.character(main_data$orighiredate_key), format = "%m/%d/%Y")
main_data$terminationdate_key <- as.Date(as.character(main_data$terminationdate_key), format = "%m/%d/%Y")
```

Figure 2.4.1 Snippet Code to change the data type for dates data

To do the changes, as.Date() function is used to transformed the data into dates. as.character() is also used to transform the data into character first. The new format for date will be yyyy-mm-dd from the old mm/dd/yyyy.

recorddate_key	birthdate_key	orighiredate_key	terminationdate_key
2006-01-01	1946-01-01	1997-07-09	2006-01-01
2006-01-01	1941-01-15	1992-07-22	2006-01-15
2006-01-01	1941-01-15	1992-07-22	2006-01-15
2006-01-01	1946-01-16	1997-07-24	2006-01-16
2006-01-01	1946-01-17	1997-07-25	2006-01-17
2006-01-01	1946-01-20	1997-07-28	2006-01-20

Figure 2.4.2 Result from formatting the date

After the changes, the date will be transformed into the chosen format.

The next pre-processing is fixing typos inside termination reason and city name variable. As “Resignation” is misspelled as “Resignaton” inside the termination reason, while City Name “New Westminster” is wrongly typed as “New Westminister”

```
#2. fixing termination reason typo
main_data$termreason_desc[main_data$termreason_desc == 'Resignaton'] <- 'Resignation'

#3. fixing city name typo
main_data$city_name[main_data$city_name == 'New Westminister'] <- 'New Westminster'
```

Figure 2.4.3 Snippet Code to fix the typo

Using the code above, all the typo mentioned before will be corrected

For the final pre-processing that will be applied towards the data, the variable “store_name” will be parsed into character, since the store name is represented by a number, but the number itself did not have any numerical value attached to it.

```
#4. Parsing the store names into character  
main_data$store_name <- as.character(main_data$store_name)
```

Figure 2.4.4 Snippet Code to change the data type for Store Name Variable

After the code has been ran, the “store_name” will be set as character valued number.

2.5 Data Transformation

Final changes that can be done is data transformation. This means altering the structure of the data, to ensure easier data manipulation for the analysis (Stitch, n.d.). This means removing any column that have no benefit being inside the dataset. The only column that fulfils this condition are between gender_short and gender_full. As there is only Male and Female inside the dataset, one of the variables is redundant and can be removed without compromising the data.

```
main_data <- select (main_data, -c(gender_short))
```

Figure 2.5.1 Snippet code to drop the column

Using the code above, the data frame will drop the gender_short column, making the variable count dropped into 17.

3.0 Questions and Analysis

3.1 Question 1 – Why people leaving their jobs?

3.1.1 Analysis 1 – Termination Reason and Age

For the first analysis, we must know what the correlation between people's age with their termination reason to know is age a deciding factor in their decision to quit the job.

```
main_data%>%
  filter(termttype_desc == "voluntary")%>%
  ggplot(aes(y= age, x=termreason_desc))+
  geom_boxplot(alpha = 0)+
  geom_jitter(alpha = 0.5, aes(colour = "orange"))+
  facet_wrap(~gender_full)+
  stat_summary(fun=mean, geom="point", shape=20, size=3, color="blue")+
  labs(title = "Correlation between Termination Reason and Age, separated by Gender",
       x='Termination Reason', y = 'Age')
```

Figure 3.1.1.1 Snippet Code for Analysis 1

The code above is to generate a boxplot with a jitter layer on top to represent the distribution between age and termination reason. A couple of filters has been put such as only the terminated employee data should be considered, also the termination type should be all voluntary, since we want to know why people are leaving their jobs in their own accord, not based on company's policy. The visualization will also be split between gender, to understand the age distribution separately by gender.

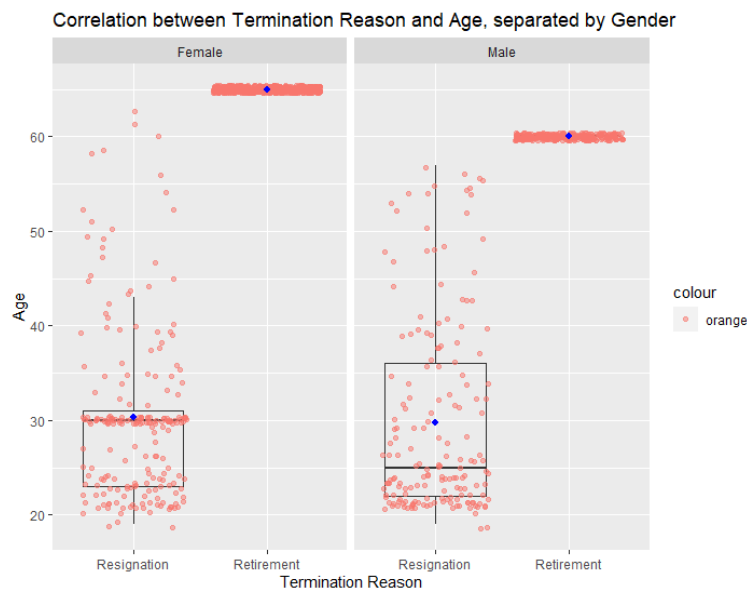


Figure 3.1.1.2 Visualization for Analysis 1

The graph above is the visualization for analysis 1. The graph also shows a red dot that represents the average for each boxplot. From the graph above, we learn that the retirement age for male and female are different, as male go retire at 60 years of age, while female counterpart go retire at age 65. Why women retirement age is later than men is usually because delayed career, usually the result from motherhood. Other reasons are women who works until nearing retirement age are someone who are probably not in a good financial position, so they need as much money as possible before being too old. They also can start their Canada Pension Plan/CPP later, which mean more monthly income after retiring (Government of Canada, n.d.).

For Resignation average, female has the average around 30 years old. This can be caused due to 29.2 years old become the average age of maternity at first birth for Canada citizen in 2015 (Provencher, Milan, Hallman, & D'Aoust, n.d.). This means that around that age, female will start motherhood, and most likely resign to focus on motherhood. On the other hand, male resignation age seems to be more evenly distributed, with a bit of cluster around 20-25 years old, as that the average age for university student (Statistics Canada, 2019), with their resignation can be caused due to they seeing the job as only a short-term position only.

3.1.2 Analysis 2 – Termination Reason and Job Title

For the next analysis, we can analyse the correlation between termination reason with Job Title to understand which job positions has the biggest turnover, with hope that the result can provide some context on the question.

```
#Analysis 2 - Termination Reason - Job Title
main_data%>%
  filter(termination_desc == "Resignation")%>%
  ggplot(aes(x=department_name, fill = job_title))+
  geom_bar()+
  labs(title = "Correlation between Termination Reason and Job Title", x='Termination Reason', y = 'Count', fill = "Job Title")
```

Figure 3.1.2.1 Snippet Code for Analysis 2

The code above will generate a bar chart to represent the Amount of Resignation from a specific job title. Each bar will represent a department inside the company, while the job will be coloured differently based on the title.

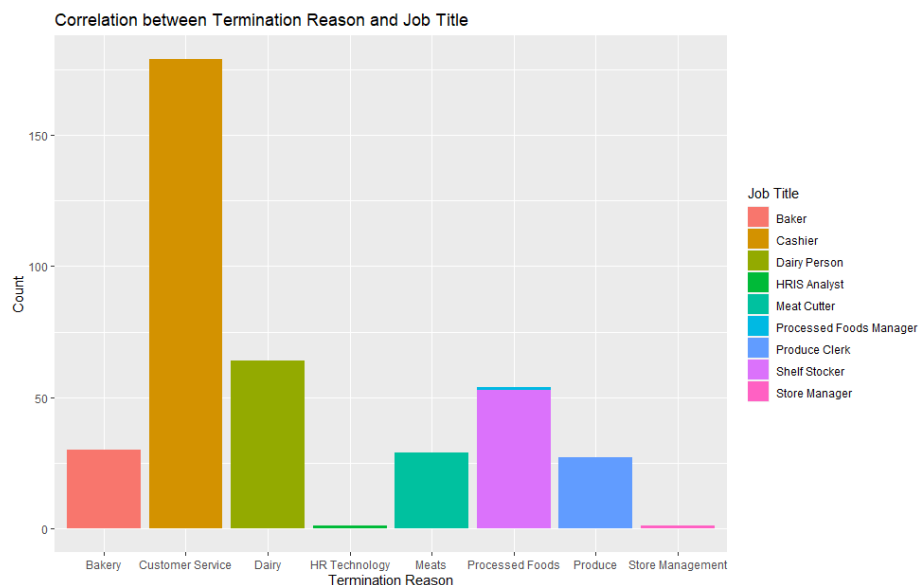


Figure 3.1.2.2 Visualization for Analysis 2

From the result above, we can conclude that between all the jobs, cashier has the highest resignation count from all other jobs. This result can be caused due to being a cashier has a low education criterion, so more people can be qualified as it. Cashier also constantly categorized as one of the lowest paying jobs in Canada with an average of USD \$12/hr rate (Talent, 2021), reinforcing the assumption that most people become a cashier only as a filler and short-term jobs during a gap of unemployment.

3.1.3 Analysis 3 – Termination Reason and City

The last analysis that can be done to help answering this question is the correlation between Termination Reason with City.

```
#Analysis 3 - Termination Reason - City
main_data%>%
  filter(termination_desc == "Resignation")%>%
  ggplot(aes(x= termination_desc, fill = job_title))+
  geom_bar()+
  facet_wrap(~city_name)+
  labs(title = "Correlation between Termination Reason and City", x='Termination Reason', y = 'Count')
```

Figure 3.1.3.1 Snippet Code for Analysis 3

To find the correlation, the code above will be used to create a visualization bar chart, with count of resignation separated per city name.

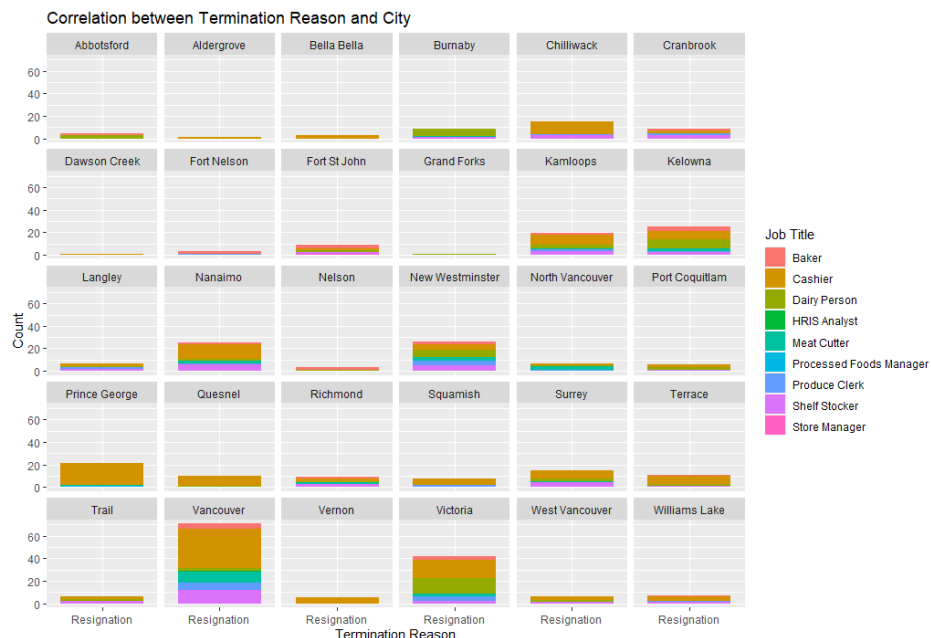


Figure 3.1.3.2 Visualization for Analysis 3

From the visualization above, we can learn that the bigger cities consistently have the highest resignation count from all the data. This is likely caused by higher job opportunities inside a highly packed city. Other reason that may cause this effect is the living cost in big cities are much higher. This means people will jump to another job should the new job's benefits outweigh the old one. On the other hand, smaller cities such as Dawson Creek where the population is much smaller, in turn less job opportunities, cheaper living cost, mean that people will usually stay on their job longer since it's already provided them with a comfortable lifestyle.

3.2 Question 2 – How is the Company current condition?

3.2.1 Analysis 4 – Comparing the workforce between 2006 and 2015

To get an understanding towards the company's workforce current condition, we can learn by comparing between the available job during the start of data recording, which in 2006, and the latest data recording, which is in 2015.

```
#Analysis 4 - Comparing the workforce between 2006 and 2015

data41<- main_data%>%
  group_by(department_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2006")%>%
  count()
data41['year']='2006'
data41

data42<- main_data%>%
  group_by(department_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2015")%>%
  count()
data42['year']='2015'
data42

data43 = rbind(data41,data42)
rm(data41)
rm(data42)

data43%>%
  ggplot(aes(fill=year, y=n, x=department_name))+
  geom_bar(position="dodge", stat="identity")+
  labs(title = "Comparing the workforce between 2006 and 2015",
       x="Department Name", y="Employee Count", fill="Year")
```

Figure 3.2.1.1 Snippet Code for Analysis 4

Using the code above, two sub data frame will be created, with the first one taking the data of the active employees from 2006 based on department name, while the second data frame will take the 2015 counterpart. After that, the 2 data frame will be combined by row to create the completed data frame ready for visualization. After combining, the two data frame will be deleted since it has no more purpose. The combined data frame will be visualized using bar chart, with the data will be compared by year.

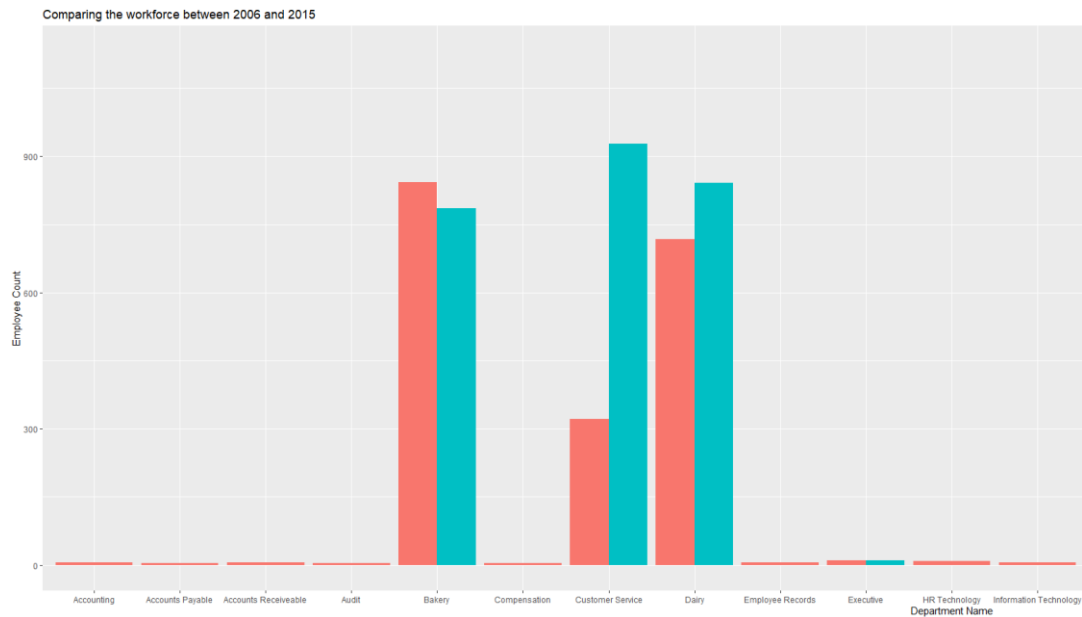


Figure 3.2.1.2 Visualization Part 1 for Analysis 4

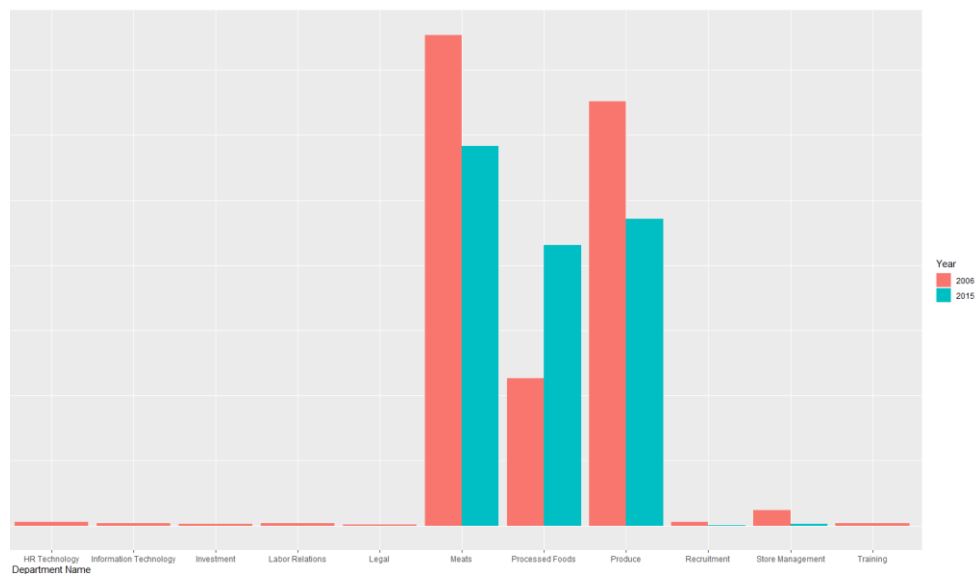


Figure 3.2.1.3 Visualization Part 2 for Analysis 4

From the visualization above, a lot of departments did not survive the 10 years gap between the 2006 and 2015, with a lot of departments having an already very small team in 2006 and disappear completely during the 2015 parts. Between the 10 years gap, there is a significant decrease in the store management, that means the store amount is also decreasing. Even HR department that should be an essential part of every company, got deleted. This can be a sign of company's restructure, as some departments are redundant and can be assimilated, such as Accounts Payable,

Compensation, and Accounts Receivable can be combined with accounting department. Recruitment and Training also can be combined into one department that handles both tasks. Labour Relations and HR Technology can be combined into HR, while HR technology and Information Technology can become one department that only focuses on technology development.

3.2.2 Analysis 5 – Comparing store condition between 2006 and 2015

After workforce condition, the next aspect that we can check are the stores condition between 2006 and 2015 to understand the company branch condition overtime.

```
#Analysis 5 - Comparing store condition between 2006 and 2015

data51<- main_data%>%
  group_by(store_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2006")%>%
  count()
data51['year']='2006'
data51

data52 <- main_data%>%
  group_by(store_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2015")%>%
  count()
data52['year']='2015'
data52

data53 = rbind(data51,data52)
rm(data51)
rm(data52)

data53%>%
  ggplot(aes(fill=year, y=n, x=store_name))+
  geom_bar(position="dodge", stat="identity")+
  labs(title = "Comparing the store condition between 2006 and 2015",
       x="Store Name", y="Employee Count", fill="Year")
```

Figure 3.2.2.1 Snippet Code for Analysis 5

Using the code above, two sub data frame will be created with the first one containing the data of active employee in each store in year 2006, while the second one containing the same data but in year 2015. The two data frame will be combined by row into the final data frame for visualization, with the two original data frames being deleted as it already

serves its purpose. The combined data frame will be visualized using a bar chart, with the data being compared by year.

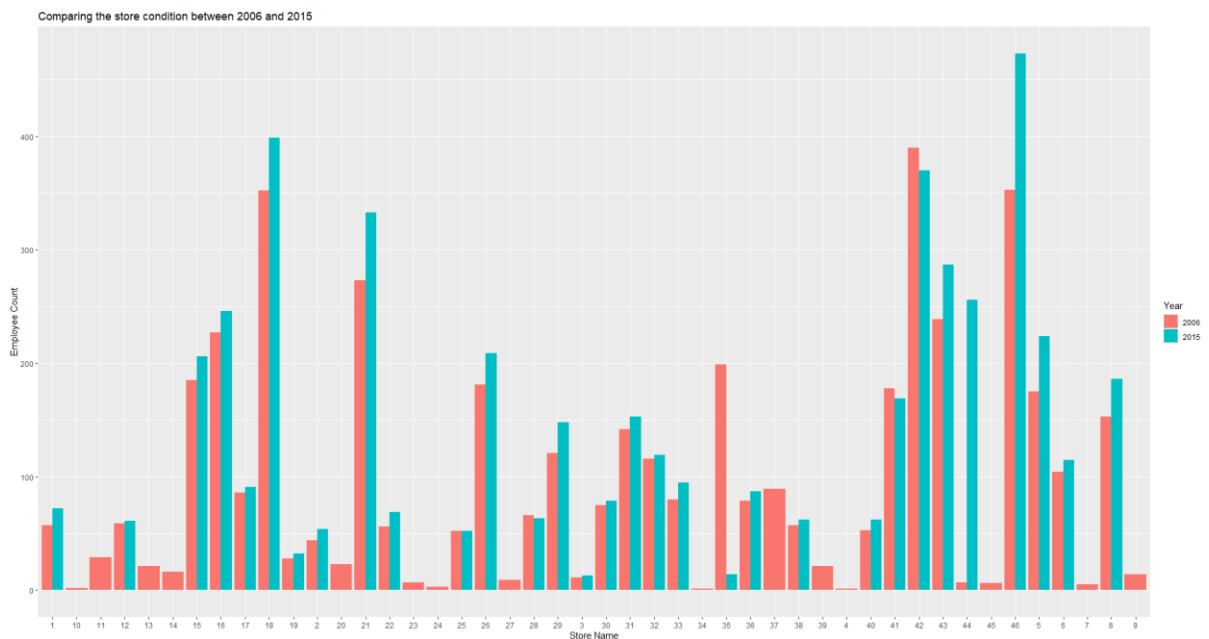


Figure 3.2.2.2 Visualization for Analysis 5

From the visualization above, an assumption can be made that if a store did not have any employee, that means the store is already closed. Based on that assumption, we can see that a lot of stores closed during the 10 years gap between 2006-2015. The reasoning can be due to the company's downsizing, as the trend is similar with Analysis 4, that the company is losing a lot of workforces, in turn also closed a lot of stores too. This move may be a planned economical move by the company.

3.2.3 Analysis 6 – Comparing city presence between 2006 and 2015

To dug deeper into the question, an analysis towards the company's presence in the city that has it stores can be done to get an understanding about the company's presence.

```
#Analysis 6 - Comparing City Presence between 2006 and 2015
data61<- main_data%>%
  group_by(city_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2006")%>%
  count()
data61['year']='2006'
data61

data62 <- main_data%>%
  group_by(city_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2015")%>%
  count()
data62['year']='2015'
data62

data63 = rbind(data61,data62)
rm(data61)
rm(data62)

data63%>%
  ggplot(aes(fill=year, y=n, x=city_name))+
  geom_bar(position="dodge", stat="identity")+
  labs(title = "Comparing the City Presence between 2006 and 2015",
       x="City Name", y="Employee Count", fill="Year")
```

Figure 3.2.3.1 Snippet Code for Analysis 6

To start the analysis, the code above will create two data frames, with each containing the data of active employee in each city, with the difference is the first one will only take the data from 2006, will the second one will take the data from 2015. These two data frames will be combined by row into one data frame. After that, the combined data frame will be visualised using bar chart, with each city presence will be compared by year.

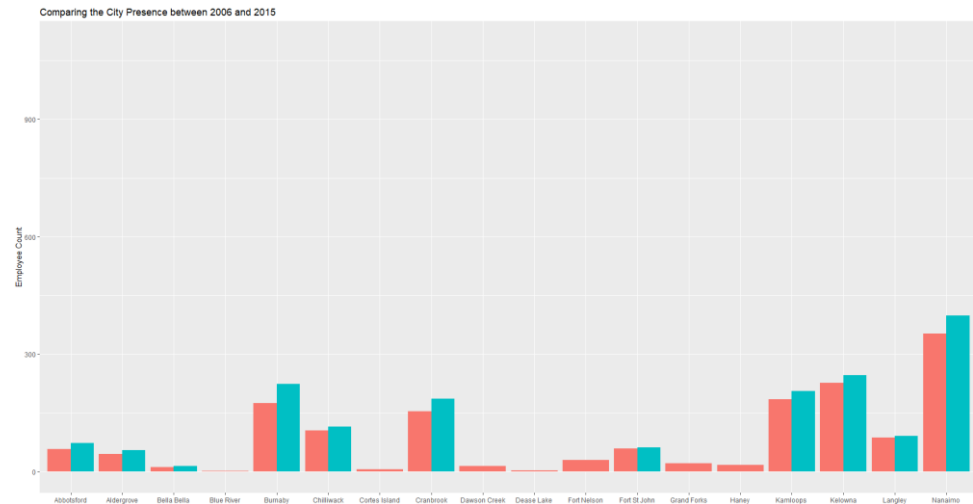


Figure 3.2.3.2 Visualization Part 1 for Analysis 6

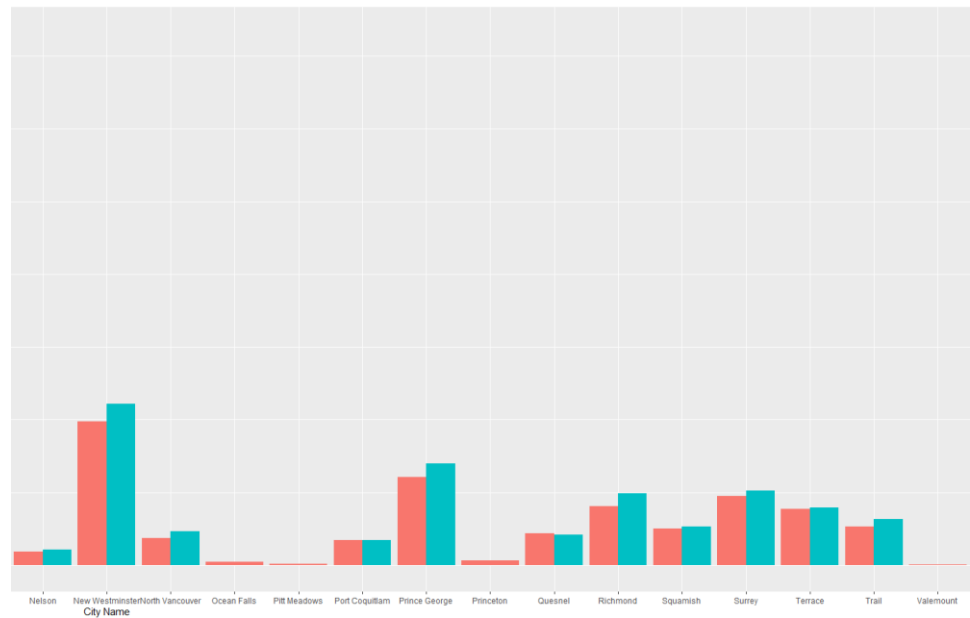


Figure 3.2.3.3 Visualization Part 2 for Analysis 6

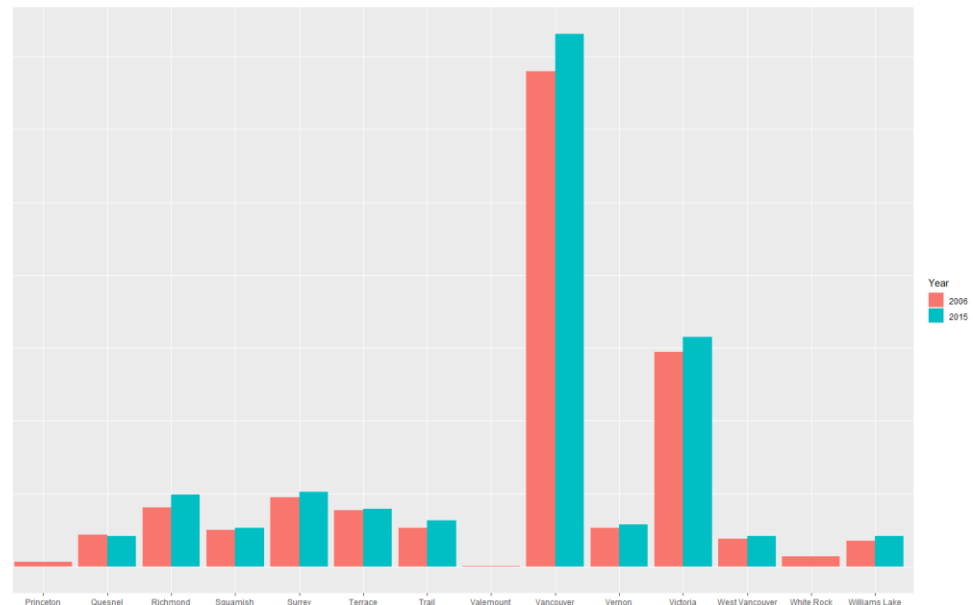


Figure 3.2.3.3 Visualization Part 3 for Analysis 6

From all visualization above, a similar assumption can be made, that if there are no active employees, then all stores in that city has been closed, and the company has no presence in the area. With this assumption, then a conclusion that by 2015, the company has lost all presence in 12 cities. This massive lost can be caused due to poor economic, losing to a competitor, or it is just not economically viable to keep a store open in the area.

3.2.3 Analysis 7 – Finding Layoff rate per year

To fully understand how the condition of the company is, the last analysis that can be done is to analyse the number of layoffs throughout the years, as it can show when the company is starting to decline.

```
#Analysis 7 - Counting Layoff Rate between 2006-2015
main_data%>%
  group_by(STATUS_YEAR)%>%
  filter(STATUS=="TERMINATED")%>%
  ggplot(aes(x=as.character(STATUS_YEAR),fill = termreason_desc))+
  geom_bar(position="dodge")+
  labs(title = "Comparing the Layoff Rate between 2006 and 2015",
       x="Year", y="Employee Count", fill="Termination Reason")
```

Figure 3.2.3.1 Snippet Code for Analysis 7

Using the code above, a bar chart will be generated that compare the termination reason amount of each year from 2006 until 2015. With this code, an understanding of when the layoff trend starts to show can be achieved.

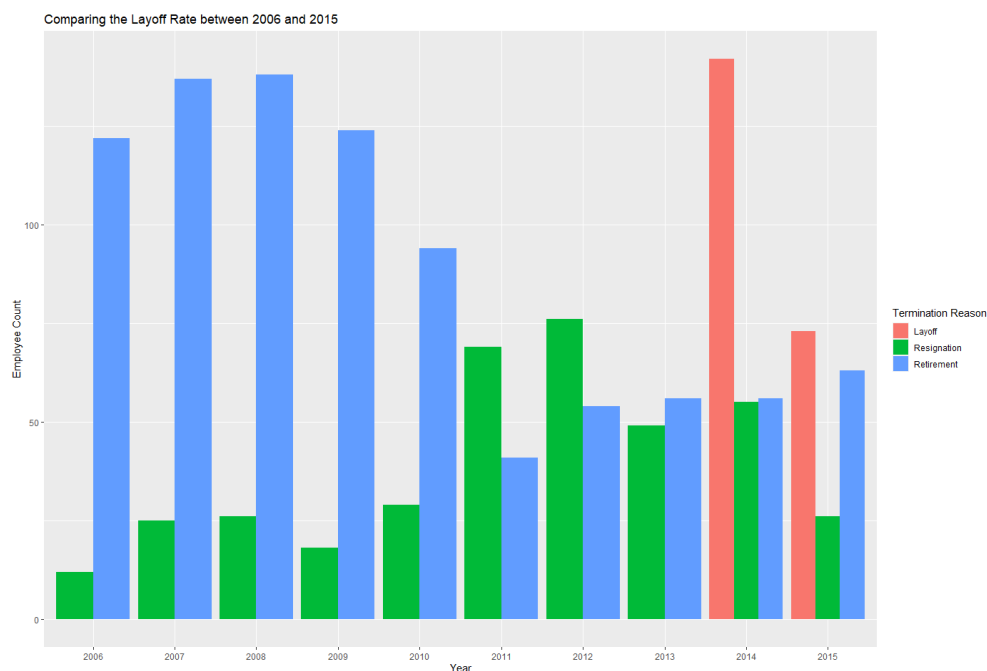


Figure 3.2.3.2 Visualization for Analysis 7

From the visualization above, a conclusion can be reached that during 2014, a massive layoff spike has emerged, with over 125 employees being laid off in a single year, with 2015 continuing the trend although with a decline in numbers to follow. This mean that starting from 2014, the

company start having a difficulty in economics, that may cause a downsizing/restructuring plan to happen. With all analysis taken into consideration, its clear that the company is doing a downsizing/restructuring even with the sacrifice of are presence, to face a difficult economic time, or to adapt to more digital world today.

3.3 Question 3 – Which Store has the highest turnover rate and why?

3.3.1 Analysis 8 – Finding the store with the highest turnover rate

For question 3, we must know which store has the highest turnover rate first, either from resignation also layoff to understand what the problem from each store is, such as managerial problem from the store, general economical problem from the area, etc.

```
#Analysis 4 - Finding the store with the highest turnover rate
main_data%>%
  filter(STATUS == "TERMINATED")%>%
  filter(termreason_desc != "Retirement")%>%
  arrange(store_name)%>%
  ggplot(aes(x=store_name, fill = termreason_desc))+
  geom_bar()+
  labs(title = "Finding the store with the highest turnover rate", x="Store name", y="Count", fill = "Termination Reason")
```

Figure 3.3.1.1 Snippet Code for Analysis 8

Using the code above, a visualization in a form of a bar chart can be created with the bar chart counting all the resigned and laid off employees by each store, to understand which store has the highest layoff number and which store has the highest resignation number. The termination reason will be color-coded to see the distribution of the reason in each store.

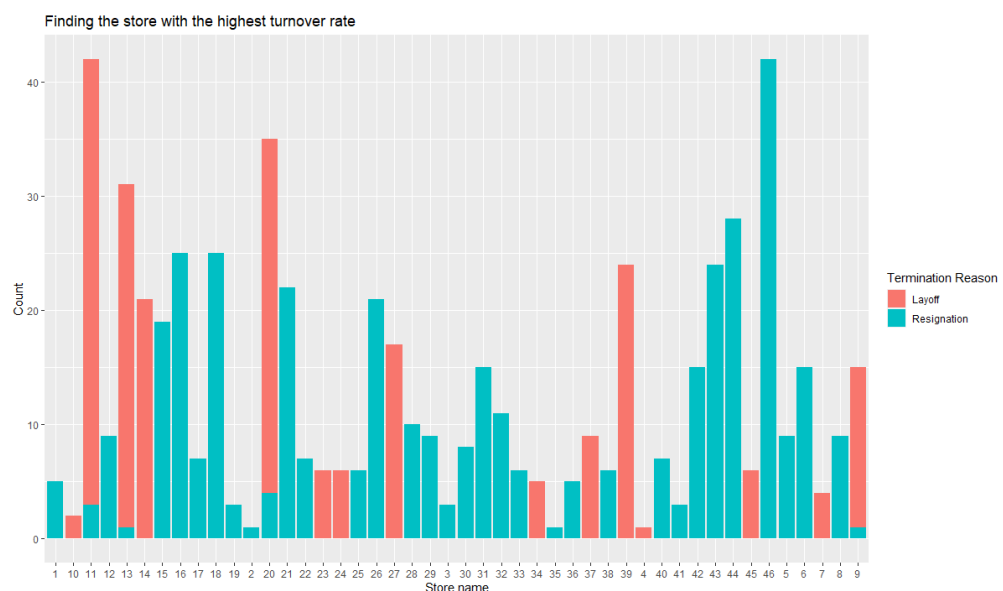


Figure 3.3.1.2 Visualization for Analysis 8

Based on the visualization above, store 11 and 46 stands out as the store with the highest layoff and resignation number respectively. From this visualization, a further analysis can be done to get a clearer understanding for this result.

3.3.2 Analysis 9 – Analysing store with most layoff's job distribution

To get a better understanding towards each store, analysis towards each store's job distribution should be done to understand which jobs has the highest turnover count with the hope to get some clarity for the situation.

```
#Analysis 9 - Finding each store's jobs distribution (Store 11)
main_data%>%
  filter(store_name == "11")%>%
  filter(termreason_desc != "Retirement")%>%
  ggplot(aes(x=job_title, fill = termreason_desc))+
  geom_bar()+
  labs(title = "Finding the store with the highest turnover rate", subtitle = "Store 11 - Fort Nelson", x="Job Title", y="Count",
        fill = "Termination Reason")

#Supporting analysis for Analysis 9 - Checking the number of active employee in store 11 by each year
main_data%>%
  filter(store_name == "11")%>%
  filter(status == "ACTIVE")%>%
  group_by(status_year)%>%
  count()
```

Figure 3.3.2.1 Snippet Code for Analysis 9

For the second part of the analysis, the code above will generate a bar chart like Analysis 4, but the difference is it is showing the count of the laid off/resigned employee from each job title and the count, specifically towards Store 11 which is in Fort Nelson. A supporting analysis will also be made by checking the number of active employees in the store by each year.

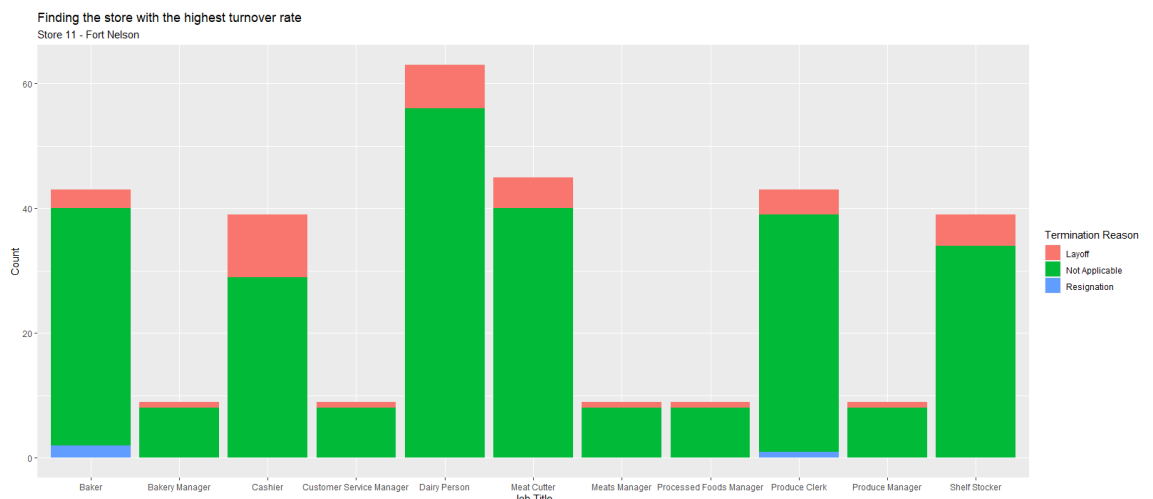


Figure 3.3.2.2 Visualization for Analysis 9

STATUS_YEAR	n
<int> <int>	
2006	29
2007	33
2008	32
2009	34
2010	35
2011	35
2012	38
2013	39

Figure 3.3.2.3 Supporting Analysis for Analysis 9

With the resulting analysis, an understanding that all positions in store 11 got laid off by the company. This understanding is also reinforced by the supporting analysis that Fort Nelson store also has been closed during 2014, with no employee left hired assigned to the store. This means the high layoff is due to the company's choice to close the branch. This closing is most likely caused by the company downsizing/restructuring that has been discovered from the previous question.

3.3.3 Analysis 10 – Analysing store with most resignation's job distribution

```
#Analysis 10 - Finding each store's jobs distribution (Store 46)
main_data%>%
  filter(STATUS == "TERMINATED")%>%
  filter(store_name == "46")%>%
  filter(termination_desc != "Retirement")%>%
  ggplot(aes(x=job_title, fill = termination_desc))+
  geom_bar()+
  labs(title = "Finding the store with the highest turnover rate", subtitle = "Store 46 - Victoria", x="Job Title", y="Count",
        fill = "Termination Reason")

#Supporting analysis for Analysis 10 - Checking the number of active employee in store 46 by each year
main_data%>%
  filter(store_name == "46")%>%
  filter(STATUS == "ACTIVE")%>%
  group_by(STATUS_YEAR)%>%
  count()
```

Figure 3.2.3.1 Snippet Code for Analysis 10

For the store with most resignation, store 46 located in Victoria, the analysis will be conducted using this code to generate a bar chart like Analysis 9. The same supporting analysis will also be generated to know how many active employees in the store by each year.

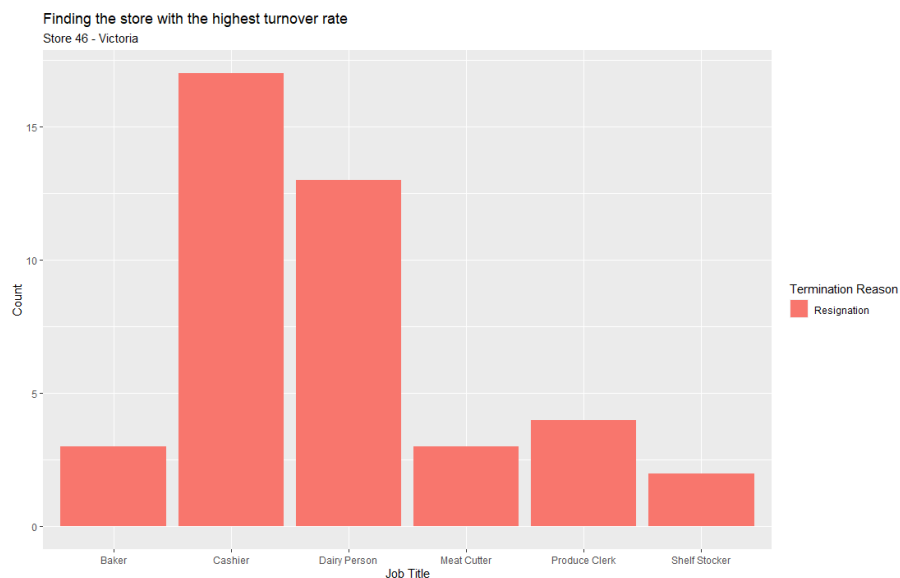


Figure 3.2.3.2 Visualization for Analysis 10

STATUS_YEAR	n
1	2006 353
2	2007 374
3	2008 405
4	2009 424
5	2010 448
6	2011 461
7	2012 470
8	2013 488
9	2014 477
10	2015 473

Figure 3.2.3.3 Supporting Analysis for Analysis 10

After the visualization, a conclusion can be made that Store 46 located in Victoria is still active during 2015, with no recorded layoff. This may

be due to the location of the store in the province's capital, which mean the store's traffic can be considered very busy. For the job with resignation's distribution analysis, all the job titles with a resignation record are usually a short-term job for a big city, such as Cashier, Produce Clerk, Shelf Stocker, that have a low-skill ceiling and usually can be used as temporary jobs, based on analysis 2 that has been discussed above. Other than the reason above, Resignation also can be caused by bad managerial control inside the store, which can be investigated further by the human relation department of the company.

3.4 Question 4 – When is the best time for hiring spree?

3.4.1 Analysis 11 – Amount of Resignation per Month (Per Year)

To prevent a lot of sudden resignation, an analysis has been done to show what months that have the highest count of people resigning.

```
#Analysis 1 - Resignation Count - Per Month (Per Year)
main_data%>%
  filter(termreason_desc == "Resignation")%>%
  mutate(termination_month = format(format(terminationdate_key, format = "%m")))%>%
  mutate(termination_year = format(format(terminationdate_key, format = "%Y")))%>%
  ggplot(aes(x= termination_month))+
  geom_bar()+
  facet_wrap(~termination_year)+
  labs(title = "Amount of Resignation per Month (Per Year)", x='Month', y = 'Count')
```

Figure 3.4.1.1 Snippet Code for Analysis 11

The code above will generate a bar chart that represent the distribution of people resigning by the months, with a separate chart for each year from 2006 – 2015.

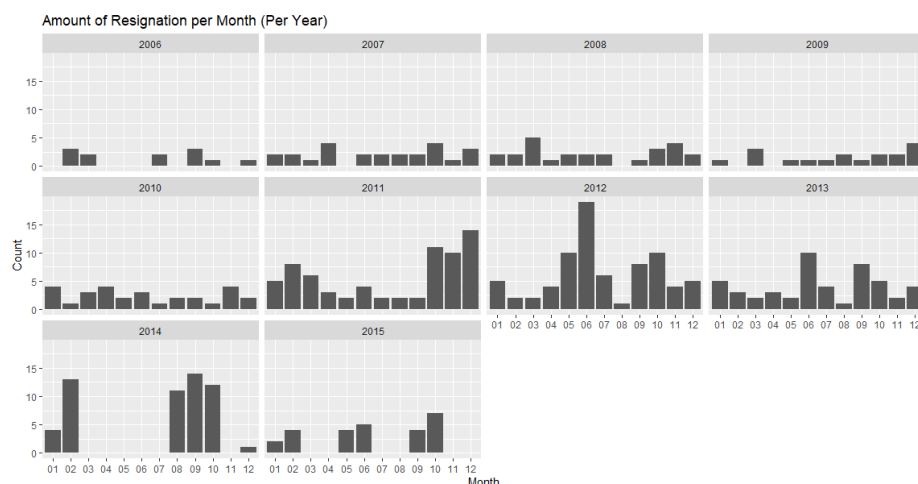


Figure 3.4.1.2 Visualization for Analysis 11

From the analysis above, we can conclude that a lot of resignation happens around September until November, as the visualization above shows that mostly on those months, the resignation count goes up, especially 2011 onwards. In 2012, there is a massive spike of resignation during 2012 that can be affected from the increase in job vacancies during June 2012 in Canada (Statistics Canada, 2012).

3.4.2 Analysis 12 – Amount of Resignation per Month (Average)

As a continuation for Analysis 11, The analysis can be averaged by combining all year's data into one, which can become a conclusion for the question.

```
#Analysis 2 - Resignation Count - Per Month (Average)
main_data%>%
  filter(termreason_desc == "Resignation")%>%
  mutate(termination_month = format(format(terminationdate_key, format = "%m")))%>%
  ggplot(aes(x= termination_month))+
  geom_bar()+
  labs(title = "Amount of Resignation per Month (Average)", x='Month', y = 'Count')
```

Figure 3.4.2.1 Snippet Code for Analysis 12

The Code above will generate a bar chart like Analysis 13, but all year's data will be combined into one, creating an average bar chart for the question.

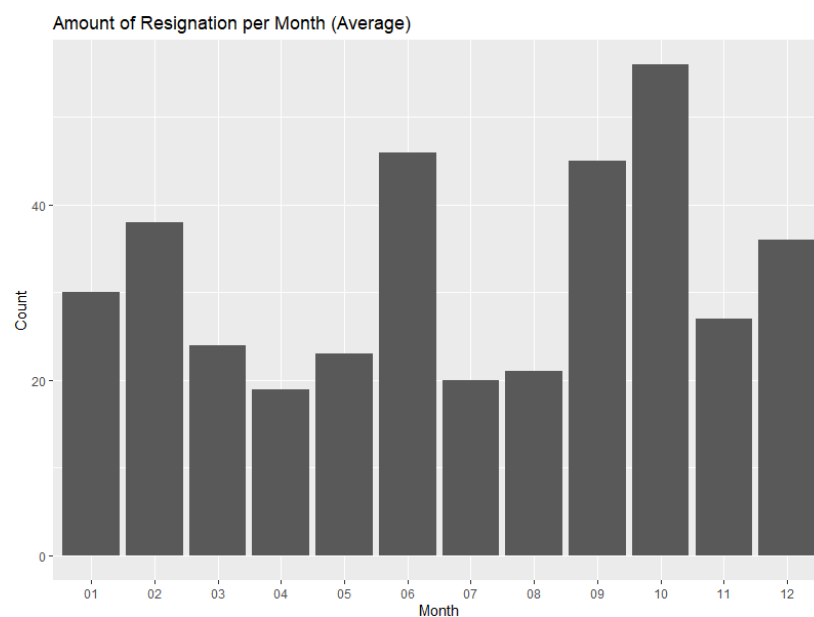


Figure 3.4.2.2 Visualization for Analysis 12

The graph shows that the months with most resignation are June, September, and October. September and October can be the result of Canada University's primary intake that started on September-December (Team Leverage Edu, 2021), so a lot of high school graduates that took a short-term job will be quitting around that time.

4.0 Extra Feature

4.1 Extra Feature 1 - Comparison Bar Chart

In Analysis 4, 5, and 6, to create a bar chart that compares between 2 fixed conditions, a comparison bar chart can be created. To make sure the data is inserted correctly, and only relevant data is taken into consideration, the first step is to create two data frames with only the necessary data.

```
data51<- main_data%>%
  group_by(store_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2006")%>%
  count()
data51['year']='2006'
data51

data52 <- main_data%>%
  group_by(store_name)%>%
  filter(STATUS == "ACTIVE")%>%
  filter(STATUS_YEAR == "2015")%>%
  count()
data52['year']='2015'
data52
```

Figure 4.1.1 Snippet Code for Extra Feature 1

With the code above, data61 data frame will be grouped by each store name. The taken data will only the one with “ACTIVE” status and “2006” status year. The data will then be counted using the count() method, which means the data will be counted by each city name. After that, a new variable called “year” can be inserted with all rows containing the value 2006 for creating the bar chart. The same process happened towards data62 data frame, with changes towards the filter being “2015” status year and all rows under variable “year” being filled with 2015 value.

```
data53 = rbind(data51,data52)
rm(data51)
rm(data52)
```

Figure 4.1.2 Snippet Code for Extra Feature 1

After both data frame has been created, the data frame will be combined by the row to create a new data frame containing all values from the two. In this case, the data63 data frame will be the final data frame that will be visualized, while the original two data frame will be deleted since it fulfilled its function.

```
data53%>%  
  ggplot(aes(fill=year, y=n, x=store_name))+  
  geom_bar(position="dodge", stat="identity")+  
  labs(title = "Comparing the store condition between 2006 and 2015",  
        x="Store Name", y="Employee Count", fill="Year")
```

Figure 4.1.3 Snippet Code for Extra Feature 1

After that, the data frame is ready for visualization. The visualization will be based on the store name on its X-axis, while the count of each city will be on its Y-Axis. The bar will also be split into two by the year, so by 2006 in red coloured bar and 2015 in blue coloured bar. After that, the bar chart position will be “dodge”, meaning the bar will not be stacked, instead display side by side to create the comparison look. Since the data will be directly compared towards the Y-axis, stat “identity” needs to be used. Finally, the data will be titled, with each axis and fill being named for easier understanding using labs() function.

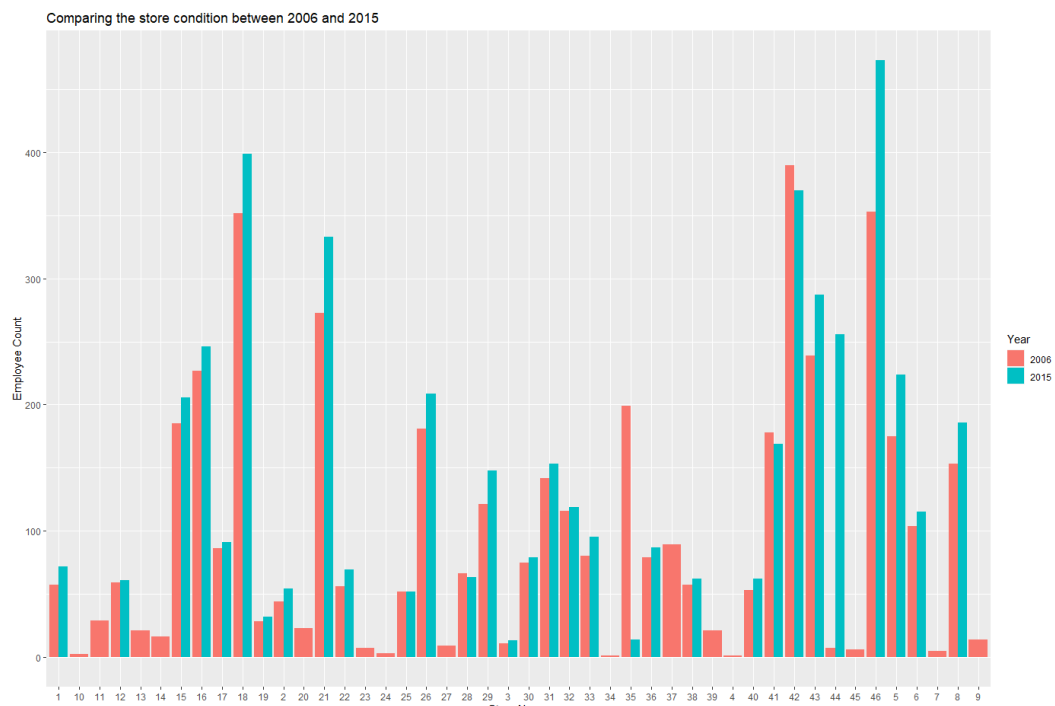


Figure 4.1.4 Result from Extra Feature 1

4.2 Boxplot with Jitter overlay and mean position

In Analysis 1, the boxplot visualization is chosen so the age number can be shown with its boundary and median. But to show the distribution more clearly, a jitter plot is much more reliable and functional. The mean position is also added separately so a better understanding can be reached using the mean.

```
main_data%>%
  filter(termttype_desc == "voluntary")%>%
  ggplot(aes(y= age, x=termreason_desc))+
  geom_boxplot(alpha = 0)+
  geom_jitter(alpha = 0.5, aes(colour = "orange"))+
  facet_wrap(~gender_full)+
  stat_summary(fun=mean, geom="point", shape=20, size=3, color="blue")+
  labs(title = "Correlation between Termination Reason and Age, separated by Gender",
       x='Termination Reason', y = 'Age')
```

Figure 4.2.1 Snippet Code for Extra Feature 2

The code above is to create the visualization. Since the analysis focus on the age distribution of voluntarily terminated people, the data will be filtered so only termination type “Voluntary” will be used. Next, the plot will contain X-axis with data from termreason_desc variable, while the Y-axis took the data from age variable. After that, the boxplot will be created with alpha on 0 value or fully nontransparent, while the jitter plot will be alpha = 0.5 or half transparent. The colour of the jitter point is also changed to be orange. These changes are made so the boxplot can still be visible through the jitter plot. The jitter plot will also be coloured in orange for easier identification. Last, the mean position will use stat_summary() method, choosing mean to show, with geom = “point” to pick the geometric object “point” for the display, shape 20 which is a circle, size 3, and color blue for easier identification using labs function.

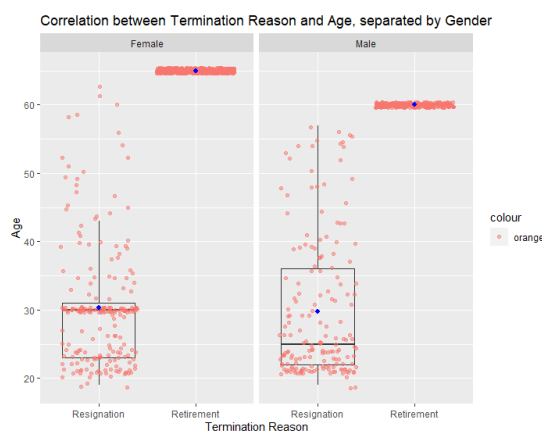


Figure 4.2.2 Result from Extra Feature 2

Conclusion

In Conclusion, a baseline condition of the company can be made that the company is facing decline by the condition of its employees, sudden high layoff rate, closure of several departments, number of stores that closed, and loss of presence in several cities ever since 2014 onwards. By the condition mentioned above, the company has been taking steps towards downsizing since 2014, since a lot of stores has been closed and a lot of layoffs has happened. From the data above, a recommendation for a full restructuring of the company is necessary to keep competing in currently advancing digital world, where online grocery shopping has become more common by the day. The company restructure also can be in form of creating a more division between departments, since in 2006, a lot of departments that is working are too specific, with a couple of departments can be combined into one without much problem.

A company restructure is very likely and can be good thing, since it may give the company a good fighting chance towards old and new competitors that may have embraced the new digital landscape, especially online shopping, into their company focus. After the downsizing, the company may use when to hire analysis and why people are leaving their jobs analysis to understand and how they can improve their workforce's needs.

References

- Burns, E. (15 January, 2021). *Data Cleaning in R Made Simple*. Retrieved from Towards Data Science: <https://towardsdatascience.com/data-cleaning-in-r-made-simple-1b77303b0b17> [Accessed 16th November 2021]
- GeeksforGeeks. (29 June, 2021). *Data Preprocessing in Data Mining*. Retrieved from GeeksforGeeks: <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/> [Accessed 16th November 2021]
- Government of Canada. (n.d.). *CPP retirement pension: Overview*. Retrieved from Government of Canada: <https://www.canada.ca/en/services/benefits/publicpensions/cpp.html> [Accessed 16th November 2021]
- Provencher, C., Milan, A., Hallman, S., & D'Aoust, C. (n.d.). *Fertility: Overview, 2012 to 2016*. Retrieved from Statistics Canada: <https://www150.statcan.gc.ca/n1/pub/91-209-x/2018001/article/54956-eng.htm> [Accessed 16th November 2021]
- Statistics Canada. (18 September, 2012). *Job vacancies, three-month period ending in June 2012*. Retrieved from Statistics Canada: <https://www150.statcan.gc.ca/n1/daily-quotidien/120918/dq120918a-eng.htm> [Accessed 16th November 2021]
- Statistics Canada. (5 November, 2019). *Postsecondary graduates, by location of residence at interview and level of study*. Retrieved from Statistics Canada: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3710003101> [Accessed 16th November 2021]
- Stitch. (n.d.). *What is data transformation: definition, benefits, and uses*. Retrieved from Stitch Data: <https://www.stitchdata.com/resources/data-transformation/> [Accessed 16th November 2021]
- Talent. (2021). *Cashier average salary in Canada 2021*. Retrieved from Talent: <https://ca.talent.com/salary?job=cashier> [Accessed 19th November 2021]
- Team Leverage Edu. (19 November, 2021). *Upcoming Intakes in Canada for 2022-2023*. Retrieved from Leverage Edu: <https://leverageedu.com/blog/intakes-in-canada/> [Accessed 16th November 2021]