INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

KHOR ZHEN WIN – TP055619

APD2F2109CS(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 submissions will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld.**

4 **Cases of plagiarism will be penalized.**

5 **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.**

6 **You must obtain 50% overall to pass this module.**

Table of Contents

1.0 Introduction

The dataset that will be discussed in this report is the employee attrition of an organization. The purpose is to investigate any trends and relations involving any issues in human resources management. Meaningful insight is to be produced from this dataset to identify any underlying issues found in this organization. The general reasoning of workers leaving the workforce of this organization should be able to be identified from the analysis. The assumptions made will be in accordance with the relevancy of this dataset and general assumptions will be made if there is a lack of clarification.

1.1 Assumptions

- The external factors happening at the time of collection of this data set will be accepted as valid factors affecting the attrition rate. (Economy status, political stability)
- Dataset is collected specific to this organization only
- The analytic tool used (RStudio) is reliable and accurate
- The data set is the entire population of the organization where it is sufficient to represent the entire organization in terms of attrition rate and is not a sample data set
- This organization is assumed to be based in Canada and references will be made based on the geographical location

2.0 Data Exploration

2.1 Data Import

```
4  # -------------------------Installing Packages-----------------------------------
5
6  # Install the package to be used
7  # Load the package
8
9  install.packages("ggplot2")
10 library(ggplot2)
11
12 install.packages("dplyr")
13 library(dplyr)
14
15
```

Figure 1: Installing & Loading Utility Packages to Help in Analysis

Before importing the data, helpful utility packages should be installed first to help in the later stage of mutating, transforming, and analysing the data. After installing the packages, they need to be loaded into the current RStudio session.

```
17 # -------------------------Importing Data-----------------------------------------
18
19 # Storing the csv file into a local tibble
20 employee_attrition = read.csv(file="D:\\Program Files\\R\\data\\employee_attrition.csv")
21 employee_attrition
22
```

Figure 2: Loading the Data File into RStudio

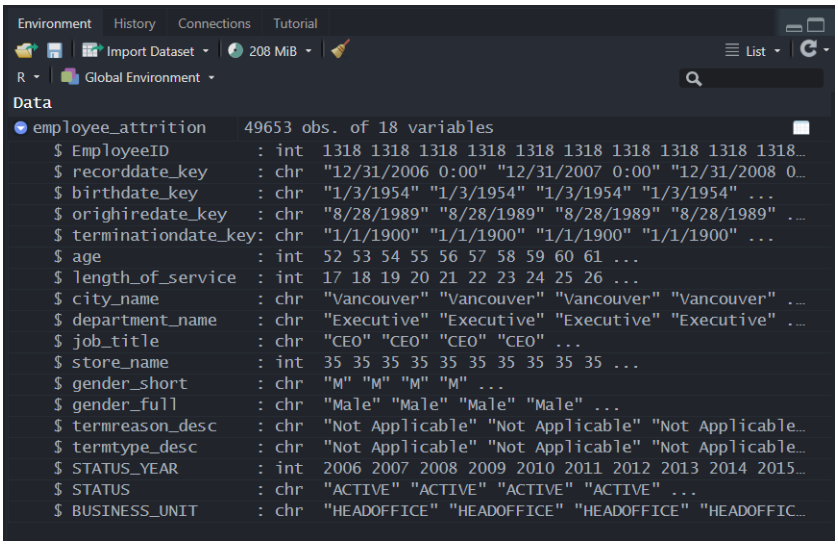


Figure 3: Local Variable That Holds the CSV Contents in a Data Frame

After importing the csv file into a variable, the csv will be stored as a data frame and can be now processed in RStudio.

2.2 Null Values

```
R 4.1.1 · C:/WINDOWS/system32/
> # Checking for any empty values in any columns
> apply(employee_attrition, 2, function(x) any(is.na(x)))
        EmployeeID      recorddate_key       birthdate_key    orighiredate_key
             FALSE               FALSE               FALSE               FALSE
terminationdate_key                 age  length_of_service           city_name
             FALSE               FALSE               FALSE               FALSE
    department_name           job_title          store_name       gender_short
             FALSE               FALSE               FALSE               FALSE
       gender_full     termreason_desc       termtype_desc         STATUS_YEAR
             FALSE               FALSE               FALSE               FALSE
            STATUS       BUSINESS_UNIT
             FALSE               FALSE
>
```

Figure 4: Checking for Any Null Values of All Columns

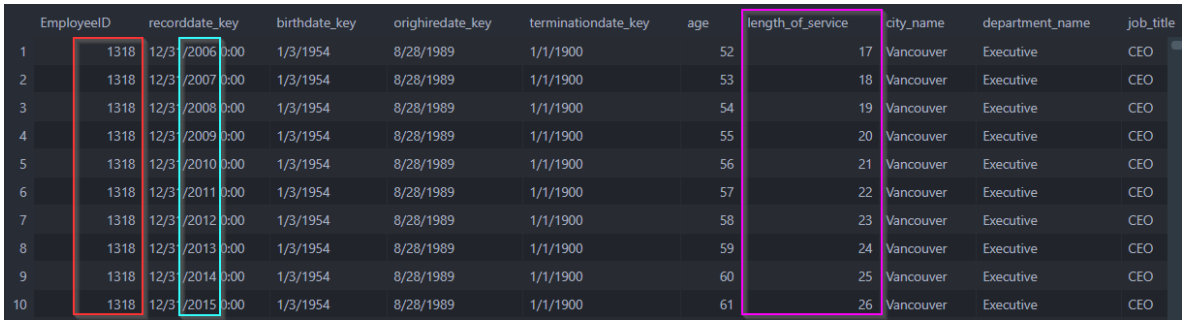
From the code used, it has been identified that there are no missing values in this data set. Therefore, there is no need to handle missing values during pre-processing.

2.3 Duplicate Data

```
Console  Terminal ×  Jobs ×
R 4.1.1 · C:/WINDOWS/system32/
> # Checking for duplicates with the main identifier
> employee_attrition %>% group_by(EmployeeID) %>% summarise(n = n(),)
# A tibble: 6,284 x 2
   EmployeeID     n
        <int> <int>
 1       1318    10
 2       1319    10
 3       1320    10
 4       1321    10
 5       1322    10
 6       1323    10
 7       1325    10
 8       1328    10
 9       1329    10
10       1330    10
# ... with 6,274 more rows
>
```

Figure 5: Checking for Duplicate Data Using EmployeeID

Assuming the main identifier used for this data set is EmployeeID, this column will be checked for duplicated records as each row should be unique. After exploring, the data is seen to have consistent repetitions among each employee. Therefore, an analysis must be made by looking at the raw data to draw an inference on how to drop the duplicate rows.



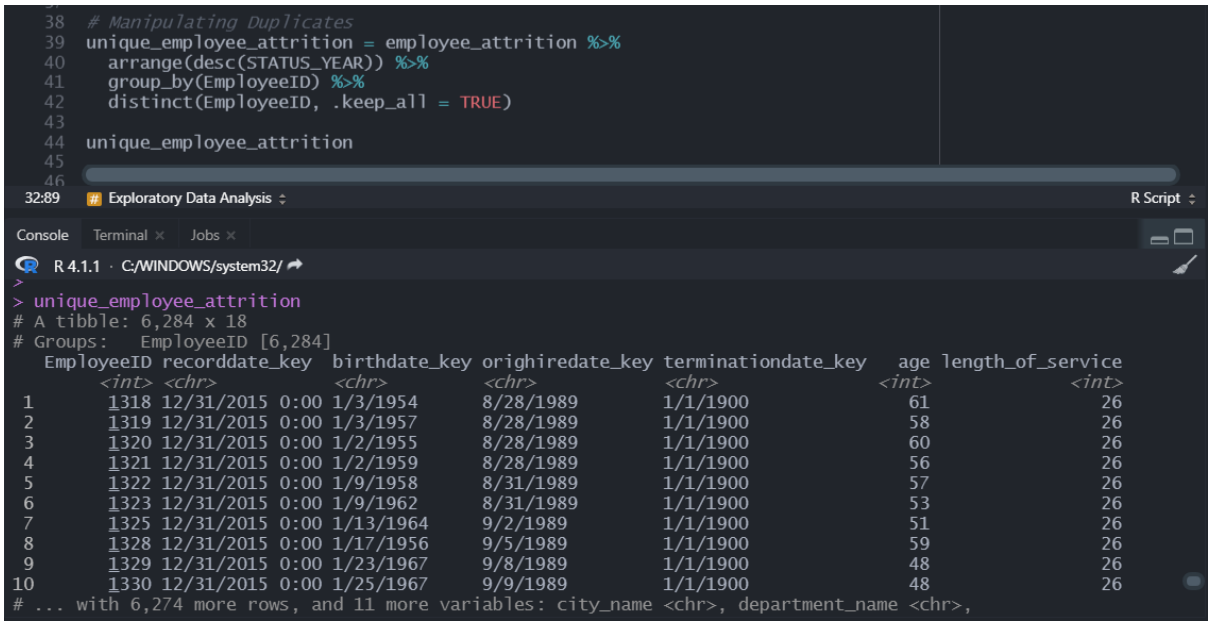
	EmployeeID	recorddate_key	birthdate_key	orighiredate_key	terminationdate_key	age	length_of_service	city_name	department_name	job_title
1	1318	12/3 /2006 0:00	1/3/1954	8/28/1989	1/1/1900	52	17	Vancouver	Executive	CEO
2	1318	12/3 /2007 0:00	1/3/1954	8/28/1989	1/1/1900	53	18	Vancouver	Executive	CEO
3	1318	12/3 /2008 0:00	1/3/1954	8/28/1989	1/1/1900	54	19	Vancouver	Executive	CEO
4	1318	12/3 /2009 0:00	1/3/1954	8/28/1989	1/1/1900	55	20	Vancouver	Executive	CEO
5	1318	12/3 /2010 0:00	1/3/1954	8/28/1989	1/1/1900	56	21	Vancouver	Executive	CEO
6	1318	12/3 /2011 0:00	1/3/1954	8/28/1989	1/1/1900	57	22	Vancouver	Executive	CEO
7	1318	12/3 /2012 0:00	1/3/1954	8/28/1989	1/1/1900	58	23	Vancouver	Executive	CEO
8	1318	12/3 /2013 0:00	1/3/1954	8/28/1989	1/1/1900	59	24	Vancouver	Executive	CEO
9	1318	12/3 /2014 0:00	1/3/1954	8/28/1989	1/1/1900	60	25	Vancouver	Executive	CEO
10	1318	12/3 /2015 0:00	1/3/1954	8/28/1989	1/1/1900	61	26	Vancouver	Executive	CEO

Figure 6: Raw Data in CSV Format

After looking at the granular data, it is inferred that each row contains the same data but is only incrementing certain values. These values are the recorded date, the age of the employee, their length of service, and a few others. From this exploration, we can conclude to keep only the latest record for each employee. It would have the latest age and length of service which would make the older records irrelevant and outdated, so they will be dropped.

3.0 Data Manipulation

3.1 Dropping Duplicate Rows



```
# Manipulating Duplicates
unique_employee_attrition = employee_attrition %>%
  arrange(desc(STATUS_YEAR)) %>%
  group_by(EmployeeID) %>%
  distinct(EmployeeID, .keep_all = TRUE)

unique_employee_attrition
```

```
32:89  Exploratory Data Analysis                                  R Script
Console  Terminal  Jobs
R 4.1.1 · C:/WINDOWS/system32/
> unique_employee_attrition
# A tibble: 6,284 x 18
# Groups:   EmployeeID [6,284]
   EmployeeID recorddate_key  birthdate_key orighiredate_key terminationdate_key   age length_of_service
        <int> <chr>           <chr>         <chr>            <chr>               <int>             <int>
 1       1318 12/31/2015 0:00 1/3/1954      8/28/1989        1/1/1900               61                26
 2       1319 12/31/2015 0:00 1/3/1957      8/28/1989        1/1/1900               58                26
 3       1320 12/31/2015 0:00 1/2/1955      8/28/1989        1/1/1900               60                26
 4       1321 12/31/2015 0:00 1/2/1959      8/28/1989        1/1/1900               56                26
 5       1322 12/31/2015 0:00 1/9/1958      8/31/1989        1/1/1900               57                26
 6       1323 12/31/2015 0:00 1/9/1962      8/31/1989        1/1/1900               53                26
 7       1325 12/31/2015 0:00 1/13/1964     9/2/1989         1/1/1900               51                26
 8       1328 12/31/2015 0:00 1/17/1956     9/5/1989         1/1/1900               59                26
 9       1329 12/31/2015 0:00 1/23/1967     9/8/1989         1/1/1900               48                26
10       1330 12/31/2015 0:00 1/25/1967     9/9/1989         1/1/1900               48                26
# ... with 6,274 more rows, and 11 more variables: city_name <chr>, department_name <chr>,
```

Figure 7: Keeping Only Latest Values of Each Employee

By using the STATUS YEAR as the comparison column, the year of the record is arranged into descending order. This is because the latest status year is assumed to be the latest row of data for that employee. The first occurrence of the unique employee ID will be kept as it is now in descending order, and the rest of the rows will be dropped. Meaning that the dropped records is the old data of employees from previous years.

3.2 Changing Column Names

Original Column Name	Renamed Columns	Column Description
EmployeeID	Employee_ID	Employee identity code
recorddate_key	Record_Date	Employee record date (Year of data)
birthdate_key	Birth_Date	Employee birth date
orighiredate_key	Hired_Date	Employee hired date
terminationdate_key	Termination_Date	Employee termination date
age	Age	Employee age
length_of_service	Service_Length	Length of service in year
city_name	City	City Name
department_name	Department	Department Name
job_title	Job_Title	Job Title
store_name	Store_Code	Store Code
gender_short	Gender	Gender in short
gender_full	Gender_Full	Gender in full
termreason_desc	Termination_Reason	Reason of termination
termtype_desc	Termination_Type	Type if termination
STATUS_YEAR	Status_Year	Year of status
STATUS	Status	Employee status in company
BUSINESS_UNIT	Business_Unit	Business unit of employee

Renaming the original column headers would make it easier for analysis as well as having more consistent and standardized naming conventions for the overall dataset. This would make pre-processing and analysis easier in the future.

4.0 Data Transformation

4.1 Termination Year + Month

```
72  # Extracting termination year and adding it into a new column
73  unique_employee_attrition$Termination_Year =
74    format(as.Date(unique_employee_attrition$Termination_Date, "%m/%d/%Y"), "%Y")
75
76  # Extracting termination month and adding it into a new column
77  unique_employee_attrition$Termination_Month =
78    format(as.Date(unique_employee_attrition$Termination_Date, "%m/%d/%Y"), "%m")
79
```

Figure 8: Extracting Terminated Year & Month into A New Column

By transforming the date given into the year and month of termination, the classification of year would be easier. This would make further analysis much easier in the later by using this column.

Status	Business_Unit	Termination_Year	Termination_Month
ACTIVE	HEADOFFICE	1900	01
ACTIVE	HEADOFFICE	1900	01
ACTIVE	HEADOFFICE	1900	01
ACTIVE	HEADOFFICE	1900	01
ACTIVE	HEADOFFICE	1900	01
ACTIVE	HEADOFFICE	1900	01
ACTIVE	HEADOFFICE	1900	01
ACTIVE	HEADOFFICE	1900	01
ACTIVE	STORES	1900	01

Figure 9: New Column for Termination Year and Month

4.2 Hired Year + Month

```
80  # Extracting hired year and adding it into a new column
81  unique_employee_attrition$Hired_Year =
82    format(as.Date(unique_employee_attrition$Hired_Date, "%m/%d/%Y"), "%Y")
83
84  # Extracting hired month and adding it into a new column
85  unique_employee_attrition$Hired_Month =
86    format(as.Date(unique_employee_attrition$Hired_Date, "%m/%d/%Y"), "%m")
87
```

Figure 10: Extracting Hired Year & Month into A New Column

By transforming the date given into the year and month, analysing the column by the classification of year would be easier. This would make further analysis much easier in the later by using this column.

Termination_Year	Termination_Month	Hired_Year	Hired_Month
1900	01	1989	08
1900	01	1989	08
1900	01	1989	08
1900	01	1989	08
1900	01	1989	08
1900	01	1989	08
1900	01	1989	09
1900	01	1989	09

Figure 11: New Column for Hired Year and Month

4.3 Age Group

```
77  # Classifying age into ranges
78  min(unique_employee_attrition$Age)
79  max(unique_employee_attrition$Age)
80
81  unique_employee_attrition =
82    mutate(unique_employee_attrition, Age_Group =
83             case_when(Age >= 18  & Age <= 24 ~ '18-24',
84                       Age >= 25  & Age <= 34 ~ '25-34',
85                       Age >= 35  & Age <= 44 ~ '35-44',
86                       Age >= 45  & Age <= 54 ~ '45-54',
87                       Age >= 55  ~ '55 Above'))
88
```

Figure 12: Classifying Age into Groups

Firstly, the minimum and maximum age is needed to be checked to classify the boundaries of the groups. A new column is added to avoid replacing the original value, so that it can still be used. The age ranges are classified into 5 major range groups based on the minimum and maximum range of age in this data set. The output in the data frame can be seen below.

```
> sample(unique_employee_attrition)
# A tibble: 6,284 x 20
# Groups:   Employee_ID [6,284]
   Gender_Full Hired_Date Gender Department       Age_Group Status Service_Length
   <chr>       <chr>      <chr>  <chr>            <chr>     <chr>           <int>
 1 Male        8/28/1989  M      Executive        55 Above  ACTIVE             26
 2 Female      8/28/1989  F      Executive        55 Above  ACTIVE             26
 3 Female      8/28/1989  F      Executive        55 Above  ACTIVE             26
 4 Male        8/28/1989  M      Executive        55 Above  ACTIVE             26
 5 Male        8/31/1989  M      Executive        55 Above  ACTIVE             26
 6 Male        8/31/1989  M      Executive        45-54     ACTIVE             26
 7 Female      9/2/1989   F      Executive        45-54     ACTIVE             26
 8 Female      9/5/1989   F      Executive        55 Above  ACTIVE             26
 9 Female      9/8/1989   F      Store Management 45-54     ACTIVE             26
10 Female      9/9/1989   F      Meats            45-54     ACTIVE             26
```

Figure 13: Sample of The New Age_Group Column

4.4 Transforming Employees Data Set into ACTIVE & TERMINATED

```
105 terminated_employees = subset(unique_employee_attrition,
106                               Status != "ACTIVE")
107
108 active_employees = subset(unique_employee_attrition,
109                             Status == "ACTIVE")
110
```

Figure 14: Code to Create New Data Frames Only for Terminated & Active Employees

Two new data frames that consists only of TERMINATED and ACTIVE employees have been created in this section. This is so that further analysis can be done where the situations and insights are only applicable for either active or terminated employees only.

5.0 Data Visualization

5.1 Question 1

Which Time Period Has the Highest Terminated Employees and Why?

5.1.1 Analysis 1 – Termination by Month

What is the distribution of the termination reason by month?

```
112 ggplot(terminated_employees,aes(x=Termination_Month, fill=Termination_Reason)) +
113   geom_bar() +
114   labs(title = "Distribution of Termination Reason in Terminated Employees", x = "Month", y = "No. Of Employees")
115
```

Figure 15: Code to Plot Bar Graph of Terminated Employees by Terminated Month, Stacked by Termination Reason

The aim for this visualization is to view the distribution of only the terminated employees by month. This plot would show which month the employees would leave the organization by resignation or retirement. It should also show when employees are laid off, which signifies when human resources make their major financial related decisions.

Using the ggplot2 package, the data set that only contains terminated employees is used in this plot. A stacked bar chart was plotted where the x axis is populated with the Termination Month and stacked with the Termination Reason. A stacked bar chart was used here as it better represents and separates two categorical values in one visualization.

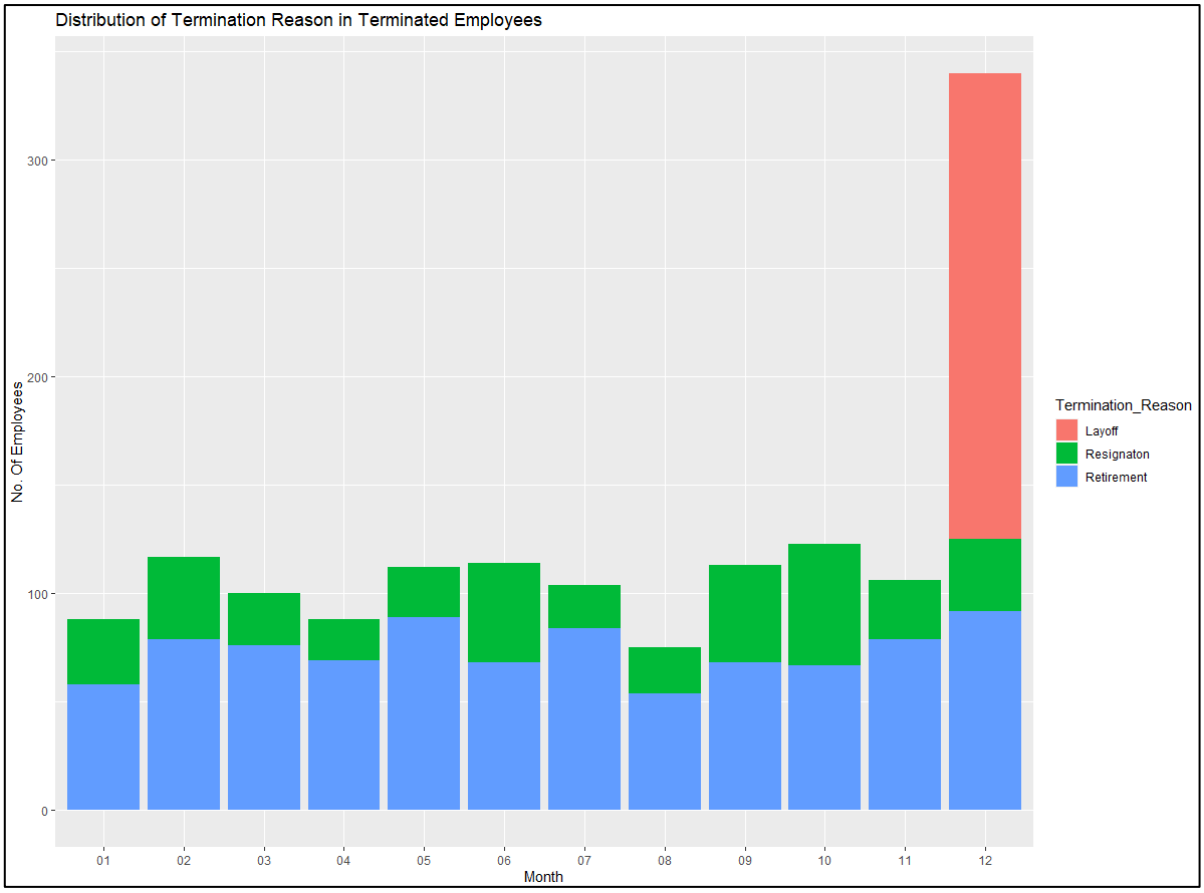


Figure 16: Bar Graph of Terminated Employees by Terminated Month, Stacked by Termination Reason

From the graph above, we can identify those layoffs happened during the month of December. Using this identification, we can assume that this organization is seen to make major business decisions at the end of the year. Assuming December is the end of the accounting year for this organization, termination related decisions are based off their financial health of that accounting year (Investopedia, 2021).

Layoffs are seen to only happen in December while resignations and retirements are evenly distributed across all the months. This reinforces the statement that layoffs happen in times of bad financial health where cost-cutting occurs in terms of staff reduction. The equally distributed resignations and retirements shows no significant pattern on the preferred month to leave. From this, it cannot be identified if there is any relation during the start, middle, or the end of the year (CFI Education Inc, 2021).

An organization could also encourage an employee to resign on their own accord. Tending to an employee to encourage them to leave voluntarily is to help the employee to better understand it would be in their best interest to do so. If the organization knows that it's not in the best financial health, encouraging their employees to seek opportunities elsewhere might be a better option when they have mouths to feed. However, this approach is not suitable for every situation and is only a hypothesis without integral evidence of this specific organization (McCarthy, 2019).

Layoffs are usually not the best solution in terms of cutting cost. Organizations can encourage voluntary retirement as discussed previously, especially tentatively to people of old age. Voluntary retirement is effective in cost reduction since these employees often have the highest paycheck. Furthermore, different avenues of cutting cost can be seen such as considering a virtual office as well as cutting back on travel cost and bonuses (CFI Education Inc, 2021).

5.1.2 Analysis 2 – Termination by Year

What is the distribution of the termination reason by year?

From the previous analysis done on termination by months, this organization lays off employees in efforts of cutting costs by downsizing. To further justify this hypothesis, a further analysis needs to be done on the terminations by year. This would provide some insight on external factors such as economic and political stability at that time that might have an influence on the economy that would drive this organization to lay off their employees.

```
124  # Termination Reason distribution by Year
125  ggplot(terminated_employees,aes(x=Termination_Year, fill=Termination_Reason)) +
126    geom_bar() +
127    labs(title = "Distribution of Termination Reason in Terminated Employees"
128         , x = "Year", y = "No. Of Employees")
129
```

Figure 17: Code to Plot Stacked Bar Graph of Terminated Employees by Terminated Year

The aim for this visualization is to view the distribution of the terminated employees by year. The plot should further show what was the reason of termination each year. This plot should be able to show the trend of either an increase or decrease of terminations in a time frame. Besides that, the visualization should show which year the organization started laying off employees.

Using the ggplot2 package, the terminated_employees data set is used in this plot. A stacked bar chart was plotted where the Termination Year is plotted on the x axis while stacking them with the Termination Reason. A stacked bar chart was used here as it better represents and separates two categorical values in one visualization.

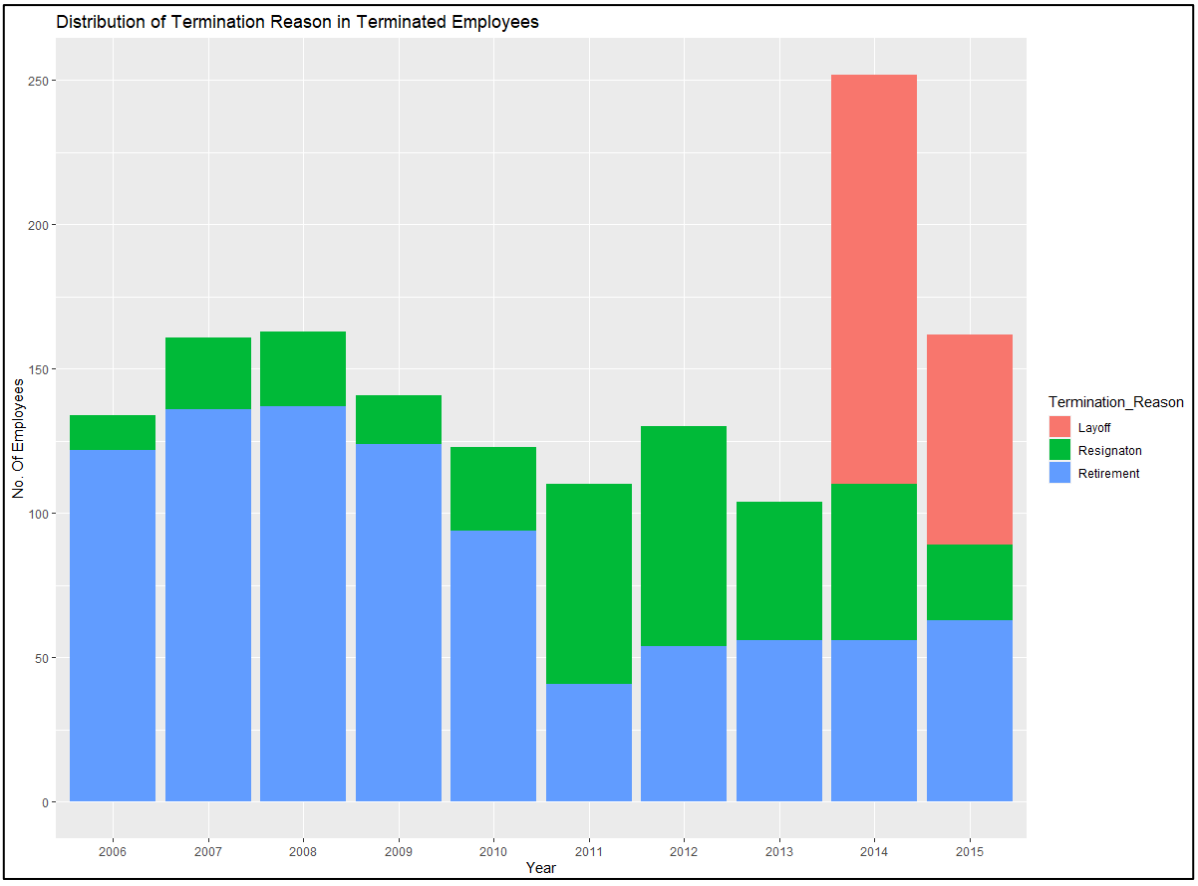


Figure 18: Stacked Bar Graph of Terminated Employees by Terminated Year

From the graph above, layoffs only occurred during 2014 and 2015. This narrows the research scope to be done on these two years to investigate the cause of the downtrend of financial health for this organization. Besides that, we can also see an increase in resignations from 2011 to 2015 which should also be an area of consideration towards the later research. The geographical location of this organization, which is assumed to be based in Canada, will be used as a point of reference in the justification research.

The GDP of Canada from the years 2011 to 2016 can be seen from the line graph below. There are 2 steep declines to be highlighted which is between year 2011 to 2012 and year 2014 to 2015. The decline from 2011 to 2012 reflects the higher resignation rates during those years while the decline from 2014 to 2015 reflects why the organization started to lay off employees (Investopedia, 2021).

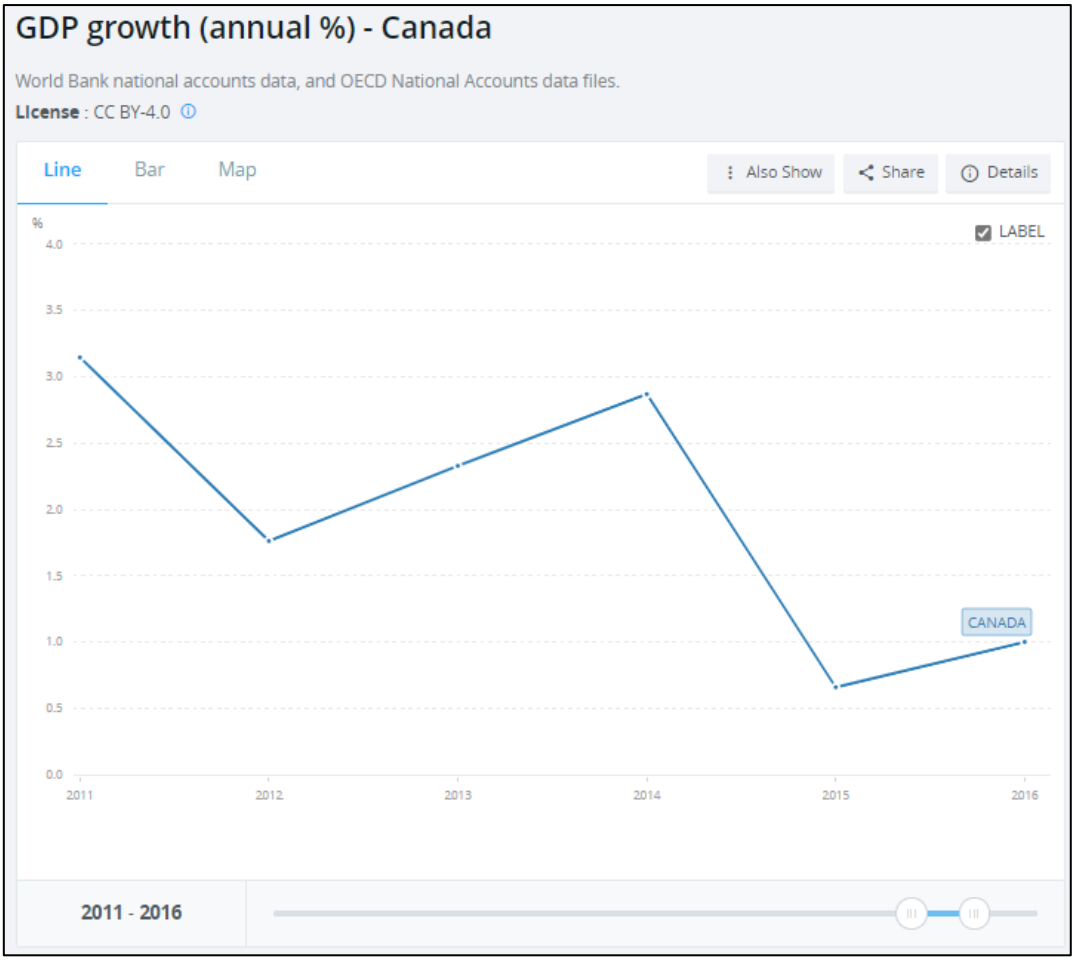


Figure 19: Annual GDP of Canada from 2011 to 2016
Source: (The World Bank Group, 2021)

Weak growth or negative GDP signals that the economy of that country is doing poorly. This often brings about job cuts when the economy is in recession where economic activities are reduced. During a recession, organizations face increased costs, a decrease in revenue, and increased debts. To pull through this period, organizations will start to lay off workers to cut costs as employees consume most of the organizations' budget distribution. Terminating employees provides them a quick solution to solve their financial crisis (Investopedia, 2021).

As discussed previously, terminating employees is not the most ideal solution where other options are available to minimize cost such as transportation, infrastructure, and various other overhead variables. However, a costing factor that can be considered is the number of new hires. If new employees can be hired while layoffs are still apparent, it shows an irrational decision made by Human Resources.

5.1.3 Analysis 3 – Over Hiring Affecting Cost

Is over hiring contributing to the termination rate?

```
132  # Distribution of hired employees
133  unique_employee_attrition %>%
134    group_by(Hired_Year) %>%
135    summarise(n = n(), ) %>%
136    ggplot(aes(x=Hired_Year, y=n, group=1)) +
137      geom_line(color="red") +
138      geom_point() +
139      labs(title = "Employees Hired per Year"
140           , x = "Year", y = "No. Of Employees")
```

Figure 20: Code to Plot Line Graph of No. of Employees being Hired Distributed by Year

The aim for this visualization is to view the trend of the hired employees by year. The plot should show the size of the company through the amount of workforce they have. This can show the hiring trend of human resources and when they stopped hiring.

Using the ggplot2 package, the data set that contains all the active and inactive employees is used in this plot. The data set was then grouped by the year hired which produced the number of employees that are hired per year. This information was then plotted where the Years are on the x-axis and number of employees on the y-axis. A line graph was utilized as this plot helps visualize a timeline of activities where the upward or downward trend can be seen.

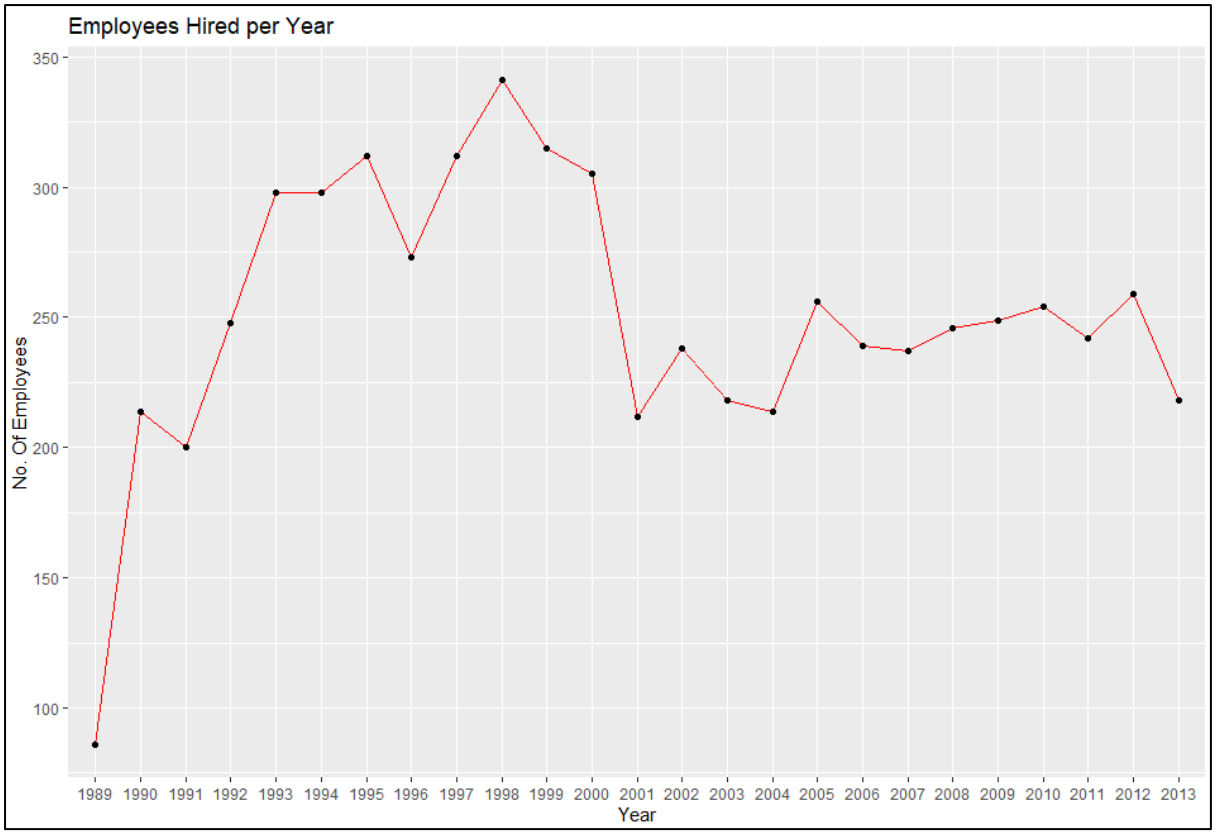


Figure 21: Line Graph for Distribution of Employees Hired per Year

From this graph we can identify that the organization has an even distribution of hiring employees throughout 2001 onwards. This organization also aggressively hired during 1989 to 1993 to upscale their workforce and showed a slight dip in year 2000. However, the organization stopped hiring after 2013. This shows that during the years of recession in the company where layoffs started, they didn't continue hiring.

Therefore, it is seen that over hiring was not a factor that contributed to the cost cutting of this organization. It would have been an illogical move to continue hiring when the organization is laying off employees, as it is not a cost-effective strategy where it would cost more to conduct the entire interviewing and onboarding process.

5.1.4 Conclusion – Question 1

It is found that there was a decrease in the GDP of Canada in the year 2014 to 2015 which caused the organization to face a recession period. This period forced the organization to downsize their company workforce during these years by laying off some employees. It is seen that after 2013, the company also stopped hiring any more employees during the recession period which concludes that over hiring is not one of the factors that affect to the high layoff rate. Therefore, the main conclusion that can be drawn from this question is that the negative economic growth of Canada was the main factor that influenced the layoff rate during 2014 to 2015.

This organization is seen to make the major financial decisions during December. Instead of choosing any other month in the year, they only laid off employees during December which can lead to a conclusion that December is the end of an accounting term for this organization. Human resources require a further contingency plan to bounce back from this recession period on specific sectors that will be discussed in further questions. However, based off the organization it may not suite them to lay off employees during December especially during the holiday period where also where new projects may start for the following year. A setback in employees might affect business relations for new projects in the future during the month of January.

5.2 Question 2

Is there any discrimination when terminating employees?

5.2.1 Analysis 1 – Discrimination by Gender

Are the employees being terminated because of their gender?

```
155  # Pie Chart for Termination by Gender
156  terminated_employees %>%
157    group_by(Gender_Full, Termination_Reason) %>%
158    summarise(n = n(), ) %>%
159    ggplot(aes(x=0, y=n, fill=Gender_Full)) +
160      geom_col(position="fill") +
161      facet_wrap(~Termination_Reason) +
162      coord_polar(theta ="y", start=0) +
163      theme_void() +
164      labs(title = "Gender Distribution of Terminated Employees") +
165      guides(fill = guide_legend(title = "Gender"))
166
167
```

Figure 22: Code to Plot Pie Charts of Terminated Employees Sorted by Gender

The aim for this visualization is to view the distribution of the terminated employees by gender. This plot would evidently show if there was any discrimination if there is an overwhelming majority in the pie chart.

Using the ggplot2 package, the dataset that contains terminated employees is used in this plot. Each pie chart is distributed with the values of gender, and then split into 3 different pie charts for the termination reason. The pie chart was used here as this graph is most suitable to represent two to three categorical values as it would show the distribution amongst those categories.

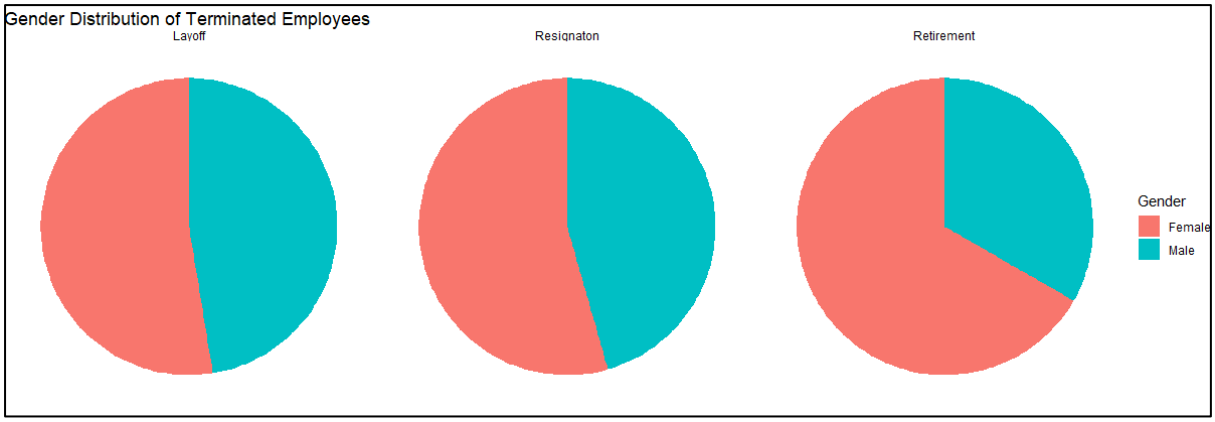


Figure 23: Bar Graph of Terminated Employees Sorted by Gender

The layoffs have a similar spread where there is no major difference in the range of values between the two genders. This shows that Human Resources is NOT gender bias when making their decisions in laying off employees.

This analysis was done as it is a known issue where some organizations execute wrongful terminations based on the employee's gender. Examples of these discriminatory cases include some termination due to pregnancy, maternity leave, or even the insurance burden on fertility treatments. It is good to see that this is not one of the problems faced by this organization as it is commonly illegal to have this misconduct in a workplace (Weisberg Cummings P.C., 2021).

What can be recommended to human resources is to promote their employees to engage human resources as a friend if they ever feel discriminated based on their identity. Usually, sensitive matters such as these tend to be ignored and suffered through while it affects mental health (Equal Rights Advocates, 2019).

5.2.2 Analysis 2 – Discrimination by Age

Are the employees being terminated because of their age?

```
154 ggplot(terminated_employees,aes(x=Age_Group, fill=Termination_Reason)) +
155   geom_bar() +
156   labs(title = "Distribution of Termination Reason in Terminated Employees by Gender"
157       , x = "Age Group", y = "No. Of Employees")
158
```

Figure 24: Code to Plot Bar Graph of Terminated Employees by Terminated Month Sorted by Age Group

The aim for this visualization is to view the distribution of the terminated employees by age group. This plot should show if there were any discrimination if there is an overwhelming majority of layoffs in any age groups.

Using the ggplot2 package, the dataset that contains only terminated employees is used in this plot. A stacked bar chart was plotted where the x axis is populated with the Age Group values while stacking them with the Termination Reason values. A stacked bar chart was used here as it better represents and separates two categorical values in one visualization.

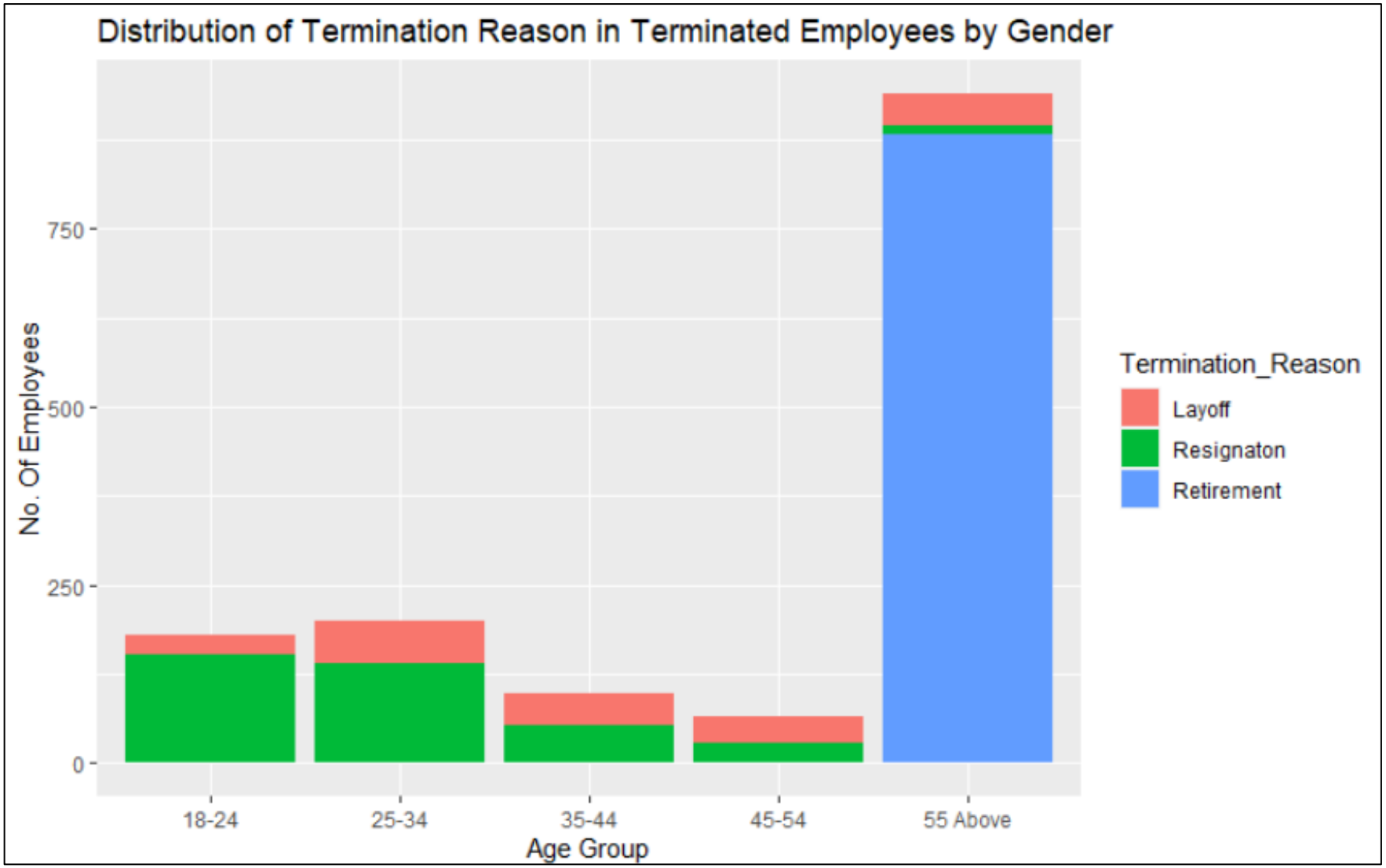


Figure 25: Bar Graph of Terminated Employees Sorted by Age Groups

[1]

The graph has a similar spread for layoffs. This shows that Human Resources is NOT age bias when making their decisions in layoffs.

Retirement is seen to only take place in the older age group which is expected where older employees are more likely to retire even before reaching the retirement age. The inconsistency found from this analysis is that younger employees tend to resign more compared to the older employees. One can argue that younger employees tend to change jobs often to continuously seek different experiences and cultures. This assessment argues that employers simply need to accept high turnover rates among young employees (Fallon, 2021).

Another reason for higher terminations in the younger age group that has been surveyed from the general youth is that there wasn't a healthy work life balance in the organization. The high resignation numbers between youths can be explained from them feeling overworked and underappreciated at work (YPulse, 2021).

Human Resources of this organization must investigate this as a possible issue that contributes to the high attrition rate of the employees. By taking pre-emptive steps, human resources can better handle this type of issue if it ever were to arise in the future. These steps may promote a higher retention rate among young employees. This is because young employees tend to prioritize working culture rather than the outcome such as salaries and bonuses.

5.2.3 Conclusion – Question 2

The human resources of this organization can be seen to not have any discriminatory influence when laying off employees. However, it is still beneficial for human resources to promote an open culture if they feel discriminated at any given time. This promotes a better bond and safe environment in the organization which has a chain effect in retaining their employees.

5.3 Question 3

Which location has the most terminated employees?

5.3.1 Analysis 1 – Geographical Distribution

What is the geographical spread, and does it show any trend?

The aim of this visualization is to locate the geographical distribution of this organization and where is their concentration of employees. The distribution should show if the employees are concentrated in a certain area of the country and if they have small branches in rural areas surrounded by rivers and mountains.

```
185  #Analysis1
186  # Geographical distribution of Cities
187  library(sf)
188  library(raster)
189  library(tmaptools)
190  library(leaflet)
191  library(leaflegend)
192
193
194  cityname =
195    unique_employee_attrition %>%
196    mutate(City = paste(City,", Canada", sep="")) %>%
197    group_by(City) %>%
198    rename(query=City) %>%
199    summarise(count = n(), ) %>%
200    arrange(desc(count))
```

Figure 26: Libraries and data frame to be used for this analysis

The data frame prepared is grouping the number of employees by their cities and counting how many employees are there in each city.

```
202  information = geocode_OSM(cityname$query,as.sf = TRUE)
203  information = left_join(information, cityname, by="query")
204
205  symbols = makeSizeIcons(
206            values = information$count,
207            shape = 'circle',
208            color = 'purple',
209            fillColor = 'purple',
210            opacity = .5,
211            baseSize = 10
212          )
213
214  leaflet(information) %>%
215    addTiles() %>%
216    addMarkers( icon = symbols,
217                popup = information$query) %>%
218    addLegendSize(
219      values = information$count,
220      color ='purple',
221      opacity = 0.5,
222      shape = 'circle',
223      orientation = 'horizontal',
224      breaks =5,
225    )
```

Figure 27: Code to Plot Geographical Map with Sized Icons based on Employee Count

The code above is firstly getting the open street map values which contain the coordinates of locations. The values are specified to only take from the cities that are available in the employee dataset which in this case are cities in Canada. This data frame is then joined with the previous data frame to obtain the number of employees in each city. Then different sized icons are made based on that frequency count to show the concentration of employees in each city. The geographical map is then plotted with the coordinates and the sized icons are placed on top of that map to show the concentration. A legend is also provided to show the concentration number based on the icon size.

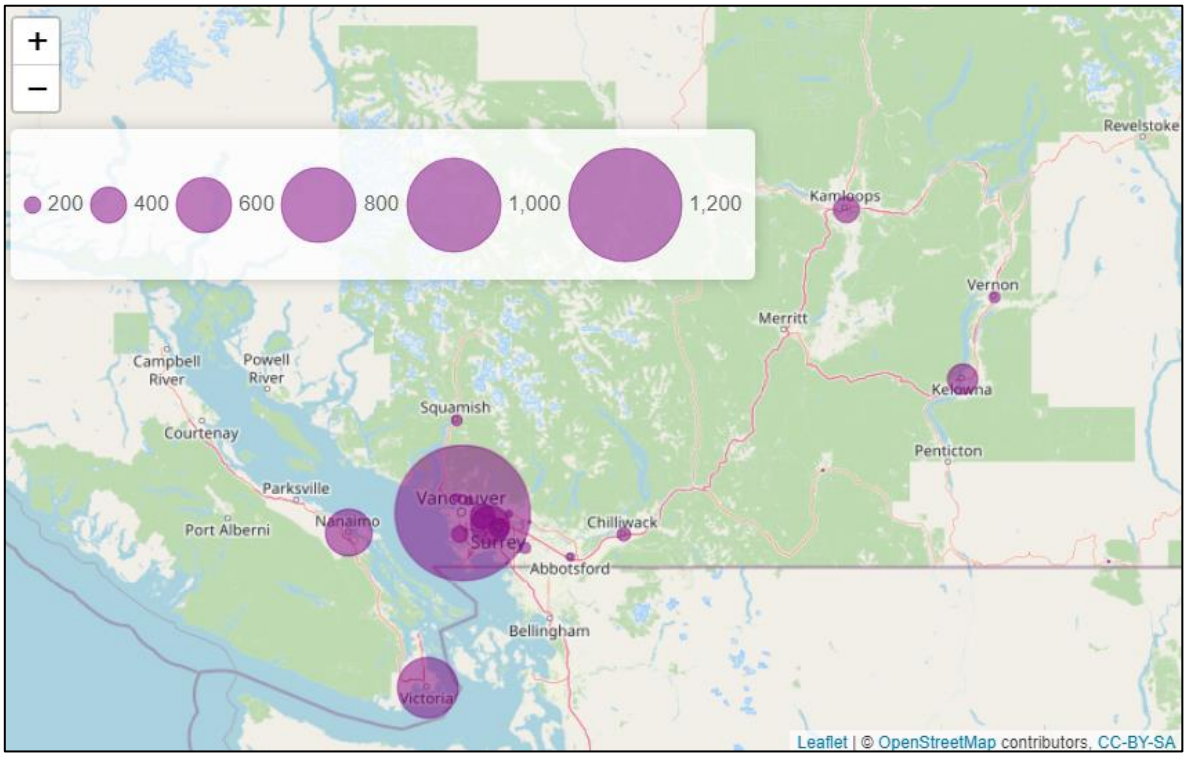


Figure 28: Geographical Spread of the Concentration of Employees in Their Cities

From this graph, it is seen that the circles get smaller as it gets further away from the main developed city areas. Take Fort Nelson for example, it is located rather north of the concentrated cities such as Vancouver. Referring to figure 30, the wide view proves that the concentration of employees is located around Vancouver City.

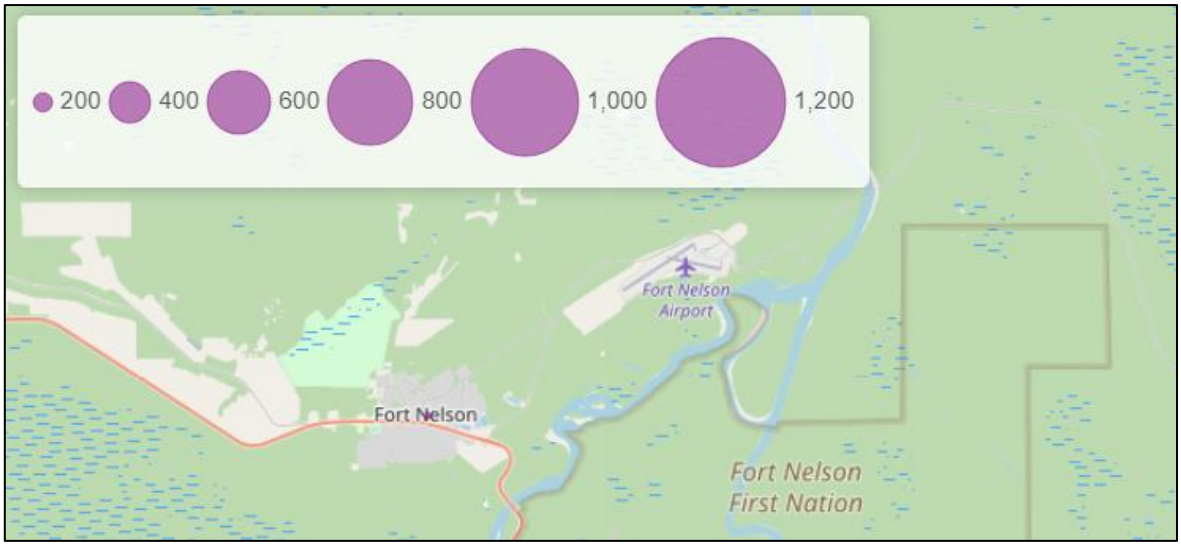


Figure 29: Concentration of Employees at Fort Nelson

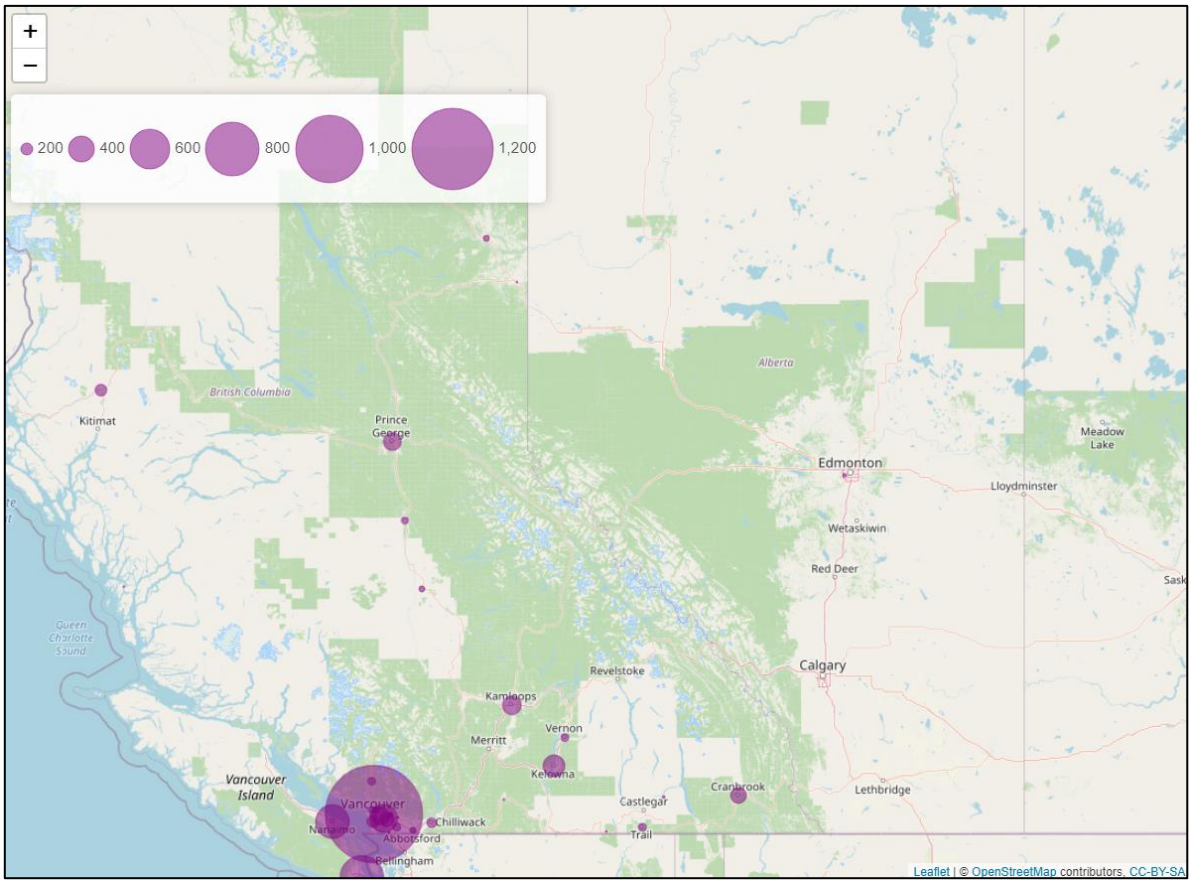


Figure 30: Zoomed Out View of Canada

5.3.2 Analysis 2 – Termination by City

What is the relationship between terminated employees and their city?

```
ggplot(unique_employee_attrition, aes(y = fct_infreq(City), fill=Termination_Type)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Distribution of Termination Type by City"
       , x = "No. Of Employees.", y = "City")

```

Figure 31: Code to Plot Vertical Bar Graph for Distribution of Employees by City

The aim for this visualization is to view the distribution of all the employees by their city. The base of operations for this organization and where is their main workforce located in should be able to be identified through this plot. This would also show which store that has shut down due to the recession period.

Using the ggplot2 package, the dataset that contains both active and terminated employees is used in this plot. A vertical bar chart was plotted where the cities are plotted on the y axis while the count is on the x axis. A vertical bar chart was used here due to the large number of categories in this column where the labels for each category would look cleaner when plotted on the y axis instead of the x axis.

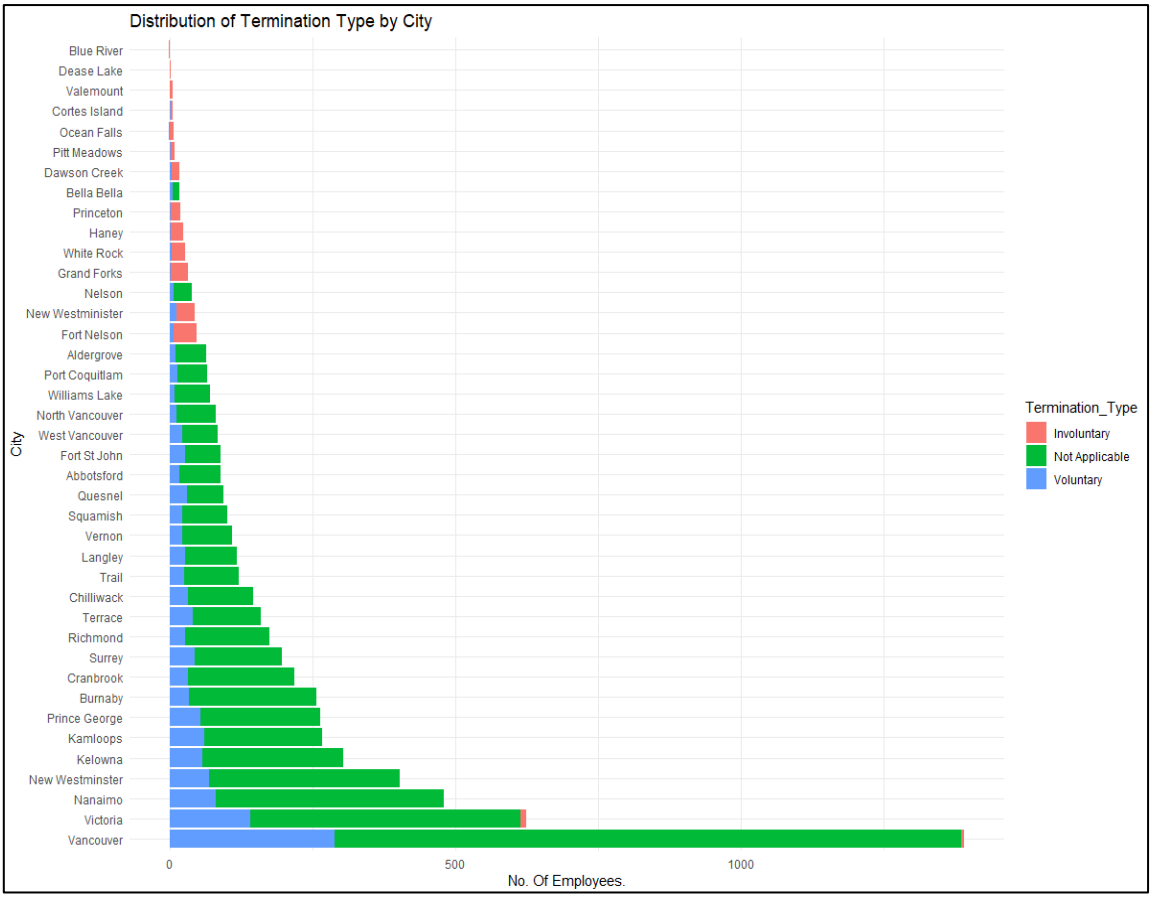


Figure 32: Vertical Bar Graph for Distribution of Employees by City

The organization is based heavily in Vancouver where majority of their workforce is located and centralized around that city. After further research, Vancouver is found to be one of the most populated cities in Canada at that time in 2016, being ranked 8th in the city population ranking in Canada. However, the layoffs are seen to only take place in the less populated cities such as New Westminster which is ranked 81st in the city population list of Canada in 2016 (todoCanada, 2017).

The layoffs of the employees seem to take place only at locations which has a relatively smaller workforce. Human resources have decided to terminate employees in these rural areas where the population is lower. This puts the employees in a tougher position during a recession period where job opportunities are already low in lower concentrated cities.

Layoff rates differ across geographical locations to a certain extent covering factors including transportation and availability of consumers. It is evident that an increase in population would lead to an increase in the probability of a better consumer market. To fortify the organizations stability, it is better to focus on the main branches of operation by downsizing and reallocating resources from the smaller branches (Ci, Morissette, & Schellenberg, 2016).

What human resources could have done to lessen the impact of unemployment was to offer a merger program with other branches in other cities. Resignations and retirements of a certain branch would open positions that are available to other employees to fill. These positions can then be filled with employees from other branches that have been planned to shut down. Having something like a transfer programme would lighten the overall cost of hiring and the livelihood of the current employees.

5.3.3 Analysis 3 – Termination by Store

What is the relationship between terminated employees and the store location?

From the previous analysis done on termination by city, this organization seems to cease their base of operations in smaller branches where there are fewer employees. To further justify this hypothesis, a further analysis needs to be done on the terminations by store. This would provide some justification on why they shut down smaller branches in rural areas.

```
# Termination Type by Store Code
unique_employee_attrition %>%
  mutate(Store_Code = paste("A",Store_Code , sep="")) %>%
  group_by(Store_Code, Termination_Type) %>%
  summarise(count = n(), ) %>%
  ggplot(aes(y =reorder(Store_Code, -count), x=count, fill=Termination_Type)) +
    geom_col() +
    theme_minimal() +
    labs(title = "Distribution of Termination Type by Store Code"
         , x = "No. Of Employees.", y = "Store Code")

```

Figure 33: Code to Plot Vertical Bar Graph for Distribution of Employees by Store

The aim for this visualization is to view the distribution of all the employees by their store. The plot should then further show the active stores for this organization, and which store no longer has any employees. This would then match then match the theory of termination of employees in rural areas found in the previous analysis.

Using the ggplot2 package, the dataset that contains both active and terminated employees is used in this plot. A vertical bar chart was plotted where the y axis has the store code values while the count is on the x axis. A vertical bar chart was used here due to the large number of categories in this column where the labels for each category would look cleaner when plotted on the y axis instead of the x axis.

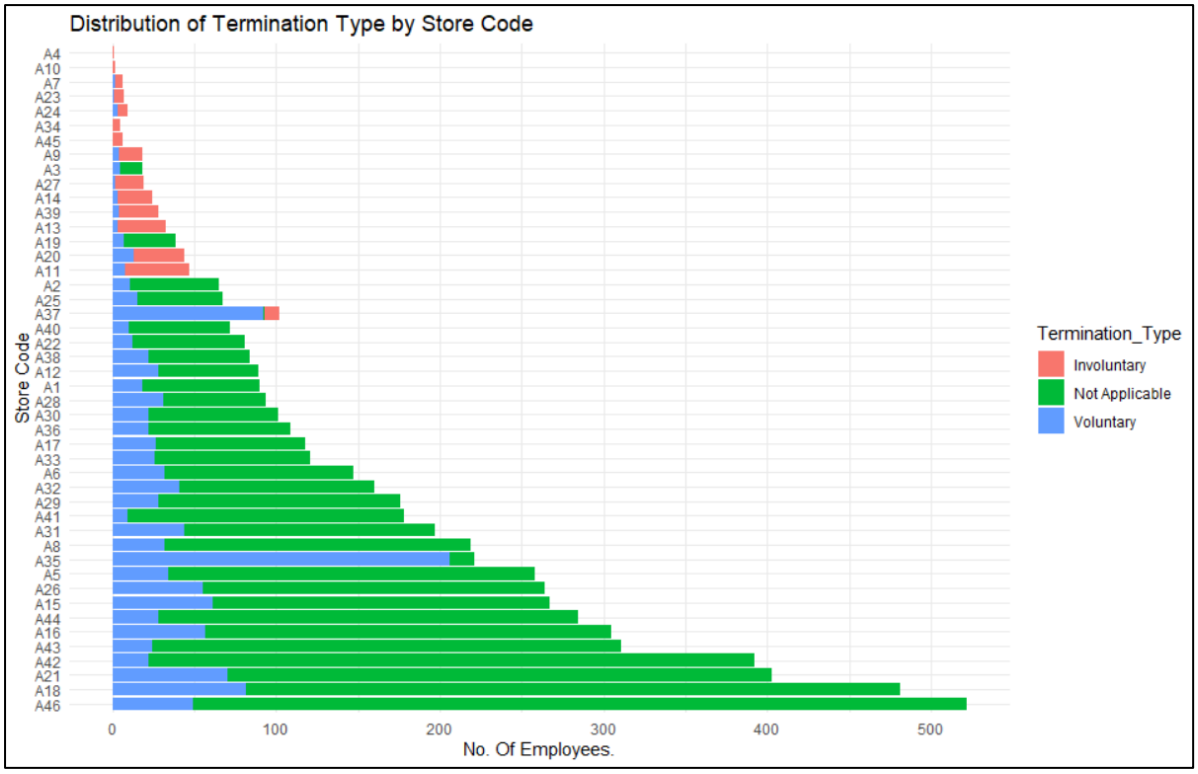


Figure 34: Vertical Bar Graph for Distribution of Employees by Store

From the graph above, we can identify stores with lower number of employees have a trend where the employees get laid off, and the store is being shut down. These stores have more involuntary terminations compared to others which justifies why human resources terminated the employees in those store branches to cease operations in that store. The reasonable explanation is in accordance with the theory that the organization is downsizing as a strategic financial move of survival during the recession period.

To reinforce the statement that small branches with low number of employees are in rural areas, the location of each store code will be extracted to check for this theory. The store will be matched to its city to see if the hypothesis holds true.

```
203 store_city =
204 unique_employee_attrition %>%
205   mutate(Store_Code = paste("A",Store_Code , sep="")) %>%
206   group_by(Store_Code, City) %>%
207   summarise(count = n(), )
208
```

Figure 35: Code to Check Which City the Store is In

The source code is taking the data set that contains all the employees and grouping the store code with the city. This code then allows us to see which store code is in which city.

	Store_Code	City	count
33	A19	Nelson	39
34	A13	Grand Forks	33
35	A39	White Rock	28
36	A14	Haney	24
37	A27	Princeton	19
38	A3	Bella Bella	18
39	A9	Dawson Creek	18
40	A24	Pitt Meadows	9

Figure 36: Stores That Belong in the City of White Rock & Haney

After referring to the graph in figure 27, the termination rates that are high in stores with lesser employees indeed hold true to the fact that they are located at rural areas. From this we can conclude that this organization closed their smaller branches as a strategy approach to cut costs where these smaller stores may not be generating enough revenue to keep it running.

5.3.4 Conclusion – Question 3

The location that has the most terminated employees is seen to be the stores that are located at rural areas where there is a lower population and is further away from the main cities. This justifies human resources decision of closing the stores that generate lesser cashflow due to those cities having lower number of residents. To summarise, the lesser the population, the lesser the revenue generated. This has led to the decision where it is better to refocus the resources onto the main store branches. However, these closed stores were originally opened in these specific areas as they were seen as a good market opportunity. Human Resources should re-look into the closed stores and their locations after the recession period as a recovery plan for their organization.

5.4 Question 4

Which department has the most terminated employees?

5.4.1 Analysis 1 – Termination by Business Unit

What business unit has the most terminations?

```
# Termination By Business Unit
ggplot(unique_employee_attrition, aes(x = Business_Unit, fill=Termination_Type)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Distribution of Termination Type by Business Unit"
       , y = "No. Of Employees.", x = "Business_Unit") +
  geom_text(
    stat = "count",
    aes(label = ..count..),
    position = position_stack(vjust = 0.5))

```

Figure 37: Code to Plot Bar Graph for Termination Distribution by Business Unit

The aim for this visualization is to view the distribution of the terminated employees by their business unit. This plot would evidently show the decision making of human resources when terminating employees on which general business unit they cut cost on.

Using the ggplot2 package, the dataset that has all the employees is used in this plot. A stacked bar chart was plotted where the x axis is populated with the Business Unit values while stacking them with the Termination Reason values. A stacked bar chart was used here as it better represents and separates two categorical values in one visualization.

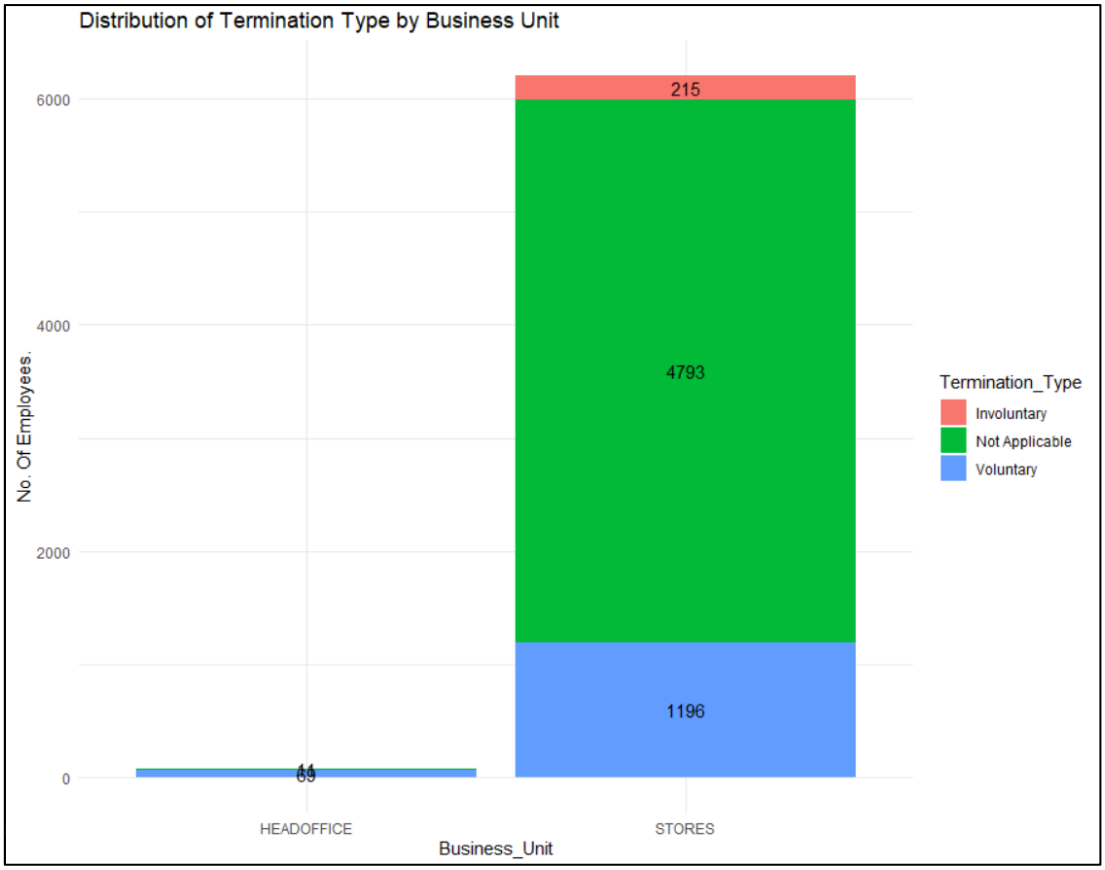


Figure 38: Bar Graph for Termination Distribution by Business Unit

From the graph, the main workforce for this organization was identified to be in the stores while there is a very low number of employees in the head office. The ratio of employees in the head office compared to all the stores could be an underlying issue that may arise in the future. This is whereby the head office does not have enough manpower to manage all their store branches efficiently and effectively due to the large ratio difference.

It can be seen only employees in the stores are getting laid off while everyone in the head office retained their jobs. Human Resources needs to review laying-off head office employees as usually they have the highest salary comparatively to the employees in stores, but they may not be the most useful. However, it can also be said that this organization had prioritized some criteria that justified why they only laid off store employees. This could be because store employees tend to be short-term contract workers compared to a long-term worker at the head office. These criteria might also account for the long-term potential these said employees might bring (SHRM, 2021).

5.4.2 Analysis 2 – Terminations by Department

What department has the most terminations?

```
277  # Termination by Department
278  ggplot(terminated_employees, aes(y = fct_infreq(Department), fill=Termination_Reason)) +
279    geom_bar() +
280    theme_minimal() +
281    labs(title = "Distribution of Termination Type by Department"
282         , x = "No. Of Employees.", y = "Department")
283
```

Figure 39: Code to Plot Vertical Bar Graph for Terminated Employees by Department

The aim for this visualization is to view the distribution of the terminated employees by their departments. This plot should show what departments does human resources need to tentatively focus more on due to their higher attrition rate.

Using the ggplot2 package, the dataset that only contains the terminated employees is used in this plot. A vertical bar chart was plotted where the y axis are the department categories while the count is on the x axis. A vertical bar chart was used here due to the large number of categories in this column where the labels for each category would look cleaner when plotted on the y axis instead of the x axis.

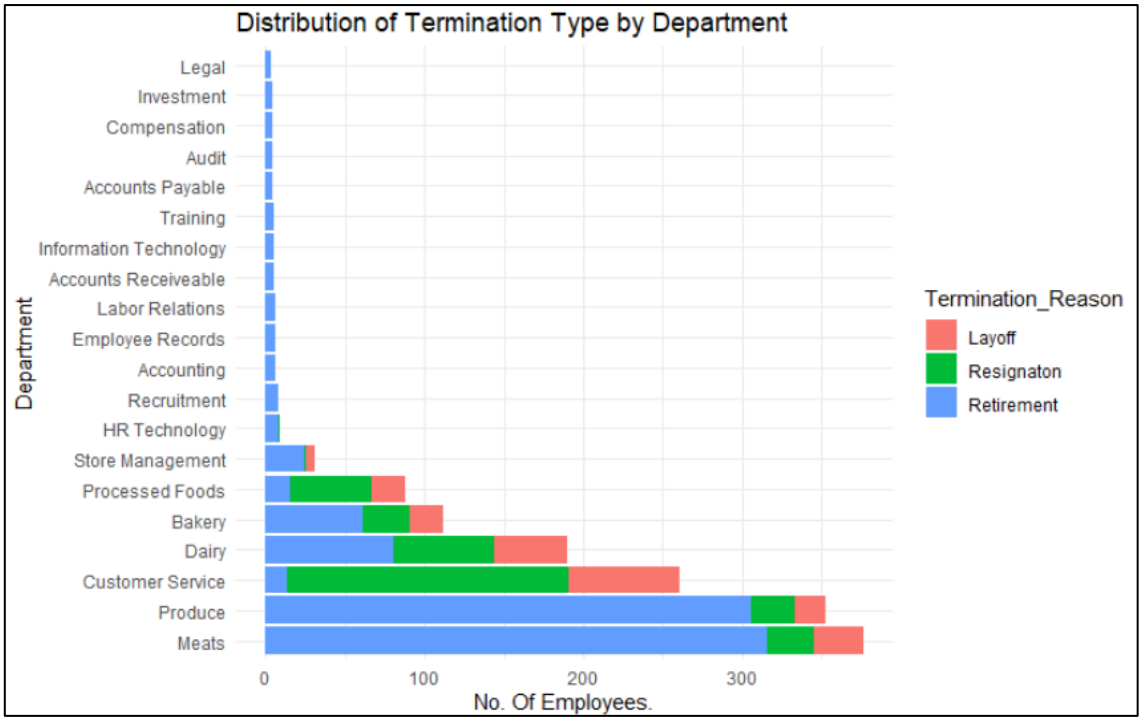


Figure 40: Vertical Bar Graph for Terminated Employees by Department

The layoff trend is seen to occur only where there is a significant number of employees in that department already. Therefore, it can be said that human resources were just downsizing the departments to cut costs. However, there is a comparatively higher number of resignations in the customer service department compared to the other departments that have high number of employees. Not only that, layoffs in the customer service department also are slightly higher than other departments. On the other hand, the amount of retires in the produce and meats are seen to be very high.

Departments with higher resignations such as in customer service can usually be explained by the job nature of that department. If the position is as a part time worker or short-term contracts, a trend of high turnovers would be evident. This is because the position in this department at the store front would usually be filled by a short-term worker for a few months before getting replaced. As short-term workers are less valuable compared to contracted employees, layoffs would be seen to be higher in those departments that have a larger proportion of shorter-term positions. This justification would be further analysed by comparing the age of the terminated employees in each department which would also justify the number of high retirements in produce and meats.

5.4.3 Analysis 3 – Departments Age Group

What are the age groups of the terminated employees in each department?

```
# Terminated Department Age
ggplot(terminated_employees, aes(y = fct_infreq(Department), fill=Age_Group)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Distribution of Age Group of Terminated Employees by Department"
       , x = "No. Of Employees.", y = "Department")

```

Figure 41: Code to Plot Vertical Bar Graph for Terminated Employees Age by Department

The aim for this visualization is to view the distribution of the age group of the departments of terminated employees. This plot should show the demographic of workers in each department and what departments tend to be popular in each age group.

Using the ggplot2 package, the dataset that only contains the terminated employees is used in this plot. A vertical bar chart was plotted where the y axis are the department categories while the count is on the x axis. A vertical bar chart was used here due to the large number of categories in this column where the labels for each category would look cleaner when plotted on the y axis instead of the x axis.

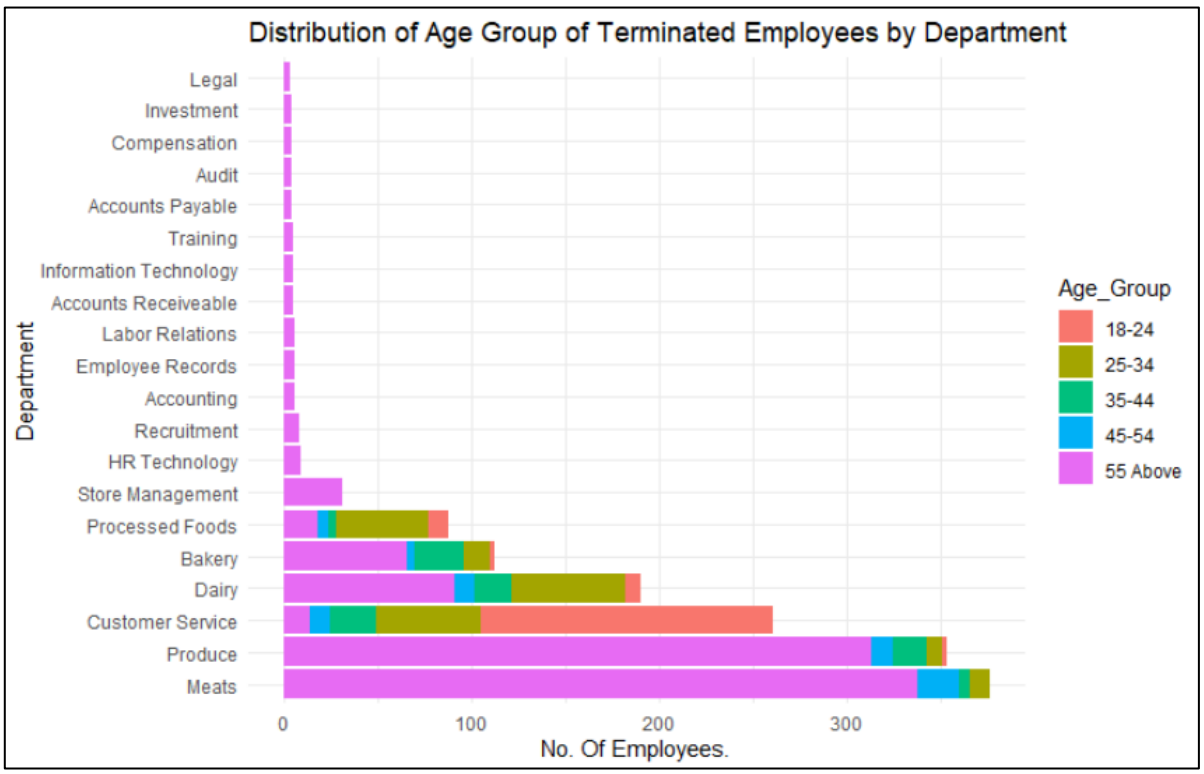


Figure 42: Vertical Bar Graph of Age Group of Terminated Employees by Department

From the above graph, it is justified from the previous analysis that the customer service department is shown to be filled with relatively younger employees which justifies the theory that the younger people tend to pick-up part-time or short-term contract jobs to fill in their gap schedule between schooling terms. It is shown that teenagers tend to pick up a part time job as a beneficial way to fill their time by learning money management skills, building self-confidence, as well as learning basic work skills and work ethics (The Center for Work Ethic Development, 2018).

The explanation for the higher aged employees working in produce and meats can only be purely theorized on the fact that younger people tend to shy away from menial and laborious jobs such as working in a manufacturing line or in a factory. Whereas older people tend to have more of a survival instinct where they do not mind the nature of the job if they have secured a position and a stable source of income. However, this theory has no concrete backing due to the lack of data where human resources may need to look at the reason for the high number of older employees in these two departments.

What human resources can do for the customer service department is to have more permanent positions rather than holding majority of part timers or short-term positions. This would increase retention rate where cost could be reduced at the hiring process. Not only that, but long-term employees enable them to grow into the position specially to suit the company's culture.

5.4.4 Analysis 4 – Active Departments Remaining

What are the active departments remaining in the organization?

```
# Active Departments Remaining
ggplot(active_employees, aes(y = fct_infreq(Department), fill=Gender_Full)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Distribution of Active Employees by Department"
       , x = "No. Of Employees.", y = "Department")
```

Figure 43: Code to Plot Vertical Bar Graph for Active Employees by Department

The aim for this visualization is to view the distribution of the departments of the remaining employees. The plot should show the distribution of employees by the remaining active departments which would show what departments got shut down. This plot should show what departments the organization decided to close and why.

Using the ggplot2 package, the dataset that only contains the active employees is used in this plot. A vertical bar chart was plotted where the y axis are the department categories while the count is on the x axis. A vertical bar chart was used here due to the larger number of categories in this column where the labels for each category would look cleaner when plotted on the y axis instead of the x axis.

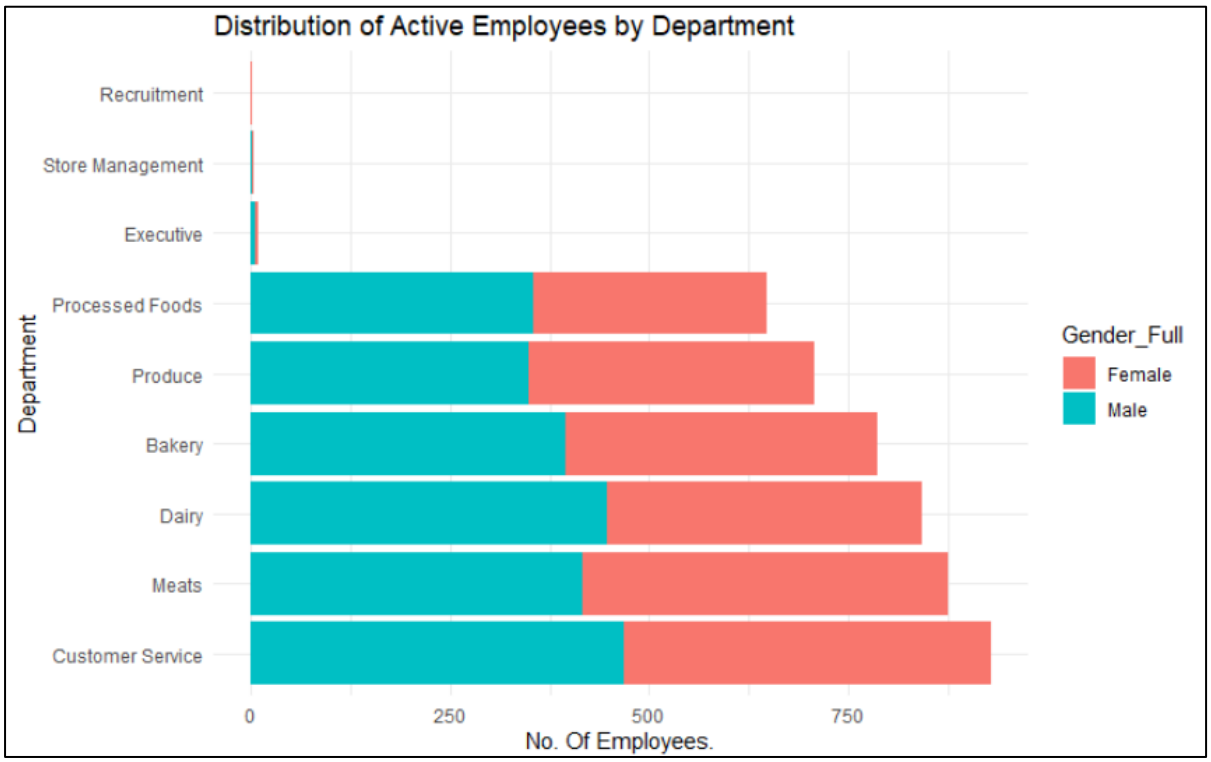


Figure 44: Vertical Bar Graph for Active Employees by Department

A total of 11 departments have been closed since the termination of employees. The organization is assumed to have started to practise outsourcing which is dedicating a certain department to be handled by a third-party external vendor. Outsourcing is very cost effective and efficient as this organization used to only have a small number of employees in the now closed departments. It is effective as the external parties are specialised and experienced in the field compared to a new internal department which is more costly (Shethna, 2020).

The effects of outsourcing enabled this organization to mainly focus on their core services which seems to be manufacturing. This not only saves costing in terms of infrastructure, but they are able to cut cost on training and recruitment as well. This would justify their reasoning behind why they opted to outsource during a time of recession (Shethna, 2020).

Human resources would only need their own department and not resort to outsourcing if the organization requires the department very frequently and long term. Unique process flows and standards set apart an organizational department compared to others. If this organization wishes to have their own internal department specified to a sector, then they would need to upscale first before anything else. In other words, remaining with outsourcing would be beneficial for them before they are ready.

An important factor that needs to be addressed is that the Human Resources department is seen to be missing from the active employee's department. If this organization is outsourcing human resources to deal with their employees, it is highly advisable for them to get an internal department. The justification would be that they are sizeable enough where they have unique issues and cultures to their company that could be misinterpreted by an outsourced party. As the attrition rates are unique to this company, an internal team would be able to best analyse and make decisions if it comes to laying off employees, and how to reduce resignations.

5.4.5 Conclusion – Question 4

The involuntarily terminated employees are seen to only occur at the stores business unit compared to the head office. Layoffs are only seen to happen where the departments already have a larger number of employees compared than the other departments. Resignations are seen to be very high in customer service due to the job positionings in nature being short-term or contingent where more younger people are seen to take up these kinds of jobs. Retirements are also seen to be very high in Produce and Meats department where the correlation found is that older people tend to handle these jobs better compared to younger people.

This organization also has downsized their departments by terminating most of the smaller departments and generalizing the larger departments job scope. This is not advisable as specific departments have specific functions to perform. Downsizing this way as a short-term solution during a recession period is already risky, this organization must revert to more specialized departments as before as general departmental functions create confusion and inefficiency in the whole organization.

5.5 Question 5

Does the length of service affect the chances of getting laid off?

5.5.1 Analysis 1 – Termination by Length of Service

Are people with shorter length of service more likely to get laid off?

```
323  involuntary_employees = subset(unique_employee_attrition,
324                                 Termination_Type == "Involuntary")
325
326  ggplot(involuntary_employees, aes(x = Service_Length)) +
327    geom_histogram(bins=7, col="red") +
328    geom_freqpoly(bins=7, color = "blue", size = 0.7) +
329    labs(title = "Distribution of length of service for laid off employees"
330         , x = "Length of Service (Years)", y = "No. of Employees")
331
```

Figure 45: Code to Prepare Laid Off Data Frame & Plotting Histogram for Laid Off Employees

The aim for this visualization is to view if the longer an employee works at the company, the less likely they are to get laid off. The plot should be able to see how many experienced workers they have laid off. This plot would evidently show the decision making of human resources when laying off employees if they value loyalty or other criteria.

Firstly, a new data frame was prepared for only the laid off employees. This means that the employees who were terminated involuntarily were separated into a new data frame. Using the ggplot2 package, the dataset that has only the laid off employees is used in this plot. A histogram and frequency polygon were plotted on top of each other using the service length as the value for the x axis. A histogram and frequency polygon were used as it better suits a continuous variable with its frequency. This enables the plot to show a better trend as these numbers are grouped together in bins which makes visualization easier to understand.

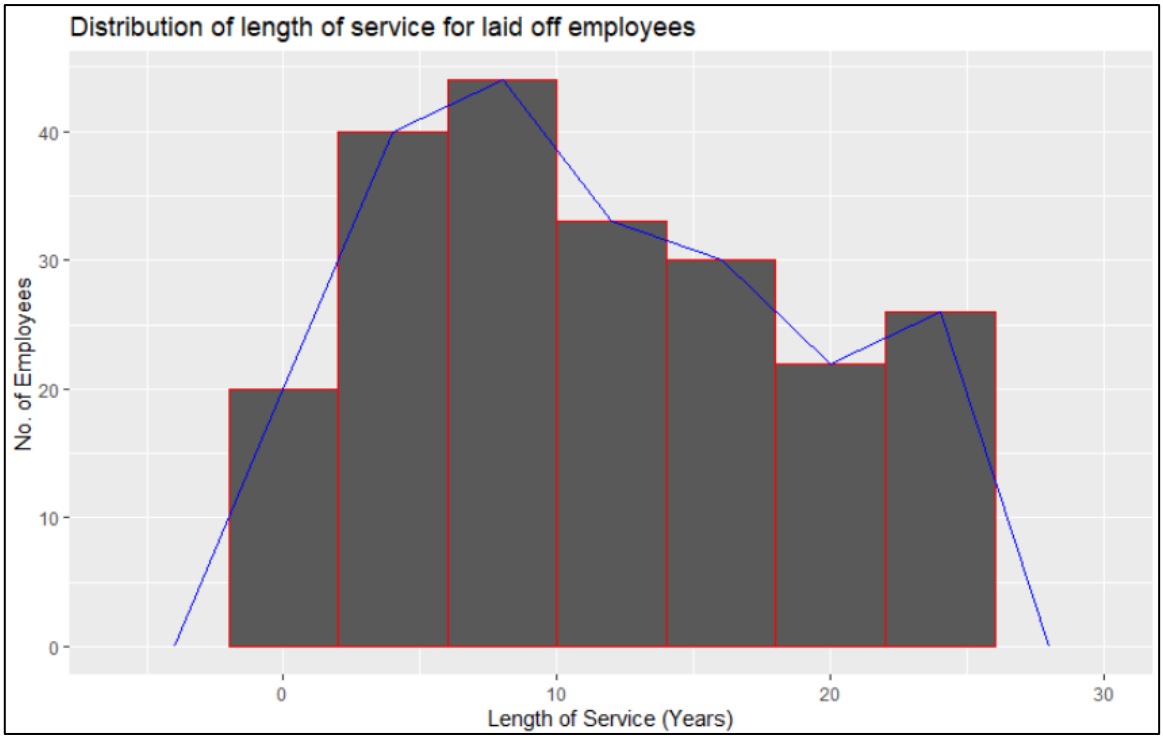


Figure 46: Histogram for Laid Off Employees based on Length of Service

From the histogram and frequency polygon above, it is seen that this organization does not practise seniority-based selection. This can be seen from the normal distribution where the terminated employees have their length of service peaked at slightly less than 10 years. Even senior employees have slightly higher number of layoffs compared to junior employees that justifies this organization has a different selection criterion when laying off employees. Based off previous analysis, this organization is most likely to practise employee status-based selection where contingent workers or short-term employees are least prioritized where they are most likely to be the target of termination. This is to ensure a greater job security for the core workers as the organization downsizes (SHRM, 2021).

Human resources should continue practising the multiple criteria ranking during the termination process as definitely its most suitable in this type of organization where there is a widespread of departments that range differently in terms of functionality. This would then prove the fairest evaluation amongst the employees especially even when the skillset of senior employees is outdated.

5.5.2 Analysis 2 – Length of Service in Departments

```
336  # Length of Service By Department in Terminated Employees
337  ggplotly(
338    ggplot(terminated_employees, aes(x = Service_Length, fill = Department)) +
339    geom_histogram(bins=7, color="black") +
340    labs(title = "Distribution of length of service for terminated employees by department"
341         , x = "Length of Service (Years)", y = "No. of Employees")
342    )
343
```

Figure 47: Code to Plot Stacked Histogram for Distribution of Length of Service for Terminated Employees by Department

Based off the previous analysis for the distribution of age in each department, the age was seen to be younger in the customer service department while employees were older in the meats and produce department. In this analysis, the aim is to find out the correlation between the length of service to draw another conclusion from the age group. This is to see if the older employees joined the organization at an earlier age and served several years before retiring or chose to retire without serving a long period.

Using the ggplot2 and plotly package, the dataset that contains only terminated employees is used in this plot. A stacked histogram is plotted with the service length as the value for the x axis stacked with the departments. A stacked histogram was plotted to easily visualize the countable difference between departments. The plotly package helps to interact and isolate individual departments for a clearer visualization on a specific value.

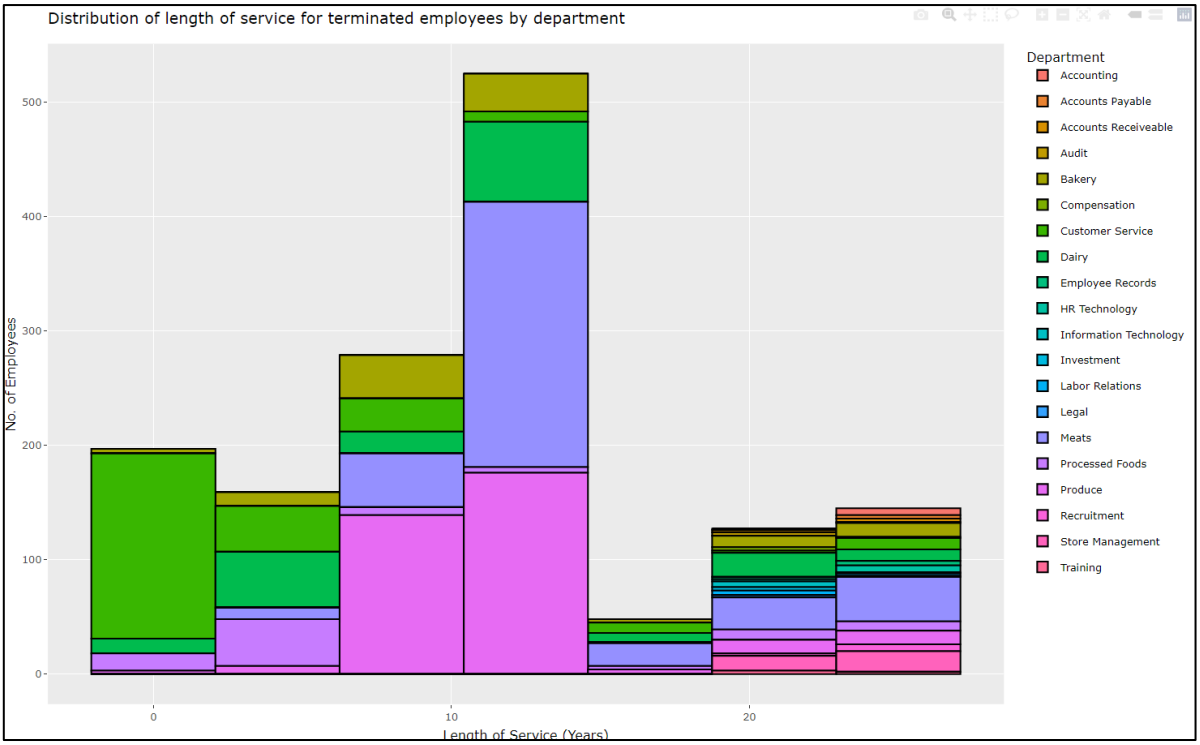


Figure 48: Stacked Histogram for Distribution of Length of Service for Terminated Employees by Department

From the graph above, it is hard to tell which departments are spread out due to the large number of departments. By utilizing plotly, there are three departments that are to be singled out and drawn a conclusion from which is customer service on its own, then meats and produce.

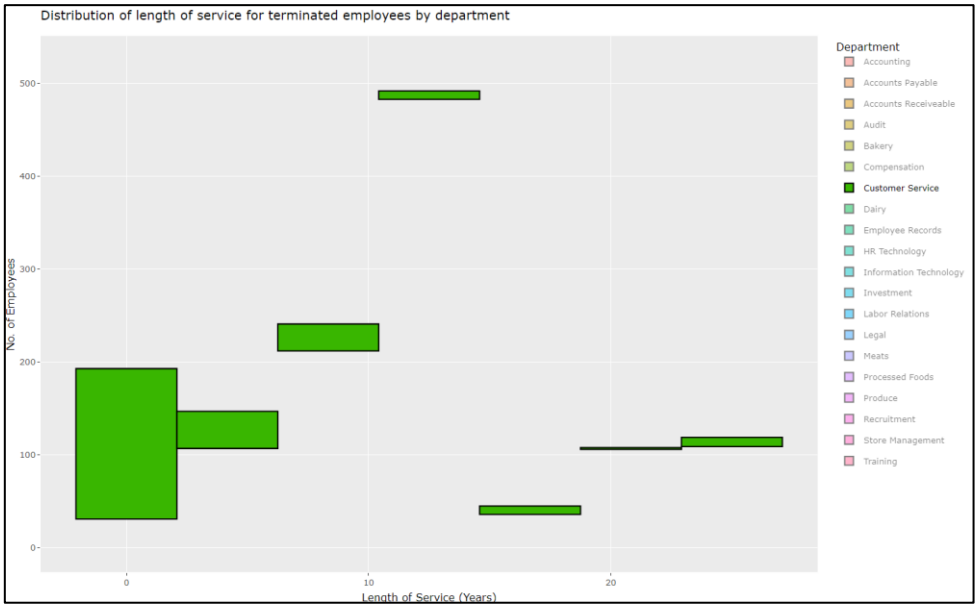


Figure 49: Distribution of Length of Service for Customer Service Department

From the distributed length of service, we can conclude that the customer service department is mostly made up of short-term employees that usually do not serve the organization for a long period of time. Human Resources should investigate longer term positions as a solution for this department and how they can retain more of the employees in this department. This includes investigating on any problems they face such as discrimination when interacting with customers, long working hours, or even strict schedules.

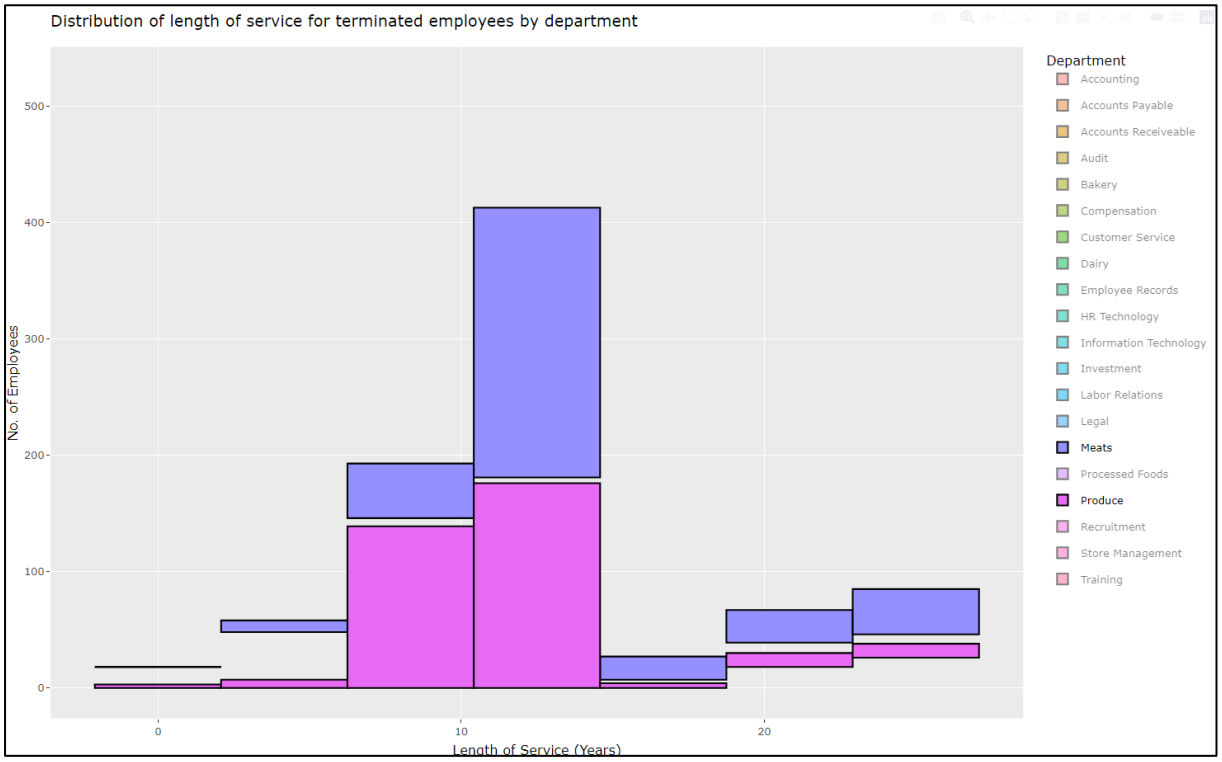


Figure 50: Distribution of Length of Service for Meats & Produce Department

For the meats and produce departments, it is generally seen that the past terminated employees usually have a length of service around 10 years. As analysed previously, majority of the age group in these department are 55 and above which also showed a many of them retired. The justification given is that older employees have been shown to stay in jobs longer as these employees tend to have a reason to come back to work such as having a family to feed or even job stability at an old age (Columbia University, 2021).

Human resources should broaden their demographic in terms of age when hiring for production lines such as the meat and produce departments. Older employees tend to retire around 10 years after working which in some sense shows that they have a “shelf-life”. Having younger employees would improve efficiency in terms of mechanical skills and would also give room to nurture them for future management positions. Upskilling them from the start would benefit the organization whereby they would know understand every level within the organization.

5.5.3 Conclusion – Question 5

It is concluded that this organization does not heavily focus on seniority-based criteria when laying off their employees. It is still unclear what criteria is selected during the termination process of this organization due to the lack of data, but it is vaguely clear that they have multiple criteria that they follow. Going forward, human resources should continue practising a multiple criterion selection during the termination process based on the relativity of the employee’s skillset and not only look at their loyalty or tenure. However so, it is still somewhat relevant to view their length of service especially if they are in crucial departments like top management or human resources as these roles require the most crucial understanding of the organizational problems.

Human Resources should also broaden their demographic age range for the customer service, meats, and the produce departments. Having a diverse demographic allows the department to complimentary enhance their employees. Older employees can share their experiences with the younger employees while younger employees are able to easily adapt to new technologies and are more mechanically proficient where they are able to benefit the department overall. However, human resources will have to address some if any disrespect occurs due to the age difference especially if the older employees feel that they are superior to the younger employees and therefore do not need to learn from them.

6.0 Extra Features

6.1 Extra Feature 1 – Writing Files

```
# Saving The Data Frames as new CSV files

write.csv(unique_employee_attrition,
          "D:\\Program Files\\R\\data\\NewData\\unique_employee_attrition.csv",
          row.names = FALSE)

write.csv(terminated_employees,
          "D:\\Program Files\\R\\data\\NewData\\terminated_employees.csv",
          row.names = FALSE)

write.csv(active_employees,
          "D:\\Program Files\\R\\data\\NewData\\active_employees.csv",
          row.names = FALSE)

write.csv(involuntary_employees,
          "D:\\Program Files\\R\\data\\NewData\\involuntary_employees.csv",
          row.names = FALSE)
```

Figure 51: Code to Write to new CSV files

The write.csv function is a built-in utility function in R. This function will write the contents of the data frame into a new csv file. The syntax is including the variable name, followed by the destination path. This extra feature was done because pre-processing has already been conducted on the original data set where the "cleaned" datasets can be utilized by future analysts. This would reduce some pre-processing time required by future analysts if they were to ever analyse this dataset.

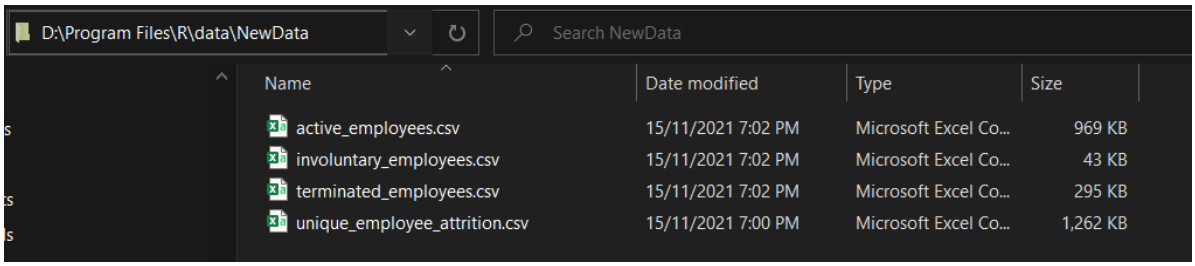


Figure 52: Newly Created CSV Files

6.2 Extra Feature 2 – Pie Chart

```
# Pie Chart

unique_employee_attrition %>%
  group_by(Gender_Full) %>%
  summarise(n = n(), ) %>%
  ggplot(aes(x=0, y=n, fill=Gender_Full)) +
  geom_col(position="fill") +
  coord_polar(theta ="y", start=0) +
  theme_void() +
  labs(title = "Gender Distribution of All Employees") +
  guides(fill = guide_legend(title = "Gender"))

```

Figure 53: Code to Plot Pie Chart for Gender

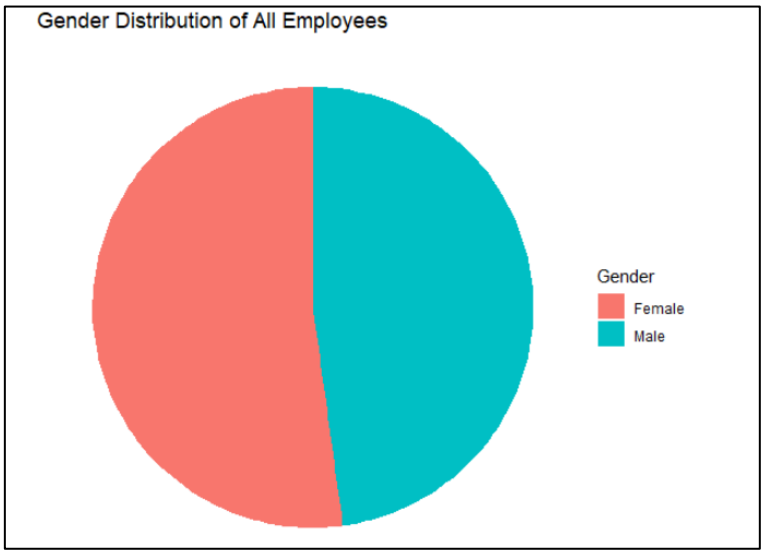


Figure 54: Pie Chart for Gender Distribution

The other extra technique that was covered in this documentation is the usage of pie chart as an extra visualization technique. A pie chart was used in a previous analysis at question 2 (refer to section 5.2.1) to visualize the distribution of the Gender column. The pie chart used is a mutation of the bar plot from the ggplot2 library where the polars were wrapped into a circle.

The value of the gender column only contains two unique values which are Male and Female. The pie chart aids to visualize a part-to-whole comparison where it is clearer on how much a slice would contribute to the overall picture. This would be hard to visualize with a bar chart as the comparison only differs in the proportions of height. Displaying less than five unique values would be more appealing and effective in a pie chart when displaying it as a whole distribution.

6.3 Extra Feature 3 – Geographical Map

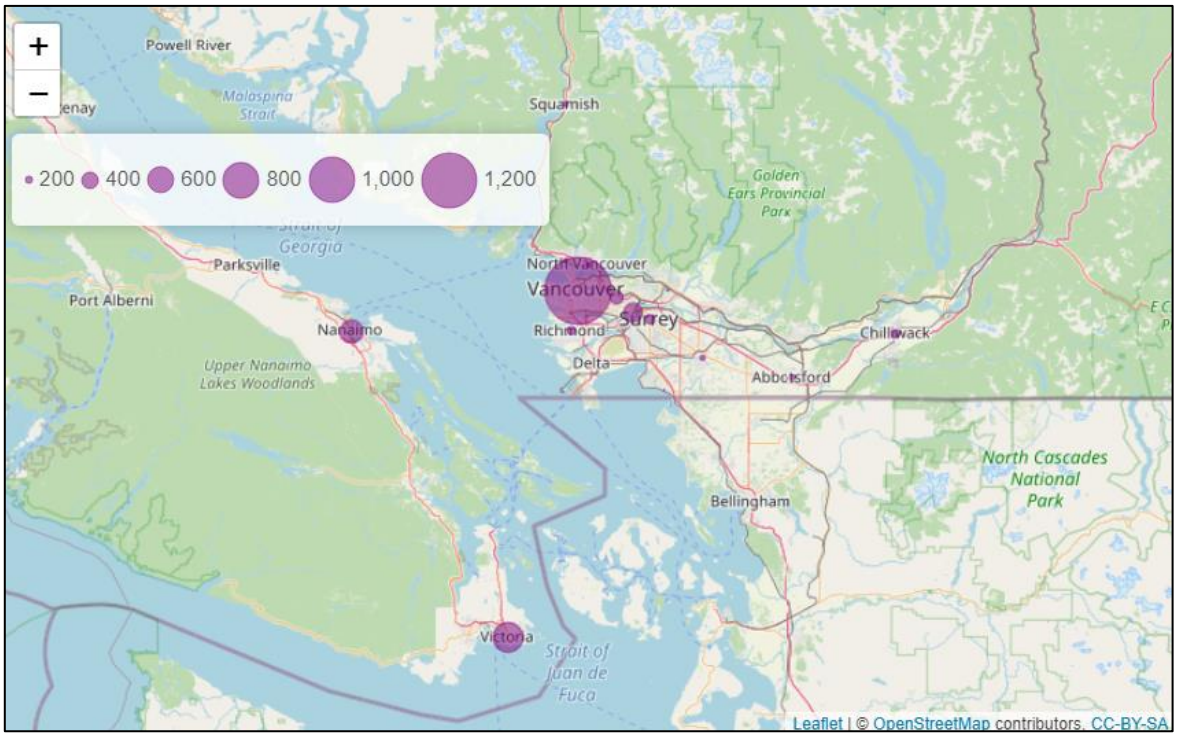


Figure 55: Geographical Map of The Distribution of Employees in Canada

The geographical map is another extra technique that was utilized that was used in question 3 (refer to 5.3.1). As previously documented in the analysis section, the code was already explained that coordinates of the cities were needed to plot the concentration of employees in this graph. The concentration of employees is plotted across the geographical map to aid in viewing the population of the cities. Instead of comparing the population of employees in each city, it would help to see if there are any geographical indifferences between the cities. It would cover issues such as greater distance for travelling or due to the city being in a rural area with mountains and rivers that transportation is costly for the organization.

By viewing the distance and the geographical nature of the city, it would help in making a conclusion for the analysis regarding employees in the city. This geographical plot is especially helpful to analyse datasets or columns that focuses on the location of the organization that other types of plots couldn't.

6.4 Extra Feature 4 – Plotly Library

The plotly library provides an interactive plot where specific values and categories can further be analysed by hovering the cursor on the bar or isolating a category to view its proportion. This library was used in question 5 (refer to 5.5.2) to view the length of service in a histogram by isolating the departments.

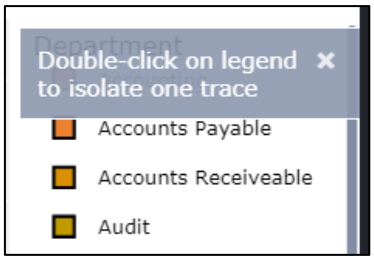


Figure 56: Isolating a variable using plotly library

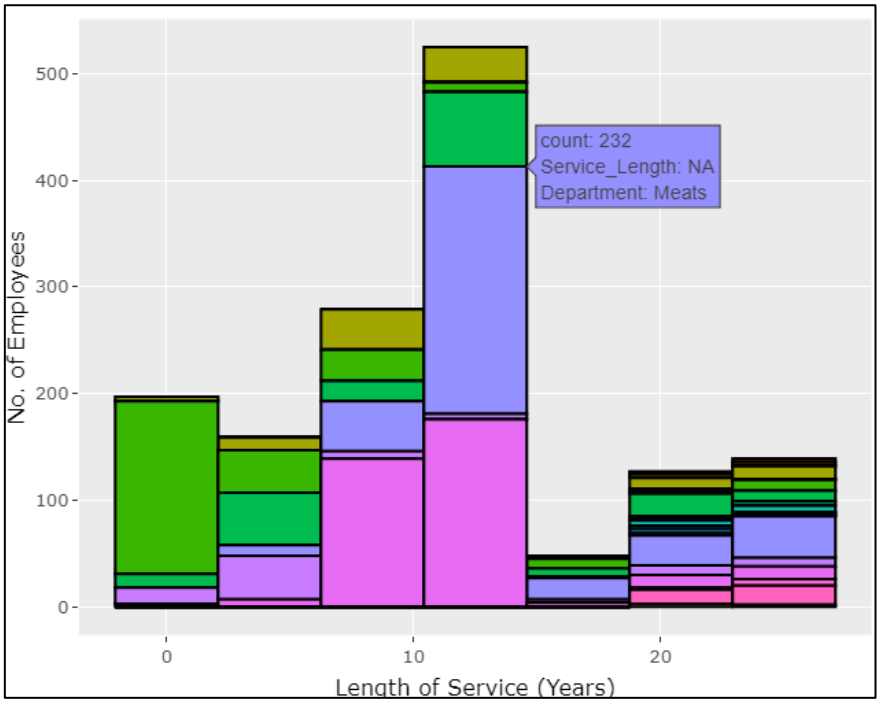


Figure 57: Hover to View Details Using Plotly Library

This package was used to help to easily identify stacked variables in a histogram or bar chart that has many categories. In this case there were many departments that made it difficult to see the spread especially when some colours were similar. Isolating the departments made it clearer and evidently easier to see compared to the overall graph proportions (refer to figure 50). Without using this library, the overall stacked histogram would be very hard to analyse by the individual departments.

6.5 Extra Feature 5 – Esquisse Library

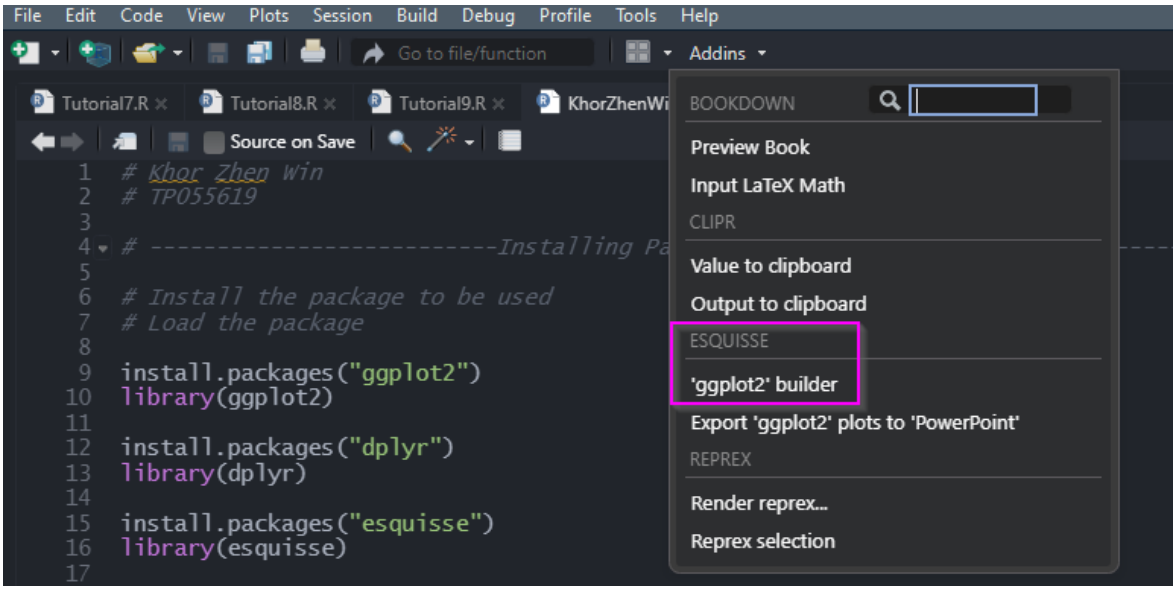


Figure 58: Esquisse Library Builder

This library is a helper to easily plot visualizations using a drag and drop GUI menu. The wizard is intuitive so that the analysts need not learn the syntax of ggplot2 library. This library would help automatically generate the type of plots based on the nature of the column values. For example, placing a categorical value will automatically generate a bar chart as shown in the figures below.

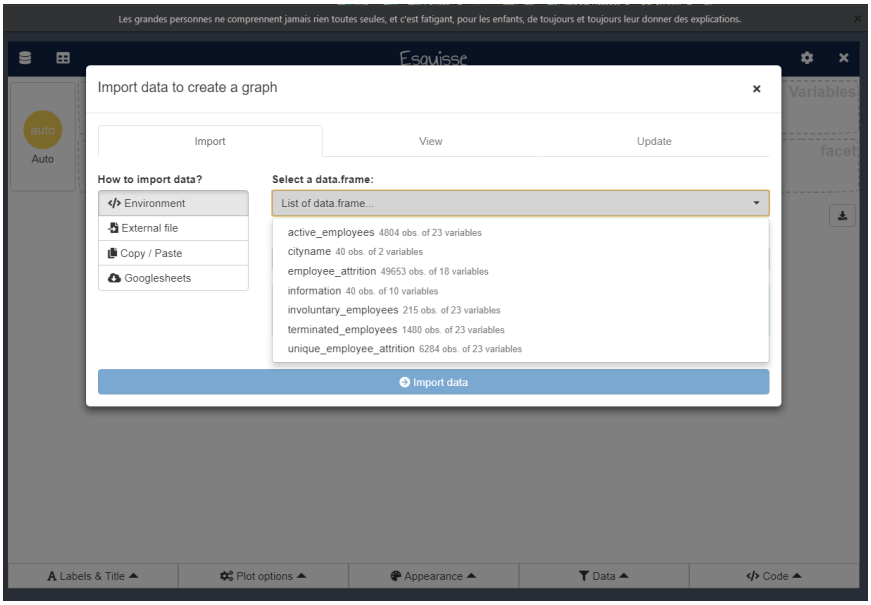


Figure 59: Choosing a Data Frame to Plot

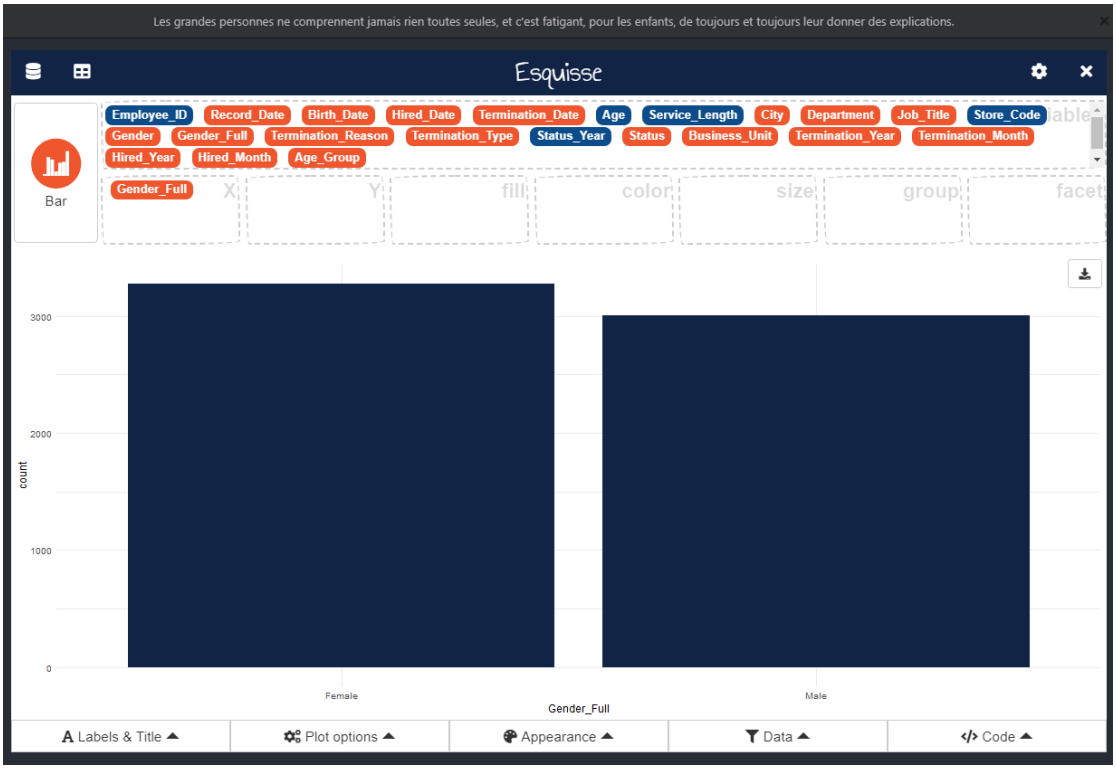


Figure 60: Bar Graph Automatically Plots After Dragging a Variable into the X Column

This library would help analysts to focus on the justifications and theories when researching compared to focusing on coding syntaxes of plotting. This would save more time and make it more efficient in the analytical cycle rather than focusing on the code to plot.

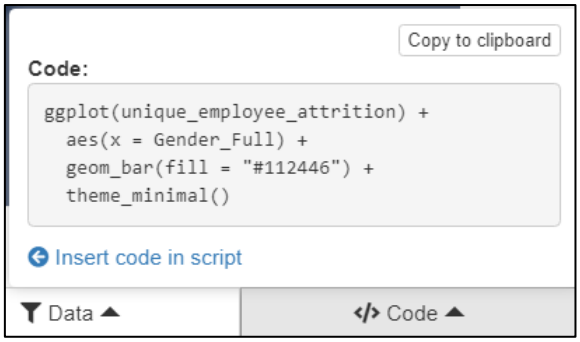


Figure 61: Code Generated by Esquisse Library

7.0 Conclusion

Based off all the analysis done, a general conclusion can be made about the attrition rate about this organization. The date, location, and the department of the employees were some of the dependent factors on the rate of attrition. Factors such as the age, gender and the length of service was found to not be of any relation to the rate of attrition. Various techniques were used to explore, pre-process, transform, and visualize the data which mostly included bar charts due to this data set having mostly categorical variables. Extensive research about the economical and geographical distribution of Canada was done to justify the theories stated in this report. The research done was to solidify the reasoning and justification of the views stated in this report.

8.0 References

CFI Education Inc. (2021). *Laid Off - Definition, Reasons, Alternatives to Layoffs*. Retrieved from Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/careers/jobs/laid-off/

Ci, W., Morissette, R., & Schellenberg, G. (27 June, 2016). *Hires and Layoffs in Canada's Economic Regions: Experimental Estimates, 2003 to 2013*. Retrieved from Statistics Canada: https://www150.statcan.gc.ca/n1/pub/11-626-x/11-626-x2016060-eng.htm

Columbia University. (30 September, 2021). *10 Advantages of Retaining and Hiring Older Workers:*. Retrieved from Columbia Public Health: https://www.publichealth.columbia.edu/research/age-smart-employer/advantages-older-workers

Equal Rights Advocates. (2019). *Gender Discrimination at Work*. Retrieved from Equal Rights Advocates: https://www.equalrights.org/issue/economic-workplace-equality/discrimination-at-work/

Fallon, N. (13 October, 2021). *Solving the Mystery of Gen Y Job Hoppers*. Retrieved from Business News Daily: https://www.businessnewsdaily.com/7012-millennial-job-hopping.html

Hale, J. (30 August, 2018). *7 Data Types: A Better Way to Think about Data Types for Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/7-data-types-a-better-way-to-think-about-data-types-for-machine-learning-939fae99a689

Investopedia. (16 April, 2021). *How Does Financial Accounting Help Decision-Making?* Retrieved from Investopedia: https://www.investopedia.com/ask/answers/041515/how-does-financial-accounting-help-decision-making.asp

Investopedia. (21 March, 2021). *Unemployment and Recession—What's the Relation?* Retrieved from Investopedia: https://www.investopedia.com/ask/answers/032515/why-does-unemployment-tend-rise-during-recession.asp

McCarthy, D. (4 December, 2019). *Learn How to Get an Employee to Quit*. Retrieved from The Balance Careers: https://www.thebalancecareers.com/how-to-coach-an-employee-out-of-a-job-2275942

Shethna, J. (2020). *How Does Outsourcing Reduce Cost?* Retrieved from Educba: https://www.educba.com/how-does-outsourcing-reduce-cost/

SHRM. (2021). *What criteria should be used in selecting employees for a workforce reduction?* Retrieved from Society for Human Resource Management (SHRM): https://www.shrm.org/resourcesandtools/tools-and-samples/hr-qa/pages/selectingemployeesforlayoff.aspx

The Center for Work Ethic Development. (19 September, 2018). *Why It's Important for Young People to Have Part-Time Jobs*. Retrieved from The Center for Work Ethic Development: https://workethic.org/why-its-important-for-young-people-to-have-part-time-jobs/

The World Bank Group. (2021). *GDP growth (annual %) - Canada*. Retrieved from The World Bank Data: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2016&locations=CA&start=2011&view=chart

todoCanada. (8 February, 2017). *Population in Canada's Cities – 2016 Census*. Retrieved from To Do Canada : https://www.todocanada.ca/population-canadas-cities-2016-census/

Weisberg Cummings P.C. (2021). *EXAMPLES OF WRONGFUL TERMINATION BASED ON DISCRIMINATION*. Retrieved from Weisberg Cummings P.C.: https://www.weisbergcummings.com/guide-employee-discrimination/chapter-7-what-is-wrongful-termination/

YPulse. (19 October, 2021). *How Gen Z & Millennials Are Fueling The Great Resignation*. Retrieved from YPulse: https://www.ypulse.com/article/2021/10/19/how-gen-z-millennials-are-fueling-the-great-resignation/