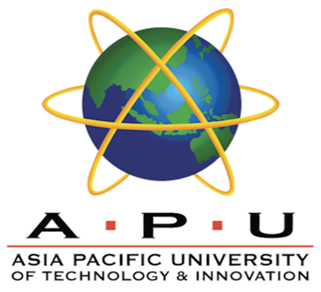


****

**INDIVIDUAL ASSIGNMENT**

**TECHNOLOGY PARK MALAYSIA**

**CT127-3-2-PFDA**

**PROGRAMMING FOR DATA ANALYTICS**

**APD2F2109CS (DA)**

**NAME: Yee Zhi Ying**

**TP NUMBER: TP055495**

**LECTURER: Mr. Amardeep Singh A/L Uttam Singh**

**HAND-OUT DATE: 4TH October 2021**

**HAND-IN DATE: 22TH NOVEMBER 2021**

**WEIGHTAGE: 50%**

**Table of Contents**

[**1.0 Introduction** 3](#_Toc88512454)

[**2.0 Data Exploration** 4](#_Toc88512455)

[**2.1 List All Column Names** 5](#_Toc88512456)

[**2.2 Sample Dataset & Number of rows and columns** 5](#_Toc88512457)

[**2.3 Data Structure of Dataset** 6](#_Toc88512458)

[**2.4 Missing Values in Dataset** 7](#_Toc88512459)

[**2.5 Distribution of Numeric Attributes** 7](#_Toc88512460)

[**2.5.1 Histogram of EmployeeID** 8](#_Toc88512461)

[**2.5.2 Histogram of Age with Active and Terminated Employees** 9](#_Toc88512462)

[**2.5.3 Histogram of Length of Service and Store Name** 10](#_Toc88512463)

[**2.5.1 Line Chart of Active and Terminated Employee Status Year** 11](#_Toc88512464)

[**2.6 Distribution of Character Attributes** 12](#_Toc88512465)

[**2.6.1 Bar Chart of Gender** 12](#_Toc88512466)

[**2.6.2 Bar Chart of Employee Status and City Name** 13](#_Toc88512467)

[**2.6.3 Bar Chart of Business Unit** 14](#_Toc88512468)

[**2.6.4 Bar Chart of Department Name and Job Titles** 14](#_Toc88512469)

[**2.6.5 Bar Chart of Termination Type and Termination Reason** 16](#_Toc88512470)

[**3.0 Data Transformation** 17](#_Toc88512471)

[**3.1 Rename Column** 17](#_Toc88512472)

[**3.2 Transform Date Columns** 18](#_Toc88512473)

[**3.3 Replace Values** 19](#_Toc88512474)

[**3.4 Create Age Group** 19](#_Toc88512475)

[**3.5 Split Dataset** 20](#_Toc88512476)

[**3.6 Remove Duplicate EmployeeID** 21](#_Toc88512477)

[**3.7 Recalculate Length of Service** 23](#_Toc88512478)

[**3.8 Eliminate Unwanted Columns** 23](#_Toc88512479)

[**4.0 Data Manipulation & Data Visualization** 25](#_Toc88512480)

[**4.1 Area 1: What is the Annual Attrition Rate?** 26](#_Toc88512481)

[**4.1.1 Hired employees in 2014 and 2015** 28](#_Toc88512482)

[**4.2 Area 2: Reason of Attrition by Department and Job Title** 30](#_Toc88512483)

[**4.2.1 Termination Type by Termination Reason** 30](#_Toc88512484)

[**4.2.2 Termination Reason by Department Name** 31](#_Toc88512485)

[**4.2.3 Termination Reason by Top 5 Job Title with Department Name** 33](#_Toc88512486)

[**4.3 Area 3: Analysis on jobs by Gender and Age** 38](#_Toc88512487)

[**4.3.1 Age Group by Gender** 38](#_Toc88512488)

[**4.3.2 Gender by Age** 39](#_Toc88512489)

[**4.3.3 Gender and Age Group with Termination Reason** 40](#_Toc88512490)

[**4.3.4 Gender and Age with Job Titles** 42](#_Toc88512491)

[**4.4 Area 4: Analysis of different job types and its length of service** 44](#_Toc88512492)

[**4.4.1 Distribution of Length of Service** 44](#_Toc88512493)

[**4.4.2 Average Length of Service for different Departments and Jobs** 46](#_Toc88512494)

[**4.4.3 Length of Service and Age by Year and Termination Reason** 47](#_Toc88512495)

[**4.5 Area 5: Analysis of the Job Environment** 50](#_Toc88512496)

[**4.5.1 City Name and Store Name** 50](#_Toc88512497)

[**4.5.2 City Name and Department by Business Unit** 51](#_Toc88512498)

[**4.5.3 City Name by Termination Reason** 53](#_Toc88512499)

[***4.5.3.1 Drill down on Fort Nelson with laid-off Job Titles*** 54](#_Toc88512500)

[***4.5.3.2 Drill down on Vancouver with retired and resigned Job Titles*** 55](#_Toc88512501)

[**5.0 Extra Features** 57](#_Toc88512502)

[**5.1 Treemap with Treemapify package** 57](#_Toc88512503)

[**5.2 Fct\_reorder() in Forcats package** 58](#_Toc88512504)

[**5.3 Cut() and Difftime() function** 60](#_Toc88512505)

[**6.0 Conclusion** 61](#_Toc88512506)

[**References** 62](#_Toc88512507)

[**Appendix** 64](#_Toc88512508)

[**A. Statistics of Robbery in British Columbia 2005 - 2014** 64](#_Toc88512509)

# **1.0 Introduction**

In this data-driven technology era, businesses acquire the use of data analytics to gain useful insight and make accurate business decisions from comprehensive analytical reports. With data analytics, businesses are able to enhance customer experience, streamline operations, mitigate and handle potential risk and etc.

In this analysis, data analytics techniques will be used to explore and study the reason of employees leaving the company from the given dataset, Attrition.csv. All techniques that are used in data exploration, data transformation, data manipulation, and data visualization will be detailed described, followed with R programming code snippets and different graphs that justify the final findings after the analysis. From the dataset, it is assumed that the data collected is from a big company that operates with various franchises around the country. Therefore, the dataset is very useful in finding potential patterns or relationships to study the problem domain.

In data analytics, there are different types of technologies and methods used to analyse data with different purpose. The methods will be used are Descriptive Analytics to look into what had happened in the past with a summary statistics of all attributes of the dataset, Exploratory Analytics to study the data relationships, Diagnostics Analytics to look into the reason of why it happened.

The purpose of this analysis is **to understand and analyse the potential reasons of attrition**, so that the business is **able to make improvements in terms of employee engagement and productivity, work environments, employee benefits and etc**. Besides that, the HR department will have a clearer insight on reducing unwanted attrition and eliminating unwanted costs and resources related to training replacement employees. In addition, it also enables the HR department to understand the satisfactory of the job from the turnover rates, and key benchmarks including a descriptive profile of all the employees who have left the company by positions, departments or business units. With this analysis, the business will be able to reduce turnover rates and achieve and maintain overall growth and health of the business by improving different aspects of the job.

# **2.0 Data Exploration**

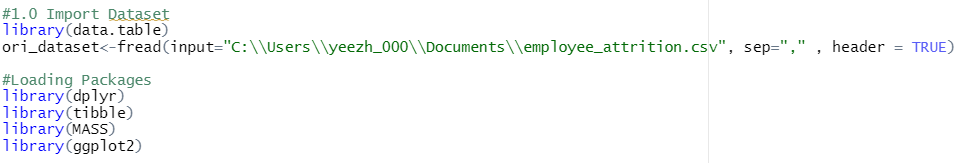
In this section, the attrition dataset will be explored with summary statistics techniques that shows the distribution of the attributes with different statistics calculation. Before that, the dataset will be loaded in R studio with the fread() function from data.table package, as it allows faster and lesser memory space usage in data imports when comparing to read\_csv() function (Jozef, 2019) . After importing the dataset, some packages will be loaded into the R studio to stand by for other operations.

Figure 1: Import dataset with fread() and loading other packages.

After that I used print() function to view the top 10 rows of the imported dataset as a tibble format, which is more neat and easier to visualize. From the output, I am only able to view some parts of the dataset, therefore the below section will be demonstrating data exploration of better understanding the whole structure of the dataset.

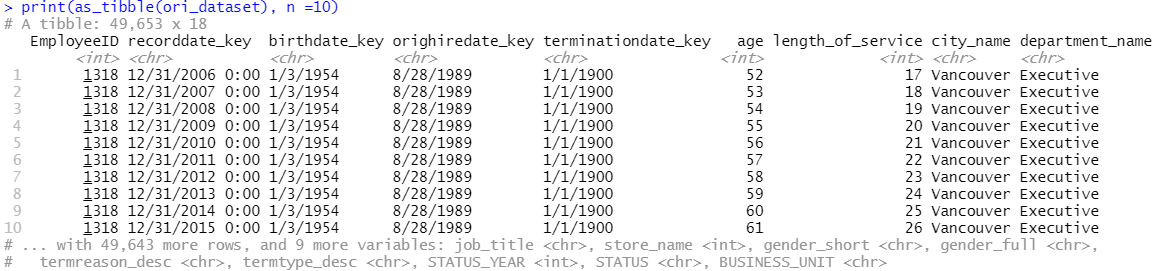


Figure 2: The top 10 rows of Attrition dataset with print as tibble function.

## **2.1 List All Column Names**

First by listing the column names of the Attrition dataset with names() function. From *Figure 3*, there is a total of 18 columns found in this dataset. The attrition dataset were recorded with attributes such as termination date, reason of termination, job titles and etc. However, the column names were found to be inconsistently labelled with capital and lower case letters, and there are duplicated columns like gender\_short and gender\_full that can be removed in the pre-processing stage.

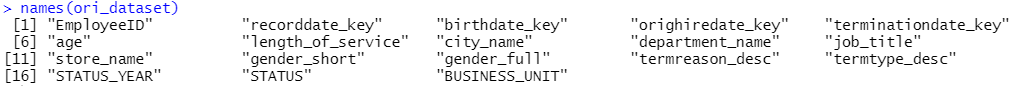


Figure 3: All column names of Attrition dataset with names() function.

## **2.2 Sample Dataset & Number of rows and columns**

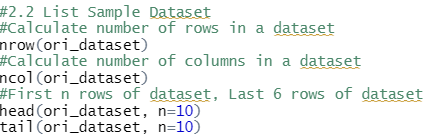
 Besides that, nrow() and ncol() functions can be used to calculate the number of rows and columns in the dataset. There are a total of 49653 numbers of rows and 18 columns recorded in the Attrition dataset. Head() and tail() function were used to view the top 10 and last 10 records of the dataset. As a result, there are duplicated records in EmployeeID, which means the dataset has recorded that particular employee in the attrition data multiple times. Additionally, some of the values are inconsistent like “Not Applicable” can be written in short as “NA”, the recorddate\_key can be stored with normal date type as “mm/dd/yyyy”, and terminationdate\_key in the first 10 rows are invalid values as the date is before the orighiredate\_key.

Figure 4: Code for finding number of rows and columns, and the top N or last N of the dataset.

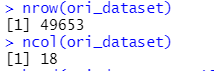


Figure 5: The result of nrow() and ncol().

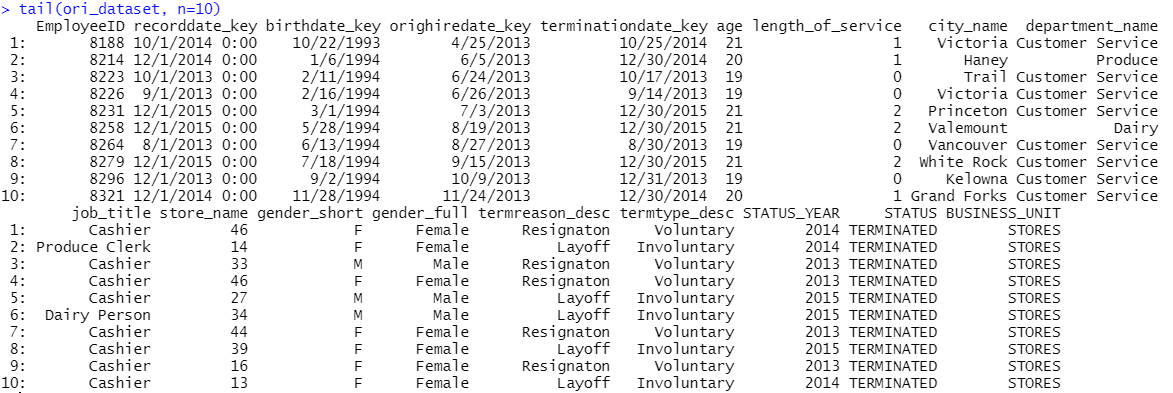
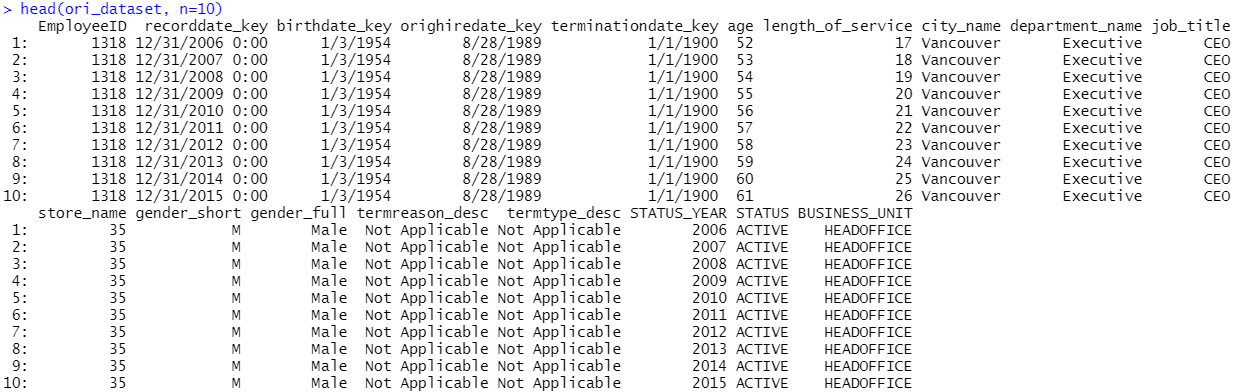


Figure 6: The result of last 10 rows of the dataset with tail() function.

Figure 7: The result of top 10 rows of the dataset with head() function.

## **2.3 Data Structure of Dataset**

 In addition, str() function is used to list the data structure of the attributes. As a result, there are a total of 5 integer variables and 13 character variables in the Attrition dataset. The date attributes like birthdate\_key were detected as character variables, but it should be recognized as date variables.

Figure 8: The code for listing data structure of the attributes in Attrition dataset.

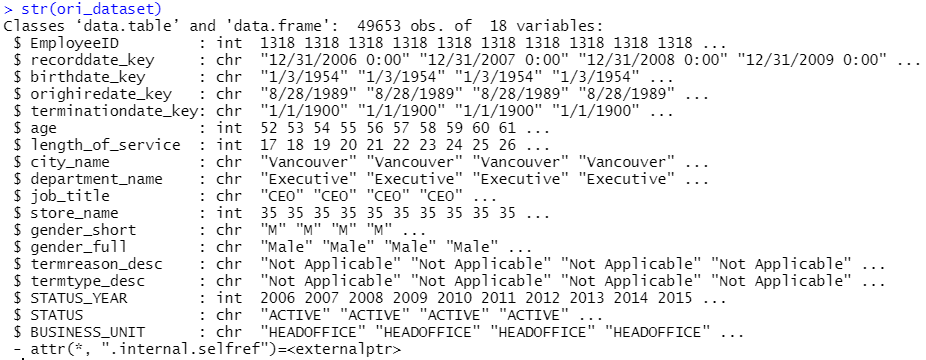


Figure 9: The data structure of all the attributes.

## **2.4 Missing Values in Dataset**

Detecting and removing missing values is also vital in data analysis, and therefore the following code is using colSums() function to calculate the sum of missing values in each column. Surprisingly, there are no missing values detected across all the columns that needed to be processed or removed.

Figure 10: Code for finding missing values in the dataset.

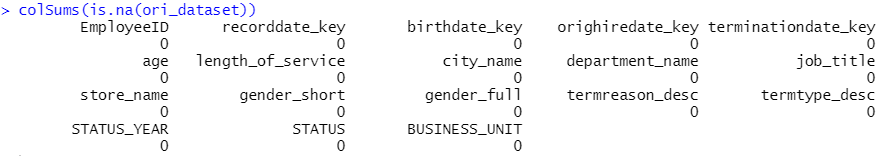


Figure 11: There is no missing values detected in the dataset.

## **2.5 Distribution of Numeric Attributes**

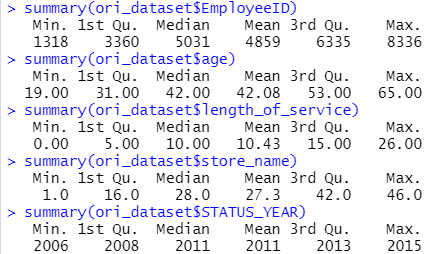
In this section, exploration on distribution of numeric attributes will be shown with the use of summary() function and suitable graphs. The summary() function can only be used for numeric attributes to calculate the min, max, median, and percentiles, and the result is stated in *Figure 12*. According to *Figure 12*, the **average age of employees is 42 years old**, and the **average length of service is 10 years**. Besides that, the status of employee in **this dataset are recorded between year 2006 and year 2015**. Other than that, employees will work in different store names which are different outlets or franchises under the company. In addition, Histogram and Line chart will be used in the following section to visualize a better sight of the distribution of each numeric attributes.

Figure 12: The summary statistics of numeric attributes.

### **2.5.1 Histogram of EmployeeID**

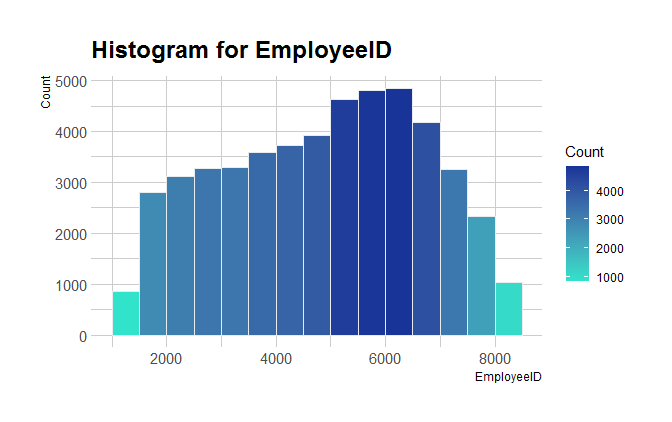
****In *Figure 13*, the data values of EmployeeID has higher frequency are distributed between **5000 and 7000**. However, as according to the graph, the histogram is not evenly distributed, which also mean that the **duplicated EmployeeID** are among the **higher frequency of EmployeeID groups**.

Figure 13: The histogram for EmployeeID.

### **2.5.2 Histogram of Age with Active and Terminated Employees**

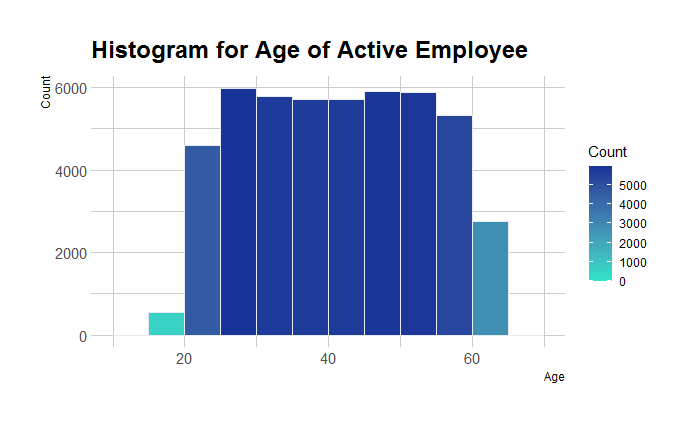
****In *Figure 14*, the **age of terminated employees are mostly 60 and above**, which also means that these employees are voluntarily leaving the company due to retirement. However, a **spike is noticed in the 20-30 age group**, which can indicate that some of the employees are college students or fresh graduates seeking for a better job from another company. In *Figure 15*, the age distribution of active employees are evenly distributed across 20 to 60 years old. There are roughly **3000 employees who are still working under this company that are above 60 years old**.

Figure 14: Histogram for Age of Active Employee.

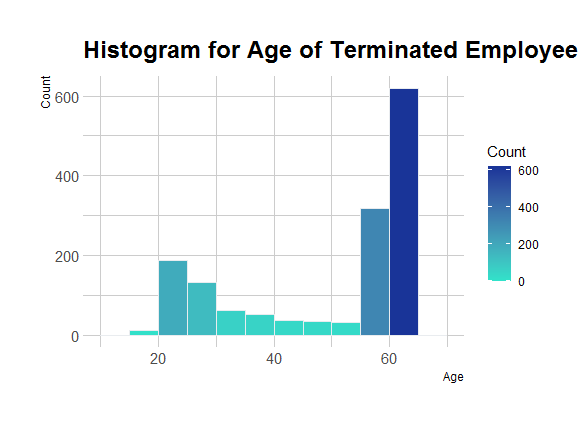
****

Figure 15: Histogram for Age of Terminated Employee.

### **2.5.3 Histogram of Length of Service and Store Name**

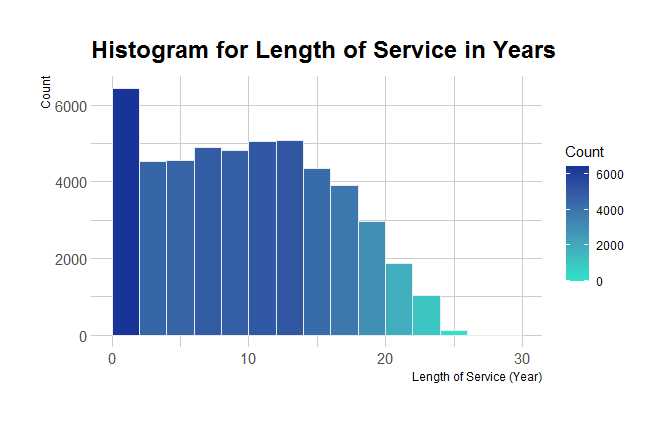
****Most of the employees only **have less than 2 years of length of service** in the company. In short, as the years of service increases, there is a high chance where employees might leave the company because of old age or retirement (*Figure 16*). In addition, **the group of 40 to 42 store number has the highest number of employees working** in it (*Figure 17*).

Figure 16: Histogram for Length of Service in Years.

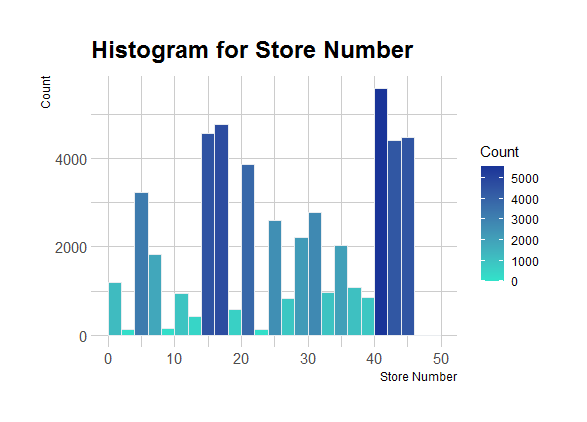
****

Figure 17: Histogram for Store Number.

### **2.5.1 Line Chart of Active and Terminated Employee Status Year**

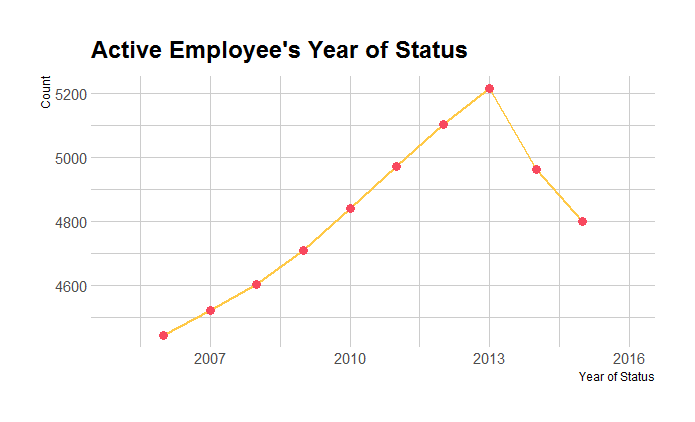
Furthermore, when compared from two line charts (*Figure 18 & Figure 19*), there is a **spike of attrition in year 2014**, and at the same time the number of active employees dropped. However in 2015, both figures have a slightly dropped in year 2015.

Figure 18: The Line chart for status year of active employees.

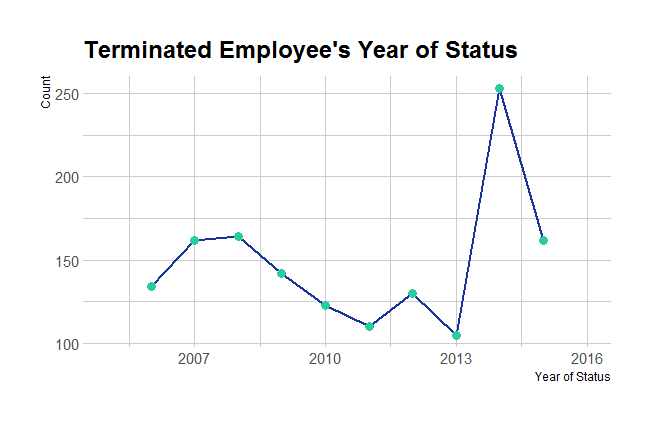


Figure 19: The Line chart for status year of terminated employees.

## **2.6 Distribution of Character Attributes**

### **2.6.1 Bar Chart of Gender**

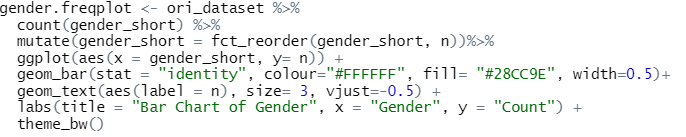
 Similarly, character attributes will be explored with the count() function and bar charts to view the distribution of the data values. Besides that, the fct\_reroder() function is used to reorder the values, and the hrbthemes package is used to customize themes and background colour for different graphs. From the code in *Figure 20*, it generated the Bar Chart as in (*Figure 22*) and the similar code is used to generate other bar charts in the following section. According to the bar chart below, it shows that **there are more** **Female employees than Male employees in the company**.

Figure 20: Sample Code for generating Bar chart for Gender.

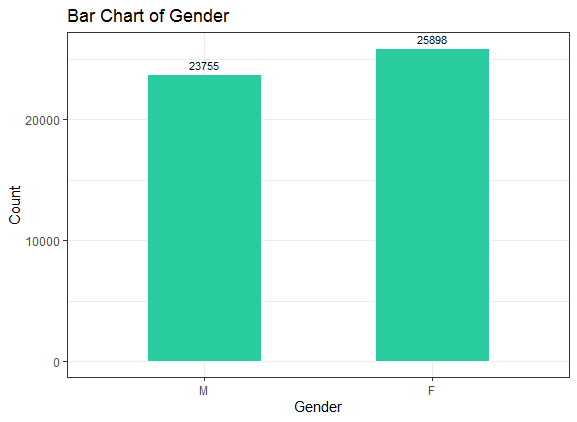


Figure 21: Bar Chart of Gender

### **2.6.2 Bar Chart of Employee Status and City Name**

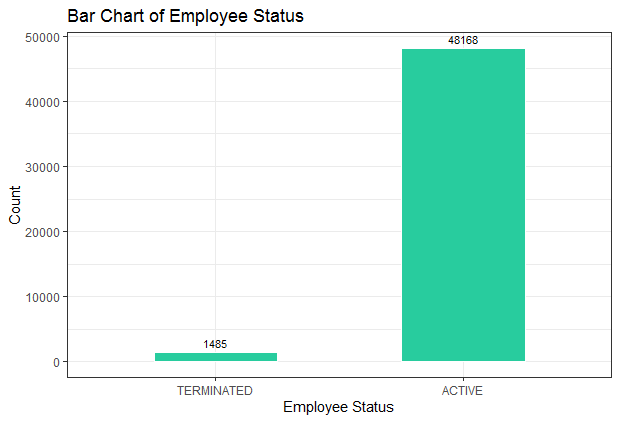
In *Figure 22*, it shows that there are **48168 employees who are still active**, whereas there are only **a small proportion of employees who have been terminated** their service in the company. In *Figure 23*, it shows that most of the city names are from **British Columbia, Canada**, which can indicate that the dataset is from a company located in this country. The city that has the **highest number of employees is Vancouver**, with a total of 11211 employees.

Figure 22: Bar Chart of Employee Status

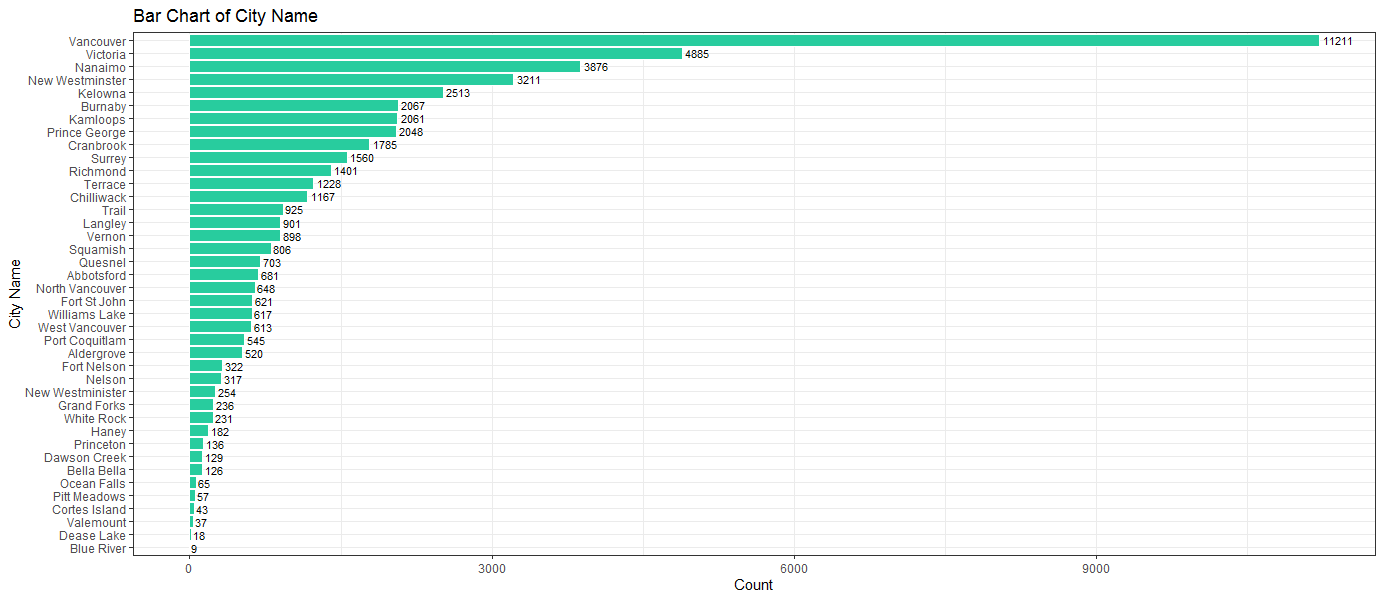


Figure 23: Bar Chart for City Name

### **2.6.3 Bar Chart of Business Unit**

Based on *Figure 24*, business units of the company are separated into **Head Office and Stores**, where employees under Head Office are mostly top-level management, and employees under stores are more towards customer service or front-end employees that serves the customers such as cashier, customer service, and so on. As similar to most business organizational structure, most of the top-level management have lesser number of people in it that manages all different departments or segments of the business. In this case, these segments are stores and the departments under it.

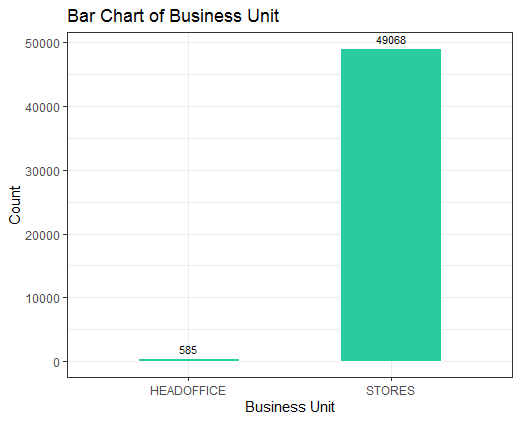
****

Figure 24: Bar Chart of Business Unit.

### **2.6.4 Bar Chart of Department Name and Job Titles**

The **departments** that has the higher frequency count of employees are from **Meats, Dairy, Produce, Bakery, and Customer Service** (*Figure 25*), and the **job titles** that has the higher frequency count of employees are **Meat Cutter, Dairy Person, Produce Clerk, Baker, and Shelf Stocker** (*Figure 26*). Firstly, it indicates that the company is **a retail business that sells produce, meat and dairy products in the grocery industry**, and hence each store may need these jobs to serve the customers. Besides that, these departments and these jobs are mostly highly repetitive, which can lead to demotivation and boredom, and hence has a higher chance of attrition.

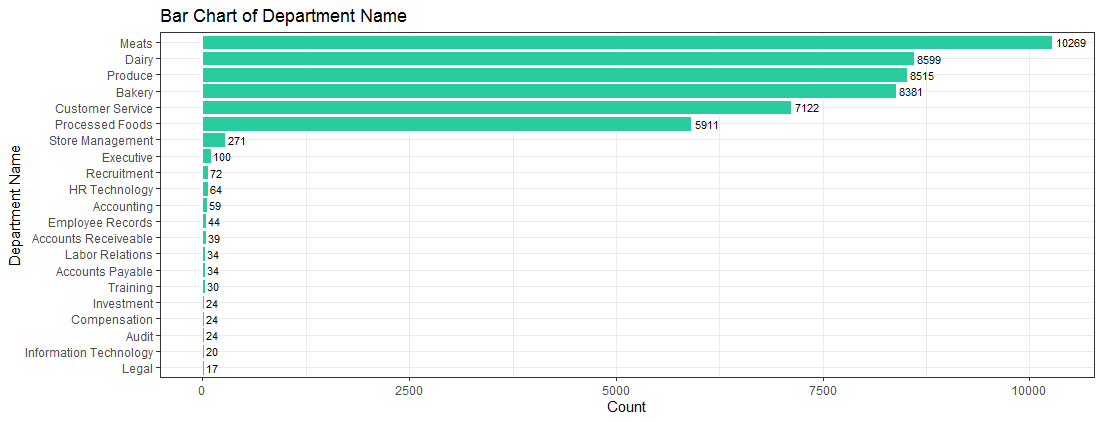


Figure 25: Bar Chart of Department Name.

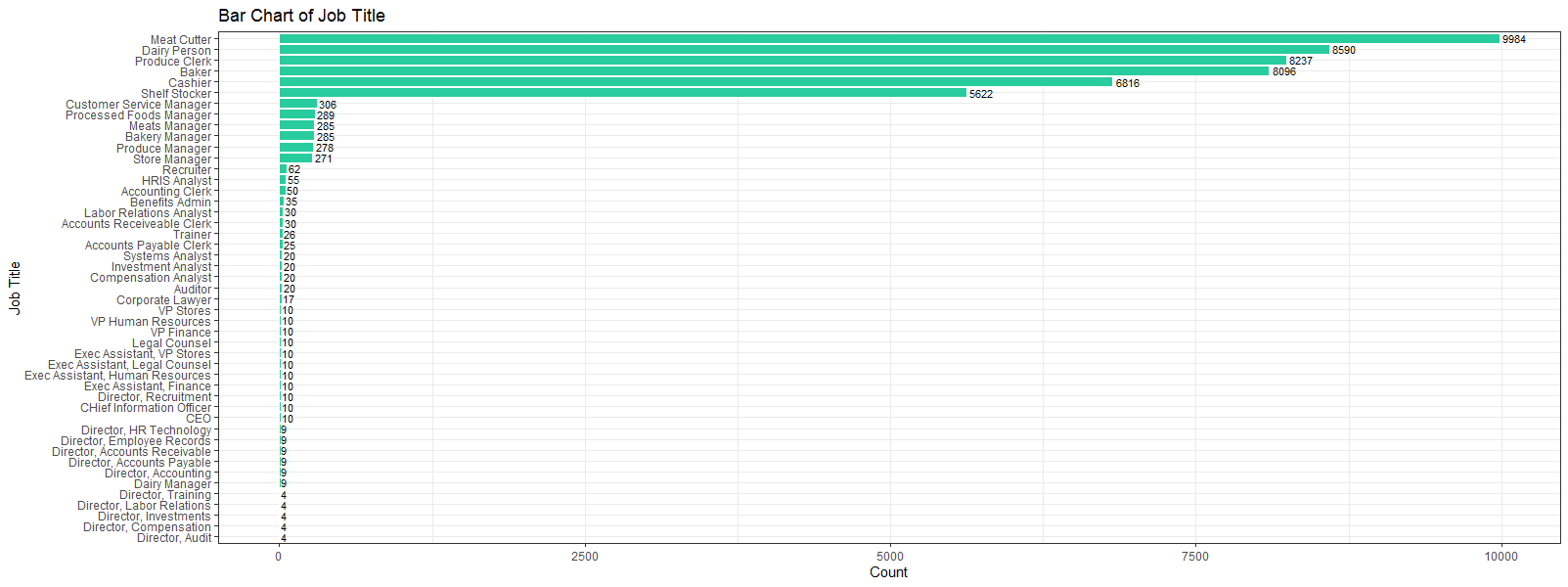


Figure 26: Bar Chart of Job Title.

### **2.6.5 Bar Chart of Termination Type and Termination Reason**

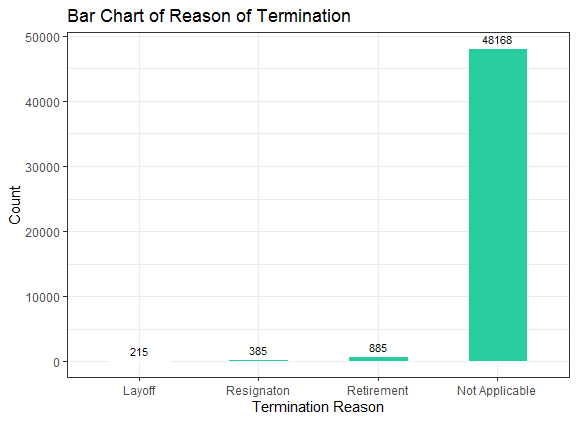
****In *Figure 28 & Figure 27*, it shows that both of the attributes have a high frequency of data values that are Not Applicable. In this case, it can be replaced into a global constant value, NA, under data transformation. The **types of termination** are **Voluntary and Involuntary**, where most of the employees are voluntarily leaving the company. The main **termination reason** for terminated employees are **Layoff, Resignation, and Retirement**, where retirement has the highest count among the three.

Figure 27: Bar Chart of Termination Reason.

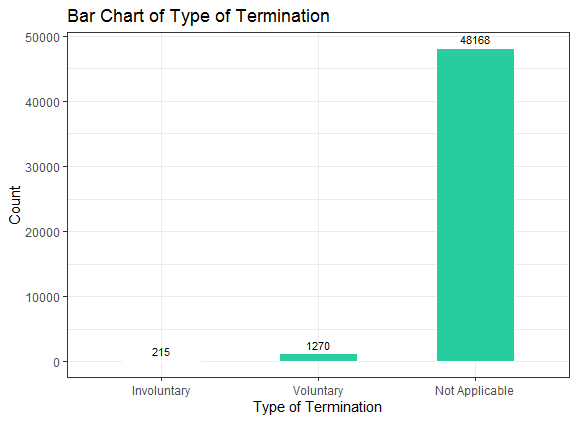


Figure 28: Bar Chart of Type of Termination.

# **3.0 Data Transformation**

## **3.1 Rename Column**

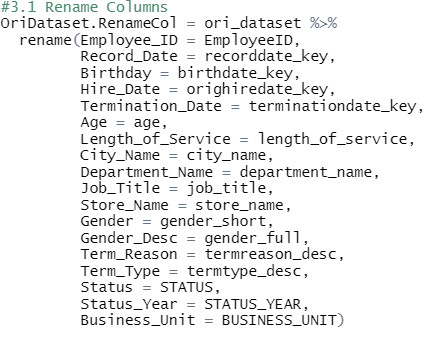
Based on *Figure 3*, the column names in the original dataset is inconsistent, hence the first step of data transformation is to rename all the columns into consistent string patterns. In *Figure 29*, all the column names have been renamed with the rename() function, and the output is checked by running the str() function and the result is shown in *Figure 30*.

Figure 29: Renaming column from the original dataset with rename() function.

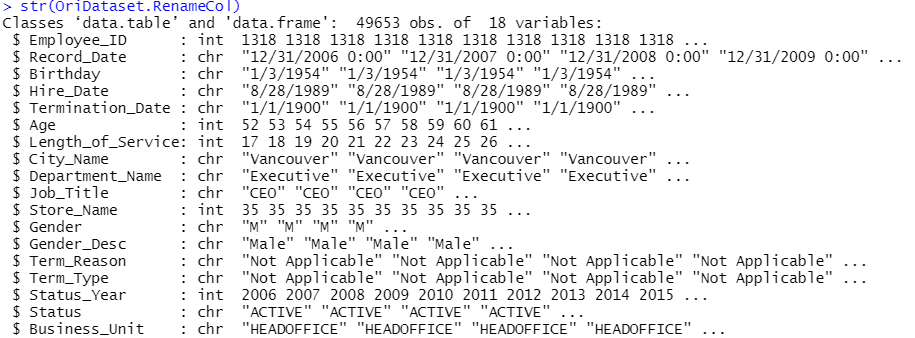


Figure 30: After renaming all the column, str() function is used to check the output.

## **3.2 Transform Date Columns**

After that, some of the date attributes were transformed from character to date variables with as.Date() function. In *Figure 31*, the “OriDataset.RenameCol” is act as the input of the mutate() function, where it allows to create, modify and delete variables by preserving the existing values in a data frame (Wickham, 2021). After modifying the date columns into Date datatype, 4 new columns were added, which are Hire\_Year, Hire\_Month, Term\_Year, and Term\_Month. These new columns is to store the year and month of the Hire\_Date or Termination\_Date with the use of the format() function. When adding a new column, the add\_column function is used, and it can also state the position where the new column can be added. As shown in *Figure 31*, the columns were added after Hire\_Date and Termination\_Date by stating the “.after” argument. Similarly, the output dataset is verified and shown with the str() function and the view() function to view the full dataset in a table format (*Figure 32 & Figure 33*).

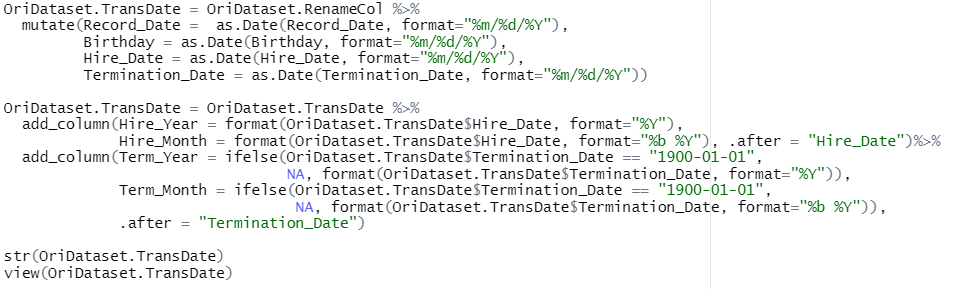


Figure 31: Transform all date attributes from character to date datatype, and added 4 new columns.

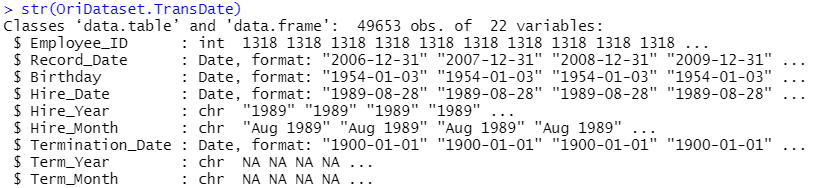


Figure 32: The output dataset is verified with str() function.

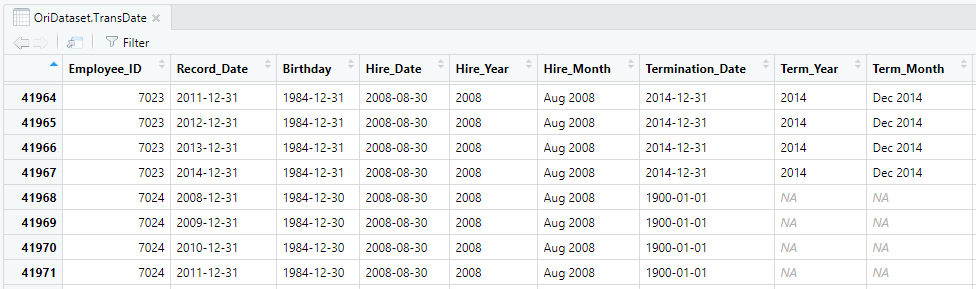


Figure 33: The whole dataset is visualized by running view() function.

## **3.3 Replace Values**

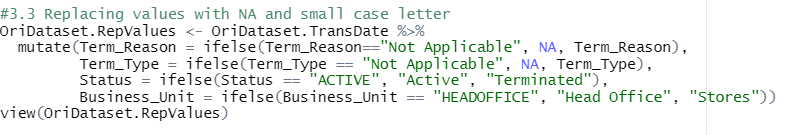
In the original dataset, there are some data values that need to be replaced to provide a more consistent and clean dataset. In *Figure 34*, data values were replaced with global constant, NA, and sentence case letters with ifelse statements. For example, the values in Term\_Reason will change to “NA”, if only it is equals to “Not Applicable.

Figure 34: Replacing values with ifelse() and mutate() in multiple columns.

## **3.4 Create Age Group**

Other than that, age group is created to classify the employees age into different working age groups. This is done with the cut() function, where it useful to categorize values of a continuous attribute like Age into different levels of a factor. The break argument is to specify the number of interval levels that we want to separate our data, and the labels argument is to label the categorical levels (Gonzalez, 2021). In *Figure 35*, Age\_Group is created by breaking the age attribute into 4 different levels, which are 15-24 is Student, 25-34 is Adult, 35-54 is Upper-Middle and 55 and above are Senior employees.

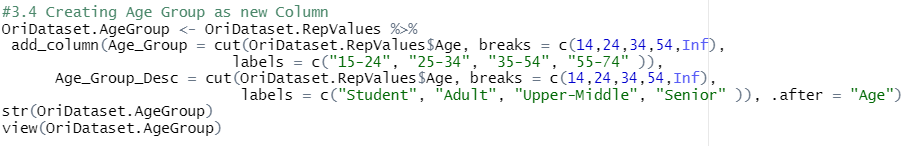


Figure 35: Age\_Group is added to classify the employees age into different Age Group.

## **3.5 Split Dataset**

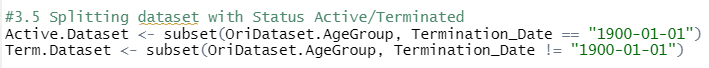
 As the main analysis is to determine and analyse the reason of attrition, and hence the dataset can be split into two parts which are Active employee dataset (*Figure 37*) and Terminated employee dataset (*Figure 38*). The terminated employee dataset will be used to further analyse on the length of service and drill down on different variables during the data manipulation stage. In *Figure 36*, the subset function is used to create a subset based on a condition (DataCamp Team, 2020). The Termination\_Date can be used as the condition to subset the active and terminated employees, as the active employees will have null values, 1900-01-01, as the termination date. This is also because of the active employees are still working in the company, and therefore there will not be any Termination\_Date in the records.

Figure 36: Splitting the dataset into two parts with the subset() function.

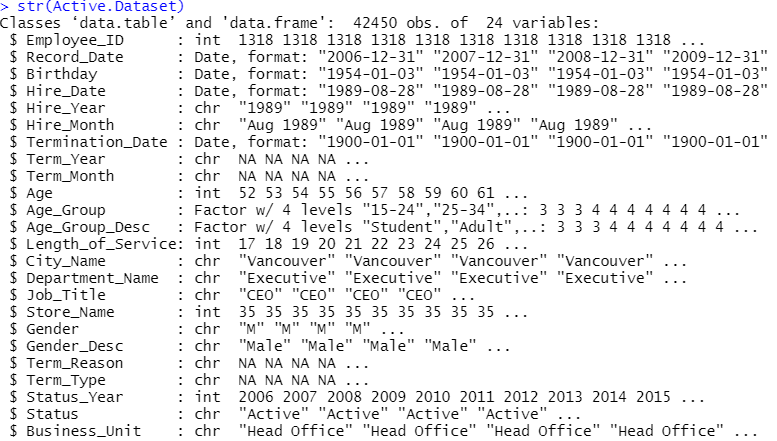


Figure 37: Active Employee Dataset has 42450 records and 24 attributes.

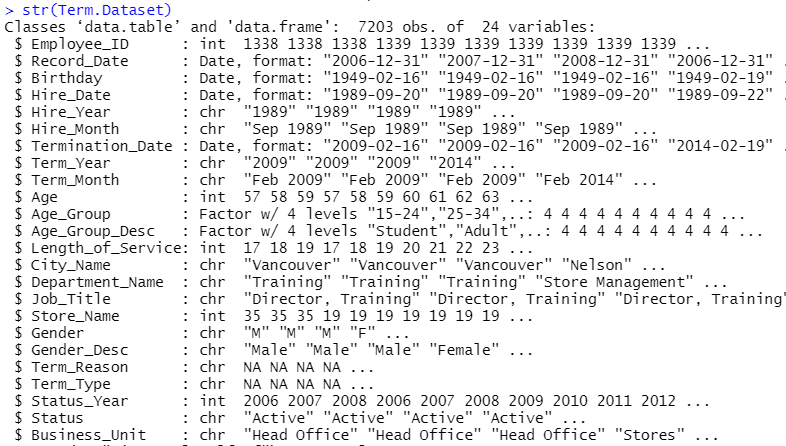
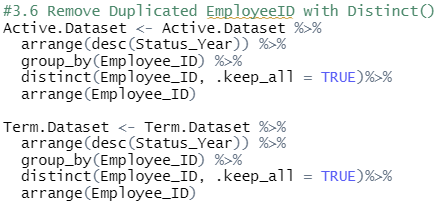


Figure 38: Terminated Employee Dataset has a total of 7203 records and 24 variables.

## **3.6 Remove Duplicate EmployeeID**

After creating two new subsets of the dataset, duplicate EmployeeID will be removed from the two subsets with the code in *Figure 39*. First is to arrange the status year by descending order and group by EmployeeID to group the data based on the latest status year of each employee. Next, the distinct() function is used to get the distinct values from the groups, and finally the EmployeeID is arrange in ascending order with the arrange function. The “.keep\_all” argument is to keep all other columns when finding the distinct values of the EmployeeID column.

Figure 39: Removing duplicated EmployeeID from the two datasets.



After that, I used the str() function to show the output and the count() function (*Figure 41*) to verify that there is no duplicated EmployeeID in both sets of data. It turns out when grouping the EmployeeID in *Figure 39*, the structure of the dataset has been transformed into a grouped dataframe *Figure 40*. Therefore, the two sets of data is then transformed back into a regular dataframe with data.frame() function (*Figure 42*). After removing all the duplicated EmployeeID, the active employees has a total of 4799 records and the terminated employees has only 1485 records as shown in *Figure 43*.

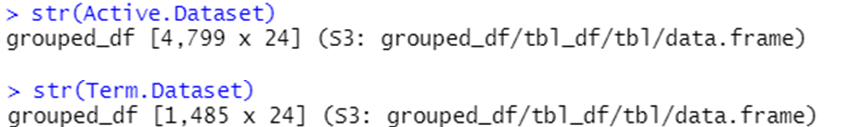


Figure 40: The two datasets have been transformed into grouped dataframe in Figure 39.

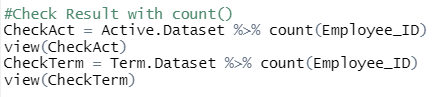


Figure 41: Using count to check the result after removing the duplicated EmployeeID.

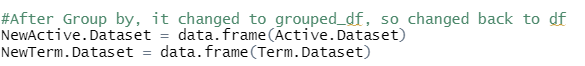
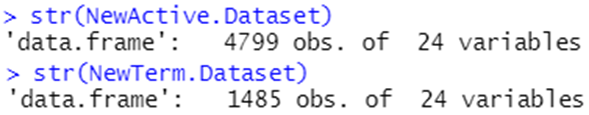


Figure 42: Transform the two datasets into dataframe with data.frame.

Figure 43: Final output of eliminating duplicated EmployeeID and transformed back to data frame.



## **3.7 Recalculate Length of Service**

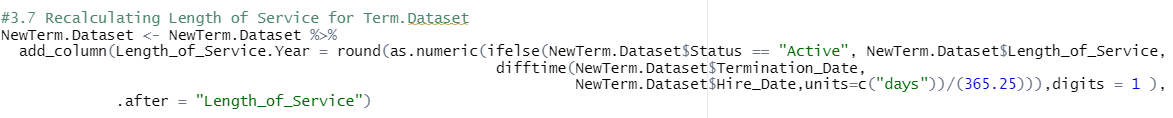
 In the Termination dataset, the length of service is recalculated into total years with months with the difftime() function in *Figure 43*. Difftime allows us to find the total days between two date columns by stating “days” in the units arguments. After finding the total days, the total years is calculated by dividing 365.12 days and it is round off to 1 decimal point. In short, these a new column is created as the Length\_of\_Service.Year and it is added to the Terminated Employee Dataset to show the length of service in a more detailed perspective based on days, months and years.

Figure 44: Recalculating Length of Service in Termination Dataset (DataScience Made Simple, 2021).

## **3.8 Eliminate Unwanted Columns**

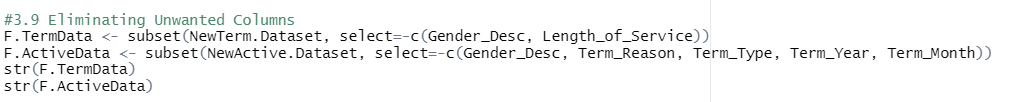
The final step of the data transformation is to eliminate unwanted columns found in both of the datasets. In this case, the subset function is used to drop unwanted columns by specifying “-” in the select argument. As a result, the two final datasets, F.TermData and F.ActiveData, are created as the final output of the data transformation phase.

Figure 45: Eliminating unwanted columns with the subset function.

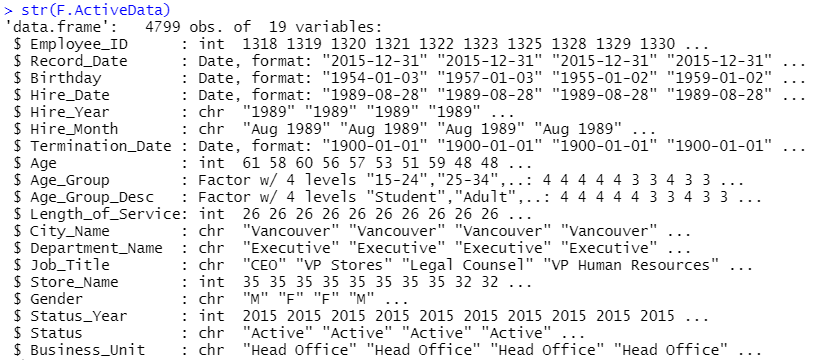


Figure 46: The final output after data transformation for Active employee dataset.

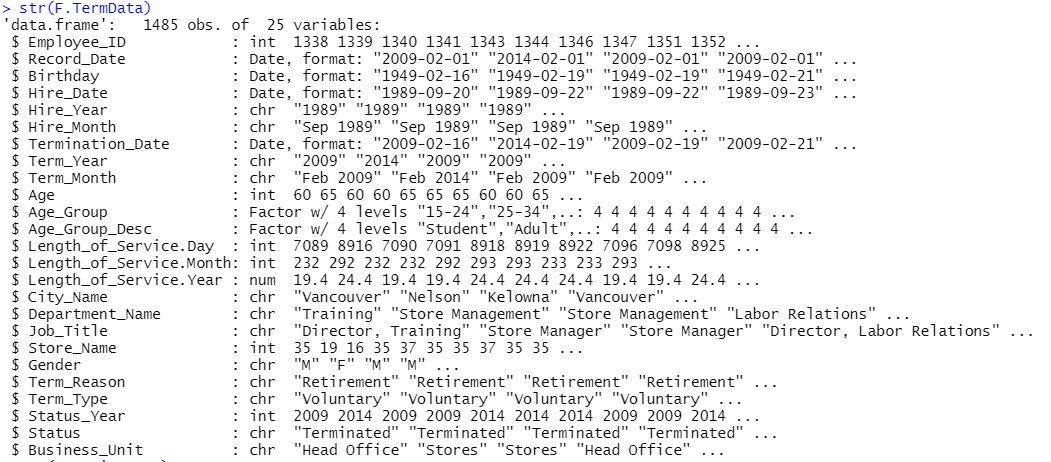


Figure 47: The final output after data transformation for Terminated employee dataset.

# **4.0 Data Manipulation & Data Visualization**

In this section, the two datasets will be analysed and manipulated with different R functions and visualized in different plots to show the analysis findings. Different analysis areas will also be stated in the following section followed with the code snippets of generating the analysis findings. Before that, here are some introduction on what and how to calculate Attrition Rate.

 Attrition Rate is commonly known as “churn rate”, and it is one of the most important metrics for a business to look into the problems that causes employees leaving the company. The attrition rate is calculated with the total number of people who have left the company, divided by the average number of employees over a period of time (*Figure 48*). For example, 500 employees is working in 2015 and 150 has left the company, and therefore the attrition is calculated as in *Figure 49*. Attrition rate can be calculated differently based on different types of attrition, which include Voluntary Attrition, Involuntary Attrition and Demographic-specific Attrition. With this metrics, the company can easily find the root cause influencing employees leaving the company, and hence retain and nurture employees the right way (Personio, 2021).

Figure 48: Attrition Rate Formula (Thakur, 2020).

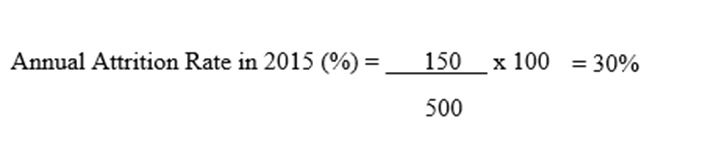


Figure 49: Example of calculating annual attrition rate.

## **4.1 Area 1: What is the Annual Attrition Rate?**

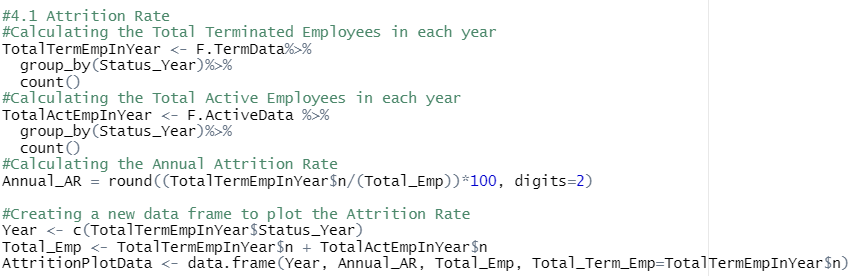
 The annual attrition rate is calculated based on the formula from *Figure 48*. First by calculating the total terminated and active employees from the two datasets that have been separated and cleaned in *3.5 Split Dataset* and *3.6 Remove Duplicate EmployeeID*, which is then group by Status Year, and count the number of employees in each year (*Figure 50*). After that, a new variable called Total\_Emp is created to store the total number of employees over the years by adding the total terminated employees with the total active employees. As a result, I am able to calculate the Annual Attrition Rate by using the total terminated employees divided by total employees, then multiply by 100%. The annual attrition rate is then round off to two digits and stored in a new variable called Annual\_AR. In order to plot a graph to show the annual attrition rate with year as the x axis, a data frame is created with the data values as shown in (*Figure 51*). In the end, the graph for “Number of Terminated Employees with Attrition Rate in 2006 – 2015” is plot with a line and bar chart with the values from the newly created data frame known as AttritionPlotData (*Figure 52 &Figure 53*).

Figure 50: The code for calculating the Annual Attrition Rate and creating a new data frame to plot a graph.

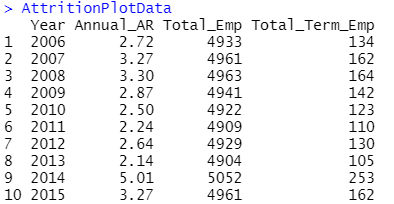


Figure 51: The data values that are stored in the newly created data frame.

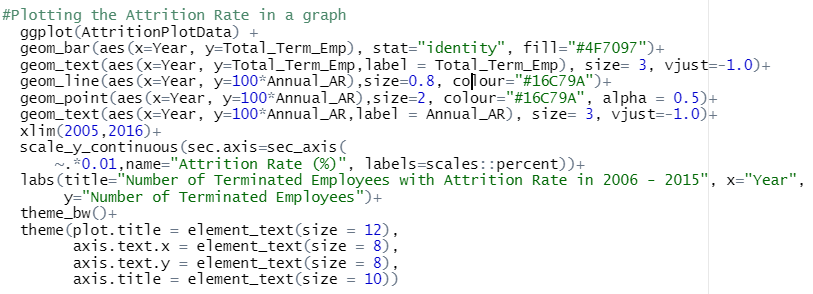


Figure 52: The code to plot a Line and Bar chart of the number of terminated employees in 2006 till 2015 with annual attrition rate (GeeksforGeeks, 2021).

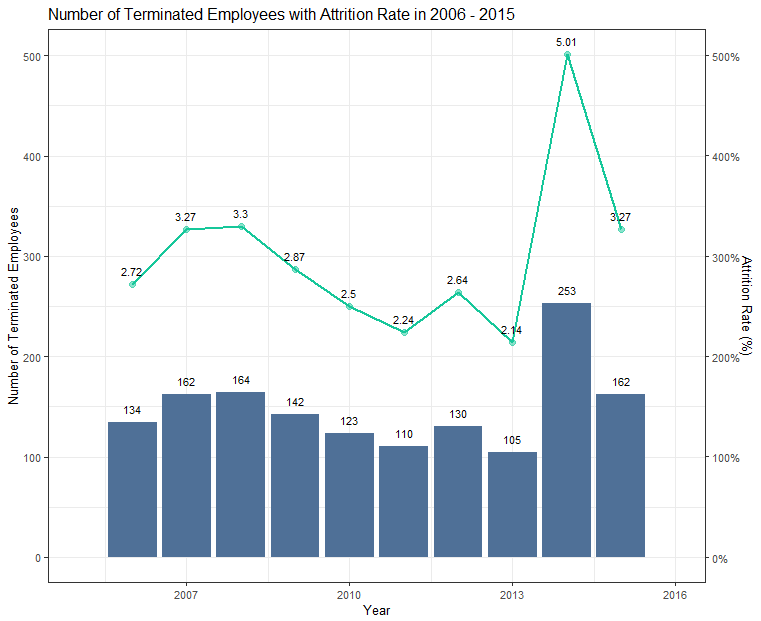


Figure 53: The Line and Bar chart of the Number of Terminated Employees in 2006 to 2015 with attrition rate.

From *Figure 53*, it shows that in **year 2014 has a spike of attrition**, which could be **involuntarily or voluntarily leaving the company**. After 2014, the number of employees who leave the company has a total of 162 employees, which is almost half of the number of employees who left the company in 2014. Therefore, we can come to a conclusion that in year 2014 something happened and it led to a spike of attrition rate, causing huge number of employees to leave the company at that time.

In this case, the year 2014 and year 2015 can be further discovered about the number of hired employees to find the potential impacts of sudden attrition. In short, most of the annual attrition rate is below 10 %, but there might be certain jobs that have high attrition rate that can be discovered and concerned.

### **4.1.1 Hired employees in 2014 and 2015**

According to (HR Profiling Solutions, 2021), impacts of high employee attrition includes lack of motivation, decreased in productivity and work quality, and has direct impacts on company profitability. Additionally, company will have to hire new employees which may also cost time and money, until then, existing staff will have huge workloads and responsibilities due to the lack of a trained workforce.

In order to find the total number of employees in year 2005 till year 2015, the original dataset is first cleaned to remove duplicate Employee IDs, then grouped by Hire Year and filter by the Hire Year which is larger and equal to 2005 (*Figure 54*).

As in *Figure 55*, it shows that **the company did not hire new employees to replace the workload of the terminated employees that left the company in year 2014** (*Figure 53*). This also shows that the company maybe having potential financial crisis of hiring new staffs, and hence some of the existing workers will be packed with huge amount of work. In this case, it also explains the meaning behind 3.27 % of attrition rate in 2015 (*Figure 53*), as some of the employees may feel burnt out and overworked, and lack of work-life balance due to the intensive amount of work to be accomplished every day. Another reason maybe because of the low salary payment, as the company have lost profitability due to a huge employee attrition, and hence, company may have to cut-cost on giving bonus to those existing employees.

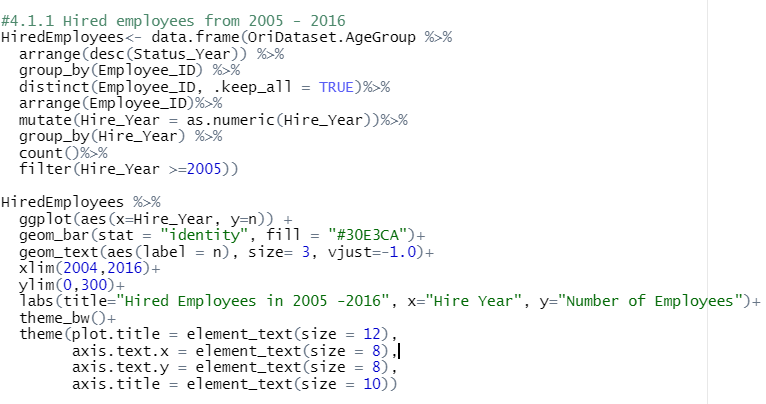
In conclusion, **existing employees in year 2014 and year 2015 will be affected by the sudden attrition in year 2014**, with the **feeling of overworked** and **a high chance of getting low paid salary**. Based on this, the termination reason will be analysed to find the reason of attrition in multivariate analysis.

Figure 54: The code of generating the bar chart of Hired employees from 2005 to 2016.



Figure 55: The Hired employees in year 2005 to year 2016.

## **4.2 Area 2: Reason of Attrition by Department and Job Title**

### **4.2.1 Termination Type by Termination Reason**

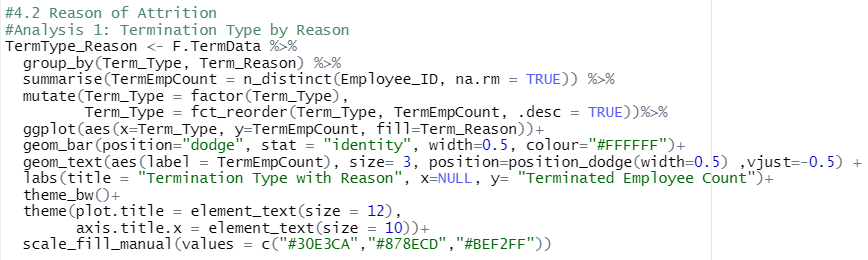
 In this section, the termination reason will be analysed by looking into different departments and jobs of the terminated employees. As in *Figure 57*, the types of termination can be categorized into Involuntary and Voluntary. The graph also shows that **reason of involuntary termination is Layoff**, whereas **reasons for voluntary termination are Retirement and Resignation.** Besides that, the bar chart also shows that **most of the terminated employees had chosen to leave the company because of retirement or resignation**, while there is a total of **215 employees who left the company involuntarily due to the reason of layoff**.

Figure 56: The code of finding the Termination type with Termination reason in a grouped bar chart.

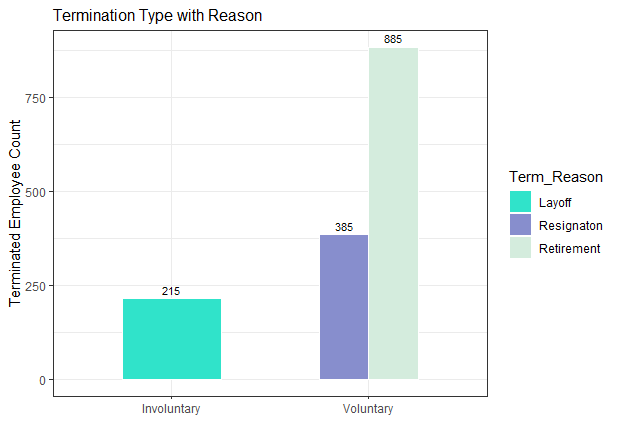


Figure 57: Termination type by Termination reason as a bar chart.

### **4.2.2 Termination Reason by Department Name**

The bar chart in *Figure 57* is then further analysed with the Department Name. In this analysis, **it will show which department has the highest attrition based on different termination reason**. In *Figure 58*, the Terminated Employee Dataset is grouped by termination reason and department name, and used n\_distinct() function to count the number of employees in each of the groups. After that, the department name is act as a factor to sort the order of the y-axis with fct\_reorder2(). Facet\_wrap() function is used to separate the graph into different termination reason for a clearer visualization. The two axis of the graph is then flipped with the coord\_flip() function and the final result is shown in *Figure 59*.

In *Figure 59*, it shows that **Customer Service department has the highest numbers of terminated employee with reasons of Layoff and Resignation**. **In Retirement, a total of 600 terminated employees were from Meats and Produce department.** Most of the employees were from departments that are related to serving customers in the store including store management. These department often need more workers as the workload is usually distributed and flowed down the organizational structure. Possible termination reason of these departments will be further analysed with the specific jobs in the following section.

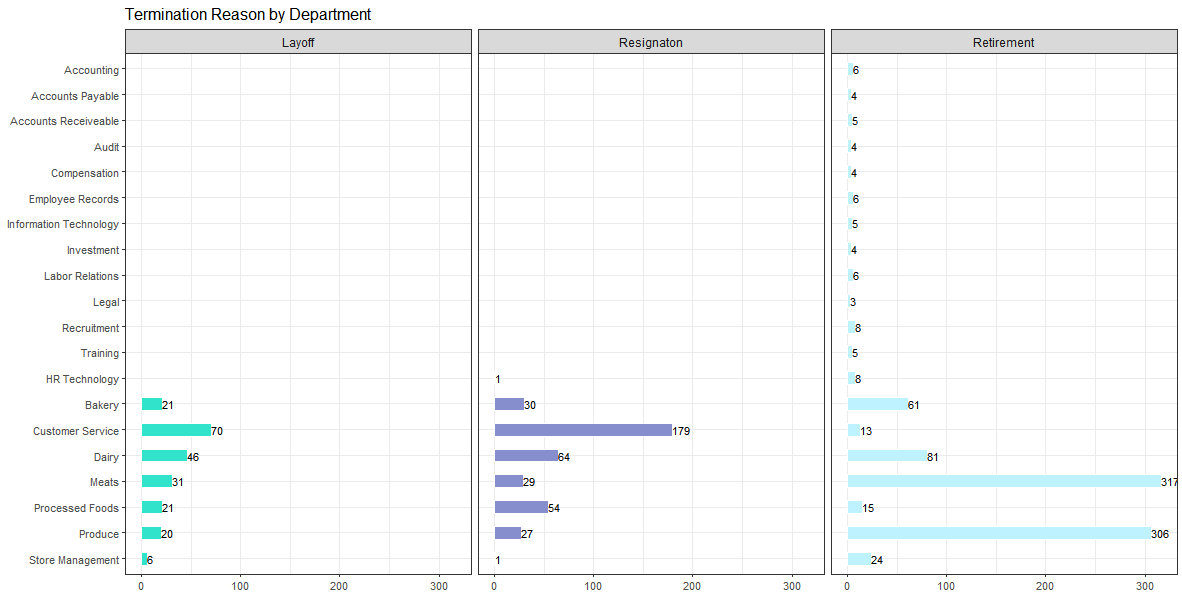
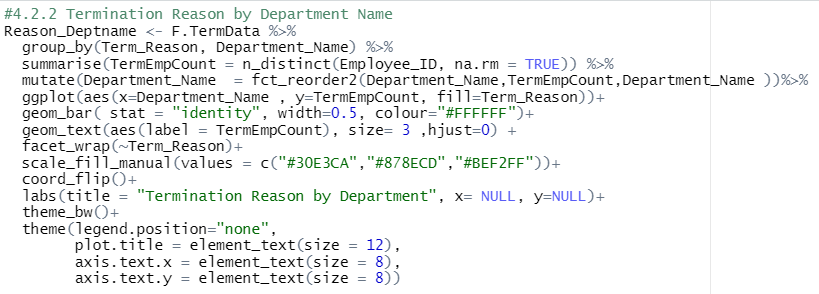
 

Figure 58: The code to find the termination reason by department name.

Figure 59: The Bar chart of Termination Reason by Department Name (R graph gallery, 2018).

### **4.2.3 Termination Reason by Top 5 Job Title with Department Name**

This section will continue the analysis of finding the relationship between termination reason and departments with specific job titles. The purpose of analysis is **to find the top 5 jobs of different departments and its termination reasons**.

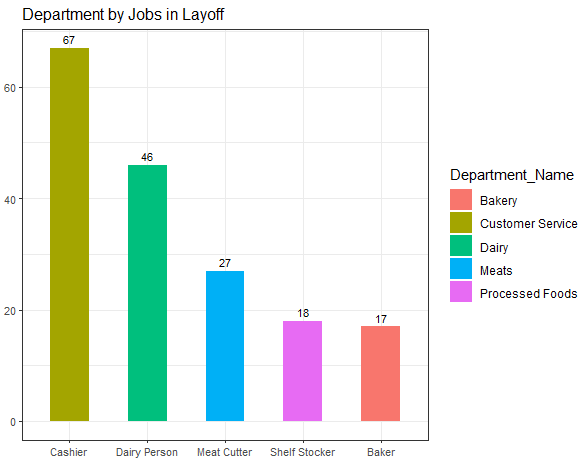
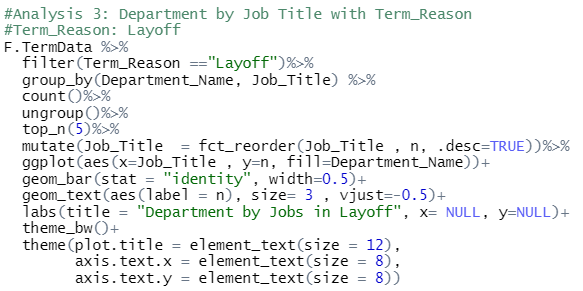
The terminated dataset were split into three different termination reason with the filter function to view a clearer visualization of the top 5 jobs with its department. The top\_n() function is used to select the top 5 job titles from the output dataset, and similar code (*Figure 60, Figure 62 & Figure 64*) were used to plot the three different bar chart (*Figure 61, Figure 63 & Figure 65*).

Figure 60: The code to find the Top 5 Jobs with department names in Layoff.

Figure 61: Bar chart of the top 5 jobs in Layoff with its Department name.

In *Figure 61*, it shows that the **top 5 jobs of laid-off employees are cashier, dairy person, meat cutter, shelf stocker and baker**. A total of 67 employees were asked to leave their job as a cashier under the customer service department, while other jobs were from departments related to production or retailing goods. For example, dairy person is to breed and milk the cattles, baker is responsible in making baked goods, and meat cutter is responsible to prepare standard cuts of meat to be sold to the customers.

In business terms, Layoff is a type of cost-saving action in business management where a company is facing financial difficulties or overstaffed problems. In some cases, an employer is able to recall laid-off employees to come back to work, which also means the layoff is temporary. However in most cases, these employees will be directly terminated based on the company situation. In conclusion to this analysis, **it shows that the company is indeed facing overstaffed problems** in the job positions from *Figure 61*, as based on the findings from *4.1.1 Hired employees in 2014 and 2015*, there isno hired employees after 2014, and this statement will be proven in the following analysis.

Another common reasons of people getting fired is about their **work performance** (Natalia, 2020). Lack of motivation and burnt out can lead to a decrease in the quality of work delivered. For example, being a cashier have to be well-mannered in serving customers but with low paid salary, while suffering health issues such as getting swollen feet because of standing too long, or having sleep problems due to working on different shifts (Howard, 2018).

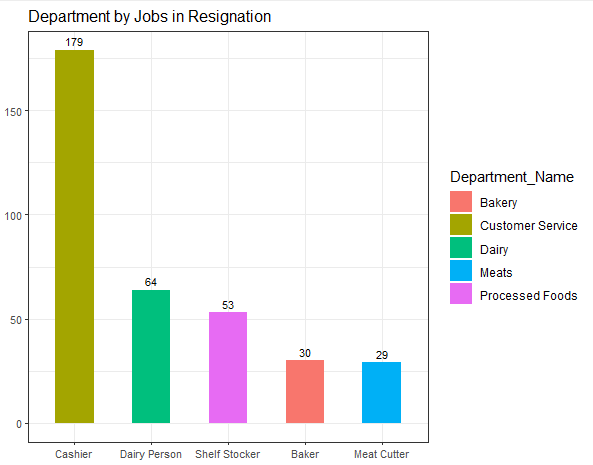


Figure 62: Bar chart of the top 5 jobs in Resignation with its Department name.

Similarly, the top 5 jobs in Resignation were as the same as the jobs in Layoff (*Figure 62*). However, **there is a total of 179 cashier employees chose to leave the company voluntarily in compared to involuntarily leaving in** *Figure 61*. In grocery industries, common reasons of attrition of **grocery store employees like cashier** are working of having **low paid salary**, **feeling unappriciated** and **little stimulation in responsiblities** (Martin, n.d.). High chances of attrition, especially cashiers will cause plague to companies in terms of having high costs to train new employees. According to (Ben, 2020) and (U.S. Bureau of Labor Statistics, 2021), it shows that some of the biggest grocery companies were desparate in retaining their minimum-wage workers due to the great amount of costs caused by huge number of cashier employees leaving the company. Some grocery companies like Costco and Walmart have committed in increasing the wages of low-paid workers to retain good employees and decrease the chances of attrition.

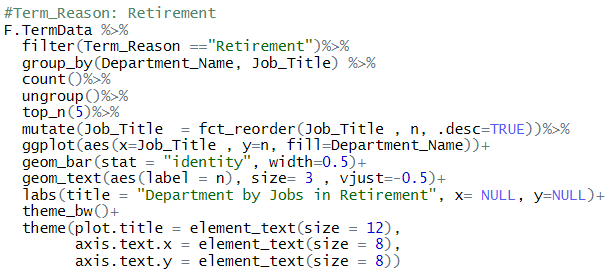
Apart from that, **store employees have been put at high risk situation of store robberies and sexual assault in the workplace**. From *A. Statistics of Robbery in British Columbia 2005 - 2014*, actual incidents of robberies have decreased drastically from 2005 to 2014, but raise 5.9% in year 2015, and most of the robbers were adults. This indicates that most probably the robbers may be armed with weapons or gun fire to threaten the store employees for robbing the grocery store. Other than that, according to a research in year 2014 about the incidents of sexual assault in workplace occurs in one-in-four Canadians (28%) and unwanted sexual contact at work occurs in one-in-seven Canadians (14%) (Angusreid, 2014). In conclusion, job insecurities of getting low paid, lack of growth and progression, being overworked and having the risks of getting robbed or sexually assualted by co-workers can be some of the reasons behind voluntarily attrition of these jobs.

Figure 63: The code to find the Top 5 Jobs with department names in Retirement.

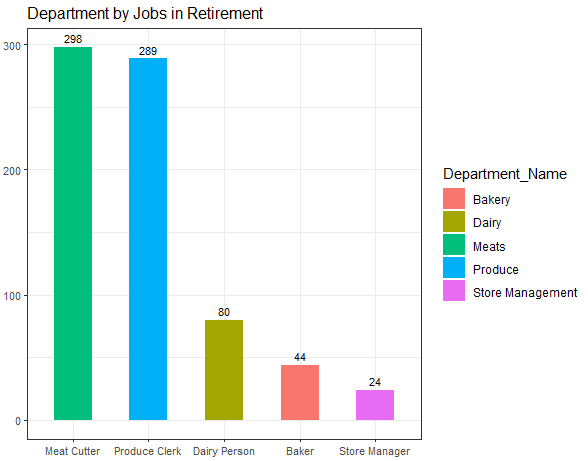
 **As for Retirement, the top 5 jobs are Meat Cutter, Produce Clerk, Diary Person, Baker and Store Manager** (*Figure 64*). Produce clerk is responsible to maintain the grocery store a clean and appealing state for its customers, like organizing items on shelves, removing expired products and etc. Store manager is responsible for daily operations of the grocery store by making sure it runs smoothly. **Employees who chose to retire were mostly Meat cutter and Produce Clerk**, with a total of 298 and 289 counts respectively.

Figure 64: Bar chart of the top 5 jobs in Retirement with its Department name.

**Retirement in the workforce is strongly related to ageing**. When a person gets old, it is possible that one person would suffer from illness that requires medical treatment physically and mentally to improve overall health status. Besides that, it is not safe for senior employees to work as heavy labour job positions, because there will be risks of getting injuries. For example, employees like meat cutter and produce clerk requires carring heavy objects, using sharp knifes to portion the meat, and climbing up the stairs to organize items on the shelves. In addition, **Canada is known as one of the most popular countries for senior citizens to stay in for retirement**. It is said that the employees in Canada have great retirement saving plans and tax-free saving accounts, and this allows senior employees to enjoy good life after retirement (Pinkasovitch, 2021).

## **4.3 Area 3: Analysis on jobs by Gender and Age**

### **4.3.1 Age Group by Gender**

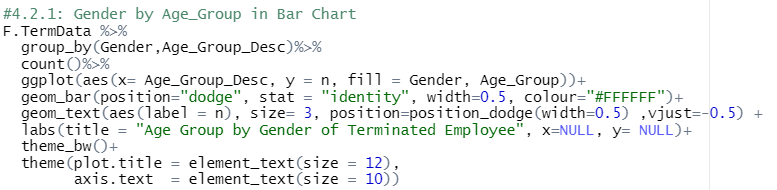
The age group of terminated employees were discovered in a bar chart with the gender attribute in *Figure 65*. In this analysis, the relationship between gender and age will be discovered and analysed. From *Figure 66,* it shows that **Female employees have a higher count in most of the age group**, while the count of **Male employees** were slightly higher in the **student age group, which is 15 to 24 years old**. This indicates that most of male student employees are prone in finding jobs during school holidays to earn some pocket money and to help support financials in the family.

Figure 65: The code of generating the bar chart of gender and age group.

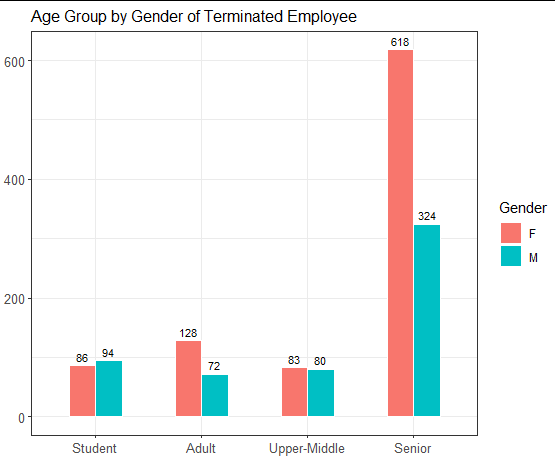


Figure 66: The bar chart of age group by gender of terminated employees.

### **4.3.2 Gender by Age**

The age of terminated employees were visualized with a histogram to view the distribution of the age of terminated employees grouped by gender. In *Figure 67*, the age attribute is plot as the x axis with geom\_histogram. The values in the x-axis can be adjusted with the breaks argument and the seq() function. The breaks argument is to set the cutoff points for the bins in a histogram plot, and the seq() function is to generate a sequence of numbers. In this case, the x-axis is manually adjusted starting from 0 to 70 years old, with an increment of 2 years apart. Then, the age is grouped by gender with the facet\_wrap() function, to view the age distribution of Male and Female employees.

As shown in *Figure 68*, **most of the terminated female employees were 60 years old and above with a total of 600 employees**, while most of the terminated male employees were in the range of 55 and 60 years old with around 300 employees. In both gender, the range of **20 to 25 years old** have the similar count of employees leaving the company. This also indicates that **students and fresh graduates may have better offers from another company** and would like to switch jobs to have a better pay to support the family financials.

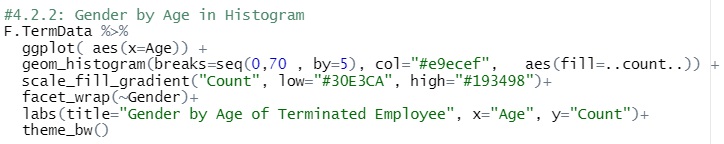
Additionally, **female employees at the age of 25 to 30 years old** are likely to quit their job **because of pregnancy**. According to (PregnancyInfo, n.d.), the average age of pregnancy have increased from 28 to 30 years old, which also means that more women are giving their first birth over the age of 30. As the pregnancy age increases, the chance of miscarriage and pregnancy complications will also increase. This proves that why there is sudden attrition at 25 to 30 years old in women employees, **where they might be a high chance of them are pregnant** and **they need to quit their jobs to take care of their health**. In this case, the graph can be further analysed with job titles and termination reasons.

Figure 67: The code for generating the histogram for age to view the distribution of the attribute by gender.



Figure 68: The Histogram of Gender by Age of Terminated Employees.

### **4.3.3 Gender and Age Group with Termination Reason**

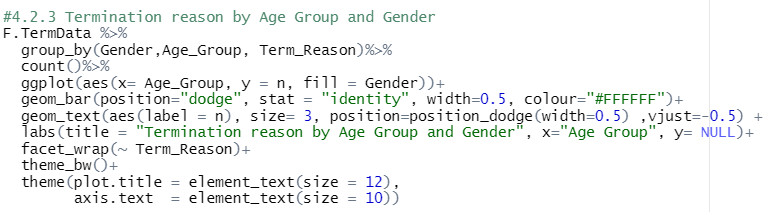
After looking into the relationship between gender and age group, multivariate analysis can be conducted on looking into the termination reason of each age group and gender attributes. Similarly, the termination dataset is grouped by gender, age group and termination reason, and use the count function to count each records within the groups. A bar chart is plot for this analysis with dodge position. This helps to separate the fill column to different bars, instead of stacking on to each other. Additionally, it is then separated in to three different termination reason with the facet\_wrap() function (*Figure 69*).

Figure 69: The code to find the relationship of termination reason by age group and gender.

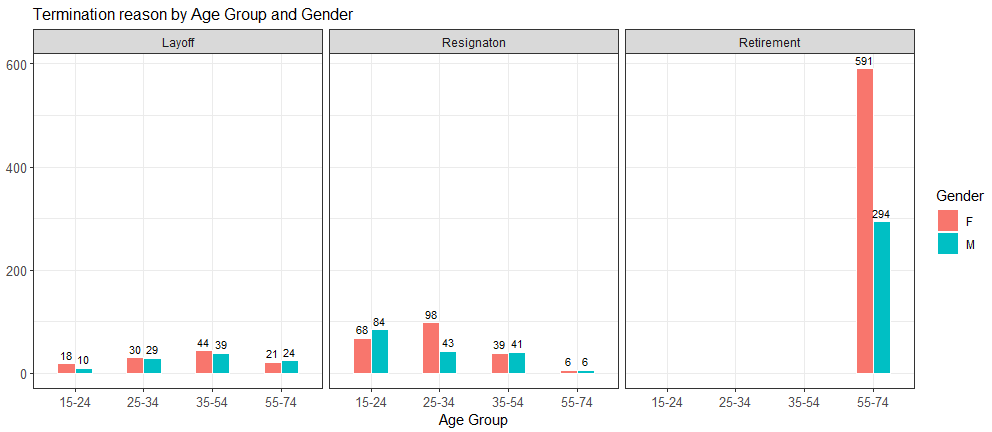
As shown in *Figure 70*, **employees who retire** from the company are mostly at the **age after 55 years old**, with a total count of 591 female employees and 294 male employees. For the **student age group**, **most of the employees chose to resign**, while **only parts of the employees were involuntarily leaving the company**.

Figure 70: The bar chart created to visualize the termination reason by age group and gender.

In addition, female employees at the age between 25 and 34 years old have a high chance of resignation, and male employees at the age between 15 and 24 years old have a high chance of resignation. In continuous to the previous analysis conclusion in *4.3.2 Gender by Age*, it is proofed that most of the female employees may be pregnant at age of 25 to 30 years old and hence they resigned to take care of their health. Additionally, it is also proved that both female and male student employees resigned because they may have received greater job opportunities with better higher paid salary.

### **4.3.4 Gender and Age with Job Titles**

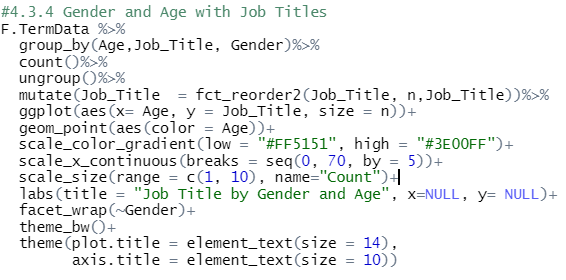
In this section, the job titles will be analysed with the gender and age attribute to find the job that has the highest attrition among different age and gender. This analysis will use a scatterplot to view the number of employees in different job titles, age and gender (*Figure 71*). In order to plot a scatterplot, the geom\_point() function from ggplot2 is used to plot the graph with age as the x-axis and job title as the y-axis. The size of each point will be the number of employees, and the colour of the point represents the age attribute with the argument color = Age. Scale\_color\_gradient() is used to change the colour of the gradient, by stating the low and high count for the age attribute. With the used of scale\_size(), the size of the point can be adjusted and the name of the legend can be modified. As a result, the scatterplot is generated as shown in *Figure 73*.

Figure 71: The code of analysis the job titles by gender and age.

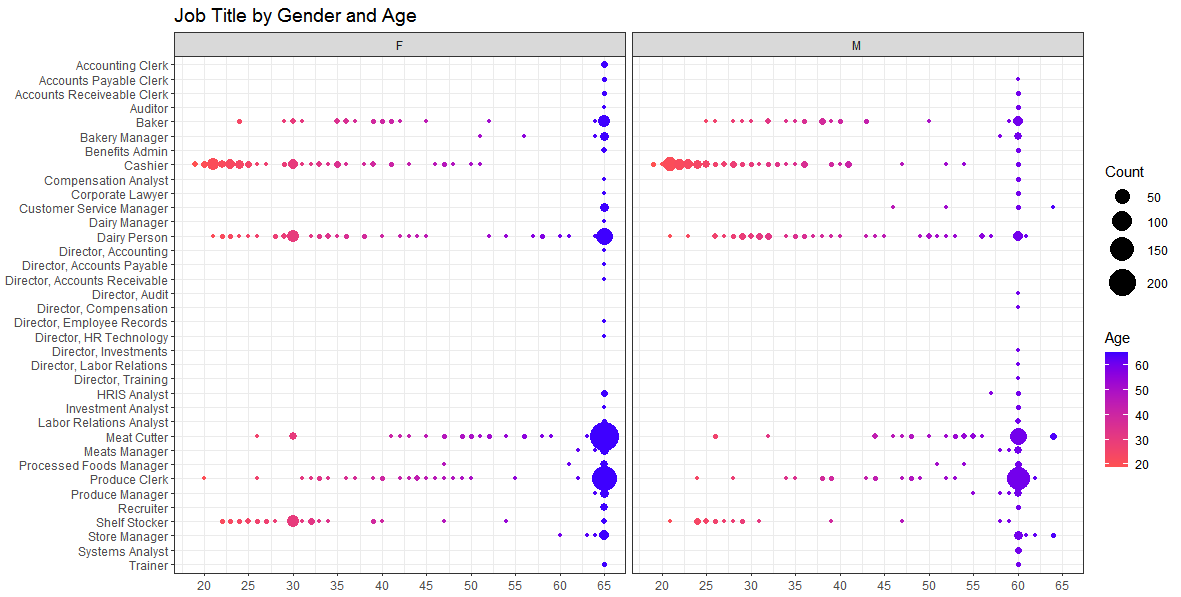
 As shown in *Figure 72*, job titles were visualized with gender and age of the employees. From the graph, it shows that **Meat cutter and Produce Clerk job positions are mostly 65 years old female**. Student age group are likely to find part-time jobs that are mostly trained on the spot, and from the analysis it shows that student employees are working as Cashier, Shelf Stocker and Dairy Person. Besides that, Female employees at the age of 30 years old will resign due to pregnancy from jobs like Meat Cutter, Shelf Stocker and Dairy Person, based on the findings from previous analysis in *4.3.3 Gender and Age Group with Termination Reason*.

Figure 72: The scatterplot chart of Job title by gender and age.

In conclusion, **most of the terminated employees were 65 years old female with resignation as the termination reason**. The **highest attrition rate** in terms of **jobs** **are Meat Cutter, Produce Clerk, Dairy Person, Baker, and Cashier**. Additionally, jobs like **cashier has a higher attrition rate** in the **student age group**, where mostly these employees are very likely to work as part-timer for a short period of time and have to continue study after the break.

## **4.4 Area 4: Analysis of different job types and its length of service**

### **4.4.1 Distribution of Length of Service**

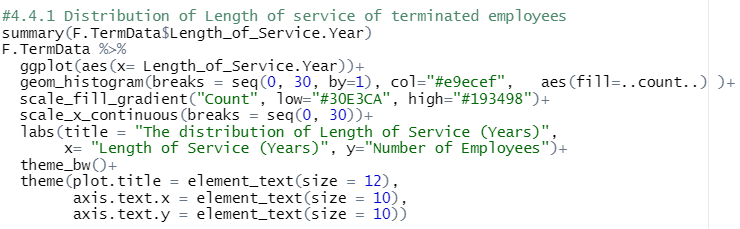
 The length of service is visualized and analysed by looking into the distribution of the data values with summary() function, and is shown in *Figure 74*. Apart from that, the length of service of terminated employees is plot in a histogram to view the full data values of the attribute (*Figure 75*). As a result, it shows that the longest length of service is 25.5 years and the shortest length of service is less than a month. The average length of service is calculated as 13.5 years across different types of job positions.

Figure 73: The code of finding the distribution of the Length of Service attribute.



Figure 74: The summary statistics of the Length of Service attribute.

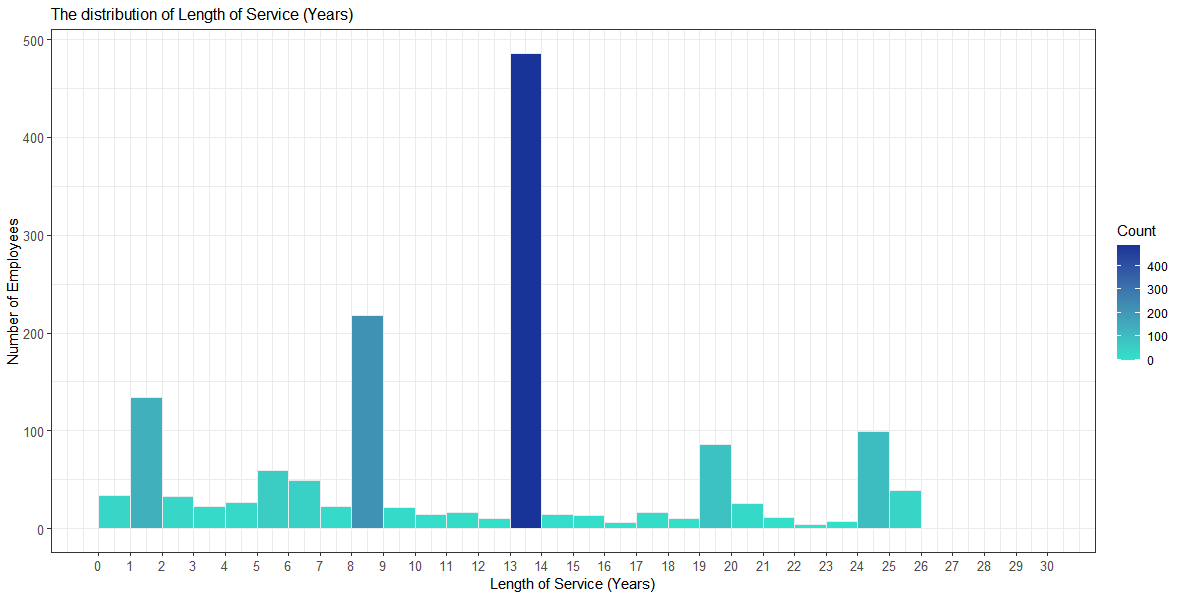
 As shown in *Figure 75*, **most of employees worked in the company for 12 to 13 years**. Besides that, **about 400 employees have worked in the company for 7 to 8 years**. Although the average length of service is 13 years, the histogram shows that the number of employees who **worked for less than 3 years** is quite high, which indicates that these **employees have left the company early to pursue other job opportunities**. In this case, the length of service can be further analysed with termination reason and job titles to find the relationship between these variables, such as which kind of jobs have the shortest or longest length of service in average, and in which year there is a high resignation in short-term employees.

Figure 75: The histogram of Length of Service for terminated employees.

### **4.4.2 Average Length of Service for different Departments and Jobs**

After analysing the distribution of length of service, this analysis will look into the average length of service in different departments and its job titles with the code in *Figure 76*. The average length of service is calculated with the mean() function and the decimal points of the output were adjusted with the round() function. With the use of geom\_bar() and fill by Job Title, the bar chart is created as shown in *Figure 77*.

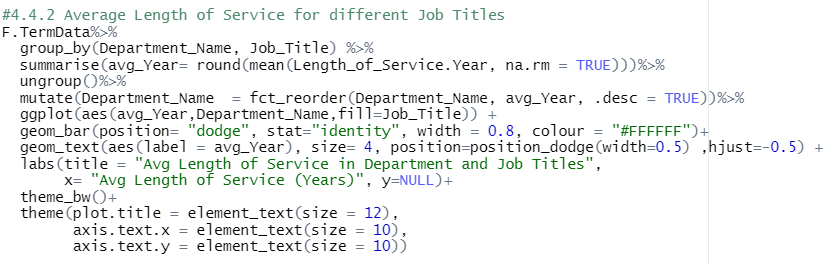


Figure 76: The code for finding the average length of service in different job titles.

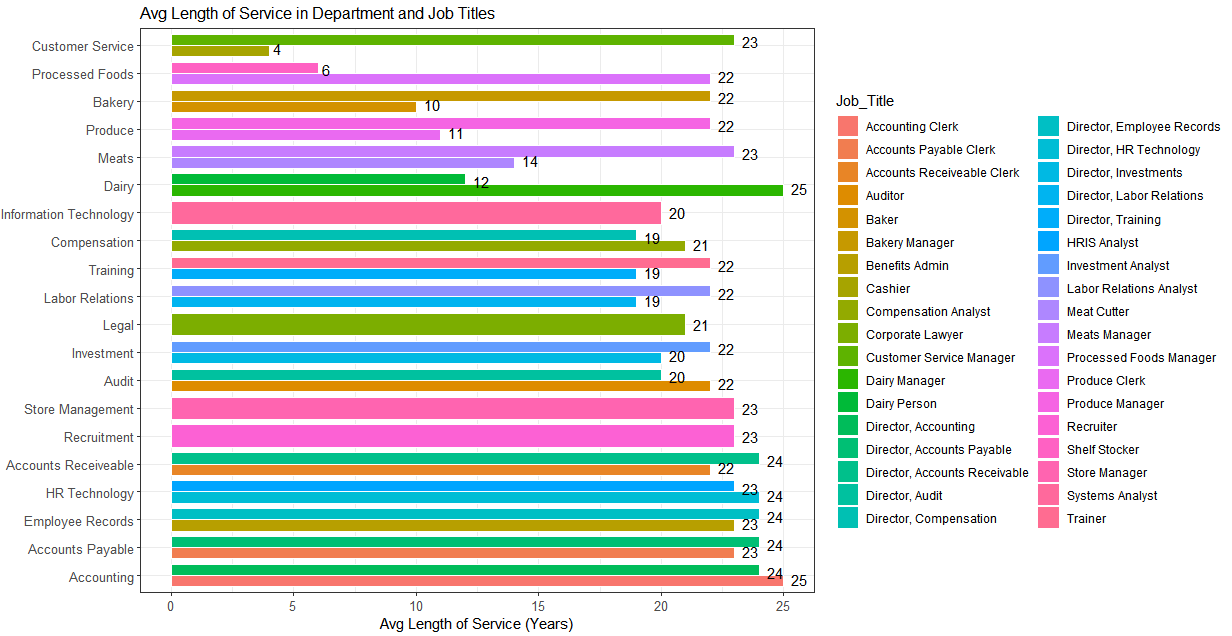


Figure 77: The bar chart of the Average Length of Service in Department and Job Titles.

From *Figure 77*, the bar chart shows that **cashier under customer service department has an average of 4 years of length of service**, this proves that most of the cashier workers are not likely to work for a long period of time. Based on the jobs that are identified in *4.2.3 Termination Reason by Top 5 Job Title with Department Name*, part of the employees worked for less than 10 to 12 years in average, like dairy person, shelf stocker, meat cutter and baker. Based on this analysis, we can conclude that most of the top-management employees have worked for an average of 20 to 25 years until they leave the company because of resignation or retirement, while only a small parts of the employees who get laid-off mostly worked for average of less than 10 to 12 years.

### **4.4.3 Length of Service and Age by Year and Termination Reason**

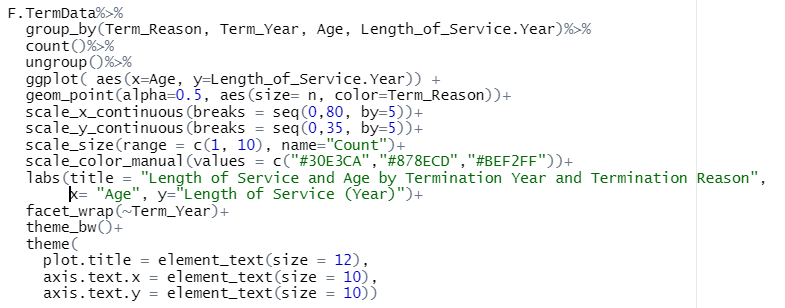
In this analysis, a bubble map is created with age as the x-axis and length of service in years as the y-axis, it is then grouped by termination reason and termination year (*Figure 78*). In this case, this analysis will also help to visualize the pattern of termination reason over the termination year with employees’ age and length of service. From that, a conclusion can be made by finding which age group has the shortest or longest length of service in years by viewing at the reasons of termination. A bubble map is created with different sizes of a coordinate by using geom\_point() function (). First, the size of the each bubble can be set with the argument size = n, and the colour of the bubbles can be set with color argument. The scale\_color\_manual() function is used to modify the colour of each bubble.

Figure 78: The code of generating a bubble map.

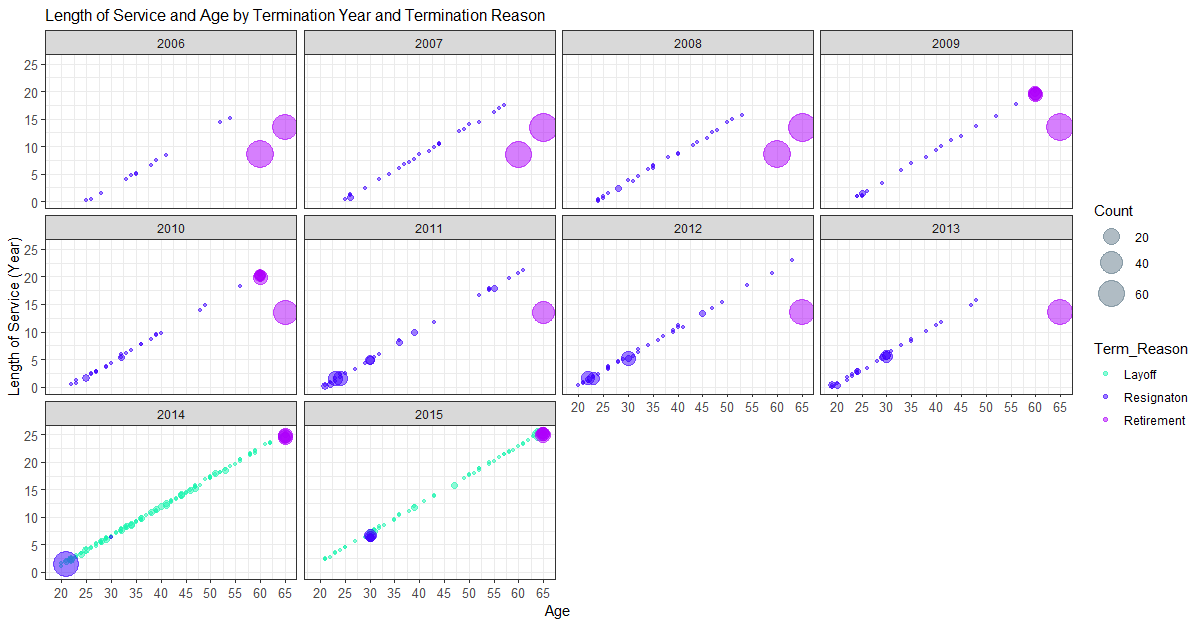
 From *Figure 79*, it shows a clear pattern of the relationship between length of service in years and age by looking into termination reason and separated into different years of termination. Starting from year 2006 until year 2015, the number of retire employees have decrease, but at the same time the length of service has increased up for up to 25 years. This means that some of the employees that were already 40 to 50 years old joined the company, and most of them served the company for only 15 to 20 years. However, in year 2014 and year 2015, employees who have worked for 25 years and above at the age of 60 years old and above chose to retire.

Figure 79: The bubble map of Length of Service (Years) and Age by Termination Year and Termination Reason.

Apart from that, resignation have increased from year 2006 to year 2014 and dropped at the year of 2015. From this graph, it shows that most of the employees who resign were 20 to 30 years old and have only serve the company for less than 5 years. This proves the findings from *Figure 62* in *4.2.3 Termination Reason by Top 5 Job Title with Department Name*, where the possible reason of attrition is because of low paid salary, feeling unappreciated and lack of growth and progression in taking on other job responsibilities, overworked and other potential risks of unpredictable incidents, and hence there is a high chance of leave for a better paid and safer job.

In addition, layoff only happens in year 2014 and year 2015, where huge number of employees were leaving invonluntarily at any age and any years of service. Based on the findings from *Figure 55*, it is proven that the company is indeed facing problems of overstaffed and financial shortage, because there is no new employees were hired after a huge number of layoff from the company. In this case, the jobs that most probably been laid off is stated in the findings from *Figure 61*.

In conclusion, the company should offer better-than-average pay and other employee benefits to retain employees in leaving the company. According to a survey by Glassdoor, 60% of employees think that employee benefits are a major factor in accepting the job offer, and 80% of employees would choose additional benefits over a pay raise. Therefore, it also shows the importance of having employee benefits in showing apprieciation and care to those well-performed employees. Examples of the must have employee benefits are healthcare insurance, holidays, performance bonus, paid leave and retirement plan. Additionally, productivity and engagement are key factors that affects work performance in the workplace. Employers can increase productivity by providing greater leaning oppourtinities with the ability to progress and grow in differnet positions in the company. In this case, employees will feel appreciated and supported in serving the company with high productivity and effort. With great benefits and higher pay than average salary and have the opportunity to grow in the company, the company is able to retain well-performance employee and increase company profitablility at the same time.

## **4.5 Area 5: Analysis of the Job Environment**

### **4.5.1 City Name and Store Name**

In this area of analysis, the job environment will be analysed with attributes such as city name, business units, store name and its termination reason and job positions. First, city name attribute is visualize with the use of a treemap. Before that, the treemapify library is installed and loaded to plot a treemap with geom\_treemap() in *Figure 80*.



Figure 80: The code in generating a treemap to view the city name with store name (Wilkins, 2021).

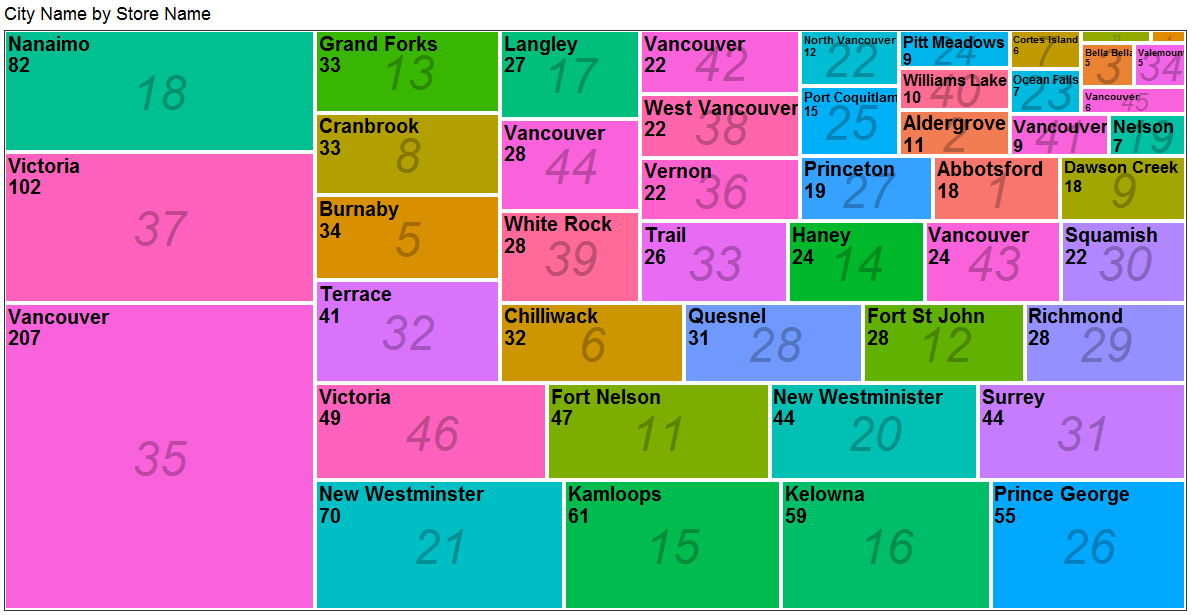
As shown in *Figure 81*, the city name and store name is visualized in a treemap with the number of terminated employees in different size of the rectangles. From the findings, **Vancouver has a total of 207 terminated employees** **and the store name is 35**, Victoria and Nanaimo has 102 and 18 terminated employees respectively. Based on the data exploratory in *Figure 23*, all these cities are from British Columbia, Canada, and Vancouver is one of a major city in Canada and it has the third largest population compared to other cities in Canada (Macrotrends, 2021). In short, the larger the population, the more job opportunities and workers there are in the city, but at the same time the competition of getting a job will also be high.

Figure 81: The tree map of city name and store name.

### **4.5.2 City Name and Department by Business Unit**

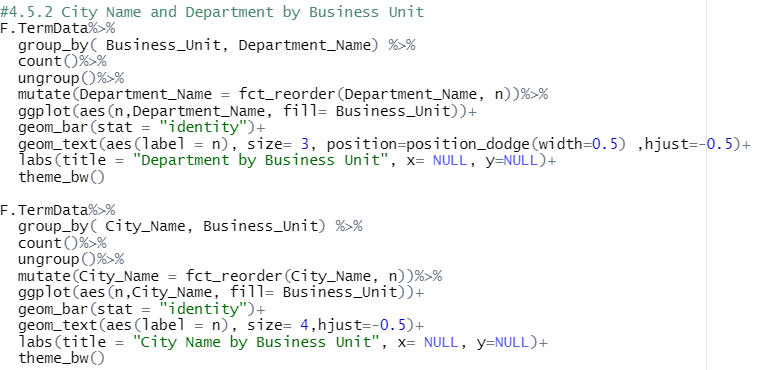
 In this analysis the city name and department will be further analysed with business unit to view the distribution of the organizational structure of the company. In *Figure 82* , there will be a bar chart to visualize the relationship of department and business unit (*Figure 83*), and another bar chart to visualize the relationship of city name and business unit (*Figure 84*) with the used of geom\_bar().

Figure 82: The code in generating city name and department by business unit.

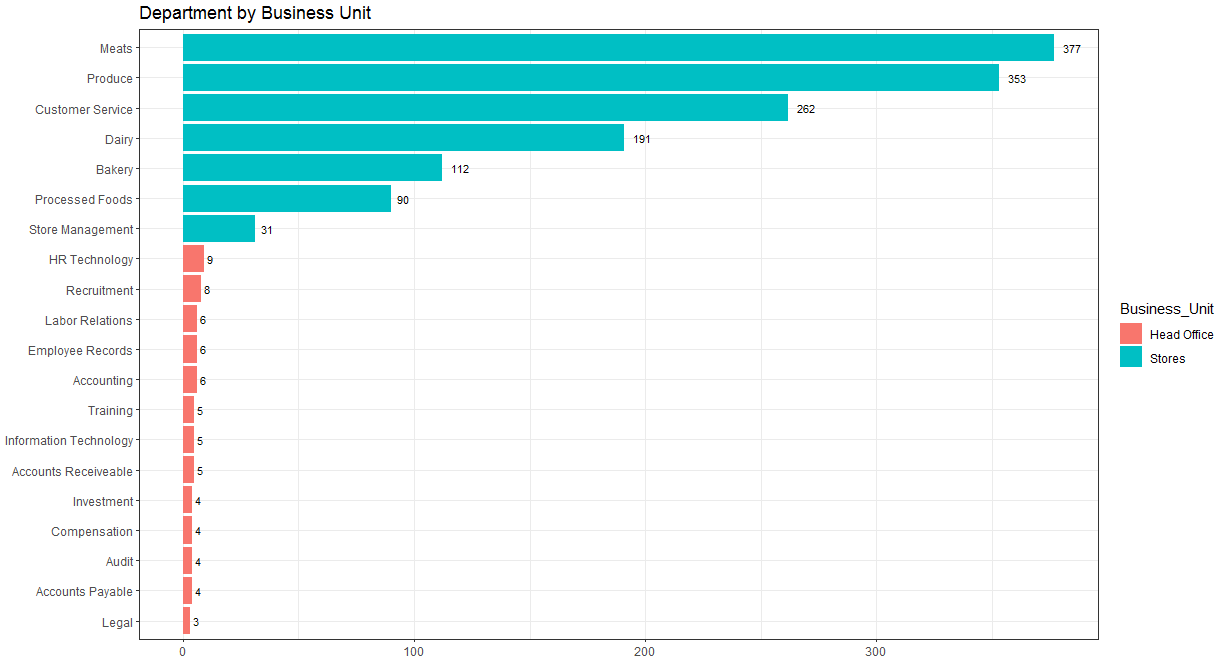
From *Figure 83*, it shows that the business unit of head office is departments of top-level management like HR, Recruitment, Information Technology, Training, Accounts & Audit, Labour Relations and etc. Departments in stores are Meats, Produce, Customer Service, Bakery, Dairy, Processed Food and Store Management. This also shows that there are more job positions in departments in stores and there are less jobs in management level, and each department has its own purpose in handling different aspect of the business operation. Positions in Head Office has larger responsibility as they are accountable in managing different departments and different employees. For example, departments like HR, recruitment, and training is responsible in recruiting and train new employees in the workforce especially the employees in store management. HR employers train store managers, and store managers are accountable and responsible in managing all grocery store employees. In short, this analysis shows a better view of the functional areas in the organization.

Figure 83: The bar chart of department name by business unit.

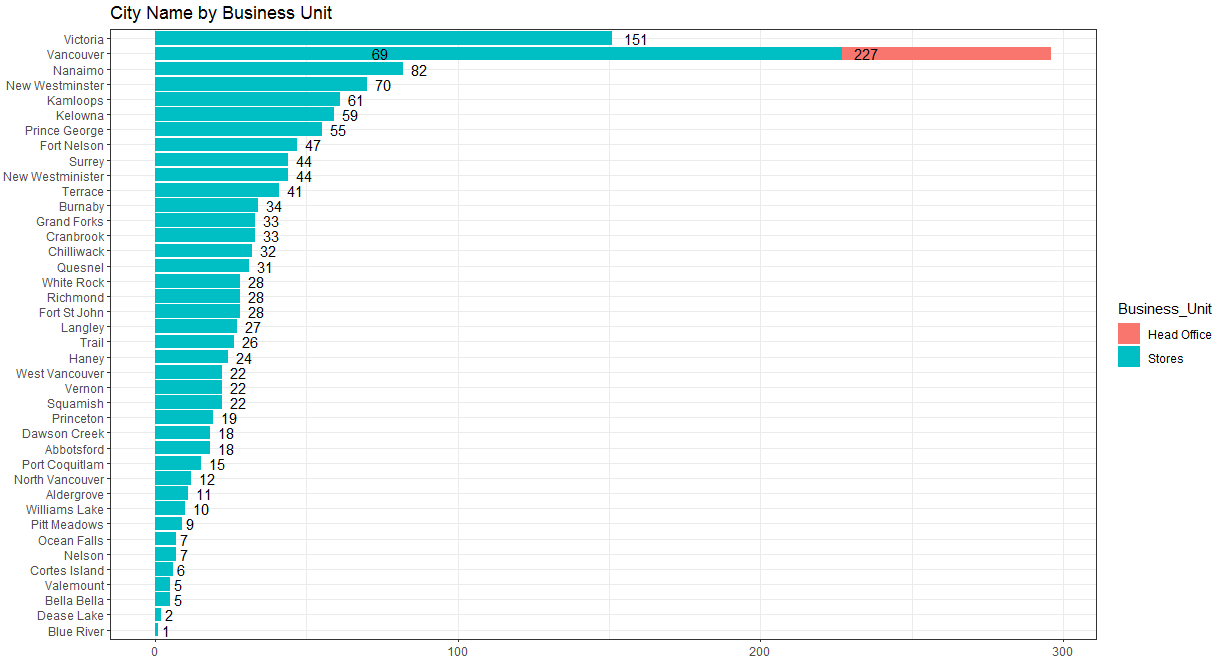


Figure 84: The bar chart of city name by business unit.

From *Figure 84*, it clearly shows that the only city that has head offices is Vancouver, with a total of 227 terminated top level management from year 2006 to year 2015. This also indicate that the head-quarter office of this company is located at the heart of British Columbia, which is store number 35, Vancouver, with a total of 227 terminated employees as top-level management, and a total of 69 grocery store employees. From this, the findings will be concluded in the next analysis, which is to find the relationship between city name and termination reason.

### **4.5.3 City Name by Termination Reason**

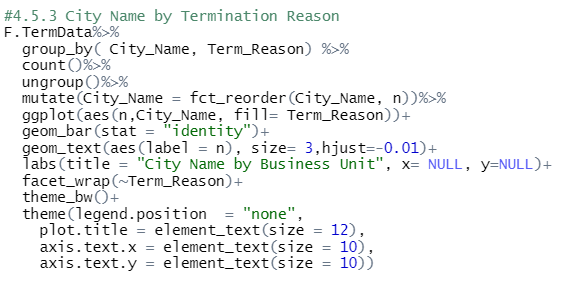
 In this analysis, the city name will be analysed with the termination reason with the code in *Figure 85* and the result is plot as a bar chart (*Figure 86*) with facet\_wrap(~Term\_Reason) to view the number of employees in each city and its reason of leaving.

Figure 85: The code in generating the bar chart of city name by termination reason.

As a result, it shows that Fort Nelson city has the highest number of laid-off employees, and Vancouver city has the highest number of resigned and retired employees (*Figure 86*). These two cities will be drilled down in the following section, in order to find the respective job titles that have been laid-off, resigned and retired.

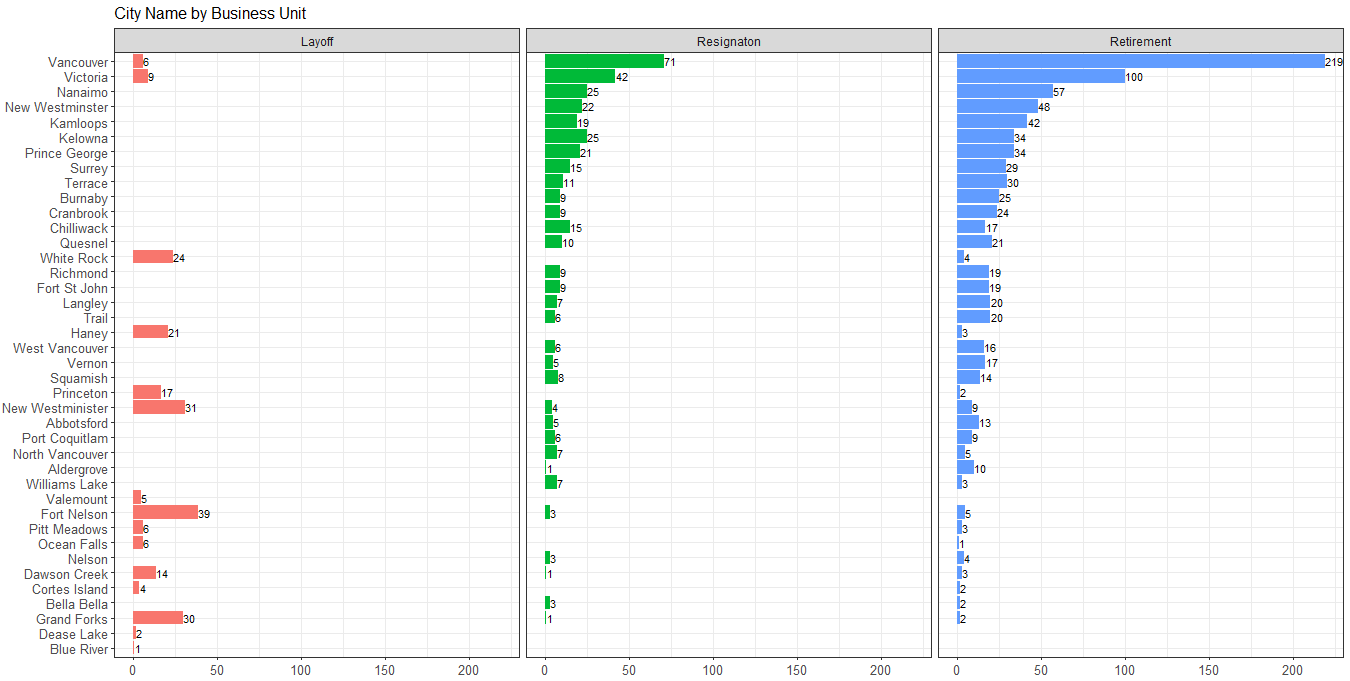
As a result, it shows that Fort Nelson city has the highest number of laid-off employees, and Vancouver city has the highest number of resigned and retired employees (*Figure 86*). These two cities will be drilled down in the following section, in order to find the respective job titles that have been laid-off, resigned and retired.

Figure 86: The bar chart of city name by termination reason.

### ***4.5.3.1 Drill down on Fort Nelson with laid-off Job Titles***

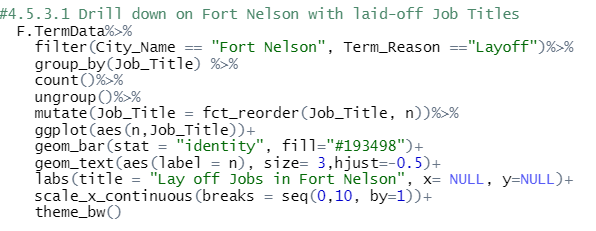
 First is to drill down on the Fort Nelson city to view the laid-off job titles with the code in *Figure 87*, and as a result a bar chart is plot in *Figure 88* .

Figure 87: The drill down code of looking into the laid-off jobs in Fort Nelson city.

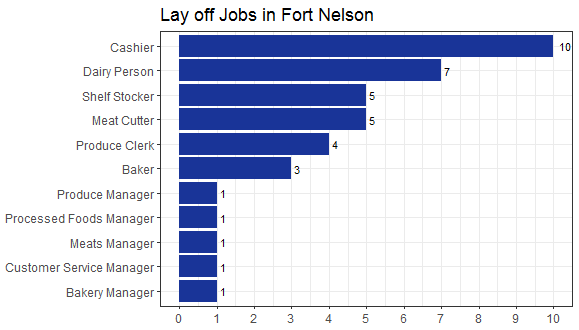
 As shown in *Figure 88*, the most laid-off jobs in this company is Cashier, Dairy Person, Shelf Stocker, and Meat Cutter, and this indicates that laid-off employees have contributed highly in the findings from *Figure 59* and *Figure 61*. To conclude the findings on Layoff terminated employees, they are mostly in the upper-middle age group, and they are mostly work as grocery store employees like cashier and dairy person. They were involuntarily terminated in year 2014 and year 2015, and most of the laid-off employees were from Fort Nelson city and they were Female employees.

Figure 88: The laid-off job titles in Fort Nelson city.

### ***4.5.3.2 Drill down on Vancouver with retired and resigned Job Titles***

In this analysis, it is to drill down on the Vancouver city to view the retired and resigned job titles based on the findings from *Figure 86*. This analysis is conducted with the code in *Figure 89* and the output is shown in *Figure 90*.

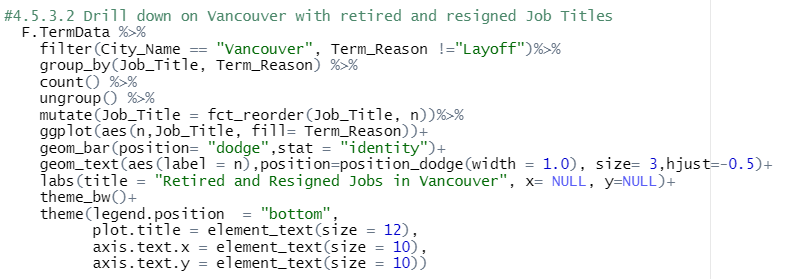


Figure 89: The drill down code of looking into the retired and resigned jobs in Vancouver city.

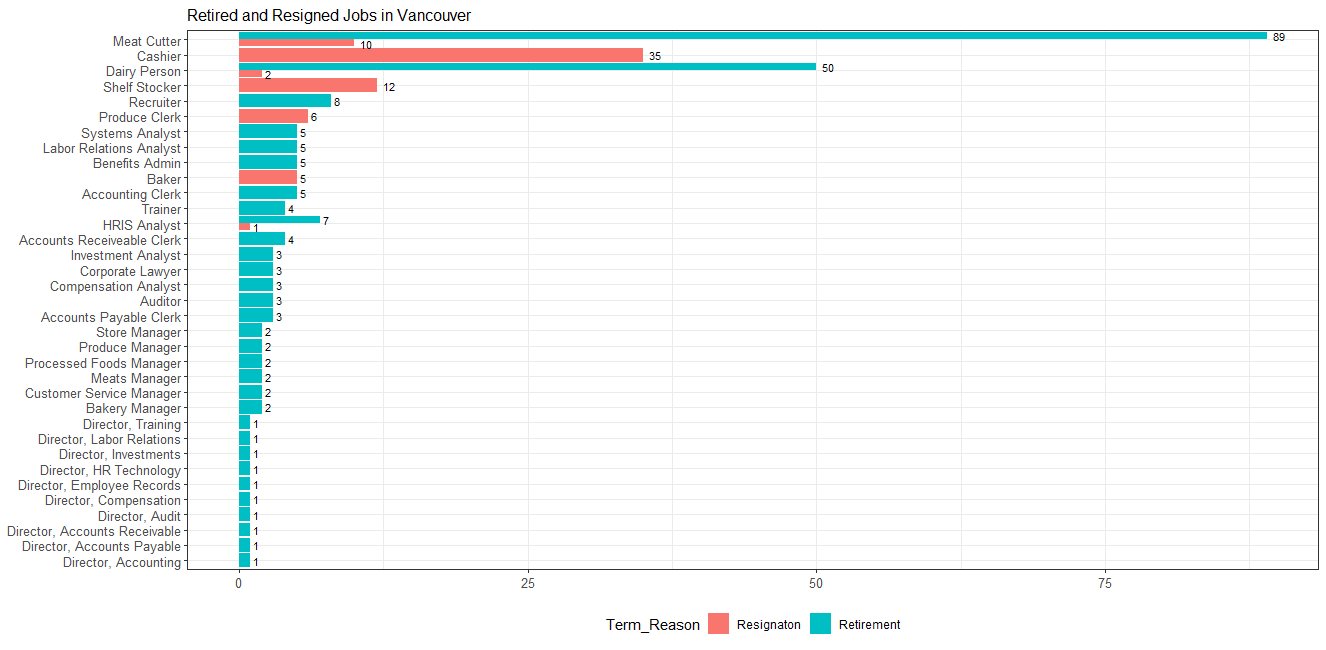
As shown in Figure 90, meat cutter contributed most of the retired employees in Vancouver with a total of 89 employees. Apart from that, a 35 cashier, 12 shelf stocker, 6 produce clerk and 5 bakers have all resigned from this city. In this graph, most of the employees were leaving the company because of retirement, and this indicates that these employees are mostly Female, age 60 years old and above and have 15 to 25 years of working experience in this company as based on the findings from *4.4.3 Length of Service and Age by Year and Termination Reason*. In addition, employees resign from the company are mostly from Vancouver city, and are mostly at the age between 20 to 30 years old. Most of the resigned employees are female and most of them have less than 10 years of working experience.

Figure 90: The retired and resigned job titles in Vancouver city.

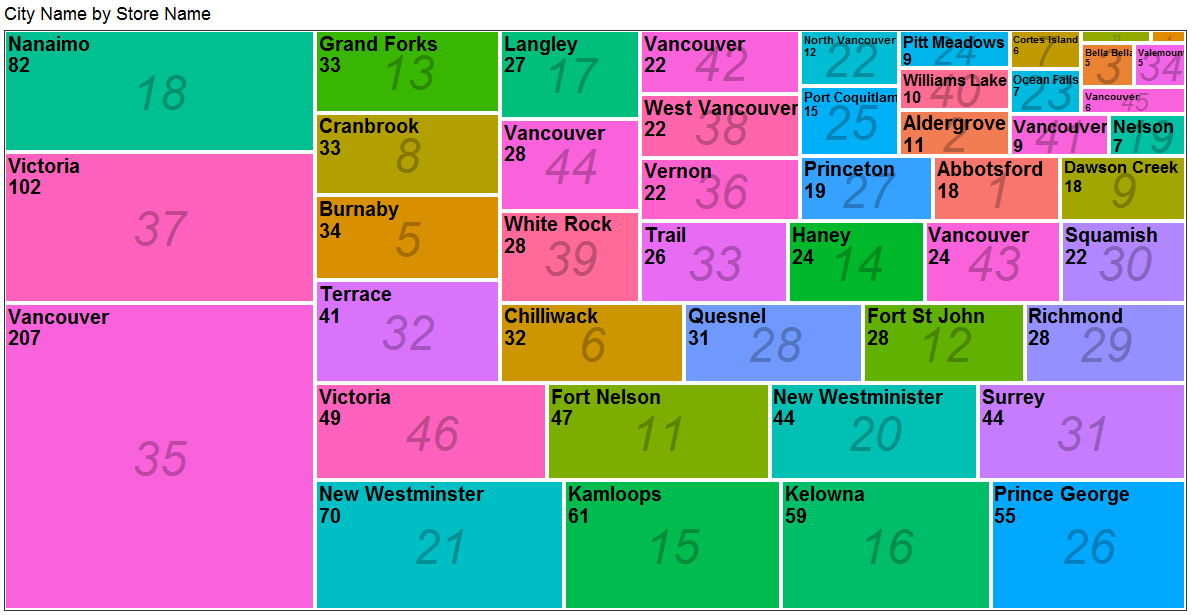
# **5.0 Extra Features**

## **5.1 Treemap with Treemapify package**

One of the extra features used in this analysis in the treemap from treemapify library (*Figure 91*). The treemap is used in the analysis of visualizing the city name and its store name with the total number of terminated employees in different sizes of rectangles (*Figure 81*). One of the main advantages of a treemap is that it allows to visualize hierarchical data in the form of nested rectangles. It can often be seen as a tree structure where the main group will be the coloured rectangles, and the subgroup will be contained inside the rectangles as leaves of the tree structure. Main benefits of using the treemap to view the categorical data is that it can optimize the use of space than bar chart with the ability of showing two categorical data values in the graph. Other than that, the treemap is able to display multiple elements together with the use of size and colour to have better visual effects in viewing the graph (Laubheimer , 2019).

Figure 91: The code in generating a treemap to view the city name with store name (Wilkins, 2021).

Figure 92: The tree map of city name and store name.



In *4.5.1 City Name and Store Name*, city name attribute is visualize with the use of a treemap. Before that, the treemapify library is installed and loaded to plot a treemap with geom\_treemap() in *Figure 80*. As shown in *Figure 81*,**Vancouver has a total of 207 terminated employees** **and the store name is 35**, Victoria and Nanaimo has 102 and 18 terminated employees respectively. With the use of the treemap, the two categorical data is clearly visualize with different aesthetics, like adding store name as subgroup with the subgroup argument, geom\_treemap\_text() allows to change the size of the main group, and geom\_subgroup\_border() is to adjust the border of each subgroup.

## **5.2 Fct\_reorder() in Forcats package**

The fct\_reorder() function to reorder a the values in ascending or descending form. This function is from the forcats package where it is created to solve common problems with factors or categorical variables (Robinson, n.d.). In this analysis, the fct\_reorder() function is mostly used in reordering the data values or x and y axis of a bar chart. For example in *4.2.3 Termination Reason by Top 5 Job Title with Department Name*, the function is used to reorder the values in a descending order to show the top 5 job titles with its department name.

First, the forcats package is installed and loaded in rstudio, before running the coding in (*Figure 93*). Then, the terminated dataset were filtered by termination reason equals to lay off with the filter function, andit is grouped by department name and job title, and the number of employees of each group were counted with the count function. After that, the top\_n function is used to select the top 5 job titles with its department from the output dataset. In order to plot the bar chart in a descending order, the values in job title were act as a factor for sorting the number of employees in a descending order with fct\_reorder() function. The “.desc” argument is stated to be true to sort the values in descending order.

In *Figure 94*, it shows that the **top 5 jobs of laid-off employees are cashier, dairy person, meat cutter, shelf stocker and baker**. A total of 67 employees were asked to leave their job as a cashier under the customer service department, while other jobs were from departments related to production or retailing goods. For example, dairy person is to breed and milk the cattles, baker is responsible in making baked goods, and meat cutter is responsible to prepare standard cuts of meat to be sold to the customers. With the use of fct\_reorder() function and the forcats library, the appearance of the graph can be improved and common factor problems can be solved.

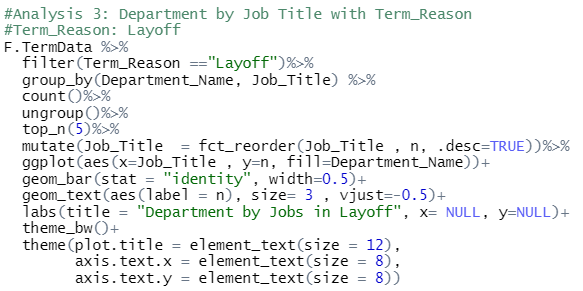


Figure 93: The code to find the Top 5 Jobs with department names in Layoff.

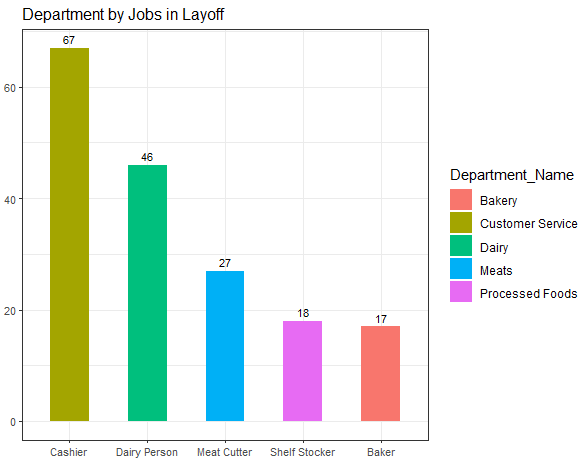


Figure 94: Bar chart of the top 5 jobs in Layoff with its Department name.

## **5.3 Cut() and Difftime() function**

Other than that, extra features such as using different built-in function like cut() and difftime() were used in this analysis. The cut() function is used to create age group where it is useful in categorizing values of a continuous attribute like Age into differnet levels of a factor in *3.4 Create Age Group*. The break argument is to specify the number of interval levels in separating the age values into different categories, and the labels argument is to label the categorical levels (Gonzalez, 2021). In *Figure 95*, Age\_Group is created by breaking the age attribute into 4 different levels, which are 15-24 is Student, 25-34 is Adult, 35-54 is Upper-Middle and 55 and above are Senior employees.

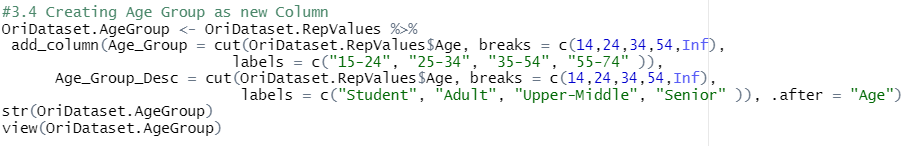


Figure 95: Age\_Group is added to classify the employees age into different Age Group.

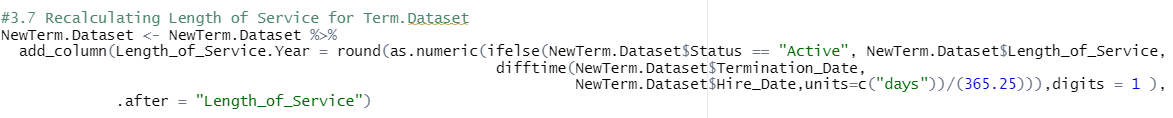
 Apart from that, the length of service is recalculated into total years with months with the difftime() function in *3.7 Recalculate Length of Service*. Difftime allows us to find the total days between two date columns by stating “days” in the units arguments. After finding the total days, the total years is calculated by dividing 365.12 days and it is round off to 1 decimal point (*Figure 96*). In short, these a new column is created as the Length\_of\_Service.Year and it is added to the Terminated Employee Dataset to show the length of service in a more detailed perspective based on days, months and years.

Figure 96: Recalculating Length of Service in Termination Dataset (DataScience Made Simple, 2021).

# **6.0 Conclusion**

In conclusion, the main goal of this analysis have been accomplished by conducting multivariate analysis in looking into different relationship of the attributes to **analyse the potential reasons of attrition**. Based on the findings, it shows that there is a sudden attrition rate in year 2014 and this is because of overstaffed problems or the company is facing financial shortage in supporting around 5000 employees per year. This is proven that there is no hiring in the year 2014 and 2015. One of the main reason of resignation is because the salary is low paid, employees feel unappreciated and there is no space for progression, as most of the jobs in resignation are short-term like cashier. In this case, the company is highly recommended to increase the salary of low paid jobs and provide more employee benefits like healthcare insurance to grocery store employees.

In addition, most of the employees chose to retire at the age of 60 years with working experience around 15 to 25 years in the company. From this statement, it shows that some of these employees shows hard work at their workplace and are willing to work at such an old age until they retire. Whereas, for employees who chose to resign and leave the company are mostly 20 to 30 years old with less than 10 years of working experience in the company. This shows that resigned employees are prune to choose for a better job opportunity with high paid salary after they graduate to support their own financials. Additionally, with the use of extra feature and functions, attributes like age group and newly calculated length of service were generated, and appearance of the graphs were more appealing and easy to ready.

# **References**

Angusreid. (2014, December). *ARI Sex Harassment at work Dec 2014*. Retrieved from Angusreid.org Web site: https://angusreid.org/wp-content/uploads/2014/12/2014.12.05-Sexual-Harassment-at-work.pdf

Ben. (2020, February 23). *Why High Cashier Turnover Rates Can Plague Companies*. Retrieved from Ninja Gig Blog Web site: https://www.ninjagig.com/blog/why-high-cashier-turnover-rates-can-plague-companies/

DataCamp Team. (2020, June 27). *Subsetting in R*. Retrieved from DataCamp Web site: https://www.datacamp.com/community/tutorials/subsets-in-r?utm\_source=adwords\_ppc&utm\_medium=cpc&utm\_campaignid=14989519638&utm\_adgroupid=127836677279&utm\_device=c&utm\_keyword=&utm\_matchtype=b&utm\_network=g&utm\_adpostion=&utm\_creative=332602034364&utm\_targ

DataScience Made Simple. (2021). *DIFFERENCE BETWEEN TWO DATES IN R BY DAYS, WEEKS, MONTHS AND YEARS*. Retrieved from DataScience Made Simple Web site: https://www.datasciencemadesimple.com/get-difference-between-two-dates-in-r-by-days-weeks-months-and-years-r-2/

GeeksforGeeks. (2021, June 24). *Combine bar and line chart in ggplot2 in R*. Retrieved from GeeksforGeeks Web site: https://www.geeksforgeeks.org/combine-bar-and-line-chart-in-ggplot2-in-r/

Gonzalez, J. C. (2021). *Cut in R*. Retrieved from R Coder Web site: https://r-coder.com/cut-r/

Howard, F. (2018, November 28). *Risks of Being a Cashier*. Retrieved from Career Trend Web site: https://careertrend.com/about-6573375-job-description--cashier-customer-service.html

HR Profiling Solutions. (2021). *Employee Turnover: Causes, Effects, and Strategies*. Retrieved from HR Profiling Solutions Web site: https://www.hrprofilingsolutions.com.au/blogs/aus-blog/employee-turnover-causes-effects-strategies/

Jozef. (2019, June 22). *How data.table's fread can save you a lot of time and memory, and take input from shell commands*. Retrieved from Jozef.io Web site: https://jozef.io/r917-fread-comparisons/

Laubheimer , P. (2019, September 29). *Treemaps: Data Visualization of Complex Hierarchies*. Retrieved from Nielsen Norman Group Web site: https://www.nngroup.com/articles/treemaps/

Macrotrends. (2021). *Vancouver, Canada Metro Area Population 1950-2021*. Retrieved from Macrotrends Web site: https://www.macrotrends.net/cities/20404/vancouver/population

Martin, M. (n.d.). *Data on Employee Turnover in the Grocery Industry*. Retrieved from Smallbusiness Chron Web site: https://smallbusiness.chron.com/data-employee-turnover-grocery-industry-18817.html

Natalia. (2020, December 23). *The top 10 reasons people get fired*. Retrieved from Perkins Asbill Web site: https://www.perkinsasbill.com/the-top-10-reasons-people-get-fired/

Personio. (2021). *What Is Your Company’s ‘Attrition Rate’ Telling You?* Retrieved from Personio Web site: https://www.personio.com/hr-lexicon/attrition-rate/

Pinkasovitch, A. (2021, April 24). *What's the Difference Between Retirement in Canada and America?* Retrieved from Investopedia Web site: https://www.investopedia.com/articles/retirement/11/difference-retiring-canada-america.asp

PregnancyInfo. (n.d.). *Age and fertility*. Retrieved from PregnancyInfo Web site: https://www.pregnancyinfo.ca/before-you-conceive/fertility/age-and-fertility/

R graph gallery. (2018). *Grouped, stacked and percent stacked barplot in ggplot2*. Retrieved from The R graph gallery Web site: https://www.r-graph-gallery.com/48-grouped-barplot-with-ggplot2.html

Robinson, E. (n.d.). *Introduction to forcats*. Retrieved from Cran r projects Web site: https://cran.r-project.org/web/packages/forcats/vignettes/forcats.html

Statistics Canada. (2021, November 19). *Incident-based crime statistics, by detailed violations, police services in British Columbia*. Retrieved from Statistics Canada Web site: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510018401&pickMembers%5B0%5D=1.1&pickMembers%5B1%5D=2.34&cubeTimeFrame.startYear=2005&cubeTimeFrame.endYear=2015&referencePeriods=20050101%2C20150101

Thakur, M. (2020). *Attrition Formula*. Retrieved from EDUCBA Web site: https://www.educba.com/attrition-formula/

U.S. Bureau of Labor Statistics. (2021, September 8). *Occupational Outlook Handbook | Cashier*. Retrieved from Bureau of Labor Statistics Web site: https://www.bls.gov/ooh/sales/cashiers.htm

Wickham, H. (2021). *Create, modify, and delete columns*. Retrieved from Tidyverse.org Web site: https://dplyr.tidyverse.org/reference/mutate.html

Wilkins, D. (2021). *Treemaps in ggplot2 with treemapify*. Retrieved from R Charts Web site: https://r-charts.com/part-whole/treemapify/

# **Appendix**

## **A. Statistics of Robbery in British Columbia 2005 - 2014**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Incident-based crime statistics, by detailed violations, police services in British Columbia** | | | | | | | | | | | | | | | | | | | | | |
| Frequency: Annual | | | | | | | | | | | | | | | | | | | | | |
| Table: 35-10-0184-01 (formerly CANSIM 252-0081) | | | | | | | | | | | | | | | | | | | | | |
| Release date: 2021-07-27 | | | | | | | | | | | | | | | | | | | | | |
| Geography: Province or territory | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | | |  | |  | |  | |  | |  | |  | |  | |  | |  |
| **British Columbia** | | | | | | | | | | | | | | | | | | | | | |
| **Total robbery** | | | | | | | | | | | | | | | | | | | | | |
| Statistics | 2005 | | 2006 | 2007 | | 2008 | | 2009 | | 2010 | | 2011 | | 2012 | | 2013 | | 2014 | | 2015 | |
|  | Number | | | | | | | | | | | | | | | | | | | | |
| Actual incidents | 5,174 | | 5,318 | 5,508 | | 5,413 | | 5,136 | | 4,901 | | 4,492 | | 4,523 | | 3,581 | | 3,392 | | 3,645 | |
|  | Rate | | | | | | | | | | | | | | | | | | | | |
| Rate per 100,000 population | 123.28 | | 125.37 | 128.36 | | 124.45 | | 116.44 | | 109.74 | | 99.78 | | 99.04 | | 77.34 | | 72.06 | | 76.31 | |
|  | Percent | | | | | | | | | | | | | | | | | | | | |
| Percentage change in rate | 1.02 | | 1.7 | 2.38 | | -3.04 | | -6.44 | | -5.76 | | -9.08 | | -0.74 | | -21.91 | | -6.83 | | 5.9 | |
|  | Number | | | | | | | | | | | | | | | | | | | | |
| Total cleared | 1,604 | | 1,650 | 1,736 | | 1,786 | | 1,739 | | 1,781 | | 1,514 | | 1,570 | | 1,309 | | 1,253 | | 1,320 | |
| Cleared by charge | 1,478 | | 1,515 | 1,619 | | 1,637 | | 1,603 | | 1,628 | | 1,375 | | 1,423 | | 1,207 | | 1,169 | | 1,243 | |
| Cleared otherwise | 126 | | 135 | 117 | | 149 | | 136 | | 153 | | 139 | | 147 | | 102 | | 84 | | 77 | |
| Total, persons charged | 1,593 | | 1,579 | 1,656 | | 1,592 | | 1,632 | | 1,555 | | 1,411 | | 1,369 | | 1,198 | | 1,132 | | 1,212 | |
|  | Rate | | | | | | | | | | | | | | | | | | | | |
| Rate, total persons charged per 100,000 population aged 12 years and over | 43.48 | | 42.54 | 44.03 | | 41.69 | | 42.08 | | 39.55 | | 35.57 | | 33.97 | | 29.29 | | 27.2 | | 28.69 | |
|  | Number | | | | | | | | | | | | | | | | | | | | |
| Total, adult charged | 1,347 | | 1,283 | 1,303 | | 1,265 | | 1,301 | | 1,215 | | 1,123 | | 1,079 | | 972 | | 941 | | 1,052 | |
|  | Rate | | | | | | | | | | | | | | | | | | | | |
| Rate, adult charged per 100,000 population aged 18 years and over | 40.37 | | 37.92 | 37.91 | | 36.17 | | 36.56 | | 33.62 | | 30.75 | | 28.99 | | 25.65 | | 24.35 | | 26.79 | |
|  | Number | | | | | | | | | | | | | | | | | | | | |
| Total, youth charged 18 19 | 246 | | 296 | 353 | | 327 | | 331 | | 340 | | 288 | | 290 | | 226 | | 191 | | 160 | |
|  | Rate | | | | | | | | | | | | | | | | | | | | |
| Rate, youth charged per 100,000 population aged 12 to 17 years | 75.21 | | 90.29 | 108.9 | | 101.8 | | 103.47 | | 106.89 | | 91.52 | | 94.21 | | 75.03 | | 64.16 | | 53.62 | |
|  | Number | | | | | | | | | | | | | | | | | | | | |
| Total, youth not charged | 64 | | 59 | 60 | | 62 | | 70 | | 96 | | 70 | | 70 | | 60 | | 22 | | 19 | |
|  | Rate | | | | | | | | | | | | | | | | | | | | |
| Rate, youth not charged per 100,000 population aged 12 to 17 years | 19.57 | | 18 | 18.51 | | 19.3 | | 21.88 | | 30.18 | | 22.24 | | 22.74 | | 19.92 | | 7.39 | | 6.37 | |
| Symbol legend: | | |  |  | |  | |  | |  | |  | |  | |  | |  | |  | |
| .. | not available for a specific reference period | | | | | | |  | |  | |  | |  | |  | |  | |  | |
|  |  | |  |  | |  | |  | |  | |  | |  | |  | |  | |  | |
| Table Corrections: | | |  |  | |  | |  | |  | |  | |  | |  | |  | |  | |
| Date | Note | |  |  | |  | |  | |  | |  | |  | |  | |  | |  | |
| 1/8/2019 | On August 1, 2019, the data for 2018 for “Total impaired driving” have been corrected. | | | | | | | | | | | | |  | |  | |  | |  | |

Table 1: Statistics of Robbery in British Columbia 2005 – 2014 (Statistics Canada, 2021).