# ASSIGNMENT

## TECHNOLOGY PARK MALAYSIA

## CT127-3-2-PFDA

## PROGRAMMING FOR DATA ANALYSIS

## APU2F2109SE

**HAND OUT DATE: 4 OCTOBER 2021**

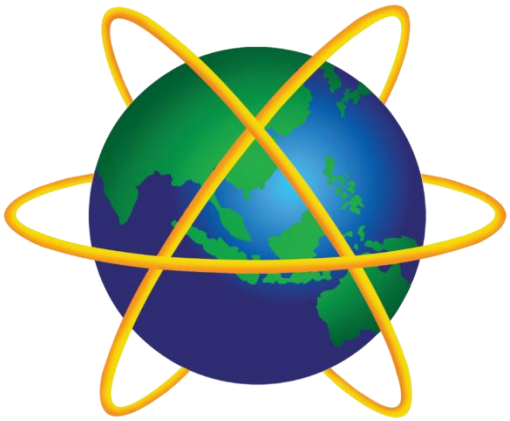
**HAND IN DATE: 22 NOVEMBER 2021**

**WEIGHTAGE: 50%**

---

**INSTRUCTIONS TO CANDIDATES:**

1. **Submit your assignment at the administrative counter.**
2. **Students are advised to underpin their answers with the use of references (cited using the American Psychological Association (APA) Referencing).**
3. **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld.**
4. **Cases of plagiarism will be penalized.**
5. **The assignment should be bound in an appropriate style (comb bound or stapled).**
6. **Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.**
7. **You must obtain 50% overall to pass this module.**




# INDIVIDUAL ASSIGNEMENT

## TECHNOLOGY PARK MALAYSIA

## CT127-3-2-PFDA

## PROGRAMMING FOR DATA ANALYSIS

## APU2F2109SE

## HAND OUT DATE: 4 OCTOBER 2021

## HAND IN DATE: 22 NOVEMBER 2021


| | | |
|---|---|---|
| **Name** | **:** | Ricky Wong Tiong Song |
| **Student ID** | **:** | TP054622 |

## Table of Contents

## Table of Figures

## 1.0 Introduction

This assignment aims to perform data analysis on a dataset containing the staff's data within an organization from 2006 to 2015. The dataset consists of 18 columns and 49654 rows, including the staff's detail, department, job title, location, working status, and termination reason. The given dataset may consist of some hidden issues in human resources management that need to be identified. Hence, the objective of this assignment is to recognize those hidden issues and provide valuable insights for the human resources management team for decision making. The data analysis will be started by identifying some questions related to the dataset and then conducting in-depth analysis to prove and answer the identified questions. All these processes will only be using the R programming language, which is a programming language specifically for statistical computing and graphics. Concepts like data Exploration, Manipulation, Transformation, and Visualization will also be used throughout the processes to achieve the objective. Overall, this report consists of 3 questions and 14 analyses regarding the dataset.

## 2.0 Assumptions

A few assumptions are made to let the reader understand the analysis process better and ensure all the justifications could make sense. Firstly, the dataset is assumed to be a dataset from a province in Canada named British Columbia because the names of the city in this dataset are the cities in British Columbia, Canada. Secondly, the statistics that will be used to support the findings throughout the analysis are mostly based in Canada.

## 3.0 Preparation

### Load Libraries

Package is the fundamental unit of shareable code in R. It is a library of prewritten code designed to accomplish some tasks or a set of tasks. Below are the packages that will be used for this data analysis project:

1. **tidyverse**: A collection of R packages designed for data science. The packages that will be used in this project are ggplot2, dplyr, and lubridate.
2. **scales**: Provide the internal scaling infrastructure used by ggplot2 and the tools to override the default breaks, labels, transformation, and palettes (scales.r-lib.org, n.d.).

```
# ================================== Load libraries ==================================
library(tidyverse)
library(lubridate)
library(scales)
```

*Figure 1:Load libraries*

### Data Import

The code above is to import the dataset that will be used in this data analysis project. Firstly, set the working directory to the folder location of the dataset excel file named "**employee_attrition.csv**". Secondly, import the data into the R environment using the `read.csv` function. The dataset has been assigned to a variable named `data`, which will act as the clean dataset.

```
# ================================== Import Data ==================================
# Set the working directory
setwd("D:/Desktop/Degree-Y2/SEM-1/PFDA/Assignment/PFDA-Assignment")
# clean dataset
data <- read.csv("employee_attrition.csv")
```

*Figure 2: Data Import*

### Data Pre-processing

A crucial step before everything starts is to do data pre-processing. This is to ensure that the dataset we are dealing with is in an appropriate format and clean. Besides, it can also help us prevent unexpected errors during our analysis and ensure our analysis results are more accurate. To do so, some useful functions can be used to examine our dataset, for example, `str()` and `summary()`.

`str()`: To examine the internal structure of an R object.

`summary()`: A generic function that shows the summaries of the dataset.

```
> str(data)
'data.frame':   49653 obs. of  18 variables:
 $ EmployeeID         : int  1318 1318 1318 1318 1318 1318 1318 1318 1318 1318 ...
 $ recorddate_key     : chr  "12/31/2006 0:00" "12/31/2007 0:00" "12/31/2008 0:00" "12/31/2009 0:00" ...
 $ birthdate_key      : chr  "1/3/1954" "1/3/1954" "1/3/1954" "1/3/1954" ...
 $ orighiredate_key   : chr  "8/28/1989" "8/28/1989" "8/28/1989" "8/28/1989" ...
 $ terminationdate_key: chr  "1/1/1900" "1/1/1900" "1/1/1900" "1/1/1900" ...
 $ age                : int  52 53 54 55 56 57 58 59 60 61 ...
 $ length_of_service  : int  17 18 19 20 21 22 23 24 25 26 ...
 $ city_name          : chr  "Vancouver" "Vancouver" "Vancouver" "Vancouver" ...
 $ department_name    : chr  "Executive" "Executive" "Executive" "Executive" ...
 $ job_title          : chr  "CEO" "CEO" "CEO" "CEO" ...
 $ store_name         : int  35 35 35 35 35 35 35 35 35 35 ...
 $ gender_short       : chr  "M" "M" "M" "M" ...
 $ gender_full        : chr  "Male" "Male" "Male" "Male" ...
 $ termreason_desc    : chr  "Not Applicable" "Not Applicable" "Not Applicable" "Not Applicable" ...
 $ termtype_desc      : chr  "Not Applicable" "Not Applicable" "Not Applicable" "Not Applicable" ...
 $ STATUS_YEAR        : int  2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 ...
 $ STATUS             : chr  "ACTIVE" "ACTIVE" "ACTIVE" "ACTIVE" ...
 $ BUSINESS_UNIT      : chr  "HEADOFFICE" "HEADOFFICE" "HEADOFFICE" "HEADOFFICE" ...
```

*Figure 3:Structure of the data before data cleaning*

```
> summary(data)
   EmployeeID   recorddate_key     birthdate_key      orighiredate_key   terminationdate_key      age        length_of_service  city_name          department_name     job_title
 Min.   :1318   Length:49653       Length:49653       Length:49653       Length:49653        Min.   :19.00   Min.   : 0.00    Length:49653       Length:49653       Length:49653
 1st Qu.:3360   Class :character   Class :character   Class :character   Class :character    1st Qu.:31.00   1st Qu.: 5.00    Class :character   Class :character   Class :character
 Median :5031   Mode  :character   Mode  :character   Mode  :character   Mode  :character    Median :42.00   Median :10.00    Mode  :character   Mode  :character   Mode  :character
 Mean   :4859                                                                              Mean   :42.08   Mean   :10.43
 3rd Qu.:6335                                                                              3rd Qu.:53.00   3rd Qu.:15.00
 Max.   :8336                                                                              Max.   :65.00   Max.   :26.00
   store_name     gender_short       gender_full        termreason_desc    termtype_desc       STATUS_YEAR      STATUS          BUSINESS_UNIT
 Min.   : 1.0   Length:49653       Length:49653       Length:49653       Length:49653       Min.   :2006   Length:49653       Length:49653
 1st Qu.:16.0   Class :character   Class :character   Class :character   Class :character   1st Qu.:2008   Class :character   Class :character
 Median :28.0   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2011   Mode  :character   Mode  :character
 Mean   :27.3                                                                             Mean   :2011
 3rd Qu.:42.0                                                                             3rd Qu.:2013
 Max.   :46.0                                                                             Max.   :2015
```

*Figure 4: Summary of the data before data cleaning*

Based on the result of str() and summary(), some variables need to be formatted, and some meaningless variables need to be omitted. Below is a detailed explanation of the data pre-processing:

Firstly, identify the useless variables. As we can see from either *Figure 3* or *Figure 4*, two variables are used to store the gender of the employee, which are `gender_short` and `gender_full`. We don't have to have two variables to give us the same data, so gender_full will be omitted in this case. Next, the `recorddate_key` should also be omitted because we have `STATUS_YEAR` in the dataset to show the year of the record. The `birthdate_key` is another meaningless data as the primary focus of the analysis is on employee attrition, so we don't have to know the birthdate of the employees.

Secondly, identify the data that needs to be formatted or modified. Based on *Figure 3*, it seems that the `orighiredate_key` and `terminationdate_key` are supposed to be in

Date format since the values are about dates. Hence, the two variables should convert into Date format. After some exploration on the dataset, it seems like all the employees that are still active in the company, their termination date will be "1/1/1900" and it could be misleading. So, if the date is "1/1/1900", the value will be replaced by `NA`. Most of the variables in the dataset appear to be categorical variables. So, those variables can be converted into factors to be easier to view the levels and order them. The categorical variables in the dataset: `city_name`, `department_name`, `job_title`, `store_name`, `gender_short`, `termreason_desc`, `termtype_desc`, `STATUS_YEAR`, `STATUS`, and `BUSINESS_UNIT`. Below is the code implemented for cleaning up the data. The cleaned dataset is stored in a variable named "**emp**".

```
# Convert char type data into factor, for date data convert from char to date.
emp <- data %>%
  mutate(
    orighiredate_key = as.Date(orighiredate_key, format = "%m/%d/%Y"),
    # 1/1/1900 means still working so set those as NA
    terminationdate_key = ifelse(terminationdate_key == "1/1/1900", NA, terminationdate_key),
    terminationdate_key = as.Date(terminationdate_key, format = "%m/%d/%Y"),
    city_name = as.factor(city_name),
    department_name = as.factor(department_name),
    job_title = as.factor(job_title),
    store_name = as.factor(store_name),
    gender_short = as.factor(gender_short),
    # typo. "Resignaton" -> "Resignation"
    termreason_desc = ifelse(termreason_desc == "Resignaton", "Resignation", termreason_desc),
    termreason_desc = as.factor(termreason_desc),
    termtype_desc = as.factor(termtype_desc),
    STATUS_YEAR = as.factor(STATUS_YEAR),
    STATUS_YEAR = ordered(STATUS_YEAR, c(2006,2007,2008,2009,2010,2011,2012,2013,2014,2015)),
    STATUS = as.factor(STATUS),
    BUSINESS_UNIT = as.factor(BUSINESS_UNIT)
  ) %>%
  # Remove meaningless attributes
  select(everything(), -c(gender_full, recorddate_key, birthdate_key))
```

*Figure 5:Data Transformation*

Lastly, check the structure and the data again before moving on to the data exploration.

```
> str(emp)
'data.frame':   49653 obs. of  15 variables:
 $ EmployeeID         : int  1318 1318 1318 1318 1318 1318 1318 1318 1318 1318 ...
 $ orighiredate_key   : Date, format: "1989-08-28" "1989-08-28" ...
 $ terminationdate_key: Date, format: NA NA ...
 $ age                : int  52 53 54 55 56 57 58 59 60 61 ...
 $ length_of_service  : int  17 18 19 20 21 22 23 24 25 26 ...
 $ city_name          : Factor w/ 40 levels "Abbotsford","Aldergrove",..: 35 35 35 35 35 35 35 35 35 35 ...
 $ department_name    : Factor w/ 21 levels "Accounting","Accounts Payable",..: 10 10 10 10 10 10 10 10 10 10 ...
 $ job_title          : Factor w/ 47 levels "Accounting Clerk",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ store_name         : Factor w/ 46 levels "1","2","3","4",..: 35 35 35 35 35 35 35 35 35 35 ...
 $ gender_short       : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ termreason_desc    : Factor w/ 4 levels "Layoff","Not Applicable",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ termtype_desc      : Factor w/ 3 levels "Involuntary",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ STATUS_YEAR        : Ord.factor w/ 10 levels "2006"<"2007"<..: 1 2 3 4 5 6 7 8 9 10 ...
 $ STATUS             : Factor w/ 2 levels "ACTIVE","TERMINATED": 1 1 1 1 1 1 1 1 1 1 ...
 $ BUSINESS_UNIT      : Factor w/ 2 levels "HEADOFFICE","STORES": 1 1 1 1 1 1 1 1 1 1 ...
```

*Figure 6: Structure of the data after cleaning up*

```
> summary(emp)
   EmployeeID   orighiredate_key     terminationdate_key       age
 Min.   :1318   Min.   :1989-08-28   Min.   :2006-01-01   Min.   :19.00
 1st Qu.:3360   1st Qu.:1995-06-02   1st Qu.:2010-06-23   1st Qu.:31.00
 Median :5031   Median :2000-03-31   Median :2013-05-31   Median :42.00
 Mean   :4859   Mean   :2000-09-04   Mean   :2012-09-26   Mean   :42.08
 3rd Qu.:6335   3rd Qu.:2005-10-13   3rd Qu.:2014-12-30   3rd Qu.:53.00
 Max.   :8336   Max.   :2013-12-11   Max.   :2015-12-30   Max.   :65.00
                                     NA's   :42450
 length_of_service           city_name            department_name          job_title
 Min.   : 0.00     Vancouver      :11211   Meats           :10269   Meat Cutter  :9984
 1st Qu.: 5.00     Victoria       : 4885   Dairy           : 8599   Dairy Person :8590
 Median :10.00     Nanaimo        : 3876   Produce         : 8515   Produce Clerk:8237
 Mean   :10.43     New Westminster: 3211   Bakery          : 8381   Baker        :8096
 3rd Qu.:15.00     Kelowna        : 2513   Customer Service: 7122   Cashier      :6816
 Max.   :26.00     Burnaby        : 2067   Processed Foods : 5911   Shelf Stocker:5622
                   (Other)        :21890   (Other)         :  856   (Other)      :2308
   store_name    gender_short       termreason_desc         termtype_desc
 46     : 4422   F:25898      Layoff        :  215   Involuntary   :  215
 18     : 3876   M:23755      Not Applicable:48168   Not Applicable:48168
 42     : 3827                Resignation   :  385   Voluntary     : 1270
 21     : 3211                Retirement    :  885
 43     : 2896
 16     : 2513
 (Other):28908
  STATUS_YEAR          STATUS         BUSINESS_UNIT
 2013   : 5320   ACTIVE    :48168   HEADOFFICE:  585
 2012   : 5231   TERMINATED: 1485   STORES    :49068
 2014   : 5215
 2011   : 5082
 2010   : 4963
 2015   : 4961
 (Other):18881
```

*Figure 7: Summary of the data after cleaning up*

## 4.0 Data Exploration

Data exploration is the first step in data analysis, where the analyst will explore the dataset to get a general idea of the data. Several techniques can be used for data exploration, such as data visualization and statistical analysis like mean, standard deviation, regression, etc. The visuals or statistics could help the data analyst identify patterns and relationships between variables and help them construct questions about the data.

### Treasure Hunt 1

Since the primary focus of the analysis is about employee attrition, simple data exploration is applied to have a general idea of why employees chose to leave the company and have a general understanding of the termination trend over the 10 years.

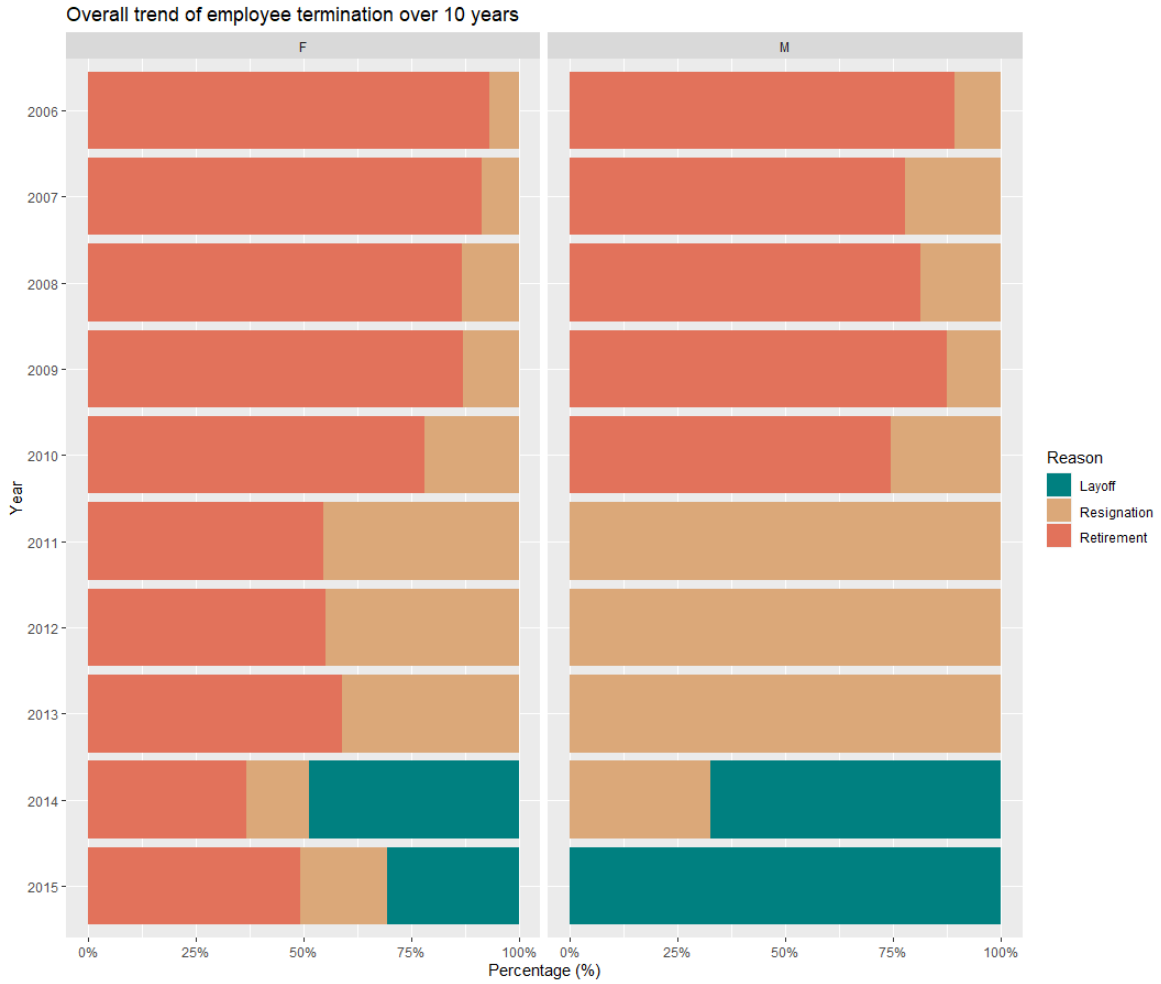



*Figure 8:Overall trend of employee termination over 10 years*

**Source code (Figure 8)**:

```
# Treasure Hunt 2 - Have idea on the trend
emp %>%
  filter(STATUS == "TERMINATED") %>%
  group_by(STATUS_YEAR) %>%
  ggplot(aes(STATUS_YEAR, fill = termreason_desc)) +
  geom_bar(position = "fill") +
  scale_x_discrete(limits = rev) +
  scale_y_continuous(labels = scales::percent) +
  facet_wrap(~gender_short) +
  coord_flip() +
  labs(title = "Overall trend of employee termination over 10 years", fill = "Reason") +
  xlab("Year") +
  ylab("Percentage (%)") +
  scale_fill_manual(values = c("Layoff" = "#008080",
                               "Resignation" = "#dba879",
                               "Retirement" = "#e2725b"))
```

*Figure 9: Source code for Figure 8*

Based on *Figure 8*, we can see that:

1. Retirement is the primary reason that the employees leave the company over 10 years.
2. Over the 10 years, the company will likely have more older female workers than older male workers because many older female workers will retire from the job every year.
3. The case where the employee left the company due to layoff only occurred in 2014 and 2015.

These findings have led to these questions:

1. What age did the employee retire at?
2. What are the reasons for the layoffs?

## Treasure Hunt 2

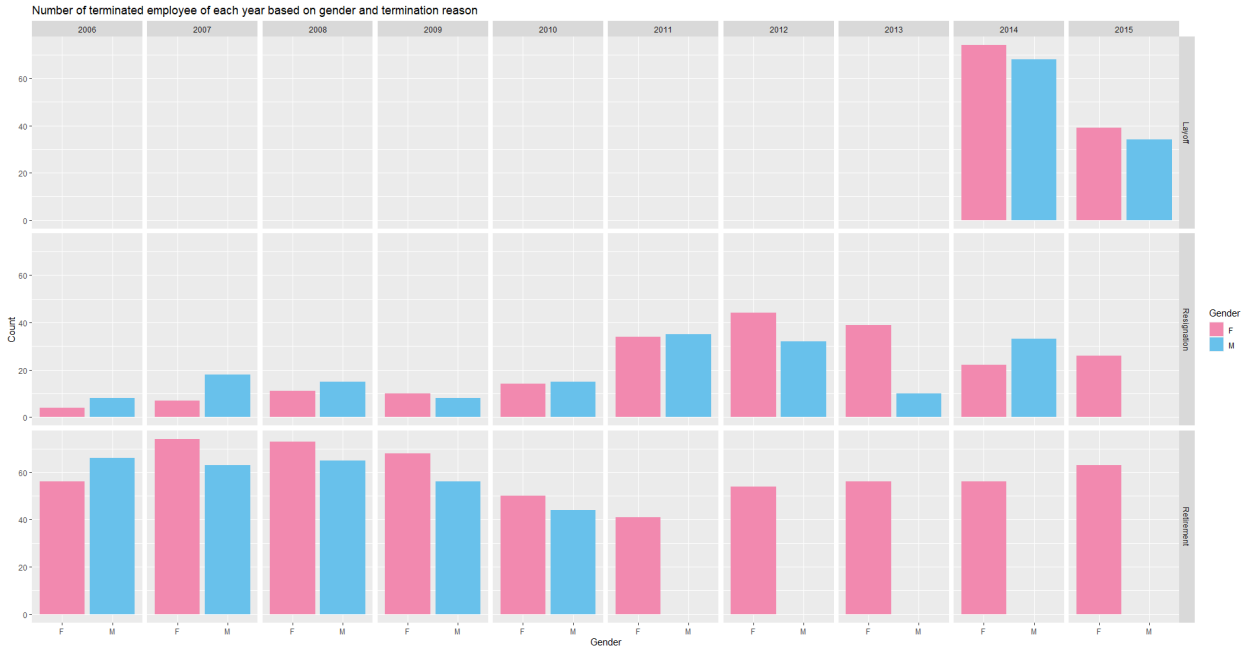



*Figure 10: Number of terminated employees of each year based on gender and termination reason*

## Source code (Figure 10):

```
# is there any relationship between gender and termination rate?
emp %>%
  filter(
    STATUS == "TERMINATED",
    ) %>%
  group_by(gender_short) %>%
  ggplot(aes(gender_short, fill = gender_short)) +
  geom_bar() +
  facet_grid(termreason_desc~STATUS_YEAR) +
  labs(title = "Number of terminated employee of each year based on gender and termination reason", fill = "Gender") +
  xlab("Gender") +
  ylab("Count") +
  scale_fill_manual(values = c("F" = "#F289AF",
                               "M" = "#68C1EB"))
```

*Figure 11: Source code for Figure 10*

*Figure 10* is another interesting finding on the employee attrition dataset. The graph shows the number of terminated employees over the 10 years based on gender and termination reasons. Below are the findings:

1. It proves that layoff only occurred in 2014 and 2015 once again, and it seems like there are slightly more female employees laid off than male employees.
2. Starting from 2011, the number of resignations in both genders increased, and after that, the number of resignations in female employees started to surpass the number of

resignations in male employees. The number of resignations in male employees dropped sharply in 2013 and dramatically in 2014 again.

3. The graph also shows the issue of uneven age distribution because female employees mainly contribute to retirement. It means the company has more old female workers compared to male workers in the past 10 years.

These findings have led to these questions:

1. What are the reasons for resignation?

## Question 1: What age did the employee retire at?

This question is designed to investigate the retirement age of the employees. As we discovered in our data exploration, retirement is the primary reason that employees leave the company. Besides, the company has a significant number of female employees left the company due to retirement every year. Hence, this question is crucial to find out the retirement age of the employees so that the company can be well-prepared for the retirement wave in the future. From the perspective of the younger employees, the retirement of the older employees could be good news because it means they might be able to get a promotion and perhaps get a raise on their salary. But for the company, it is a significant loss. When older employees retire from the job, it means they are losing experts and experience-based knowledge. It can almost feel like they're taking the company with them. According to Insperity, there are 8 tips for the company to prepare for retirement waves (Turner, n.d.):

1. **Prevent knowledge gaps.** It is the main reason to prepare for retirement waves. The company should ensure that older employees can share their experiences with their juniors or the company before leaving.
2. **Don't undervalue older employees.** The company should provide retirement benefits for their employees to match their contributions over the years. It's the company's job to discuss with the employee to know their plan and opinion on their retirement.
3. **Cross-train employees.** Training needs to be done so that the knowledge will not just leave the company as they leave. It is to prevent any knowledge silos after the employee retires.
4. **Consider alternatives to full retirement.** Some employees might want to retire ASAP once they reach the retirement age, but some do not. For those who would like to stay in the company for a few years again, the company can perhaps let them switch to part-time or other work arrangements.
5. **Plan succession across all departments.** The company needs to investigate each department of the company to know which department will be affected the most due to employee retirement. Then plan to place or hire new younger employees to be prepared to replace their position.
6. **Manage across generations.** The company needs to encourage older employees to mentor younger employees. Younger employees should also be encouraged to learn as much as they can from their seniors.

7. **Make annual assessments.** The company needs to investigate which department or jobs will have a big wave of employee retirement.
8. **Don't wait till they're out the door.** Knowledge transfer will consume a lot of time and effort, so it requires long-term planning to ensure the transition is smooth. The company also might spend much time discussing with their employees' retirement planning. Hence, it's not something that can be done in two weeks before the employee retires.

## Analysis 1-1: Find the relationship between age and termination reason

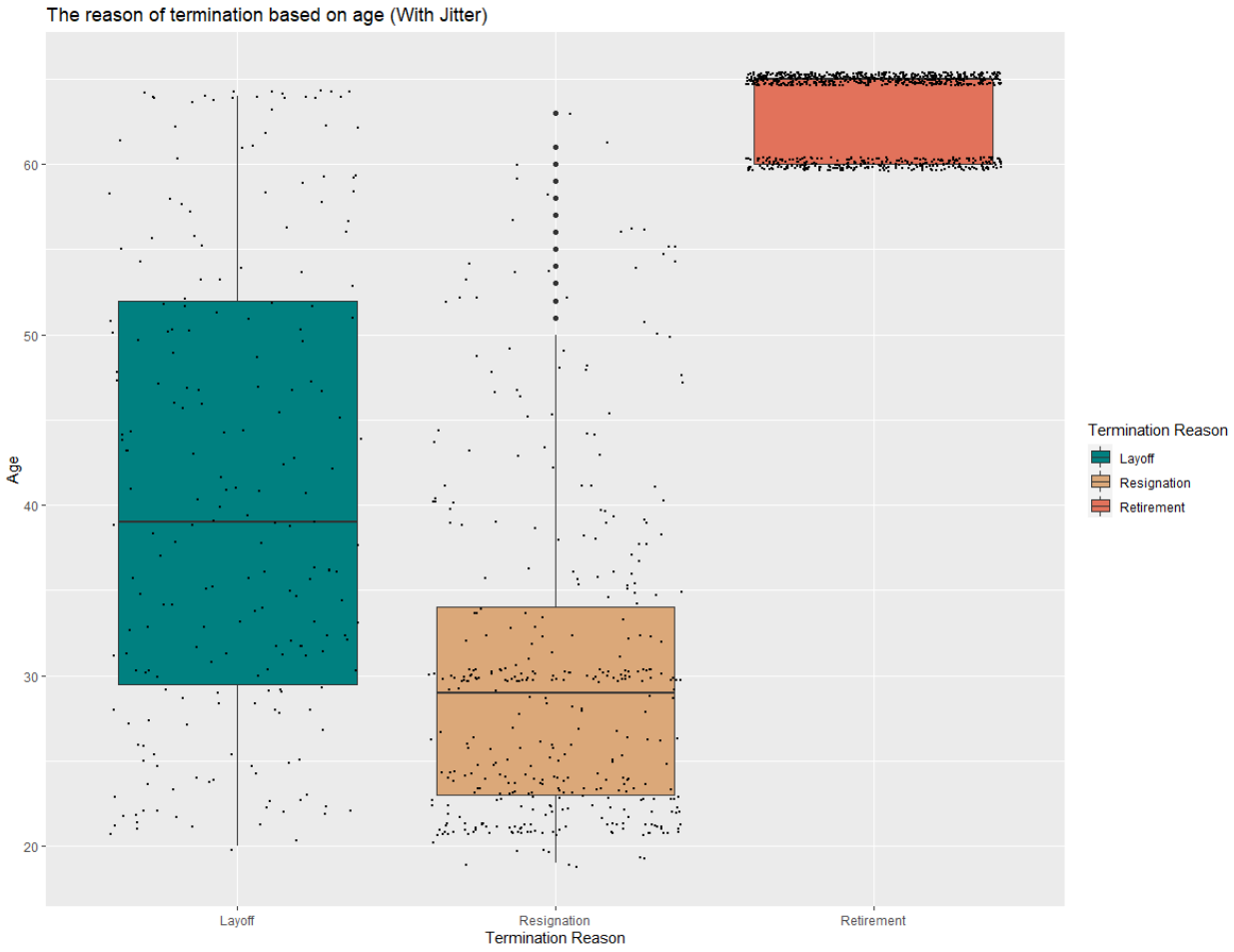



*Figure 12: The reason for the termination based on age (with Jitter)*

## Source code (Figure 12):

```
# Find the relationship between age and termination reason
age_terminated <- emp %>%
  drop_na(termreason_desc) %>%
  filter(STATUS == "TERMINATED") %>%
  group_by(age)

## boxplot with jitter
ggplot(data = age_terminated, mapping = aes(x = termreason_desc, y = age)) +
  geom_boxplot(mapping = aes(fill = termreason_desc)) +
  labs(title = "The reason of termination based on age (With Jitter)", fill = "Termination Reason") +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  xlab("Termination Reason") +
  ylab("Age") +
  scale_fill_manual(values = c("Layoff" = "#008080",
                               "Resignation" = "#dba879",
                               "Retirement" = "#e2725b"))
```

*Figure 13: Source code for Figure 12*

**Explanation**

*Figure 12* shows the relationship between age and termination reason. Based on the graph, we can observe that retirement has a bimodal distribution. The dots are mainly distributed in 2 groups, where are 60 years old and 65 years old. It means that when the employees reach 60 or 65 years old, they will choose to retire from their job. According to the official site of the Government of Canada, the minimum age to qualify for the Canada Pension Plan (CPP) is 60 years old (Government of Canada, 2021). The legal age to receive the pension is 65, but they can choose to receive it as early as 60 or as late as 70. Hence, it's not a surprise that most of the employees in the company will retire from the company by 60 years old or older. By knowing the retirement age of the employees, the company can start to investigate the employees that might retire from their job in the next 5 to 10 years.

## Analysis 1-2: Find the number of active employees that will be going to retire soon

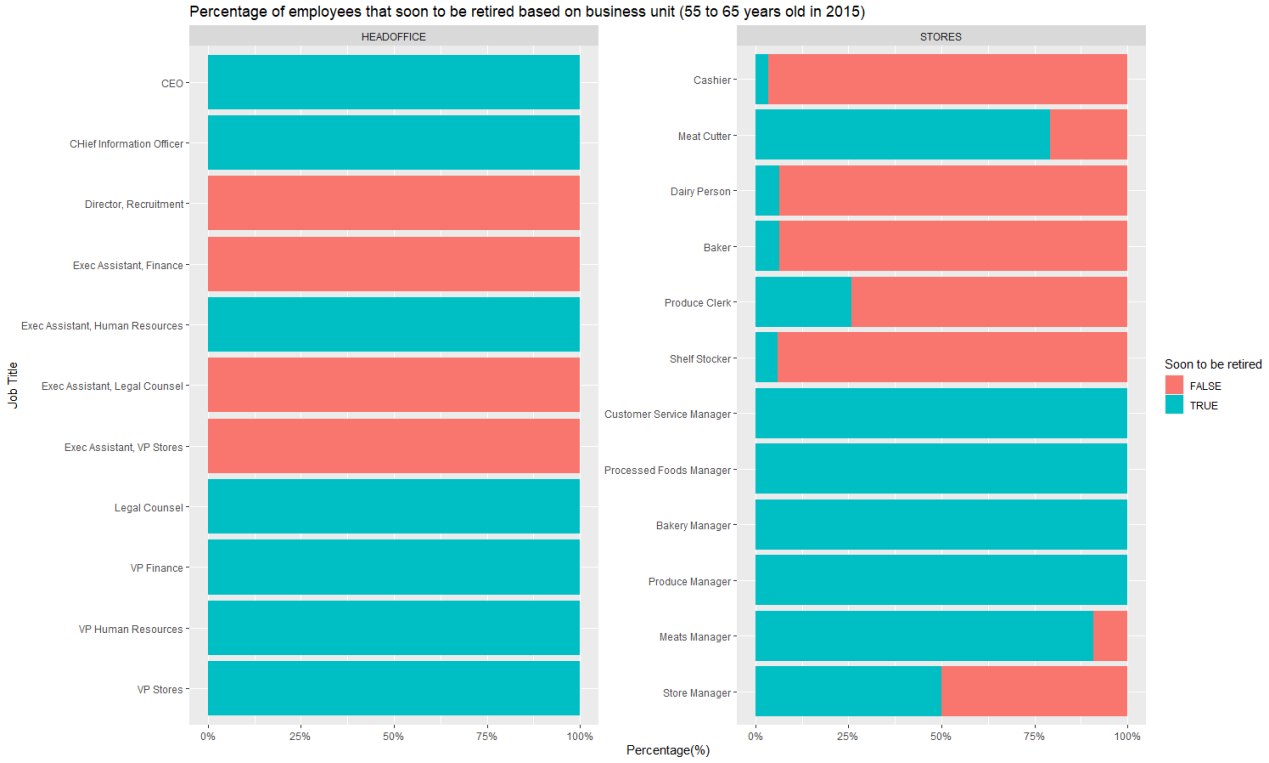



*Figure 14: Percentage of employees that might retire soon based on business unit*

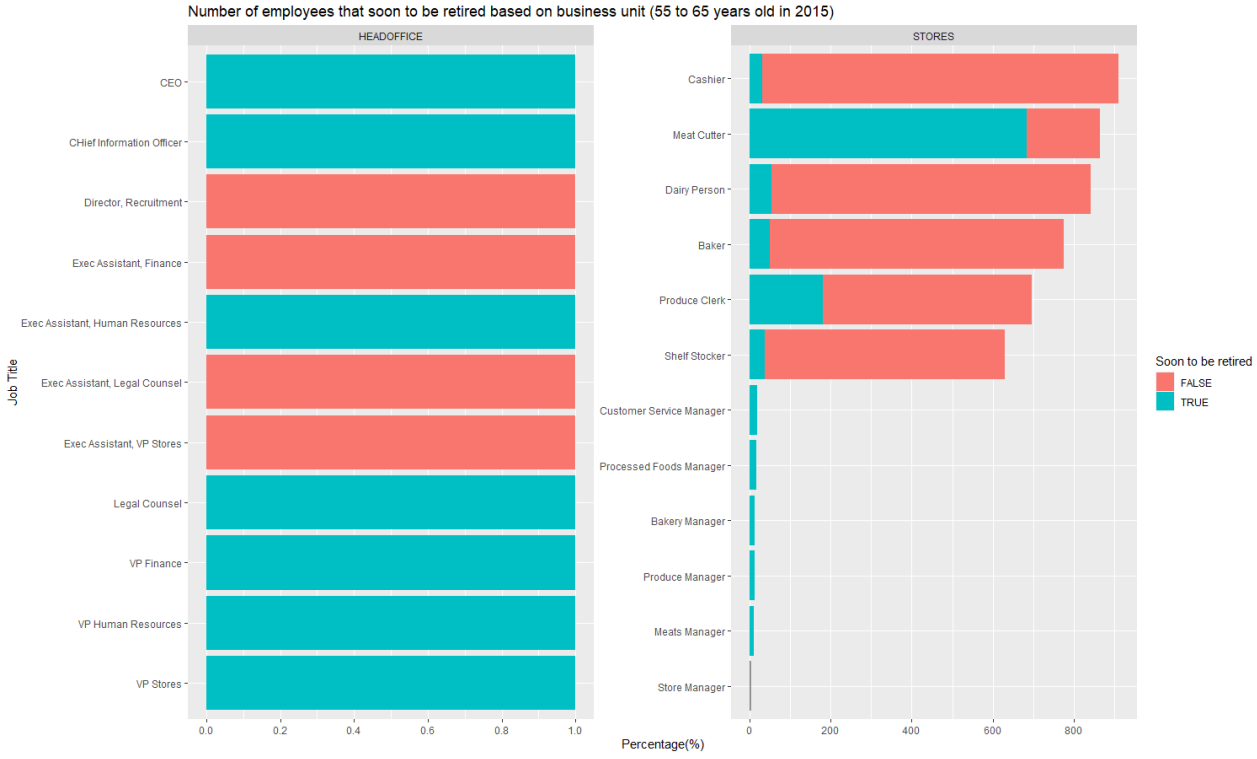



*Figure 15: Number of employees that might retire soon based on business unit*

**Source code (Figure 14):**

```r
# Find the number of active employees that will going to retire soon
retire_soon <- emp %>%
  filter(
    STATUS_YEAR == 2015,
    STATUS == "ACTIVE"
  ) %>%
  group_by(job_title) %>%
  mutate(
    might_retired_soon = ifelse((age >= 55 & age <= 65), TRUE, FALSE)
  )

ggplot(retire_soon, mapping = aes(x = fct_rev(fct_infreq(job_title)), fill = might_retired_soon)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Percentage of employees that soon to be retired based on business unit (55 to 65 years old in 2015)", fill = "Soon to be retired") +
  xlab("Job Title") +
  ylab("Percentage(%)") +
  coord_flip() +
  facet_wrap(~BUSINESS_UNIT, scales = "free")
```

*Figure 16: Source code for Figure 14*

**Source code (Figure 15):**

```r
ggplot(retire_soon, mapping = aes(x = fct_rev(fct_infreq(job_title)), fill = might_retired_soon)) +
  geom_bar() +
  scale_y_continuous(breaks= pretty_breaks()) +
  labs(title = "Number of employees that soon to be retired based on business unit (55 to 65 years old in 2015)", fill = "Soon to be retired") +
  xlab("Job Title") +
  ylab("Percentage(%)") +
  coord_flip() +
  facet_wrap(~BUSINESS_UNIT, scales = "free")
```

*Figure 17: Source code for Figure 15*

**Explanation**

As we know from the previous analysis, the employees are likely to retire when they are 60 or 65 years old. Hence, the company can figure out the employees who will retire within 5 to 10 years, which means finding out the active employees in the range of 55 to 65 years old. *Figure 14* shows the percentage of employees that will be retired soon in 5 to 10 years based on the business unit. *Figure 15* displays the number of employees that will be going to be retired soon.

Firstly, let's look at the employees in HEADOFFICE. Within 5 or 10 years, there will be 7 out of 11 employees that might retire. As we examine the job titles, we can see that these jobs belong to the company's top management. These positions are highly crucial as they control and oversee the entire organization. When a top management level employee has decided to retire, the company might face a substantial cultural shift (J.Kain, 2021). It will undeniably have a massive impact on the company from top to bottom, especially for positions like CEOs.

Given the current situation, the company should start the selection process to find potential candidates with essential qualities to lead the company's future. The process will be time-consuming as there will be lots of discussion among the board members, discussion about the

company's future, interviewing potential candidates, and so on. If they can't find any suitable candidate within the company, they might need to choose candidates out of the company, which will be another lengthy process. The company will also need to plan how to onboard the new CEO so that there will be a smooth transition and help the new CEO understand the current situation. These processes and procedures require lots of time and effort, so planning 5 years or 10 years before will be sufficient.

By examining the statistics on the employees in STORES, we can also see that the company might also face a vast wave of retirement. The job that will be affected the most is the meat cutter because 75% of the meat cutter will retire soon. Either *Figure 14* or *Figure 15* also shows the distribution of age in each job. There is an uneven distribution of age in meat cutters, which will be a severe issue soon. As mentioned before, when the older employees are leaving, the company is losing experience and knowledge. Those experienced meat cutters will be a significant loss for either the company or the store where they worked.

Meat cutter, the job title speaks itself; their job is to cut, trim, and package meats. Although it doesn't require work experience or university qualification to become a meat cutter, this job depends on the skills and experience they gain throughout their career. From making a simple cut, removing bones to knowing food safety, it is all about skills and experience. It usually takes up to 2 years of on-the-job training to be considered a highly skilled meat cutter (Colleges, 2021). Besides, from the customer's perspective, they might prefer to buy meat processed by those experienced meat cutters. Hence, it also might affect the sales of the stores. To prevent the worst things happen, the company should start hiring new meat cutters to replace the older employees when they retire. Other than that, to avoid the knowledge gap, the company should start to organize training led by those older meat cutters to share their knowledge and experience with younger and less experienced employees.

## Analysis 1-3: Find which city will be affected the most due to the meat cutter's huge retirement wave

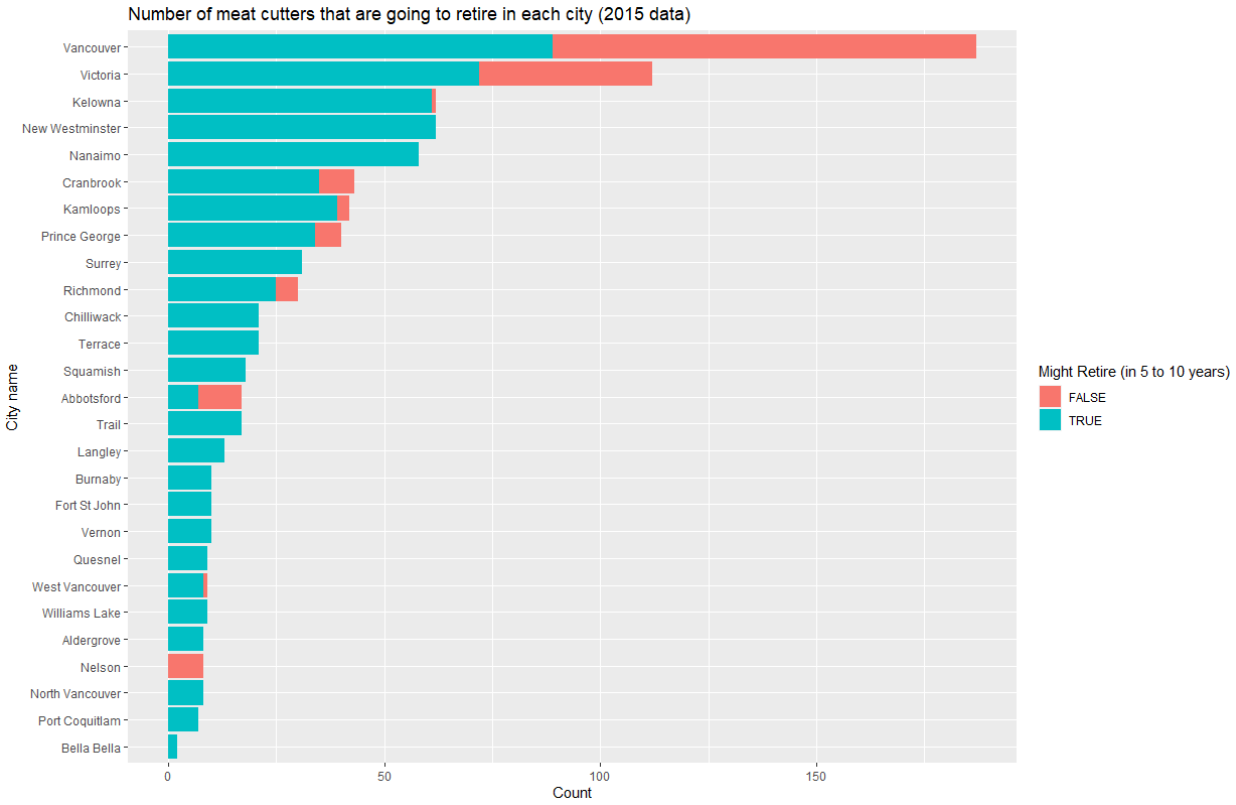



*Figure 18: Number of meat cutters that are going to retire in each city (2015)*

## Source code (Figure 18):

```
stores_retire_soon %>%
  filter(
    job_title == "Meat Cutter"
  ) %>%
  ggplot(aes(fct_rev(fct_infreq(city_name)))) +
  geom_bar(aes(fill = might_retired)) +
  labs(
    title = "Number of meat cutters that are going to retire in each city (2015 data)",
    fill = "Might Retire (in 5 to 10 years)") +
  xlab("City name") +
  ylab("Count") +
  coord_flip()
```

*Figure 19: Source code for Figure 18*

**Explanation**

*Figure 18* shows the number of meat cutters that will retire based on the cities. The order of the bars is according to the numbers of the meat cutters, which means Vancouver has the most significant number of meat cutters, and Bella Bella has the smallest number of meat cutters. Bars in turquoise color represent 55 to 65 years old employees and are expected to retire in 5 to 10 years. On the other hand, bars in red color represents the employees that are under 55 years old. The graph has shown that in the next 5 to 10 years, the number of meat cutters will decrease significantly and reach an extent that some of the city no longer has any meat cutters. All cities on the list have more than 50% of the employees who will retire soon except for Vancouver, Abbotsford, and Nelson. The worst case is that some stores' meats departments might even be forced to shut down because there will not be any meat cutters left. Besides, this graph also displays the highly uneven ratio of young employees and old employees for meat cutters. There are more older meat cutters compared to younger meat cutters. This analysis again proves the necessity to hire new meat cutters. It's not just trying to retain the expertise knowledge of those older meat cutters but also ensuring the meats department will not vanish when all those older workers retire.

**Analysis 1-4: Find which store will be affected the most due to the meat cutter's huge retirement wave**

**The number of stores in cities that has meat cutters (2015):**

| | city_name | num_of_store |
|---|---|---|
| 1 | Vancouver | 4 |
| 2 | Abbotsford | 1 |
| 3 | Aldergrove | 1 |
| 4 | Bella Bella | 1 |
| 5 | Burnaby | 1 |
| 6 | Chilliwack | 1 |
| 7 | Cranbrook | 1 |
| 8 | Fort St John | 1 |
| 9 | Kamloops | 1 |
| 10 | Kelowna | 1 |
| 11 | Langley | 1 |
| 12 | Nanaimo | 1 |
| 13 | Nelson | 1 |
| 14 | New Westminster | 1 |
| 15 | North Vancouver | 1 |
| 16 | Port Coquitlam | 1 |
| 17 | Prince George | 1 |
| 18 | Quesnel | 1 |
| 19 | Richmond | 1 |
| 20 | Squamish | 1 |
| 21 | Surrey | 1 |
| 22 | Terrace | 1 |
| 23 | Trail | 1 |
| 24 | Vernon | 1 |
| 25 | Victoria | 1 |
| 26 | West Vancouver | 1 |
| 27 | Williams Lake | 1 |

*Figure 20: Number of stores in cities that has meat cutters*

**Source code (Figure 20):**

```
emp %>%
  filter(
    STATUS_YEAR == 2015,
    job_title == "Meat Cutter",
    STATUS == "ACTIVE"
  ) %>%
  group_by(city_name) %>%
  summarise(
    num_of_store = length(unique(store_name))
  ) %>%
  arrange(desc(num_of_store)) %>%
  View()
```

*Figure 21: Source code for Figure 20*

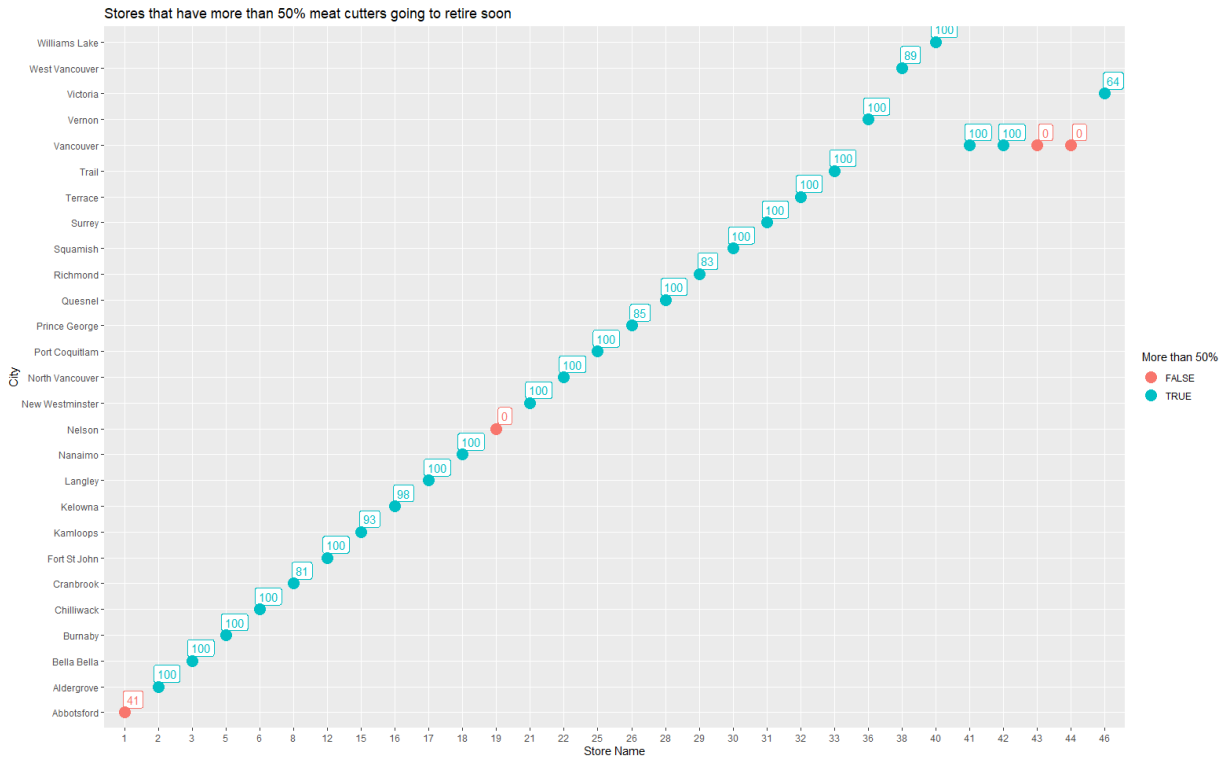



*Figure 22: The cities that have meat department in their store*

## Source code (Figure 22):

```
store_mc_retire_50 <- stores_retire_soon %>%
  filter(job_title == "Meat Cutter") %>%
  group_by(store_name) %>%
  mutate(
    total = n(),
    count = sum(might_retired == TRUE),
    percentage = round((count/total) * 100, 0)
  )

ggplot(store_mc_retire_50, aes(city_name, store_name, color = (percentage >= 50))) +
  geom_label(aes(label = percentage), nudge_x = 0.5, nudge_y = 0.25) +
  geom_point(size = 5) +
  labs(
    title = "Stores that have more than 50% meat cutters going to retire soon",
    color = "More than 50%") +
  xlab("City") +
  ylab("Store Name") +
  coord_flip()
```

*Figure 23: Source code for Figure 22*

**Explanation**

This analysis aims to determine the stores affected the most due to the vast retirement wave of meat cutters. *Figure 20* shows the number of stores in cities that had the meats department in 2015. Based on the output, the company had 30 stores that have meats department in 2015. Besides, Vancouver has 4 stores with a meats department, while the other cities on the list only have 1 store. The previous analysis shows that besides Vancouver, Abbotsford, and Nelson, the other cities will lose more than 50% of meat cutters and might be forced to shut down because there are no meat cutters left. But the truth is not like this. It turns out that the stores (41,42,43,44) in Vancouver will have 2 stores which are 41 and 42 will not have any meat cutter left due to the huge retirement wave. *Figure 22* shows the cities that have meat departments in their store and the store name. The turquoise points represent the stores with more than 50% meat cutters that are going to retire. Contrarily, points in red represent the stores that have less than 50% meat cutters that will retire in the next 5 to 10 years. This graph shows the impact on each store in the next 5 to 10 years when those older meat cutters retire from the job. The company should start hiring new meat cutters to ensure the stores will not be forced to be shut down one by one soon.

**Analysis 1-5: Find the number of middle-level management employees that are going to retire.**

Middle-level management is the liaison between top management and the rest of the company. General managers, branch managers, and department managers are some of the positions that belong to middle-level management. In this case study, these positions only appear in one business unit, which is STORES. It is simple to differentiate between ordinary employees and middle-level management due to the job titles. Those positions that have the keyword "Manager" belong to middle-level management. A new column called "level" has been added to the dataset to differentiate them in the dataset. If the job title belongs to middle-level management, their level will be "Manager"; else, it will be "Employee". Below is the code to implement it.

```
store_manager_level = c("Store Manager", "Produce Manager", "Processed Foods Manager", "Meats Manager", "Customer Service Manager", "Bakery Manager")

stores_retire_soon <- emp %>%
  filter(
    STATUS == "ACTIVE",
    BUSINESS_UNIT == "STORES",
    STATUS_YEAR == 2015
  ) %>%
  mutate(
    level = ifelse(job_title %in% store_manager_level, "Manager", "Employee"),
    might_retired = ifelse((age >=55 & age <= 65), TRUE, FALSE)
    ) %>%
  group_by(job_title)
```

*Figure 24: Source code to create a dataset that separates the employee level in STORES*


*Figure 25: Percentage of stores employee that will be going to retire in 5 to 10 years (based on level)*

**Source code (Figure 25):**

```
ggplot(stores_retire_soon, aes(job_title, fill = might_retired)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  facet_wrap(~level, scales = "free") +
  labs(
    title = "Percentage of STORES employee that will going to retire in 5 to 10 years (Based on level)",
    fill = "Might Retire (in 5 to 10 years)") +
  xlab("Job Title") +
  ylab("Percentage (%)") +
  coord_flip()
```

*Figure 26: Source code for Figure 25*

| level | count | retire_rate | percentage |
|---|---|---|---|
| Employee | 4715 | 1038 | 22 % |
| Manager | 73 | 70 | 96 % |

*Figure 27: Overall percentage of retirement in the next 5-10 years based on level*

**Source Code (Figure 27):**

```
stores_retire_soon %>%
  group_by(level) %>%
  summarise(
    count = n(),
    retire_rate = sum(might_retired == TRUE),
    percentage = paste(round((retire_rate/count) * 100, 0),"%")
  ) %>%
  View()
```

*Figure 28: Source code for figure 27*

**Explanation**

This analysis shows that within 5 to 10 years, nearly 96% of the manager-level employee who works in STORES will retire. It is a serious issue because the store or department will be leaderless once they retire from the job. These managers are responsible for overseeing the store or department's operations, providing leadership to subordinates, making sure everything runs smoothly, and more. A leaderless team is a disaster; everything will be disorganized, decisions can't be made, and no vision. In the end, the sales of those stores will drop, and the employees might choose to resign from the job to find another place to work. The worst case is that the store will be completely shut down and lead to layoffs.

Given the current situation, the company should start choosing and training new managers for the stores and department. It can be done in two ways: The first option, let the team decide their manager. Employees know what they need; they recognize the person who is suitable to lead the team. When the employees can choose their managers, it will result in having a happier workforce. A happier workforce will also lead to better productivity and lower the turnover rate. Once the employees have chosen the perfect person, the company should provide that person with adequate training so that he/she knows how to lead the team to achieve the company's mission and vision. The second option, let the company appoint. This is just like what has been mentioned in the previous analysis; the company should conduct the selection process to choose the perfect person that fits the position. But this method has a big drawback which the selected person is not approved or satisfied by the team. When the team is not satisfied with the choice, they might choose to resign and seek other opportunities.

**Conclusion (Question 1)**

Based on the findings from 5 analyses that have been performed for this question, employees will retire from the company when they are 60 or 65 years old. By knowing that, an analysis 1-2 had shown us that in the next 5 to 10 years, the company would face a huge retirement wave. Meat cutter is one of the jobs that will contribute a lot to the huge retirement wave. This is a severe issue as it will directly affect their business, and some stores might even be forced to shut down. Hence, the company must start to hire new meat cutters and allow those older meat cutters to train the younger meat cutters. This is to ensure meat cutter will still be one of the jobs available in the company and prevent the knowledge gap. As mentioned, a meat cutter typically needs to spend at least 2 years and above to become a skillful meat cutter, so 5 to 10 years is sufficient to prepare for this huge retirement wave.

The analysis also showed that the retirement wave would also include the employees from top management level like CEO, Chief Information Officer, Vice President of Human Resources, Vice President of Stores and more might retire soon. This is also a crucial issue as these positions are vital in maintaining the company's operation. The board members should start conducting top management employees' selection process to select potential candidates to replace these positions. As mentioned earlier, choosing new top management employees is not easy as it needs to have a lot of discussion and procedure, and choosing a new CEO is even more difficult and time-consuming.

Next, we also can see that many middle management level employees might be going to retire soon. As a liaison between top management and workers, they are also crucial to the company's development. The selection for those positions also needs to be done as soon as possible. Different from choosing top management-level employees, the company could give a chance to the employees to decide their leader as they know what they need the most.

Lastly, when talking about retirement, we shall not forget about the employees' benefits after they retire. The company should take the initiative to ask older employees regarding their retirement plans and try to assist them in their retirement planning.

## Question 2: What are the reasons for resignation?

The purpose of this question is to figure out the motives that caused resignation over the years. A resignation is an act of giving up a job or position formally or officially. Many reasons could lead to employee resignation, whether family issues, career development, company culture, etc. Not everyone has the same reason that makes them resign from the job. By examining the reasons for resignation, the company can gain information on how they should treat their employee, help their employees, have a chance to reflect on their mistakes, and most importantly, know how to retain their employees. Balance Careers has listed down the top 10 reasons why employees resign from their jobs (Heathfield, 2021):

1. Not having a good relationship with the boss
2. Bored by the work itself
3. Relationships with coworkers are awful
4. Fewer opportunities to use their skills and abilities
5. The contribution of their work to the organization's business goals
6. They need to work independently most of the time
7. Meaningless job
8. Low recognition of employee's performance
9. The organization's financial status is not stable
10. The culture of the organization

## Analysis 2-1: Find the relationship between age and resignation

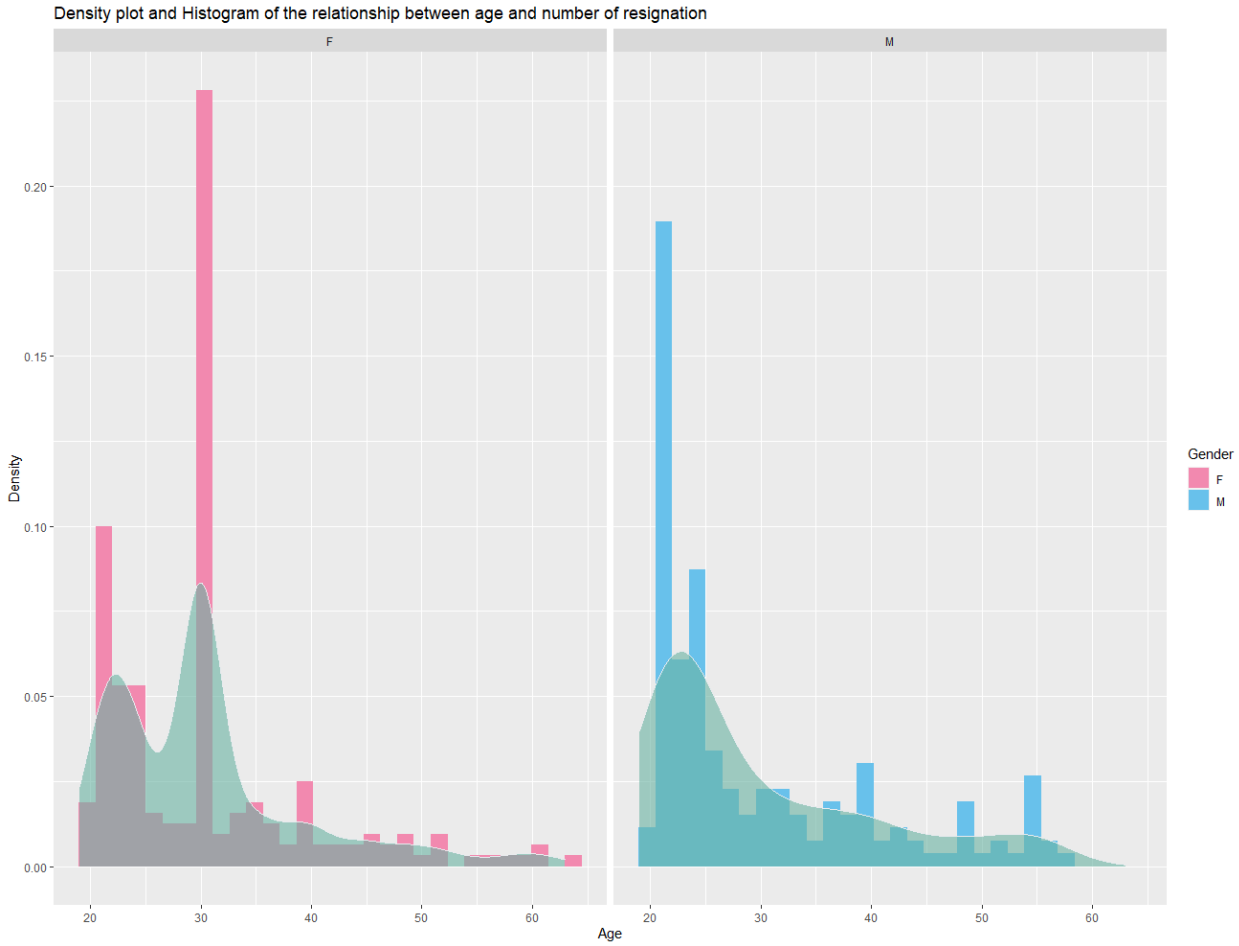



*Figure 29: Density plot and Histogram of the relationship between age and number of resignations*

## Source code (Figure 29):

```r
# ---------------------------------- Analysis 1 ----------------------------------
# Find the relationship between age and resignation
age_resign <- emp %>%
  filter(
    termreason_desc == "Resignation"
  ) %>%
  group_by(as.factor(age))

ggplot(age_resign, aes(age)) +
  geom_histogram(aes(y = ..density.., fill = gender_short), bins = 30) +
  geom_density(fill = "#69b3a2", color = "#e9ecef", alpha = 0.6) +
  facet_wrap(~gender_short) +
  labs(title = "Density plot and Histogram of the relationship between age and number of resignation", fill = "Gender") +
  xlab("Age") +
  ylab("Density") +
  scale_fill_manual(values = c("F" = "#F289AF",
                               "M" = "#68C1EB"))
```

*Figure 30: Source code for Figure 29*

**Explanation**

*Figure 29* shows the relationship between age and number of resignations in a density and histogram plot. In the graph above, we can observe a bimodal distribution in female employees where the highest peak is somewhere in the bin of 30 years old, and the second-highest peak is somewhere between 20 – 25 years old. On the other hand, Male employees' peak is somewhere in between 20 – 25 years old.

According to CollegeOnomic, 23 is the average age of college graduates for those who start college at about 18 years old (Subramanian, n.d.). Depending on the course or their degree, the time to complete their degree will be different. This might be one of the reasons that employees around 20-25 years old have a higher resignation rate might be their studies. Humans have different priorities and responsibilities in every stage of life. Usually, a student's prioritization would be school, family, sports, extracurriculars, etc. Hence, it will be hard to balance their work and school, so they will choose to resign from the job and focus on their studies. Another assumption is that these employees might want to gain some working experience before they graduate; hence they only work for the company for a short period. These could be in the form of an internship, temporary summer jobs, and so on. Research done by Highfliers concludes that graduates without any previous work experience are most likely to be unsuccessful during the selection process and have a relatively lower chance to receive the job offer regardless of their academic achievements or the school they attended (Highfliers, 2020).

Other than that, employees between 20-25 years old might also represent those fresh graduates that just entered the workforce. LinkedIn survey shows that young employees do more job-hopping than any other generation (Berger, 2016). A survey conducted by Robert Half, the largest specialized staffing firm, shows that 63% of young adult (18 – 34 years old) employees admitted that job-hopping is beneficial for their career as it could provide them short-term advantages. Hence, these conclude that job-hopping at the age of 20-something is the norm. Below are some of the factors that caused the young employee job-hopping trend:

1. **They don't feel like they're learning** (Paul, 2012)**.** Research conducted by consultant Beth N. Carver, who focused on researching exit interviews, concludes that the lack of training opportunities and lack of mentorship in the workplace are the two main motives that lead to young employees' job-hopping trend.

2. **New opportunity arises.** Younger employees are job-hopping because they find a better opportunity for their career growth. Unlike older employees, young employees have flexibility and mobility as they haven't got into married life. Hence, they can explore around and grab new opportunities as much as they discover.
3. **They take jobs that are the wrong fit.** This factor might be caused by multiple reasons such as the organization environment, company culture, no opportunity to utilize their skills and knowledge, the job doesn't match what is expected, etc.
4. **Unsure about their career path.** After graduating from school, they might spend lots of time switching jobs or experiencing new jobs until they find a job that matches their dream.

**Analysis 2-2: Find the relationship between age of first marriage and resignation**

**Explanation**

According to Statistics Canada, the most recent data(2008) reveals that the average age of first marriage for men and women is rising (Statistics Canada, 2013). The data only focused on the young adults within 20-34 years old, and by examining the trend, it is expected that the average age of first marriage will slowly grow to around 30 years old or older in the future. A report indicates that 43% of women quit their jobs after getting into married life to raise their families (Alley, 2016). Hence, these statistics have explained why *Figure 29* shows females around 30 years old resigned from the company because this group of employees represented those married females who chose to resign from the job to take care of the family.

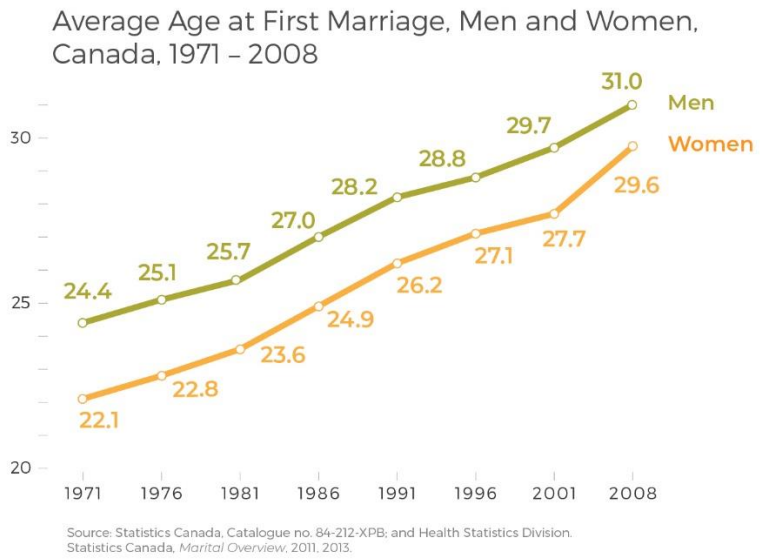



*Figure 31: Other statistics on the average age at first marriage in Canada*

Again, multiple reasons caused married women to leave the workforce. Below are some of the reasons:

1. **The company is not friendly toward married women.** It is difficult for married women to balance work and family as they have extra responsibilities to take care of in their life. They can't go back to the 100-hour workweeks. If the company doesn't provide any benefits for married women, it will force them to resign from the job.

2. **The needs of taking care of the family.** After married, they have more things to take care of, such as their children, especially newborns and babies, and their own or husband's family. This has forced them to leave the workforce even though it is the best job for them.
3. **Their family does not support them.** "If a woman wants a happy marriage, she has to leave behind her career and submit herself to being a good wife and mother." (Essig, 2012) We can't deny that such thought still exists in the current generation to force women to stay at home after marriage, even though it sounds like what we hear in our grandmother's generation. Due to this, they were forced to resign from their job to devote 100% of their time to the family.
4. **Their husband can financially support the family.** When they find out that their husbands can financially support the family, they might choose to resign and stay at home to take care of the family. Fathers are most likely to be promoted or get a raise, so it makes them earn more after marriage, researchers said (Alley, 2016).

### Analysis 2-3: Find the relationship between department and resignation

| | BUSINESS_UNIT | total_resignation |
|---|---|---|
| 1 | HEADOFFICE | 1 |
| 2 | STORES | 384 |

*Figure 32: Number of resignations in each BUSINESS UNIT*

*Figure 32* displays the number of resignations in HEADOFFICE and STORES over 10 years. It turns out that resignation most likely happened in STORES as HEADOFFICE only has one case over 10 years. Hence, HEADOFFICE will be excluded in the later analysis as it is no longer one of the focuses.

```
emp %>%
  filter(
    termreason_desc == "Resignation"
  ) %>%
  group_by(BUSINESS_UNIT) %>%
  summarise(
    total_resignation = n()
  ) %>%
  View()
```

*Figure 33: Source Code for Figure 32*

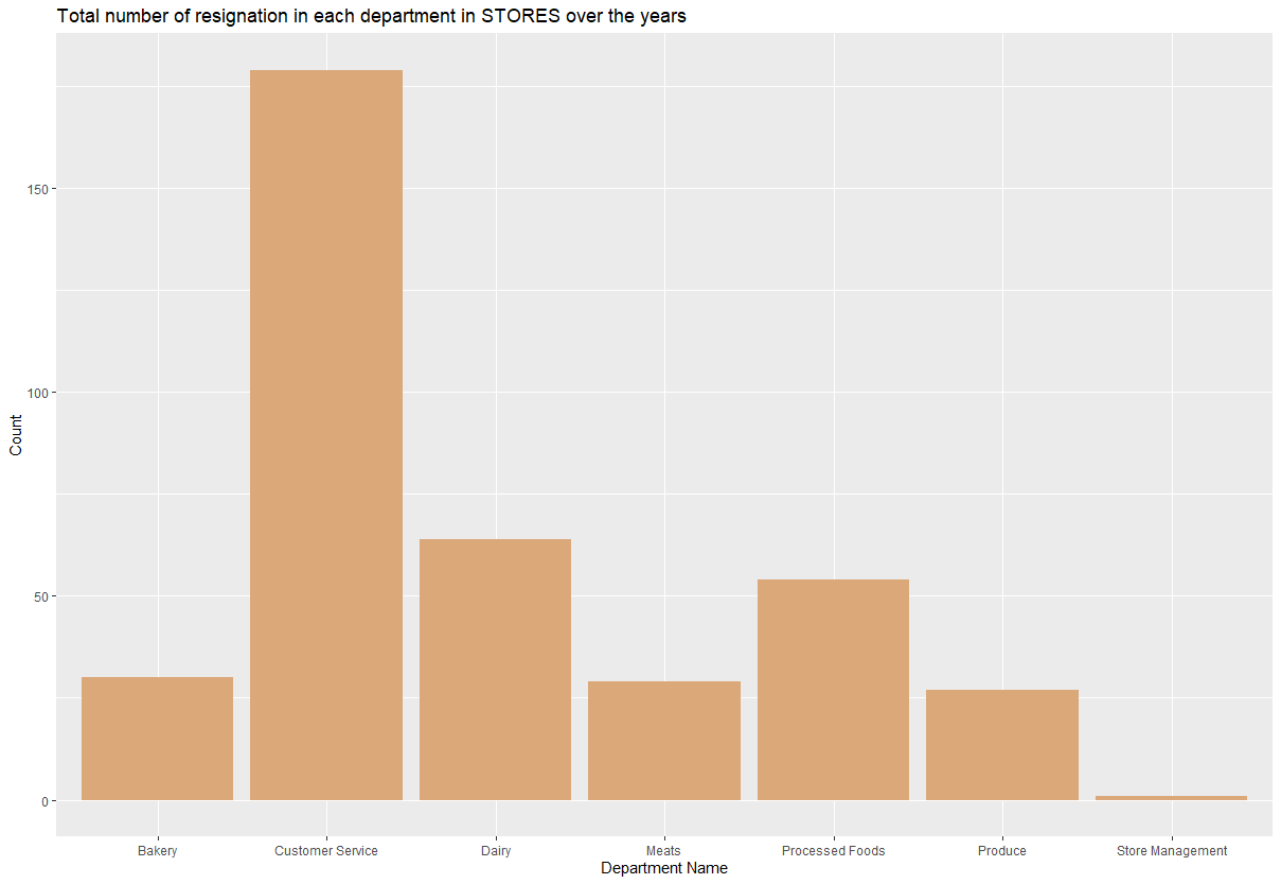



*Figure 34: Total number of resignations in each department in STORES over the years*

**Source code (Figure 34):**

```
store_resignation <- emp %>%
  filter(
    termreason_desc == "Resignation",
    BUSINESS_UNIT == "STORES"
  ) %>%
  group_by(department_name)

ggplot(store_resignation, aes(department_name)) +
  geom_bar(fill = "#dba879") +
  labs(title = "Total number of resignation in each department in STORES over the years") +
  xlab("Department Name") +
  ylab("Count")
```

*Figure 35: Source code for Figure 34*

**Explanation**

*Figure 34* shows that the Customer Service department has the highest number of resignations compared to other departments, followed by Dairy and Processed Food. As shown in *Figure 36*, two jobs belong to the Customer Service department: Customer Service Manager and Cashier. Based on the graph, the Cashier is the only contributor to the number of resignations in Customer Department over the years.

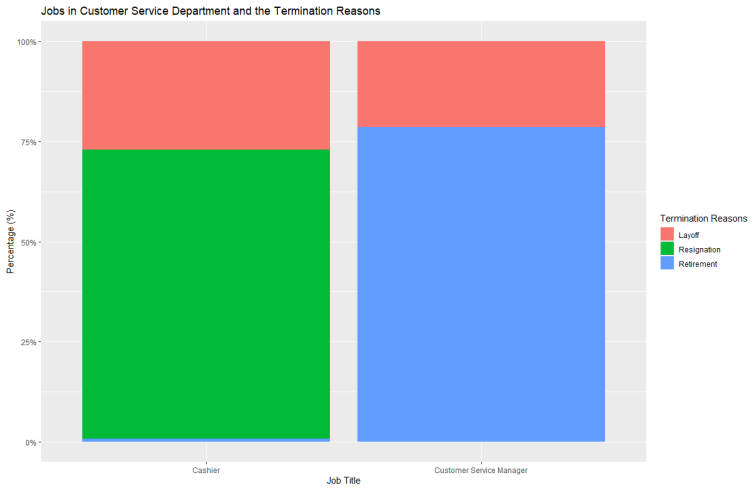



*Figure 36: Jobs in Customer Service Department and the termination reasons*

**Source code (Figure 36):**

```
emp %>%
  filter(
    department_name == "Customer Service",
    termreason_desc != "Not Applicable"
  ) %>%
  group_by(STATUS_YEAR) %>%
  ggplot(aes(job_title, fill = termreason_desc)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Jobs in Customer Service Department and the Termination Reasons", fill = "Termination Reasons") +
  xlab("Job Title") +
  ylab("Percentage (%)")
```

*Figure 37: Source code for Figure 25*

According to a report done by Agile Craft, cashiers are at 97% risk of being impacted by automation (Forrest, 2018). It is the highest among the rest of the professions in the report. Several economists also indicate that cashiers will most likely disappear in the next decade and will take over by automation (Wenger, 2019). Job security is one of the crucial things when it comes to our job. It is also one of the factors for safety needs in Maslow's Hierarchy of Needs. Not feeling safe and fearing being replaced is unhealthy for employees' mental health and might lead to lower productivity. This could be one of the reasons that cashiers are resigning from the company as the automated system is slowly replacing their jobs, meaning their job security has been threatened. Big companies like Giant, Walmart, and others have started to install self-checkout lines. Walmart even planned to go further by going entirely self-checkout by the end of 2021, meaning fewer cashiers will be in their stores. Hence, instead of waiting for the company to eliminate themselves soon, the employees chose to resign to look for another opportunity.

## Analysis 2-4: Find the relationship between the number of cashiers and resignation

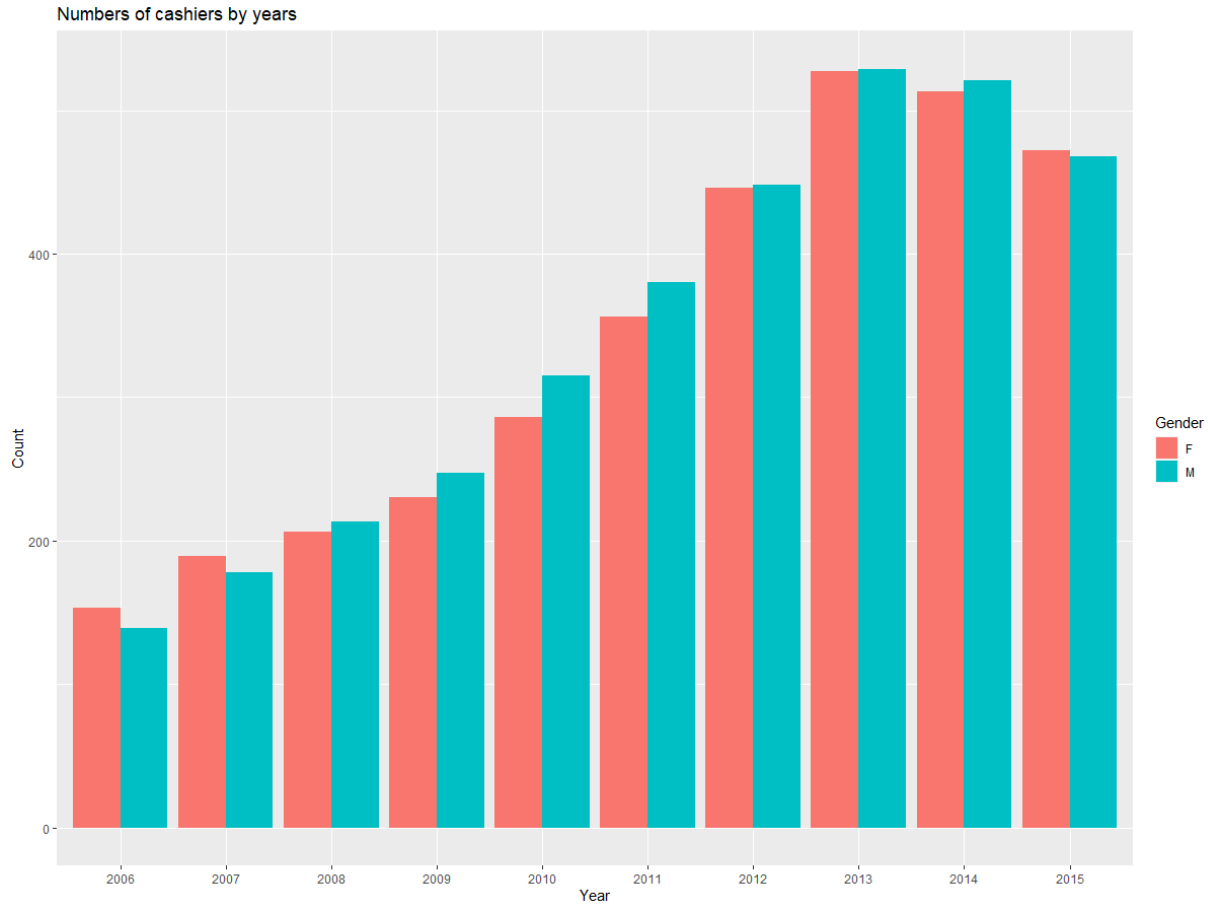



*Figure 38: Number of cashiers by years*

## Source code (Figure 38):

```
emp %>%
  filter(
    job_title == "Cashier"
  ) %>%
  group_by(STATUS_YEAR) %>%
  ggplot(aes(STATUS_YEAR, fill = gender_short)) +
  geom_bar(position = "dodge") +
  labs(title = "Numbers of cashiers by years", fill = "Gender") +
  xlab("Year") +
  ylab("Count")
```

*Figure 39: Source code for Figure 38*

**Explanation**

Cashiers contribute a lot in the number of resignations might also lead by another issue. *Figure 38* shows the number of cashiers every year. As we can see, the number of cashiers is increasing every year. The number of cashiers only started to decrease in 2014. Before 2014, the reasons why cashiers resigned from the job might be highly competitive in the job. The job can consider as an oversaturated position which makes them hard to stand out. In the end, the employees will choose to resign from the job.

A suggestion for the company is that if the company is planning to replace cashiers with an automated system, they should start hiring fewer cashiers to save labor costs and invest in better technology. Contrarily, if the company is not planning to implement the automated system in their stores, they also need to try not to hire too many cashiers in the future so that the positions will not be oversaturated.

**Analysis 2-5: Find the relationship of the class of the city and the number of resignations**

Cities can be classified into 3 main categories based on the Degree of Urbanization: cities, towns, and rural areas. Degree of Urbanization classifies the cities based on the numbers of the inhabitants or the density of inhabitants per $km^2$. Below are the types of settlements in the Degree of Urbanization (Dijkstra et al., 2020):

1. **Cities –** The population needs to have at least 50,000 inhabitants.
2. **Town –** The population needs to reach at least 5,000 inhabitants and below 50,000 inhabitants.
3. **Rural area –** Any city that has the number of population that is lesser than 5,00 inhabitants.

For this analysis, the number of populations of each city will be based on the statistics in 2020.

| Number of Population (2020) | |
|---|---|
| **City Name** | **Population** |
| Vancouver | 600,000 |
| Terrace | 19,443 |
| Nanaimo | 84,905 |
| Nelson | 9,813 |
| Kelowna | 125,109 |
| Victoria | 289,625 |
| Kamloops | 68,714 |
| Fort St John | 17,402 |
| Surrey | 394,976 |
| Vernon | 47,274 |
| Quesnel | 13,788 |
| Chilliwack | 77,000 |
| Dawson Creek | 10,802 |
| Squamish | 58,549 |
| New Westminster | 85,590 |
| Port Coquitlam | 63,508 |
| Cortes Island | 1,050 |
| Burnaby | 202,799 |
| Bella Bella | 1,106 |
| Cranbrook | 18,610 |
| Williams Lake | 14,168 |
| Trail | 9,707 |
| Prince George | 65,558 |
| Richmond | 182,000 |
| Grand Forks | 59,166 |
| West Vancouver | 42,694 |

| | |
|---|---|
| Abbotsford | 151,683 |
| Aldergrove | 12,363 |
| Langley | 23,606 |
| North Vancouver | 48,000 |
| White Rock | 66,450 |
| New Westminster | 58,549 |
| Fort Nelson | 3,366 |
| Haney | 21,041 |
| Valemount | 1,021 |
| Ocean Falls | 50 |
| Princeton | 30,681 |
| Dease Lake | 400 |
| Pitt Meadows | 17,410 |
| Blue River | 877 |

```r
# city population
emp_with_pop <- emp %>%
  mutate(
    population = case_when(
      city_name == "Vancouver" ~ 2581000,
      city_name == "Terrace" ~ 12700,
      city_name == "Nanaimo" ~ 392651,
      city_name == "Nelson" ~ 52900,
      city_name == "Kelowna" ~ 127380,
      city_name == "Victoria" ~ 408883,
      city_name == "Kamloops" ~ 80376,
      city_name == "Fort St John" ~ 22283,
      city_name == "Surrey" ~ 580360,
      city_name == "Vernon" ~ 108,
      city_name == "Quesnel" ~ 9879,
      city_name == "Chilliwack" ~ 83788,
      city_name == "Dawson Creek" ~ 12178,
      city_name == "Squamish" ~ 21273,
      city_name == "New Westminster" ~ 82590,
      city_name == "Port Coquitlam" ~ 63508,
      city_name == "Cortes Island" ~ 1050,
      city_name == "Burnaby" ~ 232755,
      city_name == "Bella Bella" ~ 1106,
      city_name == "Cranbrook" ~ 21308,
      city_name == "Williams Lake" ~ 10753,
      city_name == "Trail" ~ 552,
      city_name == "Prince George" ~ 82290,
      city_name == "Richmond" ~ 17250,
      city_name == "Grand Forks" ~ 59166,
      city_name == "West Vancouver" ~ 43805,
      city_name == "Abbotsford" ~ 161581,
      city_name == "Aldergrove" ~ 15000,
      city_name == "Langley" ~ 27000,
      city_name == "North Vancouver" ~ 89767,
      city_name == "White Rock" ~ 5751,
      city_name == "New Westminister" ~ 82590,
      city_name == "Fort Nelson" ~ 3366,
      city_name == "Haney" ~ 21041,
      city_name == "Valemount" ~ 1021,
      city_name == "Ocean Falls" ~ 50,
      city_name == "Princeton" ~ 30681,
      city_name == "Dease Lake" ~ 400,
      city_name == "Pitt Meadows" ~ 19717,
      city_name == "Blue River" ~ 877
    ),
    city_class = case_when(
      population < 5000 ~ "rural",
      population >= 5000 & population < 50000 ~ "town",
      population >= 50000 ~ "city"
    ),
    city_class = as.factor(city_class),
    city_class = ordered(city_class, c("city", "town", "rural"))
  ) %>%
  select(everything(), -c(population))
```

*Figure 40: Create a new dataset with population numbers and classify the city*

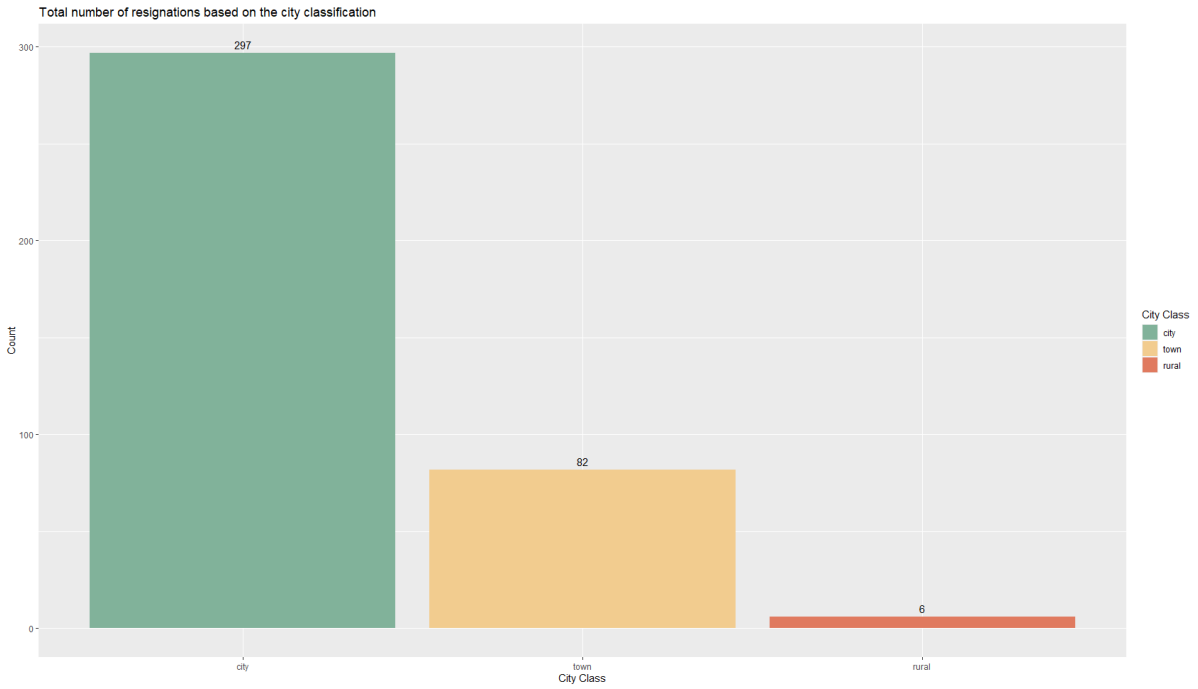



*Figure 41: Total number of resignations based on the city classification*

**Explanation**

*Figure 41* shows the total number of resignations over the years based on the city classification. Employees that work in the city have the highest number of resignations compared to the other. The x-axis represents the city class, classified based on the population's number and the type of settlements in the Degree of Urbanization.

Compared to towns and rural areas, cities always have more job opportunities, and most high-paying job openings are available there (Tucker, n.d.). When there is a tremendous amount of job opportunity, meaning that the choices will be more. Other than that, more job opportunities also lead to higher job-hopping rates. Research has indicated that job-hopping is a big issue in urban areas and cities (Larasati & Aryanto, 2020). If the employee does not satisfy the benefits or the salary they received in the company, they can simply just resign from the job and look for another opportunity if they have the capability. Multiple articles show that job-hopping has advantages too. E.g.: (Jobstore, n.d.); (Resume Professional Writers, n.d.); (Indeed, 2021). Below are some of the advantages of job-hopping:

1. **Get a higher salary.** This is the primary reason for job hop. When the employee changes their job often, the salary could also increase.

2. **Good for career growth.** Switching jobs often allows the employees to land a higher-level job at another company. The employees can grow their career quickly without spending years on getting a promotion.
3. **Gain more experience and knowledge.** Job-hopping could allow the employees to gain more experience and knowledge. Working in different companies provides them not only experiences and knowledge but also resources. All of these are highly beneficial for their career development.
4. **Have a wider network.** Having a broader network is a crucial thing in career development. Job-hopping allows the employees to meet new people and expand their network.

But when it comes to town and rural areas, job opportunity is lesser than in urban areas or cities. They only have a few choices and were forced to stick with them as there are no other choices. Hence, they will stay on a job longer than the employees who work in cities, which means a lower resignation rate.

**Conclusion (Question 2)**

After knowing the findings from 5 analyses from this question, a conclusion is made that a few possible factors might lead to employee resignation.

| No. | Analysis | Findings |
|---|---|---|
| 2-1 | Find the relationship between age and resignation | Both male and female employees that are around 20-25 years old have a higher resignation rate. This might be due to the job-hopping that usually happens to employees around that age range. Another assumption is that this group of employees might represent those students who interned in the company before for a short period. |
| 2-2 | Find the relationship between age of first marriage and resignation (A continuous analysis of Analysis 2-1) | Based on the result in analysis 2-1, female employees around 30 years old also have a high resignation rate. The average age of the first marriage in Canada is around 30 years old, so these employees might represent those female employees who just got into marriage life and chose to resign from their job to take care of the family. |
| 2-3 | Find the relationship between department and resignation | The Customer Service department has the highest resignation rate among the others. There are two job positions under the Customer Service department: Customer Service Manager and Cashiers. The Cashier is the only job that contributes to the enormous number of resignations in the Customer Service department. This might be caused by the automation that is slowly replacing the job of Cashier. Due to the fear of being replaced, cashiers are resigning to look for another opportunity. |

| | | |
|---|---|---|
| 2-4 | Find the relationship between the number of cashiers and resignation<br>(A continuous analysis of Analysis 2-3) | The number of cashiers in the company is increasing every year except in 2014 and 2015. This might be one of the reasons why cashiers are leaving the company because the competition is getting bigger. |
| 2-5 | Find the relationship of the class of the city and the number of resignations | The number of resignations in the city is more significant than in towns and rural areas. This might be due to the high number of job opportunities in the city, which leads to job-hopping. Employees in towns and rural areas have a lower resignation rate because they only have a few choices and fewer job opportunities. |

## Question 3: What are the reasons for layoffs?

The purpose of this question is to figure out what are the reasons that caused layoffs over the years. The definition of being laid off is that the employee might be terminated temporary or permanent because of some reasons or issues that is related to the business. This can be in a form of just one employee being laid off or a group of workers at the same time. When the company is trying to terminate a big group of people, we can examine that group of employees to identify is there any discriminations exists in the layoffs. Discrimination in layoffs is illegal and should be avoided. Below are some of the reasons that the employees would be laid off:

1. Cost reduction
2. Staffing redundancies
3. Relocation
4. Merger or buyout
5. Downsizing
6. Outsourcing
7. Automation
8. Going out for business

Some of the common discrimination when comes to layoffs:

1. Age discrimination
2. Gender discrimination
3. Religion discrimination
4. Nationality discrimination
5. Race discrimination
6. Disability discrimination

## Analysis 3-1: Find which job has the highest number of layoffs

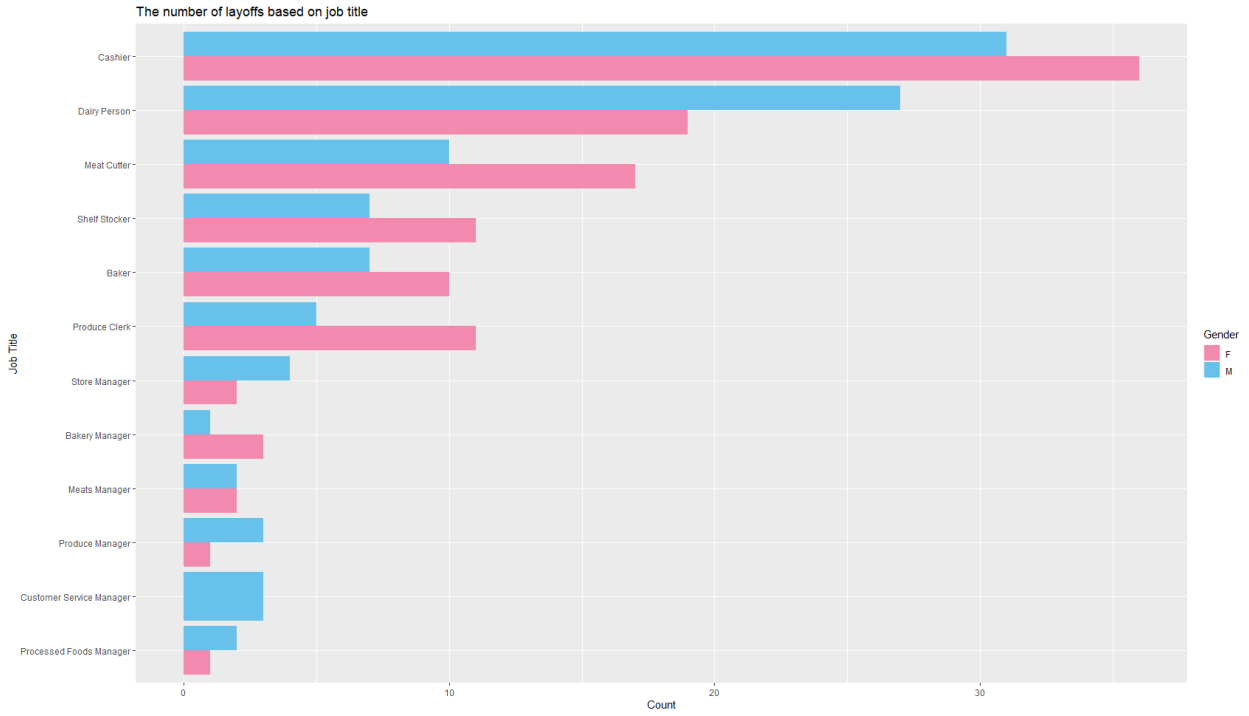



*Figure 42: The number of layoffs based on job title and categorized by gender*

## Source code (Figure 42):

```
emp %>%
  group_by(STATUS_YEAR) %>%
  filter(
    termreason_desc == "Layoff",
  ) %>%
  ggplot(aes(fct_rev(fct_infreq(job_title)), fill = gender_short))+
  geom_bar(position = "dodge") +
  labs(title = "The number of layoffs based on job title", fill = "Gender") +
  xlab("Job Title") +
  ylab("Count") +
  scale_fill_manual(values = c("F" = "#F289AF",
                               "M" = "#68C1EB")) +
  coord_flip()
```

*Figure 43: Source code for Figure 42*

This analysis is to find out the number of layoffs according to the jobs. By examining *Figure 42*, Cashier has more layoffs than the other position, followed by Dairy Person and Meat Cutter. The colors of the bar represent gender; light blue represents male, and pink represents female. Based on the graph, the reason Cashier, Dairy Person, and Meat Cutter have high numbers of layoffs might be due to automation. With automation, the number of employees has reduced because there is no longer a need for that many people.

As robotics and artificial intelligence advance, some of the jobs will slowly be replaced by those technologies. The Cashier is one of the jobs that will get a huge impact and might even vanish soon. Self-check-out is a trend right now. Companies like Amazon already use innovative technology to achieve 100% cashier-less (Thorbecke, 2020). The shoppers can just walk in the store, take what they need, and leave the store! The payment will be made automatically as they walk out of the store without interacting with a cashier. So, the company in this case study might have invested and started using self-check-out machines in their stores, so they chose to lay off many cashiers to avoid staff redundancies. As mentioned in the previous question, a report by Agile Craft stated that cashiers are at 97% risk of being impacted by automation (Forrest, 2018), among other jobs. In another report mentioned earlier, economists say that cashiers will disappear in the next decade due to automation (Wenger, 2019). Hence, this explains why Cashier has the highest layoffs among the other jobs in the graph.

As for Dairy Person, automation is also taking its place. Technology is becoming progressively crucial in Canadian dairy farming (Dairy Central, 2020). Dairy farms are slowly adopting technologies to assist them in harvesting milk and producing dairy products. Using those technologies, the quality of the milk and dairy products improved, and human cost was reduced. The daily job of a dairy person includes taking care of the cows. The use of sophisticated software allows the dairy person to know the real-time data about their cows like stress, behavior, location, hygiene, health, and more in a glance without using several hours figure out (Dairy Central, 2020). These are not the only few advantages, obviously, but we can see that automation in this field is beneficial for the company. Hence, there is no need to have many dairy persons anymore, so that the company chose to layoffs dairy person.

Robotics and automation also come to help in meat cutting, meat processing, and such. Meat cutting can be considered a dangerous job in food manufacturing for both the meat cutters and the product safety. By using technology in meat processing, the efficiency will be increased, and it also can ensure the safety of the workers. For Meat Cutters, it's a similar case to Dairy Person. The graph's result might show that the company is slowly importing machines for the meat cutting or processing operations. Hence, there is many meat cutters have been laid off due to automation.

## Analysis 3-2: Are there any age discrimination layoffs that happened in the company?

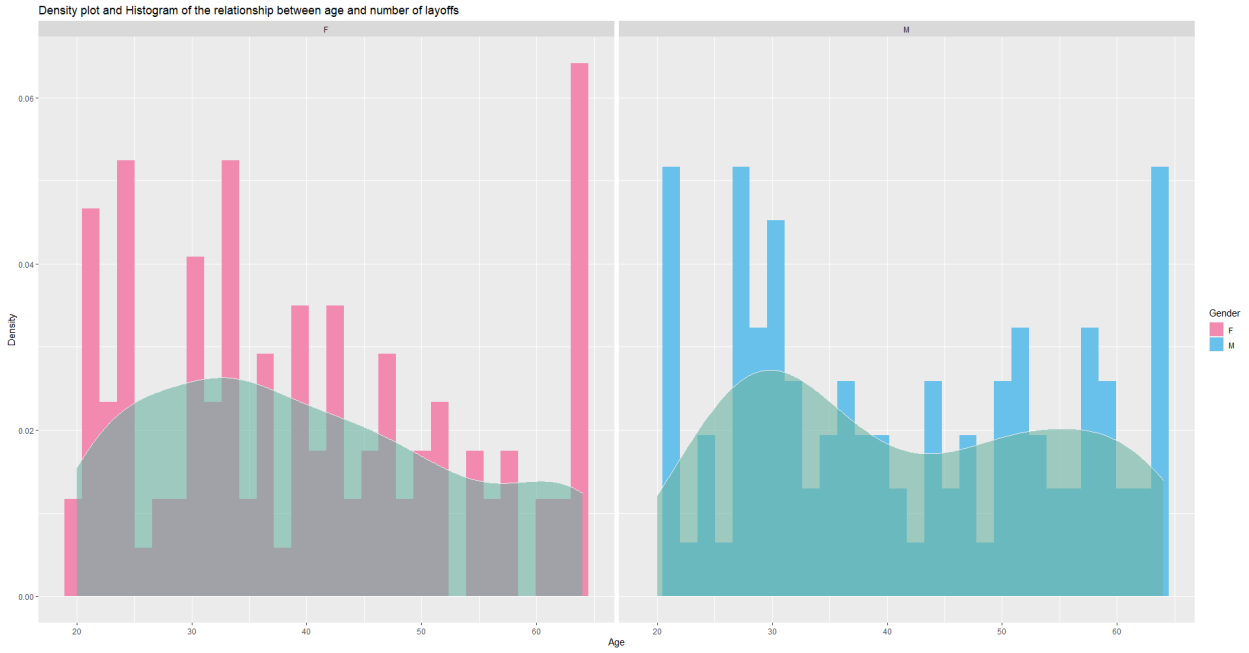



*Figure 44: Density plot and Histogram of the relationship between age and number of layoffs*

## Source code (Figure 44):

```
# Find the relationship between age and layoffs
age_layoff <- emp %>%
  filter(
    termreason_desc == "Layoff"
  ) %>%
  group_by(as.factor(age))

ggplot(age_layoff, aes(age)) +
  geom_histogram(aes(y = ..density.., fill = gender_short), bins = 30) +
  geom_density(fill = "#69b3a2", color = "#e9ecef", alpha = 0.6) +
  facet_wrap(~gender_short) +
  labs(
    title = "Density plot and Histogram of the relationship between age and number of layoffs",
    fill = "Gender") +
  xlab("Age") +
  ylab("Density") +
  scale_fill_manual(values = c("F" = "#F289AF",
                               "M" = "#68C1EB"))
```

*Figure 45: Source code for Figure 44*

## Justification

Discrimination in layoffs is entirely possible. The employees might be laid off due to their race, color, religion, sex, disability, age, and more. The purpose of this analysis is to determine whether there is a sign of age discrimination that existed in the layoff cases. *Figure 44* displays the density and histogram plot of the relationship between age and the number of layoffs. Starting

from female employees, layoffs mostly happened on female employees around 30 to 35 years old, as the values are concentrated over that range. Male employees, around 30 and 55 years old are most likely to experience a layoff. Another finding from the graph is that both male and female employees 65 years old have the highest number of layoffs in both facets.

In both facets, we can see those young employees around 30± years old seem to have a high layoff rate regardless of gender. This could be a trick that the company may use to cover up older employees' large-scale layoffs (HR.org, n.d.). In the retirement analysis, we have acknowledged that most employees will retire from the job either at 60 or 65 years old. It is possible that employees already at the age of 60-65 still have any retirement plan, so they just stay in the company until the employers are forced to lay off them. Canada does not have any law that restricts employees above 60 or 65 years old must retiring. It's just that most of the people will retire within that age range. So, it's all about personal choice and their plan. From the company's perspective, they might want to eliminate those older employees to welcome young employees equipped with the latest skills and knowledge. Hence, the company will choose to lay off those older employees. But to avoid age discrimination, the company will also lay off some younger employees to let older employees feel that the company discriminates against older employees.

**Analysis 3-3: Are there any pregnancy discrimination layoffs that happened in the company?**

**Justification**

Female employees around 30 years old might also face another kind of discrimination, which is pregnancy discrimination. According to Statistics Canada, the latest data (2016) shows that the average age of mothers at childbirth has been over the age of 30.8 years old, and it has been rising steadily (Statistics Canada, 2018). This statistic has matched what we have discovered in *Figure 44*. Usually, the termination of pregnant women in the workplace might be due to several reasons. Firstly, the company might grant pregnant or afterbirth women a break from work to temporarily lay them off. After they get enough rest and completely recover, they are still able to return to the workforce. Secondly, the company is just discriminating against pregnant or afterbirth women. Some companies do not welcome pregnant or afterbirth women as they think their productivity will be lower than male employees. Other than that, most employers will think that those pregnant and after birth women will be more focused on the family instead of work, so they choose to lay them off. But due to lack of information such as the type of layoff (temporary/permanent), it is hard to conclude that the company has pregnancy discrimination.

## Analysis 3-4: Are there any gender discrimination layoffs that happened in the company?

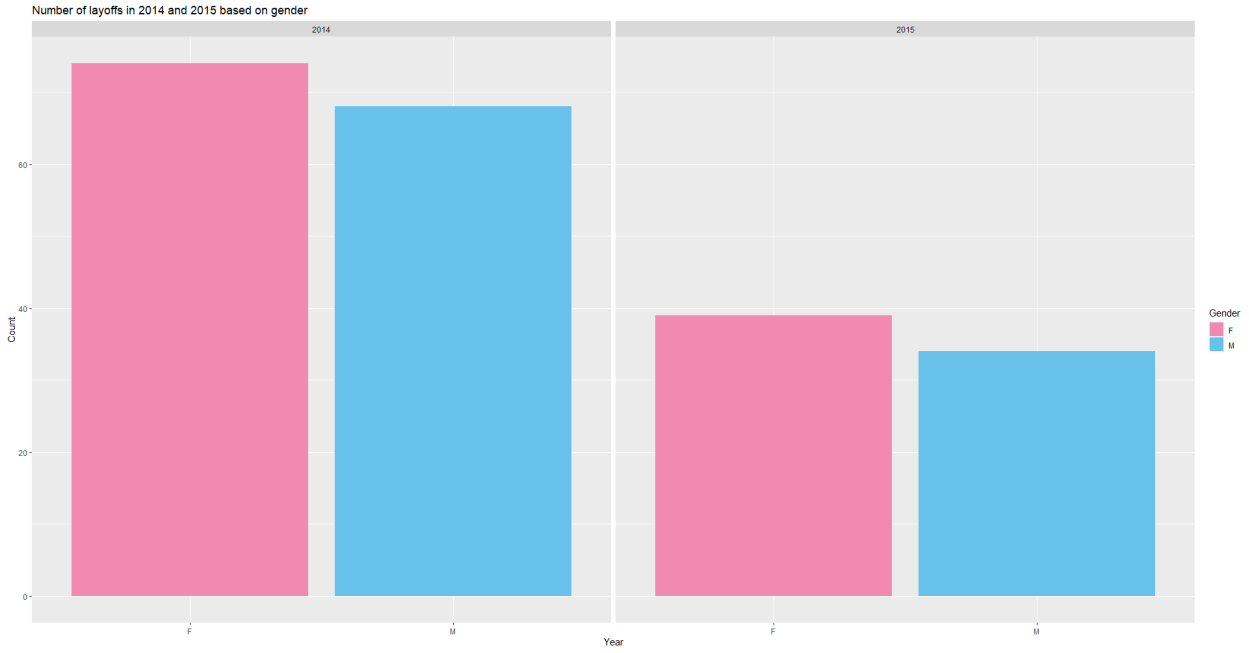



*Figure 46: Number of layoffs in 2014 and 2015 based on gender*

## Source code (Figure 46):

```
# Find the relationship between gender and layoffs
emp %>%
  filter(
    termreason_desc == "Layoff"
  ) %>%
  group_by(gender_short) %>%
  ggplot(aes(gender_short, fill = gender_short)) +
  geom_bar() +
  facet_wrap(~STATUS_YEAR) +
  labs(title = "Number of layoffs in 2014 and 2015 based on gender", fill = "Gender") +
  xlab("Year") +
  ylab("Count") +
  scale_fill_manual(values = c("F" = "#F289AF",
                               "M" = "#68C1EB"))
```

*Figure 47: Source code of Figure 46*

**Justification**

This analysis aims to discover whether there is a sign of gender discrimination by examining the number of layoffs based on gender. By observing *Figure 46*, the number of females laid off for the past 10 years is higher than males, but this is not a severe problem as it only surpasses a little. Hence, we can say that gender no sign proves that the employer in this company will lay off employees based on their gender. This could be a good sign for the company as the employee will not feel discriminated against by the gender as they are being laid off.

**Conclusion (Question 3)**

Based on the 4 analyses on the cause of layoff, a conclusion is that those who experienced laid off might be due to automation and staff redundancies. Besides, the result of the analysis shows that there may be age discrimination in layoffs.

| No. | Analysis | Findings |
|---|---|---|
| 3-1 | Find which job has the highest number of layoffs | Cashier, Dairy Person, and Meat Cutter are the jobs that have a high layoff rate. This might show that automation is slowly taking its place, and several articles show that these fields are slowly being replaced by robotics and automation. |
| 3-2 | Are there any age discrimination layoffs that happened in the company? | Based on the result, employees who are either female or male at around 30 years old have a high layoff rate. Besides that, something unusual is that employees at 65 years old, regardless of gender, have the highest layoff rate compared to the others. This abnormal finding could be a technique the company used to cover up many layoffs in older workers. Canada does not have any laws that restrict employees should retiring from work after reaching a certain age. Hence, there is a sign of age discrimination in layoffs. |
| 3-3 | Are there any pregnancy discrimination layoffs that happened in the company? | Canadian women tend to have their first child around the age of 30. The result shows that females around that age have a higher layoff rate, so it might be pregnancy discrimination in layoffs. But, the dataset doesn't provide information on whether the layoff is temporary or permanent, so we only can |

| | | |
|---|---|---|
| | | suspect that the company might have pregnancy discrimination against pregnant or afterbirth women. |
| 3-4 | Are there any gender discrimination layoffs that happened in the company? | Based on the graph, female employees seem to have a higher layoff rate. But when compared with males' layoff rate, it only has a slight difference. To conclude, the company does not have any gender discrimination in layoffs. |

## Extra Features

**Extra Feature 1:** `fct_infreq()`

`fct_inreq()` is a function from the `forcats` package, which is one of the packages in `tidyverse`. This function helps to reorder the factor levels according to the frequency of the observations. Using this function in the visualization could increase the readability of the graph and help us do observation better. Below is the difference between using it in the graph and without using it.

*Without the function:*

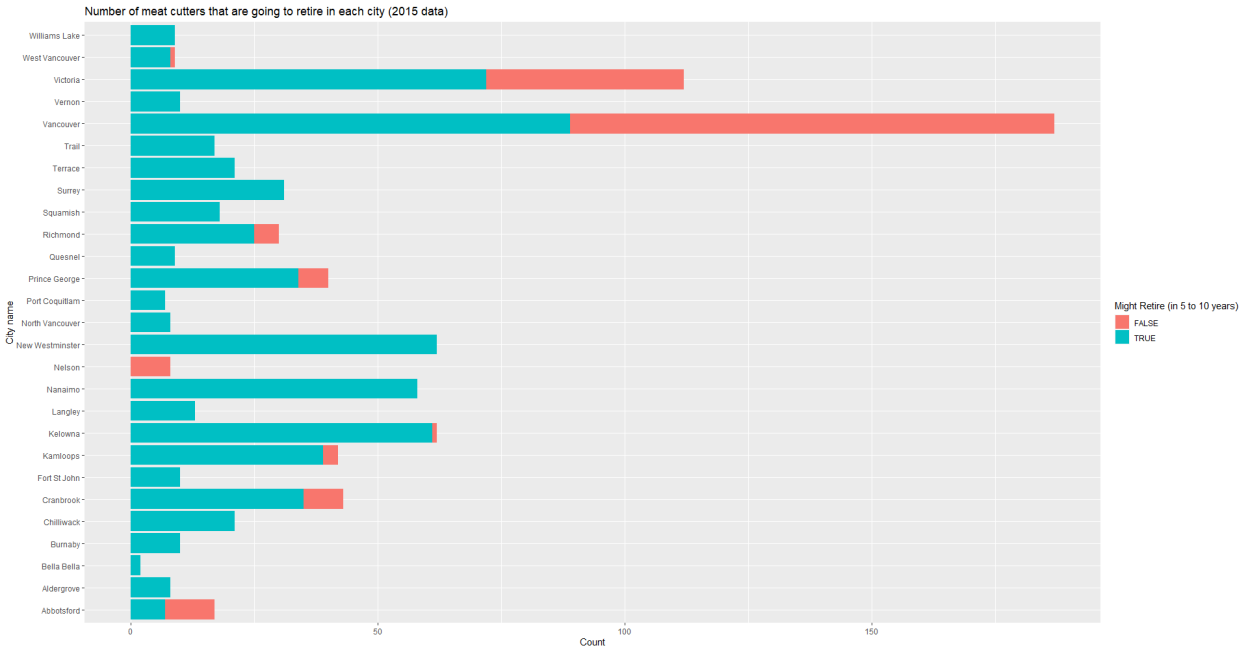



*Figure 48: Without using fct_infreq()*

As shown in *Figure 48*, the result is not following the frequency of the observations but follows the reverse alphabetical order of the y-axis. The graph is messy, and hard to determine the ranking. Although it is obvious to identify the one with the highest number, some seem to have the same number of observations, but we can't be sure. For example, the New Westminster and Kelowna in the graph seems to have the same number of observation, and one of them will be the third-highest in the graph, but I can't be sure.

*With the function:*

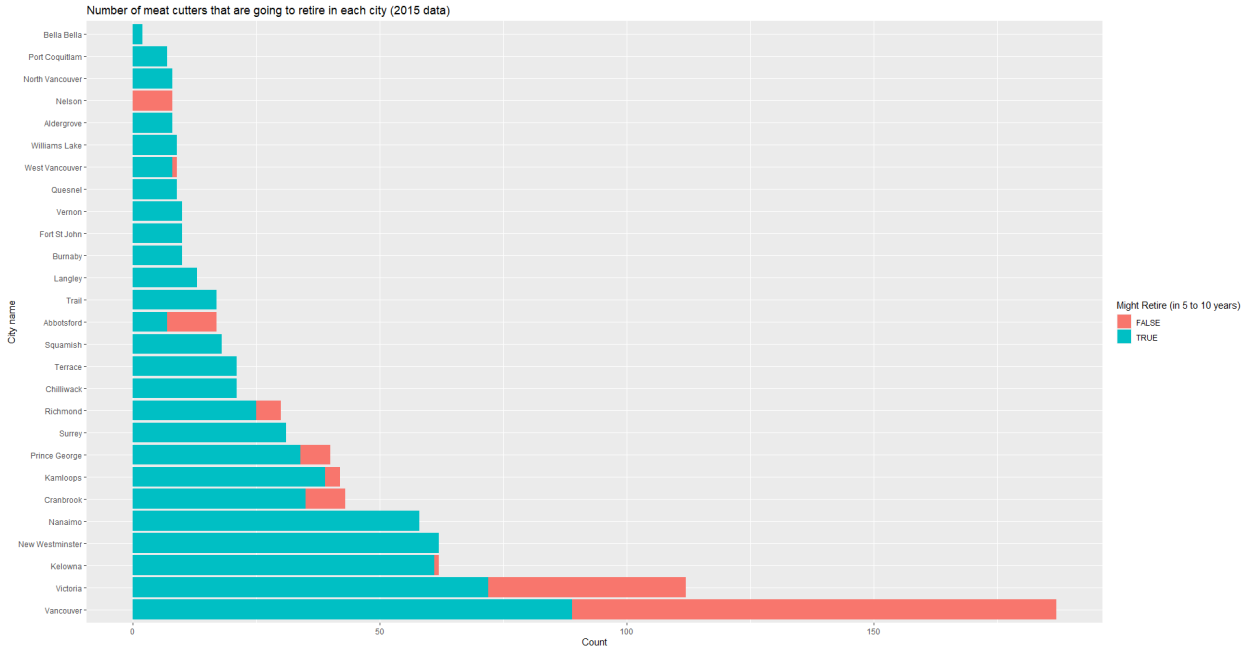


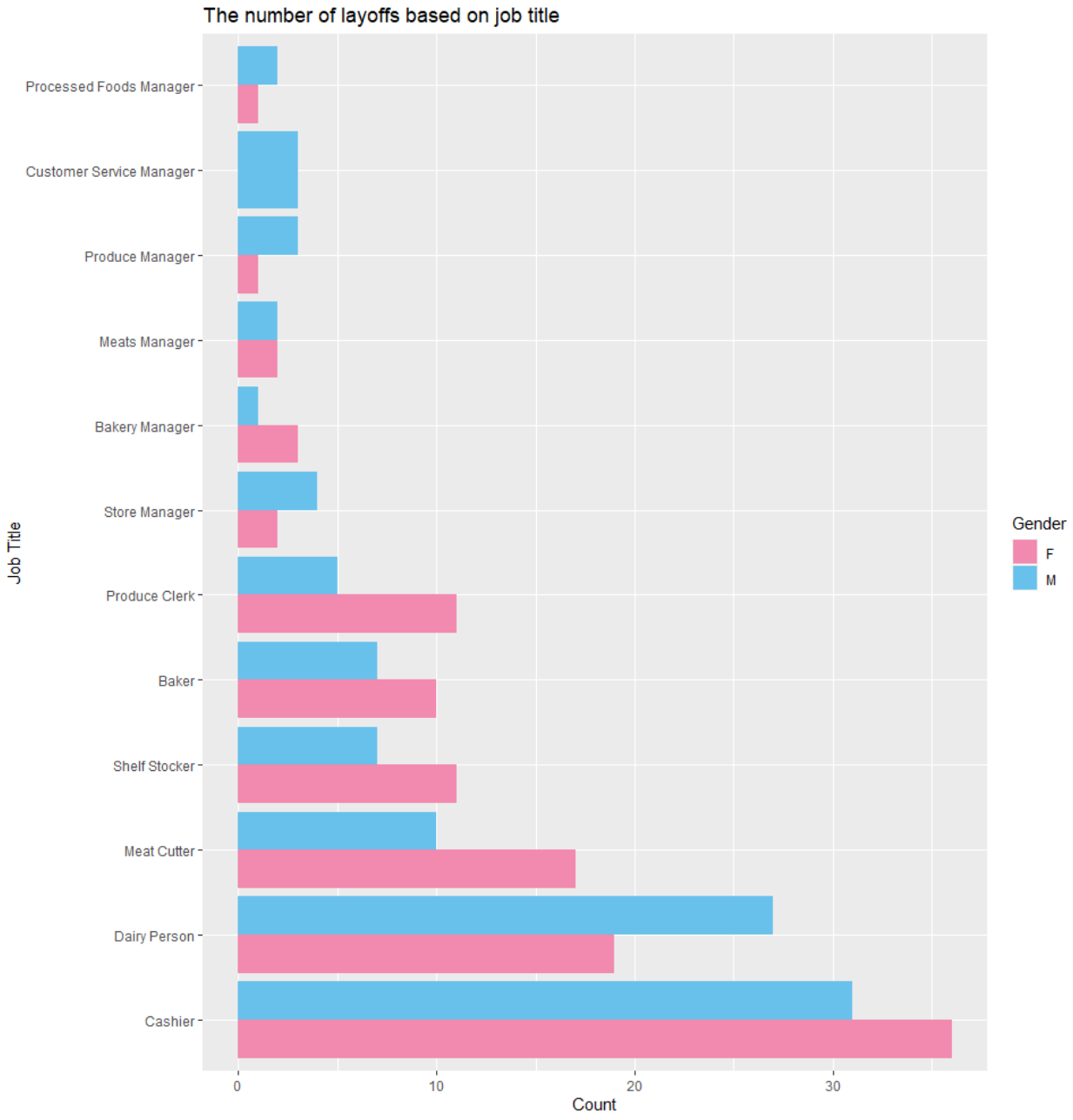

*Figure 49: With the help of fct_infreq*

*Figure 49* shows the graph that uses fct_infreq to reorder the factor according to the frequency. By default, the one with the lowest frequency will be at the top, and the one with the highest frequency will be placed at last. It is evident that the readability of the graph has increased; we can see the ranking more clearly, which could assist us in discovering further questions.

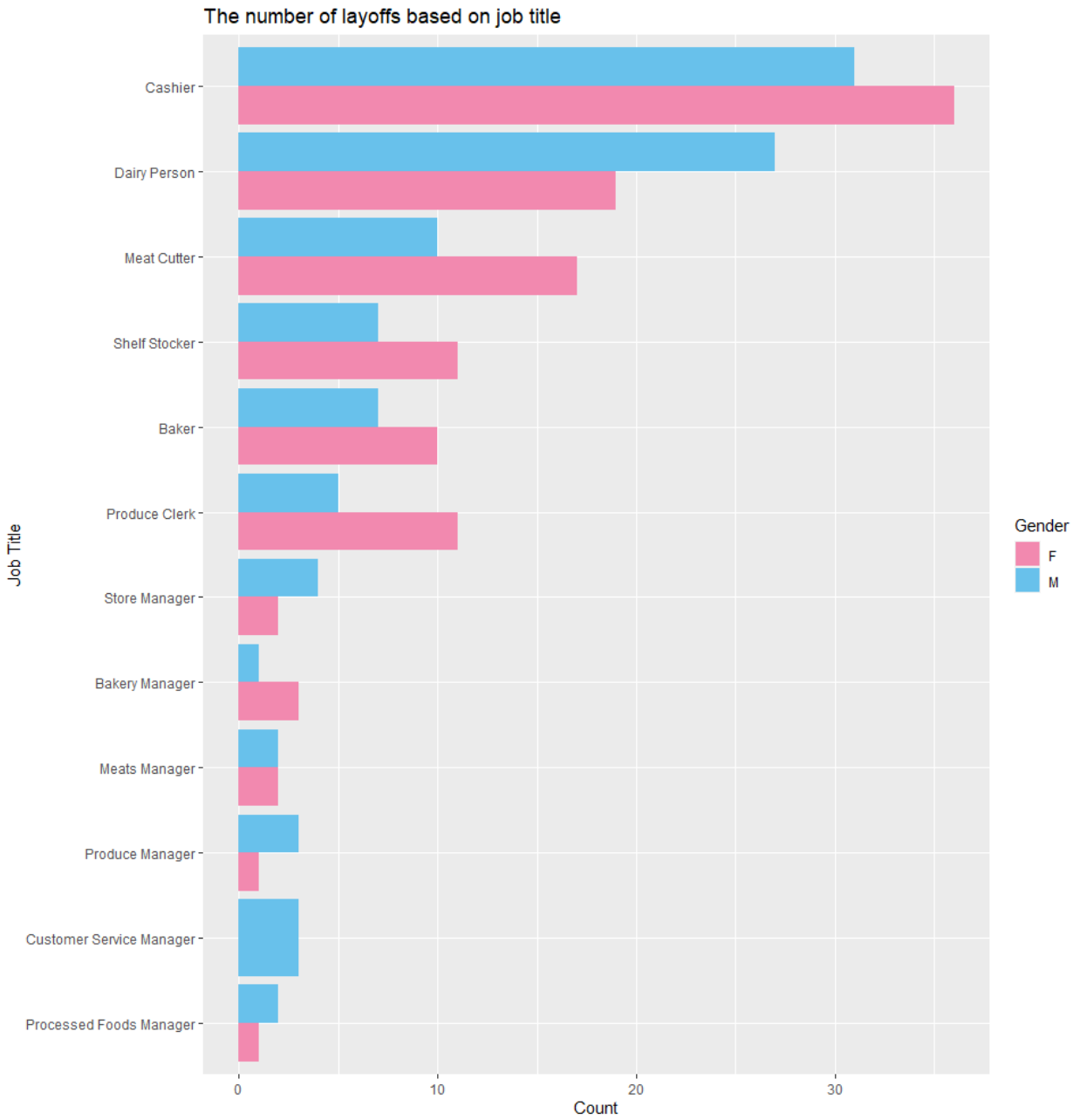
**Extra Feature 2:** `fct_rev()`

`fct_rev()` is another function from the `forcats` package which helps to reverse the order of the factor levels. It is useful when plotting a factor. Below shows the difference of using it in the plot and without using it.

*Without the function:*



*With the function:*



**Extra Feature 3:** `facet_wrap(scales = "free")`

Scales is one of the arguments that can be used in the facet_wrap function. By using scales = “free”, which will set the x axis and y axis to be free. What this means is that bars without any values will be omitted from the graph and the scales of the x axis and y axis will fit to values that fit with the facet. Below is the comparison:

*Without the argument:*

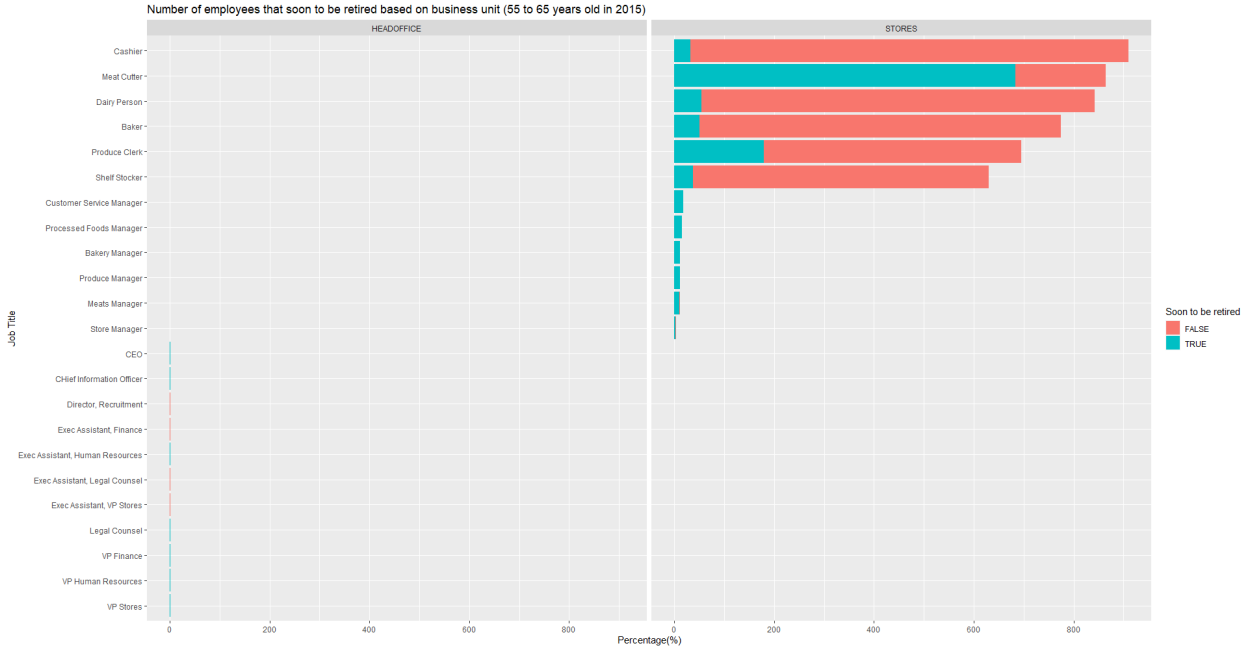



*Figure 50: Not using the scales argument in the facet_wrap*

As we can see from the graph, in both facet, there are some empty observations. The scales of the x axis are not suitable also because the values in HEADOFFICE are mostly 1. This makes the graph hard to read and could be consider as a bad graph.

*With the argument:*

As we can see, it is better compared to what we have seen in *Figure 50*. The scales of the x-axis fit properly to the values of the observations. This has improved the readability of the graph and provide more insights of the graph.

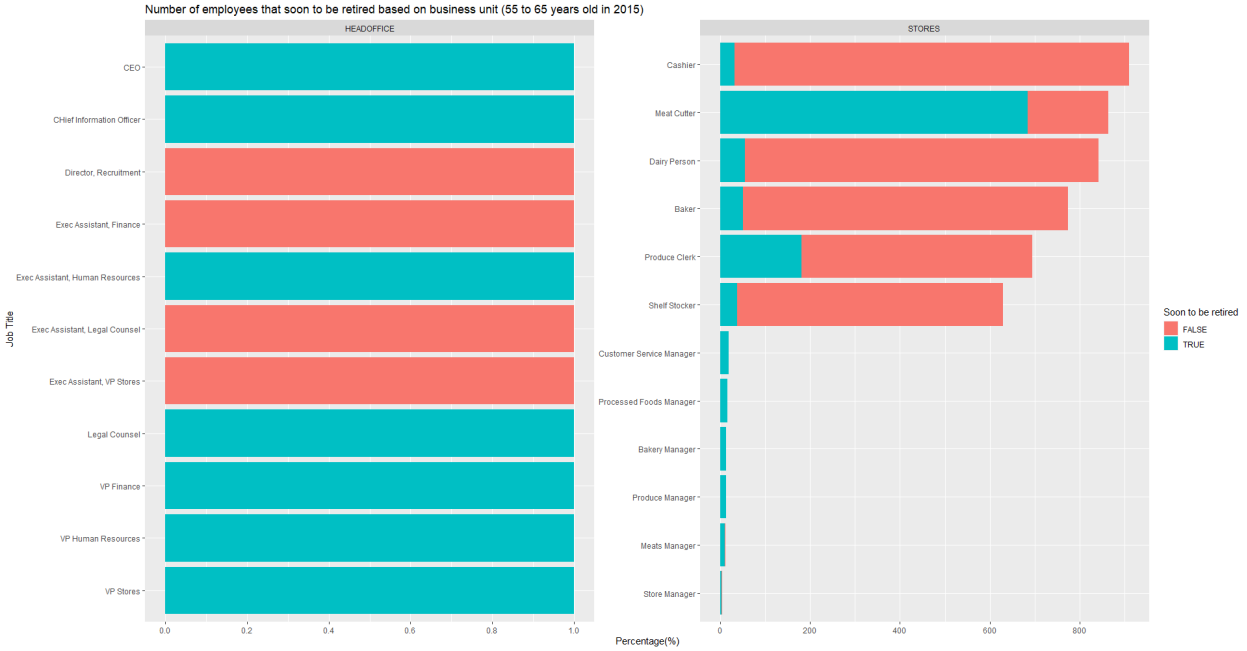



*Figure 51: With the help of scales argument in facet_wrap*

## Conclusion

The purpose of this assignment was to figure out what are the hidden issues in human resources management. Based on the conducted analysis, it can be concluded that there are multiple factors and reasons that cause retirement, resignation, and layoffs. Issues like job-hopping, automation seems to be the factors that affected the most on employee attrition. Throughout the analysis, some suggestions have also been provided for the company for making better decisions. This assignment is definitely a rewarding and fun experience for me as I get to apply what I have learned in this assignment. Besides that, I also gained some knowledge on human resource management, which has opened my eyes to the things that happened in the workplace.

## References


Alley, R. (2016, April). *43% Women In US Quit Jobs After Marriage! - IMDiversity*. Retrieved from IMDiversity: https://imdiversity.com/diversity-news/43-women-in-us-quit-jobs-after-marriage/

Berger, G. (2016, April 12). *Will This Year's College Grads Job-Hop More Than Previous Grads? | Official LinkedIn Blog*. Retrieved from linkedin.com: https://blog.linkedin.com/2016/04/12/will-this-year_s-college-grads-job-hop-more-than-previous-grads

Colleges, B. (2021, 10 20). *Become a Meat Cutter: Education and Career Roadmap*. Retrieved from bestaccreditedcolleges.org: https://bestaccreditedcolleges.org/articles/become-a-meat-cutter.html#:~:text=It%20typically%20takes%20two%20years,large%2C%20wholesale%20pieces%20of%20meat.

Dairy Central. (2020, March 3). *How Technology is Transforming Dairy Farming in Canada*. Retrieved from Dairy Central: https://dairycentral.ca/how-technology-is-transforming-dairy-farming-in-canada/

Dijkstra, L., Hamilton, E., Lall, S., & Wahba, S. (2020, March 10). *How do we define cities, towns, and rural areas?* Retrieved from World Bank Blogs: https://blogs.worldbank.org/sustainablecities/how-do-we-define-cities-towns-and-rural-areas

Essig, L. (2012, November 27). *Does Marriage Demand Women Quit Their Jobs?* Retrieved from Forbes: https://www.forbes.com/sites/laurieessig/2012/11/27/does-marriage-demand-women-quit-their-jobs/?sh=6dc844465e68

Forrest, C. (2018, July 17). *Cashier, clerk, or secretary? There's a 95+% chance you'll lose your job to a robot*. Retrieved from TechRepublic: https://www.techrepublic.com/article/cashier-clerk-or-secretary-theres-a-95-chance-youll-lose-your-job-to-a-robot/

Glassdoor. (n.d.). *Cashier Job Description*. Retrieved from Glassdoor: https://www.glassdoor.com/Job-Descriptions/Cashier.htm

Government of Canada. (2021, 10 1). *Canda Pension Plan - Eligibility - Canda.ca*. Retrieved from Canda.ca: https://www.canada.ca/en/services/benefits/publicpensions/cpp/cpp-benefit/eligibility.html

Gutierrez, D. (n.d.). *Little School on the Prairie: The Overlooked Plight of Rural Education*. Retrieved from The Institute of Plitics at Havard Kennedy School: https://iop.harvard.edu/get-involved/harvard-political-review/little-school-prairie-overlooked-plight-rural-education

Heathfield, S. M. (2021, February 24). *Top 10 Reasons Why Employees Quit*. Retrieved from the balance careers: https://www.thebalancecareers.com/top-reasons-why-employees-quit-their-job-1918985

Highfliers. (2020). *The Graduate Market in 2020.* Highfliers.

HR.org. (n.d.). *How Companies May Use Layoffs to Hide Age Discrimination*. Retrieved from HR.org: https://www.hg.org/legal-articles/how-companies-may-use-layoffs-to-hide-age-discrimination-50105

Indeed. (2021, September 1). *What Is Job-hopping? Pros and Cons*. Retrieved from Indeed.com: https://www.indeed.com/career-advice/starting-new-job/job-hopping

J.Kain, D. (2021, July 12). *What to Expect When a Long-Term CEO Retires*. Retrieved from Kirby Bates Associates: https://kirbybates.com/blog/what-to-expect-when-a-long-term-ceo-retires/

Jobstore. (n.d.). *Advantages and Disadvantages of Job Hopping*. Retrieved from Jobstrore.com: https://www.jobstore.com/careers-blog/2017/04/17/advantages-disadvantages-job-hopping/

Larasati, A., & Aryanto, D. B. (2020). Job-Hopping and the Determinant Factors. *Advances in Social Science, Education and Humanities Research*, 54-56.

Paul, A. M. (2012, September 19). *Why Young Employees Quit Ther Jobs*. Retrieved from businessinsider.com: https://www.businessinsider.com/why-young-employees-quit-their-jobs-2012-9#:~:text=Beth%20N.,very%20mobile%2C%27%20Carver%20said.

Resume Professional Writers. (n.d.). *Job Hopping: The Pros and Cons You Must Know*. Retrieved from Resume Professional Writers: https://www.resumeprofessionalwriters.com/job-hopping-pros-cons/

scales.r-lib.org. (n.d.). *Scale Functions for Visualization - scales*. Retrieved from scales.r-lib.org: https://scales.r-lib.org/

Statistics Canada. (2013). *Maritial Overview.* Statistics Canda.

Statistics Canada. (2018, June 5). *Fertility: Overview, 2012 to 2016*. Retrieved from Statistics Canada: https://www150.statcan.gc.ca/n1/pub/91-209-x/2018001/article/54956-eng.htm

Subramanian, P. (n.d.). *What is the average age of college graduates? - collegeOnomics*. Retrieved from collegeonomics.com.

Thorbecke, C. (2020, February 26). *Amazon opens a grocery store with no cashiers*. Retrieved from abcNew: https://abcnews.go.com/Business/amazon-opens-grocery-store-cashiers/story?id=69199966

Tucker, K. (n.d.). *Are There More Jobs in Urban Areas Than Rural Areas?* Retrieved from Chron: https://work.chron.com/ecotourism-careers-2100.html

Turner, A. (n.d.). *Retiring employees: 8 tips for a smooth transition - Insperity*. Retrieved from Insperity: https://www.insperity.com/blog/retiring-employees-8-tips-smooth-transition/

Wenger, Y. (2019, May 30). *The future of work: Being a cashier is Maryland's most common occupation, but it's likely vanishing*. Retrieved from The Baltimore Sun: https://www.baltimoresun.com/business/bs-md-cashier-20190305-story.html