INDIVIDUAL ASSIGNMENT 1

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APU2F2109CS(DA)_APD2F2109CS(DA)

HAND OUT DATE: 4 OCTOBER 2021

HAND IN DATE: 22 NOVEMBER 2021

WEIGHTAGE: 50%

STUDENT NAME: RYAN MARTIN

TP NUMBER: TP058091

Table of Contents

1.0 Introduction

1.1 Overview

In this assignment, students are tasked to apply data analytics techniques to a given dataset. The dataset contains the data of employees from a certain company. The aim of this analysis is to discover hidden issues within the company's human resources management. The dataset itself consists of 49,653 records and 18 variables. All the analysis here will be done using R, a free and open-source programming language for statistical analysis and computing, and RStudio, an integrated development environment (IDE) for R.

1.2 Assumptions

Below are the assumptions made for this analysis:

1. The company is a retail company, based on the data provided.
2. The company is based in British Columbia, Canada. This is based on the fact that the city names in the dataset are all cities from this region.
3. The dataset provided is a subset of a much larger dataset, with more variables such as salary or job satisfaction, and more records.
4. Based on the dataset provided, there are no records of employees with more than 1 job, so it is assumed that there are no department transfers and no promotions (career advancement). This could be because those records are not included, as stated in point 3 of assumptions.

2.0 Pre-Analysis Setup and Processing

2.1 Library Imports

```
library(dplyr)
library(ggplot2)
library(lubridate)
```

Figure 1: R code to import libraries

The external libraries that will be used in this analysis are *dplyr*, a library for data manipulation, *ggplot2*, a data visualization library, and *lubridate*, a utility library to handle datetime formats (RStudio, n.d.). Also, some library functions are used but not included using the library function, as they are only used once.

2.2 Data Import

```
df <- readr::read_csv('employee_attrition.csv')
head(df, 20)
```

Figure 2: R code to import and view data

EmployeeID <dbl>	recorddate_key <chr>	birthdate_key <chr>	orighiredate_key <chr>	terminationdate_key <chr>	a... <dbl>
1318	12/31/2006 0:00	1/3/1954	8/28/1989	1/1/1900	52
1318	12/31/2007 0:00	1/3/1954	8/28/1989	1/1/1900	53
1318	12/31/2008 0:00	1/3/1954	8/28/1989	1/1/1900	54
1318	12/31/2009 0:00	1/3/1954	8/28/1989	1/1/1900	55
1318	12/31/2010 0:00	1/3/1954	8/28/1989	1/1/1900	56
1318	12/31/2011 0:00	1/3/1954	8/28/1989	1/1/1900	57
1318	12/31/2012 0:00	1/3/1954	8/28/1989	1/1/1900	58
1318	12/31/2013 0:00	1/3/1954	8/28/1989	1/1/1900	59
1318	12/31/2014 0:00	1/3/1954	8/28/1989	1/1/1900	60
1318	12/31/2015 0:00	1/3/1954	8/28/1989	1/1/1900	61

1-10 of 20 rows | 1-6 of 18 columns Previous 1 2 Next

Figure 3: Dataset preview

The data is imported using *readr*'s *read_csv* function. It loads the data into a data frame in R. The *head* function is used to show only a portion of the top rows of the data frame.

2.3 Preprocessing

```
df <- df %>%
  mutate(
    recorddate_key = as.Date(recorddate_key, '%m/%d/%Y'),
    birthdate_key = as.Date(birthdate_key, '%m/%d/%Y'),
    orighiredate_key = as.Date(orighiredate_key, '%m/%d/%Y'),
    terminationdate_key = as.Date(terminationdate_key, '%m/%d/%Y'),
    termreason_desc = ifelse(termreason_desc == 'Resignaton',
                             'Resignation',
                             termreason_desc),
    city_name = ifelse(city_name == 'New Westminister',
                       'New Westminster',
                       city_name),
    store_name = as.factor(store_name)) %>%
    select(-gender_short)

# df with 1 record for each employee
df2 <- df %>%
  group_by(EmployeeID) %>%
  filter(STATUS_YEAR == max(STATUS_YEAR)) %>%
  filter(!(n() > 1 & termreason_desc == 'Not Applicable')) %>%
  ungroup()
```

Figure 4: R code for data preprocessing

Here is the preprocessing done to the imported dataset. The *mutate* function from the *dplyr* library is used to change the columns in the data frame. The first 4 lines inside the mutate argument list are used to change the date columns, which were originally read as character, into a date type. Then, typos such as 'Resignaton' and 'New Westminister' are fixed, and lastly the store_name variable is changed from numeric to a factor type.

After the original data frame is cleaned, a new one is created with only 1 record for each employee in the dataset. Only the latest values will be included, as terminations are always the last record. The new data frame is created by filtering the dataset for records with the the latest *STATUS_YEAR* for each employee. Another filter is used because for terminated employees, there will be 2 records at the latest year of record, so the record with 'Not Applicable' as the termination reason will be discarded. The resulting data frame has 6,284 records.

EmployeeID <dbl>	recorddate_key <date>	birthdate_key <date>	orighiredate_key <date>	terminationdate_key <date>	a... <dbl>
1318	2015-12-31	1954-01-03	1989-08-28	1900-01-01	61
1319	2015-12-31	1957-01-03	1989-08-28	1900-01-01	58
1320	2015-12-31	1955-01-02	1989-08-28	1900-01-01	60
1321	2015-12-31	1959-01-02	1989-08-28	1900-01-01	56
1322	2015-12-31	1958-01-09	1989-08-31	1900-01-01	57
1323	2015-12-31	1962-01-09	1989-08-31	1900-01-01	53
1325	2015-12-31	1964-01-13	1989-09-02	1900-01-01	51
1328	2015-12-31	1956-01-17	1989-09-05	1900-01-01	59
1329	2015-12-31	1967-01-23	1989-09-08	1900-01-01	48
1330	2015-12-31	1967-01-25	1989-09-09	1900-01-01	48

1-10 of 6,284 rows | 1-6 of 17 columns Previous 1 2 3 4 5 6 ... 629 Next

Figure 5: View of transformed data frame

3.0 Question 1: What Factors Lead to Employee Termination?

3.1 Overview

For this question, the analyses will mostly be focusing on resignations. Retirements are caused by old age, and this is an undeniable fact. Companies cannot stop employees from retiring. Layoffs on the other hand are quite difficult to analyze because they are mostly due to the company's decision (Indeed, 2021). Resignations are done by employees of their own accord, so the factors for resignation reside in the employees themselves.

3.2 Analyses

3.2.1 Termination Reasons Overview

3.2.1.1 Termination Reasons Univariate Analysis

```
df2 %>%
  count(termreason_desc) %>%
  ggplot(aes(n, reorder(termreason_desc, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Employee by Termination Reason',
    x = 'Number of Employees',
    y = 'Termination Reason'
  )
```

Figure 6: R code to plot bar chart of termination reason

Here, a bar chart will be used to plot the count of employees for both genders. The *count* function is used to count the number of records for the specified group, which in this case is the *ter*mreason*_desc* variable. The plot itself will be drawn using the *geom_bar* function. The *stat* option is set to identity, which forces the bar plot to use the x and y axis values set using the *aes* (aesthetics) set in the *ggplot* function. The bars are also reordered from the highest value to the lowest using the *reorder* function.

The '%>%' symbol is the pipe operator, provided by *dplyr* from the *magrittr* library. It is used to bring the value of the expression on its left-hand side as an argument to the instruction (which is usually a function) on its right-hand side. The + operator is provided by *ggplot2* to ease composition of graphs.

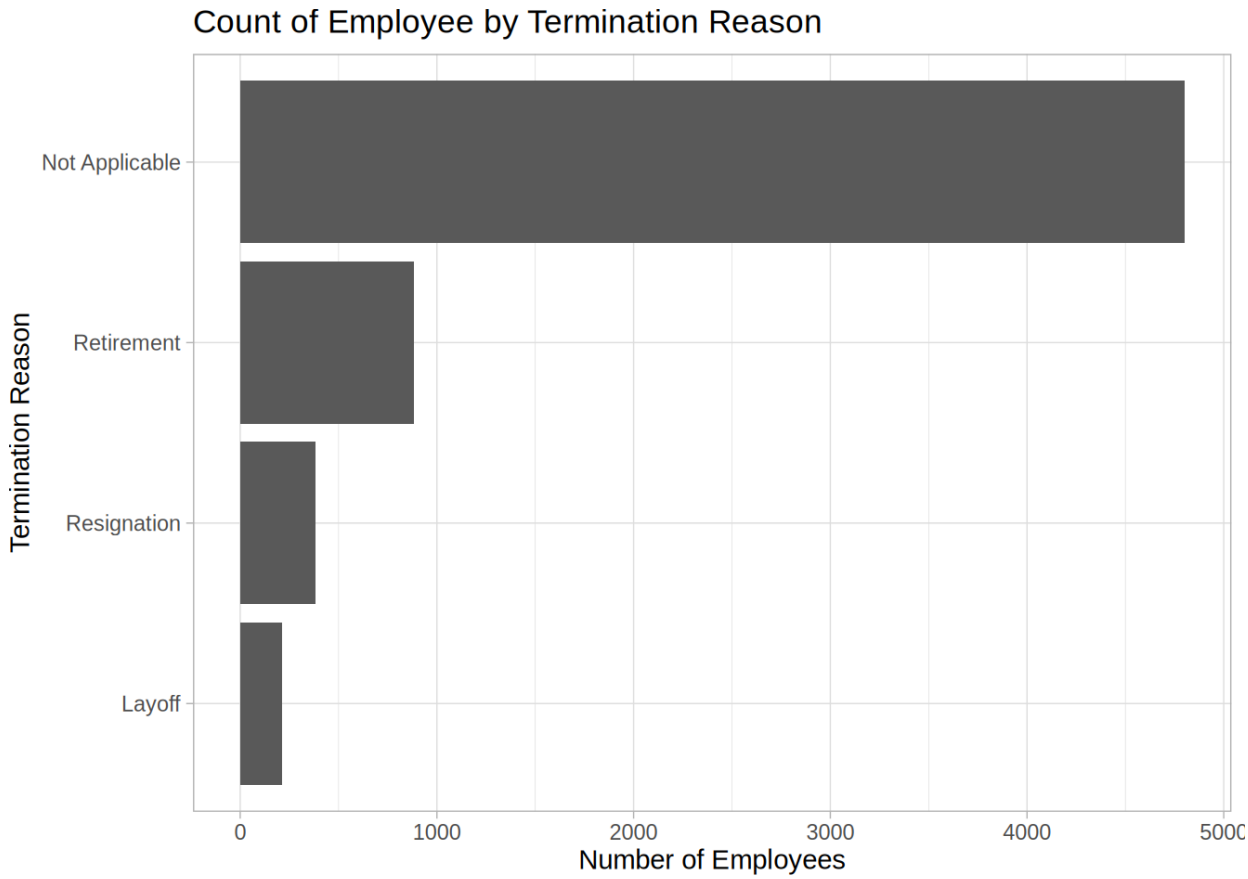


Figure 7: Bar chart of termination reason

Looks like most of the employees are still active, meaning not terminated yet. The termination reasons are the point of interest for this analysis, so having little data on them is not preferred. Layoffs have the least number of records so analyses of this should be challenging.

3.2.1.2 Termination Reasons Over the Years

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  count(STATUS_YEAR, termreason_desc) %>%
  ggplot(aes(STATUS_YEAR, n, color = termreason_desc)) +
  geom_line() +
  scale_x_continuous(breaks = seq(2006, 2015)) +
  scale_y_continuous(breaks = seq(0, 150, 25)) +
  labs(
    title = 'Count of Terminations Over The Years Grouped by Reason',
    x = 'Year',
    y = 'Termination Count',
    color = 'Termination Reason'
  )
```

Figure 8: R code to plot line plot of termination reasons

Here, the functions *scale_x_continuous* and *scale_y_continous* are used to customize the x and y axis ticks for a line graph drawn using the *geom_line* function. The *lab* function is used to add labels such as the title and the axis names to the plot. The other functions such as *filter* and *count* will be explained in later sections.

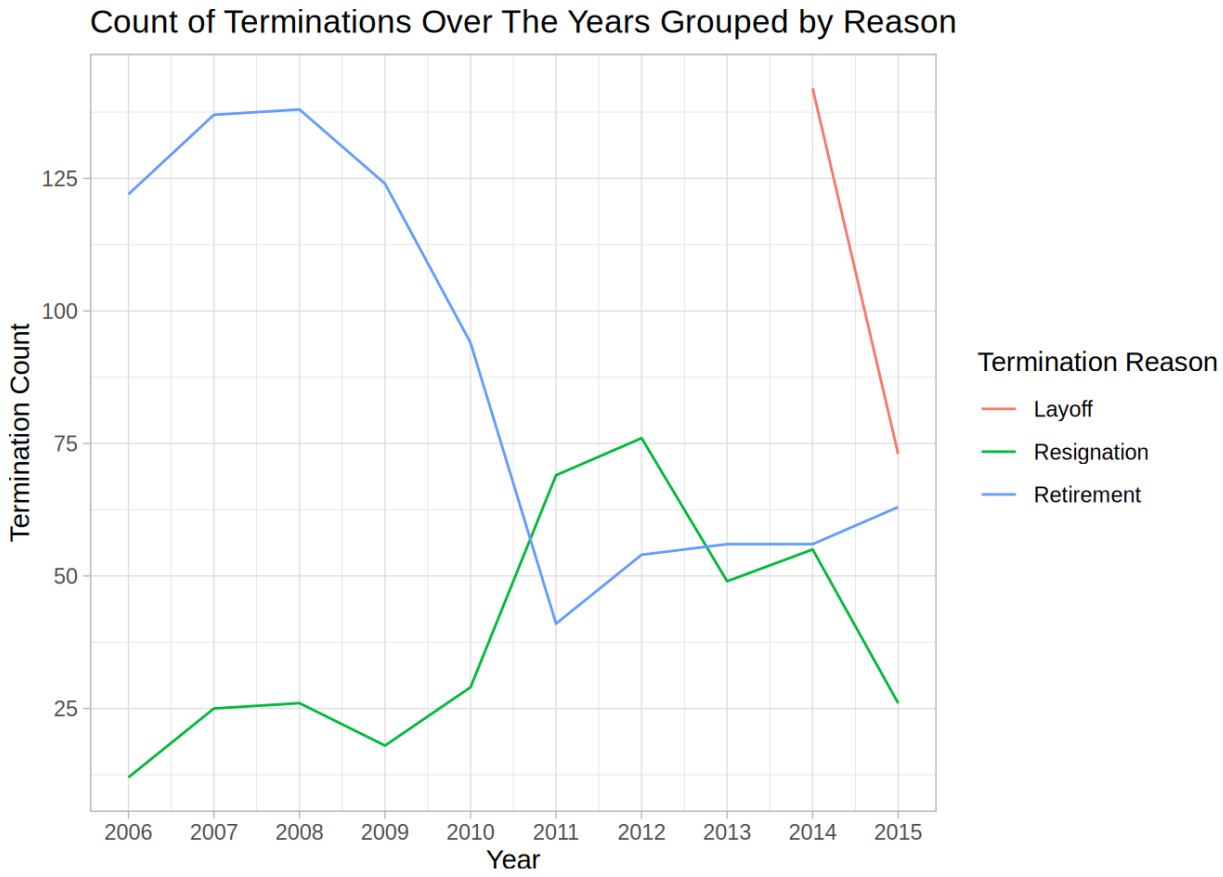


Figure 9: Line plot of termination reasons

Looking at the chart, layoffs only start in the year 2014. This could be the reason why there are so few records on them, but there's no way to explain why this happened with only the given dataset. Resignations and retirements have enough data, but there are no obvious patterns that can be seen from the chart above.

3.2.2 Age regarding Employee Termination

3.2.2.1 Age Univariate Analysis

```
df2 %>%
  ggplot(aes(age)) +
  geom_histogram(binwidth = 2) +
  scale_x_continuous(breaks = seq(20, 70, 5)) +
  labs(
    title = 'Distribution of Employee Age',
    x = 'Age',
    y = 'Count'
  )
```

Figure 10: R code to plot histogram of employee age

The first step taken in the age analysis is looking at the value distribution. Here, *ggplot's geom_histogram* function is used to generate the graph. A more specific x-axis marker is added to the plot using the *scale_x_continuous* function. It takes a sequence of numbers from 20 to 70, incremented by 5. This is achieved with the built-in *seq* function from R.

Figure 11: Histogram of employee age

Looking at the chart, age distribution seems uniform, with an unexpectedly high number of employees above the age of 60. This distribution could be different if the data is grouped using another variable, so that will be the next step in this analysis.

3.2.2.1 Age Grouped by Termination Reason

3.2.2.1.1 Age of Resignations

```
df2 %>%
  filter(termreason_desc == 'Resignation') %>%
  ggplot(aes(age)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = seq(10, 70, 2)) +
  labs(
    title = 'Distribution of Age of Employee at Resignation',
    x = 'Age',
    y = 'Number of Employees'
  )
```

Figure 12: R code to plot histogram of employees at resignation

Here, the data frame is filtered before plotting to include only the records with resignation as the termination reason. This is done using the *filter* function provided by *dplyr*.

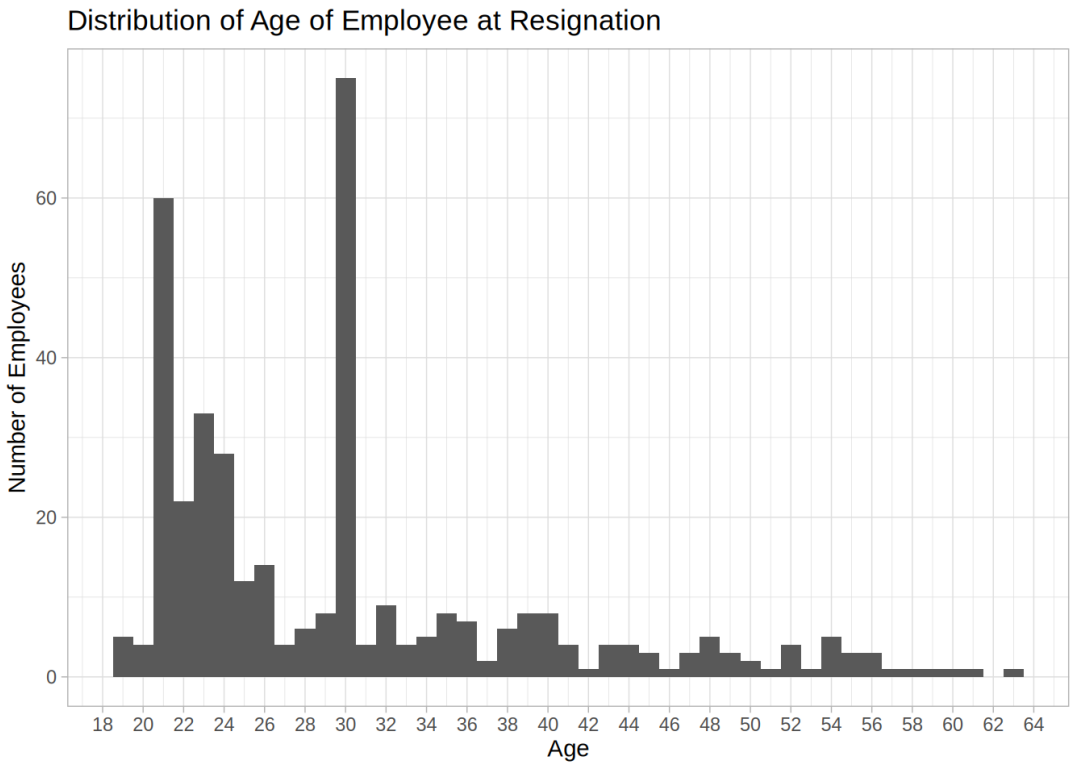


Figure 13: Histogram of employees at resignation

Unlike the previous chart, it can be seen from this one that many employees are around the ages of 20 until 30 during resignation. Age could be a factor to employee termination, but to understand why further analysis is needed. It will be covered in the later sections of this document.

3.2.2.1.2 Age of Retirements

```
df2 %>%
  filter(termreason_desc == 'Retirement') %>%
  ggplot(aes(age)) +
  geom_histogram() +
  labs(
    title = 'Distribution of Age of Employee at Retirement',
    x = 'Age',
    y = 'Number of Employees'
  )
```

Figure 14: R code to plot histogram of employees at retirement

Figure 15: Histogram of employees at retirement

Here is the graph using only the retired employees' data. This is expected as the average retirement age of workers in Canada is between 60 and 70 (Government of Canada, Statistics Canada, 2021).

3.2.2.1.3 Age of Layoffs

```
df2 %>%
  filter(termreason_desc == 'Layoff') %>%
  ggplot(aes(age)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = seq(10, 70, 2)) +
  labs(
    title = 'Distribution of Age of Employee at Layoff',
    x = 'Age',
    y = 'Number of Employees'
  )
```

Figure 16: R code to plot histogram of employees at layoff

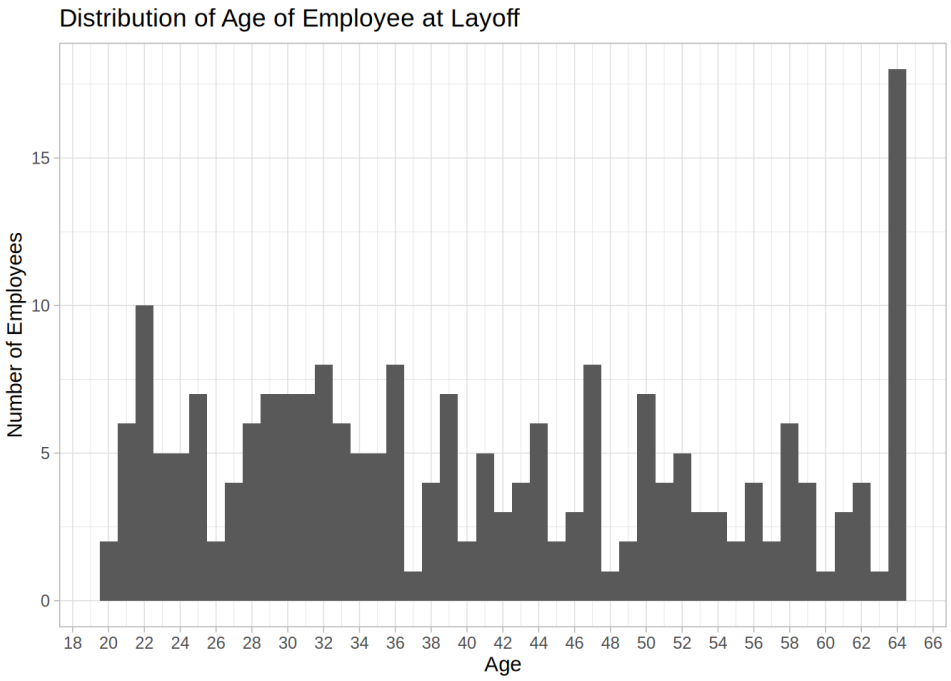


Figure 17: Histogram of employees at layoff

No obvious patterns here. However, there's an unusually high number of employees in their 60s being laid off. Even though Canada has a human right's law that protects older employees from forced layoff, it is still happening here. The cause of this cannot be determined from the given dataset alone (Legalline, n.d.).

3.2.3 Length of Service (Tenure) Regarding Employee Termination

3.2.3.1 Length of Service Univariate Analysis

```
df2 %>%
  ggplot(aes(length_of_service)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = seq(0, 30, 1)) +
  labs(
    title = 'Distribution of Employee Tenure',
    x = 'Tenure',
    y = 'Count'
  )
```

Figure 18: R code to plot histogram of employee tenure

From this point onwards, the term length of service will be shortened to tenure. Just like the previous analysis, the distribution of the values of tenure will be examined. The code is also almost identical as before, with the only difference being the variable being plotted, which is *length_of_sevice* in this case, and the labels.

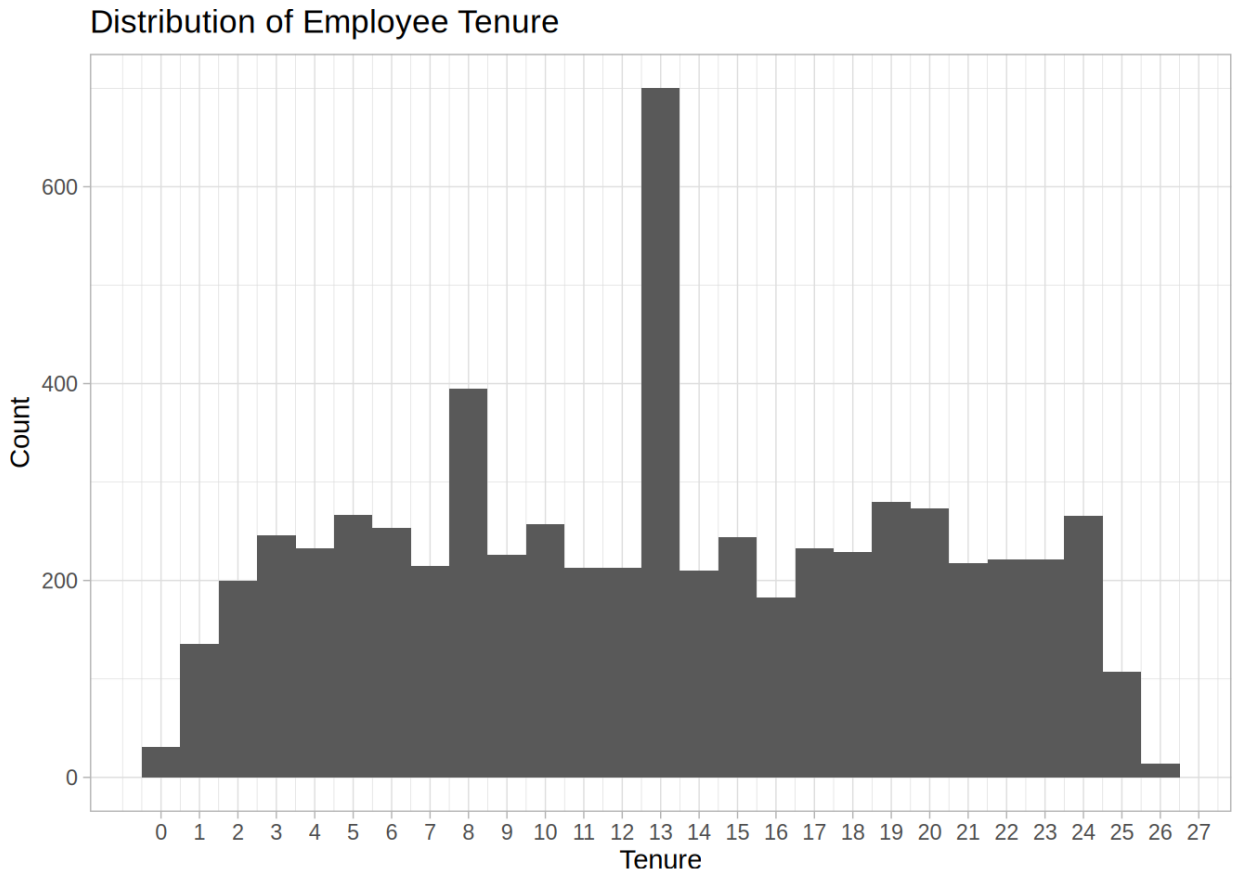


Figure 19: Histogram of employee tenure

The distribution here looks quite uniform as well, with an unusually high number of employees with 13 years of tenure. To shorten the analysis, the upcoming analyses will only focus mostly on employee resignation.

3.2.3.2 Tenure of Employees at Resignation

```
df2 %>%
  filter(termreason_desc == 'Resignation') %>%
  ggplot(aes(length_of_service)) +
  geom_histogram(binwidth = 1) +
  labs(
    title = 'Distribution of Tenure of Employee at Resignation',
    x = 'Tenure',
    y = 'Number of Employees'
  )
```

Figure 20: R code to plot histogram of employee tenure at resignation

Again, the code here is the same as the one from the age analysis, with only the plotted variable and labels being changed.

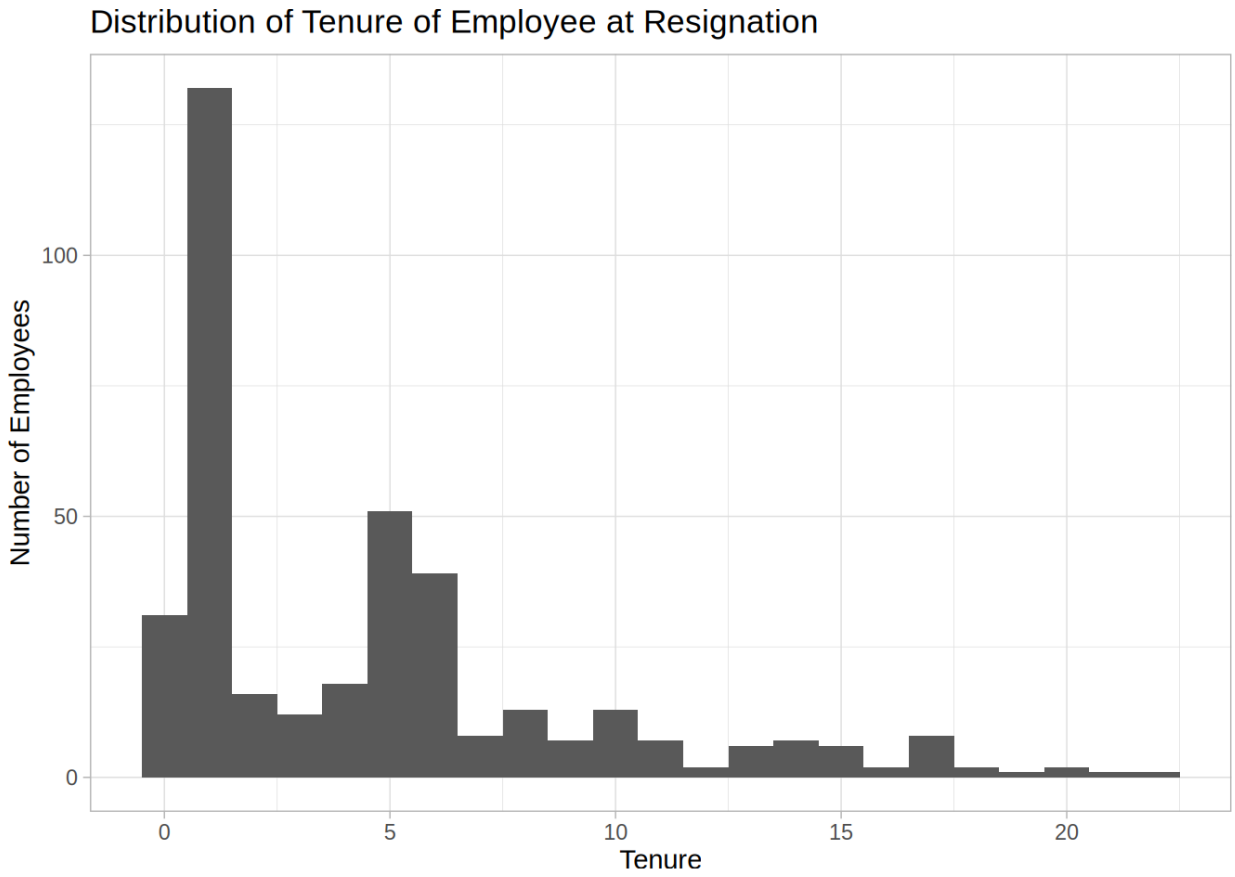


Figure 21: histogram of employee tenure at resignation

The distribution here looks like the distribution of age for resigned employees. These 2 variables might be correlated, so additional analysis of this will be done on this.

3.2.3.3 Age & Tenure Correlation

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  ggplot(aes(age, length_of_service, color = termreason_desc)) +
  geom_point() +
  labs(
    title = 'Correlation of Age with Tenure of Terminated Employees',
    x = 'Age',
    y = 'Tenure',
    color = 'Termination Reason'
  )
```

Figure 22: R code to plot scatter plot of employee age & tenure

To look at the correlation between the 2 variables, a scatterplot will be used. Here the data is filtered to only include the terminated employee, and the plot is made with the following options:

- Age as the x-axis value
- Tenure as the y-axis value
- Termination reason as the color for the dots. This is done to see the points from each termination reason.

The options are set using the *aes* parameter for the *ggplot* function. Then, *geom_point* is used to draw the scatter plot.

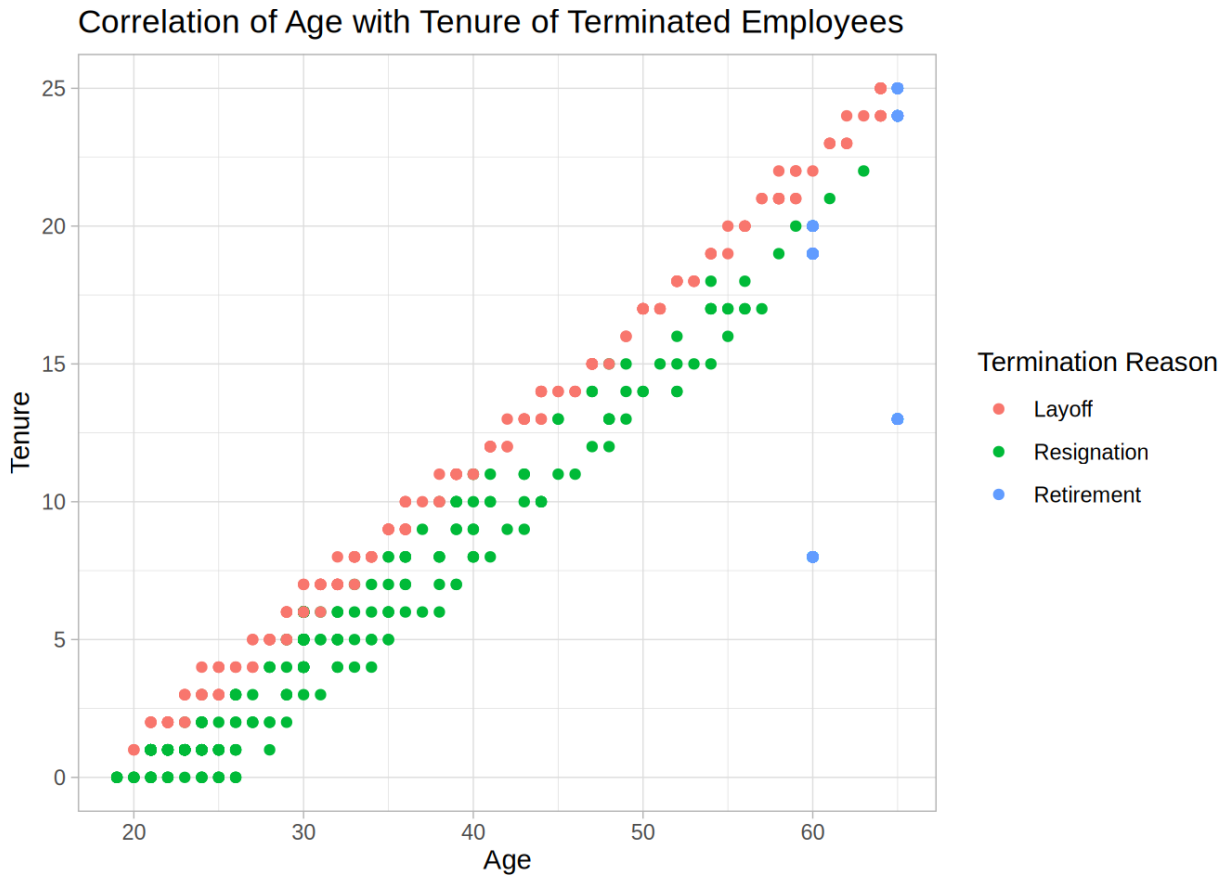


Figure 23: Scatter plot of employee age & tenure

It is apparent from the plot that these 2 variables have a strong, positive correlation. The Pearson's correlation coefficient is 0.8493077. This value can be computed using the *cor* function provided by R. The code is as shown below.

```
cor(df2$age, df2$length_of_service)

## [1] 0.8493077
```

Figure 24: R code to calculate the correlation coefficient

3.2.4 Gender Regarding Employee Termination

3.2.4.1 Gender Univariate Analysis

```
df2 %>%
  count(gender_full) %>%
  ggplot(aes(n, reorder(gender_full, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Employee by Gender',
    x = 'Number of Employees',
    y = 'Gender'
  )
```

Figure 25: R code to plot a bar chart of employee gender

The code used here is the same as in section 3.2.1.1, with the only differences being the variable of interest and the labels.

Figure 26: Bar chart of employee gender

It can be seen from this graph that there are more female employees than male. This, however, doesn't explain anything about the termination reasons.

3.2.4.2 Gender Grouped by Termination Reason

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  count(gender_full, termreason_desc) %>%
  ggplot(aes(gender_full, n)) +
  geom_bar(stat = 'identity') +
  facet_wrap(~termreason_desc) +
  labs(
    title = 'Comparison of Employee Gender Grouped by Termination Reason',
    x = 'Gender',
    y = 'Number of Employees'
  )
```

Figure 27: R code to plot bar chart of employee gender by termination

To view the count of genders in each termination reason, the data frame will need to be manipulated first. In the code above, after filtering for only terminated employees, the *count* function is used the count the number of records grouped by gender and termination reason. It is then passed into the *ggplot* function. Here, *facet_wrap* is used to create multiple plots using a grouping variable, which in this case is the termination reason.

Figure 28: Bar chart of employee gender by termination

Even after being grouped by termination reason, there are still more female employees on each group than male. The amount of retired females here could mean that they are more 'loyal' than male employees here, but further analysis is needed to verify this claim. It will be done in a later section of this document.

3.2.5 City Regarding Employee Termination

3.2.5.1 City Univariate Analysis

```
df2 %>%
  count(city_name) %>%
  ggplot(aes(n, reorder(city_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Employee by City',
    x = 'Number of Employees',
    y = 'City Name'
  )
```

Figure 29: R code to plot bar chart of city names

Again, the code used here is the same as the one in the previous section, with the only difference being the variable of interest and the labels.

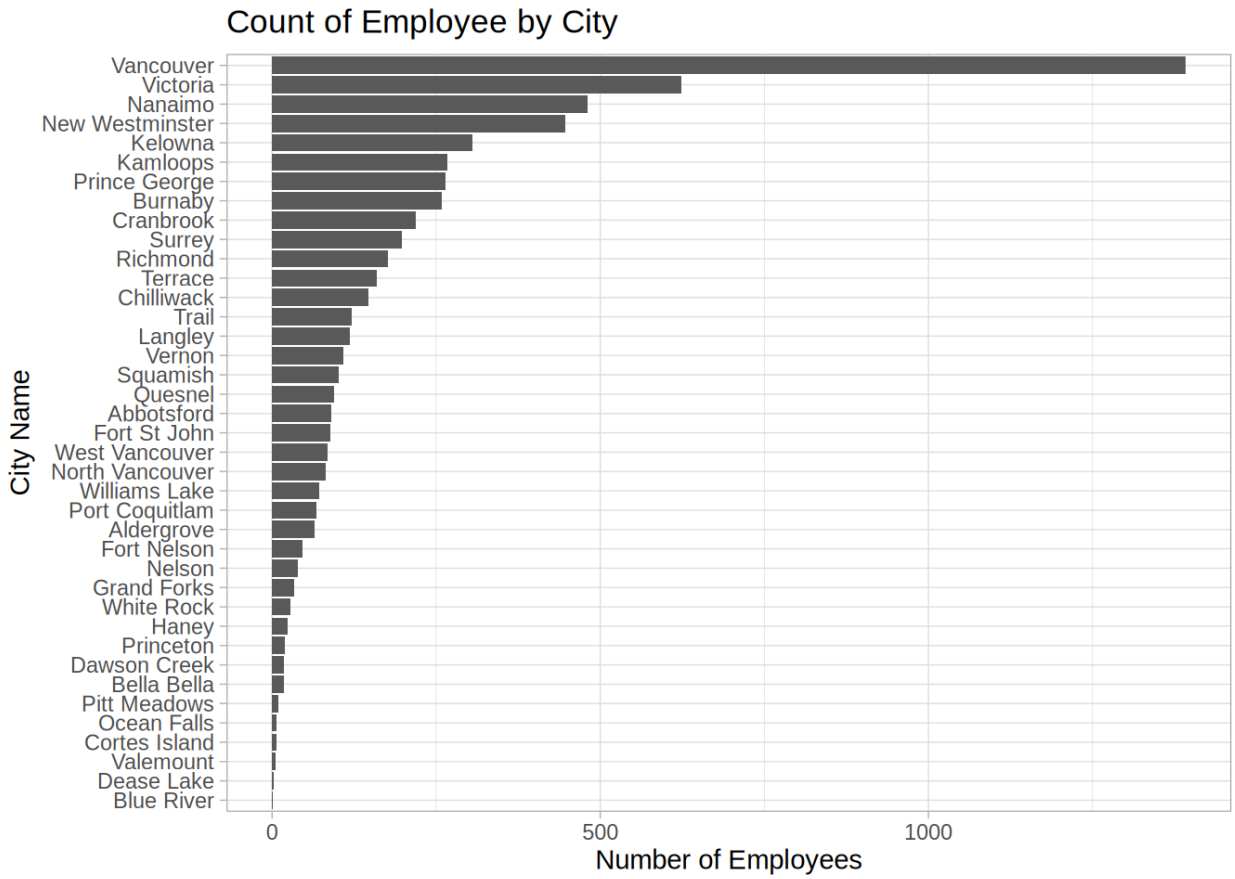


Figure 30: Bar chart of city names

Nothing interesting here. The cities with the most employees are Vancouver, which is the largest city in British Columbia, and Victoria, the capital city.

3.2.5.2 City Grouped by Termination Reason

3.2.5.2.1 Cities of Resignations & Retirements

```
df2 %>%
  filter(termreason_desc == 'Resignation') %>%
  count(city_name) %>%
  ggplot(aes(n, reorder(city_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Resignations Grouped by City',
    x = 'Number of Employees',
    y = 'City Name'
  )
```

Figure 31: R code to plot bar chart of city names of with resignations

```
df2 %>%
  filter(termreason_desc == 'Retirement') %>%
  count(city_name) %>%
  ggplot(aes(n, reorder(city_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Retirements Grouped by City',
    x = 'Number of Employees',
    y = 'City Name'
  )
```

Figure 32: R code to plot bar chart of city names of with retirements

Again, the code from the previous section is reused here to draw the plots.

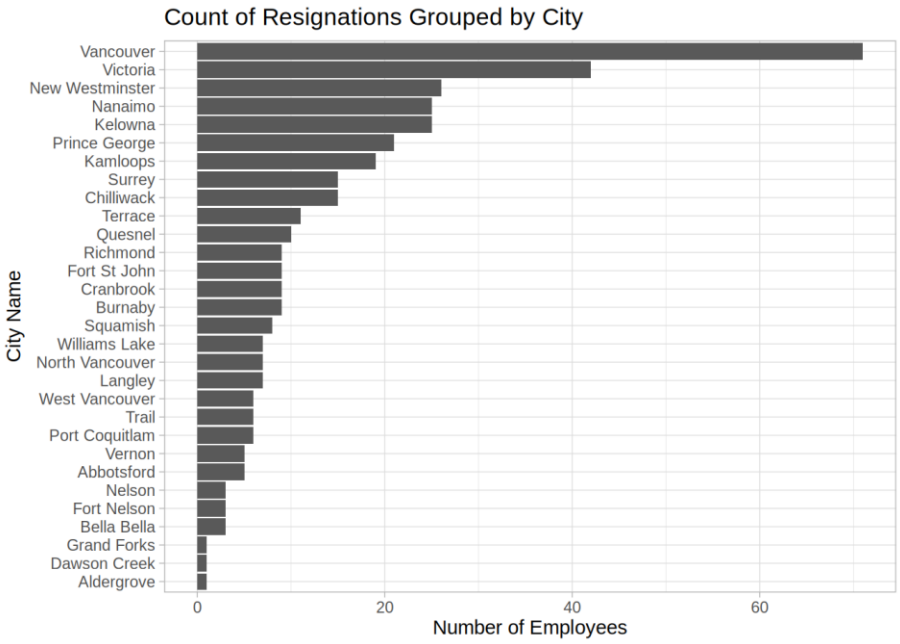


Figure 33: Bar chart of city names of with resignations

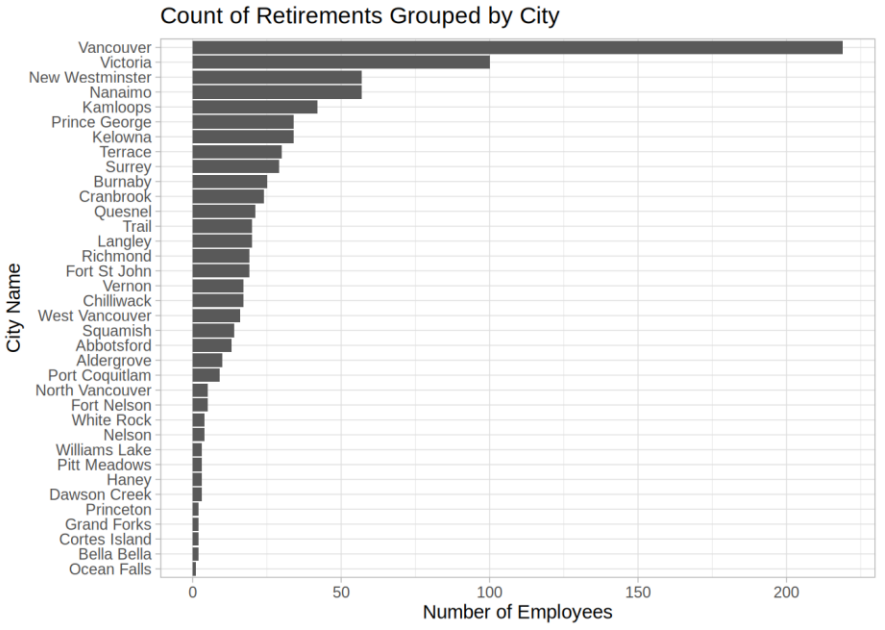


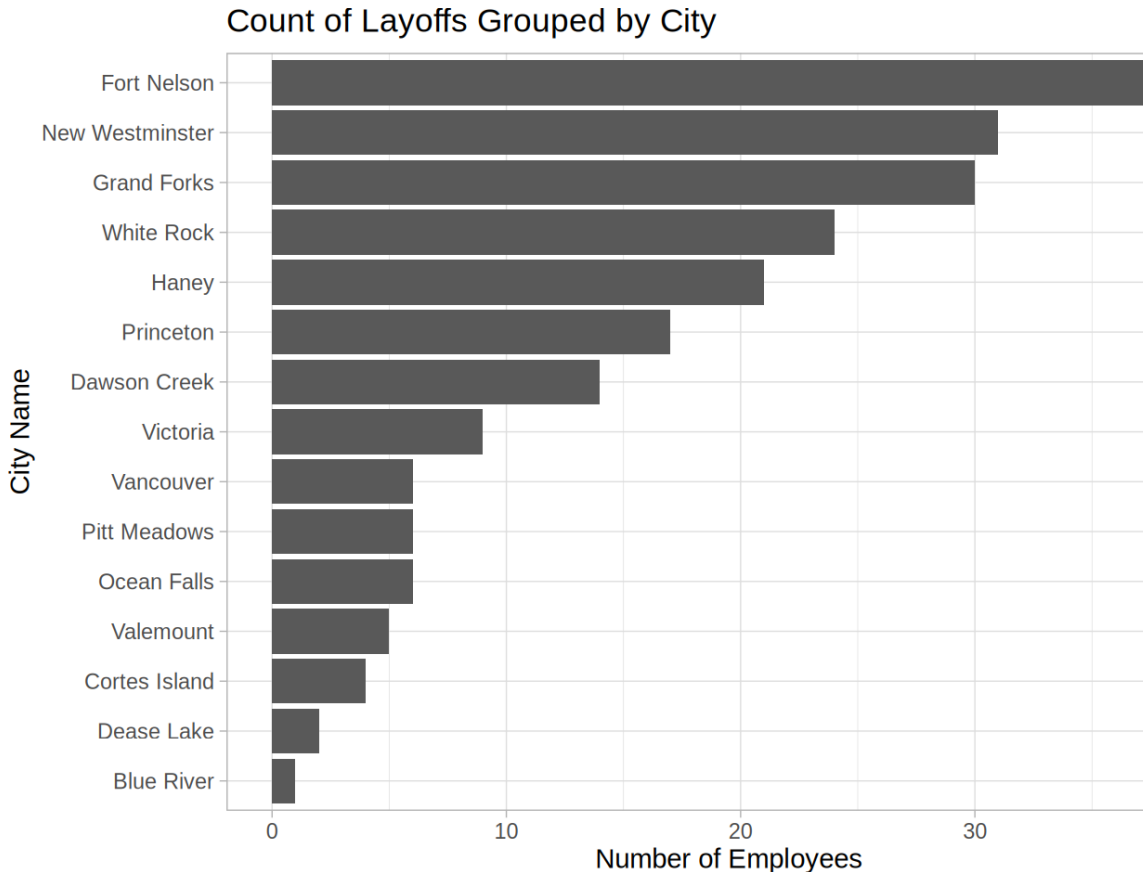
Figure 34: Bar chart of city names of with retirements

Looking at both plots for employee resignation and retirements, nothing much seems to be different than the plot in the previous section. This could mean that the city isn't one of the factors of employee termination. However, the plot for employee layoff is entirely different from these 2.

3.2.5.2.2 Cities of Layoffs

```
df2 %>%
  filter(termreason_desc == 'Layoff') %>%
  count(city_name) %>%
  ggplot(aes(n, reorder(city_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Layoffs Grouped by City',
    x = 'Number of Employees',
    y = 'City Name'
  )
```

Figure 35: R code to plot bar chart of city names of with layoffs



As seen here, the cities with the most layoffs are nowhere near the top cities, except for New Westminster. This might mean that some aspects of these cities can affect employee layoff. Further analysis is needed to try to understand why this happened.

3.2.6 Department Regarding Employee Termination

3.2.6.1 Department Univariate Analysis

```
df2 %>%
  count(department_name) %>%
  ggplot(aes(n, reorder(department_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Employee by Department',
    x = 'Number of Employees',
    y = 'Department Name'
  )
```

Figure 37: R code to plot a bar char of employee count by department

After this point, the code that is reused with different variables and labels will stop having explanations. This is to reduce the redundant lines that have been mentioned before.

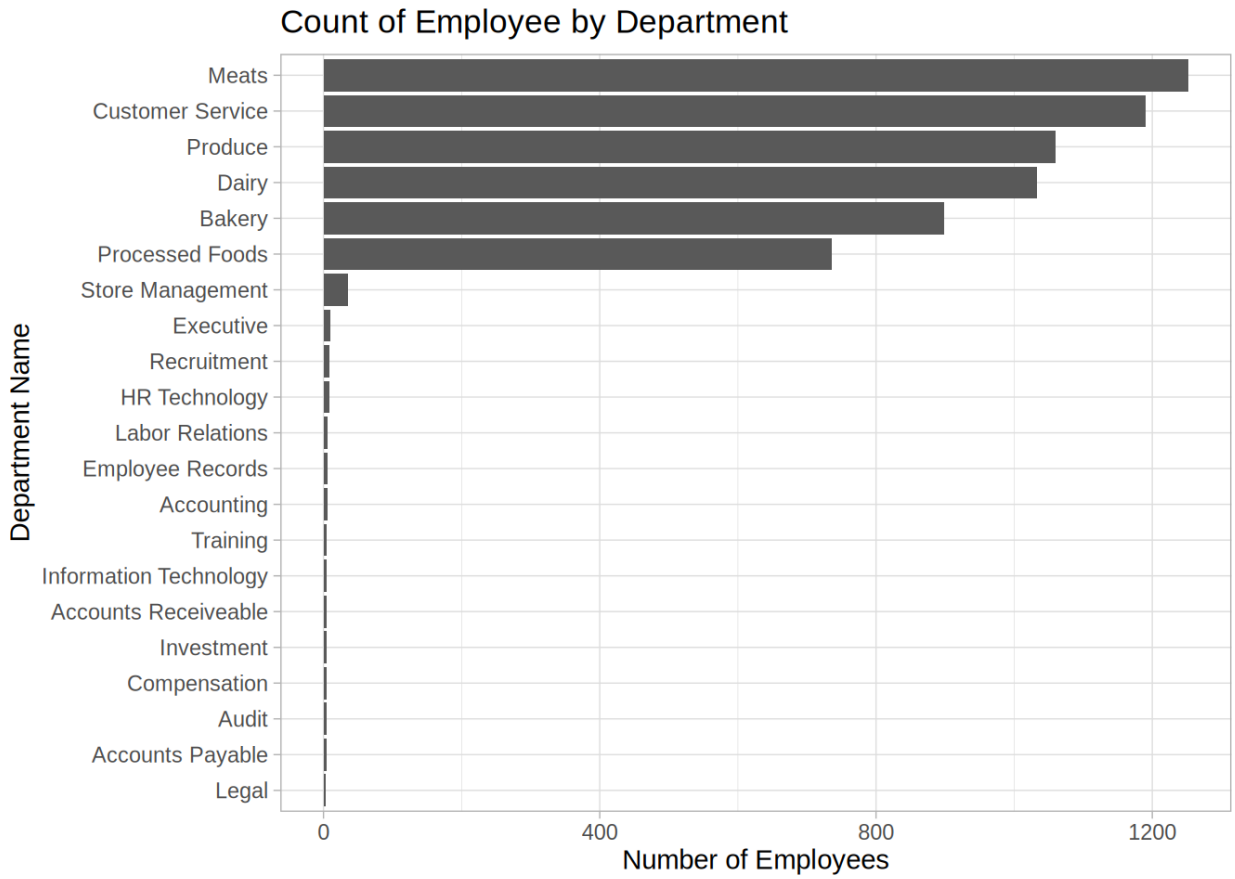


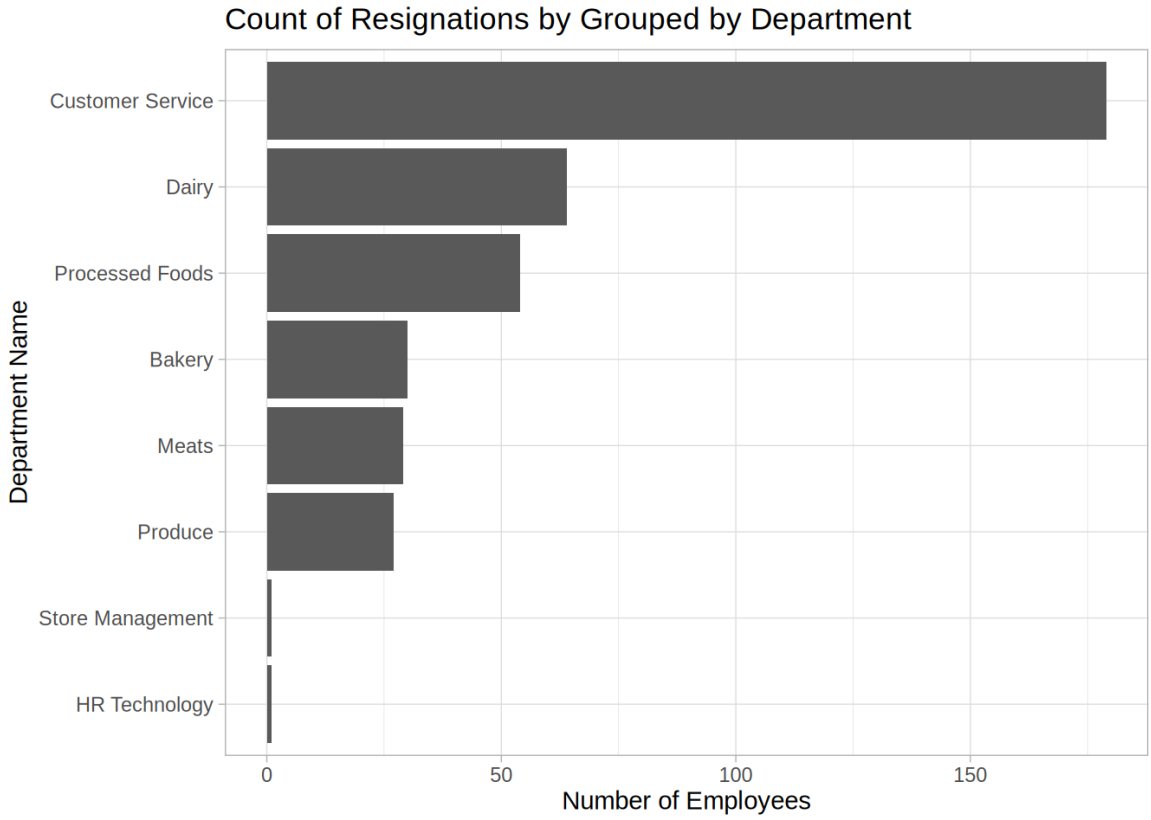
Figure 38: Bar char of employee count by department

The top 6 departments have the most employees. These departments are under the stores business unit, which has more employees than the head office.

3.2.6.2 Resignation Count Grouped by Department

```
df2 %>%
  filter(termreason_desc == 'Layoff') %>%
  count(city_name) %>%
  ggplot(aes(n, reorder(city_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Layoffs Grouped by City',
    x = 'Number of Employees',
    y = 'City Name'
  )
```

Figure 39: R code to plot a bar chart of resignations by department



From the chart above, the customer service department has an overwhelming number of resignations compared to the other departments. This could mean that departments influence resignation. However, inside each department there are individual jobs, so the numbers here might be influenced by individual jobs, rather than the whole department. Further analysis of this will be done in the later sections of this document.

3.2.7 Job Title Regarding Employee Termination

3.2.7.1 Job Title Univariate Analysis

```
df2 %>%
  count(job_title) %>%
  ggplot(aes(n, reorder(job_title, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Employee by Job Title',
    x = 'Number of Employees',
    y = 'Job Title'
  )
```

Figure 41: R code to plot a bar chart of employee count by job title

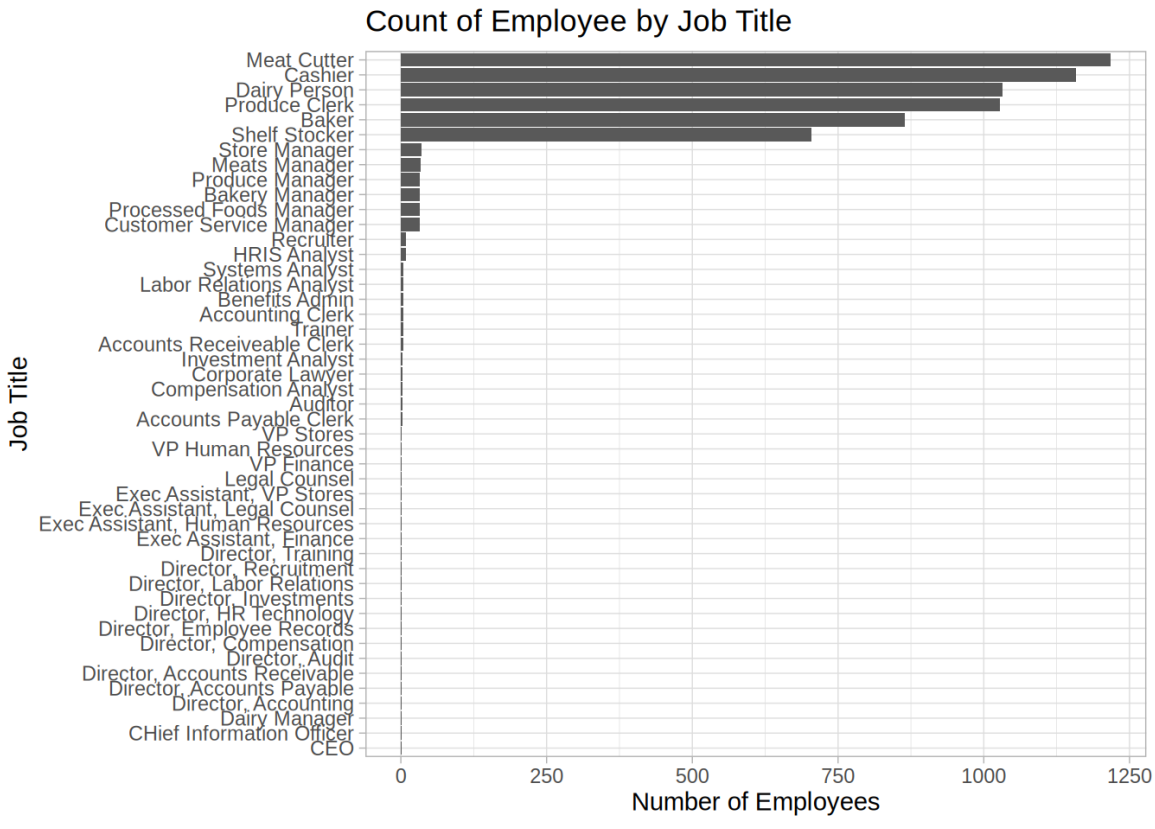


Figure 42: Bar chart of employee count by job title

There are again 6 jobs with the most employees in the bar chart. This could mean that each of the top 6 jobs here belong to each of the top 6 departments in the previous section. A further analysis of this to confirm this statement will be made in a later section of this document.

3.2.7.2 Resignation Count Grouped by Job Title

```
df2 %>%
  filter(termreason_desc == 'Resignation') %>%
  count(job_title, termreason_desc) %>%
  ggplot(aes(n, reorder(job_title, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Resignations by Grouped by Job Title',
    x = 'Number of Employees',
    y = 'Job Title'
  )
```

Figure 43: R code to plot bar chart of resignations by job title

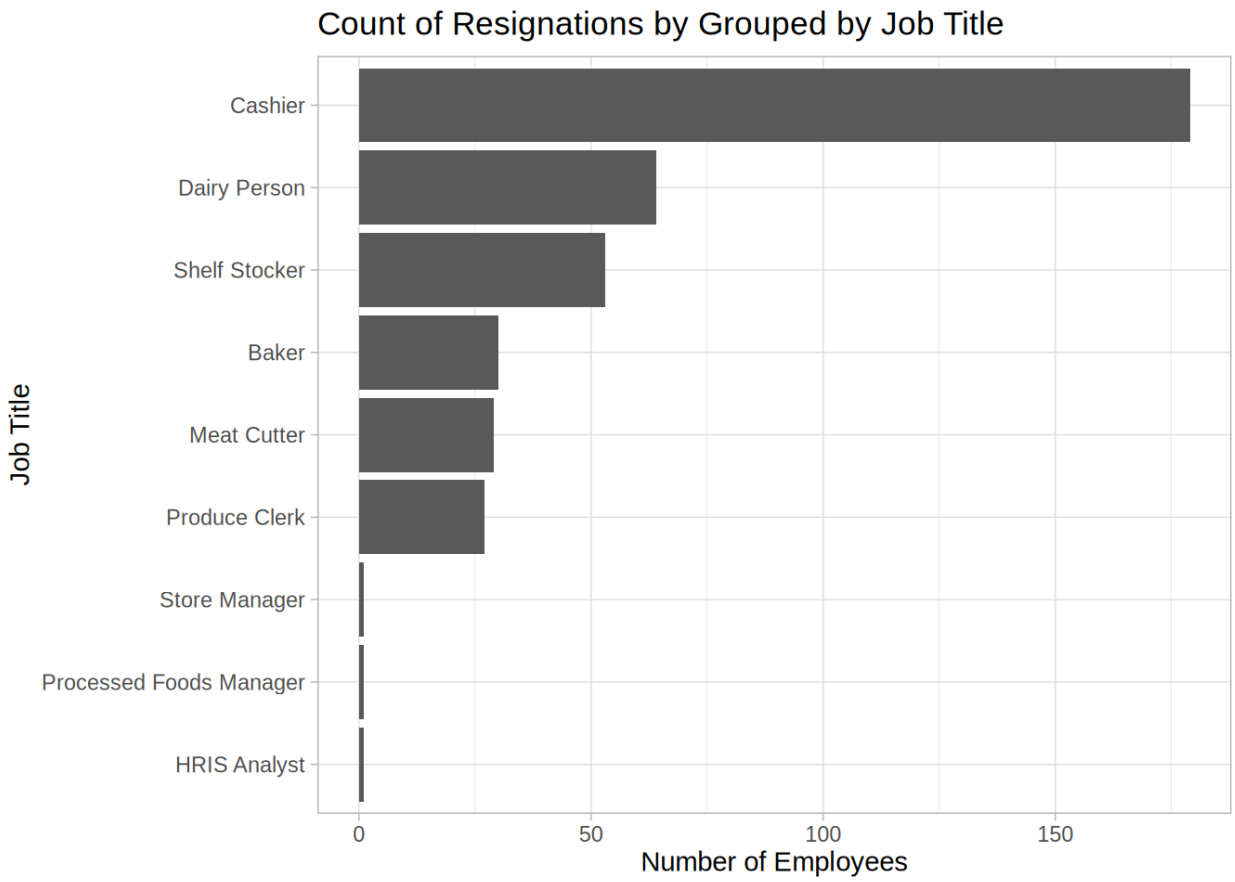


Figure 44: Bar chart of resignations by job title

The same case as the previous section. The *job_title* and *department_name* almost seem to have a one-to-one relationship. Again, further analysis of this is needed.

3.2.8 Store

3.2.8.1 Store Univariate Analysis

```
df2 %>%
  count(store_name) %>%
  ggplot(aes(n, reorder(store_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Employee by Store',
    x = 'Number of Employees',
    y = 'Store Name'
  )
```

Figure 45: R code to plot bar chart of employee count by store

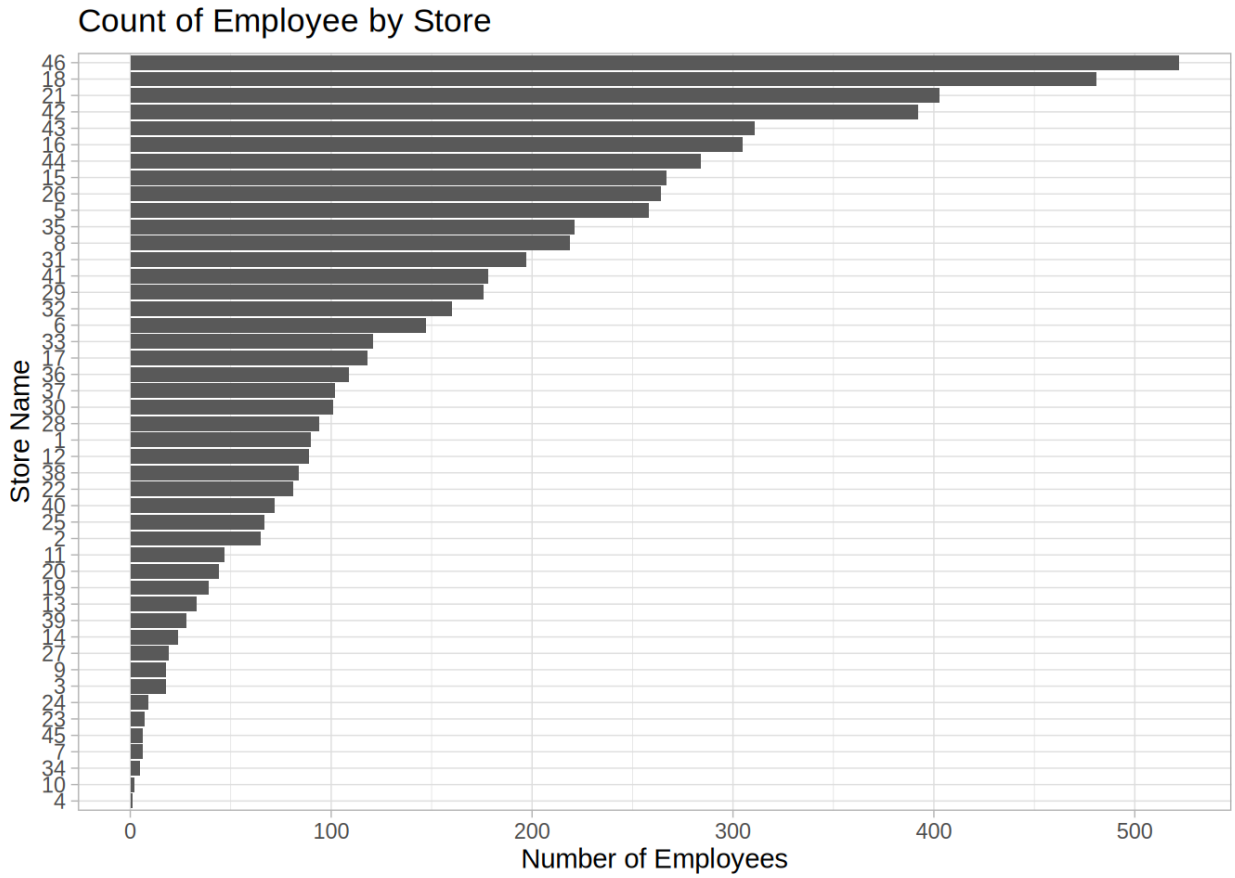


Figure 46: Bar chart of employee count by store

Looking at only the stores doesn't provide any useful insights. Also, some stores have very little employee data (some less than 10). The cause of this might be an issue with the data collection, or some other reason. This might affect analyses using this column, so I will exclude some of the stores before that. The excluded stores will be the ones with employees less than the 1st quantile value of the column.

```
# some stores have very little employees so I filtered the data to include
only
# the stores with employees above the 1st quantile of employee number.
q1 <- df2 %>% count(store_name) %>% .$n %>% quantile(.25)

store <- df2 %>%
  group_by(store_name) %>%
  filter(n() >= q1) %>%
  ungroup()
```

Figure 47: R code to filter data frame into new variable

To filter the data, the first quantile (Q1) of the number of employees grouped by stores is stored in the q1 variable. Then the data frame is grouped using the *group_by* function using the *store_name* variable, filtered to include only the stores with employees more than or equal to the 1st quantile value. The groups are then removed using the *ungroup* function and the resulting data frame is stored into the *store* variable.

3.2.8.2 Termination Reasons Grouped by Store

3.2.8.2.1 Resignations & Retirements Grouped by Store

```
store %>%
  filter(termreason_desc == 'Resignation') %>%
  count(store_name, termreason_desc) %>%
  ggplot(aes(n, reorder(store_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Resignations Grouped by Store',
    x = 'Number of Employees',
    y = 'Store Name'
  )
```

Figure 48: R code to plot bar chart of resignations by store

```
store %>%
  filter(termreason_desc == 'Retirement') %>%
  count(store_name, termreason_desc) %>%
  ggplot(aes(n, reorder(store_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Retirements Grouped by Store',
    x = 'Number of Employees',
    y = 'Store Name'
  )
```

Figure 49: R code to plot bar chart of retirements by store

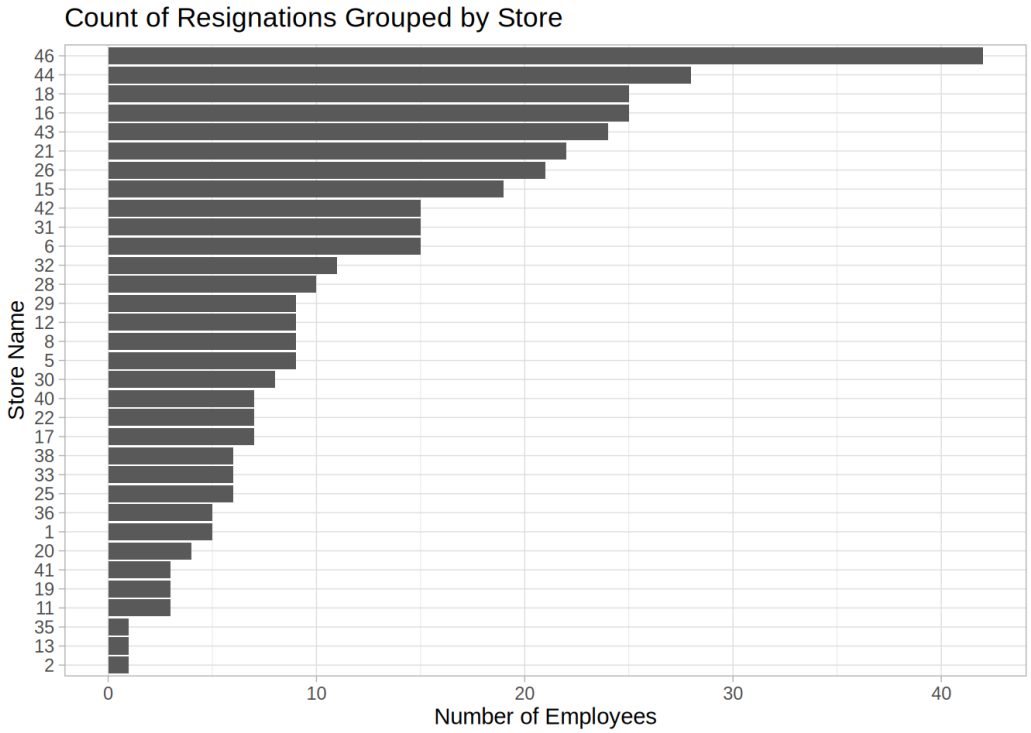


Figure 50: Bar chart of resignation by store

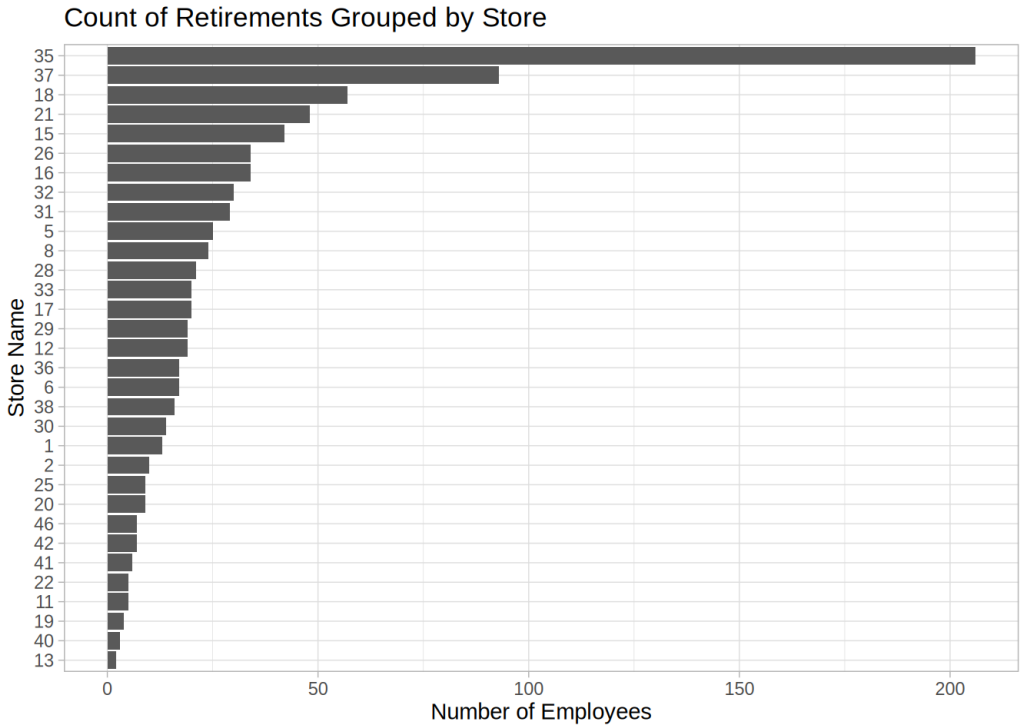


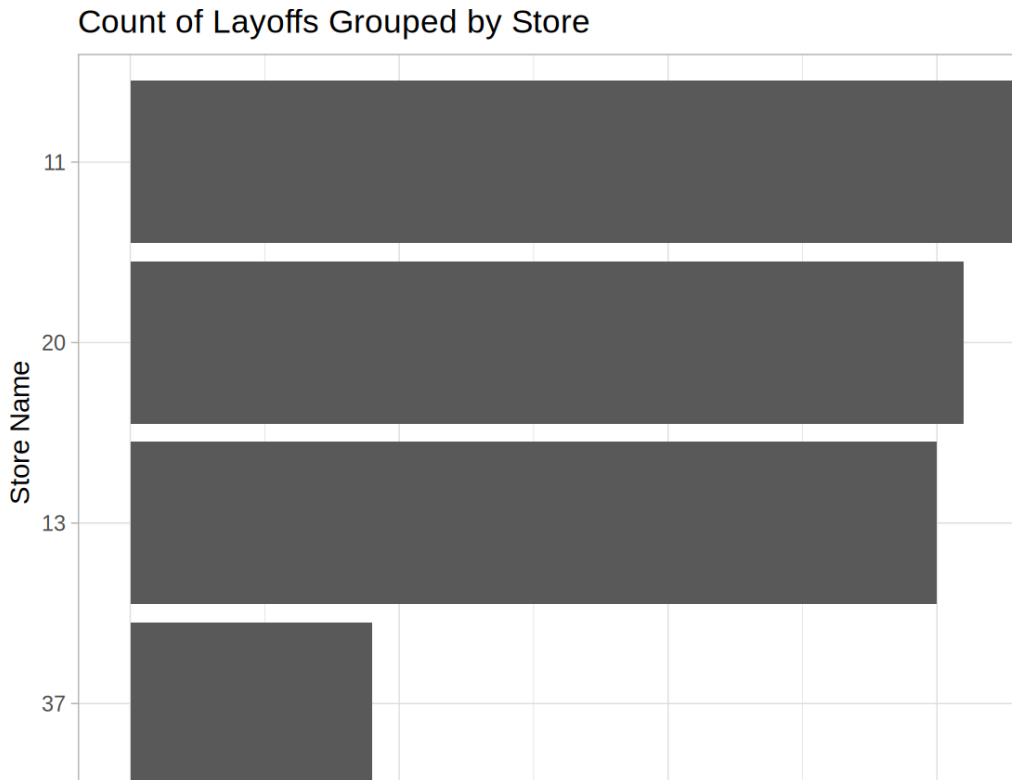
Figure 51: Bar chart of retirements by store

Still nothing useful can be determined from looking at these 2 plots. Resignation and Retirement don't seem to be affected by the store. However, the results are different if the data is filtered for layoffs.

3.2.8.2.2 Layoffs Grouped by Store

```
store %>%
  filter(termreason_desc == 'Layoff') %>%
  count(store_name, termreason_desc) %>%
  ggplot(aes(n, reorder(store_name, n))) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Count of Layoffs Grouped by Store',
    x = 'Number of Employees',
    y = 'Store Name'
  )
```

Figure 52: R code to plot bar chart of layoffs by store



There are only 4 stores with layoffs. Besides the fact that layoffs happened only on 2014 & 2015, nothing much is known about them. Further analysis of this will be done in a later section of this document to try to explain why this happened.

3.3 Answer

Based on the above analyses, there are several factors that could affect employee attrition. Besides retirement, which is caused by old age, resignation and layoff could be explained using the following variables in the dataset:

- Age
- Gender
- Job & Department
- Store

From these factors, additional analyses will be done to understand how they are affecting attrition.

4.0 Question 2: How Does Age Affect Employee Attrition?

4.1 Overview

The relationship between termination age and attrition has been discussed in the previous section. In this section, I will be focusing on the hire age, which is the age at which the employee was hired by the company. This metric would be more useful for employees, as they can focus more on certain age groups to prevent resignation early on after employment.

4.2 Analyses

4.2.1 Age During Hire

```
# new dataframe with employee age during hire
df3 <- df2 %>% mutate(hire_age = age - (STATUS_YEAR -
year(orighiredate_key)))
```

Figure 54: R code to create new dataframe with hire year column

Before the analysis, a new data frame will be created that includes the *hire_year* variable. This variable is calculated from the latest age of the employees subtracted with the difference between the latest year and the original hire year. The transformation is done using the *mutate* function.

```
df3 %>%
  ggplot(aes(hire_age)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = seq(19, 55)) +
  labs(
    title = 'Distribution of Age of Employees at Hire',
    x = 'Hire Age',
    y = 'Number of Employees'
  )
```

Figure 55: R code to plot histogram of hire age distribution

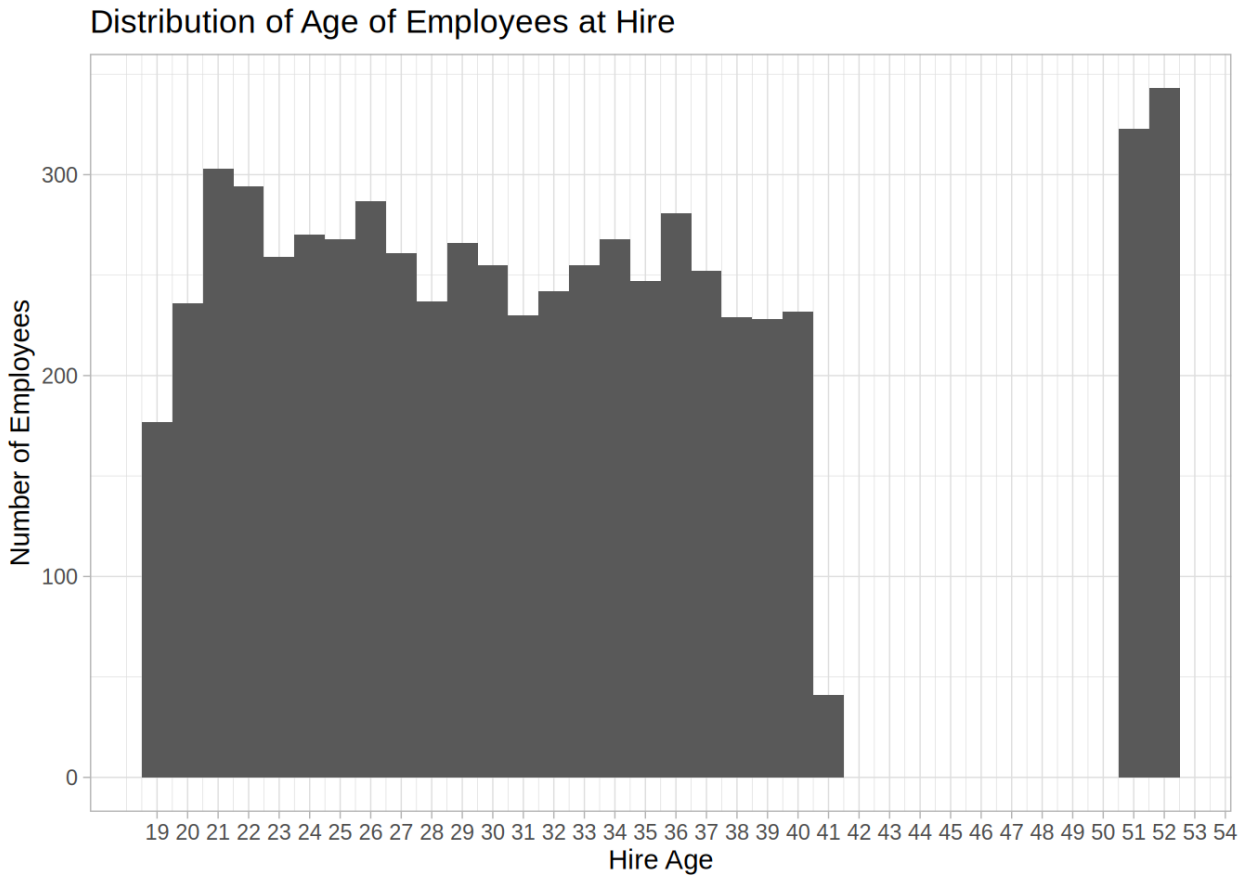


Figure 56: Histogram of hire age distribution

Most of the employees were between the age of 20 and 40 at the time of hiring. There's also a high number of employees aged 51-52, but not many employees in their 40s during hire.

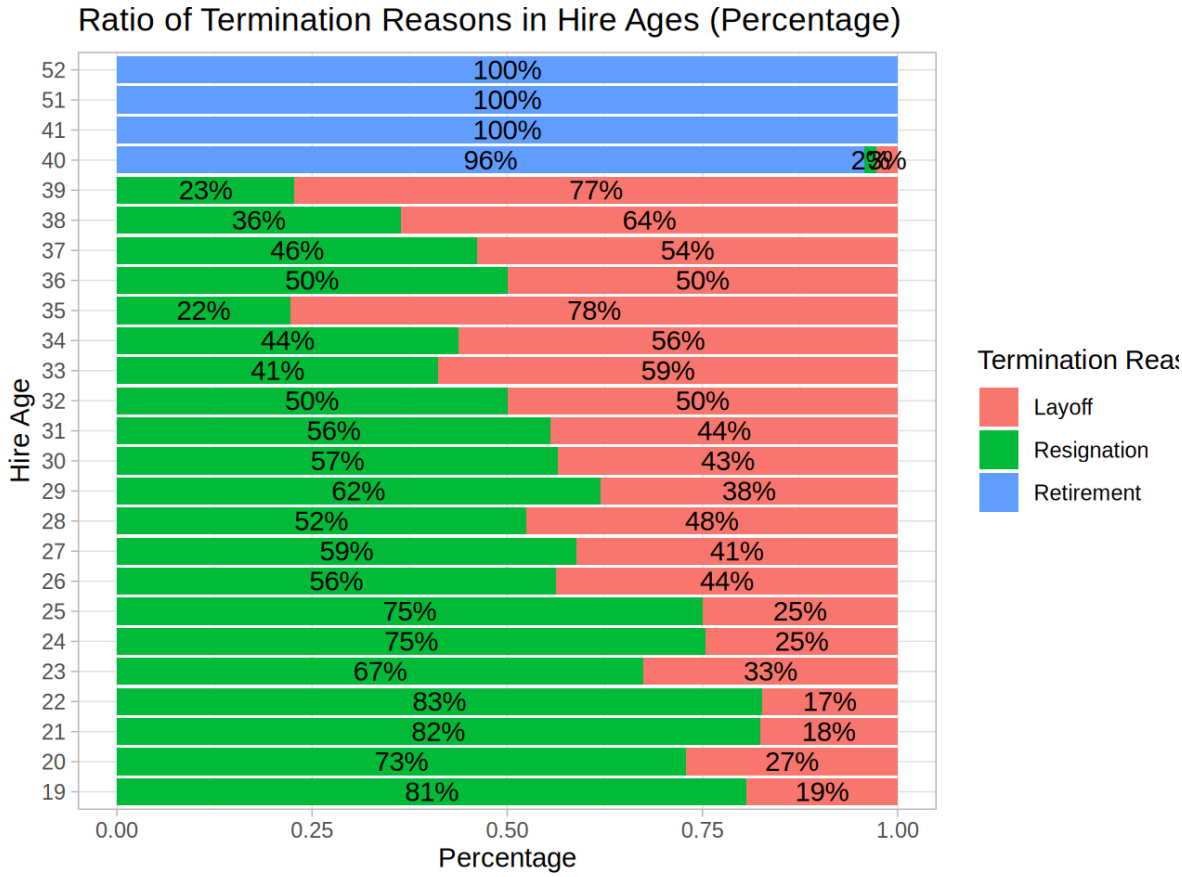
4.2.2 Termination Reasons Grouped by Hire Age

```
df3 %>%
  filter(STATUS == 'TERMINATED') %>%
  count(hire_age, termreason_desc) %>%
  group_by(hire_age) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(n, as.factor(hire_age),
             fill = termreason_desc,
             label = paste(np, '%', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5)) +
  labs(
    title = 'Ratio of Termination Reasons in Hire Ages (Percentage)',
    x = 'Percentage',
    y = 'Hire Age',
    fill = 'Termination Reason'
  )
```

Figure 57: R code to plot stacked bar chart of terminations by hire age

What the code here is trying to achieve is to plot a 100% stacked bar chart, which is used to compare the ratio of groups in categorical variables. After filtering and counting the data using the hire age and termination reason variables, the data frame is transformed using the *mutate* function. The new variables added are used to get the percentage of each group, which is the termination reason here, in each hire age.

The *fill* option is used to specify the segments of each bar, and the *label* is used to specify the labels on the chart. The *position = 'fill'* option in *geom_bar* is what makes chart a stacked bar chart. g*eom_text* is then used to place the labels, as well as move them to the center of each segment using the *position_fill* with vjust, or vertical justification to 0.5, which means middle.



When comparing the termination reasons with employee age of hire, some patterns can be seen. From the chart above, we can see that employees hired at a younger age tend to resign more. The opposite happens to employees hired at an older age, as they tend to get laid off more (excluding ages 40 and above, which are more likely to retire).

There are several potential causes for this. One of them is the lack of 'the feeling of accomplishment' by younger employees. Younger employees want to make an impact (University of Waterloo, 2020). Another is the lack of challenge or growth in the workplace. Younger employees tend to get bored with their jobs quickly if it doesn't provide them with the opportunities to showcase their skills and abilities (Ltd, 2019).

There could be other factors in play here such as the relationship with their employers or coworkers, or low salaries, but it is difficult to determine their significance using the dataset provided alone.

4.2.3 Correlation Between Hire Age and Tenure

```
df3 %>%
  filter(STATUS == 'TERMINATED') %>%
  ggplot(aes(hire_age, length_of_service, color = termreason_desc)) +
  geom_point() +
  labs(
    title = 'Correlation between Hire Age and Tenure',
    x = 'Hire Age',
    y = 'Tenure',
    color = 'Termination Reason'
  )
```

Figure 59: R code to plot scatter plot of employee hire age & tenure

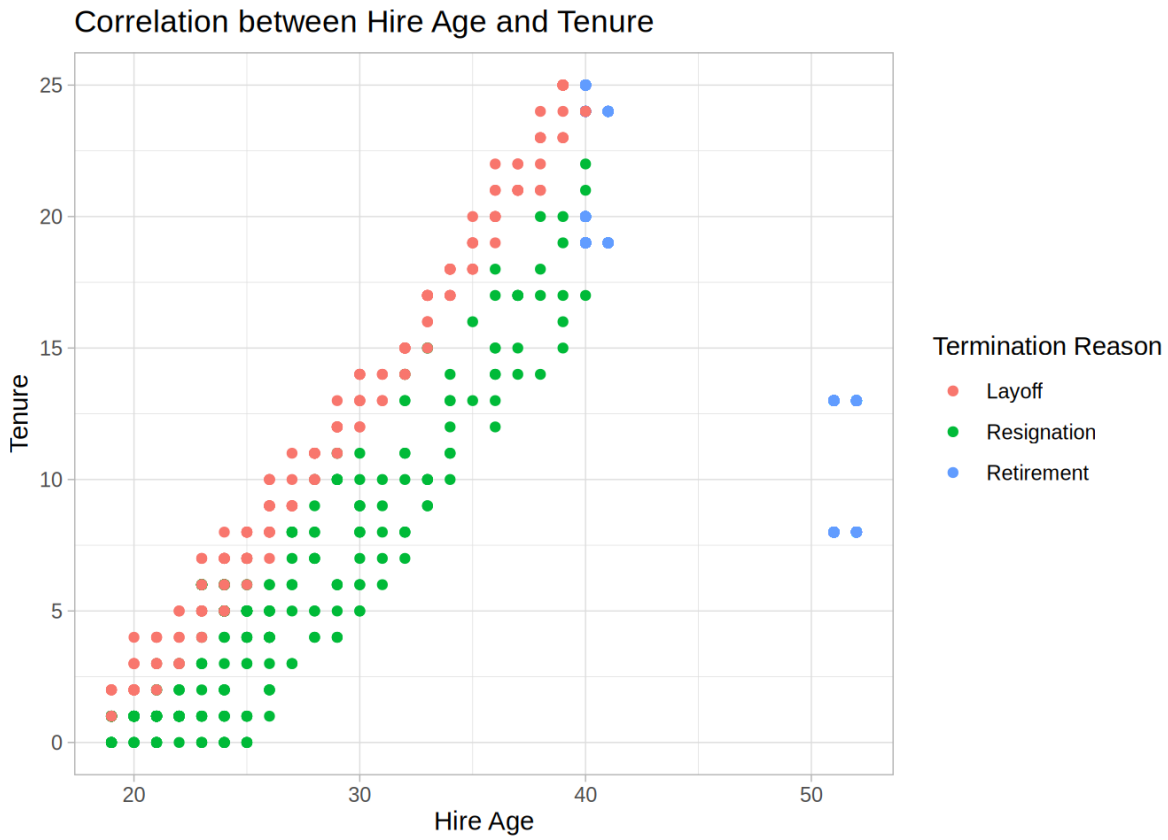


Figure 60: Scatter plot of employee hire age & tenure

Just like termination age, hire age also has a strong, positive correlation with employee tenure. The Pearson's correlation coefficient for this is lower than before, at 0.5851923, but can still be considered as a high value.

4.3 Answer

Based on the above analyses, it can be concluded that age, more specifically the age during hire, can influence employee attrition. Younger employees are more likely to resign, whereas older employees during hire are more likely to retire or get laid off. Younger employees are more likely to resign due to reasons such as boredom or lack of accomplishment.

4.4 Recommendations

Since the employees that are most likely to leave the company are younger ones, it should benefit the company to focus more on them. Since this is a retail company, it might be difficult to give them a sense of achievement or more challenges.

One thing they can do is to give them more responsibilities. The company can assign a few employees to handle an event such as a limited time event, which can be specific to a store or a city. This can provide employees with the chance for them to show their skills and give them the engagement needed to not get bored working there.

5.0 Question 3: How Does Jobs Affect Employee Attrition?

5.1 Overview

Referring to the analyses at section 3.2.7, there are several jobs that have a higher number of terminations. In particular, cashiers have a higher number of resignations than the other jobs. A deeper analysis will be conducted in this section to try to answer those questions.

5.2 Analyses

5.2.1 Ratio of Termination Reasons by Job

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  count(job_title, termreason_desc) %>%
  group_by(job_title) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(n, reorder(job_title, n),
             fill = termreason_desc,
             label = paste(np, '%', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5)) +
  labs(
    title = 'Ratio of Termination Reasons in Job Titles (Percentage)',
    x = 'Percentage',
    y = 'Job Title',
    fill = 'Termination Reason'
  )
```

Figure 61: R code to plot stacked bar chart of terminations in jobs

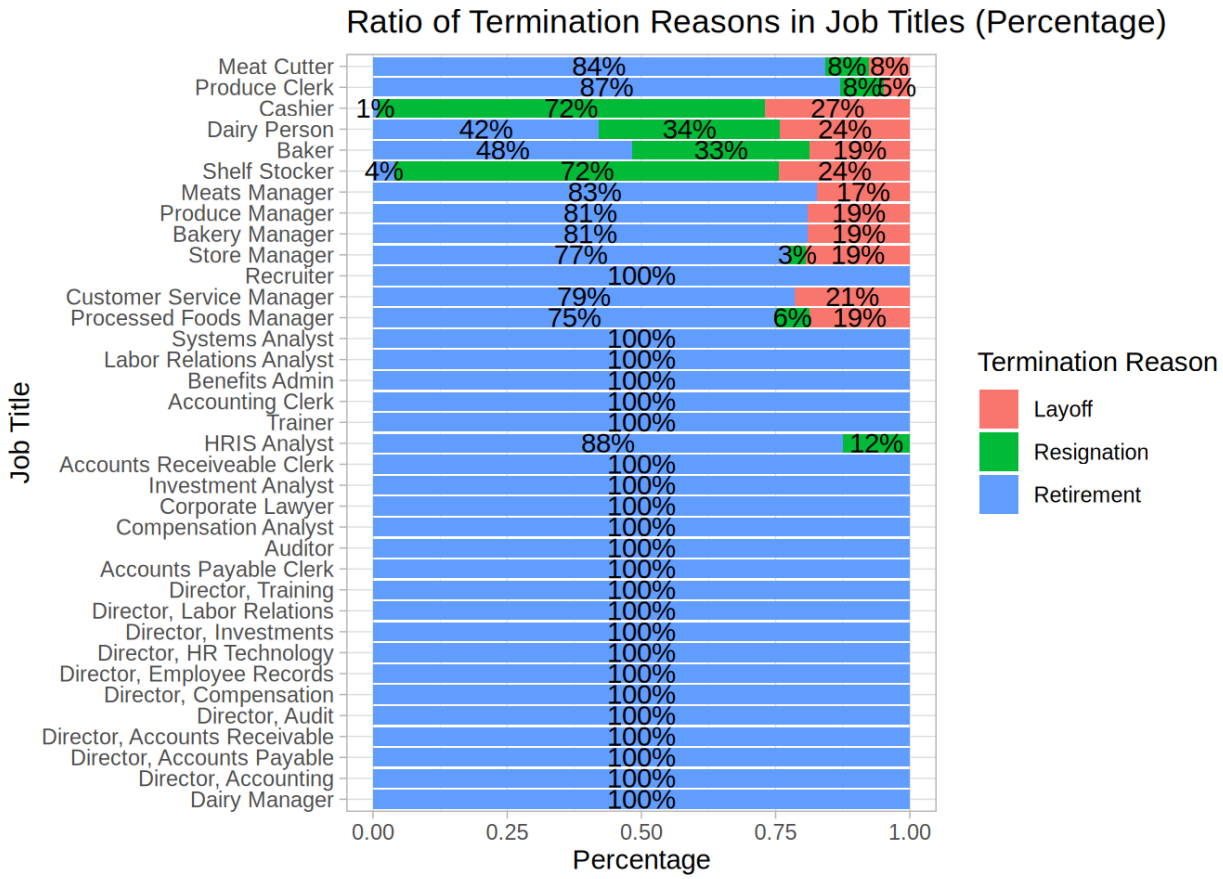


Figure 62: Stacked bar chart of terminations in jobs

By looking at the ratio of termination reasons in each job, cashiers and shelf stockers have quite a lot more resignations compared to retirements and layoffs. Following these 2 are dairy persons and bakers. Meat cutters and produce clerks have the most retirements, both in count and in ratio.

From this chart alone, it can be inferred that specific jobs, like cashiers, can be held accountable for termination. Before coming to this conclusion, verification would need to be done to make sure that the department is not responsible for this, since jobs belong to departments. It is to make sure that the jobs of interest here do not belong to the same department.

5.2.2 Ratio of Termination Reasons by Department

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  count(department_name, termreason_desc) %>%
  group_by(department_name) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(n, reorder(department_name, n),
             fill = termreason_desc,
             label = paste(np, '%', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5)) +
  labs(
    title = 'Ratio of Termination Reasons in Departments (Percentage)',
    x = 'Percentage',
    y = 'Department Name',
    fill = 'Termination Reason'
  )
```

Figure 63: R code to plot stacked bar chart of terminations in departments

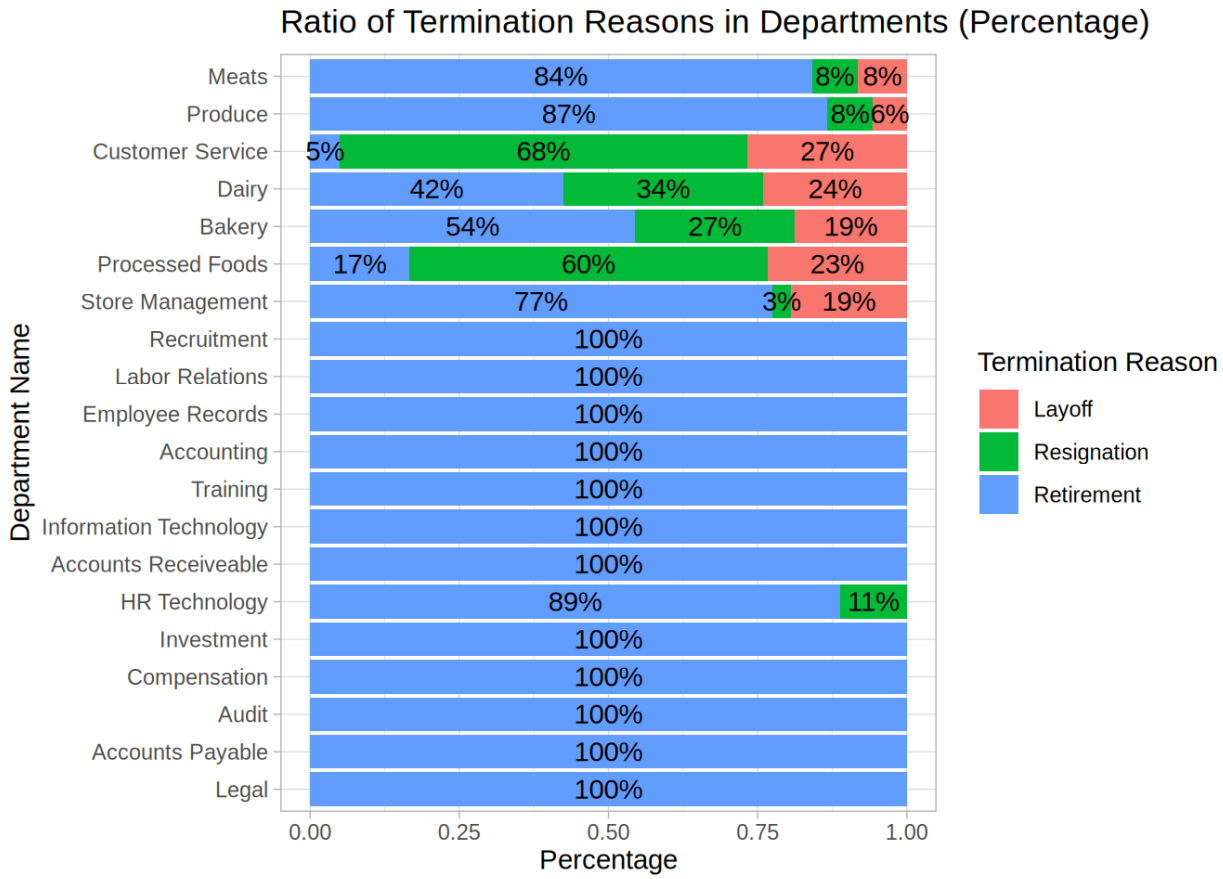


Figure 64: Stacked bar chart of terminations in departments

Here is the check to help verify the previous section's statements. The chart here looks like the previous section. The ratio of termination reasons doesn't look far off as well. The jobs from the previous section are most likely in different departments. As a side note, some of the departments have very little employees. These are the departments under the head office business unit. The ones with a lot of employees are from the stores business unit.

5.2.3 Which Jobs Belong to Which Department?

```
df2 %>% count(department_name, job_title)
```

Figure 65: R code to see which jobs belong to which departments

department_name <chr>	job_title <chr>	n <int>
Compensation	Compensation Analyst	3
Compensation	Director, Compensation	1
Customer Service	Cashier	1158
Customer Service	Customer Service Manager	32
Dairy	Dairy Manager	1
Dairy	Dairy Person	1032
Employee Records	Benefits Admin	5
Employee Records	Director, Employee Records	1
Executive	CEO	1
Executive	CHief Information Officer	1

11-20 of 47 rows Previous 1 **2** 3 4 5 Next

Figure 66: Table showing jobs and their departments (1)

department_name <chr>	job_title <chr>	n <int>
Information Technology	Systems Analyst	5
Investment	Director, Investments	1
Investment	Investment Analyst	3
Labor Relations	Director, Labor Relations	1
Labor Relations	Labor Relations Analyst	5
Legal	Corporate Lawyer	3
Meats	Meat Cutter	1218
Meats	Meats Manager	34
Processed Foods	Processed Foods Manager	32
Processed Foods	Shelf Stocker	704

31-40 of 47 rows Previous 1 2 3 **4** 5 Next

Figure 67: Table showing jobs and their departments (2)

This confirms the claim that jobs themselves affect attrition, not departments.

5.3 Answer

Based on these analyses only, there doesn't seem to be anything in the provided dataset that could explain why some jobs have higher termination rates than others. However, this can be explained by considering job satisfaction and salary. Cashiers and shelf stockers have a very high resignation rate. There is not enough information in the dataset provided to understand why this is happening, so external resources will be used.

This dataset contains data of employees from British Columbia (BC), Canada, so the external data gathered will be specific to this region. Cashiers and shelf stockers in BC have below average salary and job satisfaction. The average living cost in BC is $1,839 a month for 1 person and the average salary after taxes is $3,048 (LivingCost.org, 2021).

The average salaries in BC are $2,559 a month for cashiers and $2,600 a month for shelf stockers (Talent.com, 2021). Both are lower than the average, which may cause employees to quit their jobs to find a better paying one. Also, cashiers often deal with stressful working conditions such as rude customers and terrible bosses. This can also cause them to look for a job with better working conditions.

5.4 Recommendations

A simple solution that can help to reduce resignations in these jobs is to increase the salary. However, the raise would most likely not go above the average, as any more would not be beneficial for the company. Also raising the salary doesn't help fix the job dissatisfaction. For that, a fix could be to improve the working conditions. It can be implemented in several ways such as providing free lunch or snacks, more break times, giving awards, etc.

6.0 Question 4: How Does Gender Affect Employee Attrition?

6.1 Overview

This section aims to answer the questions from section 3.2.4. Previously, only the number of employees was analyzed. In this section, employee gender will be analyzed with other variables to try to get more insights.

6.2 Analyses

6.2.1 Ratio of Termination Reasons by Gender

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  count(gender_full, termreason_desc) %>%
  group_by(gender_full) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(x = n, y = gender_full,
             fill = termreason_desc,
             label = paste(np, '% (', n, ')', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5), size = 3) +
  labs(
    title = 'Ratio of Termination Reasons in Each Gender',
    x = 'Percentage',
    y = 'Gender',
    fill = 'Termination Reason'
  )
```

Figure 68: R code to plot stacked bar chart of termination reasons in each gender

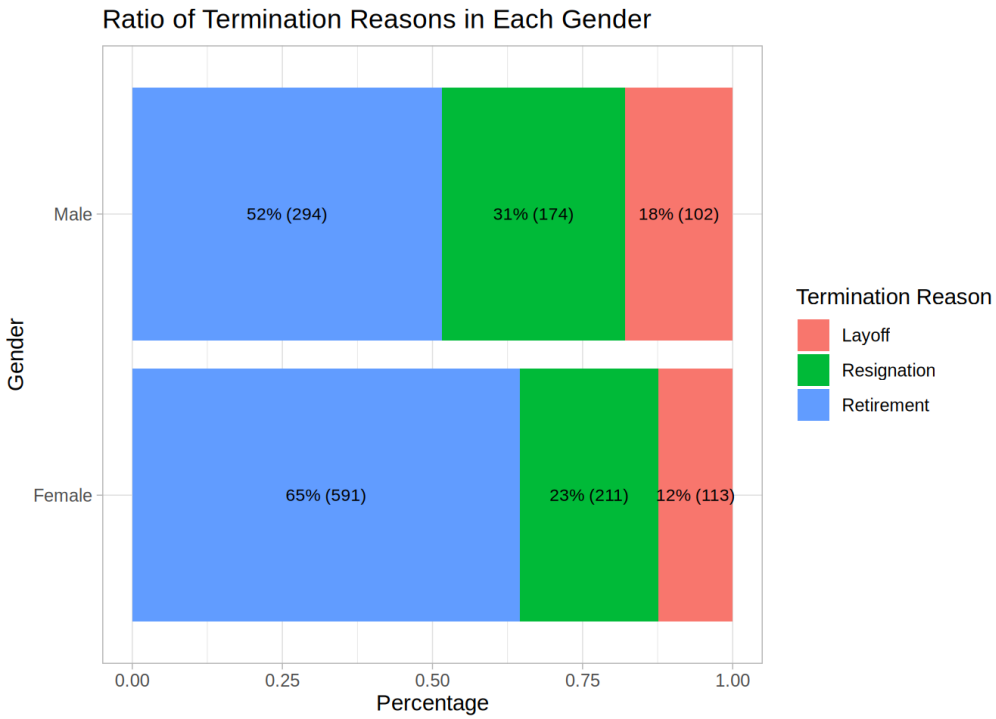


Figure 69: Stacked bar chart of termination reasons in each gender

The ratio for both male and female is quite similar, with females being a little bit more 'loyal' to the company, because of the larger ratio of retirements over resignations and layoffs.

6.2.2 Tenure by Gender

```
df2 %>%
  group_by(gender_full) %>%
  summarise(avg_tenure = sum(length_of_service) / n()) %>%
  ggplot(aes(gender_full, avg_tenure)) +
  geom_bar(stat = 'identity') +
  labs(
    title = 'Average Tenure of Employees Grouped by Gender',
    x = 'Gender',
    y = 'Average Tenure'
  )
```

Figure 70: R code to plot bar chart of average employee tenure

Here, before plotting a new variable, *avg_tenure*, the average tenure, is computed. Plotting the bar chart of the sum of tenure would be incorrect, as the number of female and male employees are different. a*vg_tenure* is computed by grouping by gender first, then dividing the sum of tenure with the number of records.

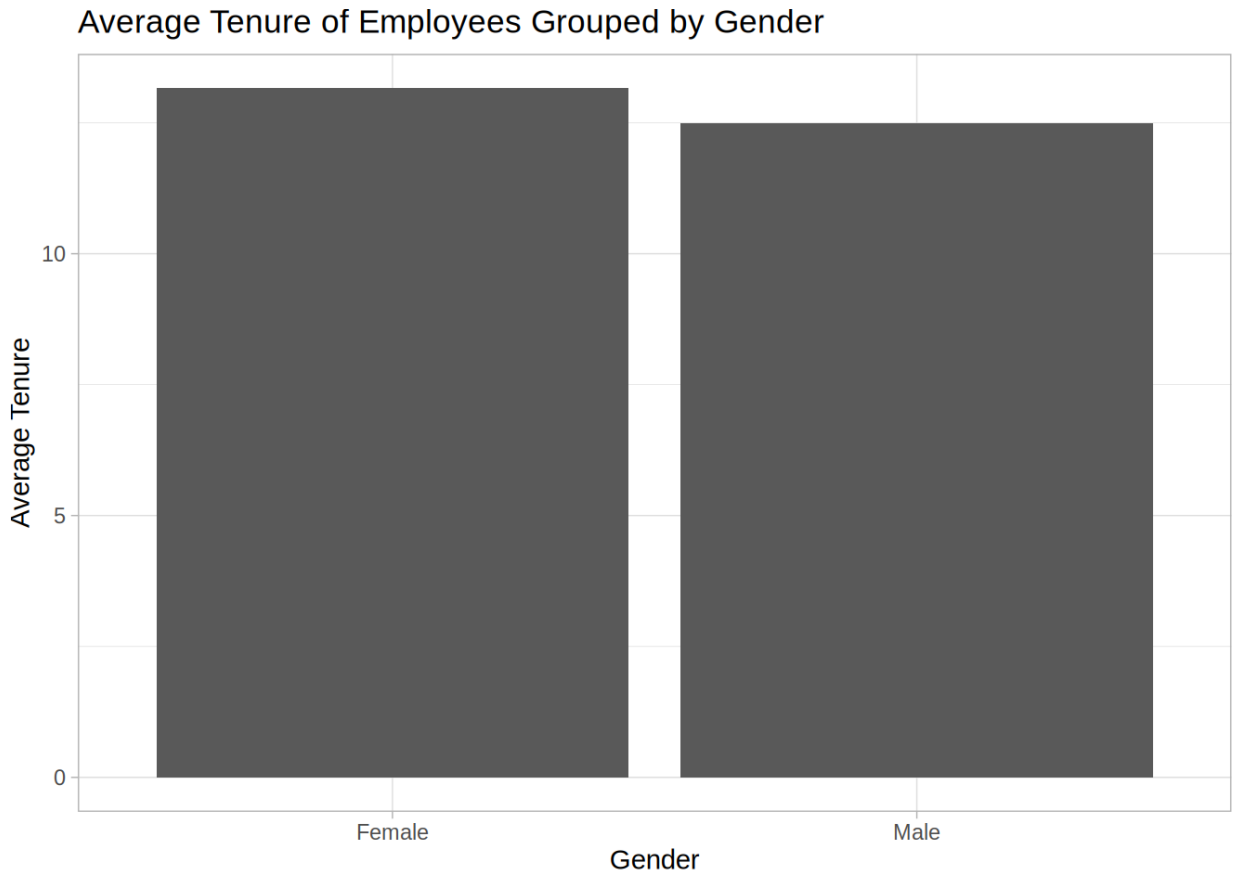


Figure 71: Bar chart of average employee tenure

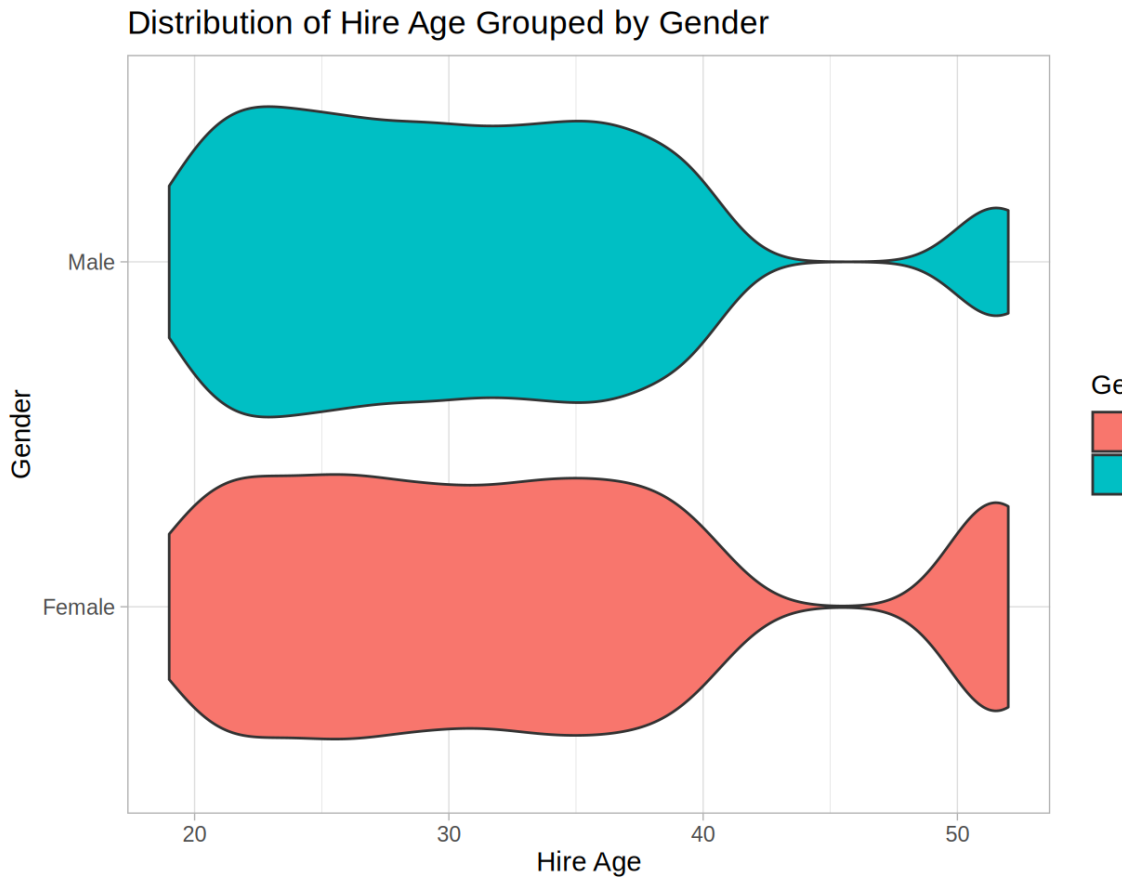
The average tenure of female employees is higher than male employees. This backs up the statement in the previous section about female employees being more loyal to the company.

6.2.3 Age of Hire by Gender

```
df3 %>%
  ggplot(aes(hire_age, gender_full, fill = gender_full)) +
  geom_violin() +
  labs(
    title = 'Distribution of Hire Age Grouped by Gender',
    x = 'Hire Age',
    y = 'Gender',
    fill = 'Gender'
  )
```

Figure 72: R code to plot violin plot of hire age distribution in genders

The function *geom_violin* is used to draw a violin plot. It is a type of plot that is like a boxplot. Its main use is to visualize distribution.



Most of the employees are hired between the age of 20 and 40, regardless of gender. This aligns with the analysis done in section 4.2.1. Gender and hire age do not seem to be connected.

6.2.4 Termination Reasons of Gender Grouped by Job Title

6.2.4.1 Cashiers & Shelf Stockers

```
df2 %>%
  filter(STATUS == 'TERMINATED' &
         job_title %in% c('Cashier', 'Shelf Stocker')) %>%
  count(gender_full, termreason_desc, job_title) %>%
  group_by(gender_full) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(x = n, y = gender_full,
             fill = termreason_desc,
             label = paste(np, '% (', n, ')', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5), size = 3) +
  facet_wrap(~job_title, ncol = 1) +
  labs(
    title = 'Ratio of Cashier & Shelf Stocker Termination Reasons by Gender',
    x = 'Percentage',
    y = 'Gender',
    fill = 'Termination Reason'
  )
```

Figure 74: R code to plot stacked bar chart of termination reason by gender of cashiers and shelf stockers

What the code here is trying to do is the same as in section 6.2.1, which is to plot a stacked bar chart of termination reasons in each gender. But this time, the chart will be additionally grouped by the job title. The data frame here is filtered to only contain records of where the job is cashier or shelf stocker. This is done using the %in% operator, which checks if the value in its left-hand side is inside of a data structure in its right-hand side. f*acet_wrap* is used again to create multiple plots. The *ncol* option is set to 1, which sets the number of columns in the combined plots to be 1.

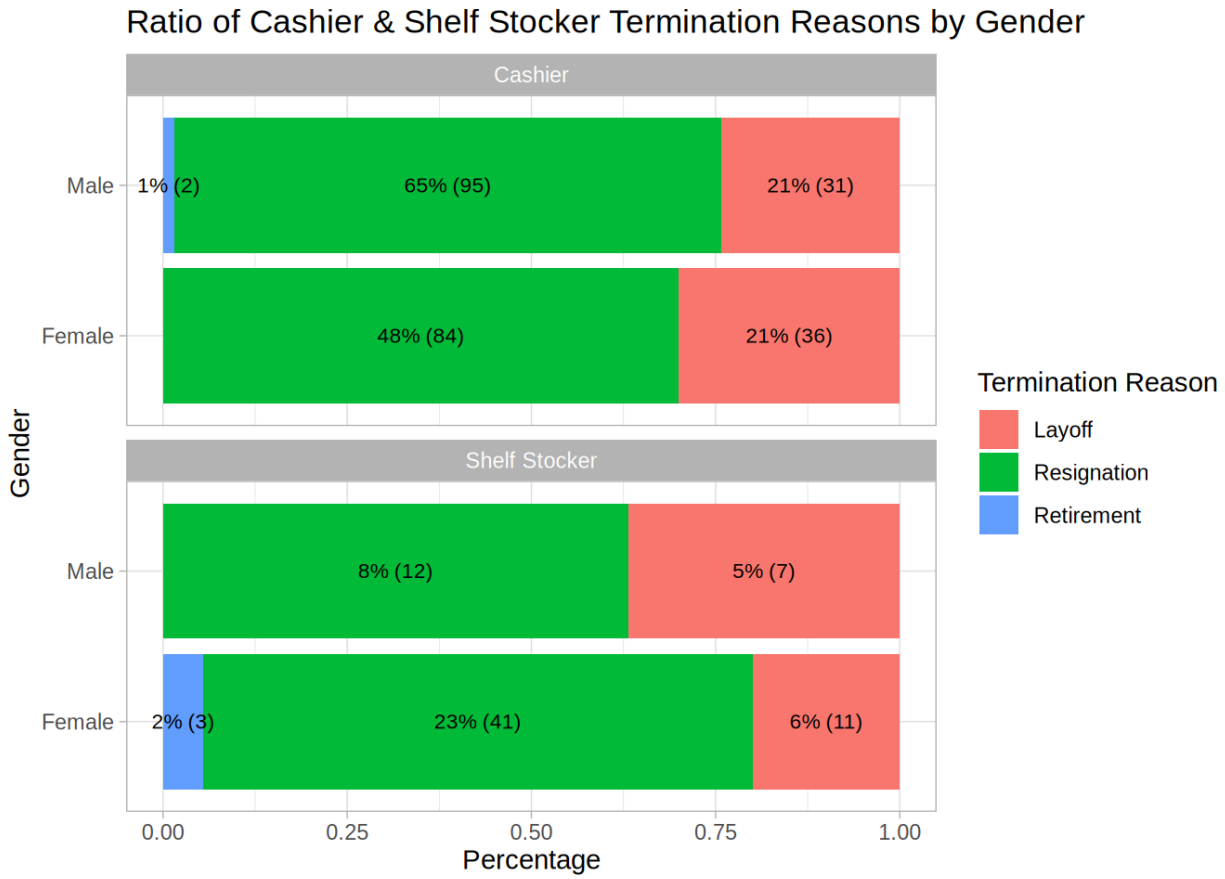


Figure 75: Stacked bar chart of termination reason by gender of cashiers and shelf stockers

Male cashiers are slightly more likely to resign than female cashiers. Shelf stockers, however, have more female employee resignations than male. This means that gender might not be affecting resignation. The results for shelf stocker ratios are the opposite of cashiers.

6.2.4.2 Top 10 Jobs

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  filter(job_title %in% (
    df2 %>% count(job_title) %>% arrange(desc(n)) %>% head(10) %>%
.$job_title
  )) %>%
  count(gender_full, termreason_desc, job_title) %>%
  group_by(job_title, gender_full) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(n, gender_full,
             fill = termreason_desc,
             label = paste(np, '%', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5), size = 3) +
  facet_wrap(~job_title) +
  labs(
    title = 'Ratio of Termination Reasons by Gender of Top 10 Jobs',
    x = 'Percentage',
    y = 'Gender',
    fill = 'Termination Reason'
  )
```

Figure 76: R code to plot gender & termination reason of top 10 jobs

Just like the previous section, the data frame is filtered to contain only a specific subset. In this code snippet, the data frame is filtered to include only the records with only the jobs that are contained in a 'subquery'. The subquery is the top 10 jobs by the number of employees. Piping into *.$n* is the same as doing *df2$n*, which returns the column specified in the right-hand side of the $ symbol.

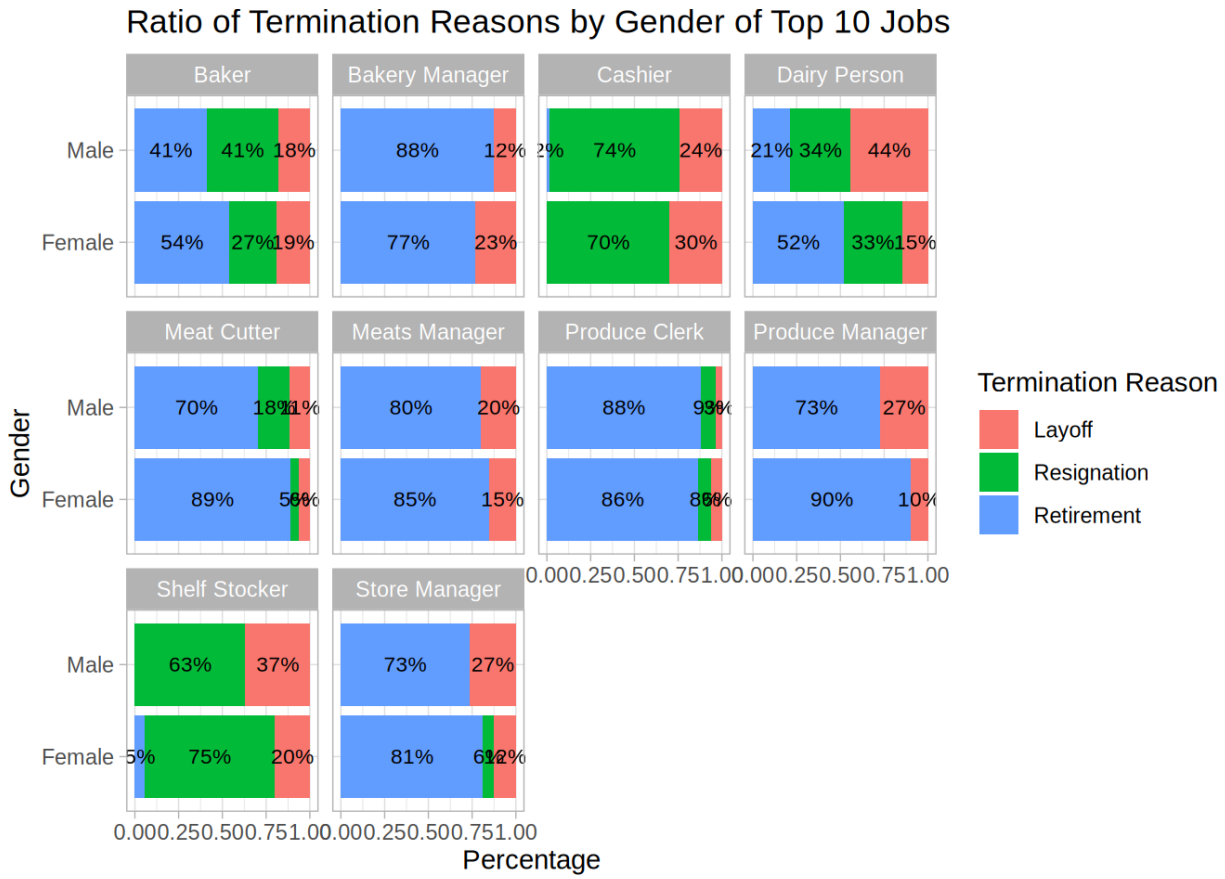


Figure 77: Stacked bar chart of gender & termination reason of top 10 jobs

There doesn't seem to be any noticeable patterns in resignation of both genders in different jobs. This analysis further strengthens the claim in the previous section about gender not influencing resignation. However, most of the jobs have more male employee layoffs then female, which supports the previous claim from section 6.2.1.

6.3 Answer

Besides the fact that female employees are less likely to resign or get laid off than male employees, there isn't anything else that can be inferred about gender. Again, I will be referring to external sources to try to explain why this happened.

A survey of 4,000 women who had left their jobs was conducted by LinkedIn in 2015. They had identified the different termination reasons of men and women. The top reasons for both men and women are "the lack of advancement opportunity" and "the dissatisfaction with senior leadership". Womens' third reason is "the dissatisfaction with the work culture/environment", while for men it's "the lack of more challenging work". It can be inferred from this that women in general values highly a good working environment and equal opportunities (Lewis, 2015).

6.4 Recommendations

A suggestion that could work for female employees is to conduct coaching programs. Referring to the analyses for question 2, these coaching programs can be targeted at younger employees, as they are the most prone to early resignation. These programs can give female employees the insight they needed about the company's working environment and job opportunities. As for male employees, a solution could be to give them more challenging tasks.

7.0 Question 5: What Causes Employee Layoff?

7.1 Overview

The layoffs in the company happened only in the years 2014 and 2015, so there isn't a lot of data on layoffs. Some clues are found in section 3.2.8, and will be analyzed further here, but it is difficult to say just from the provided dataset alone, what makes an employee more likely to be laid off. This section is an extension of the previous sections, and will focus on the store variable, which hasn't been analyzed much.

7.2 Analyses

7.2.1 Employee Status by Store

```
store %>%
  group_by(store_name, termreason_desc) %>%
  count() %>%
  group_by(store_name) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(n, reorder(store_name, n),
             fill = termreason_desc,
             label = paste(np, '%', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5)) +
  labs(
    title = 'Ratio of Active & Terminated Employees by Store (Percentage)',
    x = 'Percentage',
    y = 'Store Name',
    fill = 'Termination Reason'
  )
```

Figure 78: R code to plot stacked bar chart of terminations by store

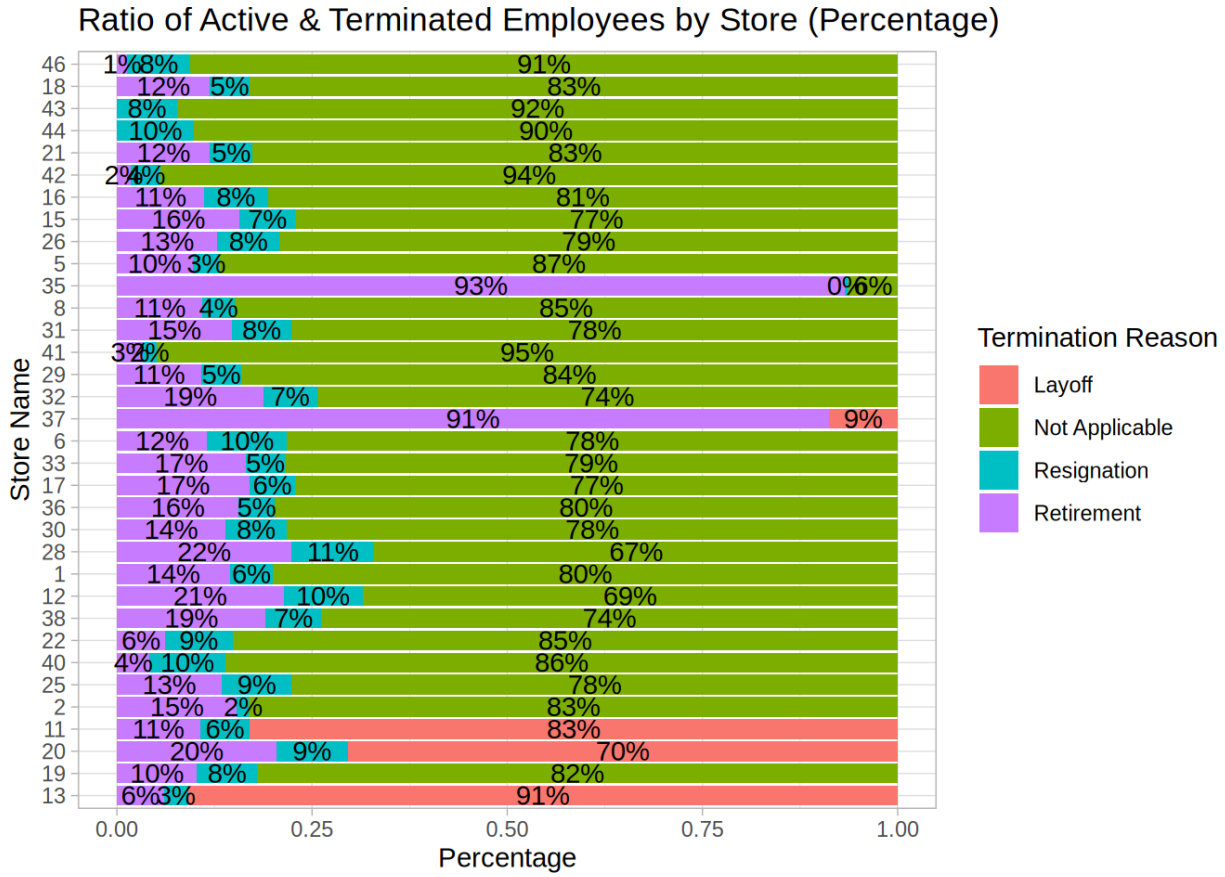


Figure 79: Stacked bar chart of terminations by store

Just like in section 3.2.8, the stores listed here have been filtered to exclude stores with very few employees. We can see that stores 11, 13, and 20 have the largest ratio of layoffs. Also, they do not have any more active employees. This could mean that the store has shut down, or they have stopped sharing their employee data. However, the case that they have stopped sharing their data is a very unlikely case. Store shutdown is believable, but there is no information in the dataset provided to support this claim. External sources also cannot be used here as there is no information about the company itself.

Another point of interest is store 37, with over 90% of their employees retired, and the rest laid off. This might be caused by the carelessness of the store's human resources department, which failed to recruit new employees to sustain the store. Store 35 also has a similar pattern to 37, with only just a few employees remaining. For the later sections, the analyses will just focus on layoffs, and on certain stores only, namely store 11, 13, and 20.

7.2.2 Which Cities Do the Stores Belong In?

```
df2 %>%
  group_by(city_name) %>%
  summarise(stores = toString(unique(store_name))) %>%
  arrange(desc(stringr::str_length(stores)))
```

Figure 80: R code to see which stores belong to which cities

Here the code is trying to concatenate the store names in the same city into a new column. The *summarise* function is used to get only the unique stores in each city using the *unique* function, then transforming them into a string using the *toString* function. To get this to work the data frame has to be grouped by the variable *city_name* first. Then it is arranged descendlingly using the length of the string, using the *str_length* function from the *stringr* package. It is arranged this way because more stores would result in a longer string.

city_name <chr>	stores <chr>
Vancouver	35, 43, 44, 41, 42, 45
New Westminster	21, 20
Victoria	46, 37
Dease Lake	10
Fort Nelson	11
Fort St John	12
Grand Forks	13
Haney	14
Kamloops	15
Kelowna	16

1-10 of 39 rows Previous 1 2 3 4 Next

Figure 81: Table of cities and their stores

It looks like most of the cities have only 1 store in them, except for the big cities, Vancouver, New Westminster, and Victoria. The stores of interest are all in different cities, so it is unlikely that the layoffs will happen due to a certain factor in these cities. This answers the question in section 3.2.5, which asks if cities played a role in employee layoff.

7.2.3 Hire Age Distribution

```
# dataframe with only stores 11, 13, 20
certain_store <- store %>% filter(store_name %in% c(11, 13, 20))
```

Figure 82: R code to filter dataframe and store into new variable

Before continuing with the analysis, a new data frame that includes only the records from the stores of interest. This is so the data frame doesn't have to be filtered again for these analyses.

```
df3 %>%
  filter(STATUS == 'TERMINATED' & store_name %in% certain_store$store_name)
%>%
  ggplot(aes(hire_age)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~store_name, ncol = 1) +
  labs(
    title = 'Distribution of Hire Age of Certain Stores',
    x = 'Hire Age',
    y = 'Number of Employees'
  )
```

Figure 83: R code to plot histogram of hire age distribution for certain stores

Here, the data frame from question 4 is used because the hire age is needed in this analysis. The data frame is then filtered to contain only the records from the *certain_store* data frame that was newly created. After that the histogram was drawn.

Figure 84: Histogram of hire age distribution for certain stores

The distributions here do not look unusual. They roughly follow the distribution of the whole dataset, which can be found in section 4.2.1.

7.2.4 Gender Distribution

```
certain_store%>%
  ggplot(aes(gender_full)) +
  geom_bar() +
  facet_wrap(~store_name)
```

Figure 85: R code to plot bar chart of gender in certain stores

Here the code is used to draw bar charts of the gender count, grouped by the stores. If only the x or y axis is specified in the *aes* function, *geom_bar* doesn't need the option *stat = 'identity'*.

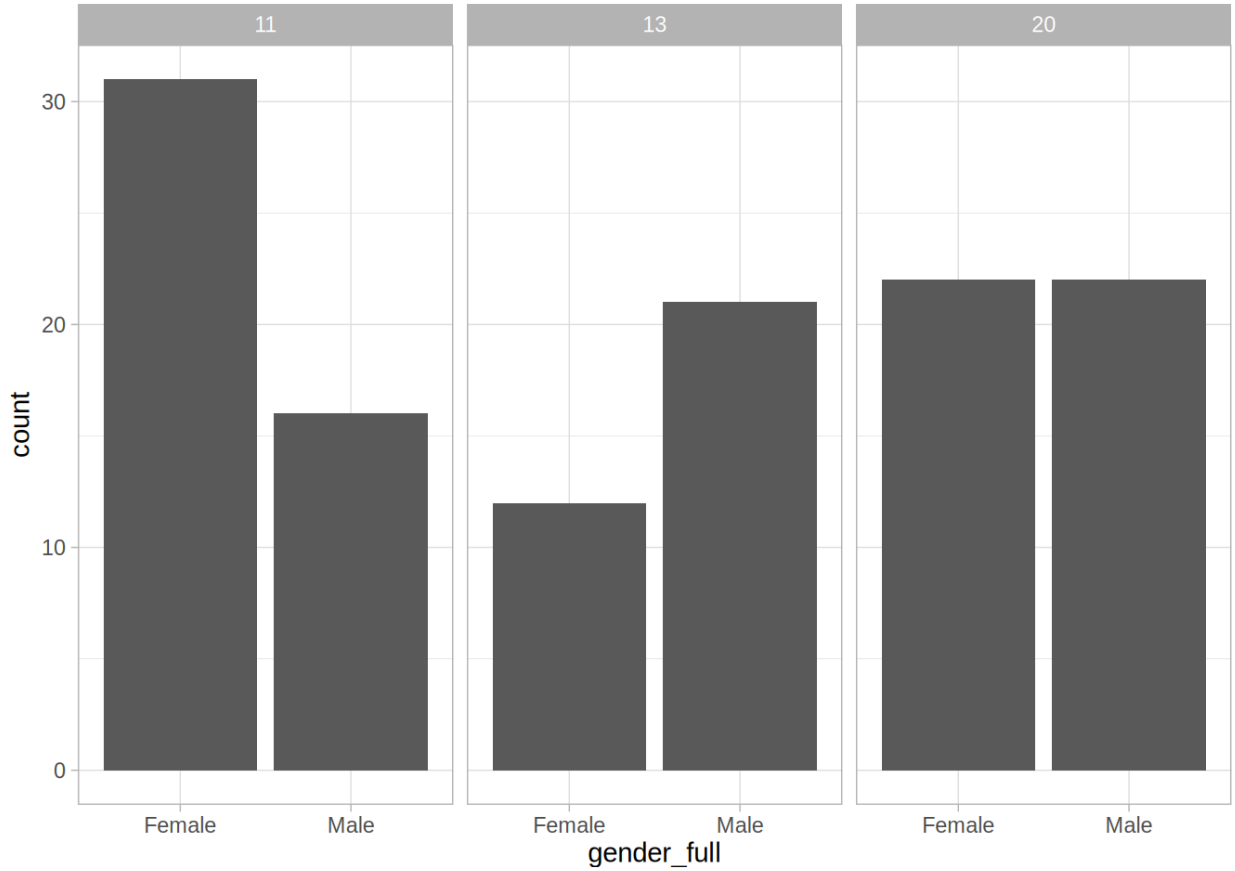


Figure 86: Bar chart of gender in certain stores

Again, here there are no noticeable patterns. The gender distribution in the stores seems to be random. There doesn't seem to be a common pattern among these stores. The information in the data provided is not enough to find insights into these stores.

7.3 Answer

Based on the analyses done in this section, it can be concluded that the factors of employee layoff cannot be determined from the dataset provided. There are several patterns found in the analyses, such as in section 4.2.2, which suggests older employees are more likely to get laid off, and in section 7.2.1, which suggests some stores are more likely to lay off their employees.

However, there are no supporting analyses that can verify those claims. Based on external sources, the factors that could cause employee layoff are mostly not inside the dataset provided (Indeed, 2021). The only exception to this is business closing, which by looking at the graph from section 7.2.1, being the only reason found in the analyses.

8.0 Extra Features

8.1 R Markdown

The entire analysis for this assignment is done using R notebooks. An R notebook is a document format that uses markdown for formatting and contains chunks of r code that can be independently executed (Yihui Xie, 2021). Markdown is a lightweight markup language that can be used to format documents using only plain text (Markdown Guide, n.d.).

R notebooks can be used inside RStudio. RStudio provides a graphical interface to insert new code chunks and to buttons to run them interactively. The output of the code chunks will be shown under the corresponding code chunk.

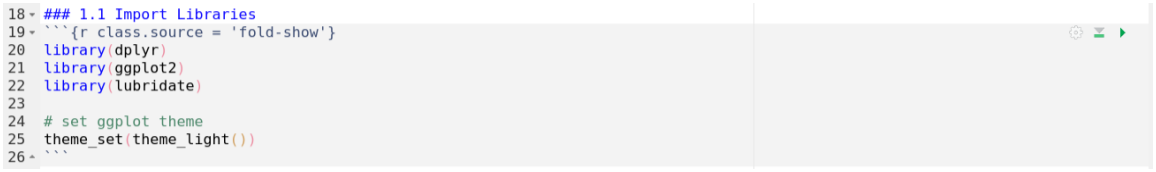


Figure 87: R code chunk in RStudio

Here is an image of a code chunk inside RStudio. The buttons on the top right of the chunk is used to run the code. R notebooks can also be exported into different document formats such as HTML, PDF, DOCX, etc. This is one of the advantages of using an R notebook.

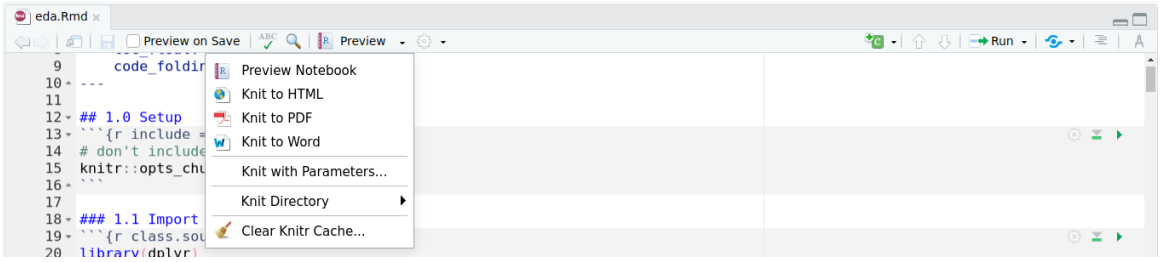


Figure 88: Preview button for rmarkdown in RStudio

The document can be exported using the preview button's dropdown menu. The output format is specified in the YAML header at the top of the R notebook. The file metadata and output settings are specified here. The document can also be rendered without using RStudio, by running a command directly in the R console or using an RScript. The command used looks like this: rmarkdown::render("<filename>", output_format = "all"), where filename is the name of the R notebook file, which has an .Rmd extension

8.2 Stacked Bar Charts

A stacked bar chart is a variation of the regular bar chart, with segments in each bar that represent a subcategory. It is used to compare categorical variables grouped with another variable. It is heavily used in this analysis.

```
df2 %>%
  filter(STATUS == 'TERMINATED') %>%
  count(gender_full, termreason_desc) %>%
  group_by(gender_full) %>%
  mutate(ng = sum(n),
         np = round(n / ng, 2) * 100) %>%
  ggplot(aes(x = n, y = gender_full,
             fill = termreason_desc,
             label = paste(np, '% (', n, ')', sep = ''))) +
  geom_bar(stat = 'identity', position = 'fill') +
  geom_text(position = position_fill(vjust = 0.5), size = 3) +
  labs(
    title = 'Ratio of Termination Reasons in Each Gender',
    x = 'Percentage',
    y = 'Gender',
    fill = 'Termination Reason'
  )
```

Figure 89: R code to plot stacked bar chart

This plot can be achieved by specifying the subgroup (the segments) using the *fill* parameter in the aesthetics (*aes*) inside the *ggplot* function. Then, the parameter *position* in the *geom_bar* function needs to be set to 'fill'. The labels for the percentage or count can be added using the *label* parameter and moved using the *position* parameter from the *geom_text* function.

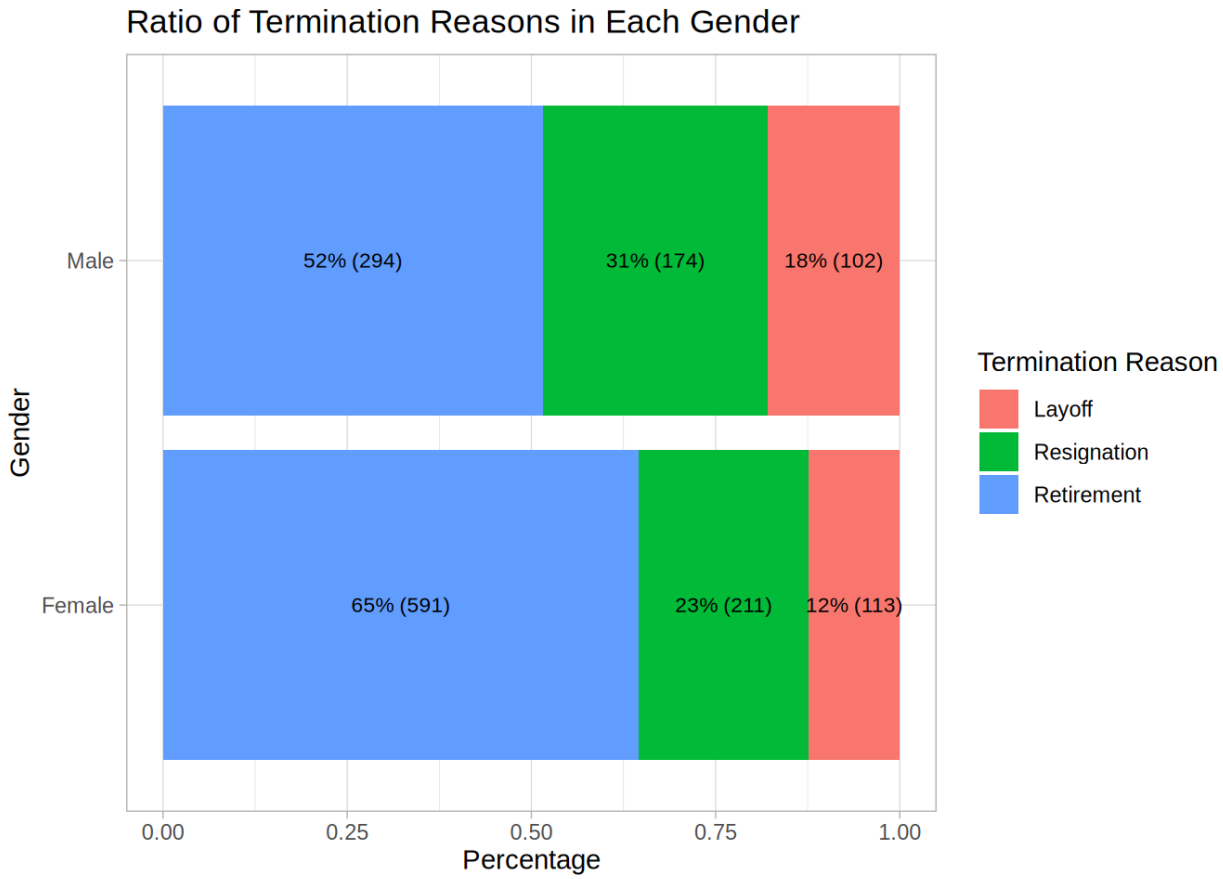


Figure 90: Stacked bar chart

Here is the result of executing the code. This is the same chart as the one used in section 6.2.1. The segments in the bars help to visualize the ratio of the subgroups of interest.

8.3 Violin Charts

Violin charts are a type of visualization that is used to view the distribution of a variable. It is often used to replace boxplots, as the shape of the violin chart is less confusing the box and lines in the boxplot. It can be drawn using the *geom_violin* function provided by the *ggplot2* package.

```
df3 %>%
  ggplot(aes(hire_age, gender_full, fill = gender_full)) +
  geom_violin() +
  labs(
    title = 'Distribution of Hire Age Grouped by Gender',
    x = 'Hire Age',
    y = 'Gender',
    fill = 'Termination Reason'
  )
```

Figure 91: R code to plot violin chart

The aesthetics needed for a violin plot is the x and y axis. In the example above, the x axis is a continuous variable, hire age, and y is a categorical variable, gender full. f*ill* is used here to give different colors to the different groups.

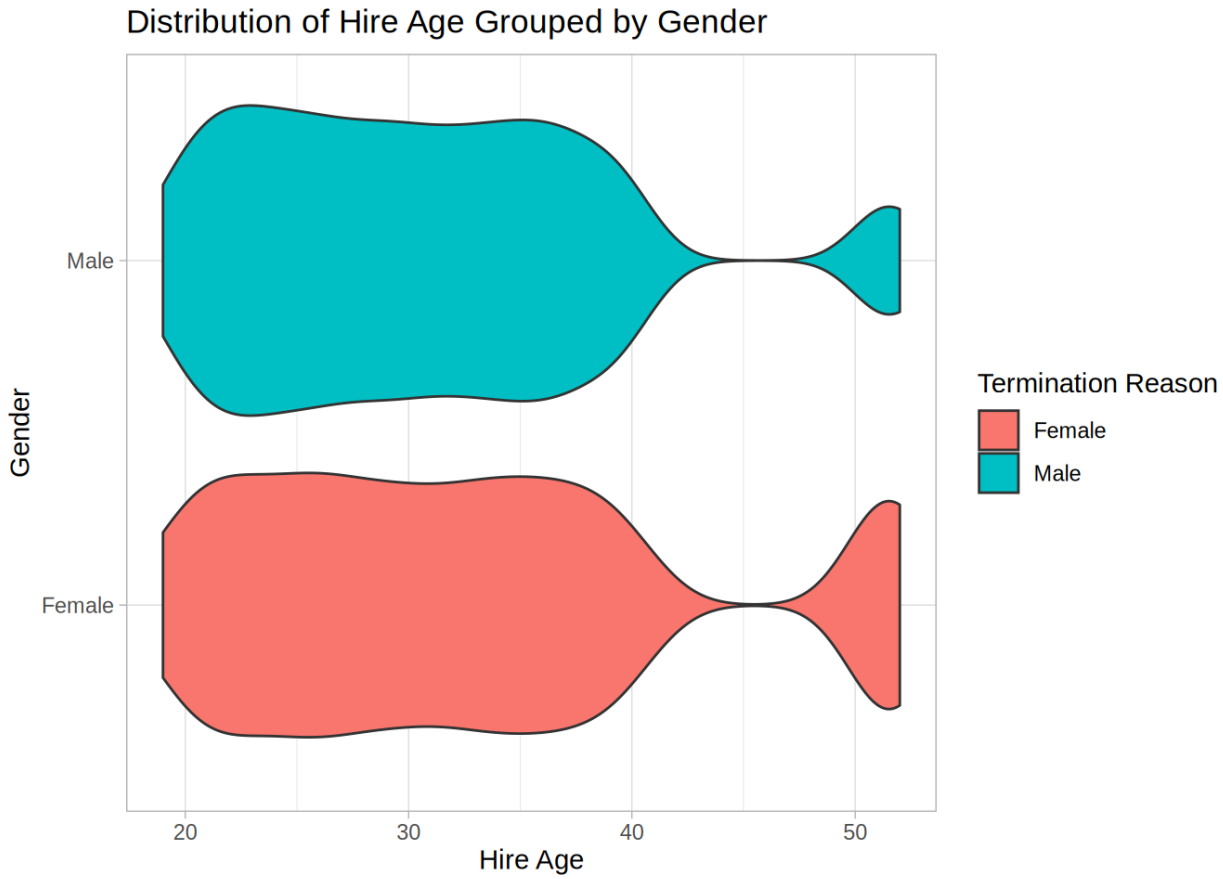


Figure 92: Violin chart

9.0 Conclusion

There are several reasons as to why employees would leave their jobs. Employee resignation can be caused by several factors, including age, job & department, and gender. In short, younger employees are more likely to resign, with more males resigning than females. Some jobs are also responsible for this, such as cashiers, which have a high resignation rate. The cause of this lies in the individual jobs themselves.

Unlike resignation, employee layoffs are harder to explain. Some stores have a very high number of layoffs, while some don't. Not enough information is available in the dataset provided to explain this. Factors such as salary, marital status, or job satisfaction could have an impact on employee attrition, but they are not included in the dataset. External resources must be used to take those factors into account.

To conclude, there can be several factors that lead to employee termination, and these factors can be used by the human resources department to make smarter decisions going forward to reduce the number of employees quitting.

10.0 References

Lewis, A. (2015, November 5). *Why women are leaving their jobs (your first guess is wrong)*. LinkedIn. Retrieved November 22, 2021, from https://www.linkedin.com/business/talent/blog/talent-strategy/why-women-are-leaving-their-jobs.

LivingCost.org. (2021, September 19). *Cost of living in British Columbia: 22 cities compared*. Livingcost.org. Retrieved November 22, 2021, from https://livingcost.org/cost/canada/bc.

Markdown Guide. (n.d.). *Getting started*. Markdown Guide. Retrieved November 22, 2021, from https://www.markdownguide.org/getting-started.

RStudio. (n.d.). *Tidyverse packages*. Tidyverse. Retrieved November 22, 2021, from https://www.tidyverse.org/packages/.

StaffCircle Ltd. (2019, May 20). *Reasons why young workers leave their jobs and look for more stable work*. Medium. Retrieved November 22, 2021, from https://medium.com/@staffcircle/reasons-why-young-workers-leave-their-jobs-and-look-for-more-stable-work-ad35e1febc01.

Statistics Canada, G. of C. (2021, January 25). *Retirement age by class of worker, Annual*. Retirement age by class of worker, annual. Retrieved November 22, 2021, from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410006001.

Talent.com. (n.d.). *Cashier salary in Canada - average salary*. Talent.com. Retrieved November 22, 2021, from https://ca.talent.com/salary?job=cashier.

Talent.com. (n.d.). *Shelf Stocker salary in Canada - average salary*. Talent.com. Retrieved November 22, 2021, from https://ca.talent.com/salary?job=shelf%2Bstocker.

University of Waterloo. (2020, April 9). *Four reasons you might be seeing young employees leave jobs (and tips for retaining them)*. HIRE Waterloo. Retrieved November 22, 2021, from https://uwaterloo.ca/hire/four-reasons-young-employees-leave-jobs.

Yihui Xie, J. J. A. (2021, April 9). *R markdown: The definitive guide*. 3.1 HTML document. Retrieved November 22, 2021, from https://bookdown.org/yihui/rmarkdown/html-document.html.