Logo, company name

Description automatically generated

INDIVIDUAL ASSIGMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APU2F2109CS(DA)

HAND OUT DATE : 13 OCTOBER 2021

HAND IN DATE : 22 NOVEMBER 2021

WEIGHTAGE : 50%

NAME : ANDI RACHEL AZRA AQILAH

TP NUMBER : TP061443

Table of Contents

[1. Introduction and Assumptions 3](#_Toc88488426)

[1.1. Assumptions 3](#_Toc88488427)

[2. Data import / Cleaning / pre-processing / transformation 3](#_Toc88488428)

[3. Why the employee wants to resign? 5](#_Toc88488429)

[3.1. Analysis 1 – How many People that resign based on age 5](#_Toc88488430)

[3.2. Analysis 2 – How many employees that resign based on year 7](#_Toc88488431)

[3.3. Analysis 3 – How many employees that resign based on their job and department 9](#_Toc88488432)

[4. Why there are a lot of layoff employee in the organization 12](#_Toc88488433)

[4.1. Analysis 1 – How many layoffs employee in 2014 and 2015 based on their job 12](#_Toc88488434)

[4.2. Analysis 2 – How many layoffs employee based on their age 12](#_Toc88488435)

[4.3. Analysis 3 – Which store that has most layoff employee 14](#_Toc88488436)

[5. Is the terminated employee amount covered by the new employee 15](#_Toc88488437)

[5.1. Analysis 1 – How many new employees based on year 15](#_Toc88488438)

[5.2. Analysis 2 – How many terminated employees based on year 16](#_Toc88488439)

[5.3. Analysis 3 – How many new employees that choose to be cashier 17](#_Toc88488440)

[6. Why some stores are closed 19](#_Toc88488441)

[6.1. Analysis 1 – How many stores that still active 19](#_Toc88488442)

[6.2. Analysis 2 – Which store has most terminated employee 19](#_Toc88488443)

[6.3. Analysis 3 – How many terminated employee by store in 2014 20](#_Toc88488444)

[6.4. Analysis 4 – What happen to stores that close in 2015 21](#_Toc88488445)

[7. Is the missing department effect the organization 23](#_Toc88488446)

[7.1. Analysis 1 – How many jobs are missing 23](#_Toc88488447)

[7.2. Analysis 2 – What happen to the director in certain departments 23](#_Toc88488448)

[7.3. Analysis 3 – Recruiter and Trainer age distribution 24](#_Toc88488449)

[7.4. Analysis 4 – Terminated Analyst by year 25](#_Toc88488450)

[8. Extra feature 27](#_Toc88488451)

[8.1. Ggplotly 27](#_Toc88488452)

[8.2. Lollipop plot 28](#_Toc88488453)

[8.3. Stringr 29](#_Toc88488454)

[8.4. Tile plot 30](#_Toc88488455)

[8.5. Violin 31](#_Toc88488456)

[9. Conclusion 32](#_Toc88488457)

[10. References 32](#_Toc88488458)

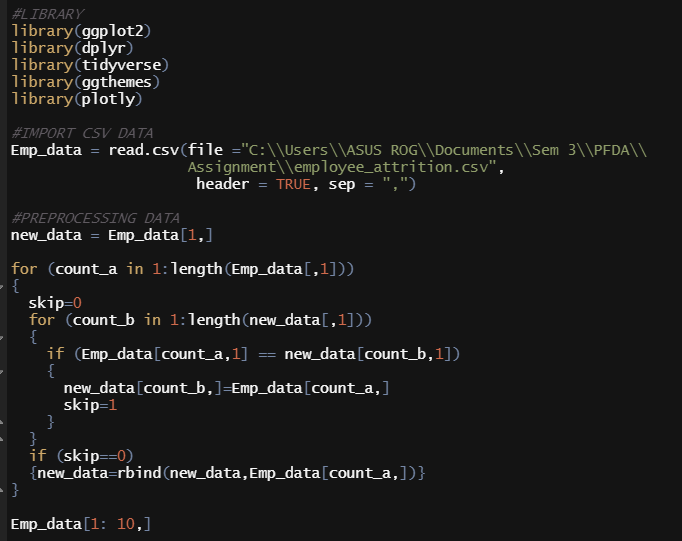
# Introduction and Assumptions

The assignment demands us to investigate a dataset including data from employees at a company in order to uncover a hidden problem in management of human resources. The dataset consists of 18 columns and 49654 rows that includes the personal detail of the staff, job department, position, location, working status, and reason of termination. The goal of this research is to analyse a given dataset by using RStudio in order to expose unseen issues in the organization and provide useful information for decision-making.

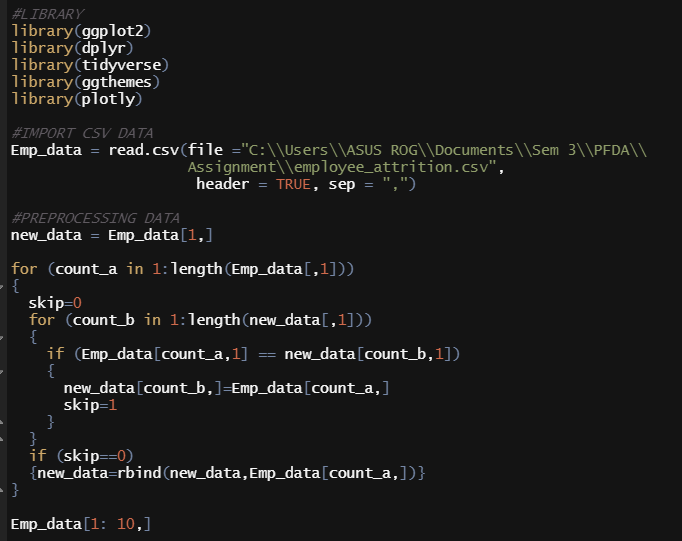
* 1. Assumptions

The assumptions that we have made for this assignment is the dataset provided is not accurate and valid due to external factors.

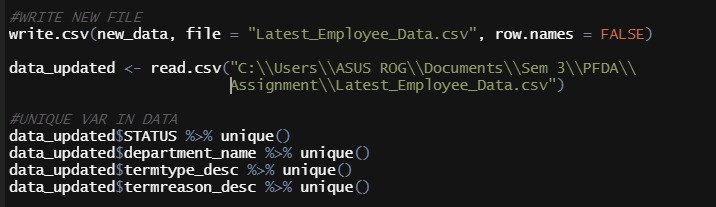
# Data import / Cleaning / pre-processing / transformation



The first thing to do is to import the excel file with format of csv to RStudio. We can use read.csv function that read the csv file and separate the data from comma. The header true is to force RStudio to accept the header as parameter.



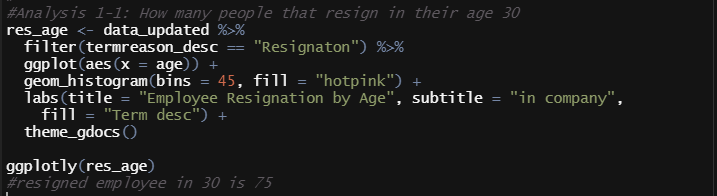
To preprocess the data, we can use nested for loop. In the syntax, new\_data have the first column of the Emp\_data. The outer loop assign count\_a = 1 to the length of the Emp\_data first row, while the inner loop assign count\_b = 1 to the length of the Emp\_data first column. For inner loop if the Emp\_data from row count\_a and column 1 equals new\_data from row count\_b and column 1 then skip = 1. The inner loop will continue until skip = 0, then the data new\_data and Emp\_data with count\_a will bind by row. The result of the looping process is new\_data.



After the new\_data is completely preprocessed, to use new data we can write a new csv file by using write.csv function. When the csv file is available, assign a new data to read the csv file.

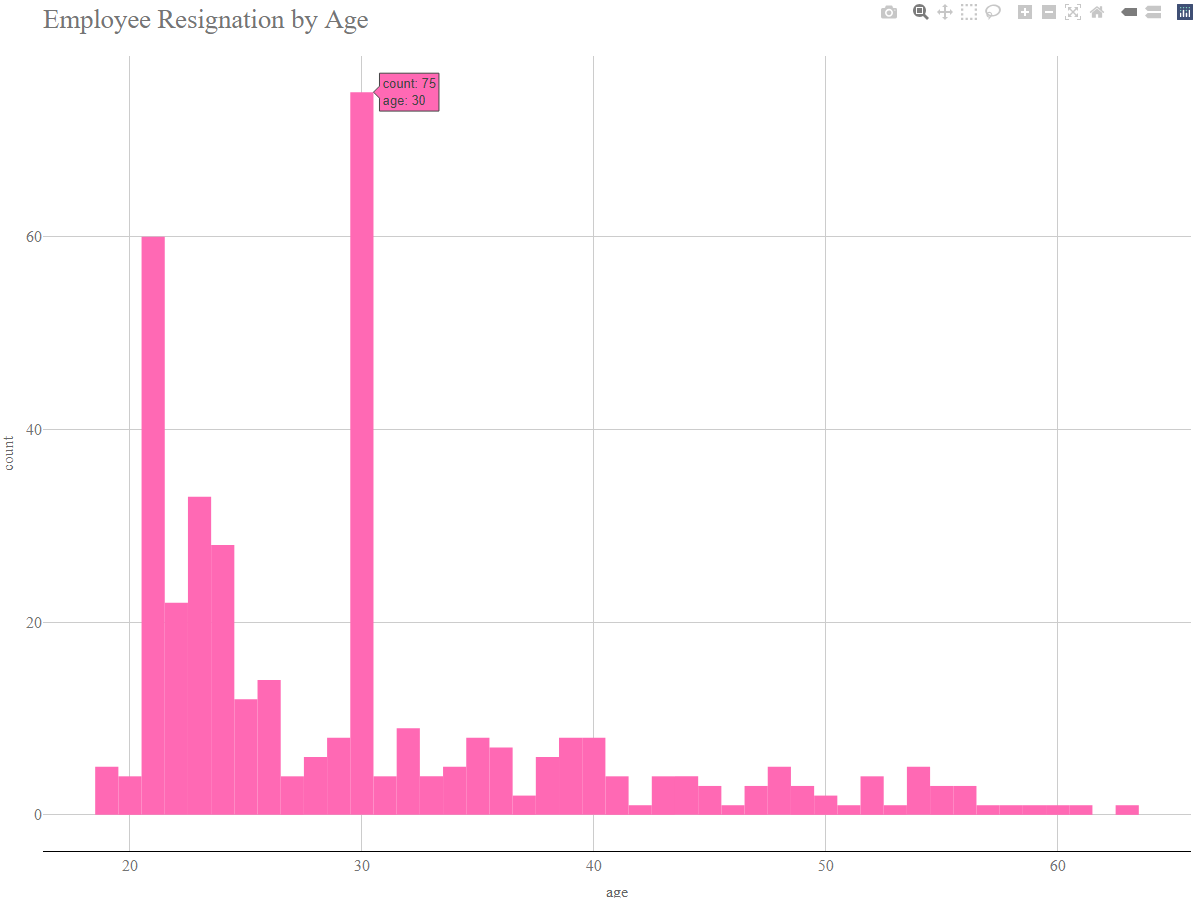
# Why the employee wants to resign?

1. Analysis 1 – How many People that resign based on age

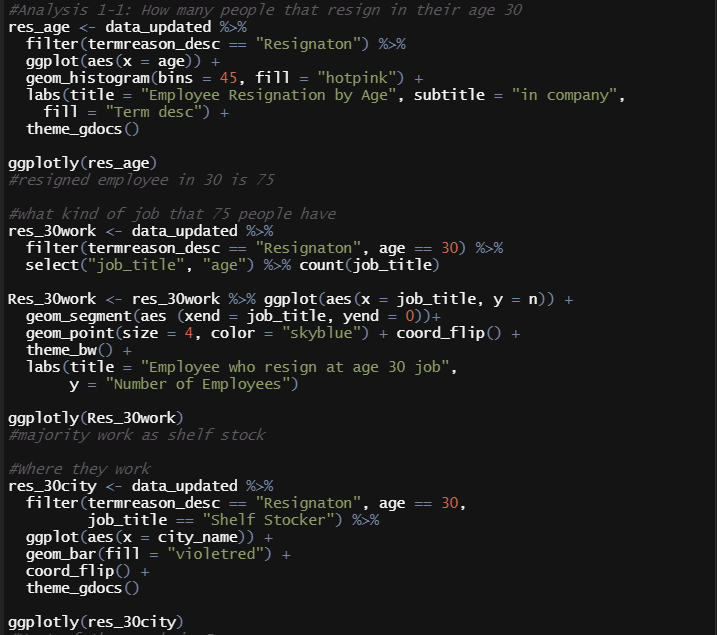




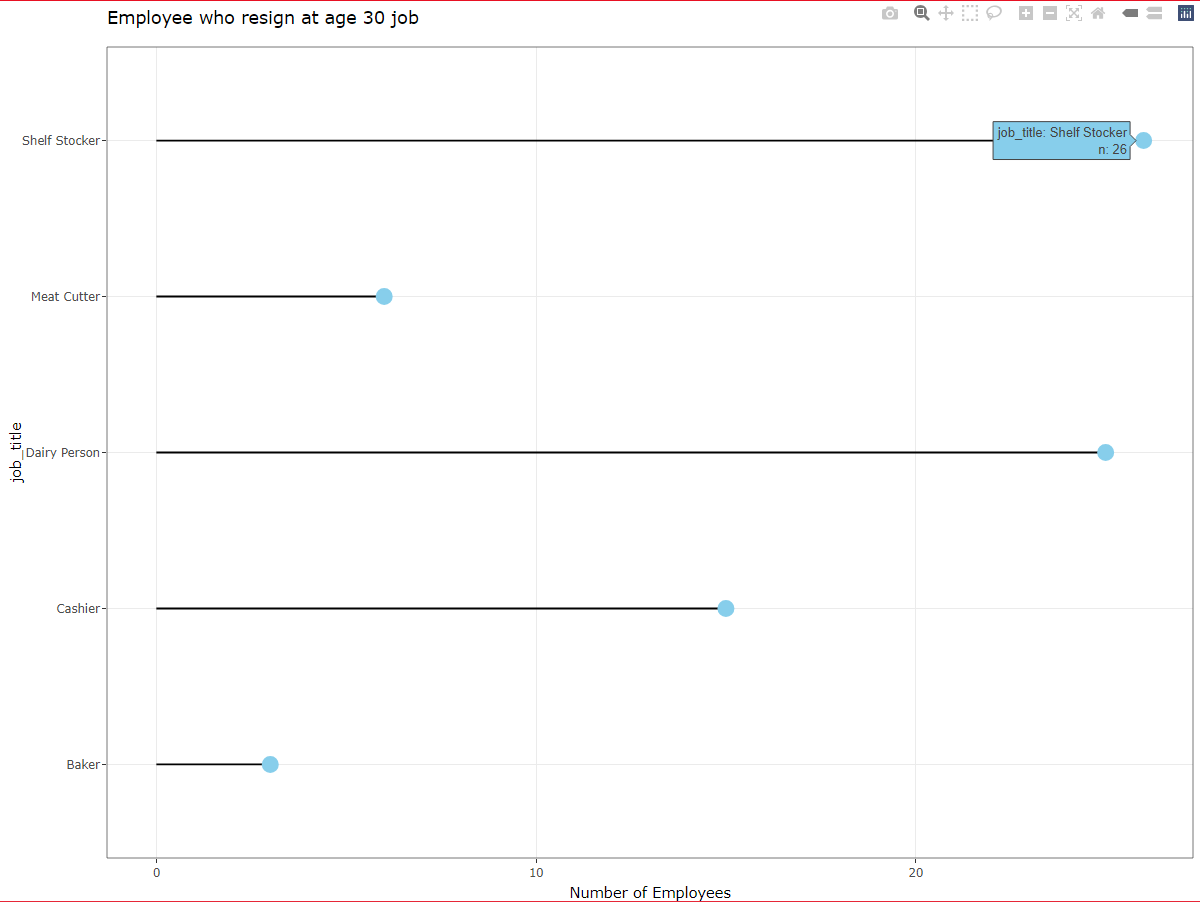
In Analysis 1, we search for people who resign at age 30. This analysis creates histogram of employee that resign by age. To create the histogram, we need to use data that has filtered by the term reason of resignation along with ggplot. We also use ggplotly to show the frequency of the certain bar.



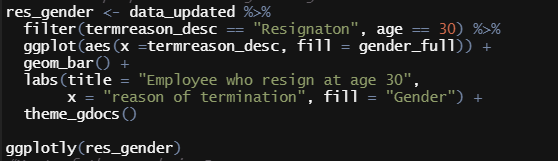
The result shows that the most employee that resign is by age 30, along with 21 years old. It is assumed that the employee that has resign by age 21 is doing a part time job for work experience, or for the salary. However, for employee that resign at age 30 is very suspicious. Thus, we can do more depth analysis about employee at age 30.



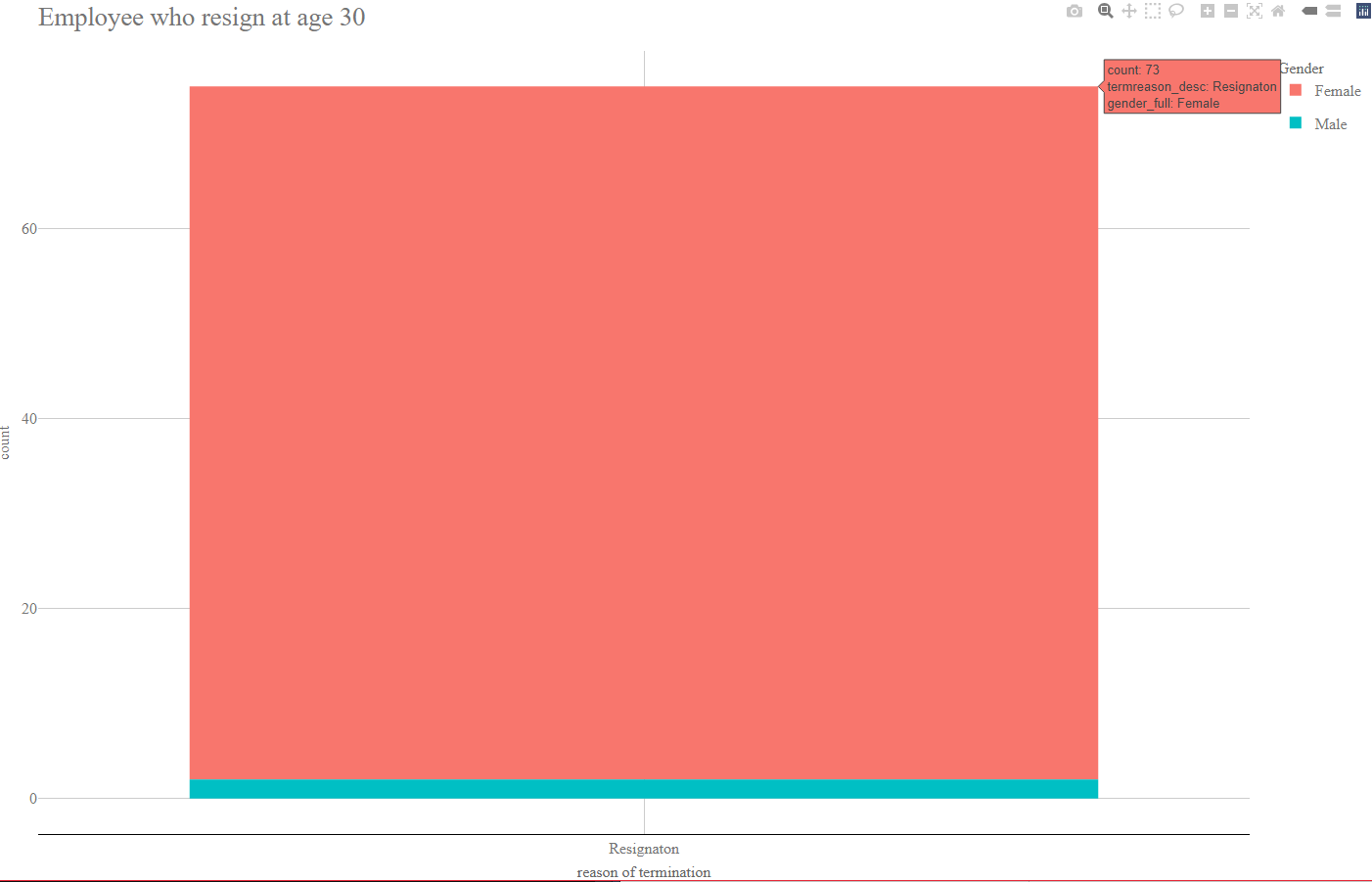
This code is to explore what kind of job that 30 years old employee has. We can get the results by create a data to filter the termination reason into resignation and the age is equal to 30. Then we select the job title and age for later to be count by job title. To visualize the data, basically we can use the data that has been processed which is res\_30work then use ggplot. To make the data easier to read, we can use lollipop plot and ggplotly to display the amount of employee in each job.



In this plot, it shows that majority of resign employee has job in Shelf Stocker and Dairy person. It is known that both of the job is physically strenuous (glassdoor, n.d.). However, we still not sure yet the reason behind the termination, therefore we analyze again the data.

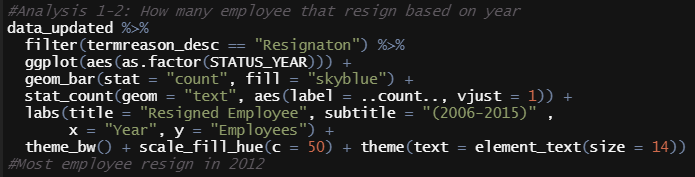


The code above is to find what gender does resigned employee has. By filter the termination reason into resignation and age into 30, then use ggplot with bar plot. We also add ggthemes and ggplotly for better visualizations.

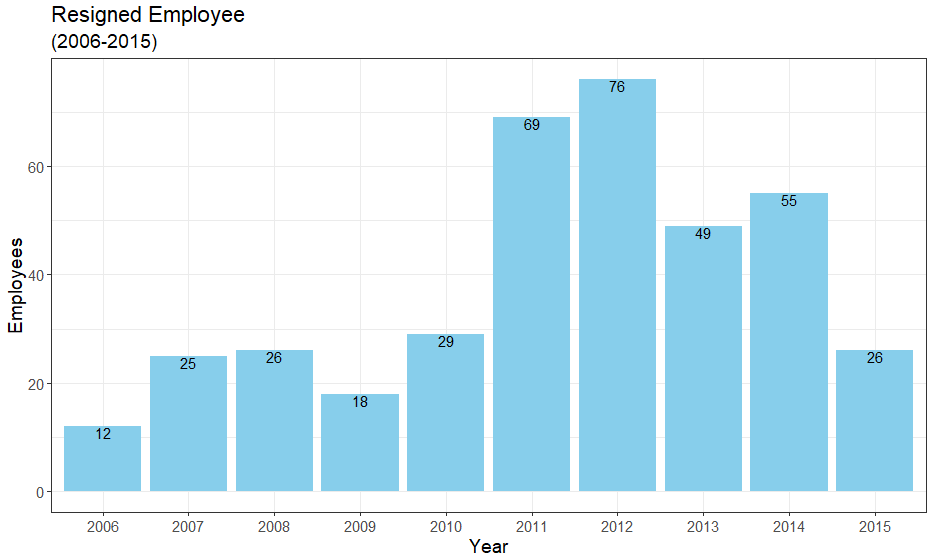


The bar plot shows us that the majority of resigned employee is female. It is interesting since female in their 30s usually is very dedicated to their work. Based on my own research, in 2015 female workforce has increase significantly. However, Governments and employers have not sufficiently adjusted to this changing landscape, which places women at a disadvantage economically (Howard, 2021).

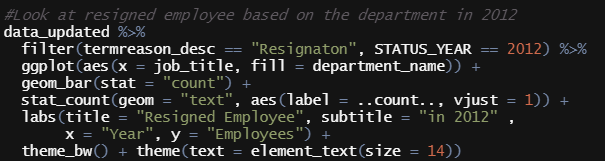
1. Analysis 2 – How many employees that resign based on year



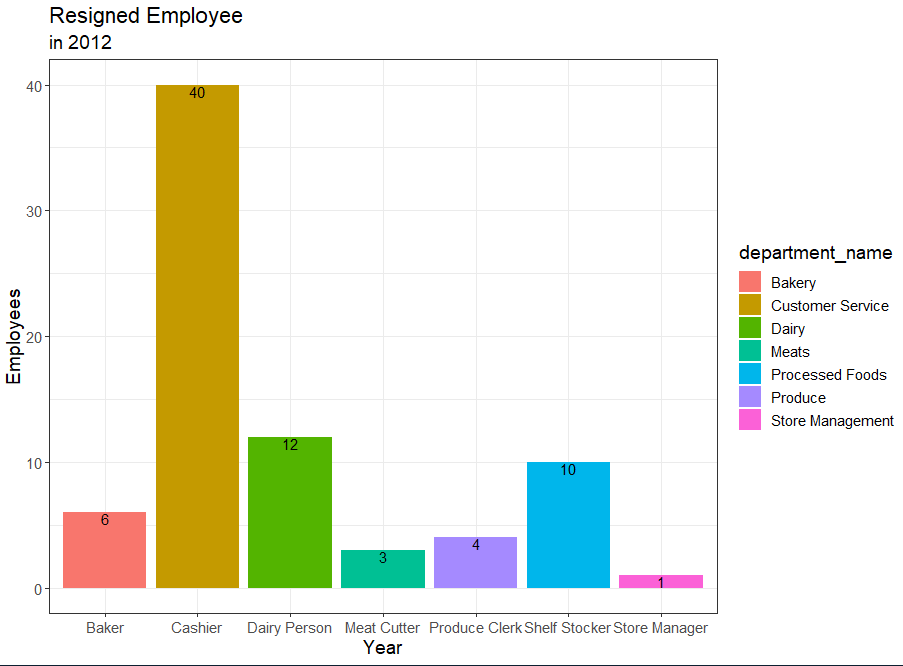
In the second analysis we want to know at what year the employees want to resign. This code will create a basic bar plot with the frequency in each year by using ggplot and stat count. We can also add labs to generate title and subtitle for a better visualization, and fill the bar with colors.



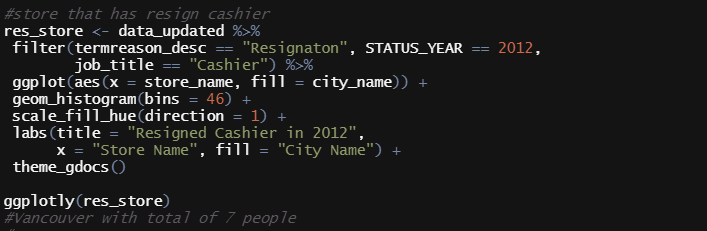
The bar plot above shows us that 2012 has the most resigned employee. There are no problems in 2012 rather than political issues. This time we will do more further analysis again.



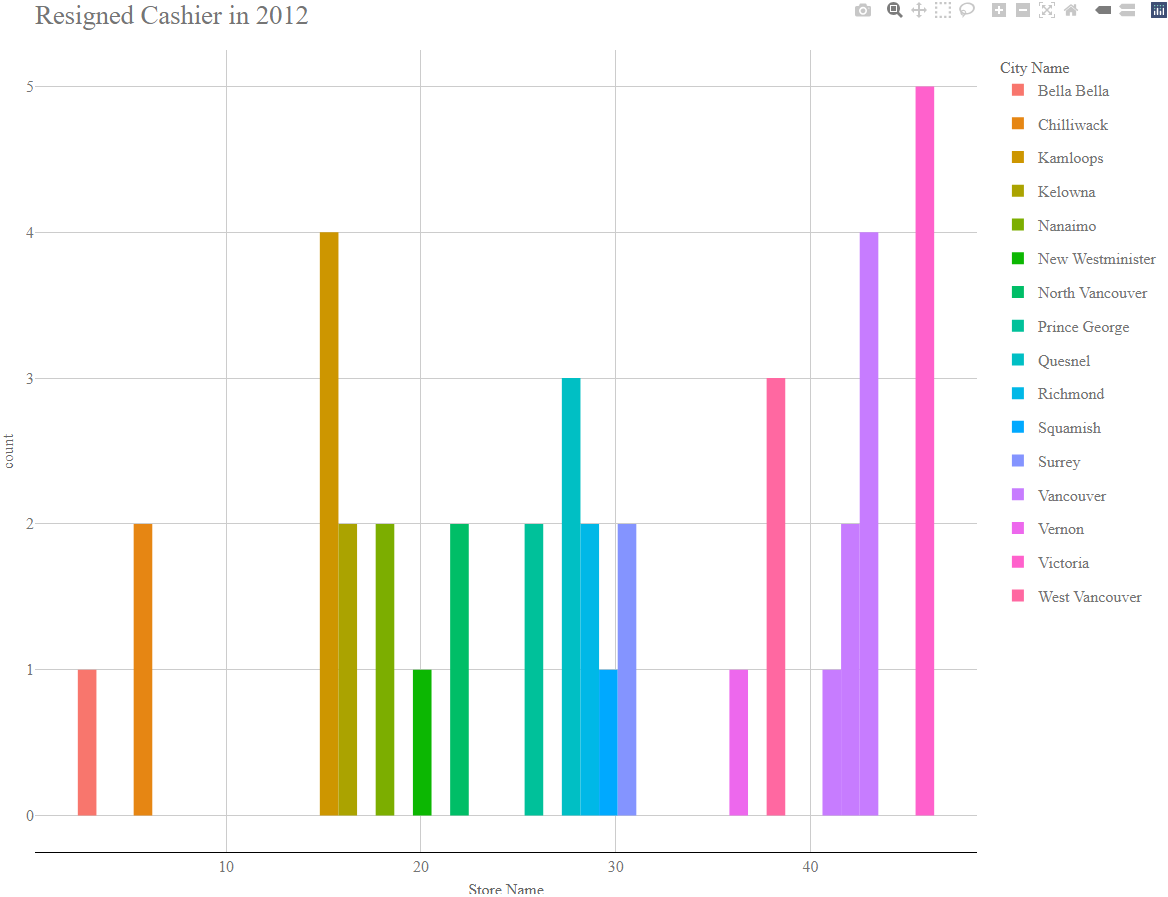
The code above generates a basic bar plot with the frequency in job by using ggplot and stat count after filtering the termination reason into resignation and status year to 2012, and also fill the bar based on department name. Then we can add labs to generate title, x and y title, and subtitles along with black and white themes for an improved visualization.



The visualizations of the bar plot illustrate that cashier have the most resigned employee in 2012 with the frequency of 40. Generally, employees who work at the cashier are usually part-time workers can quit at any time. To have a better analysis, we need to search where they work.

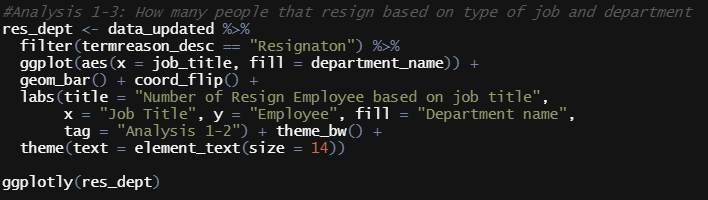


This code allows us to look how may cashier that has resign by the store name by using histogram. The code is almost the same as the code before, but we add the cashier in job title in the filter. We add the bins to 46 to look the frequency each store since the amount of store is 46.

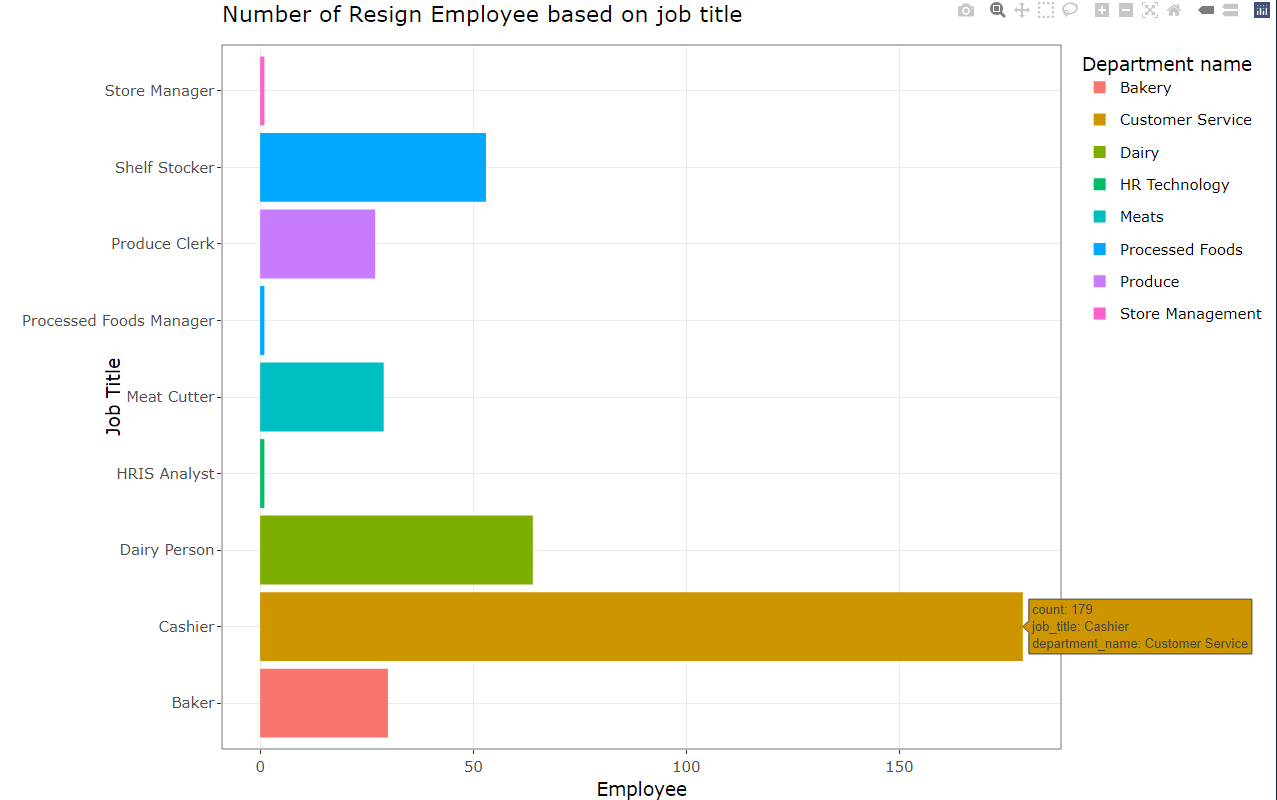


The histogram displays the frequency of resigned employee by store name. The results shown that the stores located in Vancouver is the most resigned employee, and the head office is there. It is concluded that in 2012, Vancouver has a lot of cashiers that resigned because of 2 assumptions, first because they have internal issues, and second is they just want a better job. This assumption is based on the length service in the employee is quite high, around 4-10 years.

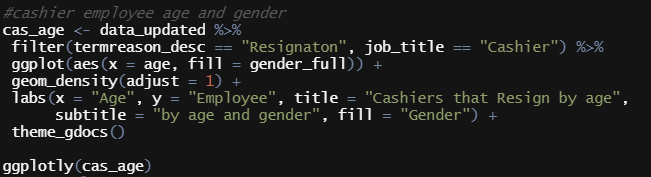
1. Analysis 3 – How many employees that resign based on their job and department



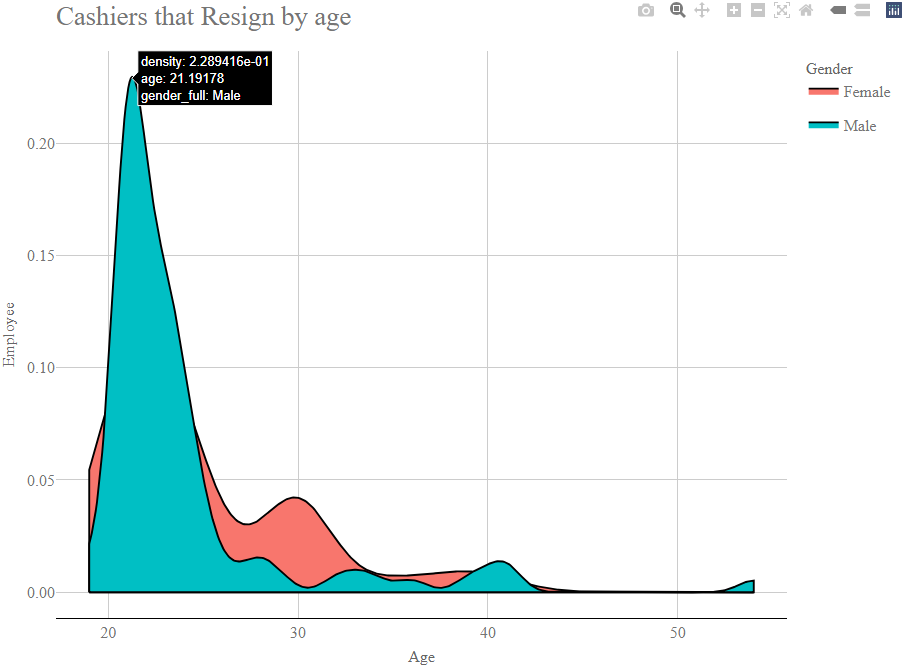
In this analysis, we look for amount of resigned employee based on their job title and department. After filter the data with the termination reason of resignation, use ggplot with the aesthetic x on job title, then fill it with the department name. Also add visualizations for bar plot and flip the bar for a better look. We can also add labs to rename the legend or add titles along with subtitles.



This bar plot confirms that cashier has the most resigned employee with the amount of 179. This data alone cannot substantiate our conclusion. Therefore, we can add another sub analysis.



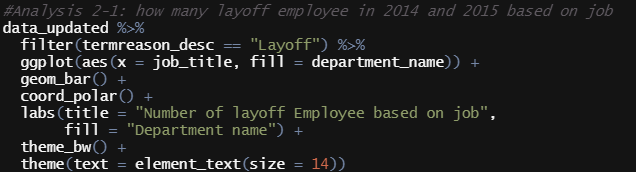
In this sub analysis, we want to know at what age these cashiers are resign by filter the termination reason and job title, then use ggplot to identify the x axis and lastly use density to visualize.



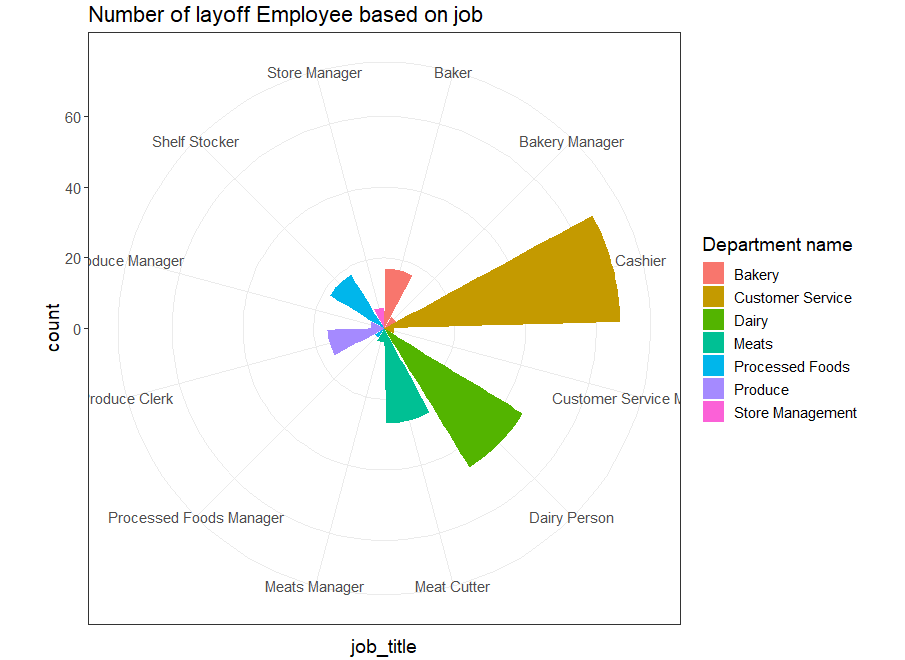
The visualization tells us that employee that work as cashier mostly resign in their 20s, and the majority is male. Base on this analysis, we know that employees at age 20 to 29 only work as cashier for work experience or for the salary. This is supported by the data. Resigned employees that work as cashier only has 0-2 length of service.

# Why there are a lot of layoff employee in the organization

1. Analysis 1 – How many layoffs employee in 2014 and 2015 based on their job

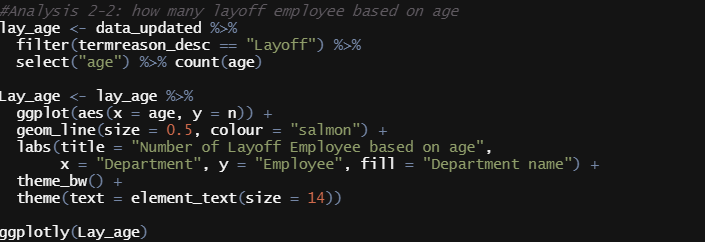


This analysis creates a circular barplot to look the amount of layoff employees based on their job by using ggplot and geom\_bar along with coord\_polar. It is also specified in 2014 to 2015 since the layoffs are only happen there.

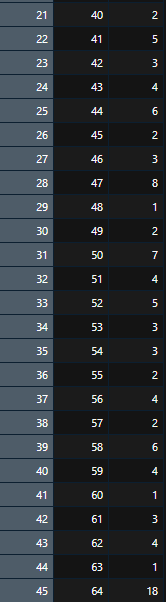
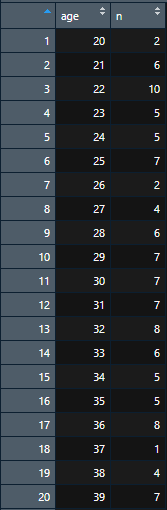


Based on the results, cashier is the most affected job in the layoff. It is possible that the cashier layoffs are based on This data raises many questions about the layoff which will be continued in the next analysis.

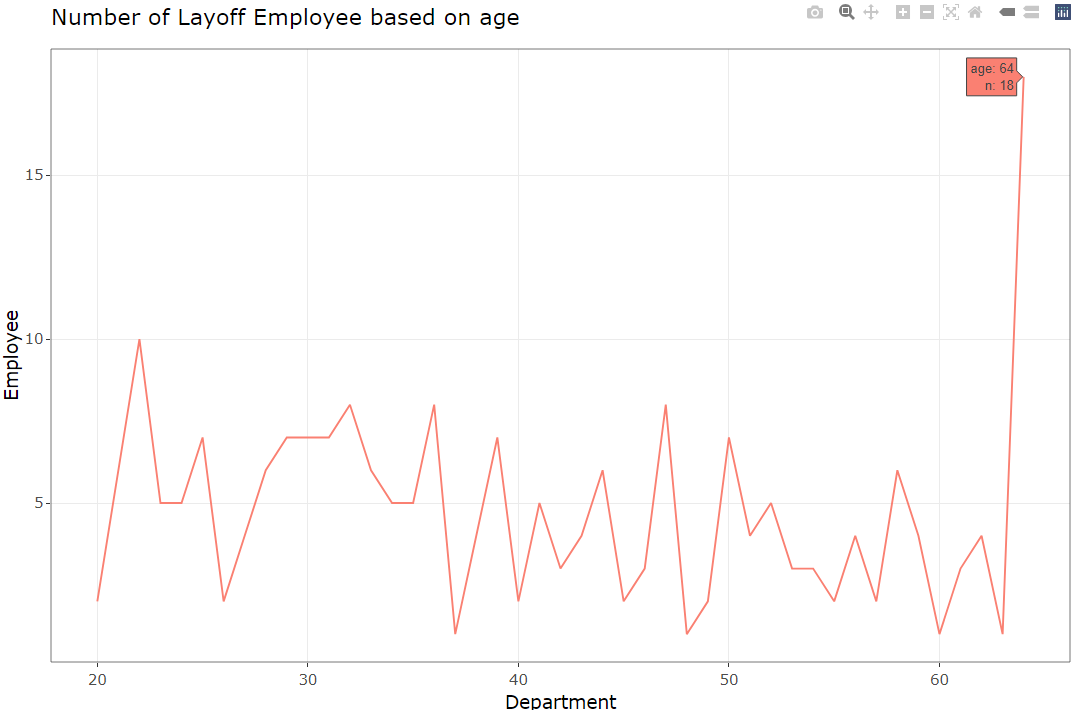
1. Analysis 2 – How many layoffs employee based on their age



The next analysis is to search how many layoffs employee based on age by creating a specific data to count the age distribution. To visualize the data, we can use line plot with ggplot and add title by using labs.

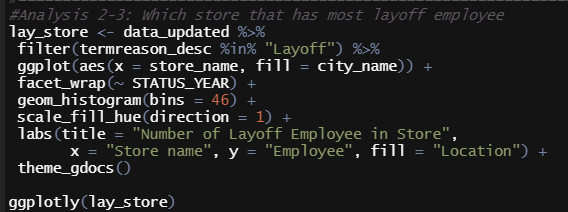


The table above is the results of the data that has been specify and counted by age. The employee with the most layoffs is at age 64. Thus, the visualization should be based on the table.

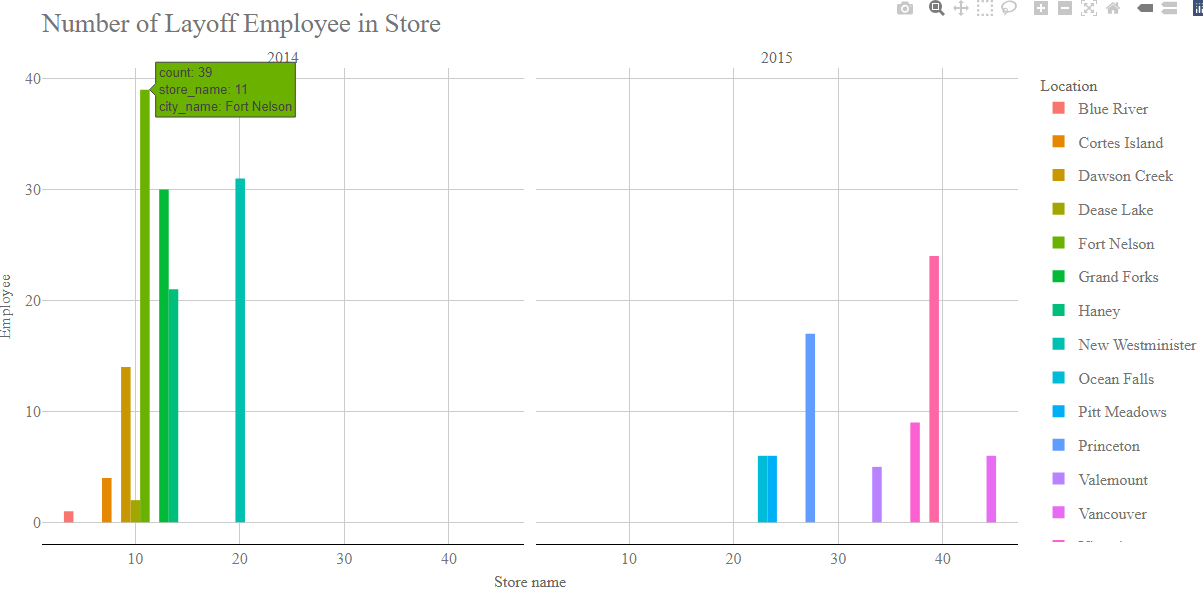


The line plot concludes that age 64 has the most layoffs employee. Thus, it is possible there are some issues in the organization that makes the employee with the age of 64 need to lay off because their length of service is high.

1. Analysis 3 – Which store that has most layoff employee



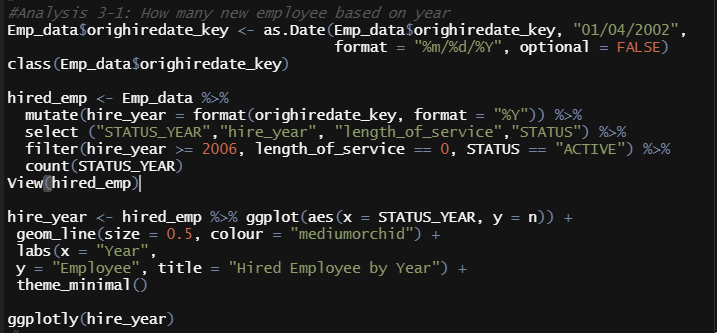
The third analysis is about the store that affected by the layoffs. It does by using filter then ggplot to create histogram that grouped by year. We also use scale fill by hue to color the histogram since the store name data is continuous.



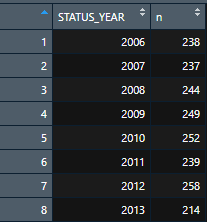
By the data visualization above, the massive layoff is in 2014 with the store name 11 and the location is in Fort Nelson. Based on our own research, the economy in 2014 is pretty weak and one of the factors is due to the snow (CBC News, 2014). It is also possible that most of the store that located far in the north is closed. Thus, employee that works in these stores are getting layoffs.

# Is the terminated employee amount covered by the new employee

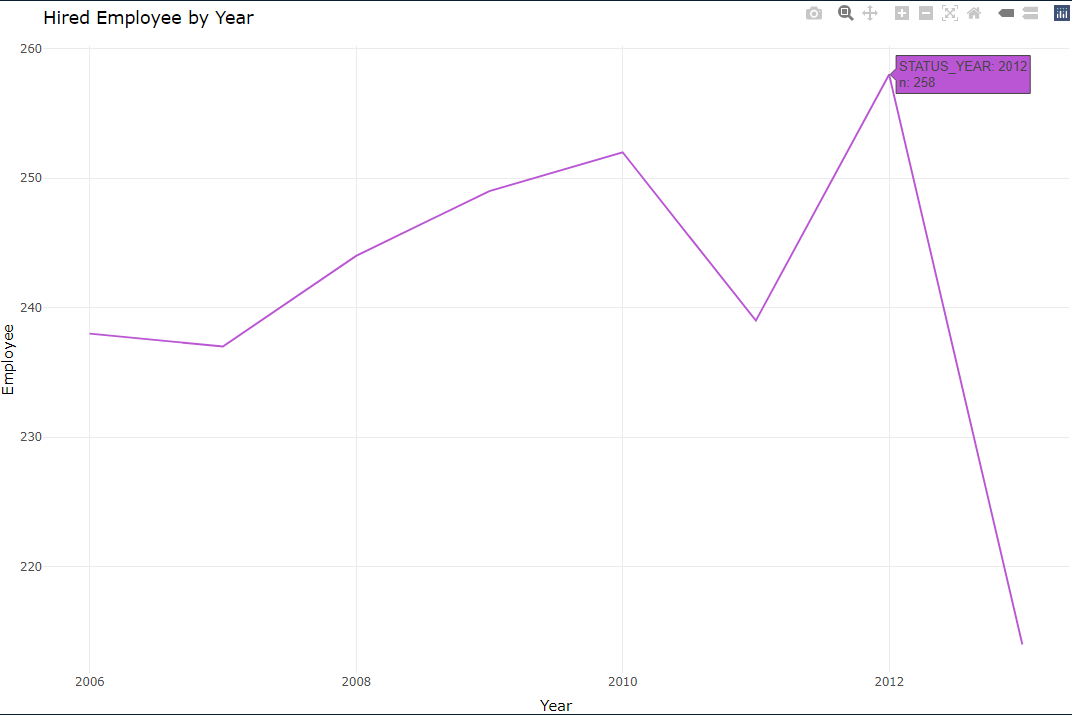
1. Analysis 1 – How many new employees based on year



The analysis will preprocess the data to find the hired employee and filtered the important parts, then visualized by using line plot in ggplot and customized the line plot with labs and ggplotly.

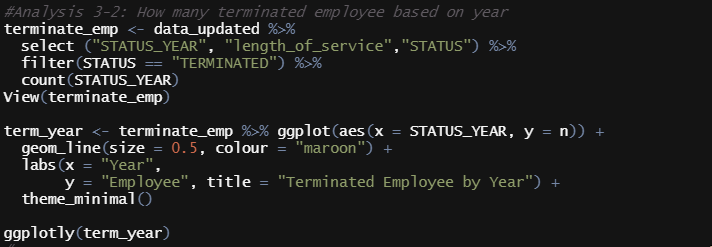


This is the result of the preprocessing part by counting the employee who hired based on the year. This data later can be compared to answer the questions.

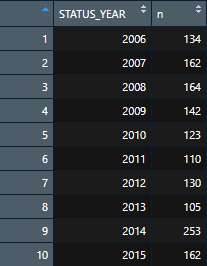


The line plot above shows that the organization only hired people until 2013. The highest number of employee acceptance was in 2012. The basic assumptions is maybe the organization just doesn’t have reasons to recruit new employee or something is wrong in the department related to recruiting.

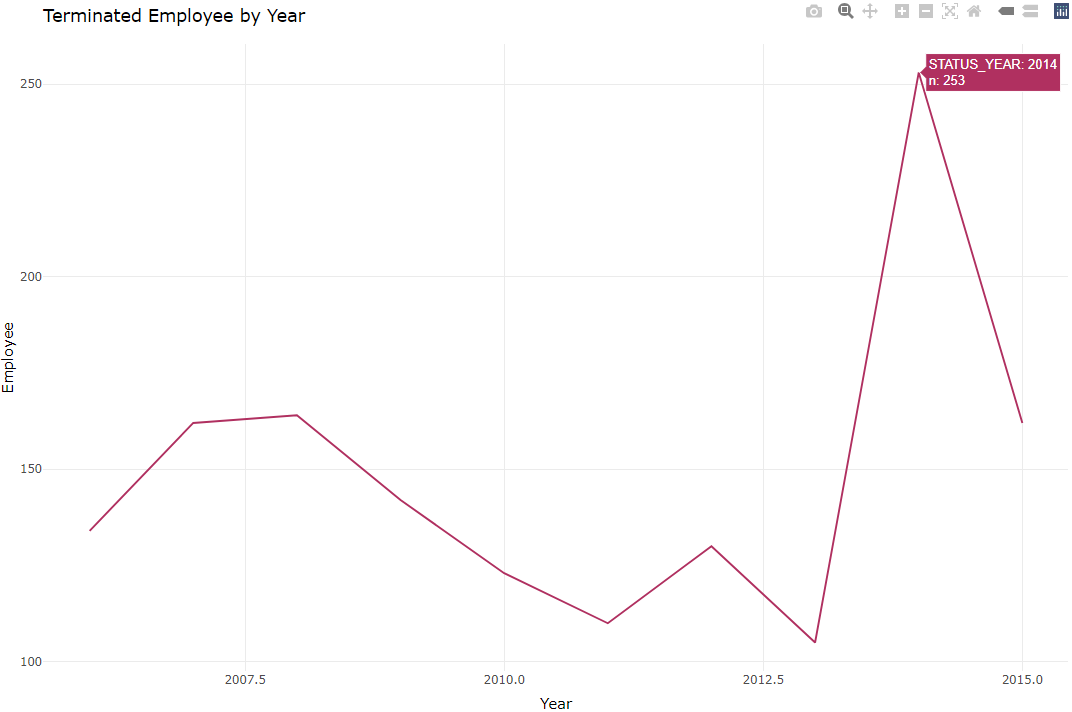
1. Analysis 2 – How many terminated employees based on year



The next analysis is to know how many terminated employees by creating a new data to be visualized later. To visualize the data, we can use ggplot then tidy it up by using labs, ggthemes and ggplotly.

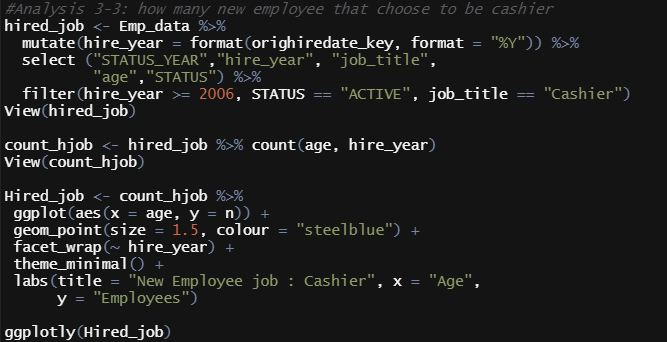


This is the data that has been create. It contains the year that the employee has terminated along with the frequency.

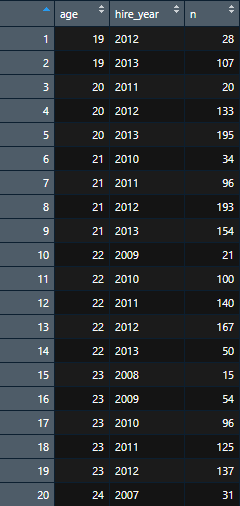


In this line plot, it is shown that employee terminated mostly at 2014. Based on the analysis 2, it is resolved that most of the terminated employee are covered by the new employee. But, if the most terminated employee is in 2014 it means that the employee termination above year 2014 is not covered, resulting the collapse of the organization.

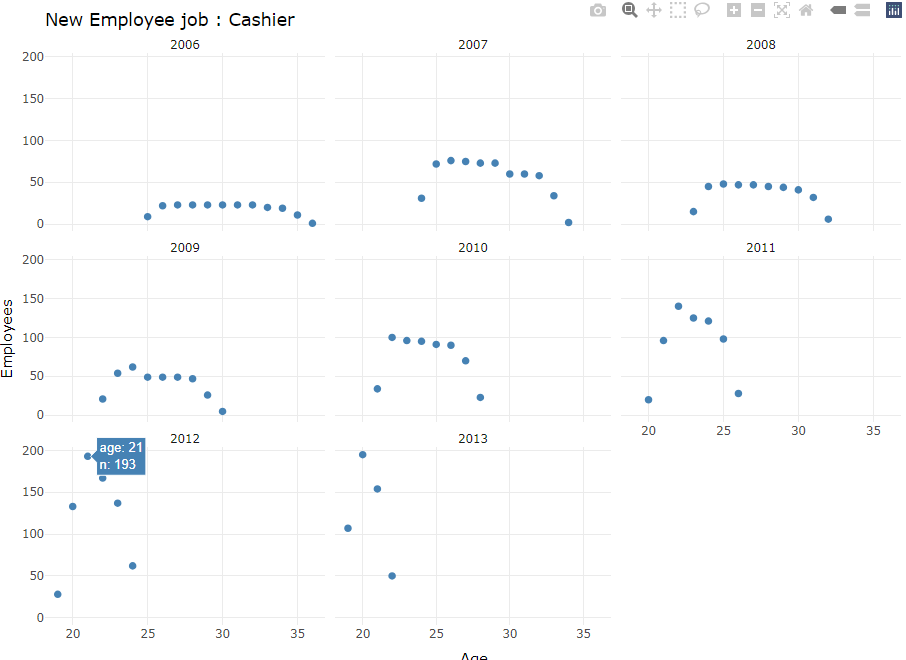
1. Analysis 3 – How many new employees that choose to be cashier



Based on the previous analysis in another question, we got a bit curious about cashier. So, in this analysis we want to know how many new employees that choose to be cashier. The first step is to create a new variable for the hired employee just like analysis 1. Then we count the frequency of the new employee by age and year that the employee hired. For visualization we use ggplot and geom point along with ggthemes, labs, and ggplotly.

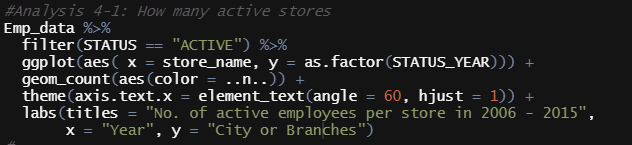
The left side is the table created after we assign a new variable while the right one is table that has generated by counting the frequency of the hired employee.



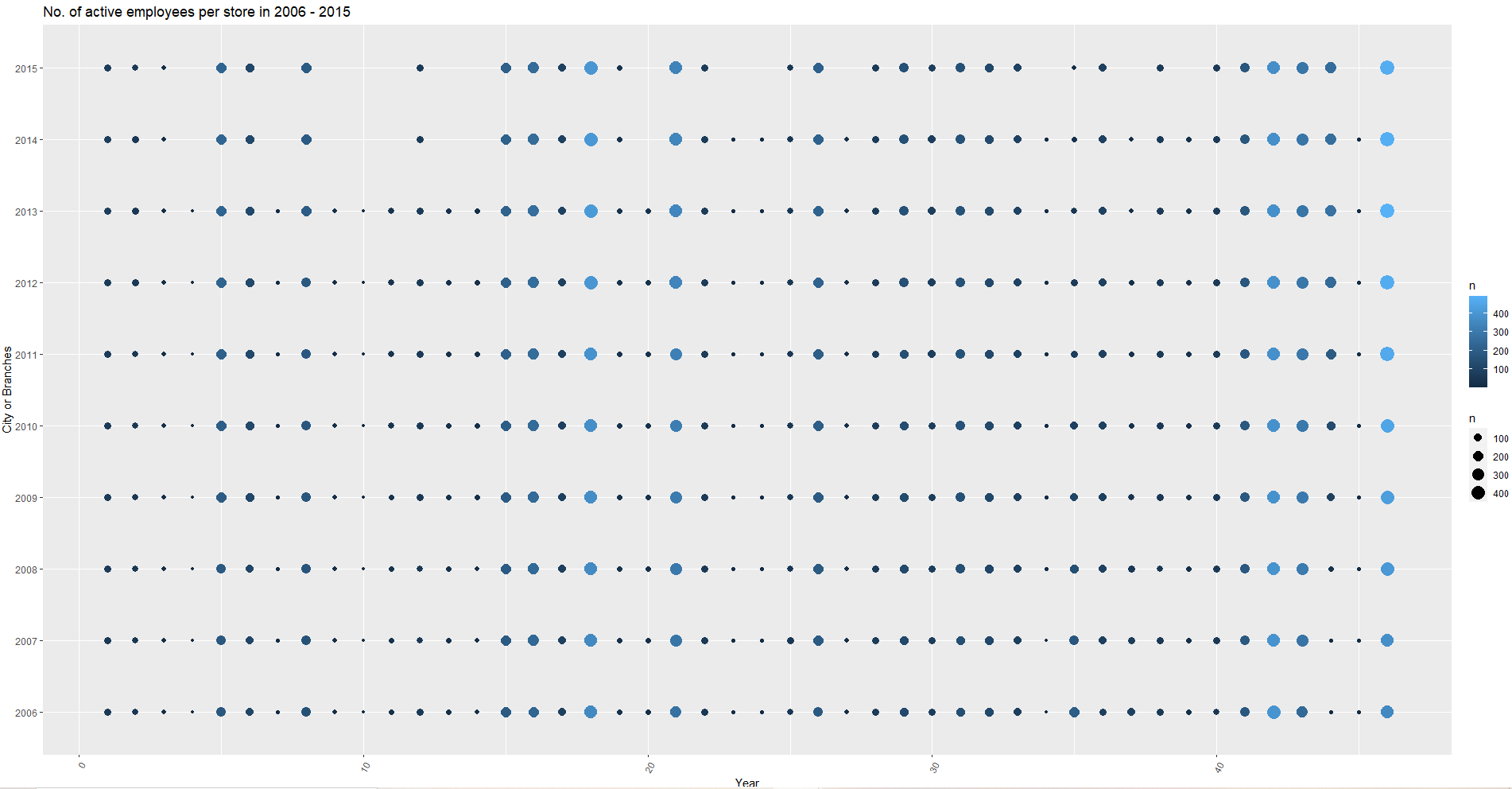
The visualization that has been categorized by year shows that cashiers are mostly hired in 2012 and 2013. Even though there are lot of termination by reason of layoff of resignation, cashier job is still in great demand for young employees.

# Why some stores are closed

1. Analysis 1 – How many stores that still active

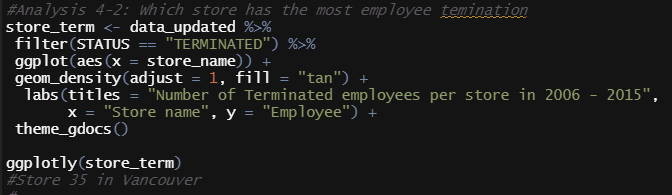


The analysis above is to know how many stores are still available. The code later will produce another variant of scatterplot in ggplot by using geom\_count function.

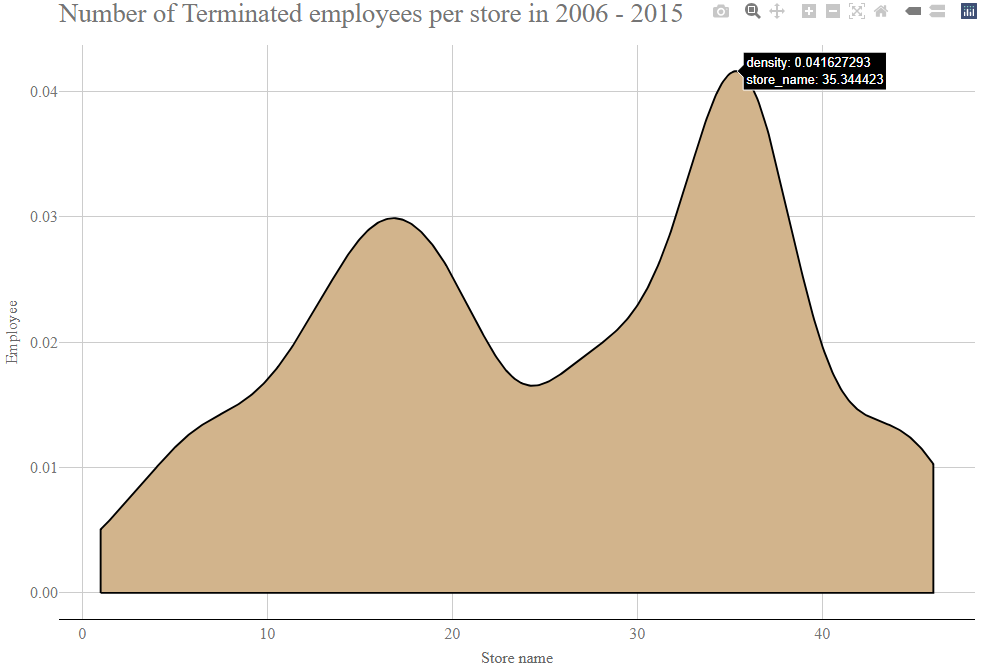


The result of the visualization is there are some stores that completely missing in 2014 and 2015. The active stores however are not affected to the closed one since the size of the dots are not changed.

1. Analysis 2 – Which store has most terminated employee

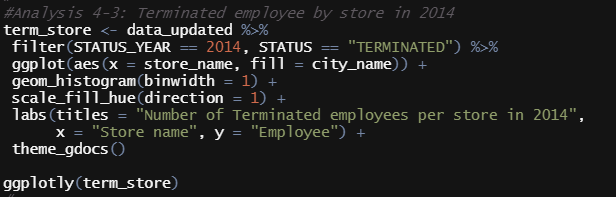


To know what happen to the missing stores, this analysis will create a density plot by using ggplot, then the visualization will be improved with labs, ggthemes, and ggplot.



The density plot above shows that the most terminated employee is in store 35 along with store 10 to 20. Store 30 Located in Vancouver which is the main store so it is excluded. However, store 10 – 20 is interesting since there are some missing stores there.

1. Analysis 3 – How many terminated employee by store in 2014

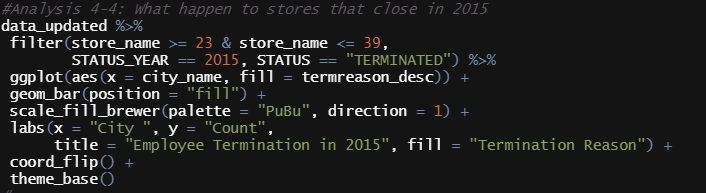


The third analysis is to look how many terminated in 2014 since there are a lot of stores missing in that year. It does create a histogram with a binwidth 1 by ggplot, then colored by scale\_fill\_hue function and do some renames with labs. It also added ggplotly for better visualizations.

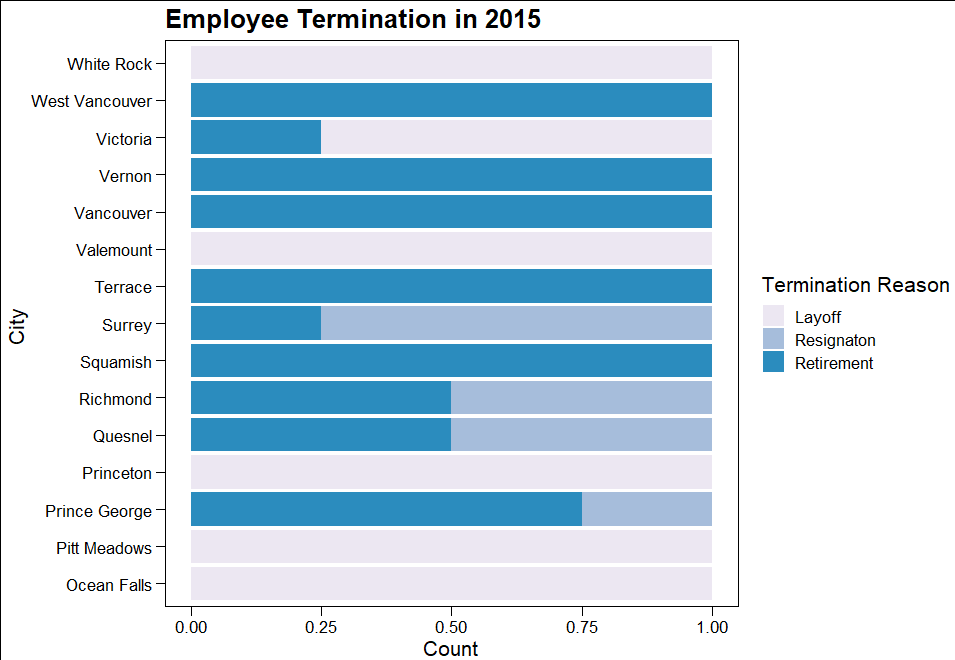


The histogram concludes that store 11 that located in Fort Nelson has the most terminated employee. Base on the main data, Fort Nelson has terminated all their employees, meaning store 11 has closed. This goes same by store 13 in Grand Forks. Based on my own research, these cities are located far from the main store which is Vancouver. The assumption that we have is the reasons of the terminated employee in 2014 is because the closing of stores that far from the main store.

1. Analysis 4 – What happen to stores that close in 2015



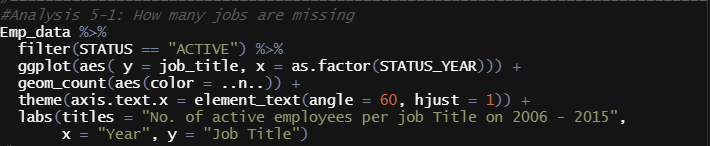
This analysis is to create a bar plot to know what happen to the closed stores in 2015. It does by using filtering the stores that missing in 2015 along with status and year, then use ggplot for the plot and scale\_fill\_brewer function and ggthemes.



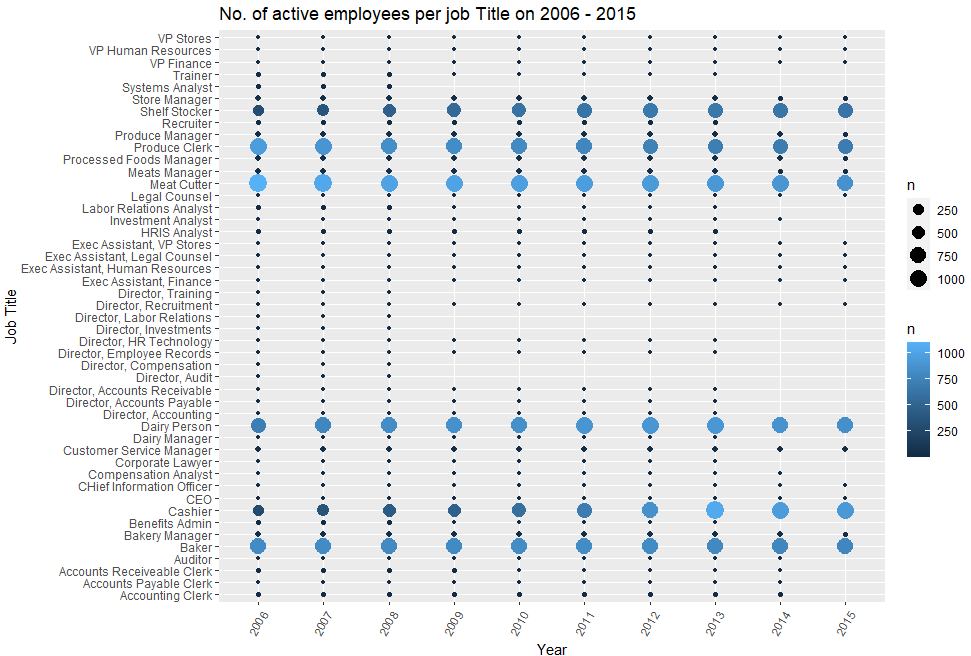
The visualization shows that store that missing due to layoff is less than retirement. These stores are closed due to the retired employee and the organization do not replacing the retired employee resulting in store closure.

# Is the missing department effect the organization

1. Analysis 1 – How many jobs are missing

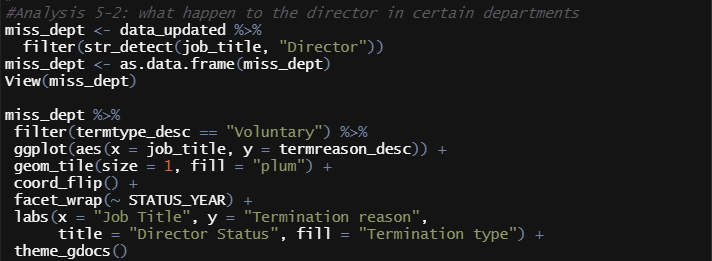


This analysis is to identify how many jobs title is missing from data in 2006 to 2015 by looking at the active employee based on job title in the scatterplot. After filter the status into active, use ggplot to visualize the data.

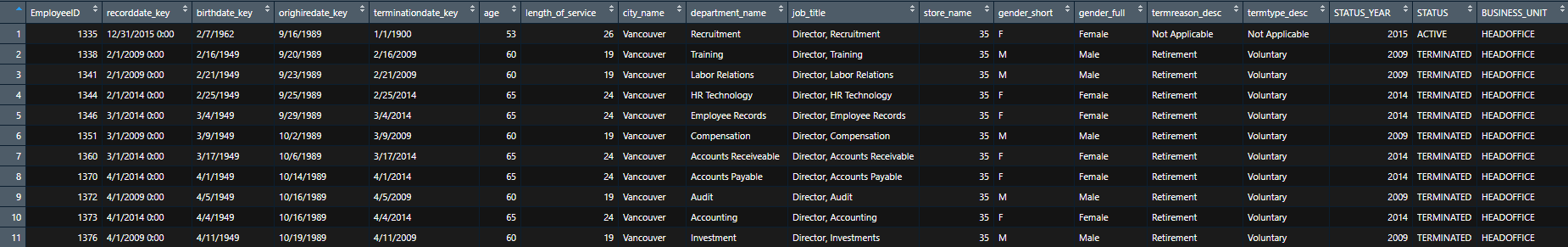


The scatterplot above summarized that there are a lot of job missing since 2009. This means that the reason of the massive termination is because there are a number of vacant job positions, thus makes the employees are not organized well.

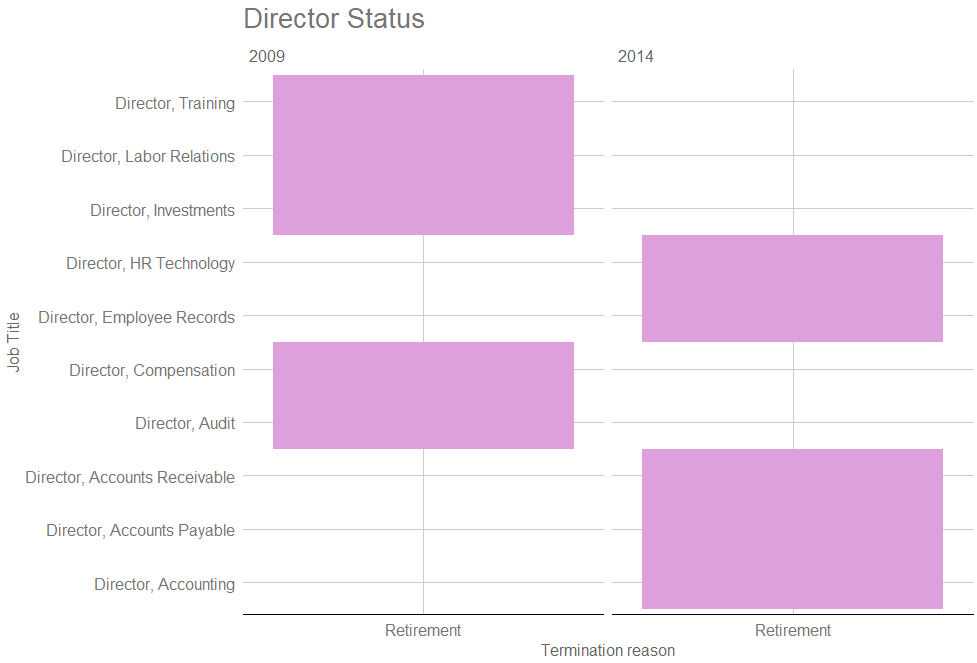
1. Analysis 2 – What happen to the director in certain departments



In this analysis we want to know what happen to the directors since there are a lot of vacant positions. Simply filter data by the job that starts with director with str\_detect function, then use ggplot to visualize the data into a tile plot.

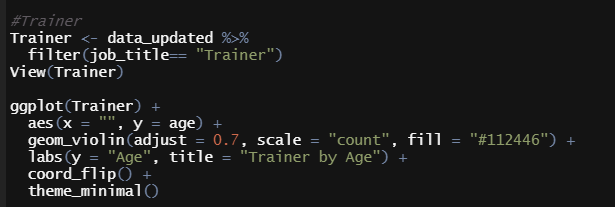
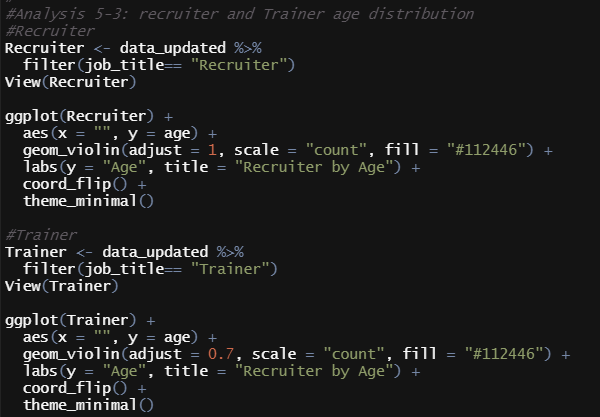


This part of data is the result of filtering by str\_detect function.

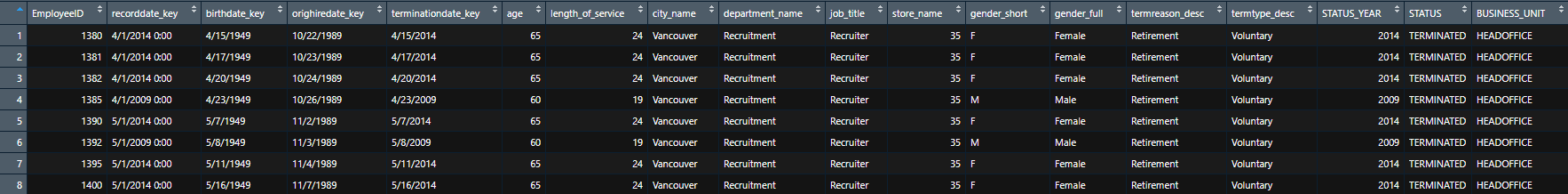


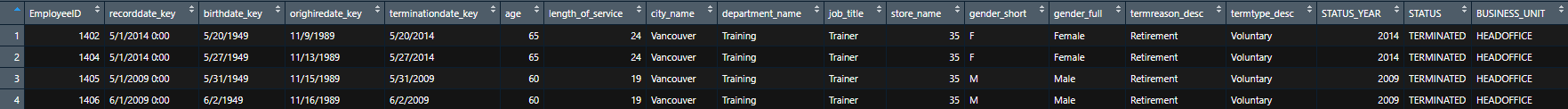
In the visualization above, it seems like all directors are gone due to retirement. It is unknown why the organization did not replace them, but this is a crucial position since directors’ role is to lead their department.

1. Analysis 3 – Recruiter and Trainer age distribution

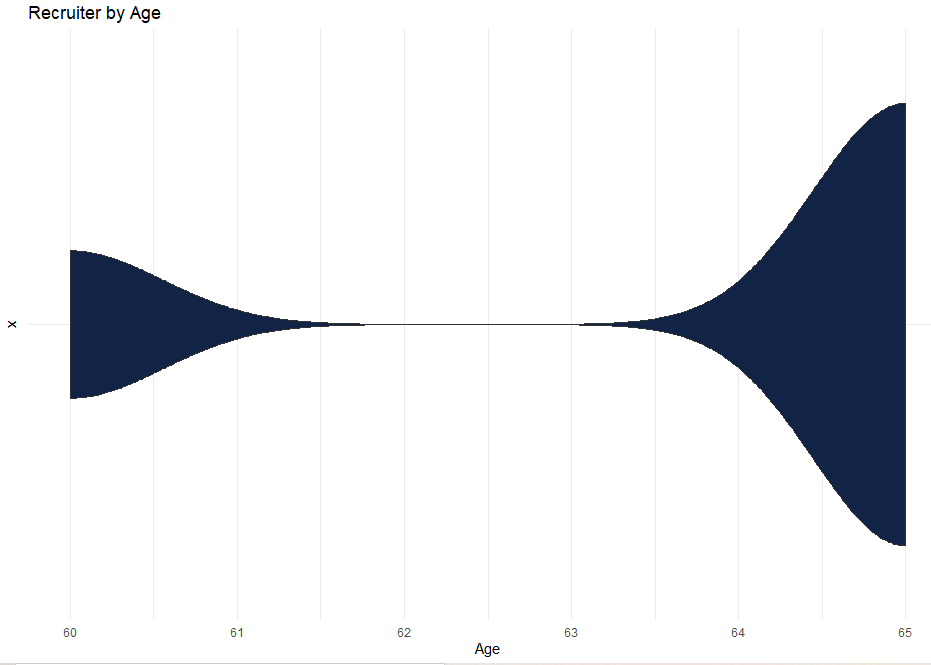
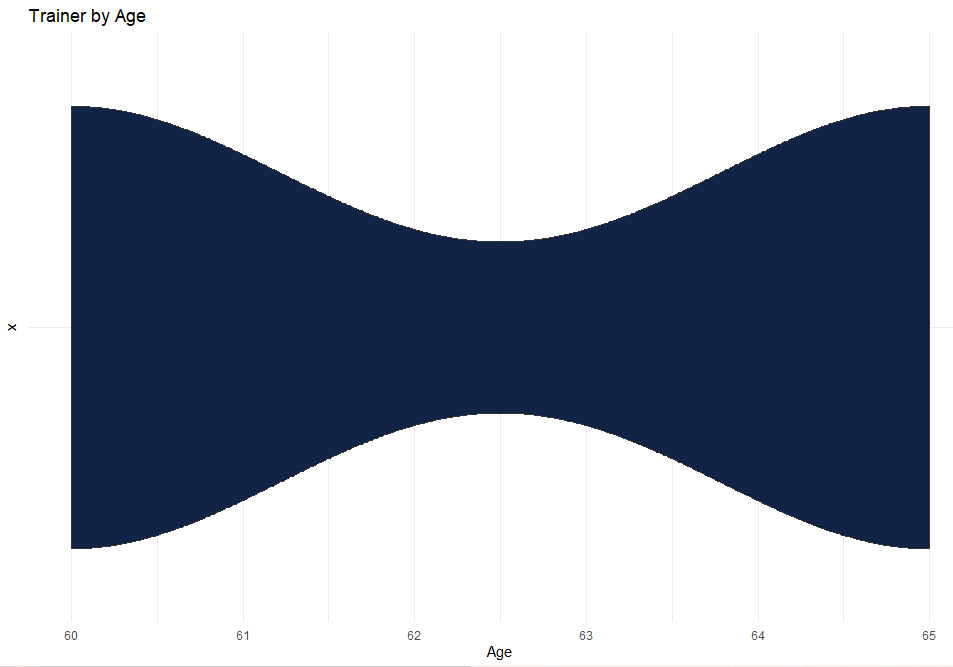


This analysis it to identify what is the age of the recruiters and trainers in the organization. First, we need to make a new data to filter the job into recruiter or trainer. Then we can visualize the data by using ggplot with violin plot.



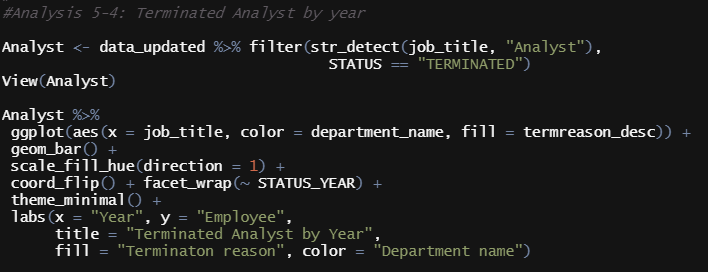


The table above is created before the visualization to get clear results with the analysis and to make sure that the visualizations are correct.

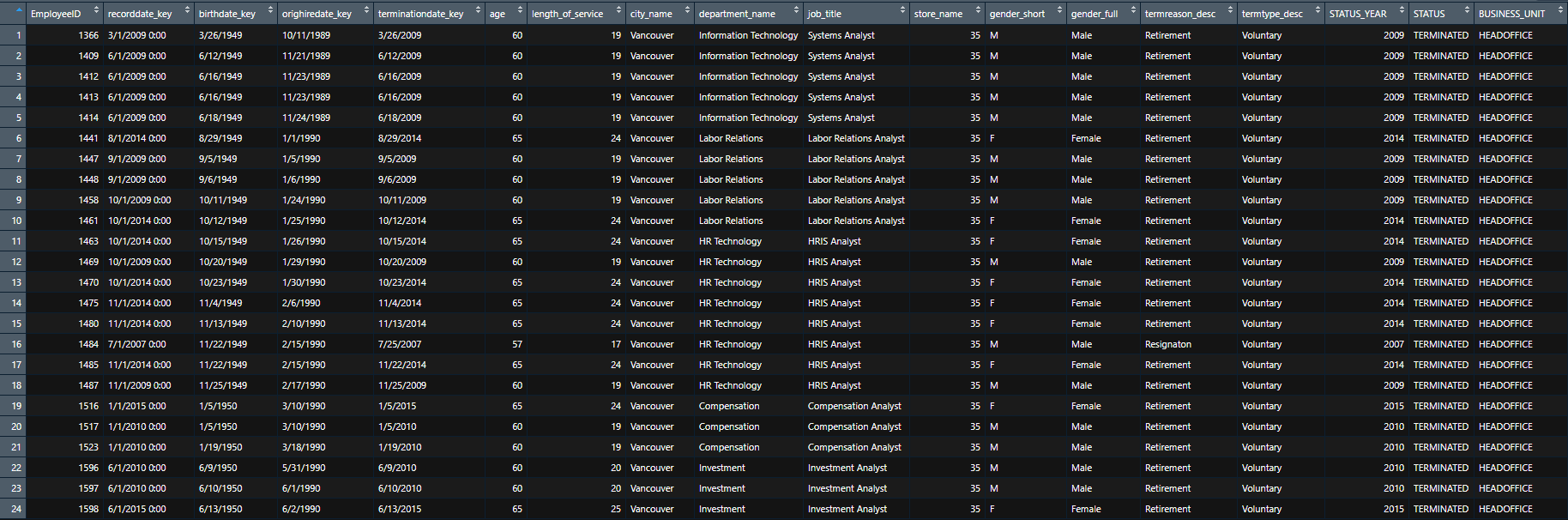
 

Based on the violin plot above, we are sure that the reason why there are no new employee in 2014 and 2015 is because the recruiter and trainers already in their retirement age which is 60 above.

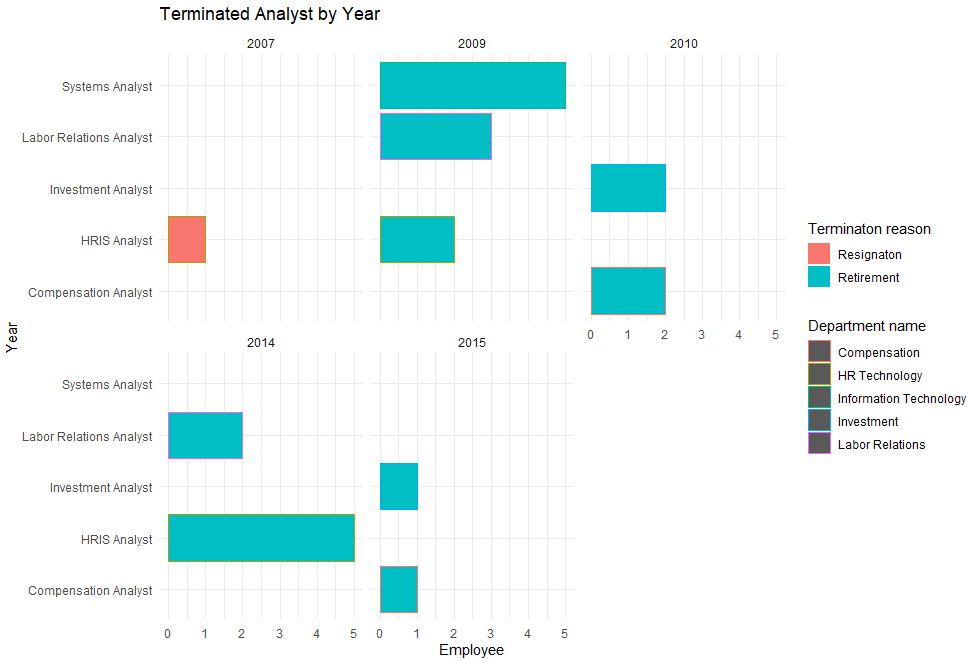
1. Analysis 4 – Terminated Analyst by year



This analysis is conducted to know how many analysts that has terminated by year. The steps are almost the same as the analysis 3 by filtering the job title into analyst with str\_detect function, then use ggplot with bar plot then use facet\_wrap function by year to split the data based on year.



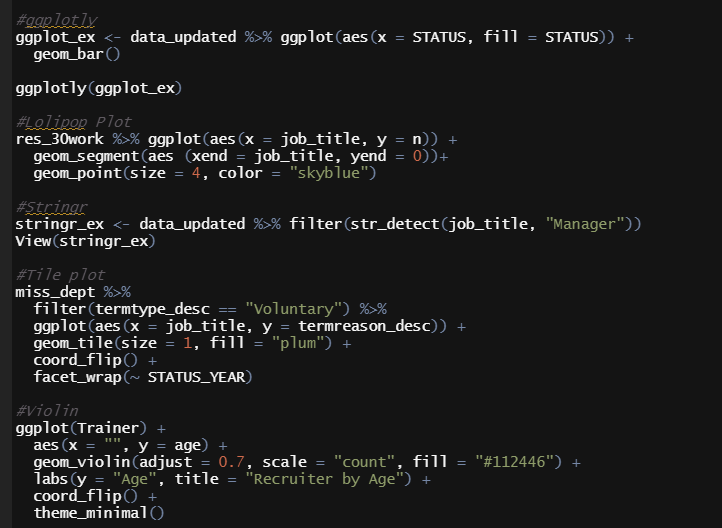
This is the table results of the filtering data with sub\_str function,



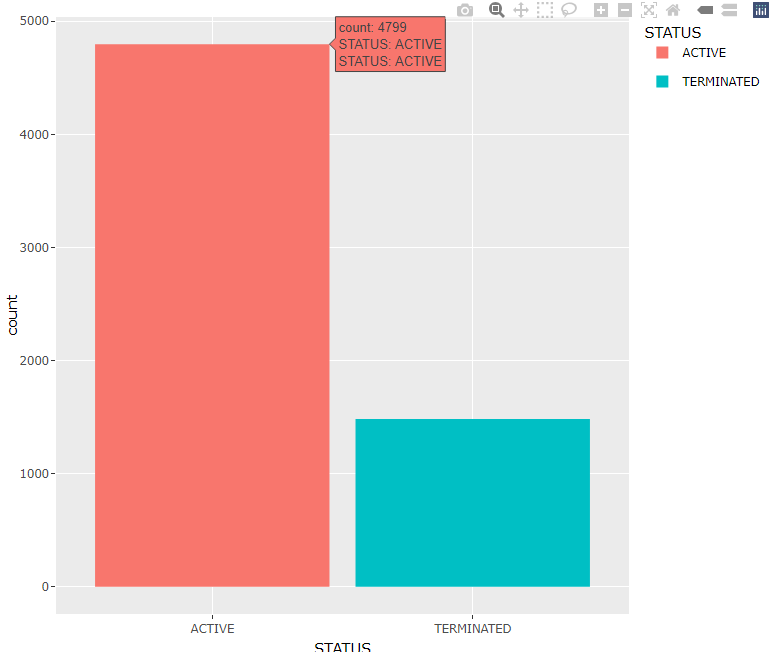
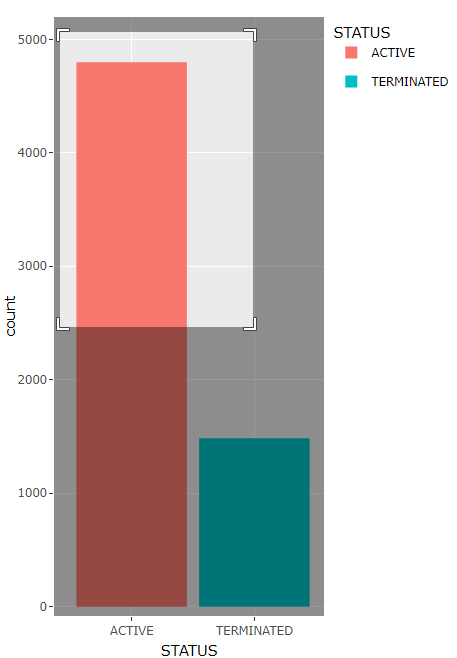
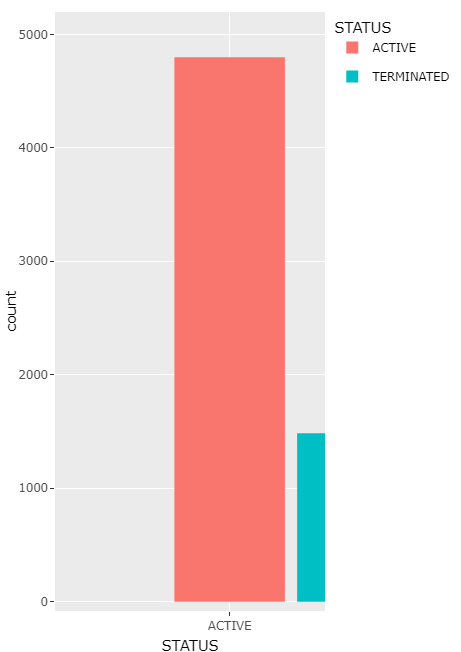
The bar plot shows that most of the analyst is missing due to retirement. This has the same case like the analysis before that the employee that has retired are not replaced into the new one. Analyst is very important since they are the ones who provide business and financial advice to the organization (indeed, 2021).

# Extra feature

1. Ggplotly

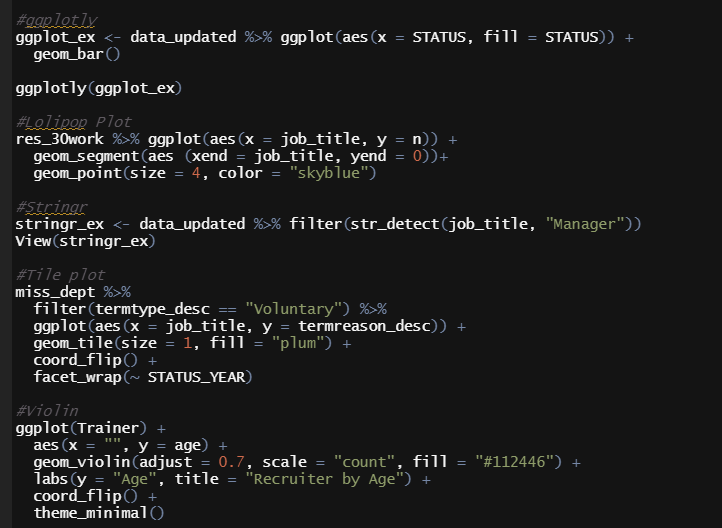


Ggplotly is a package that allow us to zoom, getting a tooltip by hovering over a marker, selecting a variable to show, and many more (Holts, Interactive charts | the R Graph Gallery, 2018). The syntax is we need to assign the data visualizations into the data, then use ggplotly function.

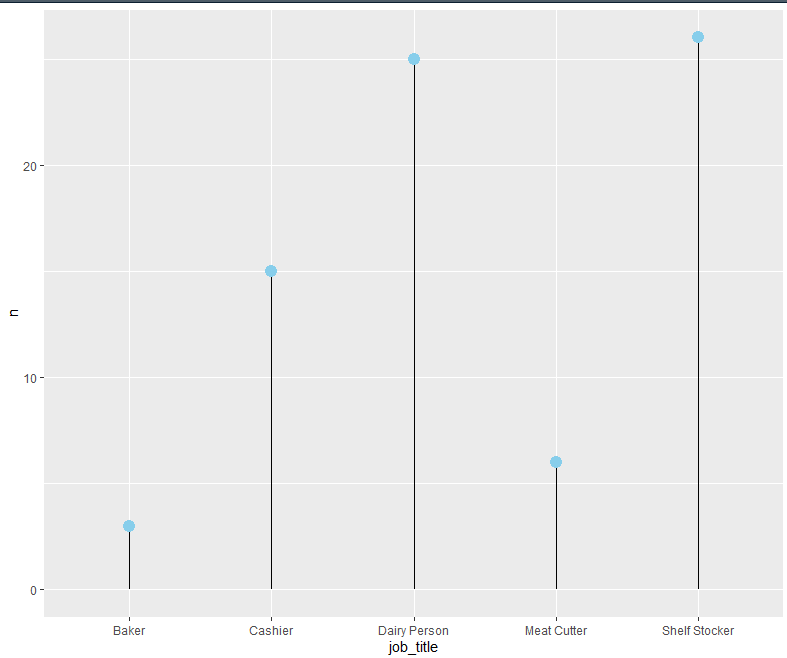


The output produced by ggplotly is interactive data that can be zoomed, panned, and showed the frequency if we select certain bars.

1. Lollipop plot

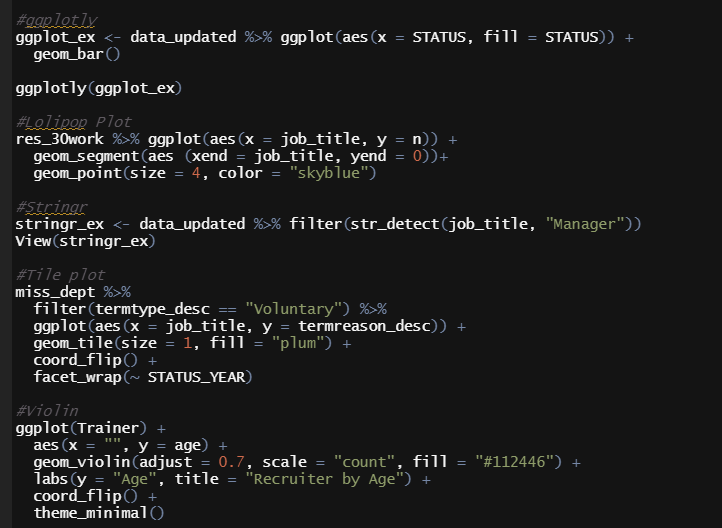


Lollipop plot is a bar plot that altered into a line and dot to demonstrate the connection between numeric and category variables (Holts, Lollipop plots, 2018).

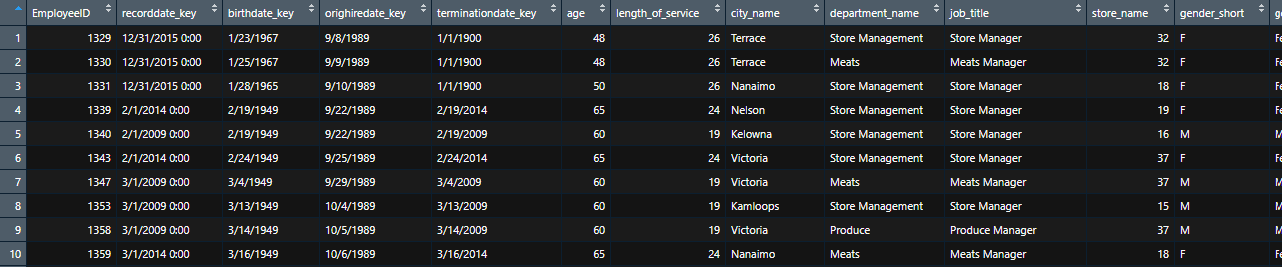


The output shows the simple version of bar chart that easy to look since if you only have 1 or 2 category value the bar plot will look like a box or rectangle that is not suitable for eyes.

1. Stringr

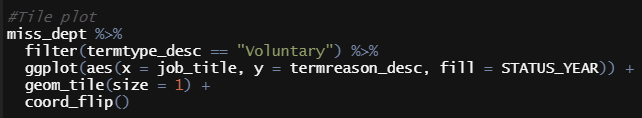


Stringr one of the packages in the tidyverse. It allows us to do many functions and one of them is str\_detect that determine whether a pattern exists in a character string (Schork, n.d.). The code above uses str\_detect inside filter function to filter job value that has analyst in the job title.

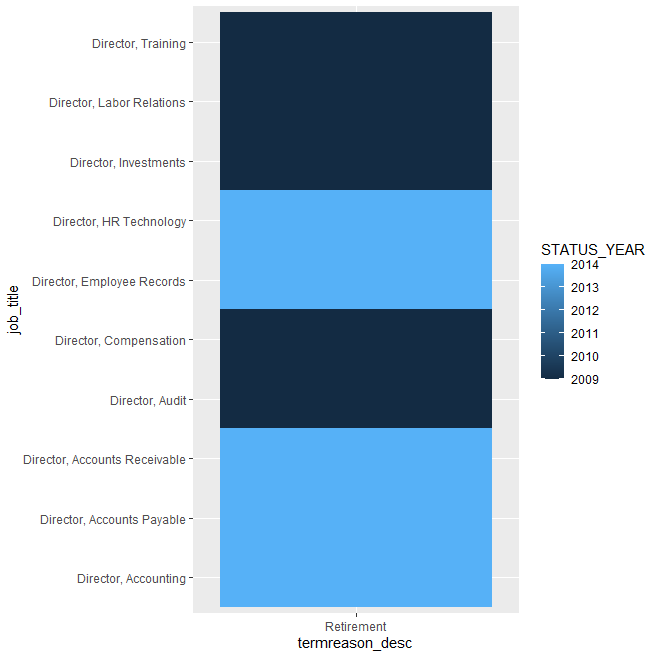


The output shows the data that has been filtered by sub\_str function. We can see that all of the job title contain manager. Sub\_str function is simple than applying job title value one by one.

1. Tile plot

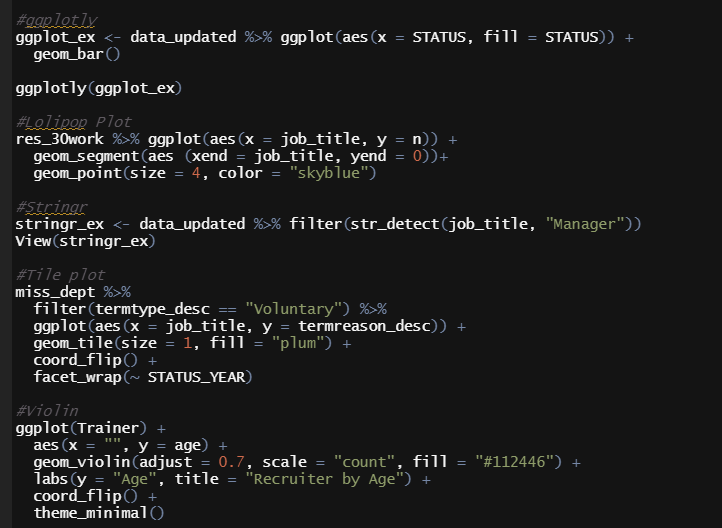


Tile plot is a plot that has rectangle shape and almost look the same as bar plot. The difference is it uses aesthetics in x and y. It is useful to represent data that has 2 categorical values.

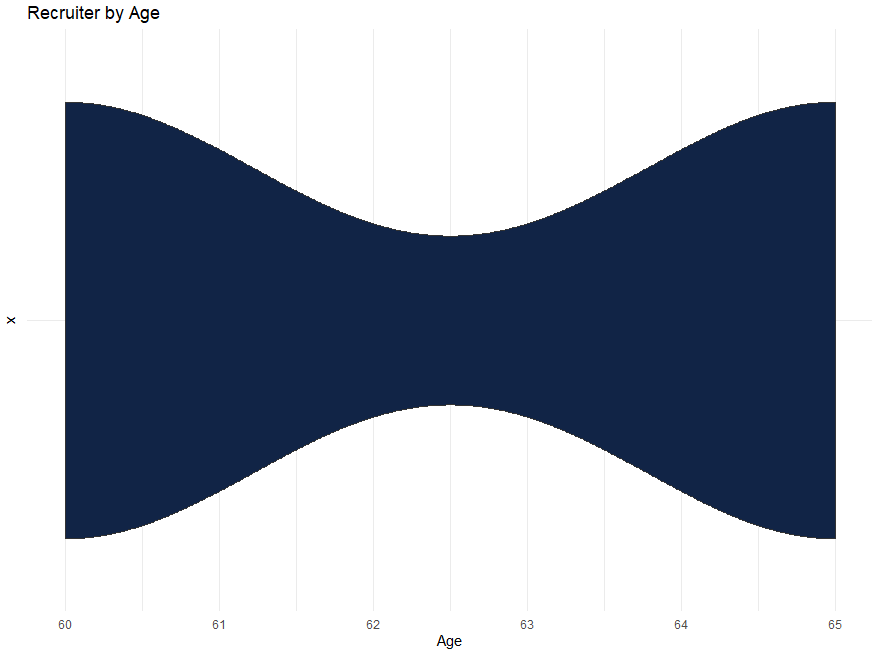


The tile plot above shows the relations of the job title and the termination reason. By using tile plot, we can see in what year the directors are retired easily.

1. Violin



Violin plot is a data visualization approach that is seldom used than boxplot. It allows us to inspect both the ranking and distribution of various groups.



The violin plot above representation of the distribution of a numeric variable for group which is recruiter. The shape shows the variable's density estimate: the larger the violin, the more data points there are in that range (Healy, n.d.). It's very similar to a boxplot, but it allows for a more thorough knowledge of the distribution.

# Conclusion

The organization is on the verge of collapse because they do not hire new employee again in 2014 – 2015. Based on the analysis, it is concluded that the Employee need to recruit more employee that can covered the crucial department or job such as directors and analyst to prevent massive resignation and layoffs in the company

# References

CBC News. (2014, May 31). *Canadian economy starts 2014 with weak GDP gain*. Retrieved from CBA bews Business: https://www.cbc.ca/news/business/canadian-economy-starts-2014-with-weak-gdp-gain-1.2659284

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

Healy, Y. (n.d.). *Violin plot*. Retrieved from from Data to Viz: https://www.data-to-viz.com/graph/violin.html

Holts, Y. (2018). *Interactive charts | the R Graph Gallery*. Retrieved from R graph Gallery: https://www.r-graph-gallery.com/interactive-charts.html

Holts, Y. (2018). *Lollipop plots*. Retrieved from R Graph Gallery: https://www.r-graph-gallery.com/lollipop-plot.html

Howard, J. (2021, June 5). *The Gender Pay Gap | Wage Gap in Canada | The Facts*. Retrieved from Canadian Women's Foundation: https://canadianwomen.org/the-facts/the-gender-pay-gap/

indeed. (2021). *Learn About Being an Analyst*. Retrieved from indeed: https://www.indeed.com/career-advice/careers/what-does-an-analyst-do

Schork, J. (n.d.). *str\_detect Function in R (stringr Package)* . Retrieved from Statistics Globe: https://statisticsglobe.com/str\_detect-function-stringr-package-in-r