Acknowledgment

I want to extend my heartfelt appreciation to Mr. Sandip Adhikari, the instructor of our module, for his superb leadership and assistance during the semester. His knowledge, advice, and helpful criticism were really helpful in assisting me in overcoming the obstacles presented by this project and achieving my objectives.

I also want to express my thanks to the instructors and staff at Infomax College for fostering a supportive learning atmosphere and giving me access to all the tools I needed to finish this project. I also want to express my gratitude to my classmates, who have continuously motivated, helped, and inspired me during this experience.

Finally, I want to thank my family and friends for their essential help. They have been an inspiration, a source of strength, and a constant source of support. Their faith in me has been crucial to my success. I am really appreciative of all of these people's support and advice, which were invaluable in helping me make my professional decision.

**Table of Contents**

[Introduction 1](#_Toc129527684)

[Data Import 1](#_Toc129527685)

[Data Cleaning 1](#_Toc129527686)

[Data Preprocessing 2](#_Toc129527687)

[Data Transformation 3](#_Toc129527688)

[Data Exploration: 3](#_Toc129527689)

[Question 1: Factors of attrition 6](#_Toc129527690)

[Analysis 1.1: Find the relationship between age and termination reason 6](#_Toc129527691)

[Analysis 1.2: How gender and termination are related 8](#_Toc129527692)

[Analysis 1.3: Find the relationship between Gender and termination type 9](#_Toc129527693)

[Analysis 1.4: Which age group is terminated at a high rate? 11](#_Toc129527694)

[Question 2: Where does attrition take place 15](#_Toc129527695)

[Analysis 2.1: Total terminated employees with different job titles with the specific reason 15](#_Toc129527696)

[Analysis 2.2 The connection between status and the name of the city. 17](#_Toc129527697)

[Analysis 2.3 Total terminated employees in each department with specific Type 19](#_Toc129527698)

[Question 3: How business can be expanded 21](#_Toc129527699)

[Output Findings: 22](#_Toc129527700)

[Analysis 3.2 Connection between the store name and the city name. 23](#_Toc129527701)

[Analysis 3.3 The connection between the Job and the name of the city. 25](#_Toc129527702)

[Question 4: What is the proportion of termination on the basis of the department? 27](#_Toc129527703)

[Analysis 4.1 The relationship between the department and termination type. 27](#_Toc129527704)

[Analysis 4.2 How the duration of service and the method of termination relate 29](#_Toc129527705)

[Extra Features 31](#_Toc129527706)

[Conclusion 32](#_Toc129527707)

[References 33](#_Toc129527708)

Table Of Figure

[Figure 1: Data Import 4](#_Toc129526153)

[Figure 2: Data cleaning 1 5](#_Toc129526154)

[Figure 4: Data Cleaning 2 5](#_Toc129526155)

[Figure: Data preprocessing 5](#_Toc129526156)

[Figure 4: Data Transformation 6](#_Toc129526157)

[Figure 5: Use of str() 6](#_Toc129526158)

[Figure 6: Output of figure 5 7](#_Toc129526159)

[7](#_Toc129526160)

[Figure 7: Use of names, nrow, ncol 7](#_Toc129526161)

[Figure 8: Output of figure 7 8](#_Toc129526162)

[Figure 9: Summary of the data 8](#_Toc129526163)

[Figure 10: Output of summary() 8](#_Toc129526164)

[Figure: Code for analysis 1.1 9](#_Toc129526165)

[Figure : Output for termination reason count 10](#_Toc129526166)

[Figure: Output of analysis 1.1 10](#_Toc129526167)

[Figure: Code of Analysis 1.2 11](#_Toc129526168)

[Figure: Output of analysis 1.2 11](#_Toc129526169)

[12](#_Toc129526170)

[Figure: Code for analysis 1.3 12](#_Toc129526171)

[Figure: Output of Analysis 1.3 13](#_Toc129526172)

[Figure: Code for analysis 1.4 14](#_Toc129526173)

[Figure: Output of analysis 1.4 15](#_Toc129526174)

[Figure: Source code for analysis 1.5 16](#_Toc129526175)

[Figure: Output of Analysis 1.5 17](#_Toc129526176)

[Figure: Code for Analysis 2.1 18](#_Toc129526177)

[Figure: Output for analysis 2.1 18](#_Toc129526178)

[Figure: Source code of analysis 2.2 20](#_Toc129526179)

[Figure: Output of analysis 2.2 21](#_Toc129526180)

[22](#_Toc129526181)

[Figure: Source code for analysis 2.3 22](#_Toc129526182)

[Figure: Output for analysis 2.3 23](#_Toc129526183)

[Figure: Source code of analysis 3.1 24](#_Toc129526184)

[Figure: Output for analysis 3.1 25](#_Toc129526185)

[26](#_Toc129526186)

[Figure: Source code for analysis 3.2 26](#_Toc129526187)

[Figure: Output for analysis 3.2 27](#_Toc129526188)

[Figure: Source code for analysis 3.3 28](#_Toc129526189)

[Figure: Output for analysis 3.3 29](#_Toc129526190)

[Figure: Source code for analysis 4.1 30](#_Toc129526191)

[Figure: Output for analysis 4.1 31](#_Toc129526192)

[Figure: Source code for analysis 4.2 32](#_Toc129526193)

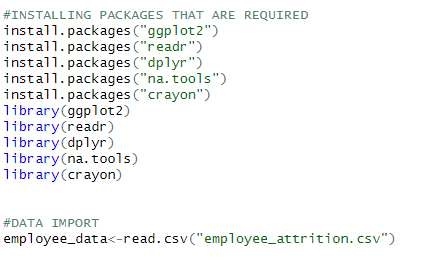
[Figure: Output for analysis 4.2 33](#_Toc129526194)

## 

# Introduction

This study aims to analyze the "Employee attrition.csv" file in-depth in order to find any hidden problems. In order to find any unseen issues that could impact the organization, this research will use the data supplied to develop a series of questions and analysis. The "Employee attrition.csv" file provides data on the workers in a dataset with 18 columns and 49653 rows. The data may be simply imported into R Studio for additional analysis because it is in CSV format. Data pre-processing is initially carried out to organize the dataset that has been provided. As there were no null values found in the data after the pre-processing, the analysis may go on to data cleaning. Columns that are not necessary are all eliminated from a set during data cleaning to make results more accurate. There were 4 questions in all, and 15 analyses were completed.

# Data Import

The process of importing external data into the R environment for further analysis is known as data import. R has a number of methods for importing data from different sources.

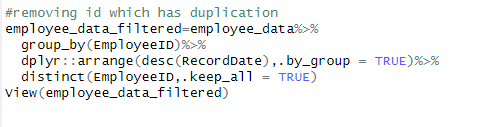
#### Figure : Data Import

With the help of the provided function, we have installed packages and read data from a CSV file and saved it as "employee\_data" which contains all that data.

# Data Cleaning

Data cleaning is a crucial step in the data analysis process. Before moving on to further in-depth research, it entails the process of finding and fixing mistakes or discrepancies in a dataset.

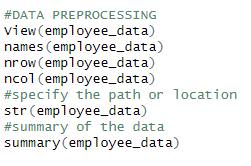
#### Figure : Data cleaning 1

On the basis of the given code the entire column having the name “gender\_short” is removed because there is the next column for gender named “gender\_full”.

#### Figure : Data Cleaning 2

And the “EmployeeID” which is duplicated is also deleted and one data which is latest is being recorded which covers all the data required for the analysis.

# Data Preprocessing

The preparation of the raw data for analysis is known as data preprocessing, which is a crucial stage in the process of analyzing the data. To prepare the data for analysis, preprocessing entails filtering, manipulating, and decreasing it.

#### Figure: Data preprocessing

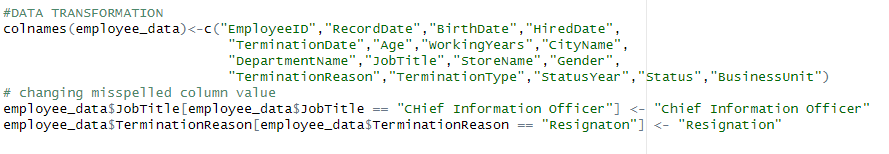
A function call in R's first line, View(employee\_data), opens a new window with a spreadsheet-like representation of the data in the employee\_data data frame. It enables interactive data exploration.

Another function call in R, names(employee\_data), retrieves the names of the columns in the employee\_data data frame.

The R function call nrow(employee\_data) returns the number of rows in the employee\_data data frame on the third line.

The R function call ncol(employee\_data) returns the number of columns in the employee\_data data frame on the fourth line.

# Data Transformation

The process of data analysis includes a crucial stage called data transformation, and R offers a large variety of tools and functions for doing so.

#### Figure : Data Transformation

The given code in figure 3 changes the names of the column using colnames().

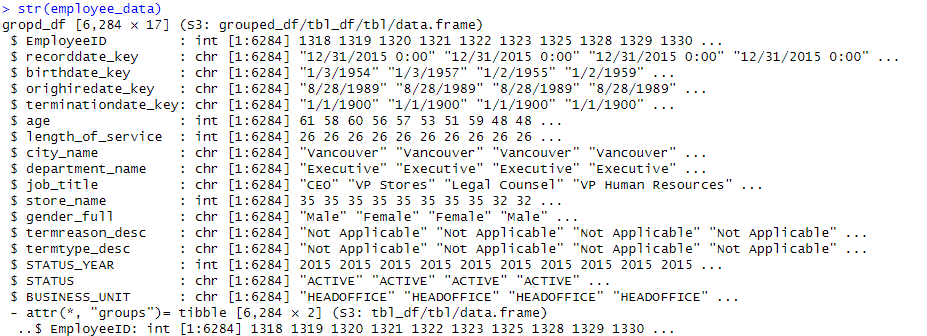
And the words which are misspelled like "CHief Information Officer" and "Resignaton" are changed into "Chief Information Officer" and "Resignation".

# Data Exploration:

**Exploration 1**

#### Figure : Use of str()

The str() function in R is used to indicate the path or directory of the source file which contains R code that you wish to run in your current R environment.

**Output**

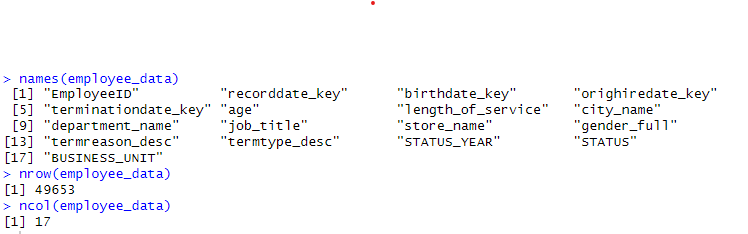
#### Figure : Output of figure 5

The str() function is used to examine an object's structure. It offers a concise and detailed overview of the internal structure of the object, including its category, measurements, and components.

**Exploration 2**

#### Figure : Use of names, nrow, ncol

The column names of a data frame are retrieved or set using the names() method. The number of rows in a data frame or matrix can be obtained using the nrow() function. The number of columns in a data frame or matrix can be obtained using the ncol() function.

**Output**

#### Figure : Output of figure 7

All of the data's titles are displayed in figure 8 so that they may be quickly identified and used to determine which titles need to be cleaned up. The number of rows and columns is the following statistic, which reveals that there are 49653 rows and 17 columns, confirming the information provided in the introduction.

**Exploration 3**

#### Figure : Summary of the data

For creating a summary of the key features of a data item, such as a vector, data frame, or matrix, in R, the summary() function is used.

**Output**

#### Figure : Output of summary()

You can see that the summary() method is an incredibly helpful tool for rapidly summarizing and comprehending the key features of R objects.

# Question 1: Factors of attrition

## Analysis 1.1: Find the relationship between age and termination reason

An essential study that may assist businesses in identifying any possible age-based discrimination is one that looks for a connection between an employee's age and the cause for their termination. In this kind of study, data on employee ages and termination causes are gathered, and statistical methods are used to look for any trends or links.

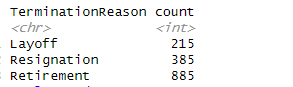
**The technique of analysis**: Data visualization and manipulation

**Source Code:**

#### Figure: Code for analysis 1.1

**Explanation:**

In order to know how many individuals were fired, quit their jobs, or retired, the code that was built first creates a filter to remove the "Not Applicable" data. Using the "geom bar" and the graph's color and labels, a bar graph is then generated. The graph's result is seen below.

**Output:**

#### Figure : Output for termination reason count

#### Figure: Output of analysis 1.1

**Explanation of output:**

The graph depicting the relationship between age and attrition stated above was created using three different indices. The ages of every employee in the company are shown below, and the colors red, green, and blue in the bar represent layoffs, resignations, and retirements, respectively.

**Summary of findings:**

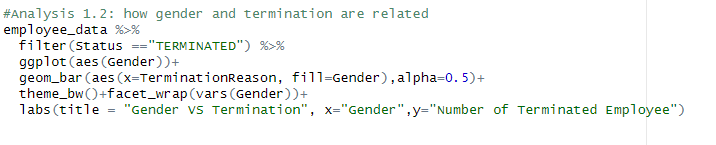
The graph demonstrates a tendency whereby persons in their 30s tend to resign at the highest rate, while those in their 60s tend to retire at the highest rate, while those in all other age groups tend to be laid off at around the same rate. According to an analysis, this tendency may result from the fact that workers in their 20s prefer to change jobs frequently because it gives them the chance to view the world from different perspectives, while workers in their 60s are more likely to retire because they have accumulated enough cost savings for their retirement plan and are getting older.

## Analysis 1.2: How gender and termination are related

The objective of this study is to determine whether there are gender inequalities in the likelihood of being fired from a job. This may be particularly crucial in businesses or professions where there has historically been a male- or female-dominated workforce, or if there is a perception of gender prejudice in recruiting practices.

**The technique of analysis**: Data visualization and manipulation

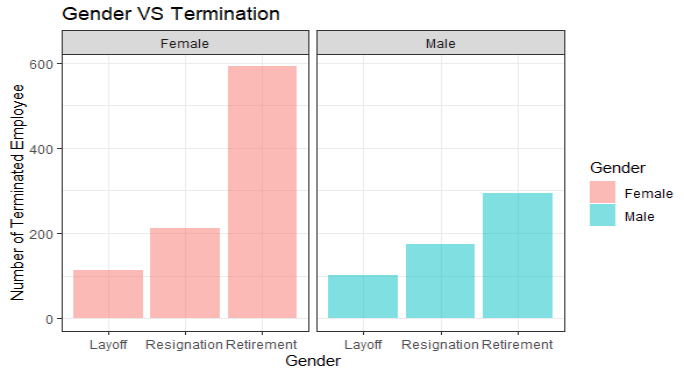
**Source Code:**



#### Figure: Code of Analysis 1.2

**Explanation:**

This system uses the ggplot2 library to create a bar chart after removing rows where the status is “Terminated” and taking employee data from a CSV file. The number of departing employees by reason is represented on the graph as pink for females and blue for males.

**Output:**

#### Figure: Output of analysis 1.2

**Explanation of Output:**

According to the graph that has been shown, both genders experience similar rates of layoffs and resignations. The data also shows that men and women had different rates of retirement, with women having a greater rate than males.

**Summary of findings:**

There seems to be a difference between the average retirement ages for men and women, with women typically retiring earlier. Women's health is one reason why these differences could exist. According to the study, women may retire earlier than males because they are more prone to have health issues as they age.

## Analysis 1.3: Find the relationship between Gender and termination type

The research on the relationship between gender and termination type, especially between voluntary and involuntary termination, is crucial since it has implications for businesses' human resource practices. Involuntary termination happens when an employer decides to sever an employee's contract due to performance, misbehavior, or other factors. Voluntary termination happens when an employee decides to quit the position of their own free will.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**

#### 

#### Figure: Code for analysis 1.3

**Explanation:**

The ggplot2 package is used in the code to build a bar chart that shows how gender and termination type are related. After eliminating any entries with the termination type "Not Applicable," it uses aes() to specify the plot's aesthetics by setting the x-axis to termination type and the fill color to gender.

**Output:**

#### Figure: Output of Analysis 1.3

**Explanation of output:**

According to the given graph, female employees experience more voluntary and involuntary terminations than male employees. This is due to the fact that, for both voluntary and involuntary termination types, the bars for female employees are higher than the bars for male employees.

**Summary of Findings**

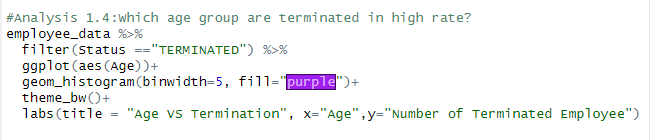
Gender and termination type are correlated, with female employees suffering more voluntary and involuntary terminations than male employees, according to the data. There may be many other contributing aspects to this gap, such as gender prejudice, discrimination, etc. In addition, women could be more willing to voluntarily quit their employment to take care of household duties or balance work and family obligations. Women may also be more susceptible to maternity leave or pregnancy-related involuntary terminations that result in job loss or a reduction in hours worked.

## Analysis 1.4: Which age group is terminated at a high rate?

Finding the age group with the highest termination rate is the main goal of study 1.4. This is a crucial subject for research because pinpointing the age group most at risk for job loss may assist businesses in creating focused retention and turnover prevention efforts. This study's goal is to obtain an understanding of the precise causes of greater termination rates in various age groups.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**

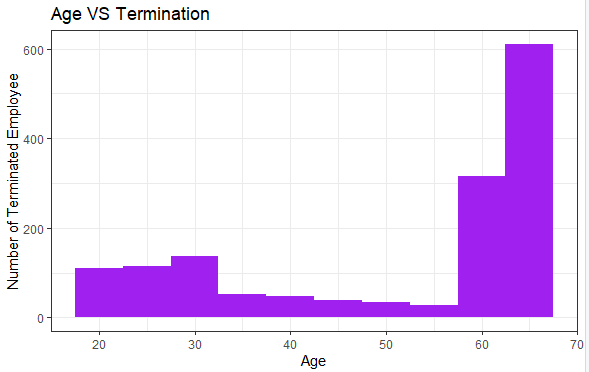


#### Figure: Code for analysis 1.4

**Explanation:**

The filter function and the Status column are used by the code to first restrict the dataset to only contain terminated workers. The ggplot function is then used to generate a ggplot object, with the Age column set as the x-axis. Using the geom histogram function, which demands a binwidth of 5 and a fill color of purple, a histogram is produced.  The plot is then given a title and axis labels using the labs() function. Overall, this code gives a visual representation of the dataset's terminated workers' age distribution.

**Output:**



#### Figure: Output of analysis 1.4

**Explanation of output:**

The code's output indicates that, in comparison to other age groups, employees above the age of 60 have a greater rate of termination. This can be observed in the graph, where the histogram is largest for those between the ages of 60 and 70 in comparison to other age groups.

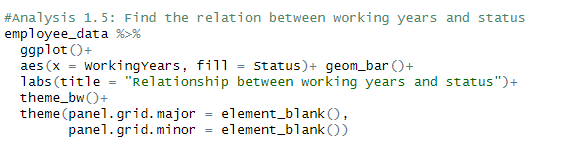
**Output Findings:**

Employee data research shows that workers over 60 experience a greater rate of termination than those in other age groups. There are several explanations for this. Older workers may struggle to execute their jobs efficiently due to health problems or physical restrictions, which is one possible explanation. They could lose productivity and quality of work as a result, which might lead to termination. Older workers may also lack modern skills or be reluctant to adopt new technologies or working methods, which might reduce their value to the organization.

**Analysis 1.5: Find the relation between working years and status**

The purpose of this study is to establish the connection between an employee's number of years of employment and their status, such as active or terminated. The aim is to determine whether there is any relationship between an employee's duration and their risk of being fired.

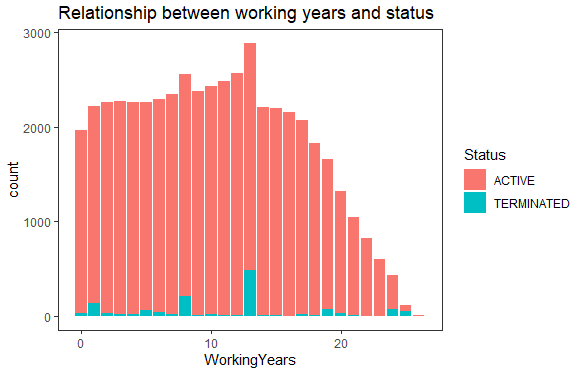
**The technique of analysis**: Data visualization and manipulation

**Source code:**

#### Figure: Source code for analysis 1.5

**Explanation**

The association between the number of working years and the status of workers is depicted in a bar chart using this code. The aes() function translates the WorkingYears variable to the x-axis and the Status variable to the fill color of the bars after the ggplot() function creates a new instance of the ggplot object. Each bar in a bar chart made with the geom\_bar() function indicates the number of employees with a specified status and a specific number of working years.

**Output:**

#### Figure: Output of Analysis 1.5

**Explanation of Output:**

According to the plot, the proportion of active employees peaks at roughly 13 years of employment and subsequently declines as the number of working year increases. Adversely, the proportion of fired workers is rather low for those with 1 to 5 years of work experience, but it steadily rises for those with more than 6 years.

**Output Findings:**

The peak in active employees at roughly 13 years of employment may indicate that this is a time when individuals may be having professional progress or consistency, which might boost job fulfillment and motivation to continue to work. But, when the number of working years climbs past this point, it's likely that some employees might start to feel burned out or start thinking about retiring. However, the rise in dismissals of workers with more than six years of experience may be a result of concerns with performance, modifications to job specifications, or even age discrimination.

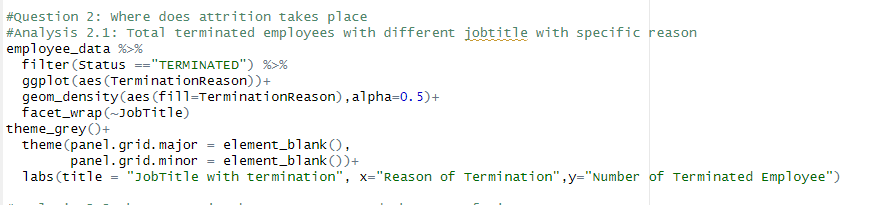
# Question 2: Where does attrition take place

## Analysis 2.1: Total terminated employees with different job titles with the specific reason

Total terminated workers with various job titles refer to the total number of employees in a specific organization that has been fired from their positions. Depending on the specific situation, the grounds for termination may include things like bad performance, a violation of business ethics or regulations, reorganization, reduction, or other organizational practices.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**



#### Figure: Code for Analysis 2.1

**Explanation:**

The ggplot2 package assists in the creation of a data visualization of the reasons for terminating employees for various job titles. The code limits the employees in the employee data frame to those who have been fired. The termination reasons are then created as a density plot, with the transparency set to 0.5 and the fill color chosen by the termination cause. The storyline is then divided up into several panels according to each job title. The main and minor grid lines are eliminated by the theme feature. The plot title and axis labels are set by the labs function. Using the density map, it's simple to compare termination reasons across job titles and see trends or patterns.

**Output:**

#### Figure: Output for analysis 2.1

**Explanation of output**

The result of this code is a data visualization that displays the distribution of reasons for terminating employees in a company for various job titles. The figure is divided into numerous panels, each of which represents a distinct job title. In each panel, a density plot showing the distribution of termination reasons is displayed. The termination cause determines the fill color of each density plot, which is overlaid. The plot's y-axis displays the density values while the x-axis lists the various termination causes. For clarity, the axis labels and title are present. By looking at the plot, we can observe how frequently each job title is terminated as well as any trends in the causes of termination across various job titles.

**Output Findings**

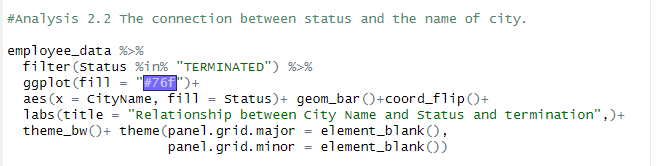
We may determine the most typical reasons for firing employees across all job titles by examining the density plots. We can also spot any changes in the distribution of termination grounds between various job titles. Also, the plot reveals irregularities or strange patterns in the distribution of termination causes that call for more research. The management may take the required actions to promote employee engagement, improve the working environment, and decrease employee turnover by knowing the reasons why people are leaving the company. This can eventually result in a rise in organizational productivity and success.

## Analysis 2.2 The connection between status and the name of the city.

Understanding the dynamics of the workforce and the elements that influence employee satisfaction in various geographical areas may be gained by analyzing the link between the employment status and the name of the city. We may spot any patterns or trends that might be affecting employee turnover and engagement in these places by analyzing the distribution of active and terminated employees across various cities.

**The technique of analysis**: Data visualization and manipulation

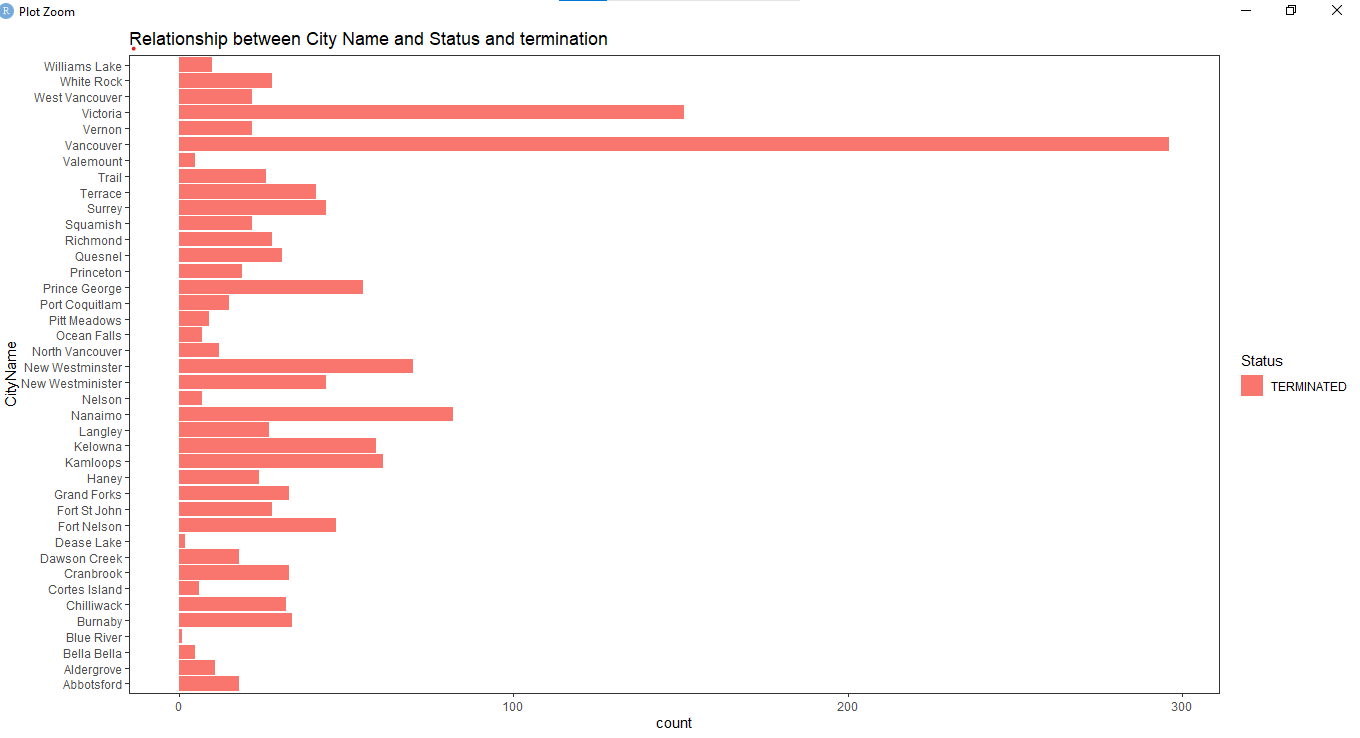
**Source Code:**



#### Figure: Source code of analysis 2.2

**Explanation:**

The data is first filtered by the algorithm to only contain workers that have the status "TERMINATED". The bar plot is then made using the CityName on the x-axis and the number of fired employees on the y-axis using ggplot2. Employee status is reflected in the color of the bars, with dismissed personnel represented by a single color.  To make the graphic more understandable, the x and y axes are reversed using the coord\_flip() function, and the plot is given a white and black theme using the theme bw() function. The plot has a title that is added using the labs() method.

**Output:**

#### Figure: Output of analysis 2.2

**Explanation of output:**

The result displays a bar chart with each bar denoting a city and its height denoting the quantity of fired workers inside that city. According to the status of the fired employees, the bars are colored (i.e., terminated or active). We can see that in some cities compared to others, there are more fired workers. In order to assist identify possible difficulties in the workplace that need to be addressed, the chart might be beneficial in identifying cities with a greater prevalence of employee terminations.

**Output Findings:**

Cities with a more competitive employment market may have a higher tendency to fire workers who fall below expectations. In contrast, localities with stricter labor rules could be less inclined to let people go for reasons outside of performance. To properly understand the variables behind the observed difference, more investigation would be required, such as looking at the particular reasons for termination and comparing them across different cities.

## Analysis 2.3 Total terminated employees in each department with specific Type

We may categorize the data by department and termination type to assess the overall number of terminated employees in each department with a certain kind. For any combination of department and termination type, this will give us a count of the employees who has been fired. By doing this, we can determine which departments have the most fired personnel as well as the most typical types of termination within each department.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**

#### 

#### Figure: Source code for analysis 2.3

**Explanation:**

Using ggplot2, this code filters out terminated workers and generates a bar plot that displays the number of terminations by category for each department. Department names are used to segment the bar plot. The plot's gridlines are removed, and the coding utilizes a gray color scheme. The plot's name is "Termination with Department Name," and the x-axis and y-axis labels on the graph are "Kind of Termination" and "Number of Terminated Employees," respectively.

**Output:**

#### Figure: Output for analysis 2.3

**Explanation of Output:**

The graph demonstrates that across all departments, there are usually more voluntary terminations than involuntary terminations. This may be the result of a number of factors, including better employment offers, discontent with the existing position, or privacy considerations.

**Output findings:**

There may be a number of reasons why there are more voluntary than involuntary terminations across all departments. Better work offers or chances might be one of the causes. Employees could freely quit their existing employees if they are presented with better job opportunities. Another cause can be dissatisfaction with the current job due to factors like low salary, poor management, a lack of promotion possibilities, or a toxic work environment. Employees may decide to leave their employment freely in these situations rather than waiting for an involuntary termination.

# Question 3: How business can be expanded

**Analysis 3.1 The connection between city and status**

We may study the employment patterns in various locations by looking at the relationship between city and status. We may learn more about the elements that might be influencing the status of employees in various places by viewing the number of active and terminated employees in each city. We could see, for instance, that some cities have a greater rate of layoffs than others, which might be a sign of problems with the regional economy or difficulties unique to a given sector.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**

#### Figure: Source code of analysis 3.1

**Explanation:**

Based on the CityName and Status columns of the employee data filtered dataset, the code creates a bar chart. The figure is made using the ggplot() tool, with each employee's status represented by the fill and the CityName on the x-axis. The geom bar() function is used to create the bar chart, while the coord flip() function is utilized to flip the x and y axes. The labs() method is used to title the plot, and the theme classic function is used to apply a traditional theme. The theme function is used to remove the major and minor grid lines from the story.

**Output:**

#### Figure: Output for analysis 3.1

**Explanation of output:**

The result is a bar graph that shows the association between the city name and the employment status of the dataset's workers. Cities are shown on the x-axis, while the number of employees is represented on the y-axis. The bars are colored blue for current employees and red for dismissed employees in accordance with the status of the employee. For additional research and decision-making, this graph may be used to examine the distribution of active and terminated personnel across various cities.

## Output Findings:

We can see from the graph that certain cities have a larger proportion of fired workers than others. This could be caused by a number of things, including high turnover rates in particular businesses or industries, a lack of employment possibilities in a particular city, or the location of a company in a less desired neighborhood. Also, we can observe that some cities have a greater percentage of employees who are actively employed, which may be a result of better employment options, a more stable economy, or stronger labor demand in such regions. Overall, this graph may be used to spot patterns in employee status across various locations and can give businesses and decision-makers information to assist them to strengthen their strategy for hiring and retaining staff.

## Analysis 3.2 Connection between the store name and the city name.

The relationship between the name of the store and the name of the city might reveal information about the locations of the stores and how well they are doing in other places. By examining this relationship, it can be able to spot trends or patterns that might guide company decisions regarding resource allocation or market focus.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**

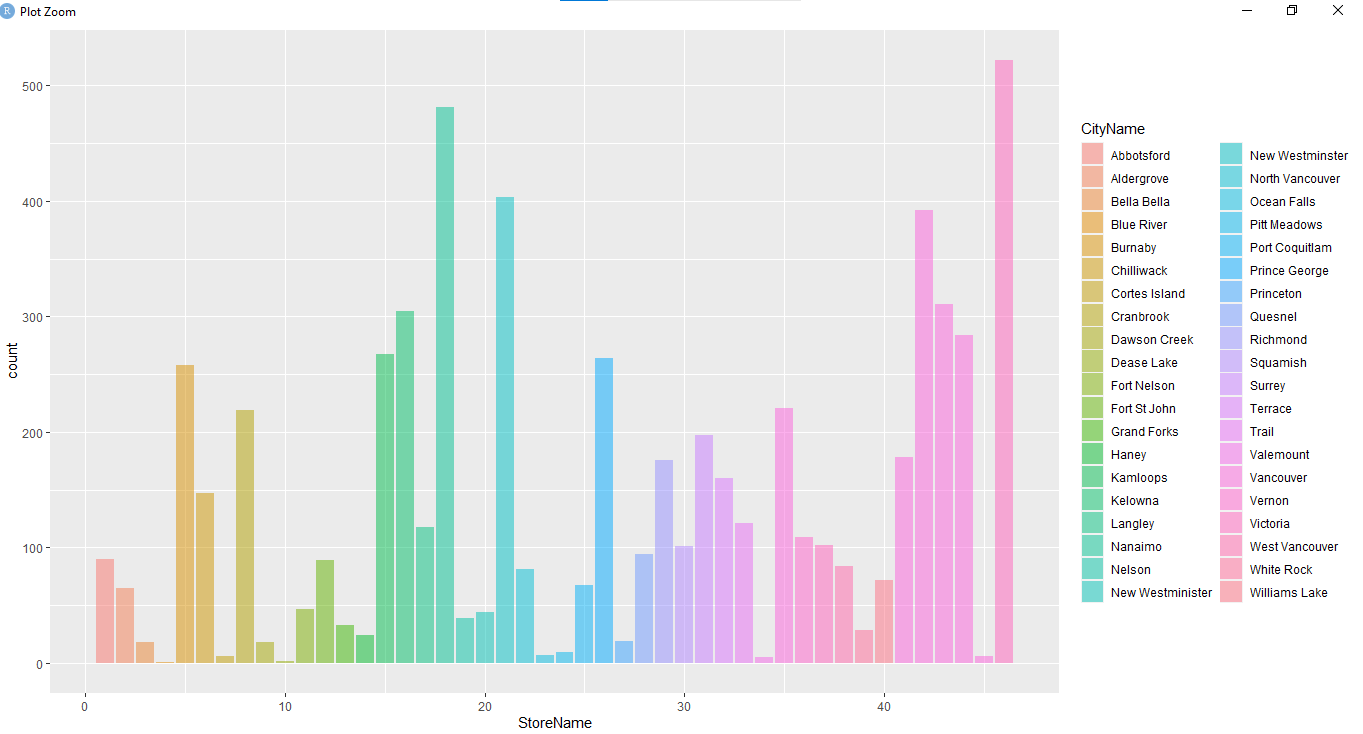
#### 

#### Figure: Source code for analysis 3.2

**Explanation:**

The employee data filtered dataset's city name and shop name relationships are shown in a bar graph created by this code. The y-axis displays the number of stores, while the x-axis lists the names of the stores. Each color on the bars denotes a separate city, and the bars are colored in accordance with the names of the cities. To make the colors just a little bit translucent, the alpha parameter is set to 0.5. The background color is changed to grey using the theme grey() function, and the grid lines are eliminated using the theme function. The graph's title and axis labels are added using the labs() function.

**Output:**



#### Figure: Output for analysis 3.2

**Explanation of output:**

The result is a bar graph that depicts the association between the name of the retailer and the name of the city. Store names are represented on the x-axis, while store counts are shown on the y-axis. The bars are colored based on the name of the city. This graph makes it easier to see how stores are distributed throughout various cities, which is helpful for future research and decision-making. We can see from the graph that certain cities have more stores than others, and some stores are spread throughout several cities. The fact that the bars are not layered allows us to view the entire number of stores in each city.

**Output Findings:**

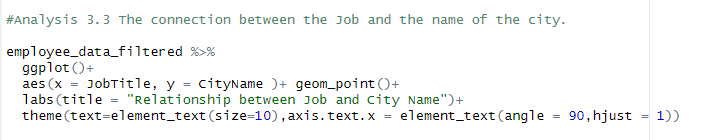
This graph shows that some stores are more well-liked in particular cities than others, while other stores may not be as well-liked in those locations. This could be the result of a variety of things, including the store's location, size, or selection of goods and services. Another result that may be made is that some cities might have a larger concentration of workers than others. This could be the result of elements like the city's employment prospects, the cost of living, or the cultural and social atmosphere.

## Analysis 3.3 The connection between the Job and the name of the city.

The association between the type of employment and the place where it is carried out is referred to as the connection between the job and the name of the city in data analysis. Understanding the workforce distribution across different places and finding possible areas for improvement in terms of employment possibilities and skill development may both benefit from this link. Organizations may gather knowledge on labor market trends and modify their recruiting and retention tactics by studying the relationship between jobs and cities.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**



#### Figure: Source code for analysis 3.3

**Explanation:**

The ggplot() function starts the plot, aes() specifies the variables to be displayed on the x and y axes, and geom\_point() generates the scatter plot in terms of code. The plot's title is added using labs(), and the plot's visual characteristics, such as font size and x-axis label angle, are changed using the theme().

**Output:**

#### Figure: Output for analysis 3.3

**Explanation of output:**

The scatterplot can be helpful for seeing trends or patterns in the data and for visualizing the distribution of job titles across several cities. In this situation, it may be used to determine which job titles are more prevalent in particular cities. The x-axis labels on the graph, which reflect the job titles, may be lengthy and many, as shown by the usage of angle and font size in the code. As a result, the graph's readability is improved by rotating and resizing the font.

**Output findings:**

The plot demonstrates how some job titles are more prevalent in some cities. Different local economies, sector specialties, or population demographics might all play a role in this. The association between job titles and city names in the dataset is usefully summarized in this figure, which may help guide additional analysis and decision-making.

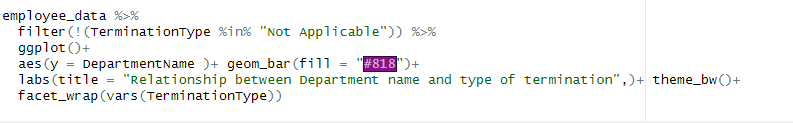
# Question 4: What is the proportion of termination on the basis of the department?

## Analysis 4.1 The relationship between the department and termination type.

The correlation between an employee's department and the form of termination (voluntary or involuntary) they go through is referred to as the relationship between department and termination type. By examining the frequency and sharing of each form of termination across several departments, this link may be studied. So, any trends or differences in the explanations for voluntary or involuntary termination across several departments can be found.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**



#### Figure: Source code for analysis 4.1

**Explanation:**

This code generates a bar chart for dismissed employees that displays the correlation between department name and kind of termination. Filtered out is the termination type of "Not Applicable". The chart is shown with a white backdrop, and the bars are filled with a grey tint. The chart is then divided into smaller charts, one for each sort of termination, and faceted according to termination type. This makes it simple to compare the distribution of department names for each form of termination.

**Output:**

#### Figure: Output for analysis 4.1

**Output justification:**

The plot shows the number of employees who voluntarily left the company in each department. The highest rates of voluntary turnover are seen in Fruits, Meats, and Customer Service. Nevertheless, only the departments responsible for department management, produce, packaged goods, meats, dairy, customer service, and baking have seen forced layoffs of employees.

**Summary of findings:**

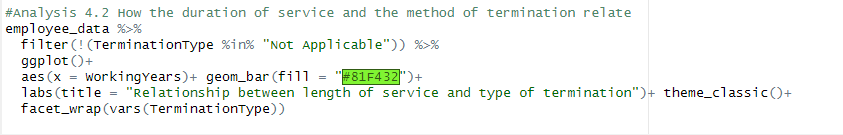
Employees in some departments may be more satisfied with their jobs and less likely to leave on their own, whereas employees in other departments may be less satisfied and more likely to resign on their own. Increased rates of both voluntary and involuntary departure may be the outcome of excessive workloads and stressful work situations. The efficiency of a department's management might potentially affect turnover rates. In comparison to higher rates in regions with poor management, worker turnover may be decreased in departments with competent managers that assist and mentor their staff. Poor performers could be more likely to be fired without cause, which would account for such departments' higher involuntary termination rates compared to other departments.

## Analysis 4.2 How the duration of service and the method of termination relate

An employee is often more likely to obtain extra perks and protections upon termination the longer they have worked for a firm. For instance, workers who have been with a business for a specific period of time may be eligible for severance pay, additional healthcare benefits, or other types of payment.

**The technique of analysis**: Data visualization and manipulation

**Source Code:**



#### Figure: Source code for analysis 4.2

**Explanation:**

With the help of ggplot, a bar plot is made to show the link between the length of service and type of termination after the employee data has been filtered to remove those who were not dismissed. The plot illustrates the proportion of working careers for workers who were terminated under each form of termination and utilizes facet\_wrap to construct distinct plots for each type of termination. The "employee\_data" data frame contains information on terminated workers, and this code examines the link between the length of service and termination type for those individuals.

**Output:**

#### Figure: Output for analysis 4.2

**Explanation of output:**

In comparison to the plot for involuntary termination, the plot for voluntary termination displays a larger count of dismissed workers for longer working years. This suggests that workers who left the firm freely are likely to have stayed with it for a longer period of time. We can spot any variations or patterns in the distribution of working years among terminated workers under voluntary and involuntary termination by viewing the data in this way. In this instance, we can observe that employees who have been with the organization for a longer amount of time tend to experience voluntary termination more frequently.

**Output Findings:**

There could be a number of reasons for the observation that workers who have been with the company for a longer time tend to encounter voluntary termination more frequently. A desire for new experiences or a change of scenery, a desire for a better work-life balance, or a sense of financial security that makes them feel more at ease leaving are a few examples of these. Changes inside the corporation, such as adjustments to the leadership or corporate culture, may also be responsible for this pattern. To ascertain the precise causes underlying the patterns in the output that have been detected, more analysis and context are required.

## Extra Features

**Labs():**

The axis labels of a graph can be created or changed using the R function labs(). It may be used to decide on titles for the graph's X and Y axes as well as its overall title. After the development of charts with the well-known R data visualization tool ggplot2, the labs() function is commonly used.

**Theme():**

The R function theme() is used to alter the visual appearance of graphs created by the popular ggplot2 R data visualization tool.

The theme() method may be used in conjunction with the ggplot() function to change the graph's visual components in a variety of ways, such as the text color and size, line thickness, tick marks, etc., to create distinctive aesthetics.

## Conclusion

In conclusion, the activities stated above highlight the important role that R programming plays in data processing and provide a variety of tools and approaches for effective and thorough solutions. Starting with data exploration and including preprocessing techniques like cleaning and standardization, the mission included every stage of the data analysis process. Also, the assignment showed how to use sophisticated data preparation methods like subsetting and reshaping, which are essential for idea creation. The visuals provided by R programming were used to quickly identify trends and associated early warning indicators. Ultimately, this project emphasizes the value of R programming as a tool for academics and data analysts who rely on information-driven judgment. Using cutting-edge data exploration, preparation, processing, andvisualization technologies can lead to in-depth insights.

# References

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag.

Hadley, W. (2017). stringr: Simple, consistent wrappers for common string operations. R package version 1.2.0. <https://CRAN.R-project.org/package=stringr>

Wickham, H., & Henry, L. (2019). tidyr: Tidy messy data. R package version 1.0.0. <https://CRAN.R-project.org/package=tidyr>

Wickham, H., & Bryan, J. (2019). readr: Read rectangular text data. R package version 1.3.1. <https://CRAN.R-project.org/package=readr>

Wickham, H., & François, R. (2016). dplyr: A grammar of data manipulation. R package version 0.7.6. <https://CRAN.R-project.org/package=dplyr>

Wickham, H. (2019). tidyr 1.0.0. Journal of Open Source Software, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

Wickham, H., & Grolemund, G. (2017). R for data science: Import, tidy, transform, visualize, and model data. O'Reilly Media, Inc

Xie, Y. (2015). Dynamic documents with R and knitr (2nd ed.). Chapman and Hall/CRC.

Kuhn, M. (2020). caret: Classification and regression training. R package version 6.0-86. <https://CRAN.R-project.org/package=caret>

Cairn.info. (n.d.). cairn.info. <https://www.cairn.info/revue-management-2016-3-page-177.htm>

Data cleaning in R. (2022, June 27). GeeksforGeeks. <https://www.geeksforgeeks.org/data-cleaning-in-r/>

Data Preprocessing in R. (n.d.). Engineering Education (EngEd) Program | Section. <https://www.section.io/engineering-education/data-preprocessing-in-r/>

Ellis, L. (2018, November 5). Explore your dataset in R. R-bloggers. <https://www.r-bloggers.com/2018/11/explore-your-dataset-in-r/>