A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

INDIVIDUAL ASSIGNMENT
TECHNOLOGY PARK MALAYSIA
CT127-3-2-PFDA
PROGRAMMING FOR DATA ANAYLSIS
APD2F2302CS

HAND OUT DATE:

HAND IN DATE:

WEIGHTAGE: 50%

Name: Ndiwe Simao Mabote

TP No: 067277

INSTRUCTIONS TO CANDITATES:

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

Table of Contents

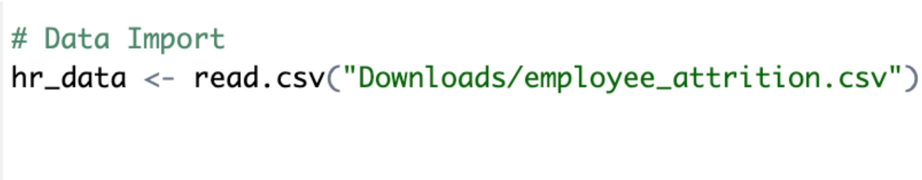
Introduction and Assumption

Data analysis is a crucial method used to uncover insights, make knowledgeable decisions, and conclude large volumes of quantitative data. It has a vital role in helping both companies and individuals describe and summarize their data effectively. In recent years the increasing availability of data and the popularity of data analysis has allowed organizations to strive to understand and extract value. As a result, data analysis has emerged as a vital tool for discovering potential hidden within data. (Coursera, 2023)

In this project, my goal is to investigate and identify the underlying issues affecting the HR department of this company by utilizing data analysis methods. To accomplish this task, I will employ various techniques including data manipulation, visualization, exploration, and transformation. This dataset contains information about employees' information that can be used to uncover the hidden challenges within the HR department. The results will be analyzed and justified using graphs, furthermore, each graph will have an R programming supporting document.

Based on the available information, it is anticipated that the challenges within the HR department may come from various factors, such as a high number of employee resignations or retirements. Additionally, it is possible that the company has experienced significant layoffs, contributing to the HR issues. To address these concerns, data manipulation techniques will be employed to extract pertinent information from the dataset. By conducting data exploration, a deeper understanding of the dataset will be obtained, enabling the identification of crucial details. Furthermore, through data transformation, key insights within the data can be revealed. Finally, employing data visualization techniques will facilitate the interpretation of our findings and enable us to draw meaningful conclusions.

Data Import

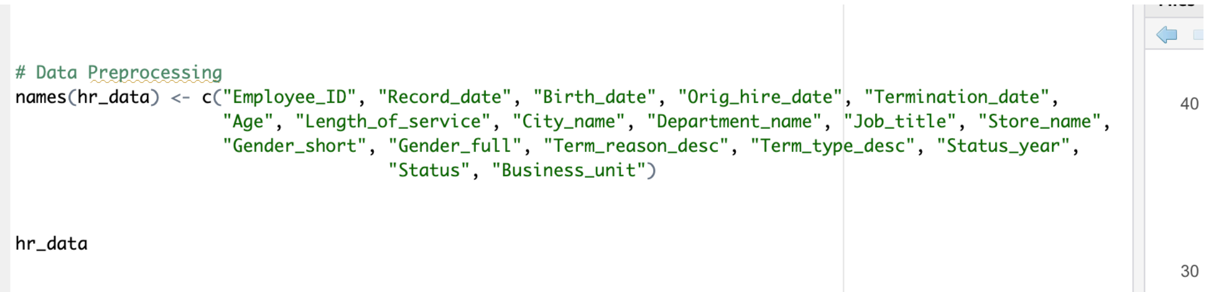


```
# Data Import
hr_data <- read.csv("Downloads/employee_attrition.csv")
```

Figure 1

As shown in Figure 1. The first step taken is to import the dataset that is going to be analyzed. I have assigned this dataset to “hr_data” so that I can access this information later. Then I loaded the hr_data to ensure to make sure that data has been successfully loaded.

Data Preprocessing

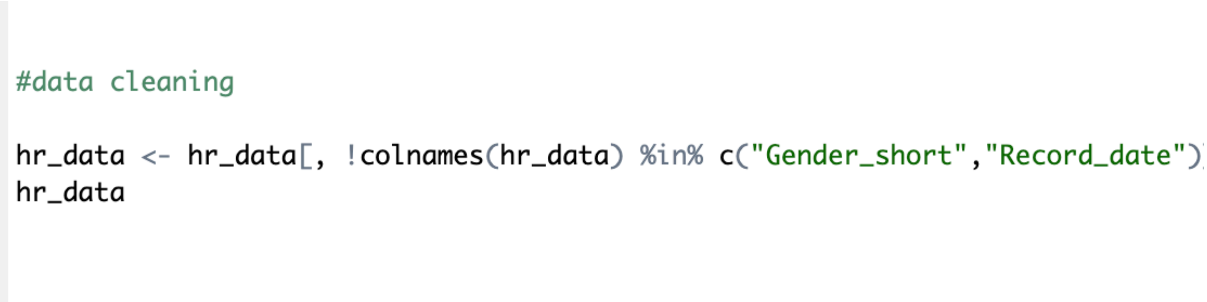


```
# Data Preprocessing
names(hr_data) <- c("Employee_ID", "Record_date", "Birth_date", "Orig_hire_date", "Termination_date",
                    "Age", "Length_of_service", "City_name", "Department_name", "Job_title", "Store_name",
                    "Gender_short", "Gender_full", "Term_reason_desc", "Term_type_desc", "Status_year",
                                    "Status", "Business_unit")


hr_data
```

Figure 2

My data preprocessing is done as shown in figure 2, as I Renamed all my columns to make them easier to accessible and more memorable. Then I loaded the data to make sure that the modifications to the headers were successfully saved and incorporated into the data.

Data Cleaning

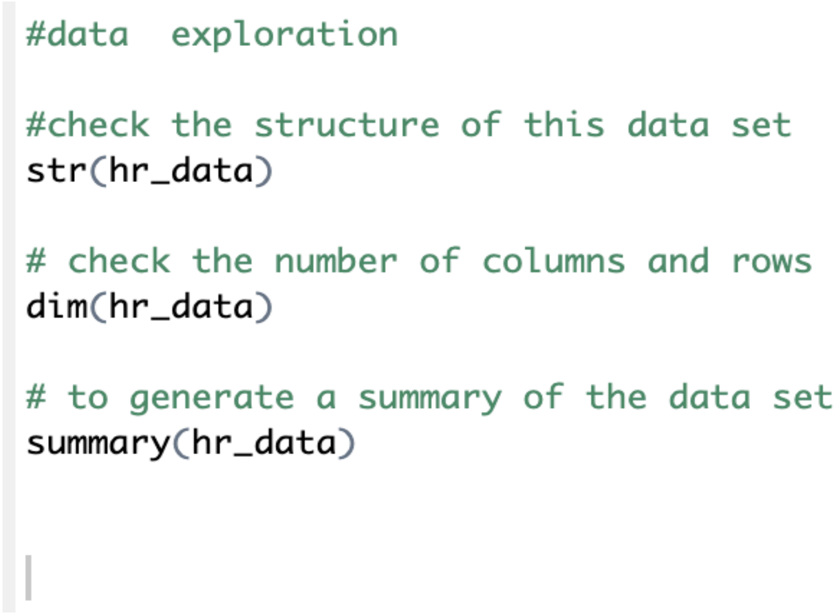


```
#data cleaning

hr_data <- hr_data[, !colnames(hr_data) %in% c("Gender_short","Record_date")
hr_data
```

Figure 3

As shown in figure 3 for my data cleaning process, I removed two columns that include gender_short as I found it redundant since there is already another column give us the exact same information and record_date as this does not give any useful information for the analysis that are going to be made.

Data Exploration



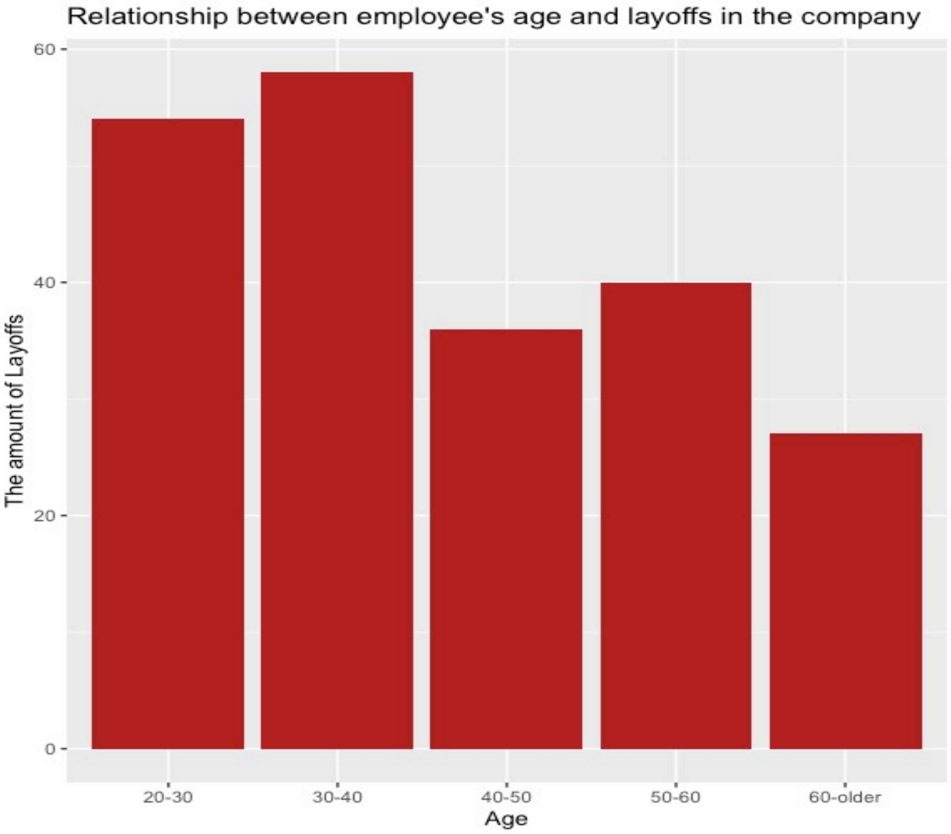
```
#data  exploration

#check the structure of this data set
str(hr_data)

# check the number of columns and rows
dim(hr_data)

# to generate a summary of the data set
summary(hr_data)
```

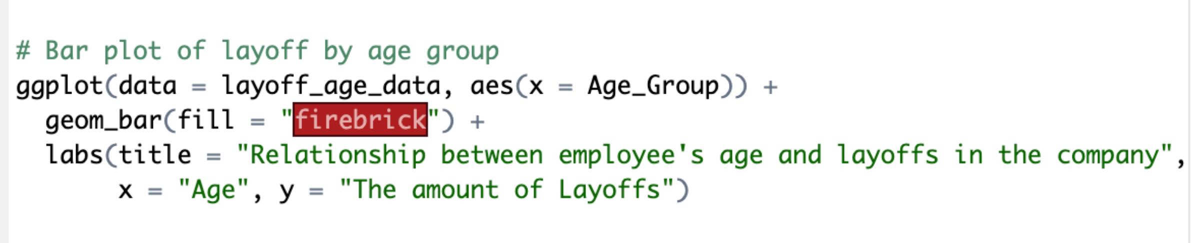
Figure 4

As shown in figure 4, I used a couple of data exploration techniques to better understand the dataset given. First, I used “str” function to check the structure of this data set, then I used the “dim” function to the check numbers of columns and rows and finally I used the “summary” so that I gain a general understanding of the dataset.

Question 1: What is the factor that affects employee's layoff?

Analysis 1-1: The relationship between employee's age and layoffs in the company

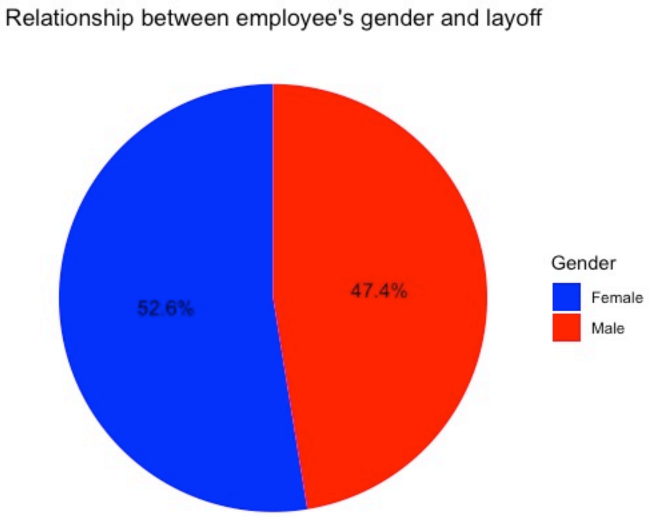


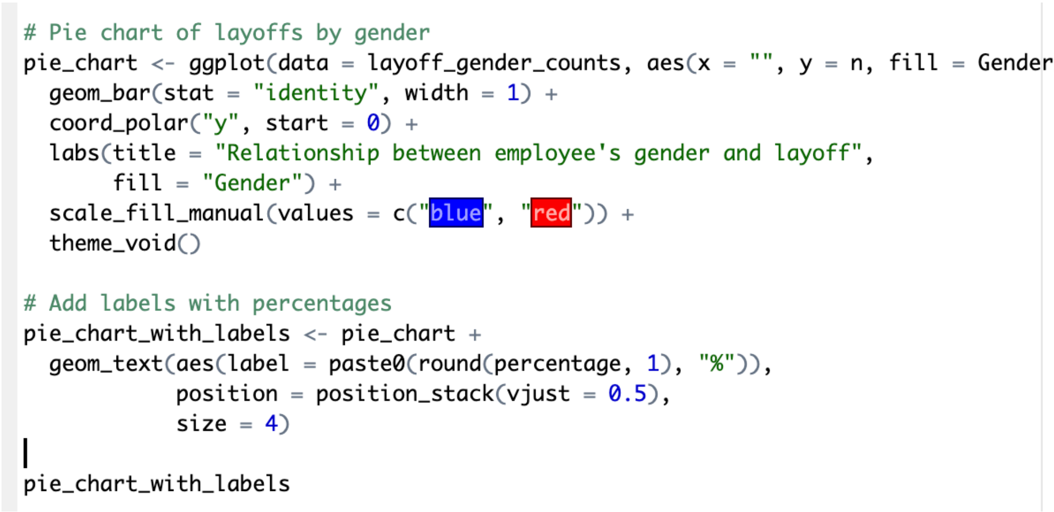


```
# Bar plot of layoff by age group
ggplot(data = layoff_age_data, aes(x = Age_Group)) +
  geom_bar(fill = "firebrick") +
  labs(title = "Relationship between employee's age and layoffs in the company",
       x = "Age", y = "The amount of Layoffs")
```

The bar plot above illustrates the relationship between employees' age and layoffs. This analysis aims to examine whether there was a slight trend favoring a specific age group during the HR department's decision-making process for layoffs. From the graph, it is evident that the HR department tended to lay off younger workers. Both the "20-30" and "30-40" age groups experienced more than 40 layoffs, while age groups above 40 years old had fewer than 40 layoffs. Although the difference in layoffs between older and younger age groups is not significant, it can still be concluded that age played a role in the HR department's decision to lay off these individuals. This raises the concern that the company should focus on retaining younger workers, as they can contribute to productivity over a longer period. However, it should be noted that the difference in layoffs is not substantial, suggesting that the HR department may have prioritized experience over longevity.

Analysis 1-2: The relationship between employee's gender and layoffs in the company

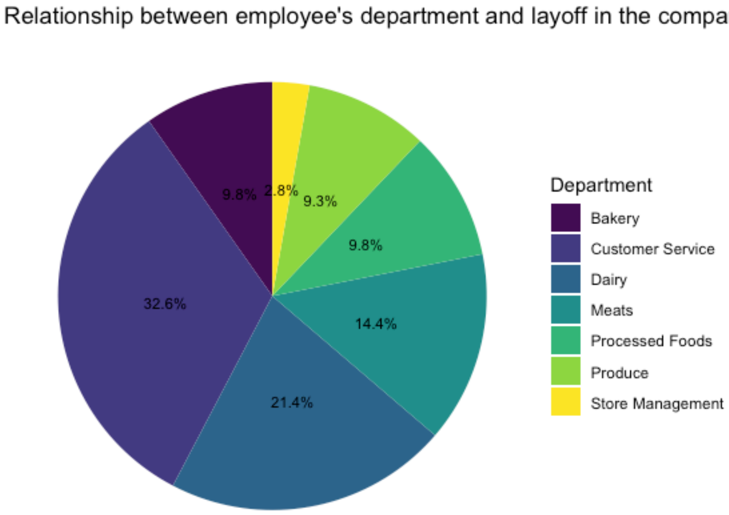


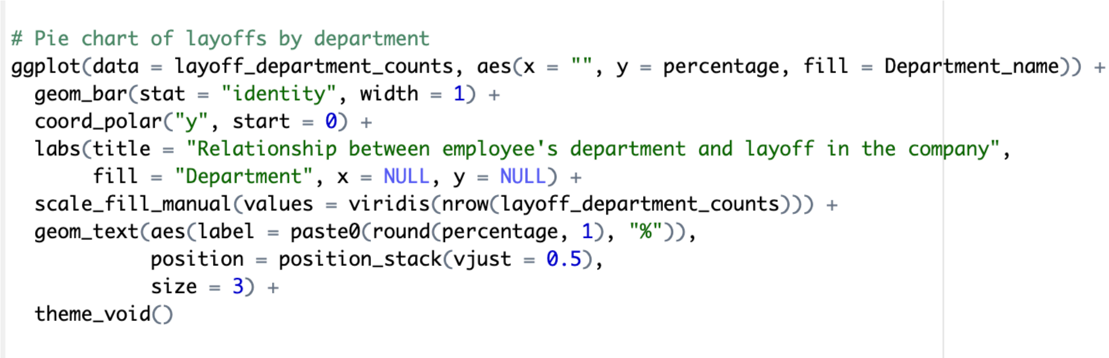


```
# Pie chart of layoffs by gender
pie_chart <- ggplot(data = layoff_gender_counts, aes(x = "", y = n, fill = Gender_full)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Relationship between employee's gender and layoff",
       fill = "Gender") +
  scale_fill_manual(values = c("blue", "red")) +
  theme_void()

# Add labels with percentages
pie_chart_with_labels <- pie_chart +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5),
            size = 4)

pie_chart_with_labels
```

As shown in the pie chart above, 52.6% of the workers that were dismissed were male and 47.4% were female. This pie chart was done to ensure that the HR department did not have bias when coming to these decisions but due to the difference in the layoffs not being significant, it can be concluded that the HR department did not have any bias when coming to this decision.

Analysis 1-3: The relationship between an employee’s department and layoff

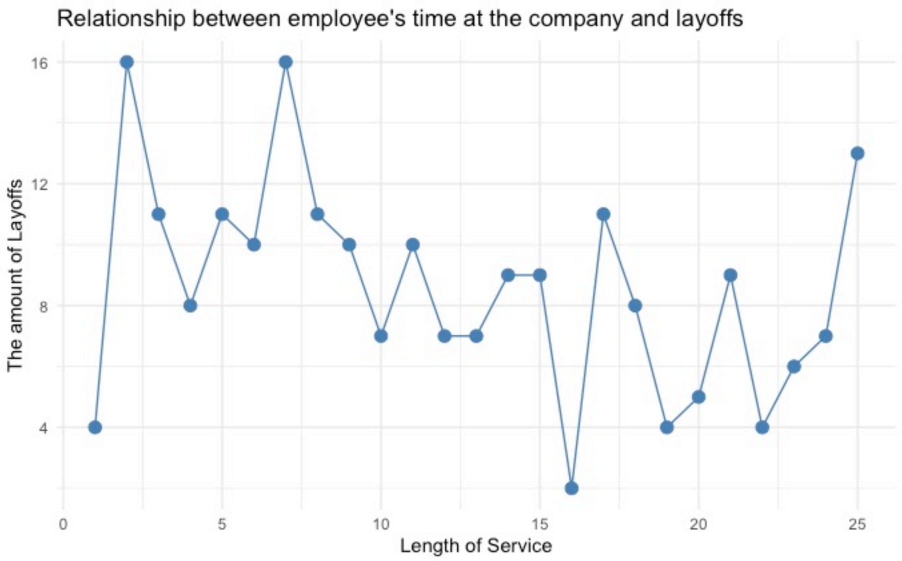


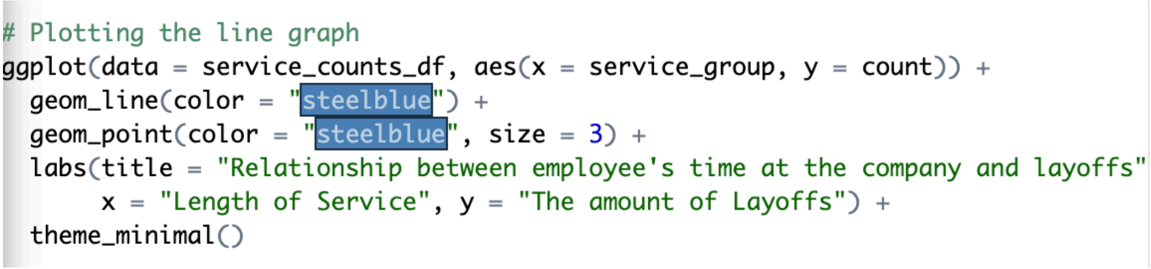


```r
# Pie chart of layoffs by department
ggplot(data = layoff_department_counts, aes(x = "", y = percentage, fill = Department_name)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Relationship between employee's department and layoff in the company",
       fill = "Department", x = NULL, y = NULL) +
  scale_fill_manual(values = viridis(nrow(layoff_department_counts))) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5),
            size = 3) +
  theme_void()
```

As shown in the pie chart above, the departments with the most layoffs include customer service at 32.6%, Dairy at 21.4%, and meats at 14.4%. This pie chart was made to examine whether there were departments that were targeted by these layoffs. From the pie chart it can be concluded that the customer service and dairy departments were the most affected by this, they were targeted by the HR department as the number of layoffs in these two departments have been substantially larger than the other departments that got laid offed. It can be concluded that HR kept the department in mind when deciding to lay off their workers.

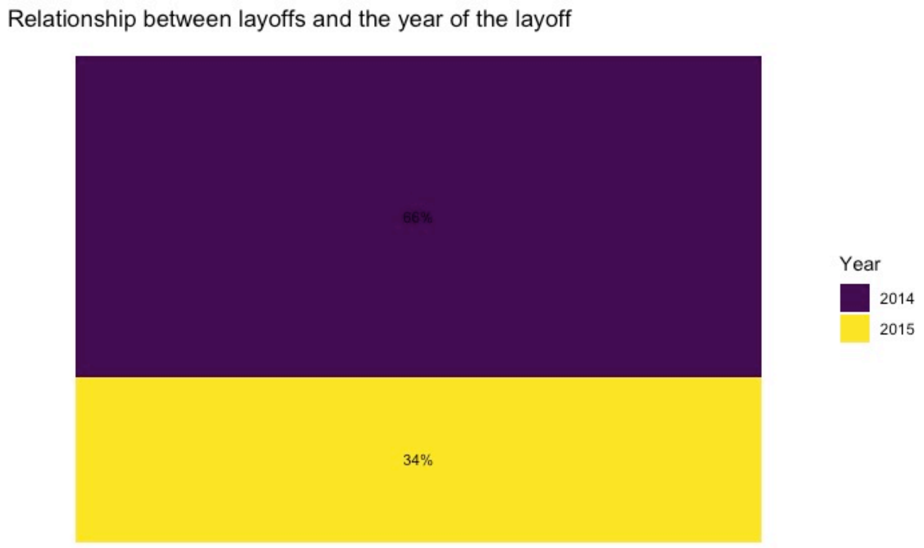
Analysis 1-4: Relationship between employee's time at the company and layoffs

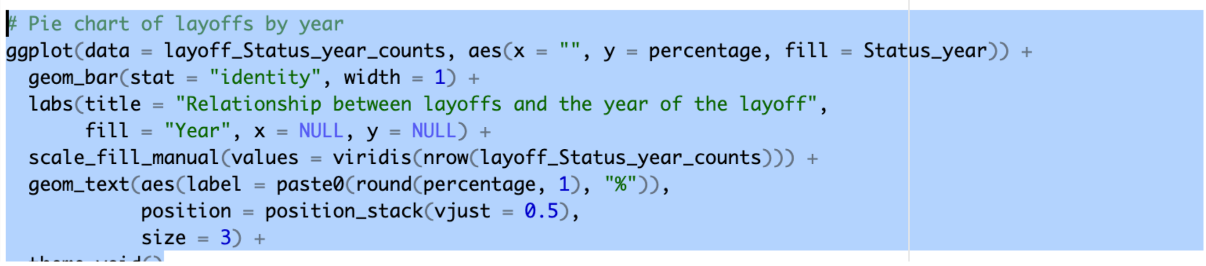



```
# Plotting the line graph
ggplot(data = service_counts_df, aes(x = service_group, y = count)) +
  geom_line(color = "steelblue") +
  geom_point(color = "steelblue", size = 3) +
  labs(title = "Relationship between employee's time at the company and layoffs",
       x = "Length of Service", y = "The amount of Layoffs") +
  theme_minimal()
```

As shown in this line graph above, workers that had been in the company had more layoffs than workers that had been there for 10 or more years but workers that had worked for more than ten years were also subject to more layoffs. This graph was done with the purpose of examining whether employees' time at the company influenced the decision of the HR department when laying off individuals. We can conclude that this was the case, and it was an excellent approach where they tended to favor people who have not been in the company for a long time, but they also did not favor new workers, they chose to keep experienced workers in the company.

Analysis 1-5: Relationship between layoffs and the year of layoffs



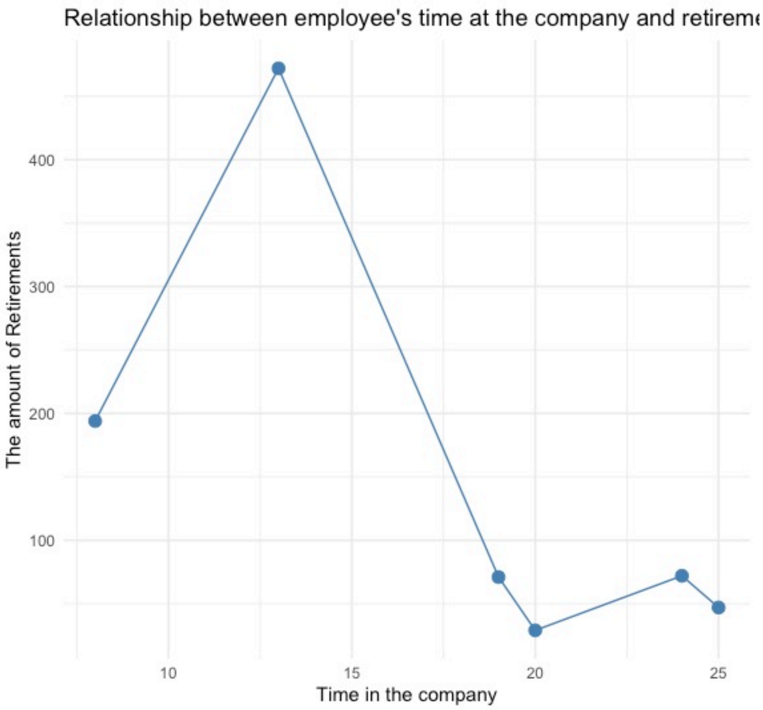


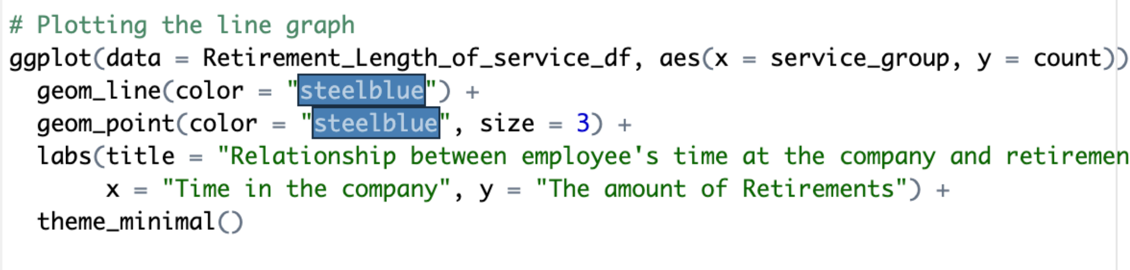
```
# Pie chart of layoffs by year
ggplot(data = layoff_Status_year_counts, aes(x = "", y = percentage, fill = Status_year)) +
  geom_bar(stat = "identity", width = 1) +
  labs(title = "Relationship between layoffs and the year of the layoff",
       fill = "Year", x = NULL, y = NULL) +
  scale_fill_manual(values = viridis(nrow(layoff_Status_year_counts))) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5),
            size = 3) +
  theme_void()
```

As this stack bar graph shows, most of the layoffs were done in 2014 as 66% of the workers were laid off this year and the rest in 2015, this is interesting as the company only has two workers which can indicate that they may be some financial problems in the companies in these years. This analysis aims to examine whether the year of the layoff played a part in the layoffs. We can conclude that this is the case as all the layoffs happened in just two years and the majority in 2014.

Question 2: What is the factor that affects employee’s retirement?

Analysis 2-1: Relationship between employee's time at the company and retirement in the company

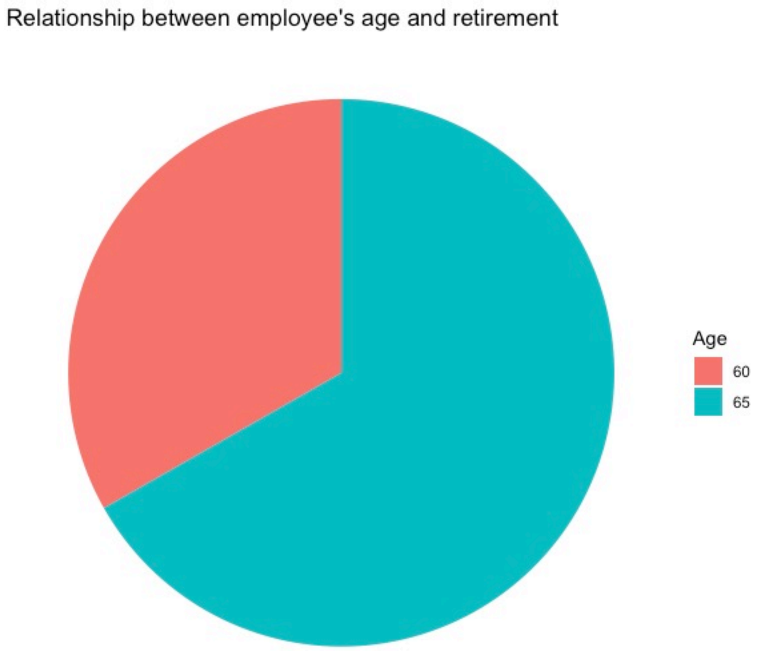


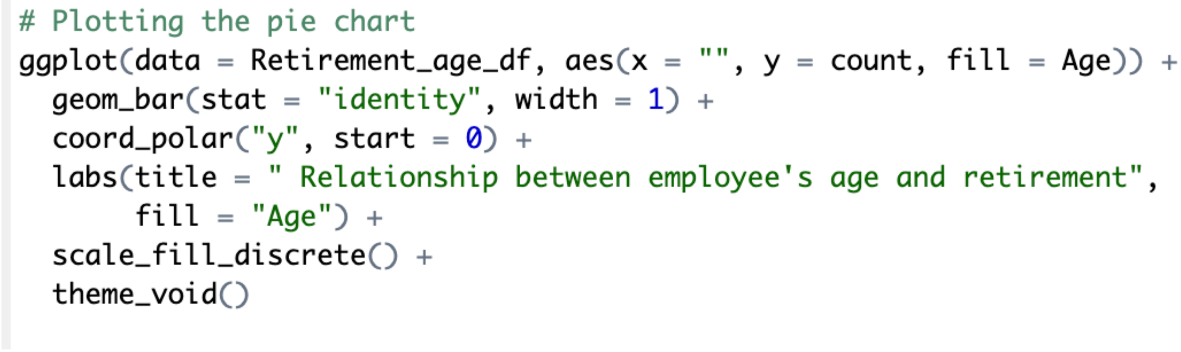


```
# Plotting the line graph
ggplot(data = Retirement_Length_of_service_df, aes(x = service_group, y = count)) +
  geom_line(color = "steelblue") +
  geom_point(color = "steelblue", size = 3) +
  labs(title = "Relationship between employee's time at the company and retirement",
       x = "Time in the company", y = "The amount of Retirements") +
  theme_minimal()
```

As the line graph above shows, most of the workers that retired worked at the company for at least 10 years but not more than 20 years. This graph aims to determine how long workers that retire in the company work. Although ten years is a long time, it is concerning that most of the workers in the company tend to retire before they have 20 years in the company, this is something that the HR department should do as it reflects that the company is either hiring old workers or that they are not keeping their young workers for a substantial amount of time.

Analysis 2-2: Relationship between employee's age and retirement in the company



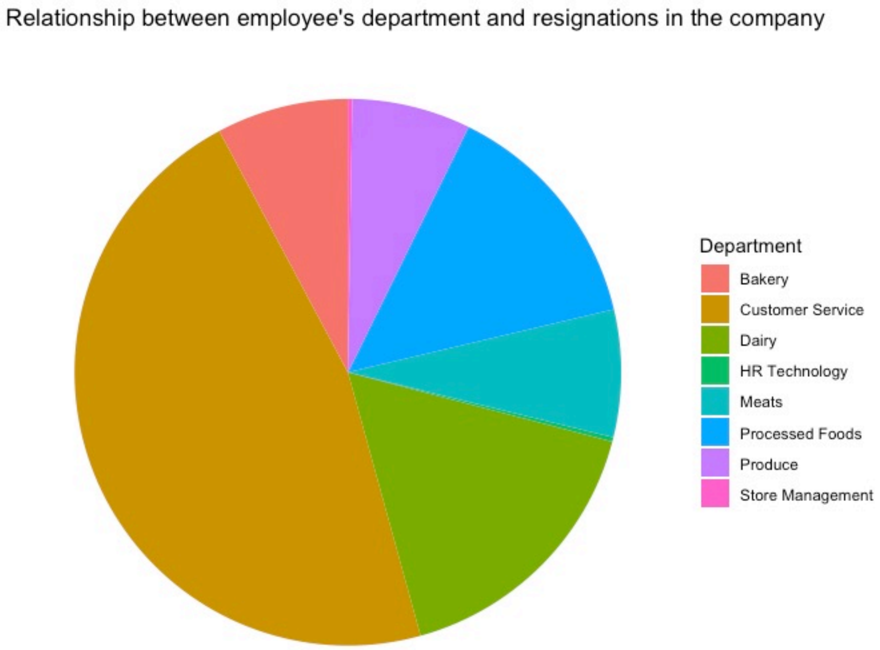


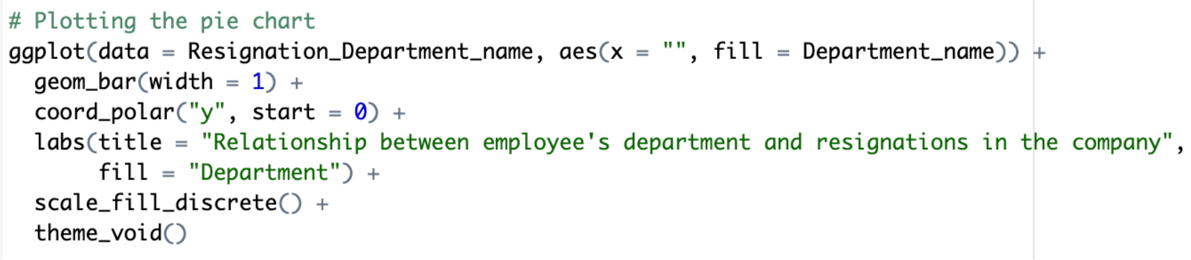
```
# Plotting the pie chart
ggplot(data = Retirement_age_df, aes(x = "", y = count, fill = Age)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = " Relationship between employee's age and retirement",
       fill = "Age") +
  scale_fill_discrete() +
  theme_void()
```

As this pie graph above shows, most of the workers in the company retire when turn 65 years old. This chart aims to determine the most common age at which the workers in this company retire. This chart by itself only shows the age at which workers retire but if it is contextualized by the previous it can see that the workers that retire in the company have been hired either in their 40s or 50s which furthers the point that the company should try to employ younger workers in the company.

Question 3: What is the factor that affects employee’s resignation?

Analysis 3-1: Relationship between employee's department and resignations in the company

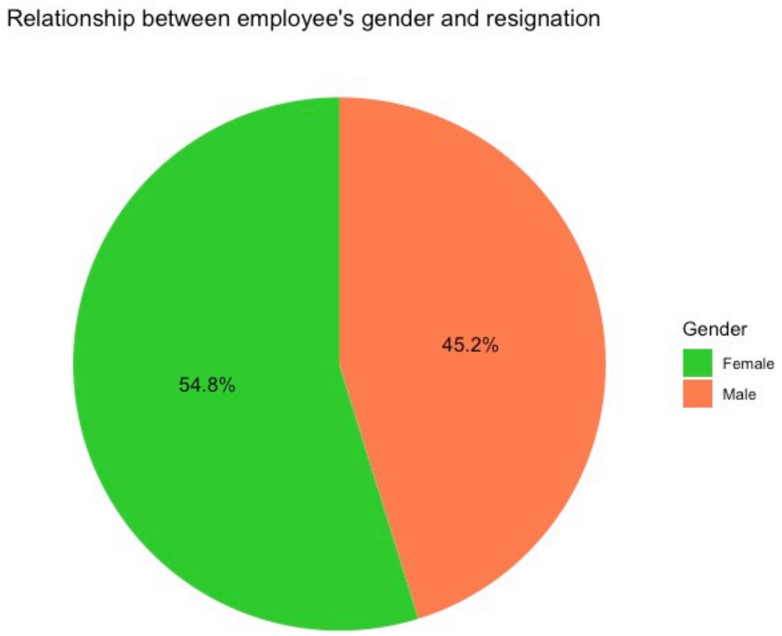


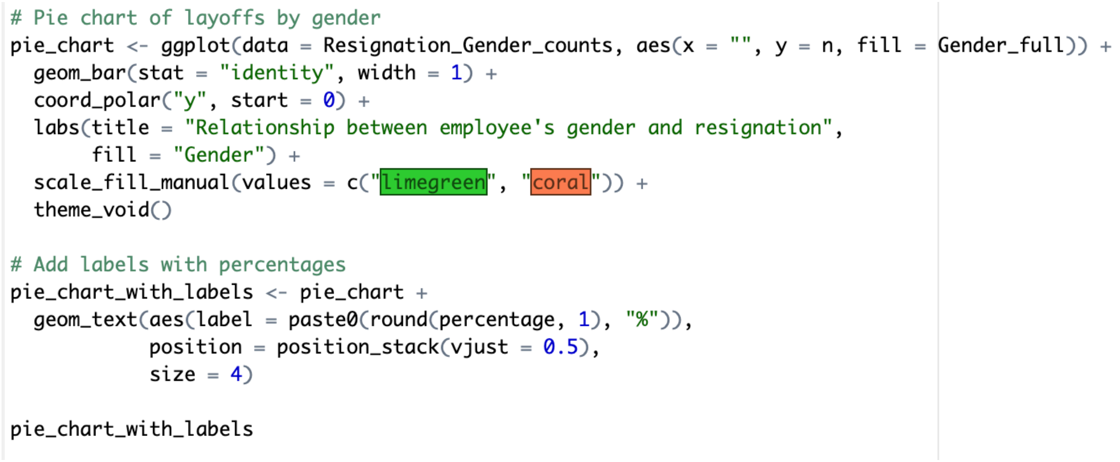


```
# Plotting the pie chart
ggplot(data = Resignation_Department_name, aes(x = "", fill = Department_name)) +
  geom_bar(width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Relationship between employee's department and resignations in the company",
       fill = "Department") +
  scale_fill_discrete() +
  theme_void()
```

As shown in this pie chart, most of the workers that decided to resign are part of the customer service department as this has the most resignations, with HR technology at second and Processed foods at third. This chart aims to illustrate which department has the most resignations, which can indicate which department needs to improve. From this graph, we can conclude that departments such as customer service, HR technology, and processed foods need to be looked at by the HR department as most of the workers that resign are working in these departments.

Analysis 3-2: Relationship between employee's gender and resignation

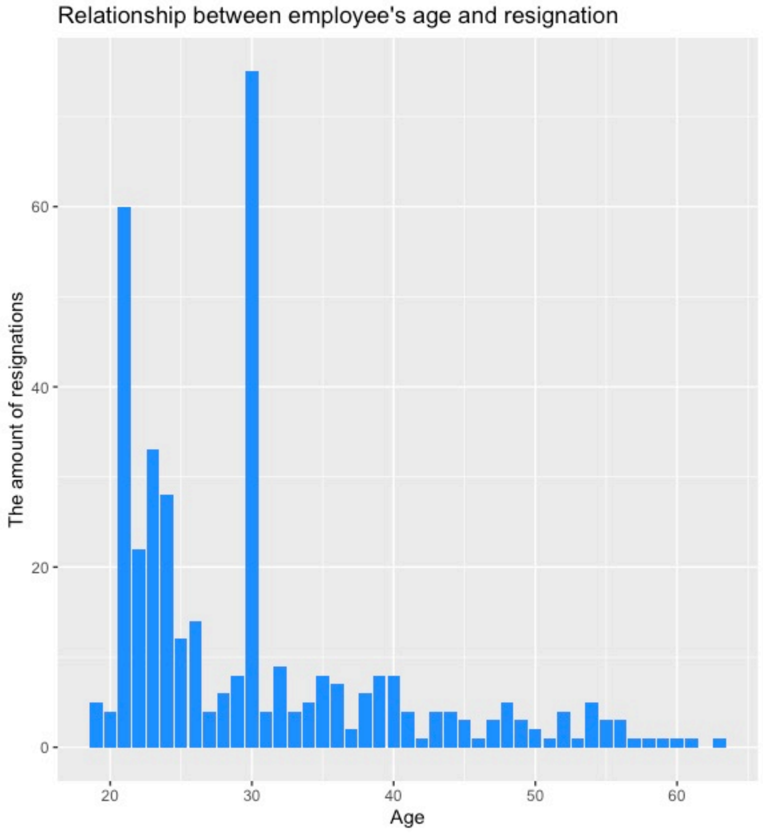


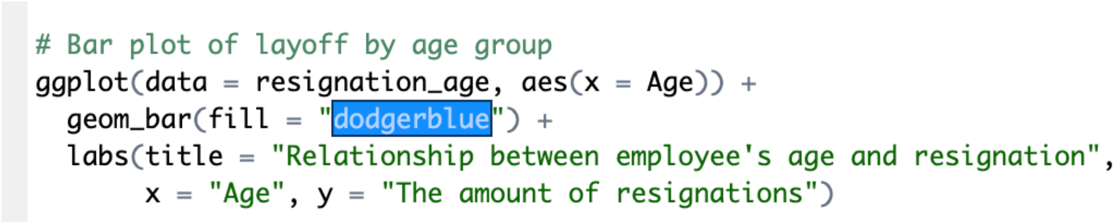


```
# Pie chart of layoffs by gender
pie_chart <- ggplot(data = Resignation_Gender_counts, aes(x = "", y = n, fill = Gender_full)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Relationship between employee's gender and resignation",
       fill = "Gender") +
  scale_fill_manual(values = c("limegreen", "coral")) +
  theme_void()

# Add labels with percentages
pie_chart_with_labels <- pie_chart +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5),
            size = 4)

pie_chart_with_labels
```

As this chart shows, most of the individuals are women with 54.8 percent which is not a substantial amount. This chart aimed to examine if there was maybe foul play in the company, because if gender was to have a substantial number of layoffs it would indicate that there is something that the HR department is either ignoring or not seeing something. This is not the case as the split between both genders is relatively even and t which means that the company has nothing to worry about in this department.

Analysis 3-3: Relationship between employee's age and resignation

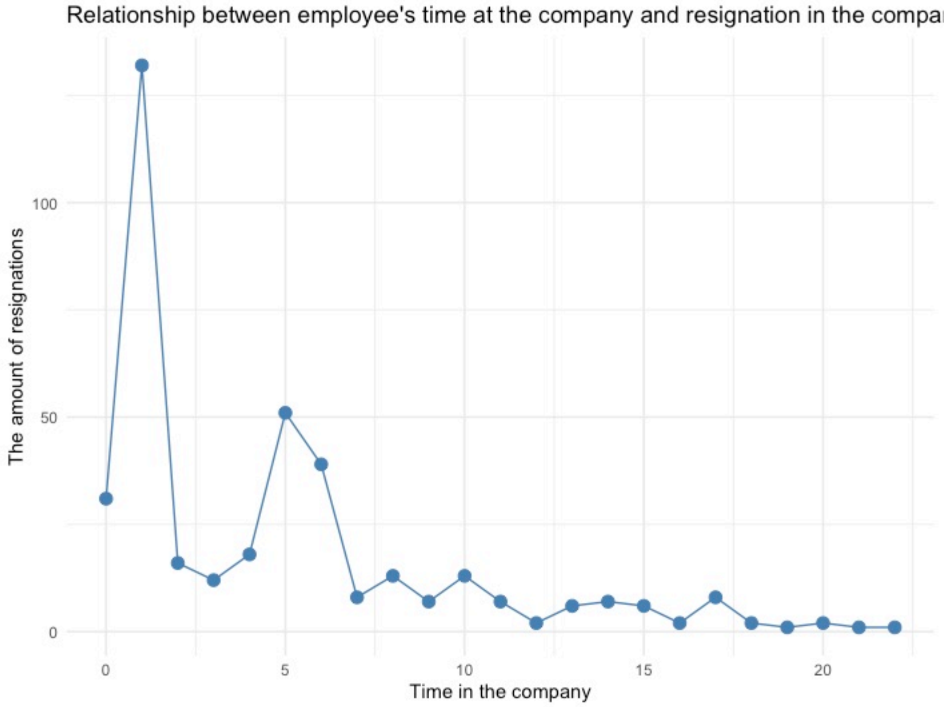


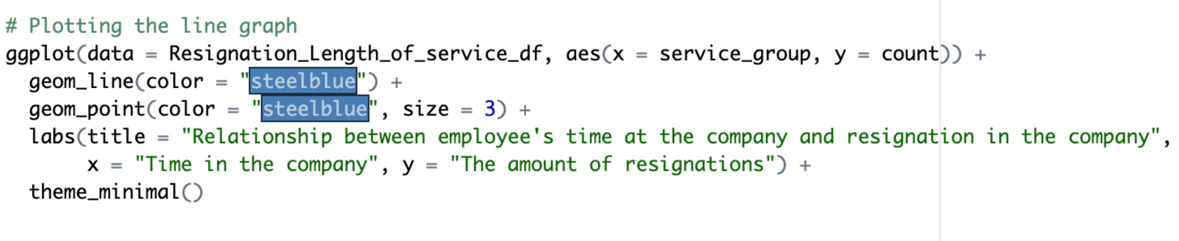


```
# Bar plot of layoff by age group
ggplot(data = resignation_age, aes(x = Age)) +
  geom_bar(fill = "dodgerblue") +
  labs(title = "Relationship between employee's age and resignation",
       x = "Age", y = "The amount of resignations")
```

As shown in the bar plot above, there is a higher frequency of retirement among workers younger than 40 years old compared to other age groups. The purpose of this graph is to show the age distribution of resigning workers. Based on the graph, a notable observation is that a significant number of resigning workers are young. This highlights an issue that the HR department should address, as efforts should be made to attract and retain more young workers in the company.

Analysis 3-4: Relationship between employee's time at the company and Resignation



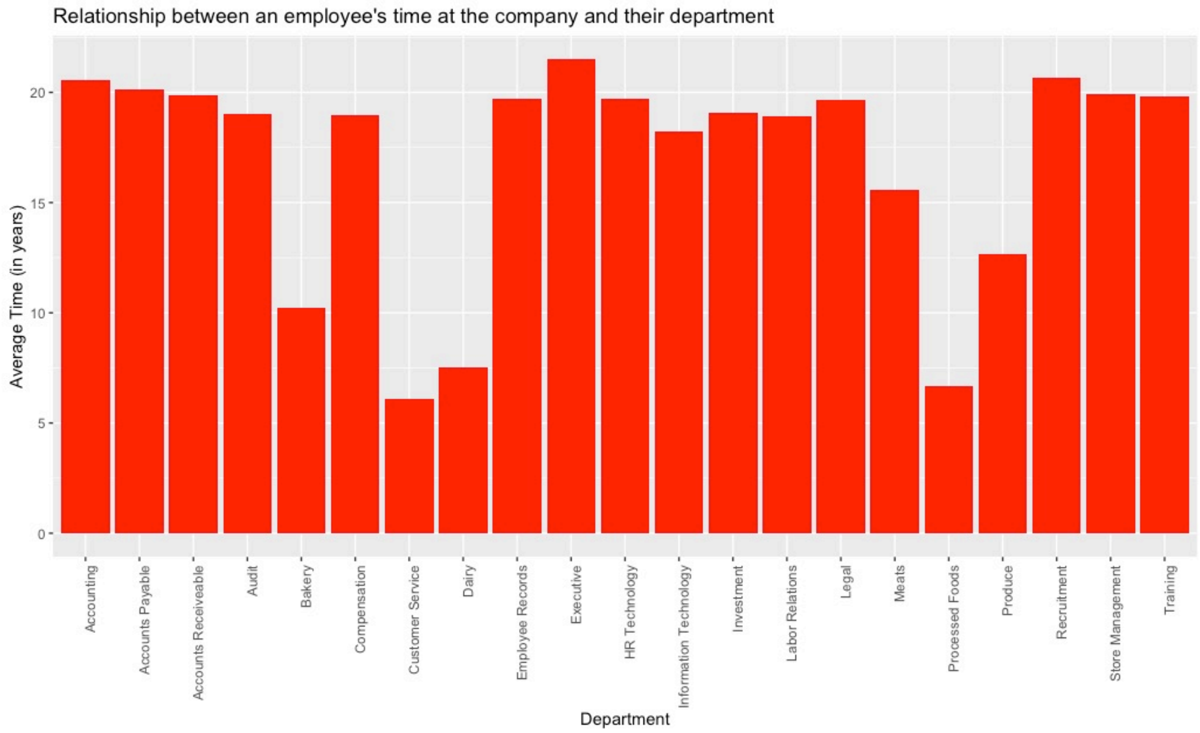


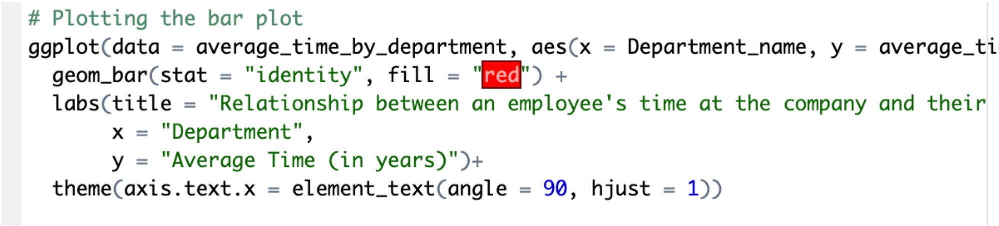
```
# Plotting the line graph
ggplot(data = Resignation_Length_of_service_df, aes(x = service_group, y = count)) +
  geom_line(color = "steelblue") +
  geom_point(color = "steelblue", size = 3) +
  labs(title = "Relationship between employee's time at the company and resignation in the company",
       x = "Time in the company", y = "The amount of resignations") +
  theme_minimal()
```

As shown line graph, most of the workers that have resigned have worked for the company for less than 5 years. The graph aims to illustrate the number of years which workers that resigned stayed in the company. We can conclude that most of the workers that resigned did not work in the company for a long time, this is something that the HR department must investigate as it is alarming, and it will lead to the company always having to hire new employees which is not conducive to growth in a company.

Question 4: What is the factor that affect employee's length of time at the company?

Analysis 4-1: Relationship between an employee's time at the company and their department

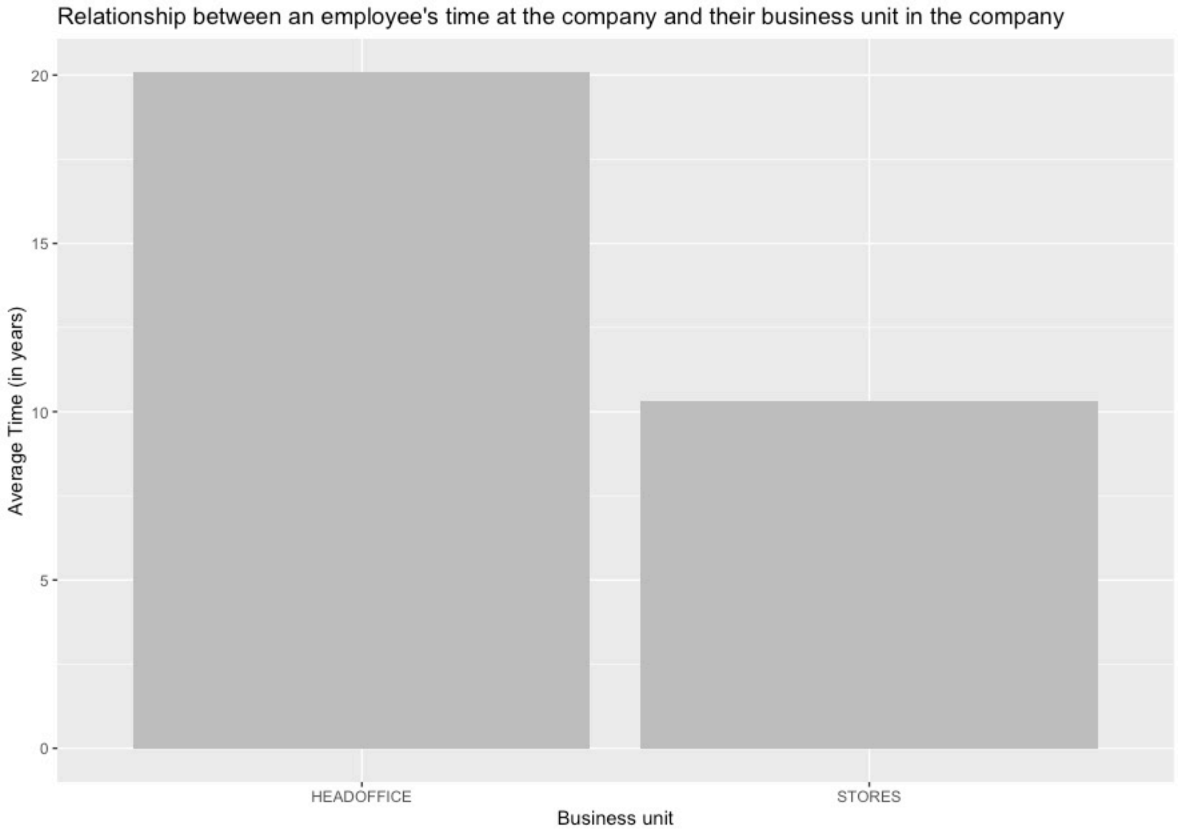


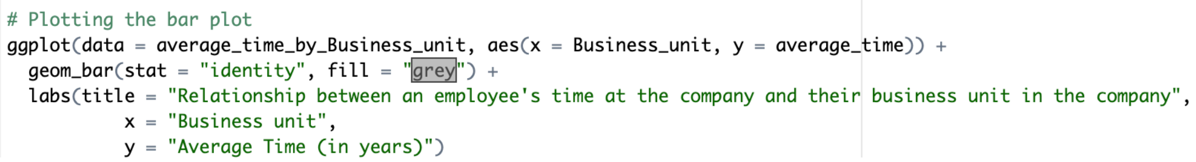


```
# Plotting the bar plot
ggplot(data = average_time_by_department, aes(x = Department_name, y = average_time)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(title = "Relationship between an employee's time at the company and their department",
       x = "Department",
       y = "Average Time (in years)")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

As shown in the bar graph above, this graph shows that most of the departments have workers that stay for more than 15 but there is a couple of departments that do not even reach ten years. This graph is aimed to show compare how long department can keep their workers for. We can conclude that processed foods, customer service, and dairy are unable to retain their workers for as long as the other departments, The HR department should investigate this and ensure that they put measures to improve these departments.

Analysis 4-2: Relationship between an employee's time at the company and their business unit in the company

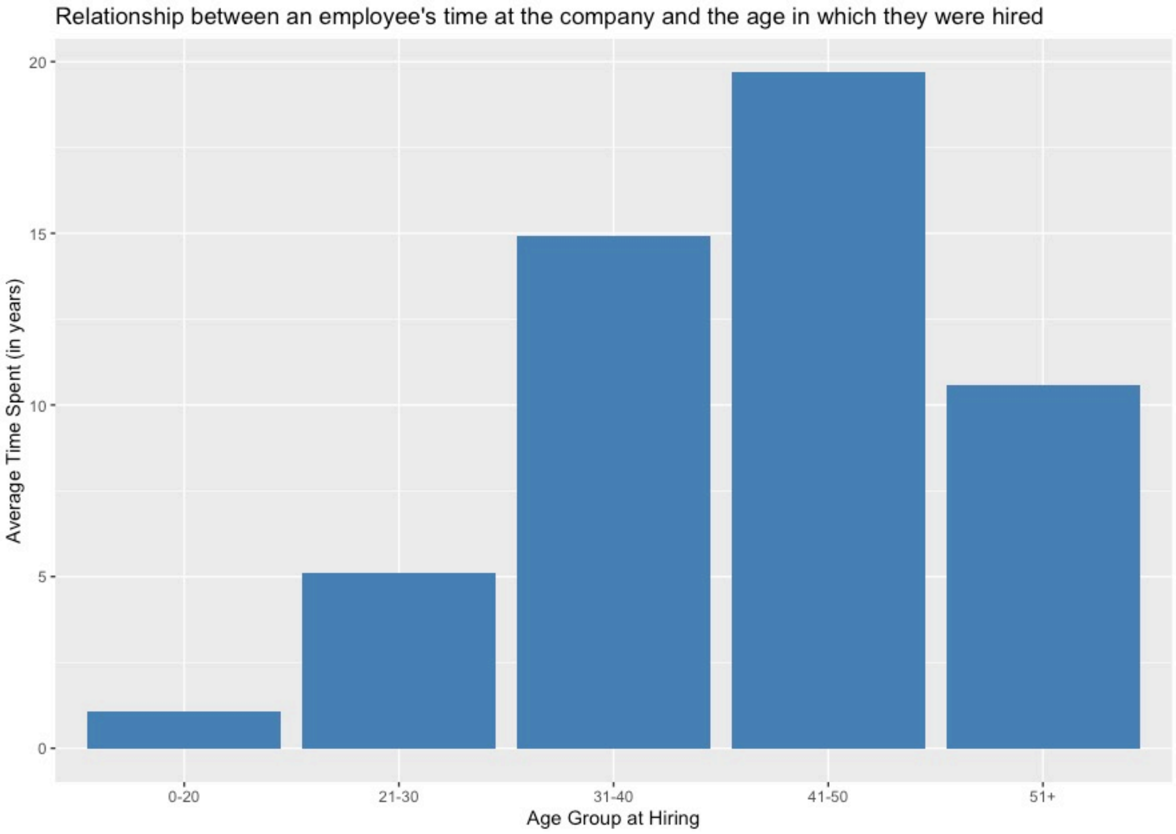


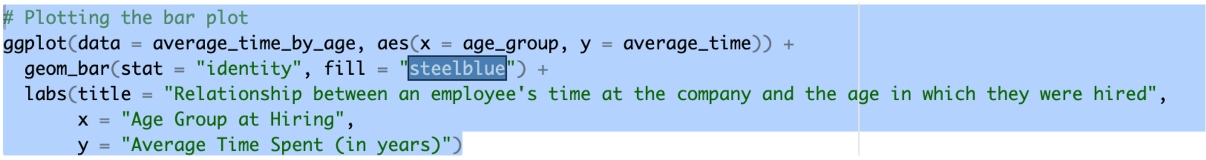


```
# Plotting the bar plot
ggplot(data = average_time_by_Business_unit, aes(x = Business_unit, y = average_time)) +
  geom_bar(stat = "identity", fill = "grey") +
  labs(title = "Relationship between an employee's time at the company and their business unit in the company",
           x = "Business unit",
           y = "Average Time (in years)")
```

As shown in the graph above, the difference between the Head office and stores in terms of the years that their workers spent working for each department is substantial. This graph aims to examine whether there is a substantial difference between the business units of the company. This graph shows that the Head office tends to stay longer than workers at the stores, this is something that the HR department should look at, as the year spans to 10 years apart, which is alarming.

Analysis 4-3: Relationship between an employee's time at the company and the age in which they were hired



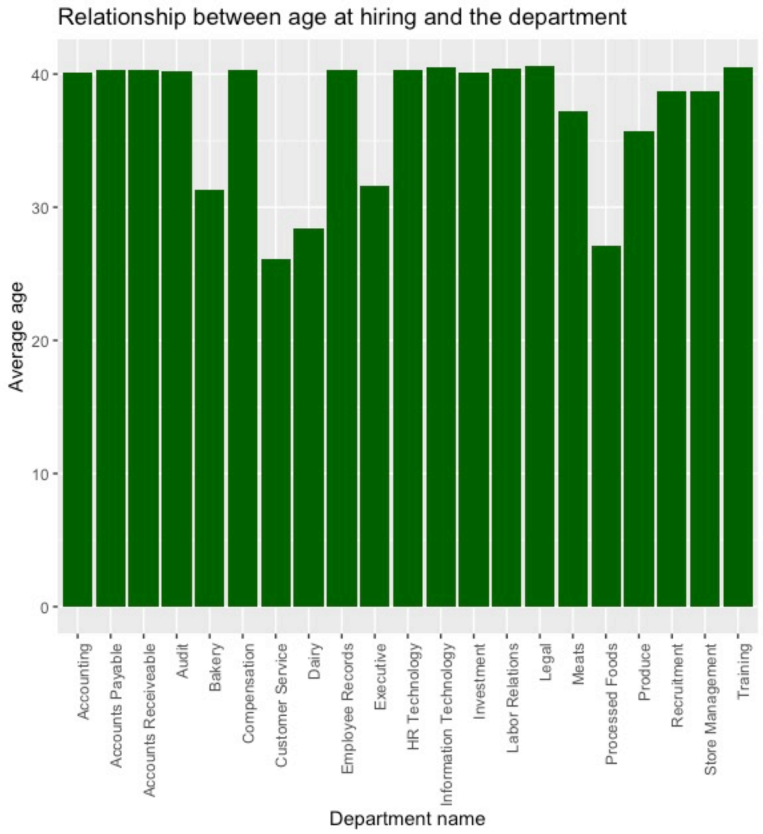


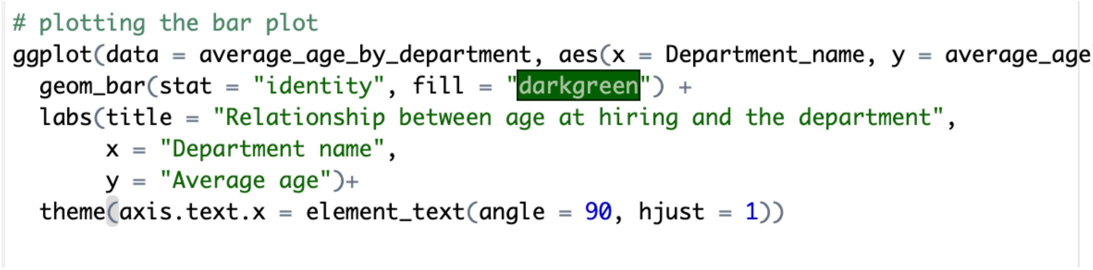
```
# Plotting the bar plot
ggplot(data = average_time_by_age, aes(x = age_group, y = average_time)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Relationship between an employee's time at the company and the age in which they were hired",
       x = "Age Group at Hiring",
       y = "Average Time Spent (in years)")
```

As shown in the graph above, most of the workers that are older than 40 tend to stay in the company while younger workers usually leave tend to stay on for less time. This graph is aimed to determine whether their young employees tend to stay. The answer to this is no, they do not, it seems that most of their young workers leave early. The HR department must look at this, as the company should make new efforts to retain its young workforce.

Question 5: Does a candidate’s age affect company hiring policy?

Analysis 5-1: Relationship between age at hiring and the department

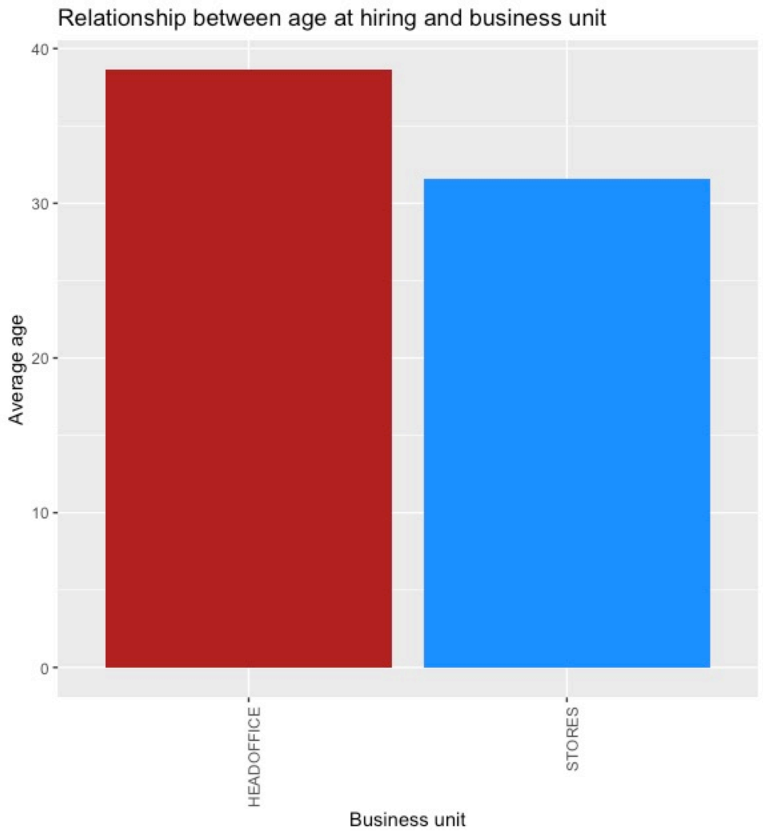


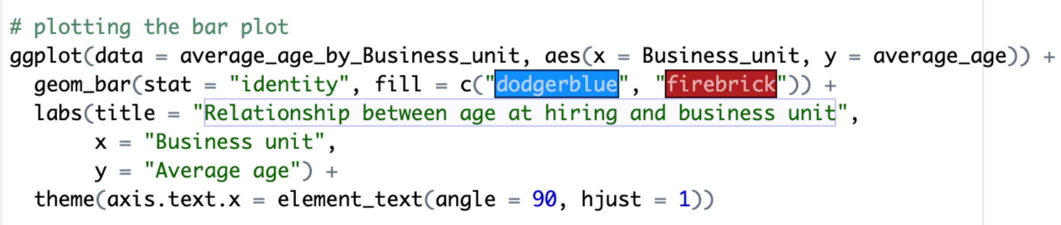


```
# plotting the bar plot
ggplot(data = average_age_by_department, aes(x = Department_name, y = average_age)) +
  geom_bar(stat = "identity", fill = "darkgreen") +
  labs(title = "Relationship between age at hiring and the department",
       x = "Department name",
       y = "Average age")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

As shown in the graph above, the most of departments hire workers around the mid to late 30s. This graph aims to examine whether the HR department favors a specific age group. It can be concluded that is indeed the case as most of the departments tend to hire an older demographic and this is concerning as the company needs to look to bring in a younger demographic of workers.

Analysis 5-2: Relationship between age at hiring and business unit



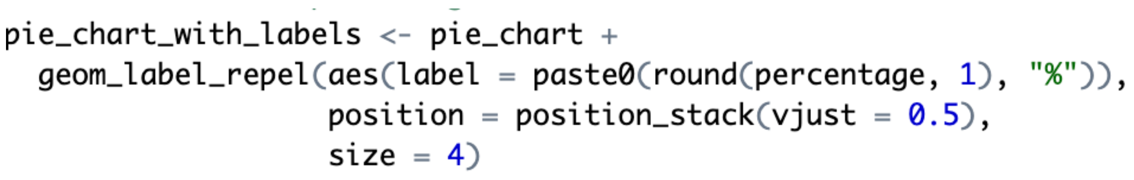


```
# plotting the bar plot
ggplot(data = average_age_by_Business_unit, aes(x = Business_unit, y = average_age)) +
  geom_bar(stat = "identity", fill = c("dodgerblue", "firebrick")) +
  labs(title = "Relationship between age at hiring and business unit",
       x = "Business unit",
       y = "Average age") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

As shown in the graph above, Head offices tend to hire workers around their late 30s while stores hire them in their early 30s. This graph aims to analyze whether the HR department tends to favor older individuals for jobs at the head office. The answer to that is yes, they do but although this is the case, it can be argued that is the right since companies tend to need older individuals to lead their company.

Additional features

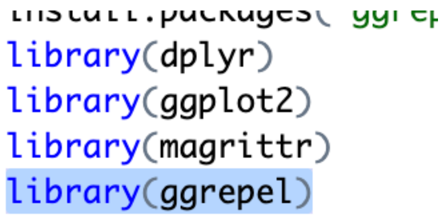
1. Geom_label_repel



```
pie_chart_with_labels <- pie_chart +
  geom_label_repel(aes(label = paste0(round(percentage, 1), "%")),
                   position = position_stack(vjust = 0.5),
                   size = 4)
```

You can add text labels to a plot using the function, and they will be automatically repelled away from other labels or overlapping points. This keeps labels from becoming too crowded and guarantees that all labels can still be read. (Turner, 2016)

It accessed using the library(ggrepel)



```
install.packages("ggrep
library(dplyr)
library(ggplot2)
library(magrittr)
library(ggrepel)
```

Conclusion

In summary, the data analysis conducted in this project aimed to investigate the underlying issues that affect the HR department of the company. By using various data analysis techniques such as data manipulation, visualization, exploration, and transformation, insights were gained from the dataset containing employee information.

The analysis revealed several key findings. In terms of layoffs, it was observed that younger employees were more likely to be laid off, indicating a potential bias towards retaining more experienced workers. Certain departments, such as customer service and dairy, were disproportionately affected by layoffs, suggesting targeted decision-making.

Regarding retirements, most employees retired after working for the company for 10 to 20 years, highlighting the need to focus on attracting and retaining younger talent. The age at which employees retired was predominantly around 65 years old.

Resignations were more frequent among younger employees, emphasizing the importance of addressing retention strategies for this demographic. Certain departments, including customer service, HR technology, and processed foods, had higher resignation rates.

The length of time employees stayed with the company varied across departments and business units. Departments like processed foods, customer service, and dairy had shorter average tenures compared to others. The HR department should investigate these departments and implement measures to improve retention.

The hiring policy showed a preference for older candidates, particularly for positions at the head office. To maintain a diverse and dynamic workforce, efforts should be made to attract and hire younger candidates.

Overall, the data analysis provided valuable insights into the challenges within the HR department, including issues related to employee layoffs, retirements, resignations, and retention. By addressing these findings, the company can make informed decisions and implement strategies to improve employee satisfaction, retention, and overall performance.

Bibliography

Coursera. (2023, May 22). *What Is Data Analysis? (With Examples).* From Coursera: https://www.coursera.org/articles/what-is-data-analysis-with-examples#

Turner, S. (2016, January 06). *Repel overlapping text labels in ggplot2: R-bloggers.* From R: https://www.r-bloggers.com/2016/01/repel-overlapping-text-labels-in-ggplot2/#:~:text=Enter%20the%20ggrepel%20package%2C%20a,Here%20it%20is%20in%20action.