



ASSIGNMENT
TECHNOLOGY PARK MALAYSIA
CT127-3-2-PFDA
PROGRAMMING FOR DATA ANALYSIS
APU2F2109CS(DA)

MARCELL AGUNG WAHYUDI
TP058650

HAND OUT DATE: 4 OCTOBER 2021

HAND IN DATE: 22 NOVEMBER 2021

WEIGHTAGE: 50%

INSTRUCTIONS TO CANDIDATES:

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

Contents

Introduction	3
Assumption	3
Data Import/Cleaning/Pre-Processing/Transformation	4
Data Analysis	5
Findings.....	20
Conclusion	21
Reference	22

Introduction

For the assignment, you are asked to explore the application of data analytics techniques to the dataset which is provided. You must study data problems related to the dataset, giving special consideration to the unique properties of the problem domain, and testing one or more techniques on it.

Your analysis should be deep and in details, also it must go further than what has already been covered in this course. You must adopt the data Exploration, Manipulation, Transformation and Visualization concepts to guide you through the solution process. It is very important to explain and justify the techniques that have been chosen.

You also may need to pre-process your data to get it into an appropriate format. The assignment should involve a number of techniques by categorize it into different criteria and a detailed exploration with the commands using in each criteria. Outline the findings, analyze them and justify correctly with an appropriate graph. Also, a supporting document is needed to reflect the graph and code using R programming concepts.

This dataset contains the data of staffs within an organization that could determine some hidden issue in human resources management. Human resource department manager assigned you to perform analysis with the given dataset to identify hidden problem in the organization and provide meaningful insight for decision making.

Assumption

It is assumed the termination date of an employee that has a value of 1/1/1900 means the employee hasn't been terminated and still active in the company.

Data Import/Cleaning/Pre-Processing/Transformation

```
#Data Import
empData <- read.csv("C:\\Users\\ASUS ROG\\Documents\\College Stuff\\3rd Semester\\(PFDA)
ProgrammingForDataAnalysis (R)\\Assignment\\employee_attrition.csv")
empData
```

The code above reads the csv file from a local save folder and saves it into 'empData' variable.

```
[empData <- read.csv("C:\\Users\\ASUS ROG\\Documents\\College Stuff\\3rd
Semester\\(PFDA) ProgrammingForDataAnalysis
(R)\\Assignment\\employee_attrition.csv")]
```

```
#Pre-Processing
#To answer the assignment questions, first we must find Staffs who have been terminated
#We can do that by filtering out employees who have termination date of 1/1/1900
#(termination date of 1/1/1900 means still in the company).
empTerminated <- empData %>% filter(terminationdate_key != '1/1/1900')
empTerminated
```

Since the assignment question asks to find the hidden issues within the human resources management, to prepare the data, first, the data is cleaned of employees who are still active (employees who have terminationdate_key of 1/1/1900)

```
#Removing duplicated Employee ID
empUnique <- empData %>% distinct(EmployeeID)
empUnique

#DataFrame for Unique Terminated Employees
empUniqueTerminated <- empTerminated %>% distinct(EmployeeID, .keep_all = TRUE)
empUniqueTerminated
```

Since the dataset has duplicates Employee ID, the code above is to remove duplicates Employee ID.

empUnique is for all Employee.

empUniqueTerminated is for Unique Terminated Employees.

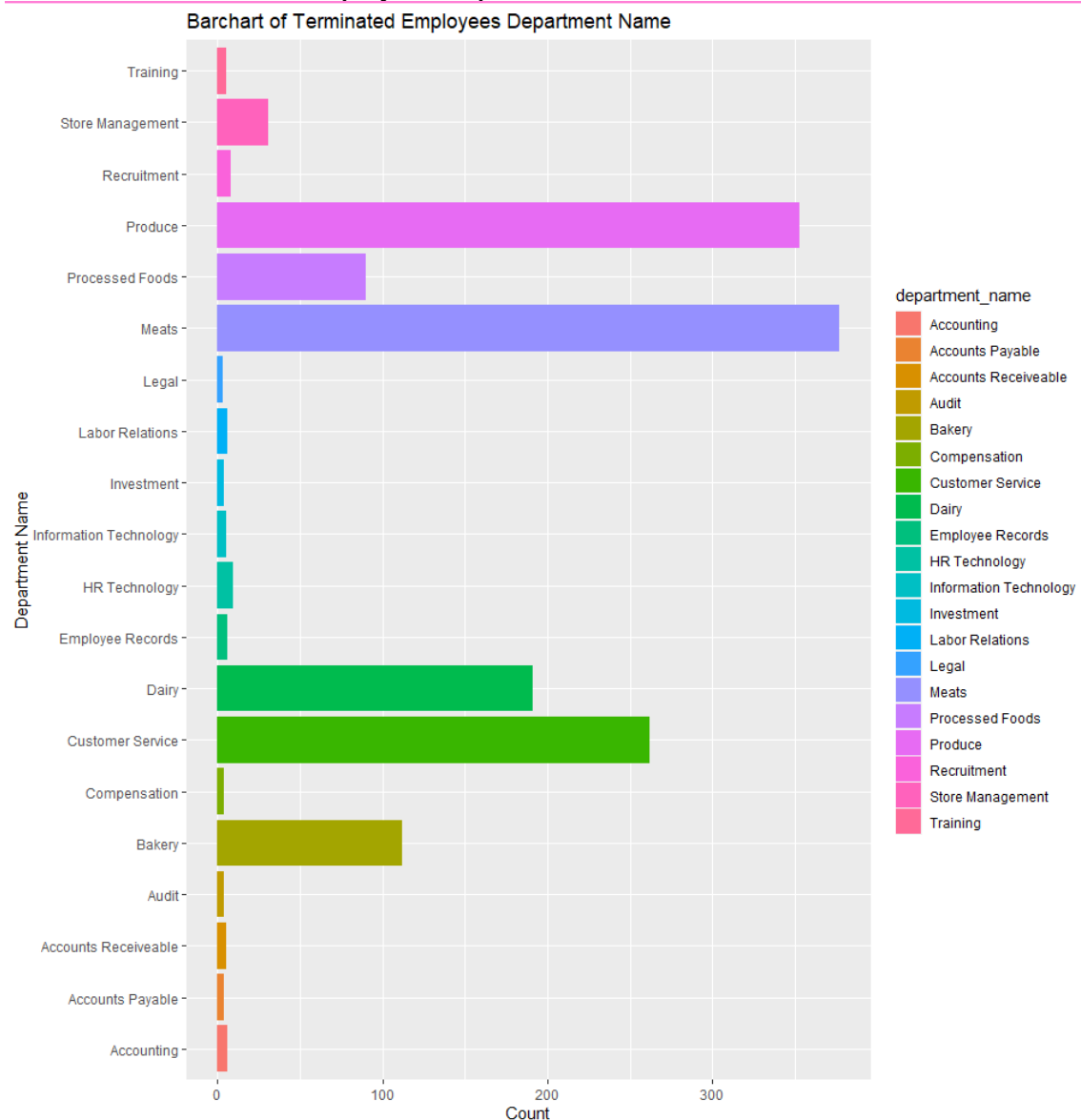
Data Analysis

Question 1. Why would staff leave the Company

Analysis 1-1: Relationship between terminated employees position and attrition

```
#Analysis 1. Analysis of one column (dependent), why would staff leave the Company  
#Analysis 1-1: Relationship between position and attrition  
  
ggplot(data = empUniqueTerminated,aes(y = department_name, fill = department_name)) + geom_bar()+  
labs(title="Barchart of Terminated Employees Department Name", x="Count", y="Department Name")
```

Barchart of Terminated Employees Department Name:



From the data visualization, it can be seen that the most terminated employees that caught attention (ordered from most to least) are from Department: Meats, Produce, Customer Service, Dairy, Bakery, etc.

This may likely be because of Job Dissatisfaction from these departments

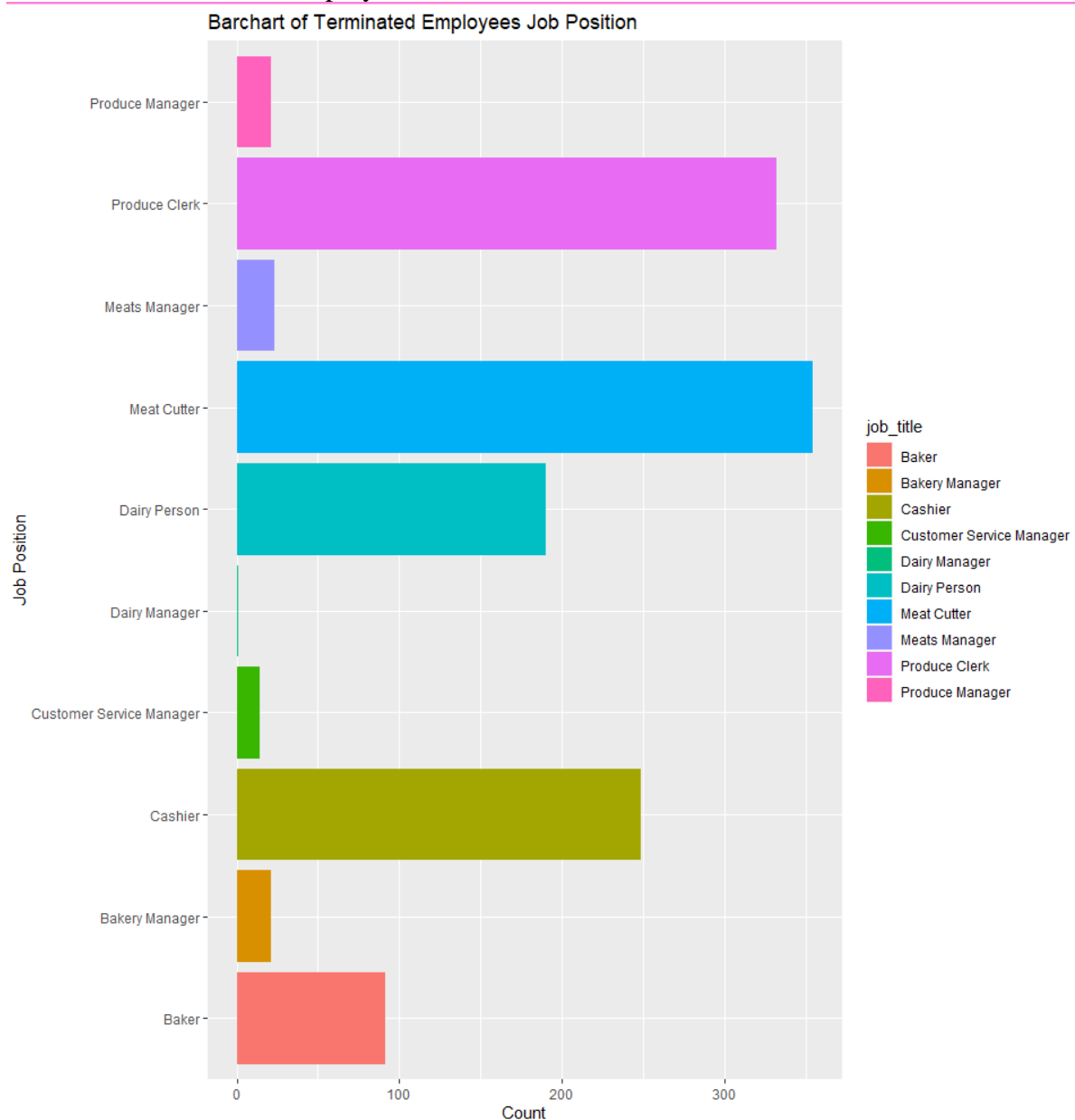
We can dig deeper and find which job positions have the most terminated employees from the previous departments through this code:

```
#Filter out other departments,
DeptNames <- c("Bakery", "Produce", "Customer Service", "Dairy", "Meats")
empUniqueTerminated2 <- empUniqueTerminated %>% filter(department_name %in% DeptNames)
empUniqueTerminated2
```

And then to visualize the data:

```
ggplot(data = empUniqueTerminated2, aes(y = job_title, fill = job_title)) + geom_bar() +
  labs(title="Barchart of Terminated Employees Job Position", y="Job Position", x="Count")
```

Barchart of Terminated Employees Job Position:



From this bar chart, it can be seen that the job position with the most terminated employees (ordered from most to least) are:

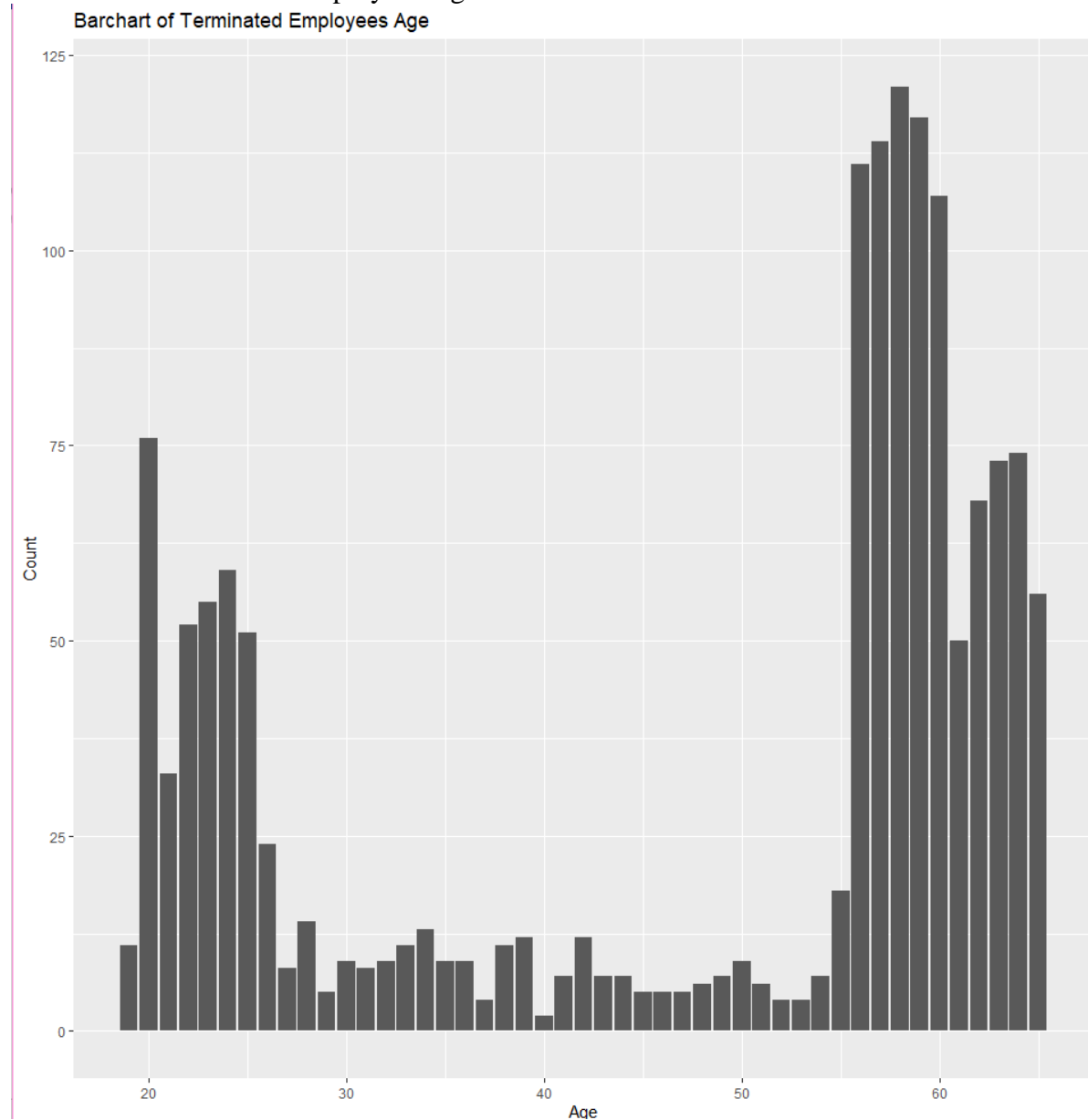
Meat Cutter, Produce Clerk, Cashier, Dairy Person, Baker, etc.

Analysis 1-2: Relationship between terminated age and attrition

Code for data visualization:

```
ggplot(data = empUniqueTerminated, aes(x = age, fill = age)) + geom_bar() +  
  labs(title="Barchart of Terminated Employees Age", x="Age", y="Count")
```

Barchart of Terminated Employees Age:



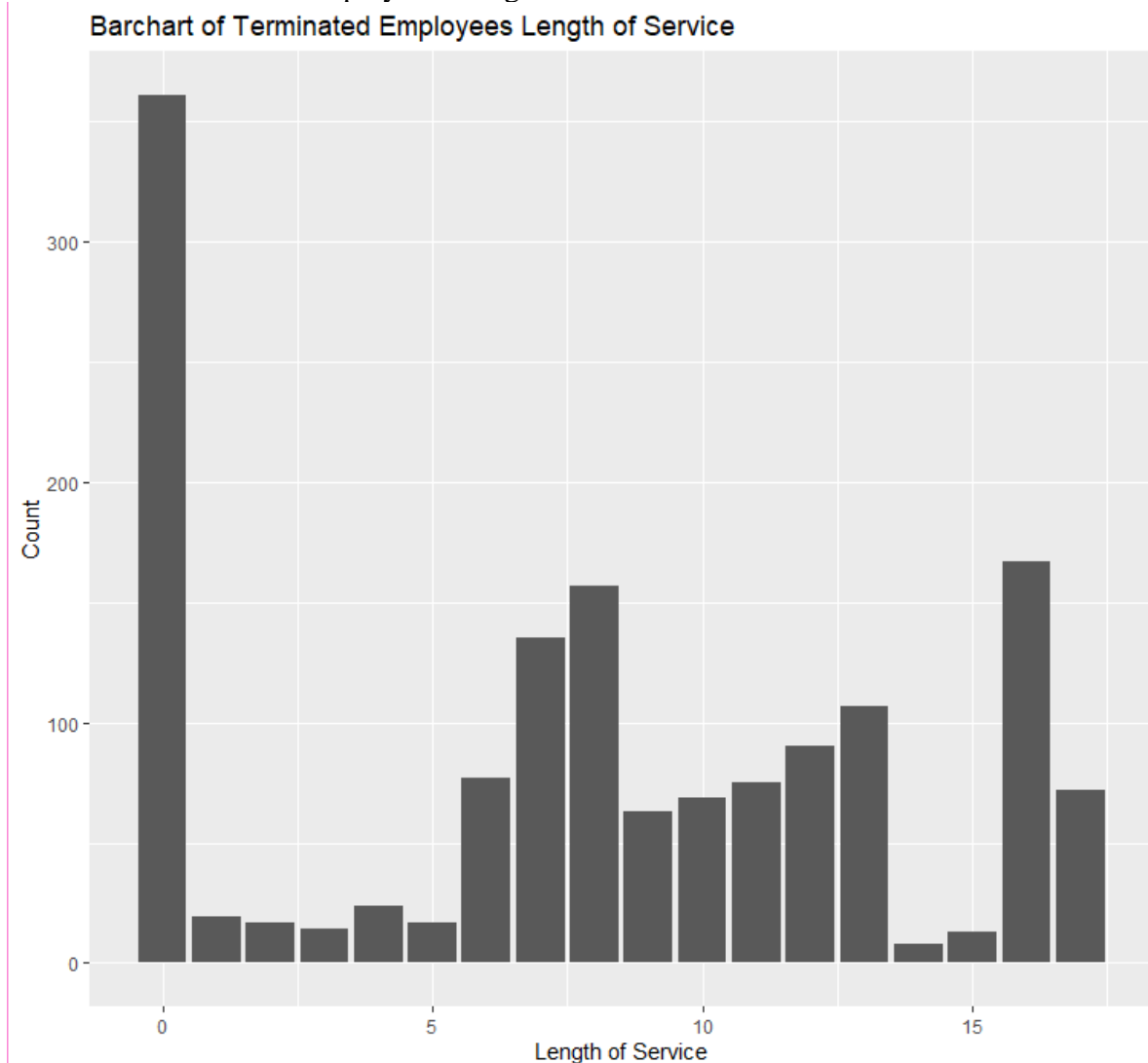
From this bar chart, it can be concluded that the most terminated employees are around 55+, which most likely be the retirement year for them.
the 2nd highest are from people in their 20s till 30 years old, above it, many employees rarely get terminated

Analysis 1-3: Relationship between length_of_service of terminated employees and attrition

Code for data visualization:

```
#Analysis 1-3: Relationship between length_of_service and attrition
ggplot(data = empUniqueTerminated, aes(x = length_of_service, fill = length_of_service)) + geom_bar() +
  labs(title="Barchart of Terminated Employees Length of Service", x="length_of_service", y="Count")
```

Barchart of Terminated Employees' Length Of Service:



From the Bar chart, it can be seen that the most terminated employees served 0 year, the second most terminated is at 16 years of service, also, from graph, it can be seen that the graph dips on the 14th and 15th year, one hypothesis is that the company might reward employees who have worked for 15 years for the company, thus motivating the employees to stay a bit longer.

Analysis 1-4: Job positions of terminated employees age 20-25

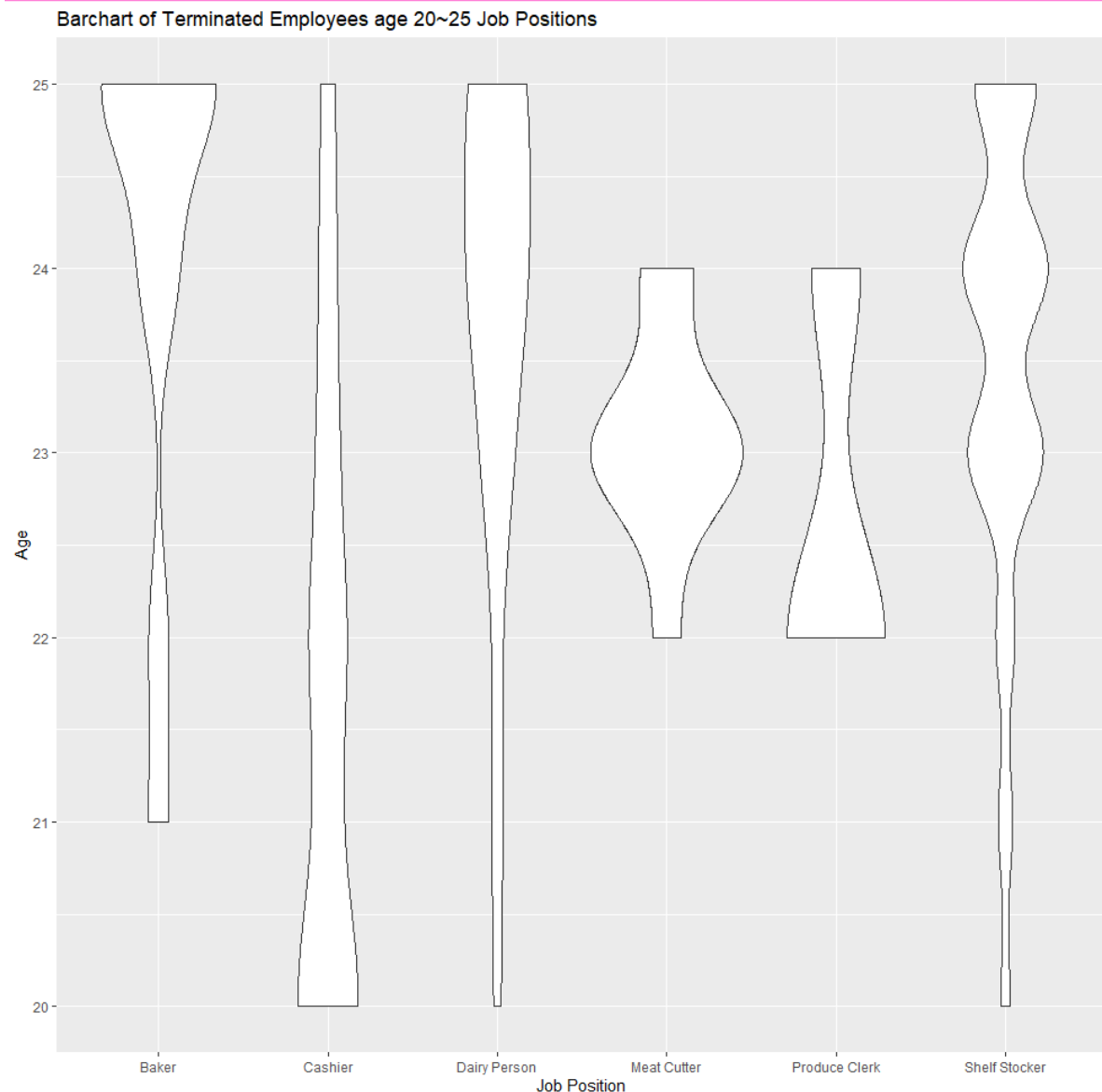
To filter only employees age 20-25, the following code is executed:

```
empYoung <- empUniqueTerminated %>% filter(age > '19' & age < '26')
```

And then, to visualize:

```
ggplot(empYoung, aes(x=job_title, y=age))+  
  geom_violin()+  
  labs(title="Barchart of Terminated Employees age 20~25 Job Positions",  
        x="Job Position", y="Age")
```

Barchart of Terminated Employees age 20~25 Job Positions:



From the visualization, it can be seen that most terminated employees age 20~25 are terminated at 20th years old as a cashier, 23rd as a Meat Cutte, and 25th as a Baker. These could be because these young people are just looking for a side-hustle

Question 2. Are there relationship between employees age and their job title?

Before analyzing, the data needs to be filtered out first, to do that, the following code will be executed:

```
empAgeJobTitle <- empUniqueTerminated %>% select(age,job_title) %>% arrange(job_title)
```

The following code will select only the attribute age and job_title from the empUniqueTerminated dataframe.

Analysis 2-1: Checking the basic statistics of the dataset attributes.

To simply check the basic statistical information of the dataset, the following code can be used: [summary()]

```
> summary(empAgeJobTitle)
      age      job_title
Min.   :19.00  Length:1485
1st Qu.:28.00  Class  :character
Median :57.00  Mode   :character
Mean   :47.69
3rd Qu.:60.00
Max.   :65.00
```

From this data, it is seen that the Min age is 19, Max is 65, and mean is 47.69 which seems normal

Analysis 2-2: Checking age stats of a specific job title.

Incase it is needed to check the min age of a specific job position, the min, max, and mean functions can be executed:

```
empBaker <- empAgeJobTitle %>% filter(job_title == 'Baker')
min(empBaker$age)
max(empBaker$age)
mean(empBaker$age)
```

Result:

```
> min(empBaker$age)
[1] 21
> max(empBaker$age)
[1] 61
> mean(empBaker$age)
[1] 45.10989
```

Question 3. Relationship between gender and termination.

To first clean the data of unnecessary informations:

```
empGenderTRD <- empUniqueTerminated %>%  
  filter(termreason_desc != 'Not Applicable') %>%  
  select(gender_full, termreason_desc) %>%  
  arrange(termreason_desc)
```

The code above takes the empUniqueTerminated and filters out 'Not Applicable' and filters out other attributes except for gender_full and termreason_desc, and the arrange() function is to arrange the data based on the termreason_desc (empGenderTRD – employee gender terminated reason desc)

Code for data visualization:

```
ggplot(empGenderTRD, aes(x=termreason_desc, fill=gender_full)) +  
  geom_bar(position="dodge") +  
  labs(title="Barchart of Relationship between Gender and Terminated Reason",  
       x="Terminated Reason", y="Frequency")
```

Resulting barchart:



From the visualization, it can be seen that way more employees retired from the company rather than resigned (this information is apart from the N/A data).

This could mean that more employees are motivated to stay at the company until their retirement.

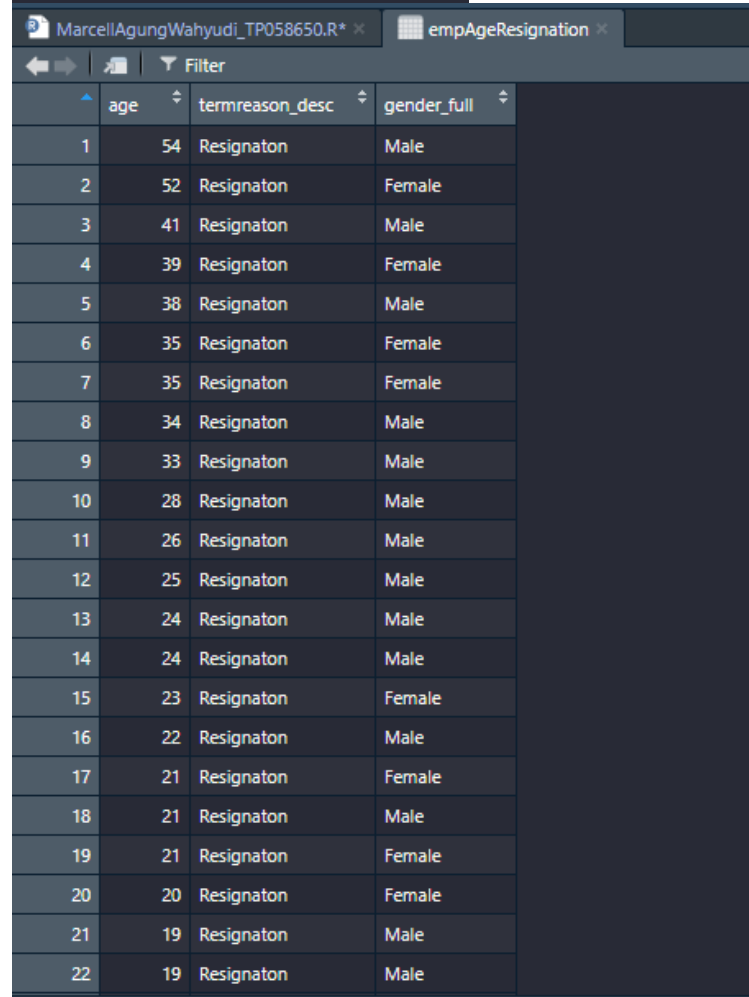
Question 4. At what age does employees resigns from the company.

Code to filter only employees who resigns, then selects only age, termreason_desc, and gender_full attributes):

```
empAgeResignation <- empUniqueTerminated %>%  
  filter(termreason_desc == 'Resignaton')%>%  
  select(age, termreason_desc, gender_full)
```

To view the Dataframe, this code can be executed:

View(empAgeResignation)



	age	termreason_desc	gender_full
1	54	Resignaton	Male
2	52	Resignaton	Female
3	41	Resignaton	Male
4	39	Resignaton	Female
5	38	Resignaton	Male
6	35	Resignaton	Female
7	35	Resignaton	Female
8	34	Resignaton	Male
9	33	Resignaton	Male
10	28	Resignaton	Male
11	26	Resignaton	Male
12	25	Resignaton	Male
13	24	Resignaton	Male
14	24	Resignaton	Male
15	23	Resignaton	Female
16	22	Resignaton	Male
17	21	Resignaton	Female
18	21	Resignaton	Male
19	21	Resignaton	Female
20	20	Resignaton	Female
21	19	Resignaton	Male
22	19	Resignaton	Male

Then to visualize the Relationship between age and resignation, a violin plot is used:

```
ggplot(empAgeResignation, aes(x=termreason_desc, y=age, fill=gender_full)) +  
  geom_violin() +  
  labs(title="Barchart of Relationship between Resignation and Age w/ Gender",  
        x="Resignation", y="Age")
```



From the visualization, it can be seen that most males and females resigns at the age of 20~30, and slowly decreases from then on.

Question 5. Which job positions laid off the most employees

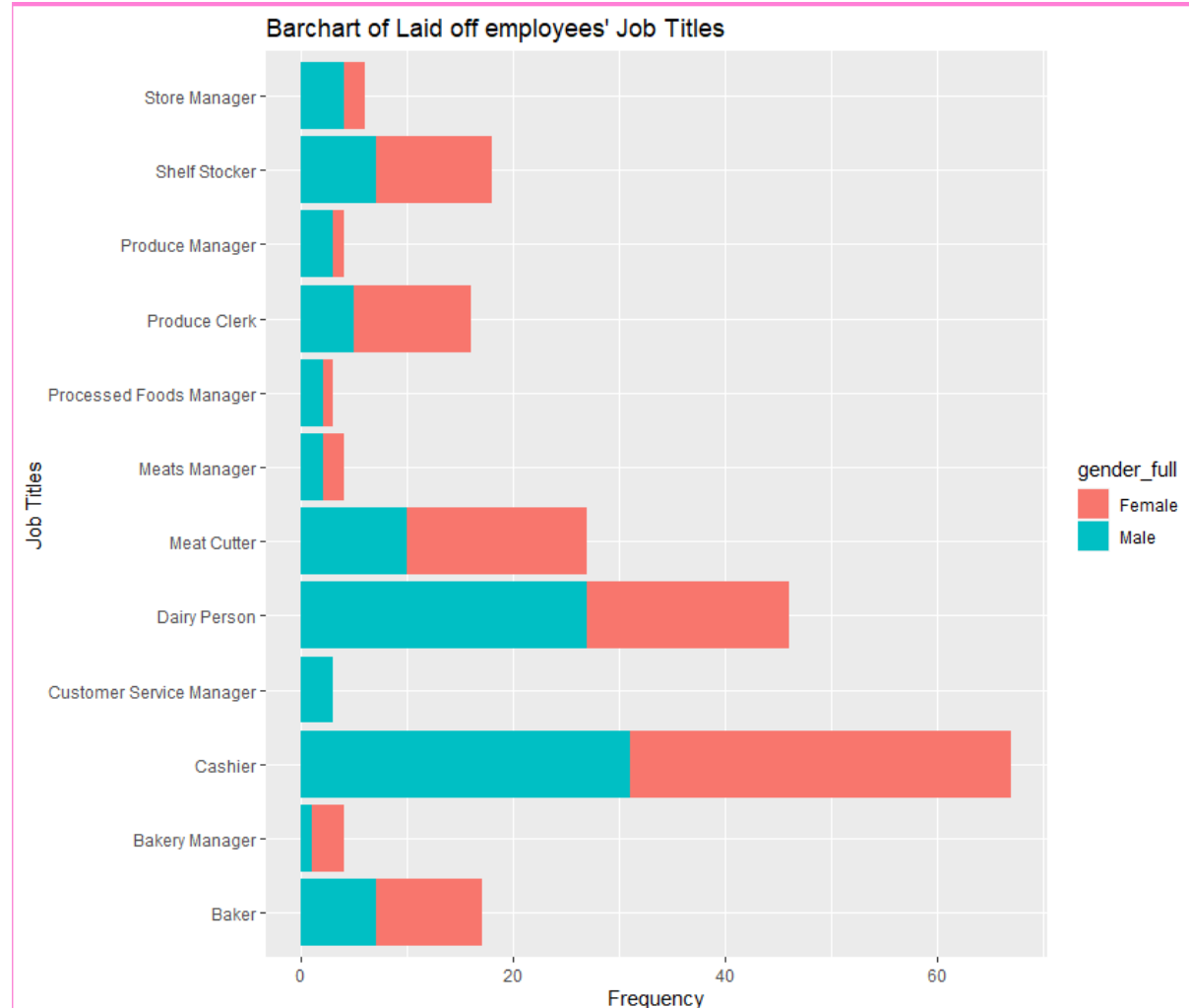
Code to filter only employees that have been laid off:

```
empLayoff <- empTerminated %>% filter(termreason_desc == 'Layoff')
```

To visualize the data:

```
ggplot(empLayoff, aes(y=job_title, fill=gender_full)) +  
  geom_bar() +  
  labs(title="Barchart of Laid off employees' Job Titles", x="Frequency", "Job Titles")
```

Result:



From the visualizaition, it can be seen that the most laid off employees both Male and Female came from the 'Cashier' job position.

Question 6. Does attrition have any connection with a specific store

Analysis 6-1: Finding which store_name has the most employee termination

To get the store name with most termination, we need to find the mode, since there are no mode functions in R, a custom function is borrowed from the internet:

```
getMode <- function(v) {  
  uniqv <- unique(v)  
  uniqv[which.max(tabulate(match(v, uniqv)))]  
}
```

Execute the getMode function to find the mode of empUniqueTerminated\$store_name

```
> getMode(empUniqueTerminated$store_name)  
[1] 35
```

Analysis 6-2: Finding most terminated job position in store 35

To do that, first we need to filter only employees who've been terminated at store 35, this can be achieved with the following code:

```
empStore35 <- empUniqueTerminated %>% filter(store_name==(getMode(empUniqueTerminated$store_name)))
```

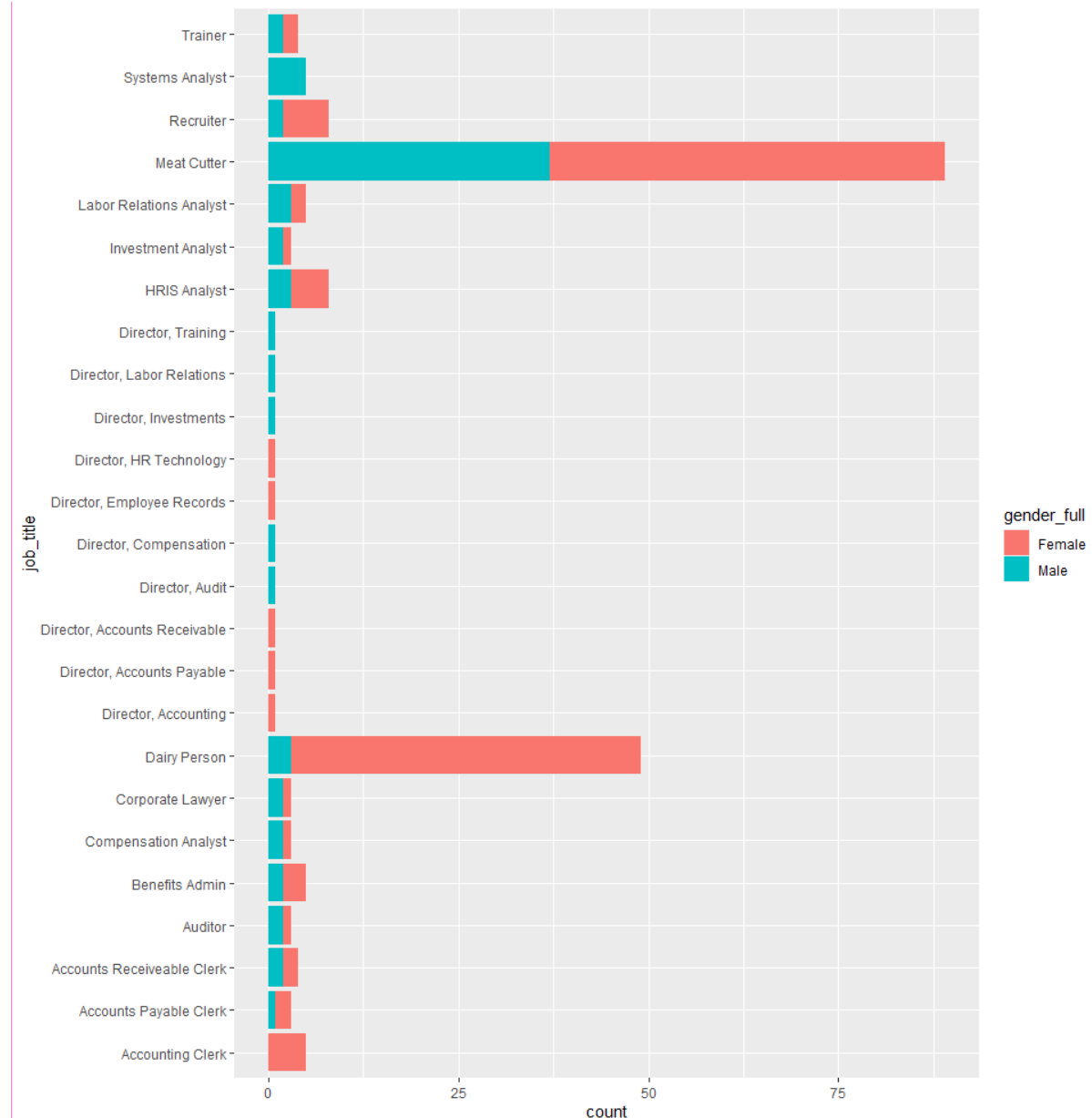
Code: [empStore35 <- empUniqueTerminated %>%

filter(store_name==(getMode(empUniqueTerminated\$store_name)))]

Then we can visualize the dataframe:

```
ggplot(empStore35, aes(y=job_title, fill=gender_full))+  
  geom_bar()
```

Result:



As it is seen, the most noticeable job positions that got terminated from store 35 is as a Meat Cutter and Dairy Person.

Analysis 6-3: Finding the age of the employees working as Meat Cutter and Dairy Person in Store 35.

To filter only employees working as Meat Cutter and Dairy Person

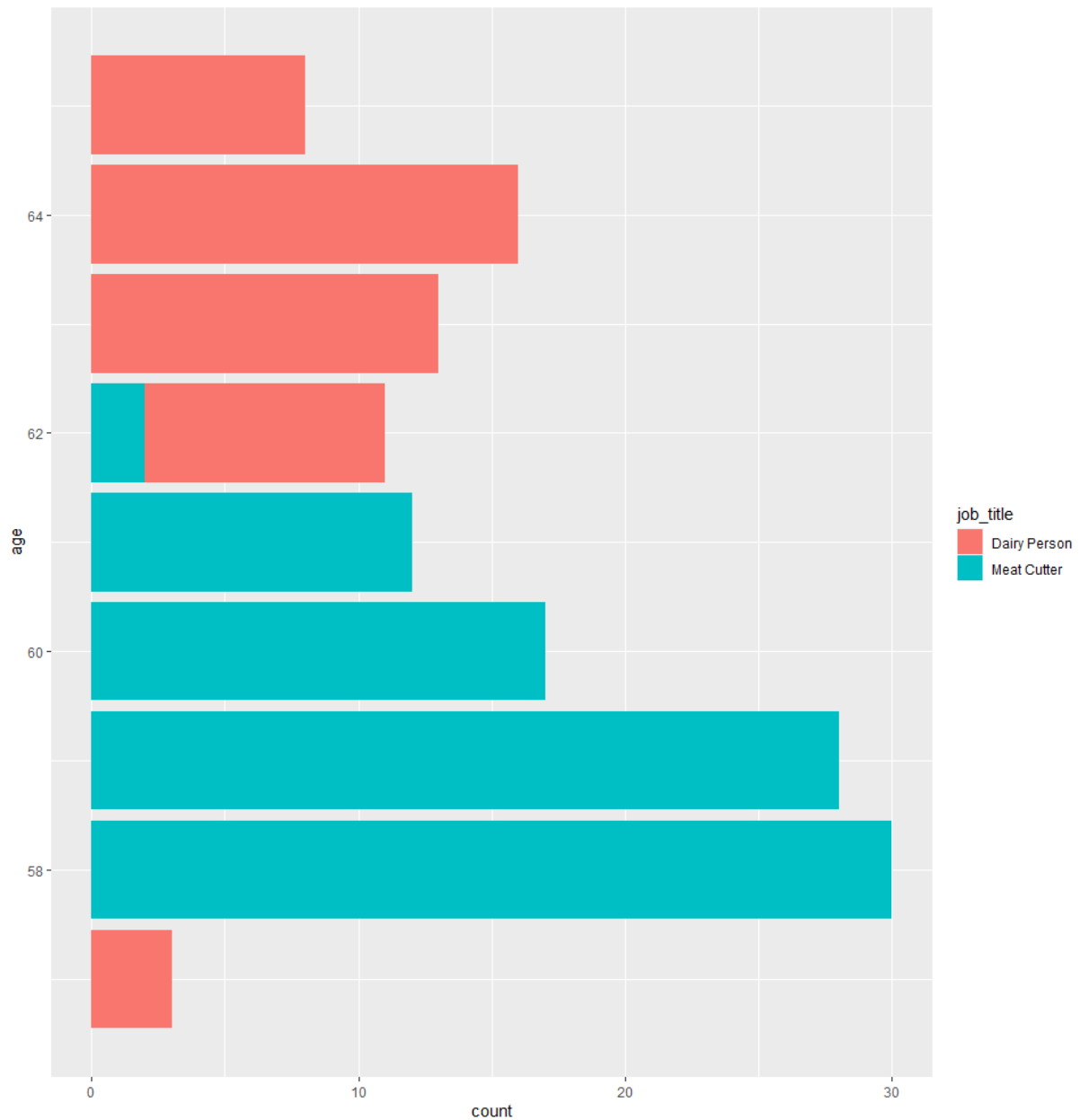
```
jobTitle35 <- empStore35 %>% filter(job_title=='Meat Cutter' | job_title=='Dairy Person')
```

Code : [jobTitle35 <- empStore35 %>% filter(job_title=='Meat Cutter' | job_title=='Dairy Person')]

To visualize the dataframe, the following code can be executed:

```
ggplot(jobTitle35, aes(y=age, fill=job_title))+
  geom_bar()
```


Result:



From the data visualization, it can be seen that the employees are terminated at around age 57~64, which could likely be their termination years, which can be simplified that store 35 have the most terminated employees because there are plenty of people who are around 60 years old, which are retiring.

Extra Features

#1. Custom Function

```
getMode <- function(v) {  
  uniqv <- unique(v)  
  uniqv[which.max(tabulate(match(v, uniqv)))]  
}
```

The following code is a custom function used to get the mode of a vector, it is taken from the internet and the source will be in the reference tab.

Code demonstration:

```
> getMode(empUniqueTerminated$store_name)  
[1] 35
```

Adding this extra feature helped find a certain mystery in the assignment, which is what store has the most termination.

#2. Violin Plot

Code to make the plot:

```
ggplot(empYoung, aes(x=job_title, y=age))+  
  geom_violin()+  
  labs(title="Barchart of Terminated Employees age 20~25 Job Positions",  
        x="Job Position", y="Age")
```

Resulting Plot:

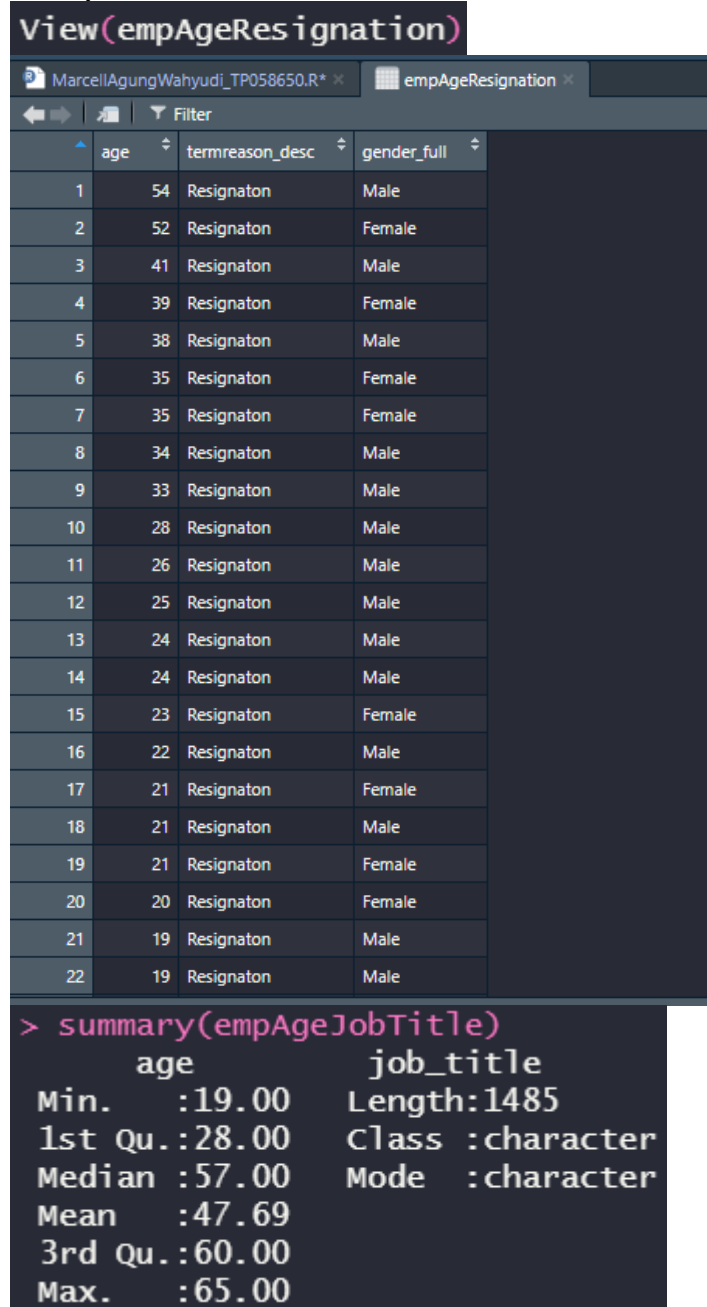


The violin plot shows more information than a bar chart. These extra information are in a form of density, which in this case represents the frequency, which then the x and y can be used to represent other values, and in this example, it represents employees job position and age.

#3. view() and summary()

Example Code:

```
View(empAgeResignation)
```



	age	termreason_desc	gender_full
1	54	Resignaton	Male
2	52	Resignaton	Female
3	41	Resignaton	Male
4	39	Resignaton	Female
5	38	Resignaton	Male
6	35	Resignaton	Female
7	35	Resignaton	Female
8	34	Resignaton	Male
9	33	Resignaton	Male
10	28	Resignaton	Male
11	26	Resignaton	Male
12	25	Resignaton	Male
13	24	Resignaton	Male
14	24	Resignaton	Male
15	23	Resignaton	Female
16	22	Resignaton	Male
17	21	Resignaton	Female
18	21	Resignaton	Male
19	21	Resignaton	Female
20	20	Resignaton	Female
21	19	Resignaton	Male
22	19	Resignaton	Male

```
> summary(empAgeJobTitle)
      age      job_title
Min.   :19.00  Length:1485
1st Qu.:28.00  Class  :character
Median :57.00  Mode   :character
Mean   :47.69
3rd Qu.:60.00
Max.   :65.00
```

These basic R function was not discussed in the module lecture, and thus included in the extra features , these features are basic but handy features to be used.

View() is used to view an object in a table format, rather than only viewing the raw data.

summary() is used to view basic statistics from all attributes of an object.

In this case, View was used to view the dataframe in a table format to check if the data filtration worked as intended or not.summary() was used to easily get the Min, Mean, Max of terminated employees' age in the dataframe.

Findings

Based on the graphs, these are the conclusions:

1. The most terminated employees that stood out are from the departments: Meats, Produce, Customer Service, Dairy, and Bakery.
2. The most terminated job positions are: Meat Cutter, Produce Clerk, Cashier, Dairy Person, and Baker.
3. The most terminated employees are roughly around 55+ which could be hypothesized that it is their retirement years. The second highest are from employees at 20~30 years old. From this sentence, it can be concluded that more people stays at the company till their retirement than resigning the company.
4. From the employees serving years, the most terminated employees served 0 year, which could mean they are an intern or they might be exploring their job options. The second highest are from employees who have served 16 years of experience, and it was found out that the graph reached it's lowest at 14th and 15th year of serving, which could be hypothesized that the company might reward employees who have reached at least 15 years of service, which could motivate the employees to stay longer.
5. Way more employees retired from the company rather than resigned, which means that more employees are motivated to stay at the company until their retirement.
6. Most males and females resigns on their 20s~30s/
7. The most laid off employees both males and females came from the 'Cashier' job position.
8. The most terminated employees came from store 35, those employees worked as Meat Cutter and Dairy Person, and they were terminated because they are retiring since they are around 57+ years old.

Conclusion

Based on the analysis, it is hypothesized that there may be job dissatisfactions which could lead the employees to leave the employee from the departments: Meats, Produce, Customer Service, Dairy, and Bakery.

The job position within the above departments with the most terminated employees are in the positions of: Meat Cutter, Produce Clerk, Cashier, Dairy Person, and Baker. This information should be investigated deeper and find out why these job positions have such high termination employee rate.

From the analysis, it is also known that the most terminated employees (outside of retirement) are from employees who are in their 20~25s, which can be hypothesized that these young people are just working for side-hustle. The job position that has the most termination of these young people are Cashier for 20 year old's. Meat Cutter for 23 year old's, and Baker for 25 year old's.

From analysis #6, it is also known that store 35 has the most employee termination, which is only because the employees are around 57+ years old, which means they are retiring.

Reference

Prabhakaran, S. (2017). *Top 50 ggplot2 Visualizations - The Master List (With Full R Code)*. Retrieved from r-statistics.co: <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>

tutorialspoint. (2021). *R - Mean, Median and Mode*. Retrieved from tutorialspoint: https://www.tutorialspoint.com/r/r_mean_median_mode.htm

Mario Castro. (2020, Sep 17). *R advanced functions that will make your life easier* [Video]. YouTube. <https://www.youtube.com/watch?v=aHiwr3-xJLM>