Covid-19 Report (COMP1013)

Ayad Siddiqui (22029605)

2024-06-04

Declaration

By including this statement, we the authors of this work, verify that:

- We hold a copy of this assignment that we can produce if the original is lost or damaged.
- We hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.
- We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (**which may retain a copy on its database for future plagiarism checking**).
- We hereby certify that we have read and understand what the School of Computing, Engineering and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

Part 1

Distribution Across Counties and Age Groups `part1_analysis.R`

Task

Write the code to analyse the distribution of COVID patients (confirmed or suspected) across counties. Write the code to investigate the distribution of the patients across age groups (e.g., 0-18, 19-35, 36-50, 51+). Visualise both the findings using the histogram. Explain your findings.

Introduction

In this part of the report, we analyze the distribution of COVID-19 patients across various counties and age groups. The analysis aims to identify patterns in the spread of the virus and determine which demographics are most affected.

Methodology and code used

Data Sources The data used in this analysis includes patient demographic information, encounter data, and condition diagnoses.

```
# Set the working directory to the location of your data files
setwd("C:/Users/User/Dropbox/Uni work/Analytics Programing/Report Project")

# Load the data using base R functions
patients <- read.csv("data/patientsUG.csv", stringsAsFactors = FALSE)
encounters <- read.csv("data/encountersUG.csv", stringsAsFactors = FALSE)
conditions <- read.csv("data/conditionsUG.csv", stringsAsFactors = FALSE)

# Display first few rows of each data frame to ensure they are loaded correctly
head(patients[, 1:3])
```

```
##       X                                    Id  BIRTHDATE
## 1  3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8 1988-03-17
## 2   532 9718334c-3289-4b1c-a017-72f3df283ab3 1951-06-13
## 3  5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2 2006-02-06
## 4  7462 c10ee469-6182-4228-ac26-21bcf2412337 1912-10-28
## 5 10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a 1948-06-24
## 6  7818 e283d725-b355-4b86-98a5-b8274e643527 1992-09-01
```

```
head(encounters[, 1:3])
```

```
##   X                                    Id                START
## 1 1 d5ee30a9-362f-429e-a87a-ee38d999b0a5 2019-02-16T01:02:32Z
## 2 2 6a74fdef-2287-44bf-b9e7-18012376faca 2019-08-02T01:02:32Z
## 3 3 8bca6d8a-ab80-4cbf-8abb-46654235f227 2019-10-31T01:02:32Z
## 4 4 821e57ac-9304-46a9-9f9b-83daf60e9e43 2020-01-31T01:02:32Z
## 5 5 681c380b-3c84-4c55-80a6-db3d9ea12fee 2020-03-02T01:02:32Z
## 6 6 9aa748b8-3b44-4e34-b7a8-2e56f2ca3ca2 2019-07-08T08:02:25Z
```

```r
head(conditions[, 1:3])
```

```
##   X      START       STOP
## 1 1 2019-02-15 2019-08-01
## 2 2 2019-10-30 2020-01-30
## 3 3 2020-03-01 2020-03-30
## 4 4 2020-03-01 2020-03-01
## 5 5 2020-03-01 2020-03-30
## 6 6 2020-02-12 2020-02-26
```

Data Filtering

Patients diagnosed with COVID-19 or suspected COVID-19 were filtered.

```r
# Filter for COVID-19 related conditions
covid_conditions_1 <- conditions[conditions$DESCRIPTION == "COVID-19", ]
covid_conditions_2 <- conditions[conditions$DESCRIPTION == "Suspected COVID-19", ]

# Combine the results
covid_conditions <- rbind(covid_conditions_1, covid_conditions_2)

# Separate confirmed COVID-19 cases
confirmed_covid <- covid_conditions[covid_conditions$DESCRIPTION == "COVID-19", ]

# Separate non-confirmed COVID-19 cases
suspected_covid <- covid_conditions[covid_conditions$DESCRIPTION != "COVID-19", ]

# Combine confirmed COVID-19 cases first, then suspected cases
covid_conditions <- rbind(confirmed_covid, suspected_covid)

# Remove duplicates, keeping the first occurrence (which will be confirmed COVID-19 if it exists)
covid_conditions <- covid_conditions[!duplicated(covid_conditions$PATIENT), ]

#Remove unnecessary
rm(confirmed_covid)
rm(suspected_covid)

# Display
head(covid_conditions[, 1:3])
```

```
##     X      START       STOP
## 5   5 2020-03-01 2020-03-30
## 12 12 2020-03-13 2020-04-14
## 24 24 2020-03-11 2020-04-15
## 31 31 2020-03-02 2020-04-07
## 38 38 2020-03-02 2020-03-18
## 48 57 2020-02-25 2020-03-13
```

Data Merging

The filtered data was merged to get a comprehensive dataset including patient demographics.

```r
# Merge conditions with encounters
covid_patients <- merge(covid_conditions, encounters, by.x = "ENCOUNTER", by.y = "Id")

# Merge the result with patients data
covid_patients <- merge(covid_patients, patients, by = "PATIENT.x", by.y = "Id")

# Display first few rows of merged data to ensure merge is done correctly
head(covid_patients[, 1:3])
```

```
##                               PATIENT.x                               ENCOUNTER
## 1 0000b247-1def-417a-a783-41c8682be022 93c3da2d-9420-49fa-94e3-7140ab9aeba1
## 2 00049ee8-5953-4edd-a277-b9c1b1a7f16b dab47020-5bd0-4ce6-ae5c-e4f1ebd04627
## 3 00079a57-24a8-430f-b4f8-a1cf34f90060 3a23144d-0dee-4dca-90ea-0ad14c1c6909
## 4 0008a63c-c95c-46c2-9ef3-831d68892019 2637f12f-5cf1-4287-9e8d-b410a8b41451
## 5 000aa2a0-e307-456f-9f71-c11ab3fc024c 20b83010-3a32-4429-98a2-52c03b3878ae
## 6 0013bde7-14fe-482f-9547-a29077b87904 eda79632-e1e4-45bd-b2f3-0fbaba608b9e
##      X.x
## 1  87722
## 2  72611
## 3  75793
## 4   7250
## 5  36196
## 6 102513
```

Analysis and Findings

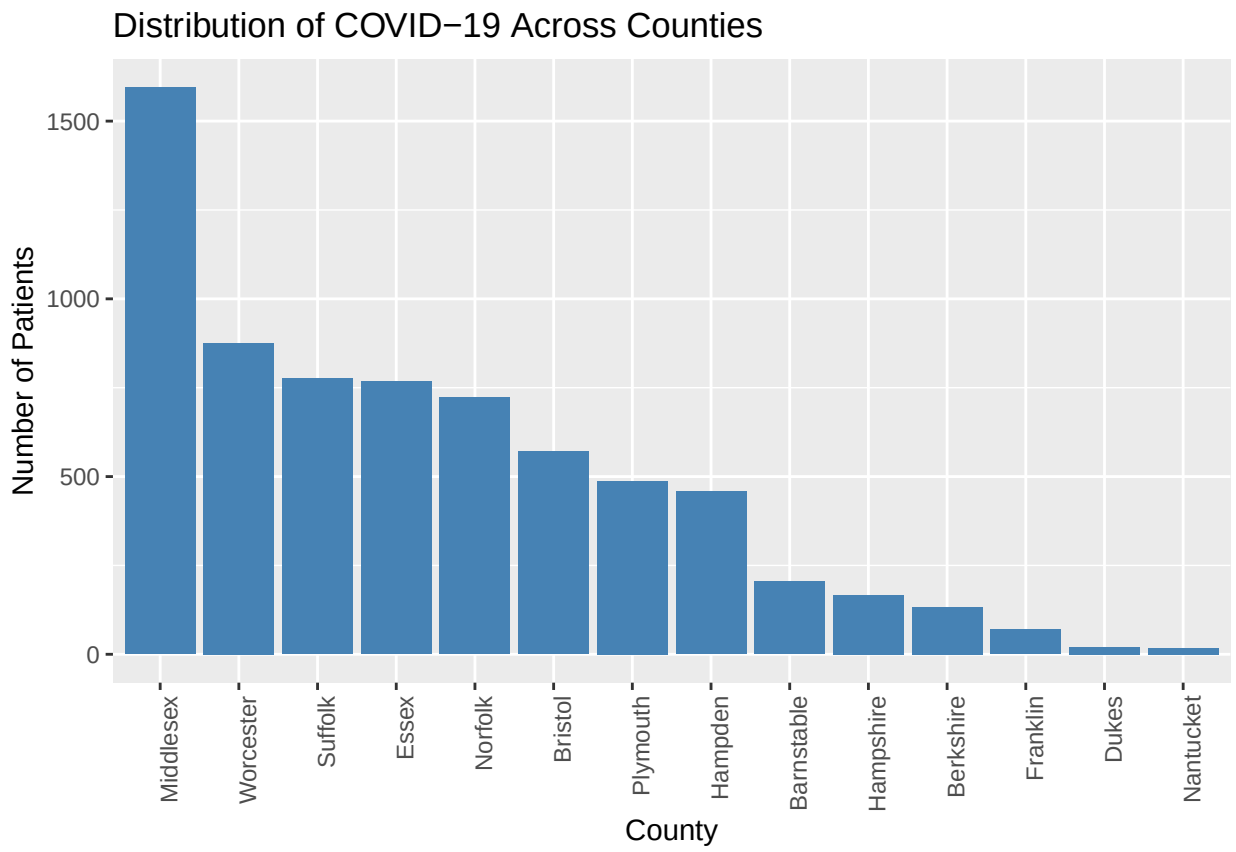
Distribution Across Counties

```r
# Frequency table of the element
county_distribution <- table(covid_patients$COUNTY)
county_distribution <- sort(county_distribution, decreasing = TRUE)

county_distribution_df <- data.frame(
  County = names(county_distribution),
  Number_of_Patients = as.vector(county_distribution)
)

#Remove the "county" part
county_distribution_df$County <- sub(" County", "", county_distribution_df$County)

print(county_distribution_df)
```

```
##        County Number_of_Patients
## 1   Middlesex               1595
## 2   Worcester                875
## 3     Suffolk                776
## 4       Essex                767
## 5     Norfolk                723
## 6     Bristol                571
## 7    Plymouth                487
## 8     Hampden                458
```

```
## 9  Barnstable                205
## 10  Hampshire                166
## 11  Berkshire                133
## 12   Franklin                 70
## 13      Dukes                 20
## 14  Nantucket                 17
```

```r
# Load necessary library
library(ggplot2)

# Create the ggplot for county distribution
ggplot(county_distribution_df, aes(x = reorder(County, -Number_of_Patients), y = Number_of_Patients)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Distribution of COVID-19 Across Counties", x = "County", y = "Number of Patients")
```



Findings: Middlesex County has the highest number of COVID-19 cases with over 1500 cases (1595). This may be an indicator of higher population density or other socioeconomic factors. Smaller counties like Franklin, Dukes, and Nantucket have significantly fewer cases, indicating a possible lower population density or better control measures in these regions.

Distribution Across Age Groups

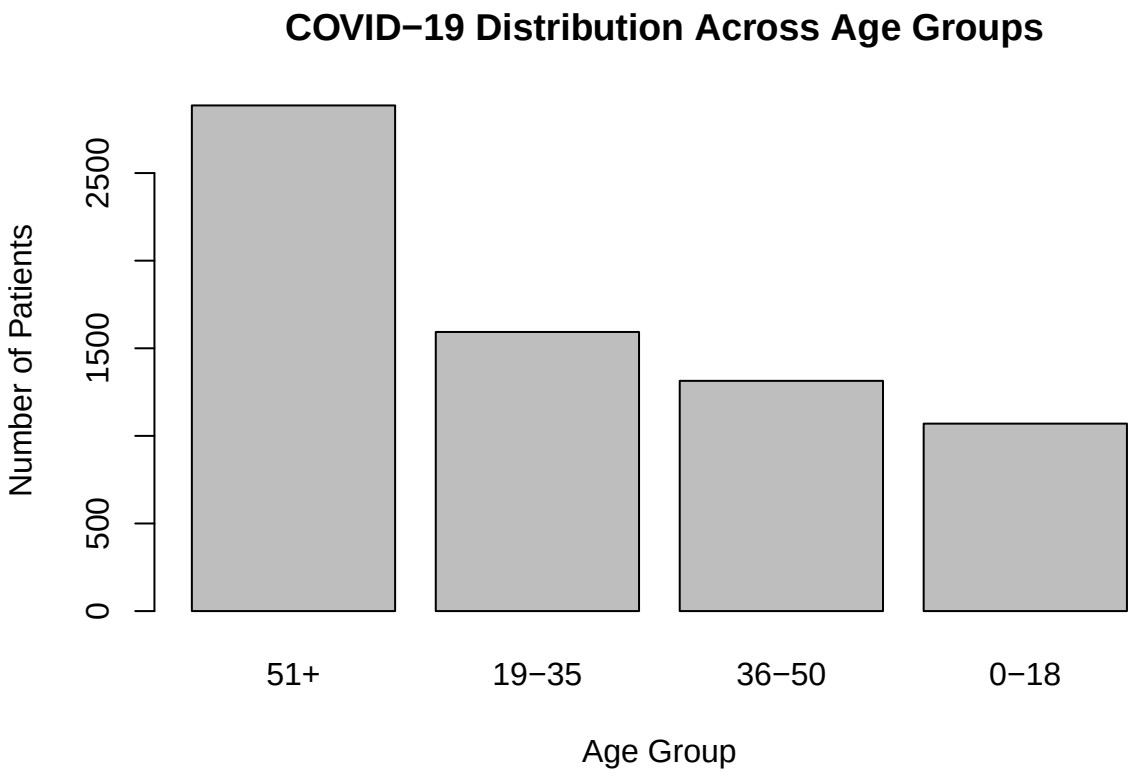
```r
calculate_age <- function(birthDate) {
  today <- Sys.Date()
  birthDate <- as.Date(birthDate)
  age <- as.numeric(difftime(today, birthDate, units = "weeks")) %/% 52.25
  return(age)
}

# Add age and age group columns
covid_patients$Age <- sapply(covid_patients$BIRTHDATE, calculate_age)
```

```r
# Define the age groups
age_breaks <- c(-Inf, 18, 35, 50, Inf)
age_labels <- c("0-18", "19-35", "36-50", "51+")

# Assign age groups to the covid_patients data frame
covid_patients$AgeGroup <- cut(
  covid_patients$Age,
  breaks = age_breaks,
  labels = age_labels
)

age_group_distribution <- table(covid_patients$AgeGroup)
age_group_distribution <- sort(age_group_distribution, decreasing = TRUE)

# Create a data frame for age group distribution
age_group_distribution_df <- data.frame(
  AgeGroup = names(age_group_distribution),
  Number_of_Patients = as.vector(age_group_distribution)
)

# Print the data frame
print(age_group_distribution_df)
```

Breaking into groups

```
##   AgeGroup Number_of_Patients
## 1      51+               2886
## 2    19-35               1593
## 3    36-50               1314
## 4     0-18               1070
```

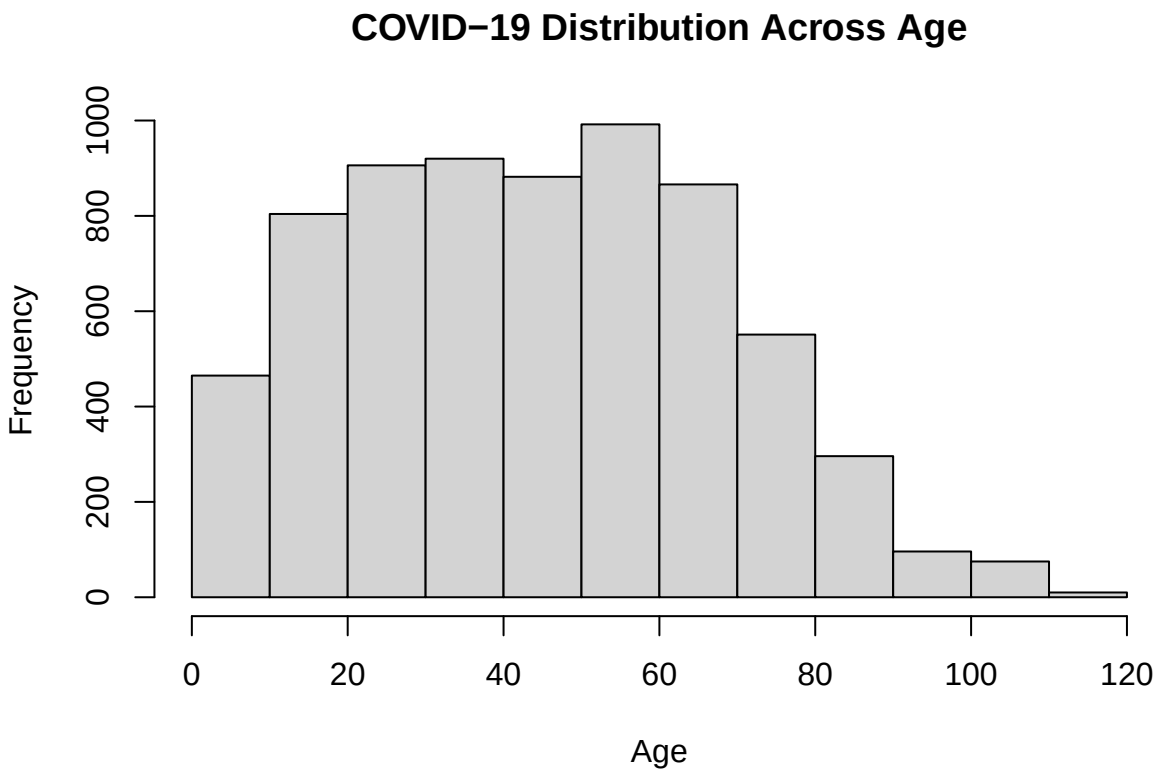
```r
# Barplot of age distribution
barplot(height = age_group_distribution_df$Number_of_Patients,
        names.arg = age_group_distribution_df$AgeGroup,
        main = " COVID-19 Distribution Across Age Groups",
        xlab = "Age Group",
        ylab = "Number of Patients")
```



Findings: The analysis indicates that older adults, particularly those over the age of 51, are more significantly affected by COVID-19. The age group distribution shows a noticeable peak around the age of 60, suggesting that middle-aged to older adults are more frequently diagnosed with COVID-19.

Age Distribution

```
# Histogram of age distribution
hist(covid_patients$Age,
     breaks = 10,  # Number of bins
     main = "COVID-19 Distribution Across Age",
     xlab = "Age",
     ylab = "Frequency")
```



Findings: The histogram of age distribution reveals a peak around the age of 60, indicating that middle-aged to older adults are more frequently diagnosed with COVID-19. The decline in cases beyond the age of 80 could be due to a lower population in this age group or higher mortality rates.

Conclusion

The results of this analysis highlight the demographic and regional disparities in the impact of COVID-19. By understanding these patterns, public health authorities can better allocate resources and design interventions to protect the most vulnerable populations.

Part 2

Common Conditions Among COVID-19 Patients `part2_analysis.R`

Task

Filter those patients in the dataset that have contracted COVID-19 or Suspected COVID-19; ; what are the top 10 most common conditions (symptoms) related to the patients? Do the conditions differ between genders? Provide a table to rank the top 10 conditions for male and female patients separately. Elaborate on the findings.

Introduction

In this part, we analyze the most common conditions (symptoms) among patients who have contracted COVID-19 or are suspected of having COVID-19. The analysis also investigates whether these conditions differ between genders.

Methodology and code used

Data Sources The data used in this analysis includes patient demographic information, encounter data, and condition diagnoses.

```
# Set the working directory to the location of your data files
setwd("C:/Users/User/Dropbox/Uni work/Analytics Programing/Report Project")

# Load the data using base R functions
patients <- read.csv("data/patientsUG.csv", stringsAsFactors = FALSE)
encounters <- read.csv("data/encountersUG.csv", stringsAsFactors = FALSE)
conditions <- read.csv("data/conditionsUG.csv", stringsAsFactors = FALSE)

# Display first few rows of each data frame to ensure they are loaded correctly
head(patients[, 1:3])
```

```
##       X                                    Id  BIRTHDATE
## 1  3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8 1988-03-17
## 2   532 9718334c-3289-4b1c-a017-72f3df283ab3 1951-06-13
## 3  5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2 2006-02-06
## 4  7462 c10ee469-6182-4228-ac26-21bcf2412337 1912-10-28
## 5 10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a 1948-06-24
## 6  7818 e283d725-b355-4b86-98a5-b8274e643527 1992-09-01
```

```
head(encounters[, 1:3])
```

```
##   X                                  Id                 START
## 1 1 d5ee30a9-362f-429e-a87a-ee38d999b0a5 2019-02-16T01:02:32Z
## 2 2 6a74fdef-2287-44bf-b9e7-18012376faca 2019-08-02T01:02:32Z
## 3 3 8bca6d8a-ab80-4cbf-8abb-46654235f227 2019-10-31T01:02:32Z
## 4 4 821e57ac-9304-46a9-9f9b-83daf60e9e43 2020-01-31T01:02:32Z
## 5 5 681c380b-3c84-4c55-80a6-db3d9ea12fee 2020-03-02T01:02:32Z
## 6 6 9aa748b8-3b44-4e34-b7a8-2e56f2ca3ca2 2019-07-08T08:02:25Z
```

```r
head(conditions[, 1:3])
```

```
##   X      START       STOP
## 1 1 2019-02-15 2019-08-01
## 2 2 2019-10-30 2020-01-30
## 3 3 2020-03-01 2020-03-30
## 4 4 2020-03-01 2020-03-01
## 5 5 2020-03-01 2020-03-30
## 6 6 2020-02-12 2020-02-26
```

Data Filtering

Patients diagnosed with COVID-19 or suspected COVID-19 were filtered.

```r
# Filter for COVID-19 related conditions
covid_conditions_1 <- conditions[conditions$DESCRIPTION == "COVID-19", ]
covid_conditions_2 <- conditions[conditions$DESCRIPTION == "Suspected COVID-19", ]

# Combine the results
covid_conditions <- rbind(covid_conditions_1, covid_conditions_2)

# Separate confirmed COVID-19 cases
confirmed_covid <- covid_conditions[covid_conditions$DESCRIPTION == "COVID-19", ]

# Separate non-confirmed COVID-19 cases
suspected_covid <- covid_conditions[covid_conditions$DESCRIPTION != "COVID-19", ]

# Combine confirmed COVID-19 cases first, then suspected cases
covid_conditions <- rbind(confirmed_covid, suspected_covid)

# Remove duplicates, keeping the first occurrence (which will be confirmed COVID-19 if it exists)
covid_conditions <- covid_conditions[!duplicated(covid_conditions$PATIENT), ]

#Remove unnecessary
rm(confirmed_covid)
rm(suspected_covid)

# Display
head(covid_conditions[, 1:3])
```

```
##     X      START       STOP
## 5   5 2020-03-01 2020-03-30
## 12 12 2020-03-13 2020-04-14
## 24 24 2020-03-11 2020-04-15
## 31 31 2020-03-02 2020-04-07
## 38 38 2020-03-02 2020-03-18
## 48 57 2020-02-25 2020-03-13
```

Data Merging

The filtered data was merged to get a comprehensive dataset including patient demographics.

```r
# Merge conditions with encounters
covid_patients <- merge(covid_conditions, encounters, by.x = "ENCOUNTER", by.y = "Id")

# Merge the result with patients data
covid_patients <- merge(covid_patients, patients, by = "PATIENT.x", by.y = "Id")

# Display first few rows of merged data to ensure merge is done correctly
head(covid_patients[, 1:3])
```

```
##                               PATIENT.x                            ENCOUNTER
## 1 0000b247-1def-417a-a783-41c8682be022 93c3da2d-9420-49fa-94e3-7140ab9aeba1
## 2 00049ee8-5953-4edd-a277-b9c1b1a7f16b dab47020-5bd0-4ce6-ae5c-e4f1ebd04627
## 3 00079a57-24a8-430f-b4f8-a1cf34f90060 3a23144d-0dee-4dca-90ea-0ad14c1c6909
## 4 0008a63c-c95c-46c2-9ef3-831d68892019 2637f12f-5cf1-4287-9e8d-b410a8b41451
## 5 000aa2a0-e307-456f-9f71-c11ab3fc024c 20b83010-3a32-4429-98a2-52c03b3878ae
## 6 0013bde7-14fe-482f-9547-a29077b87904 eda79632-e1e4-45bd-b2f3-0fbaba608b9e
##      X.x
## 1  87722
## 2  72611
## 3  75793
## 4   7250
## 5  36196
## 6 102513
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Perform a semi join to filter rows in conditions that have a match in covid_patients$PATIENT.x
covid_patient_conditions <- semi_join(conditions, covid_patients, by = c("PATIENT" = "PATIENT.x"))

# Merge with patient data to get gender
covid_patient_conditions <- merge(covid_patient_conditions, patients, by.x = "PATIENT", by.y = "Id")

# Display test as always!
head(covid_patient_conditions[, 1:3])
```

```
##                                PATIENT   X.x      START
## 1 0000b247-1def-417a-a783-41c8682be022 87717 2020-02-18
## 2 0000b247-1def-417a-a783-41c8682be022 87718 2020-02-18
## 3 0000b247-1def-417a-a783-41c8682be022 87719 2020-02-18
## 4 0000b247-1def-417a-a783-41c8682be022 87720 2020-02-18
## 5 0000b247-1def-417a-a783-41c8682be022 87721 2020-02-18
## 6 0000b247-1def-417a-a783-41c8682be022 87722 2020-02-18
```

Analysis and Findings

Top 10 Conditions for Males

```r
# Function to get top 10 conditions for a given gender
get_top_conditions <- function(gender) {
  gender_conditions <- covid_patient_conditions[covid_patient_conditions$GENDER == gender, ]
  condition_counts <- table(gender_conditions$DESCRIPTION)
  top_conditions <- head(sort(condition_counts, decreasing = TRUE), 10)
  return(data.frame(Condition = names(top_conditions), Count = as.vector(top_conditions)))
}

# Get top 10 conditions for males
top_conditions_male <- get_top_conditions("M")

# Display top 10 conditions for males
print(top_conditions_male)
```

```
##                                Condition Count
## 1                     Suspected COVID-19  3265
## 2                               COVID-19  3161
## 3                          Fever (finding)  2886
## 4                          Cough (finding)  2202
## 5                 Loss of taste (finding)  1725
## 6                        Fatigue (finding)  1271
## 7  Body mass index 30+ - obesity (finding)  1188
## 8                Sputum finding (finding)  1070
## 9                        Anemia (disorder)  1049
## 10                            Prediabetes   991
```

Findings: The top 10 conditions for males include "Suspected COVID-19," "COVID-19," "Fever (finding)," "Cough (finding)," and "Loss of taste (finding)." "Anemia (disorder)" is among the top 10 conditions for males but not for females.

Top 10 Conditions for Females

```r
# Get top 10 conditions for females
top_conditions_female <- get_top_conditions("F")

# Display top 10 conditions for females
print(top_conditions_female)
```

```
##                                Condition Count
## 1                     Suspected COVID-19  3598
## 2                               COVID-19  3487
## 3                          Fever (finding)  3202
## 4                          Cough (finding)  2472
## 5                 Loss of taste (finding)  1846
## 6  Body mass index 30+ - obesity (finding)  1420
## 7                        Fatigue (finding)  1373
```

```
## 8           Miscarriage in first trimester  1199
## 9                 Sputum finding (finding)  1190
## 10                             Prediabetes  1031
```

Findings: The top 10 conditions for females include "Suspected COVID-19," "COVID-19," "Fever (finding)," "Cough (finding)," and "Loss of taste (finding)." "Miscarriage in first trimester" is specific to females.

Conclusion

The findings highlight the importance of understanding the common comorbidities in COVID-19 patients to improve treatment and management strategies. Gender-specific differences in conditions should be considered in public health interventions.

Part 3

Analysis of Factors Influencing Hospitalization Rates for COVID-19 Patients `part3_analysis.R`

Task

Analyze the factors that might influence the hospitalization rate (ambulatory, emergency, inpatient, urgent care) for COVID-19 patients (confirmed or suspected). Focus on two factors: gender and race, to identify any trends that explain the variation in hospitalization rates.

Introduction

This report aims to analyze the factors that might influence the hospitalization rate (ambulatory, emergency, inpatient, urgent care) for COVID-19 patients (confirmed or suspected). We will focus on two factors: gender and race, to identify any trends that explain the variation in hospitalization rates.

Methodology and code used

Data Sources

The data used in this analysis includes patient demographic information, encounter data, and condition diagnoses.

```
# Set the working directory to the location of your data files
setwd("C:/Users/User/Dropbox/Uni work/Analytics Programing/Report Project")

# Load the data using base R functions
patients <- read.csv("data/patientsUG.csv", stringsAsFactors = FALSE)
encounters <- read.csv("data/encountersUG.csv", stringsAsFactors = FALSE)
conditions <- read.csv("data/conditionsUG.csv", stringsAsFactors = FALSE)

# Display first few rows of each data frame to ensure they are loaded correctly
head(patients[, 1:3])
```

```
##       X                                    Id  BIRTHDATE
## 1  3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8 1988-03-17
## 2   532 9718334c-3289-4b1c-a017-72f3df283ab3 1951-06-13
## 3  5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2 2006-02-06
## 4  7462 c10ee469-6182-4228-ac26-21bcf2412337 1912-10-28
## 5 10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a 1948-06-24
## 6  7818 e283d725-b355-4b86-98a5-b8274e643527 1992-09-01
```

```
head(encounters[, 1:3])
```

```
##   X                                    Id                 START
## 1 1 d5ee30a9-362f-429e-a87a-ee38d999b0a5 2019-02-16T01:02:32Z
## 2 2 6a74fdef-2287-44bf-b9e7-18012376faca 2019-08-02T01:02:32Z
## 3 3 8bca6d8a-ab80-4cbf-8abb-46654235f227 2019-10-31T01:02:32Z
## 4 4 821e57ac-9304-46a9-9f9b-83daf60e9e43 2020-01-31T01:02:32Z
## 5 5 681c380b-3c84-4c55-80a6-db3d9ea12fee 2020-03-02T01:02:32Z
## 6 6 9aa748b8-3b44-4e34-b7a8-2e56f2ca3ca2 2019-07-08T08:02:25Z
```

```r
head(conditions[, 1:3])
```

```
##   X      START       STOP
## 1 1 2019-02-15 2019-08-01
## 2 2 2019-10-30 2020-01-30
## 3 3 2020-03-01 2020-03-30
## 4 4 2020-03-01 2020-03-01
## 5 5 2020-03-01 2020-03-30
## 6 6 2020-02-12 2020-02-26
```

Data Filtering

We filtered the data to include only confirmed or suspected COVID-19 cases.

```r
# Filter for COVID-19 related conditions
covid_conditions_1 <- conditions[conditions$DESCRIPTION == "COVID-19", ]
covid_conditions_2 <- conditions[conditions$DESCRIPTION == "Suspected COVID-19", ]

# Combine the results
covid_conditions <- rbind(covid_conditions_1, covid_conditions_2)

# Separate confirmed COVID-19 cases
confirmed_covid <- covid_conditions[covid_conditions$DESCRIPTION == "COVID-19", ]

# Separate non-confirmed COVID-19 cases
suspected_covid <- covid_conditions[covid_conditions$DESCRIPTION != "COVID-19", ]

# Combine confirmed COVID-19 cases first, then suspected cases
covid_conditions <- rbind(confirmed_covid, suspected_covid)

# Remove duplicates, keeping the first occurrence (which will be confirmed COVID-19 if it exists)
covid_conditions <- covid_conditions[!duplicated(covid_conditions$PATIENT), ]

# Remove unnecessary
rm(confirmed_covid)
rm(suspected_covid)

# Display
head(covid_conditions[, 1:3])
```

```
##     X      START       STOP
## 5   5 2020-03-01 2020-03-30
## 12 12 2020-03-13 2020-04-14
## 24 24 2020-03-11 2020-04-15
## 31 31 2020-03-02 2020-04-07
## 38 38 2020-03-02 2020-03-18
## 48 57 2020-02-25 2020-03-13
```

Data Merging

The filtered condition data was merged with encounter and patient demographic data.

```r
# Merge conditions with encounters
covid_patients <- merge(covid_conditions, encounters, by.x = "ENCOUNTER", by.y = "Id")

# Merge the result with patients data
covid_patients <- merge(covid_patients, patients, by = "PATIENT.x", by.y = "Id")

# Display first few rows of merged data to ensure merge is done correctly
head(covid_patients[, 1:3])
```

```
##                               PATIENT.x                               ENCOUNTER
## 1 0000b247-1def-417a-a783-41c8682be022 93c3da2d-9420-49fa-94e3-7140ab9aeba1
## 2 00049ee8-5953-4edd-a277-b9c1b1a7f16b dab47020-5bd0-4ce6-ae5c-e4f1ebd04627
## 3 00079a57-24a8-430f-b4f8-a1cf34f90060 3a23144d-0dee-4dca-90ea-0ad14c1c6909
## 4 0008a63c-c95c-46c2-9ef3-831d68892019 2637f12f-5cf1-4287-9e8d-b410a8b41451
## 5 000aa2a0-e307-456f-9f71-c11ab3fc024c 20b83010-3a32-4429-98a2-52c03b3878ae
## 6 0013bde7-14fe-482f-9547-a29077b87904 eda79632-e1e4-45bd-b2f3-0fbaba608b9e
##      X.x
## 1  87722
## 2  72611
## 3  75793
## 4   7250
## 5  36196
## 6 102513
```

Analysis and Findings

Filter for Relevant Encounter Types

```r
# Filter for relevant encounter types
relevant_encounters <- covid_patients[
  covid_patients$ENCOUNTERCLASS == "ambulatory" |
    covid_patients$ENCOUNTERCLASS == "emergency" |
    covid_patients$ENCOUNTERCLASS == "inpatient" |
    covid_patients$ENCOUNTERCLASS == "urgentcare",
]

# Display first few rows of relevant encounters
head(relevant_encounters[, 1:3])
```

```
##                               PATIENT.x                               ENCOUNTER
## 1 0000b247-1def-417a-a783-41c8682be022 93c3da2d-9420-49fa-94e3-7140ab9aeba1
## 2 00049ee8-5953-4edd-a277-b9c1b1a7f16b dab47020-5bd0-4ce6-ae5c-e4f1ebd04627
## 3 00079a57-24a8-430f-b4f8-a1cf34f90060 3a23144d-0dee-4dca-90ea-0ad14c1c6909
## 4 0008a63c-c95c-46c2-9ef3-831d68892019 2637f12f-5cf1-4287-9e8d-b410a8b41451
## 5 000aa2a0-e307-456f-9f71-c11ab3fc024c 20b83010-3a32-4429-98a2-52c03b3878ae
## 6 0013bde7-14fe-482f-9547-a29077b87904 eda79632-e1e4-45bd-b2f3-0fbaba608b9e
##      X.x
## 1  87722
## 2  72611
## 3  75793
## 4   7250
```

```
## 5  36196
## 6 102513
```

Summaries of Race and Gender

```
# Summaries of race and gender
summary_by_race <- table(relevant_encounters$ENCOUNTERCLASS, relevant_encounters$RACE)
summary_by_gender <- table(relevant_encounters$ENCOUNTERCLASS, relevant_encounters$GENDER)

# Convert to data frame for plotting
summary_by_race_df <- as.data.frame(summary_by_race)
summary_by_gender_df <- as.data.frame(summary_by_gender)
colnames(summary_by_race_df) <- c("EncounterClass", "Race", "Count")
colnames(summary_by_gender_df) <- c("EncounterClass", "Gender", "Count")

# Display the data frames
head(summary_by_race_df)
```

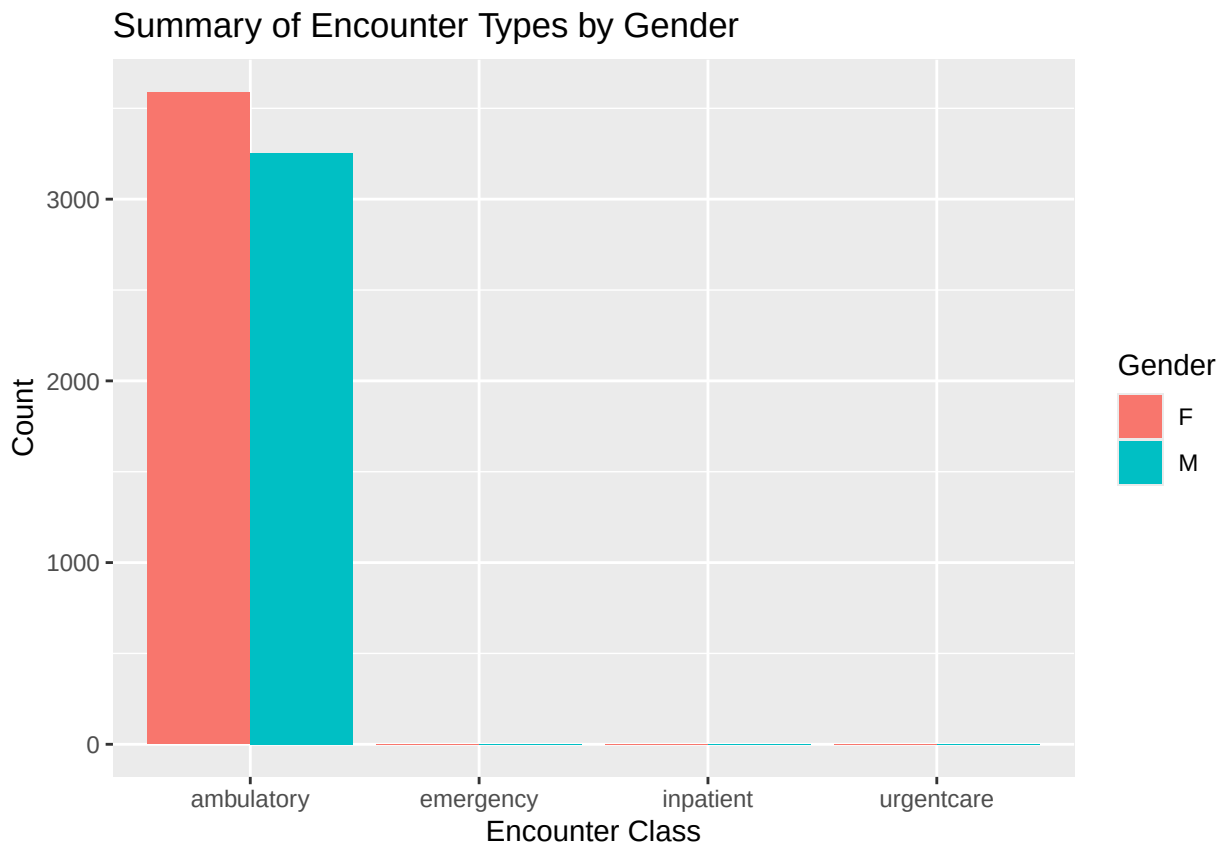
```
##   EncounterClass  Race Count
## 1     ambulatory asian   496
## 2      emergency asian     0
## 3      inpatient asian     0
## 4     urgentcare asian     0
## 5     ambulatory black   568
## 6      emergency black     0
```

```
head(summary_by_gender_df)
```

```
##   EncounterClass Gender Count
## 1     ambulatory      F  3588
## 2      emergency      F     1
## 3      inpatient      F     0
## 4     urgentcare      F     1
## 5     ambulatory      M  3254
## 6      emergency      M     1
```

Summary of Encounter Types by Gender

```
# Plot the summary of encounter types by gender
library(ggplot2)
ggplot(summary_by_gender_df, aes(x = EncounterClass, y = Count, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Summary of Encounter Types by Gender",
       x = "Encounter Class",
       y = "Count")
```

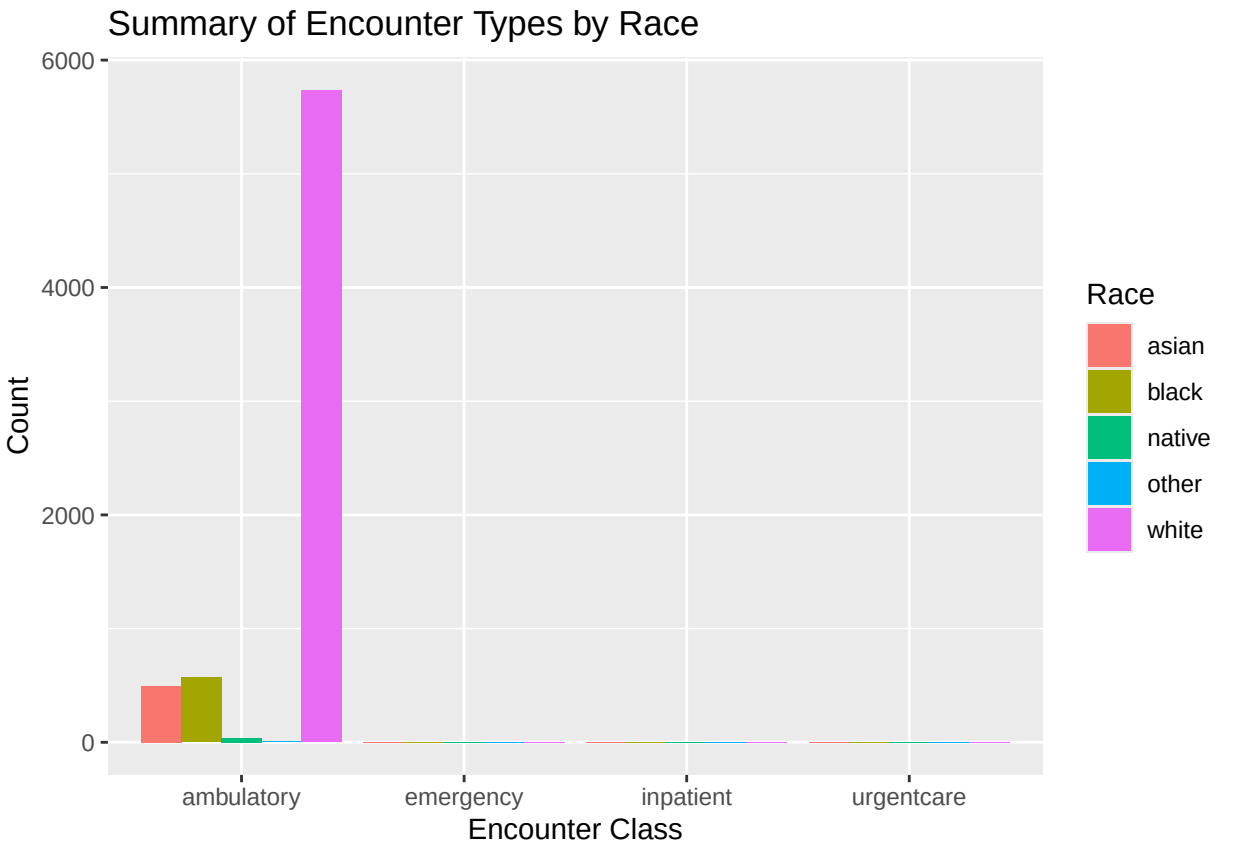


Findings: For Gender

- The vast majority of encounters for both males and females are ambulatory.
- There are very few emergency, inpatient, or urgent care encounters for either gender.
- This trend suggests that most COVID-19 patients are managed on an outpatient basis regardless of gender, with minimal need for higher levels of care.
- The vast majority of encounters for both males and females are ambulatory.
- There are very few emergency, inpatient, or urgent care encounters for either gender.
- This trend suggests that most COVID-19 patients are managed on an outpatient basis regardless of gender, with minimal need for higher levels of care.

Summary of Encounter Types by Race

```
# Plot the summary of encounter types by race
ggplot(summary_by_race_df, aes(x = EncounterClass, y = Count, fill = Race)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Summary of Encounter Types by Race",
       x = "Encounter Class",
       y = "Count")
```



Findings: For Race

- Similar to gender, the majority of encounters for all racial groups are ambulatory.
- White patients have the highest number of ambulatory encounters, followed by Black and Asian patients.
- Emergency, inpatient, and urgent care encounters are negligible across all racial groups.
- This indicates that, like gender, race does not significantly affect the type of encounter, with the majority being managed on an outpatient basis.

Conclusion

The analysis indicates that the majority of COVID-19 patients, irrespective of gender or race, are managed on an ambulatory basis. There is a minimal need for emergency, inpatient, or urgent care across these groups. These findings suggest that hospitalization rates for COVID-19 patients are not significantly influenced by gender or race in this dataset.

Part 4

Analysis of Characteristics Influencing COVID-19 Recovery `part4_analysis.R`

Task

Investigate the characteristics of patients (confirmed or suspected) who recover from COVID-19 compared to those who don't. Analyze factors such as demographics (age, gender, zip code), symptoms, and the timeline of diagnosis and recovery to understand how these factors impact recovery outcomes.

Introduction

This report investigates the characteristics of patients (confirmed or suspected) who recover from COVID-19 compared to those who don't. Factors considered include demographics (age, gender, zip code), symptoms, and the timeline of diagnosis and recovery. The aim is to analyze how these factors impact recovery outcomes.

Methodology and code used

Data Sources

The data used in this analysis includes patient demographic information, encounter data, and condition diagnoses.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
# Set the working directory to the location of your data files
setwd("C:/Users/User/Dropbox/Uni work/Analytics Programing/Report Project")

# Load the data using base R functions
patients <- read.csv("data/patientsUG.csv", stringsAsFactors = FALSE)
encounters <- read.csv("data/encountersUG.csv", stringsAsFactors = FALSE)
conditions <- read.csv("data/conditionsUG.csv", stringsAsFactors = FALSE)

# Display first few rows of each data frame to ensure they are loaded correctly
head(patients[, 1:3])
```

```
##       X                                    Id  BIRTHDATE
## 1  3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8 1988-03-17
## 2   532 9718334c-3289-4b1c-a017-72f3df283ab3 1951-06-13
```

```
## 3  5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2 2006-02-06
## 4  7462 c10ee469-6182-4228-ac26-21bcf2412337 1912-10-28
## 5 10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a 1948-06-24
## 6  7818 e283d725-b355-4b86-98a5-b8274e643527 1992-09-01
```

```
head(encounters[, 1:3])
```

```
##   X                                    Id                 START
## 1 1 d5ee30a9-362f-429e-a87a-ee38d999b0a5 2019-02-16T01:02:32Z
## 2 2 6a74fdef-2287-44bf-b9e7-18012376faca 2019-08-02T01:02:32Z
## 3 3 8bca6d8a-ab80-4cbf-8abb-46654235f227 2019-10-31T01:02:32Z
## 4 4 821e57ac-9304-46a9-9f9b-83daf60e9e43 2020-01-31T01:02:32Z
## 5 5 681c380b-3c84-4c55-80a6-db3d9ea12fee 2020-03-02T01:02:32Z
## 6 6 9aa748b8-3b44-4e34-b7a8-2e56f2ca3ca2 2019-07-08T08:02:25Z
```

```
head(conditions[, 1:3])
```

```
##   X      START       STOP
## 1 1 2019-02-15 2019-08-01
## 2 2 2019-10-30 2020-01-30
## 3 3 2020-03-01 2020-03-30
## 4 4 2020-03-01 2020-03-01
## 5 5 2020-03-01 2020-03-30
## 6 6 2020-02-12 2020-02-26
```

Data Filtering

We filtered the data to include only confirmed or suspected COVID-19 cases.

```
# Filter for COVID-19 related conditions
covid_conditions_1 <- conditions[conditions$DESCRIPTION == "COVID-19", ]
covid_conditions_2 <- conditions[conditions$DESCRIPTION == "Suspected COVID-19", ]

# Combine the results
covid_conditions <- rbind(covid_conditions_1, covid_conditions_2)

# Separate confirmed COVID-19 cases
confirmed_covid <- covid_conditions[covid_conditions$DESCRIPTION == "COVID-19", ]

# Separate non-confirmed COVID-19 cases
suspected_covid <- covid_conditions[covid_conditions$DESCRIPTION != "COVID-19", ]

# Combine confirmed COVID-19 cases first, then suspected cases
covid_conditions <- rbind(confirmed_covid, suspected_covid)

# Remove duplicates, keeping the first occurrence (which will be confirmed COVID-19 if it exists)
covid_conditions <- covid_conditions[!duplicated(covid_conditions$PATIENT), ]

# Remove unnecessary
rm(confirmed_covid)
rm(suspected_covid)
```

```r
# Display
head(covid_conditions[, 1:3])
```

```
##     X      START       STOP
## 5   5 2020-03-01 2020-03-30
## 12 12 2020-03-13 2020-04-14
## 24 24 2020-03-11 2020-04-15
## 31 31 2020-03-02 2020-04-07
## 38 38 2020-03-02 2020-03-18
## 48 57 2020-02-25 2020-03-13
```

Data Merging

The filtered condition data was merged with encounter and patient demographic data.

```r
# Merge conditions with encounters
covid_patients <- merge(covid_conditions, encounters, by.x = "ENCOUNTER", by.y = "Id")

# Merge the result with patients data
covid_patients <- merge(covid_patients, patients, by = "PATIENT.x", by.y = "Id")

# Display first few rows of merged data to ensure merge is done correctly
head(covid_patients[, 1:3])
```

```
##                               PATIENT.x                             ENCOUNTER
## 1 0000b247-1def-417a-a783-41c8682be022 93c3da2d-9420-49fa-94e3-7140ab9aeba1
## 2 00049ee8-5953-4edd-a277-b9c1b1a7f16b dab47020-5bd0-4ce6-ae5c-e4f1ebd04627
## 3 00079a57-24a8-430f-b4f8-a1cf34f90060 3a23144d-0dee-4dca-90ea-0ad14c1c6909
## 4 0008a63c-c95c-46c2-9ef3-831d68892019 2637f12f-5cf1-4287-9e8d-b410a8b41451
## 5 000aa2a0-e307-456f-9f71-c11ab3fc024c 20b83010-3a32-4429-98a2-52c03b3878ae
## 6 0013bde7-14fe-482f-9547-a29077b87904 eda79632-e1e4-45bd-b2f3-0fbaba608b9e
##      X.x
## 1  87722
## 2  72611
## 3  75793
## 4   7250
## 5  36196
## 6 102513
```

Determining Recovery Outcome

Patients were classified as "Recovered" or "Not Recovered" based on the presence of a recovery date.

```r
# Create RecoveryStatus column
covid_patients$RECOVERYSTATUS <- ifelse(covid_patients$STOP.x == "", "Not Recovered", "Recovered")

# Display the updated data
head(covid_patients[, c("PATIENT.x", "STOP.x", "RECOVERYSTATUS")])
```

```
##                               PATIENT.x     STOP.x RECOVERYSTATUS
## 1 0000b247-1def-417a-a783-41c8682be022 2020-03-25      Recovered
```

```
## 2 00049ee8-5953-4edd-a277-b9c1b1a7f16b 2020-03-24      Recovered
## 3 00079a57-24a8-430f-b4f8-a1cf34f90060 2020-03-10      Recovered
## 4 0008a63c-c95c-46c2-9ef3-831d68892019             Not Recovered
## 5 000aa2a0-e307-456f-9f71-c11ab3fc024c 2020-04-04      Recovered
## 6 0013bde7-14fe-482f-9547-a29077b87904 2020-03-30      Recovered
```

Calculating Illness Duration

The duration of illness was calculated for recovered patients.

```
calculate_age <- function(birthDate) {
  today <- Sys.Date()
  birthDate <- as.Date(birthDate)
  age <- as.numeric(difftime(today, birthDate, units = "weeks")) %/% 52.25
  return(age)
}

# Add age and age group columns
covid_patients$Age <- sapply(covid_patients$BIRTHDATE, calculate_age)

# Define the age groups
age_breaks <- c(-Inf, 18, 35, 50, Inf)
age_labels <- c("0-18", "19-35", "36-50", "51+")

# Assign age groups to the covid_patients data frame
covid_patients$AgeGroup <- cut(
covid_patients$Age,
breaks = age_breaks,
labels = age_labels
)
```

Summarize Data by Recovery Status

Summarize the data by recovery status, gender, age group, zip code, and symptoms.

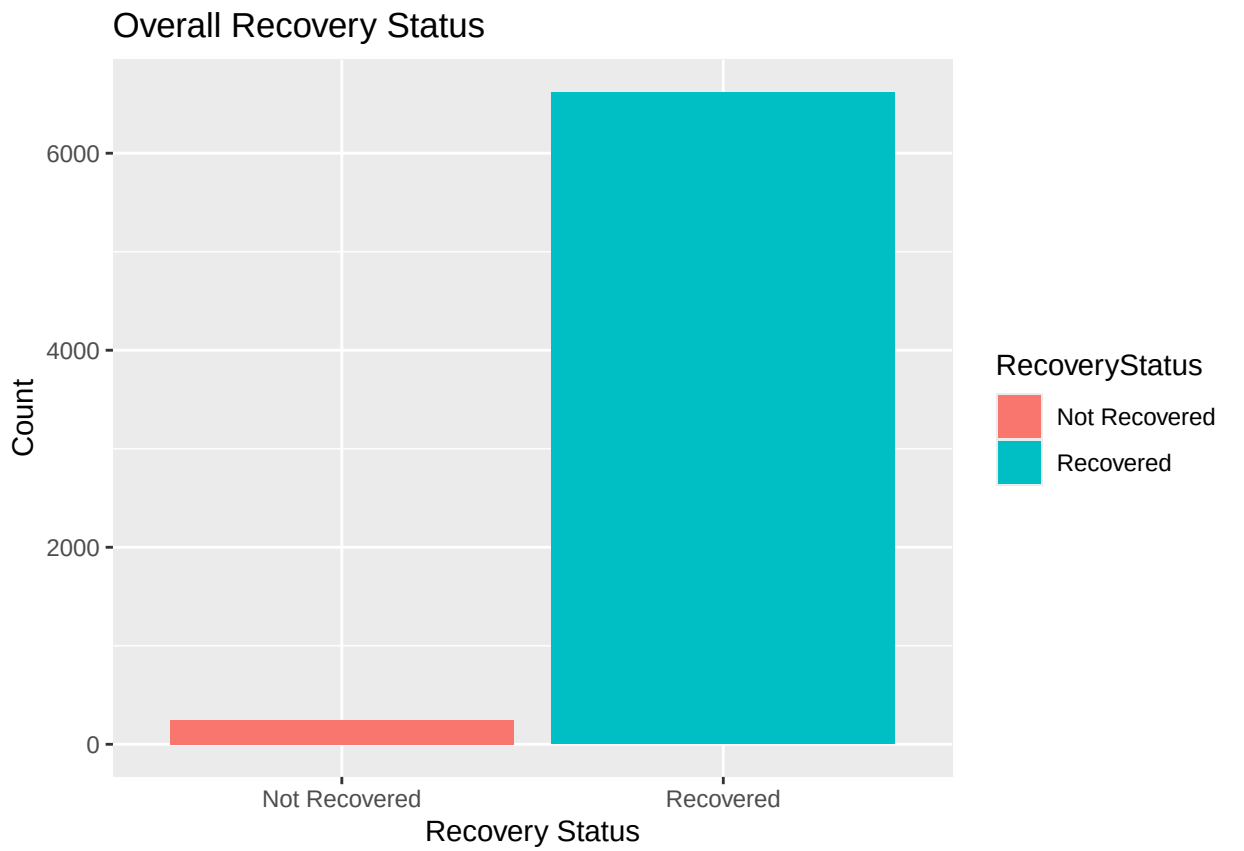
```
# Summarize data by recovery status
summary_by_recovery <- table(covid_patients$RECOVERYSTATUS)
summary_by_gender <- table(covid_patients$GENDER, covid_patients$RECOVERYSTATUS)
summary_by_age <- table(covid_patients$AgeGroup, covid_patients$RECOVERYSTATUS)
summary_by_zip <- table(covid_patients$ZIP, covid_patients$RECOVERYSTATUS)
summary_by_symptoms <- table(covid_patients$DESCRIPTION.x, covid_patients$RECOVERYSTATUS)

# Convert to data frame
summary_by_recovery_df <- as.data.frame(summary_by_recovery)
colnames(summary_by_recovery_df) <- c("RecoveryStatus", "Count")

summary_by_gender_df <- as.data.frame(summary_by_gender)
colnames(summary_by_gender_df) <- c("Gender", "RecoveryStatus", "Count")

summary_by_age_df <- as.data.frame(summary_by_age)
colnames(summary_by_age_df) <- c("AgeGroup", "RecoveryStatus", "Count")
```

```r
summary_by_zip_df <- as.data.frame(summary_by_zip)
colnames(summary_by_zip_df) <- c("Zip", "RecoveryStatus", "Count")


# Filter the data for "Recovered" status
recovered_by_zip <- summary_by_zip_df %>% filter(RecoveryStatus == "Recovered")

# Arrange the data in descending order
top_recovered_by_zip <- recovered_by_zip %>% arrange(desc(Count))

# Select the top 5 ZIP codes
top_5_zip <- top_recovered_by_zip %>% head(5)

summary_by_symptoms_df <- as.data.frame(summary_by_symptoms)
colnames(summary_by_symptoms_df) <- c("Symptom", "RecoveryStatus", "Count")
```

Visualizations

Generate visualizations for the summarized data.

```r
# Overall Recovery Status
ggplot(summary_by_recovery_df, aes(x = RecoveryStatus, y = Count, fill = RecoveryStatus)) +
  geom_bar(stat = "identity") +
  labs(title = "Overall Recovery Status",
       x = "Recovery Status",
       y = "Count")
```

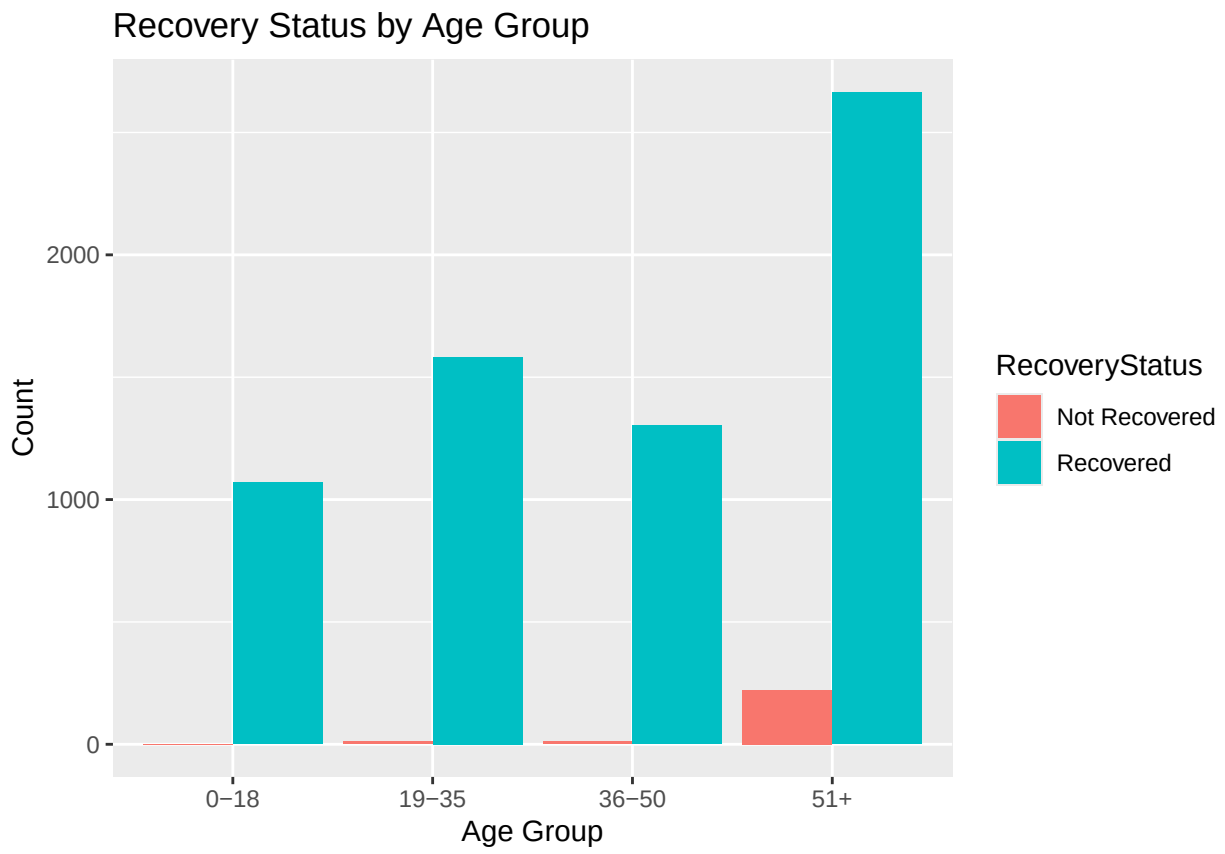



```
head(summary_by_recovery_df)
```

```
##   RecoveryStatus Count
## 1  Not Recovered   245
## 2      Recovered  6618
```

Findings: This indicates that the majority of patients, 6,618 out of the total, have successfully recovered from COVID-19, while 245 patients have not yet recovered. This suggests a relatively high recovery rate among the patients studied. The specific reasons for the differences in recovery rates could be further investigated by examining additional factors such as age, gender, pre-existing conditions, and treatment methods.

```
# Recovery Status by Age Group
ggplot(summary_by_age_df, aes(x = AgeGroup, y = Count, fill = RecoveryStatus)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Recovery Status by Age Group",
       x = "Age Group",
       y = "Count")
```

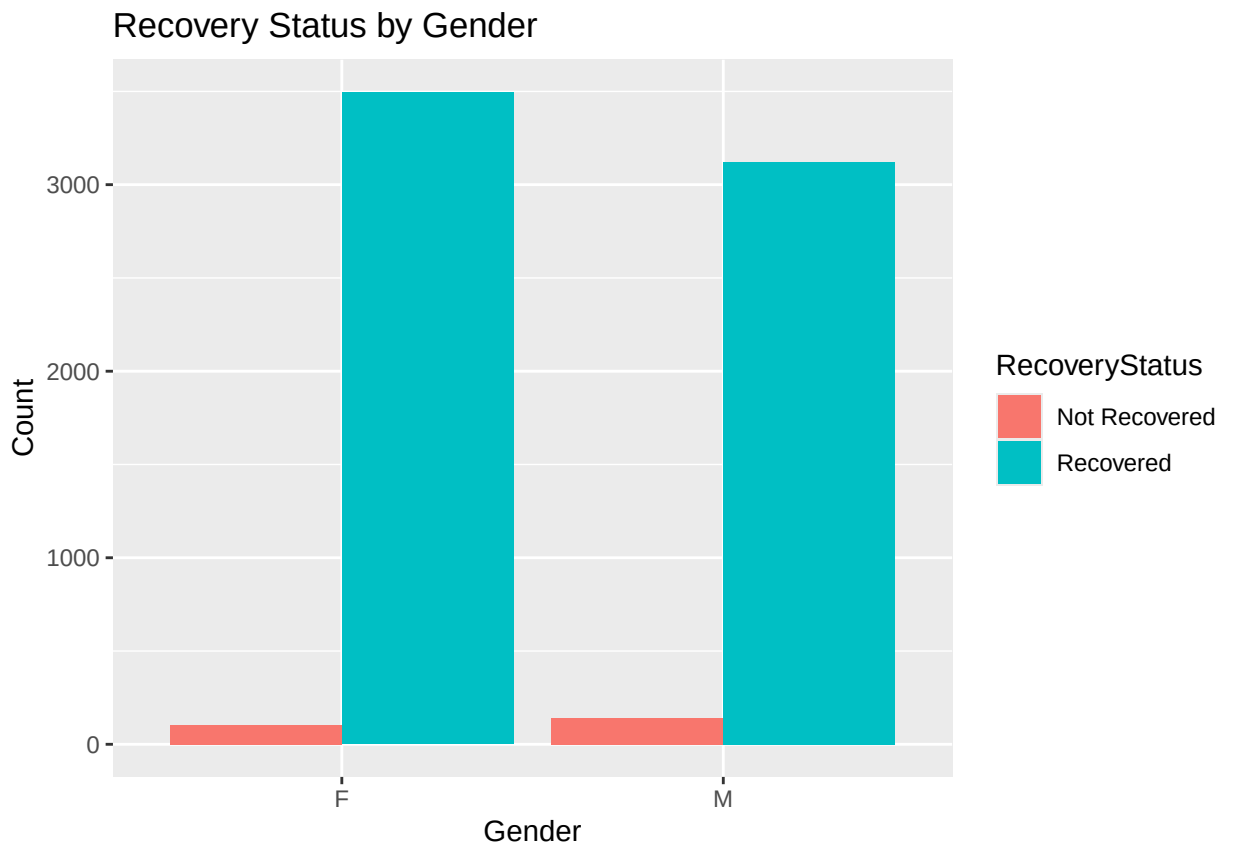



```
head(summary_by_age_df)
```

```
##   AgeGroup RecoveryStatus Count
## 1     0-18  Not Recovered     0
## 2    19-35  Not Recovered    11
## 3    36-50  Not Recovered    12
## 4      51+  Not Recovered   222
## 5     0-18      Recovered  1070
## 6    19-35      Recovered  1582
```

Findings: Patients aged 51 and older have the highest number of cases but also the highest number of recoveries, indicating that while they are heavily affected, recovery is possible with appropriate care. The 0-18 age group has zero non-recovered cases, showing high resilience or effective treatment in younger patients.

```
# Recovery Status by Gender
ggplot(summary_by_gender_df, aes(x = Gender, y = Count, fill = RecoveryStatus)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Recovery Status by Gender",
       x = "Gender",
       y = "Count")
```

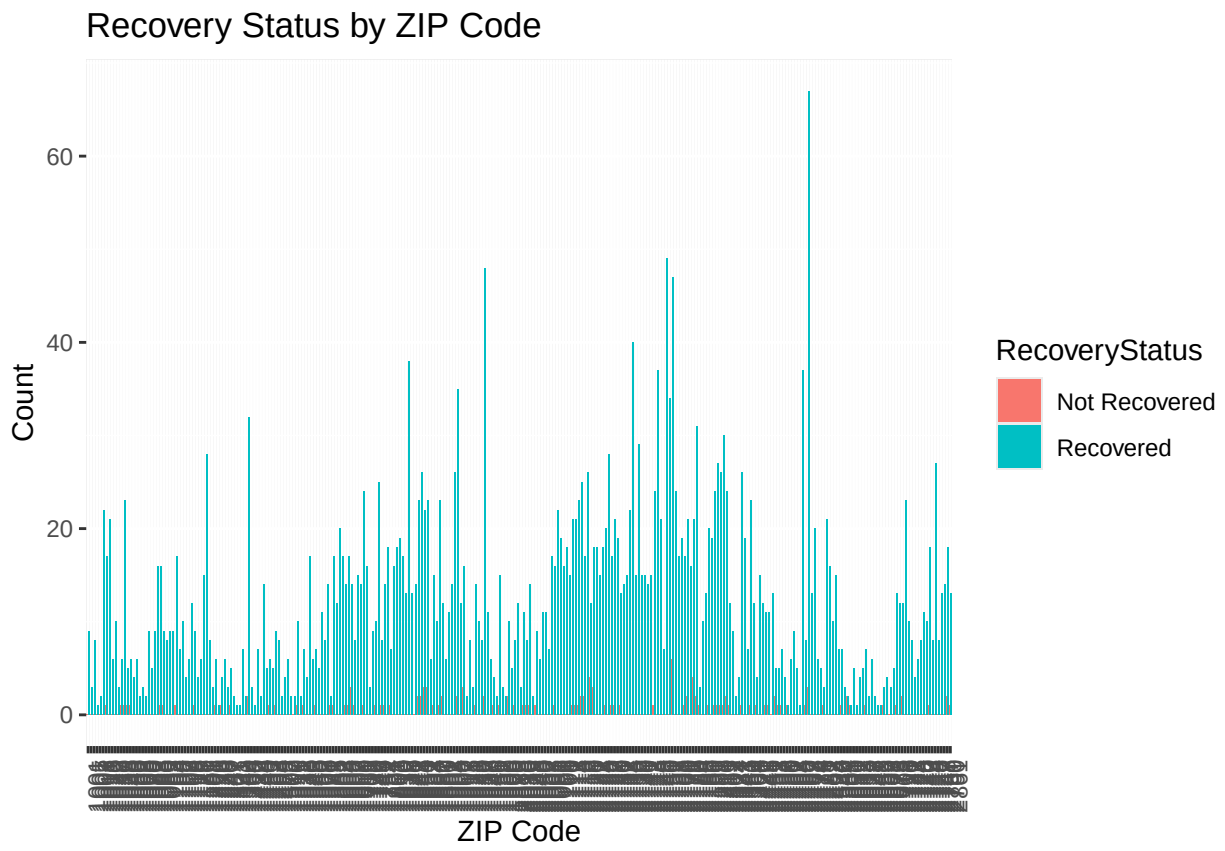



```
head(summary_by_gender_df)
```

```
##   Gender RecoveryStatus Count
## 1      F  Not Recovered   103
## 2      M  Not Recovered   142
## 3      F      Recovered  3495
## 4      M      Recovered  3123
```

Findings: More females (3495) have recovered compared to males (3123), although the number of non-recovered cases is slightly higher in males (142) compared to females (103). This may suggest a slight gender disparity in recovery outcomes, warranting further investigation into potential biological or socio-economic factors.

```
# Recovery Status by Zip Code
ggplot(summary_by_zip_df, aes(x = Zip, y = Count, fill = RecoveryStatus)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Recovery Status by ZIP Code",
       x = "ZIP Code",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

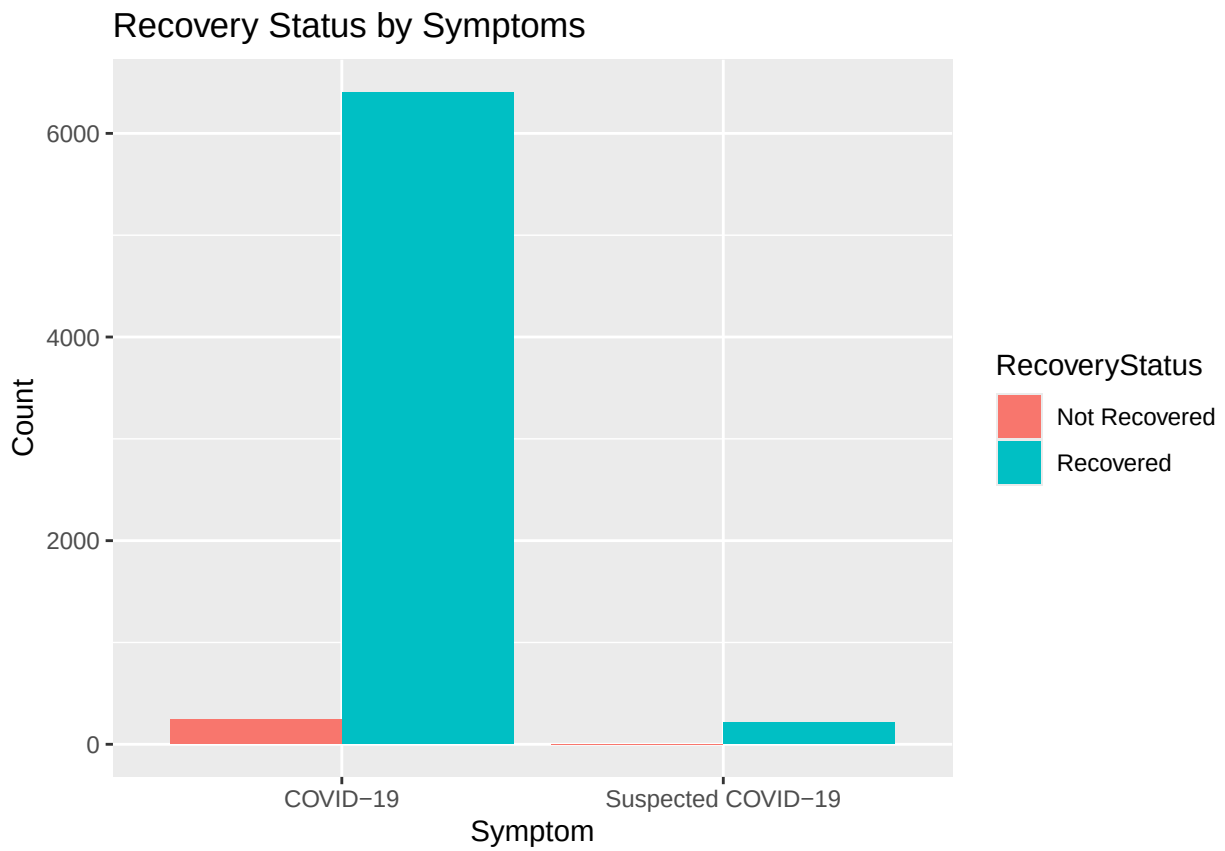



```
head(top_5_zip)
```

```
##    Zip RecoveryStatus Count
## 1 2472      Recovered    67
## 2 2151      Recovered    49
## 3 1940      Recovered    48
## 4 2155      Recovered    47
## 5 2138      Recovered    40
```

Findings: The recovery counts vary significantly across different ZIP codes, with the highest number of recoveries in ZIP code 2472. This suggests that regional factors, such as healthcare facilities and local COVID-19 policies, may impact recovery rates.

```
# Recovery Status by Symptom
ggplot(summary_by_symptoms_df, aes(x = Symptom, y = Count, fill = RecoveryStatus)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Recovery Status by Symptoms",
       x = "Symptom",
       y = "Count")
```



Findings: Patients with "COVID-19" as a symptom have both non-recovered (245) and recovered (6403) cases. "Suspected COVID-19" cases have no non-recovered cases, which may indicate either earlier intervention or less severe initial symptoms. Additionally "Suspected COVID-19 that recovered have (215)

Conclusion

The analysis reveals that age, gender, and regional factors significantly impact the recovery outcomes of COVID-19 patients. Older adults are heavily affected but show high recovery rates, females have a slight advantage in recovery outcomes, and regional differences suggest the impact of local healthcare quality and policies. The data also suggests that early and effective intervention in suspected cases can lead to better recovery outcomes.