WESTERN SYDNEY UNIVERSITY

COMP1013 – An Analysis of COVID Patient Data

Mohamad Hassan

18646447

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.

GitHub Repository: https://github.com/Qukee/COMP1013_Assignment

Table of Contents

GitHub Repository: https://github.com/Qukee/COMP1013_Assignment

Introduction

Preface

This project aims to provide a comprehensive approach to analysing COVID-19 patient data using RStudio, focusing on practical data analysis techniques. It is designed to explore the dynamics of COVID-19 through various real-world examples and statistical methods.

Graphs and Graphic Packages

In this project, simplified code is provided for figures that do not involve relatively complex code. For more intricate visualisations, detailed code is provided specifically targeting the methodology under discussion.

The main graphics packages used are **ggplot2** and **tidyverse**. These packages provide a robust framework for creating a wide range of plots and visualisations. The ggplot2 package, in particular, is great for its flexibility and extensive customisation options, making it ideal for both exploratory data analysis and presenting final results.

Accessing Data and Functions from Packages

A number of packages are automatically loaded, with their functions and datasets then available, at the start of a new R session. For functions and datasets in packages that are not already available, there is a choice between using library() or an equivalent to make all datasets and functions from the package accessible.

Three datasets were given for this project, labelled “encountersUG.csv”, “patientsUG.csv” and “conditionsUG.csv”, all which were loaded into data frames using the **‘read_csv()’** function.

Data Cleaning

Data cleaning was performed to remove any unnecessary columns and filters from the dataset. Column A in all 3 datasets was a numbered column which was not necessary to our data analysis and thus was removed/cleaned.

Naming Conventions

Starred headings (##) in the code identify a comment which is being made and is also identifiable by its green highlighted colour as follows:

```
# This code will print the words ‘test’
        Print(‘test’)
```

GitHub Repository: https://github.com/Qukee/COMP1013_Assignment

Task 1: 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.

```
# Load necessary libraries
library(reader) # for reading CSV files
library(dplyr) # for data manipulation
library(ggplot2) # for data visualisation
library(tidyr) # for tidying data

# Load the CSV files
conditions <- read_csv("Documents/conditionsUG.csv")
encounters <- read_csv("Documents/encountersUG.csv")
patients <- read_csv("Documents/patientsUG.csv")

# Inspect the data
(the glimpse function allows you to see every column in a data frame)
glimpse(conditions)
glimpse(encounters)
glimpse(patients)

# Summary of the data
summary(conditions)
summary(encounters)
summary(patients)

# Data cleaning
# Removing any unnecessary data (column A (‘...1’))
conditions <- conditions %>% select (-‘...1’)
encounters <- encounters %>% select (-‘...1’)
patients <- patients %>% select (-‘...1’)
```

I have filtered the ‘conditions’ data set to only identify COVID or Suspected COVID-19 Patients and joined that data with the Patients dataset.

```
# Identify COVID-19 Patients
(here we are creating a new DF called “covid_conditions” and filtering only by the values
“COVID” or “Suspected COVID” in the description column of the conditions dataset)
covid_conditions <- conditions %>%
filter(grepl("COVID|Suspected COVID", DESCRIPTION, ignore.case = TRUE)
# Join with ‘patients’ data to get demographic information
```

(Here we create a new DF called “covid_patients” by joining the “covid_conditions” DF with the ‘Patients” DF. The join is performed by matching Patient Id’s between datasets.)

```
covid_patients <- covid_conditions %>%
inner_join(patients, by = c(“PATIENT” = “Id”)
```

> Considering that the Patients dataset does not provide an age date, I have converted the provided birthdates into ages (whole numbers) for easier analysis then grouped them into age ranges.

```
# Create age groups
```

(Age is computed based on the current year and the BIRTHDATE. We then categorise patients into age groups: 0-18, 19-35, 36-50, and 51+.)

```
covid_patients <- covid_conditions %>%
mutate (age = as.integer(format(Sys.Date(), "%Y")) - as.integer(substr(BIRTHDATE, 1, 4)),
age_group = case_when(
age <= 18 ~ "0-18",
age <= 35 ~ "19-35",
age <= 50 ~ "36-50",
age > 50 ~ "51+"
))
```

```
# Analyse the top 10 most common conditions
top_conditions <- covid_patients %>%
count(DESCRIPTION) %>%
arrange(desc(n)) %>%
top_n(10, n)
```

```
# Analyse the distribution of COVID patients across age groups
covid_age_distribution <- covid_patients %>%
count(age_group)
```

```
# Analyse the distribution of COVID patients across counties
covid_county_distribution <- covid_patients %>%
count(COUNTY)
```

```
# Check for missing values
```

(this is to help identify any potential issues with the data)

```
colSums(is.na(conditions))
colSums(is.na(encounters))
colSums(is.na(patients))
```

```
# Plot the distribution of COVID patients across age groups (figure 1.1)
# A bar graph using ggplot2
ggplot(covid_age_distribution, aes(x = age_group, y = n)) +
geom_bar(stat = "identity", fill = "blue", color = "black") +
theme_minimal() +
labs(title = "Distribution of COVID Patients Across Age Groups", x = "Age Group", y = "Number of
COVID Patients")

# Plot the distribution of COVID patients across counties (figure 1.2)
# A bar graph using ggplot2
ggplot(covid_county_distribution, aes(x = reorder(COUNTY, n), y = n)) +
geom_bar(stat = "identity", fill = "red", color = "black") +
coord_flip() +
theme_minimal() +
labs(title = "Distribution of COVID Patients Across Counties", x = "County", y = "Number of
COVID Patients")
```

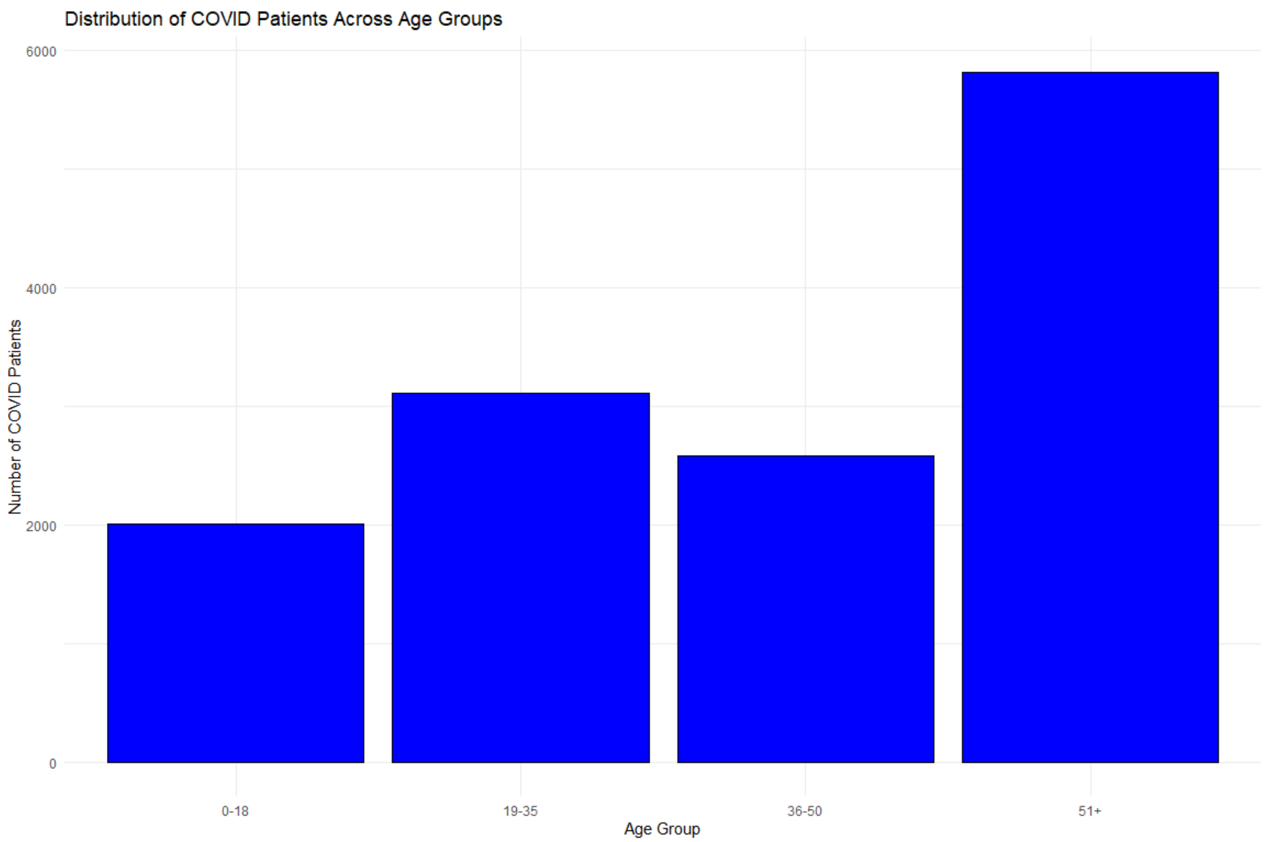

Figure 1. 1 - The Distribution of COVID Patients Across Age Groups

The Distribution of COVID Patients Across Age Groups:

The above bar graph demonstrates the relationship between the number of covid patients in our data set against their age range. We can see a glaringly obvious indication of 51+ year olds being the highest numbers of covid patients – suggesting that older adults are more vulnerable to covid-19. This increase in older patients may be attributed to a higher likelihood of having *other* underlying health conditions but is something which needs to be explored further.

The second highest number of cases comes from the 19-35 age group which could be due to higher social interaction, followed by the 36-50 age range and lastly 0-18. Overall, this graph gives us a base-level understanding of how age contributes to ones susceptibility to COVID-19.

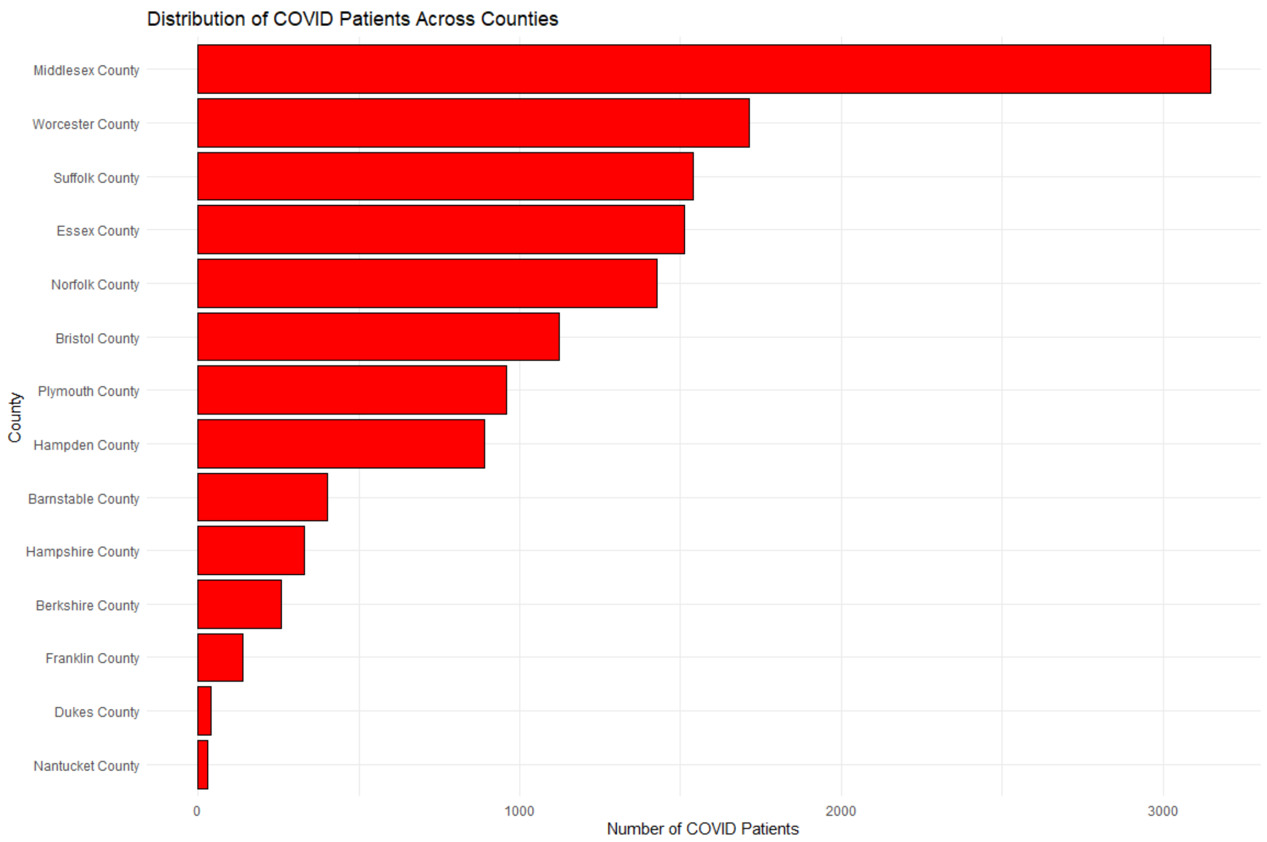

Figure 1. 2 - The Distribution of COVID Patients Across Counties

The Distribution of COVID Patients Across Counties:

Counties with the highest number of cases, such as Middlesex and Worcester also have the densest population according to the 2023 US Census. With Middlesex County have a population of 863,623 and Worcester County a population of 866,866. This indicates a higher risk of virus transmission due to closer living conditions and higher interaction rates. This builds upon our existing knowledge of COVID-19 being a highly contagious and easily transmittable disease. The variation between counties may also be indicative of differences in healthcare, public health services, community behaviour, and demographic.

Figures 1.01 and 1.02 provide a clear overview of how COVID-19 cases are distributed across various age ranges and counties.

Task 2: 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.

```
# Load necessary libraries
library(reader) # for reading CSV files
library(dplyr) # for data manipulation
library(ggplot2) # for data visualisation
library(tidyr) # for tidying data

# Load the CSV files
conditions <- read_csv("Documents/conditionsUG.csv")
encounters <- read_csv("Documents/encountersUG.csv")
patients <- read_csv("Documents/patientsUG.csv")

# Data cleaning
# Removing any unnecessary data (column A ('...1'))
conditions <- conditions %>% select (-'...1')
encounters <- encounters %>% select (-'...1')
patients <- patients %>% select (-'...1')
```

To ensure we use only datasets which are relevant to our task, I have filtered the 'conditions' data set to only identify COVID or Suspected COVID-19 Patients, then matched Patient IDs with their Gender and joined that new data set with the 'encounters' dataset to include patient symptoms (REASONDESCRIPTION).

```
# Filtering the data to look for only COVID-19 or Suspected COVID-19 patients
covid_conditions <- conditions %>%
filter(grepl("COVID|Suspected COVID", DESCRIPTION, ignore.case = TRUE))

# Join with Patients data to get Gender information
covid_conditions_with_gender <- covid_conditions %>%
inner_join(patients %>% select(Id, GENDER), by = c("PATIENT" = "Id"))

# Join with Encounters data to get 'ReasonDescription' information by matching PATIENT
number
covid_conditions_with_reason <- covid_conditions_with_gender %>%
inner_join(encounters %>% select(PATIENT, REASONDESCRIPTION), by = "PATIENT")



# Remove rows with N/A Reason Description (for data cleaning purposes)
covid_conditions_with_reason <- covid_conditions_with_reason %>%
filter(!is.na(REASONDESCRIPTION))

# Count the frequency of each condition/ReasonDescription
condition_counts <- covid_conditions_with_reason %>%
group_by(REASONDESCRIPTION, GENDER) %>%
summarise(count = n()) %>%
arrange(desc(count))

# Get top 10 conditions for each gender (Males)
top_conditions_male <- condition_counts %>%
filter(GENDER == "M") %>%
top_n(10, count) %>%
arrange(desc(count))

# Get top 10 conditions for each gender (Males)
top_conditions_female <- condition_counts %>%
filter(GENDER == "F") %>%
top_n(10, count) %>%
arrange(desc(count))

# Combine the top conditions into one table
top_conditions_combined <- full_join(
top_conditions_male %>% rename(Male_Count = count),
top_conditions_female %>% rename(Female_Count = count),
by = "REASONDESCRIPTION"
) %>%
arrange(desc(Male_Count), desc(Female_Count))

# Print the combined table of top conditions
# This table shows the top 10 conditions for both male and female patients
print(top_conditions_combined)
```

	REASONDESCRIPTION	GENDER.x	MALE COUNT	GENDER.y	FEMALE COUNT
1	Hyperlipidaemia	M	***11541***	F	***11148***
2	COVID-19	M	***1986***	F	***2018***
3	Viral Sinusitis (disorder)	M	***1864***	F	***1946***
4	Acute Bronchitis (disorder)	M	***1227***	F	***1279***
5	Anemia (disorder)	M	***1158***	F	***1197***
6	Sinusitis (disorder)	M	***846***	F	***960***
7	Acute Bacterial Sinusitis	M	***774***	F	***801***
8	Acute Viral Pharyngitis	M	***769***	F	***753***
9	Appendicitis	M	***636***	F	***570***
10	Childhood Asthma	M	***482***	F	***538***

Table 1.01 The top 10 most common conditions (symptoms) related to COVID-19 patients (top_conditions_combined)

The Top 10 Most Common symptoms related to COVID-19 Patients:

The above table lists the top 10 most common conditions/symptoms associated with COVID-19 and Suspected COVID-19 patients, separated by gender. Most conditions show a slight female predominance, except for acute viral pharyngitis and appendicitis, which are more common in males.

Hyperlipidaemia is the most prevalent condition among COVID-19 patients, followed by COVID-19 itself. These findings suggest that exposure to COVID itself is not the *biggest* threat to patients alone, rather, being exposed to COVID while having certain comorbidities (such as hyperlipidaemia) increase your chance of infection by a substantial amount.

Other less-common conditions, such as childhood Asthma may not be key drivers for COVID-19 infections however, there is still a correlation between the two.

Overall, the analysis provides valuable insights into the common conditions among COVID-19 patients, helping to understand the health profile of affected individuals and potentially guiding healthcare strategies for managing comorbid conditions.

Task 3: Write the code to analyse the factors that might influence the hospitalisation rate (ambulatory, emergency, inpatient, urgent care) for the COVID patient (confirmed or suspected) in the dataset. Any factors in the dataset, such as age, gender, zip code, marital status, race and county, can be considered. Pick 2 of the factors and explain if there is a trend that explains the variation. ***(chosen factors = age and county)***

```
# Load necessary libraries
library(reader) # for reading CSV files
library(dplyr) # for data manipulation
library(ggplot2) # for data visualisation
library(tidyr) # for tidying data

# Load the CSV files
conditions <- read_csv("Documents/conditionsUG.csv")
encounters <- read_csv("Documents/encountersUG.csv")
patients <- read_csv("Documents/patientsUG.csv

# Data cleaning
# Removing any unnecessary data (column A ('...1'))
conditions <- conditions %>% select (-'...1')
encounters <- encounters %>% select (-'...1')
patients <- patients %>% select (-'...1')
```

To ensure we use only datasets which are relevant to our task, I have filtered the 'conditions' data set to only identify COVID or Suspected COVID-19 Patients, then included patient information (gender, age, etc..) into the dataset for more context.

```
# Identify COVID-19 or Suspected COVID-19 patients
covid_conditions <- conditions %>%
filter(grepl("COVID|Suspected COVID", DESCRIPTION, ignore.case = TRUE))

# Join with patients data to get relevant information
# This will help in analysing the conditions based on age, gender and location
covid_patients <- covid_conditions %>%
inner_join(patients %>% select(Id, GENDER, AGE = BIRTHDATE, COUNTY, STATE), by =
c("PATIENT" = "Id"))
```

The Patients dataset does not have an 'age' column, rather it displays patients date of birth. This could quickly become confusing when analysing large amounts of data, so I have converted the birthdates into whole numbers for easier analysis.

```r
# Calculate age (assuming the data includes the current year for simplicity)
# Converting birthdate to age for better and easier analysis
covid_patients <- covid_patients %>%
mutate(AGE = as.integer(format(Sys.Date(), "%Y")) - as.integer(substr(AGE, 1, 4)))

# Join with encounters data to get encounter class information
# This join adds encounter class information to our dataset, which will help us identify the type
of hospitalisation
covid_encounters <- covid_patients %>%
inner_join(encounters %>% select(PATIENT, ENCOUNTERCLASS), by = "PATIENT")

# Calculate hospitalisation rates by age and county
# Filtering encounter classes to include only relevant types of hospitalisation
# Grouping by age and county to calculate the number of hospitalisations
hospitalisation_rate <- covid_encounters %>%
filter(ENCOUNTERCLASS %in% c("ambulatory", "emergency", "inpatient", "urgent care")) %>%
group_by(AGE, COUNTY) %>%
summarise(count = n()) %>%
ungroup()

# Visualise hospitalisation rates by age and gender – heat map (Figure 1.3)
ggplot(hospitalisation_rate, aes(x = AGE, y = COUNTY, fill = count)) +
geom_tile() +
labs(title = "Hospitalisation Rate by Age and County for COVID-19 Patients", x = "Age", y =
"County") +
scale_fill_gradient(low ="white", high = "red") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Visualise the trends by age – line graph (Figure 1.4)
ggplot(hospitalisation_rate, aes(x = AGE, y = count)) +
geom_line() +
labs(title = "Hospitalisation Rate by Age for COVID-19 Patients", x = "Age", y = "Count of
Hospitalisations") +
theme_minimal()
```

```
# Visualise the trends by county – bar graph (Figure 1.5)
ggplot(hospitalisation_rate, aes(x = COUNTY, y = count, fill = COUNTY)) +
geom_bar(stat = "identity") +
labs(title = "Hospitalisation Rate by County for COVID-19 Patients", x = "County", y = "Count of
Hospitalisations") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Perform regression analysis to understand impact of age and county on hospitalisation rates
model <- lm(count ~ AGE + COUNTY, data = hospitalisation_rate)

# Summary of the regression model
summary(model)

# Include an interaction term between age and county
model_interaction <- lm(count ~ AGE * COUNTY, data = hospitalisation_rate)

# Summary of the regression model with interaction
summary(model_interaction)
```

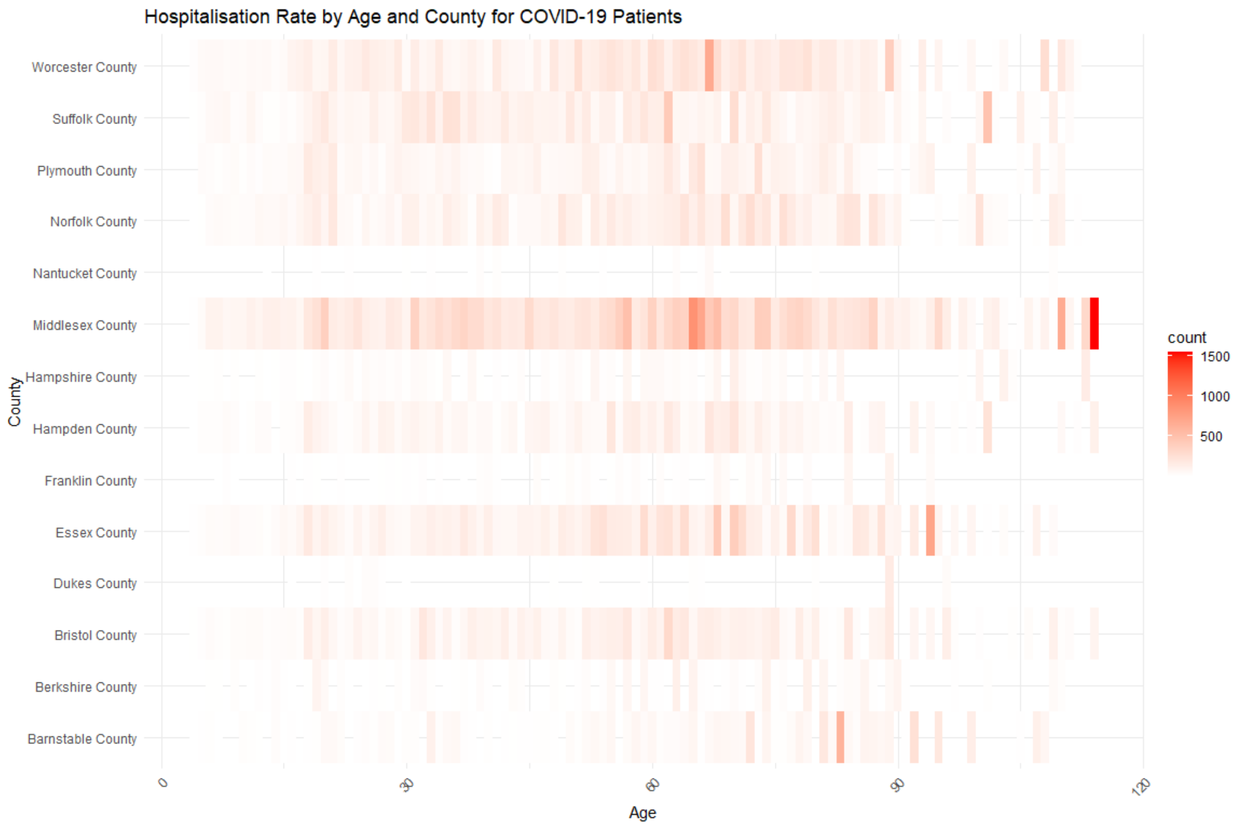

Figure 1. 3 - Heat Map of the Hospitilisation rate of COVID-19 Patients by Age and County

Hospitilisation Rate by Age and County for COVID-19 Patients (heat map):

The heatmap provides a comprehensive overview of the hospitalisation rates across different age groups and counties. The intensity of the color indicates the number of hospitalisations, with darker shades representing higher counts. This visualisation allows us to quickly identify hotspots where certain age groups in specific counties have higher hospitalisation rates. For example, Middlesex County shows a significant number of hospitalisations across various age groups, indicating a high overall impact (hospitalisation rate) amongst all age groups.

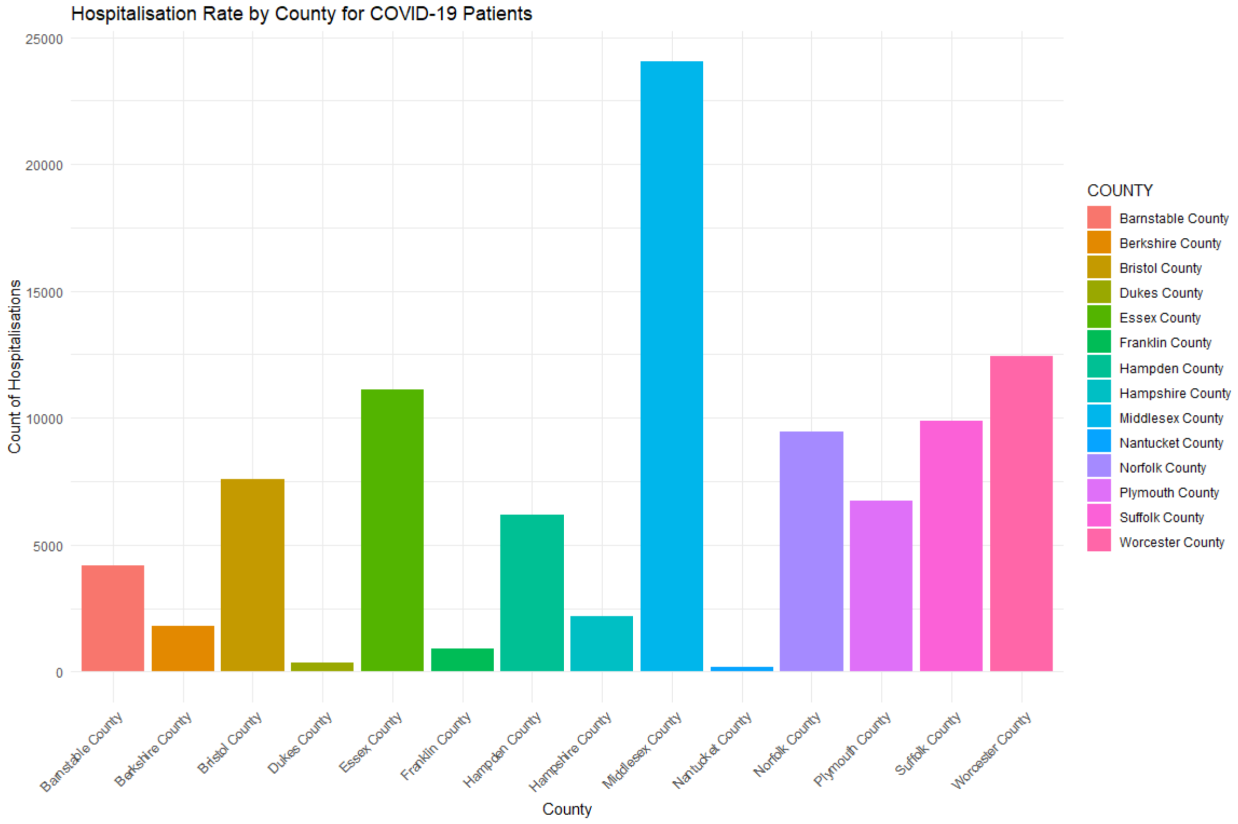

Figure 1. 4 - Bar Graph of the Hospitilisation rate by County for COVID-19 Patients

Hospitilisation Rate by County for COVID-19 Patients:

The bar graph clearly highlights the disparity in hospitalisation rates across different counties. Middlesex County has the highest number of hospitalisations, followed by Worcester and Essex Counties. This visualisation effectively showcases the geographic distribution of hospitalisations, suggesting that certain counties are more affected than others. The use of different colours for each county improves readability and allows for easy comparison.

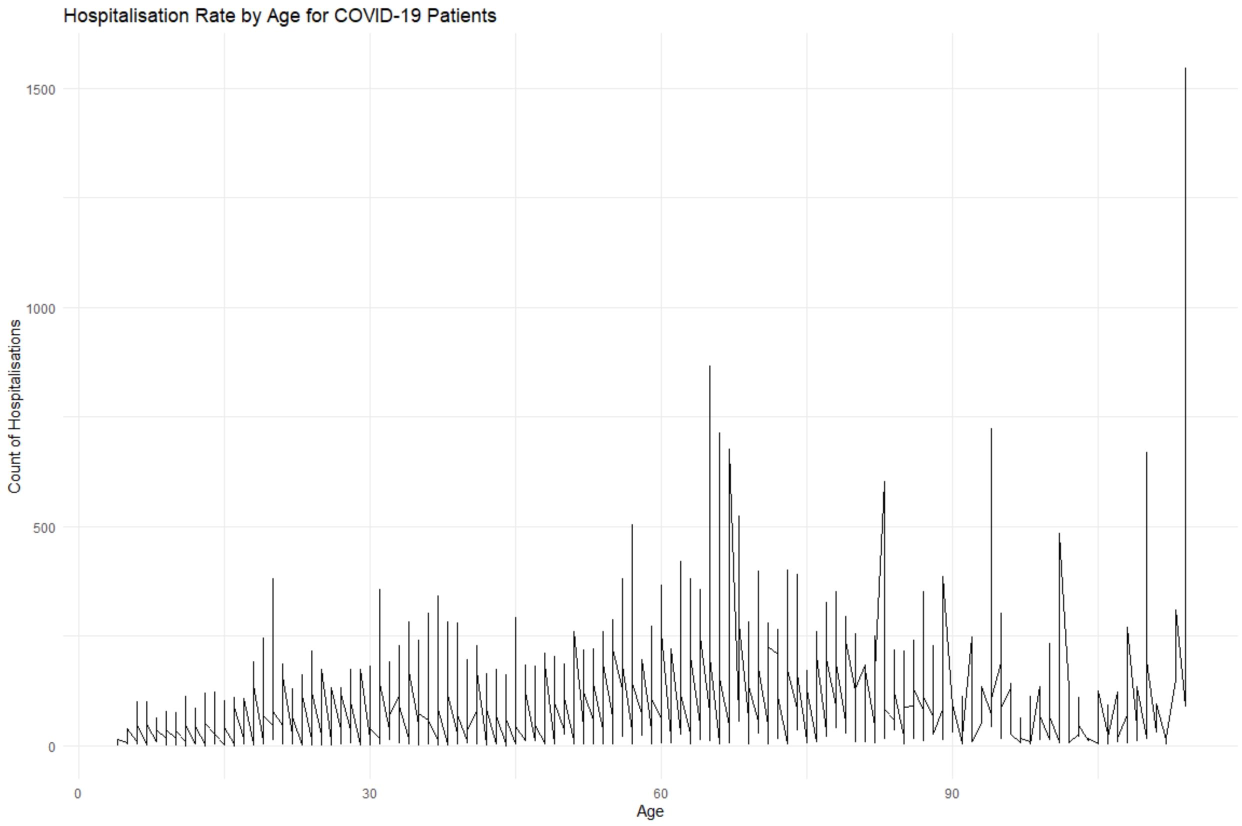

Figure 1. 5 - Line Graph of the Hospitilisation rate by County for COVID-19 Patients

Hospitilisation Rate by Age for COVID-19 Patients:

The line graph displays the hospitalisation rates across different ages. There are noticeable peaks at specific ages, indicating higher hospitalisation rates. This visualisation suggests that certain age groups, particularly older adults, have higher hospitalisation rates. The line graph helps in identifying age-related trends and outliers, providing insights into the age distribution of hospitalisations.

The Influence of Age and County on Hospitalisation Rates

The two factors, age and county, significantly influence hospitalisation rates for COVID-19 patients. The heatmap and line graph both indicate that older age groups have higher hospitalisation rates, which is consistent with the known increased vulnerability of older adults to severe COVID-19 outcomes, as outlined in task 1 (figure 1.1). The bar graph demonstrates that certain counties, such as Middlesex, have higher hospitalisation rates, potentially due to higher population density, urbanisation, and healthcare infrastructure.

Trends and Regression Analysis Explaining the Variation

- **Age**: The line graph reveals that hospitalisation rates increase with age, with noticeable peaks among older adults. This trend is likely due to the higher susceptibility of older individuals to severe illness and complications from COVID-19.
    - **Regression Analysis**: The coefficient for age is 1.20322, which is positive and highly significant (p-value < 0.001). This supports our understanding that older patients are admitted into hospital with COVID at higher rates.

- **County**: The bar graph and heatmap show that counties like Middlesex have significantly higher hospitalisation rates. This could be attributed to factors such as higher population density, socioeconomic conditions, and availability of healthcare facilities.
    - **Regression Analysis:** The regression analysis for counties is less conclusive due to the lack of detailed information about each County (population density, social behaviours, legal obligations such as social distancing) to attribute any sort of correlation to them. However, Suffolk County showed a Coefficient of -1.00734 (p-value < 0.05) suggesting that the increase in Hospitilisation rates with age is less pronounced in certain cases.

- **R-squared**: The multiple R-squared value is 0.3229, and the adjusted R-squared value is 0.3055 which means that approximately 30.55% of the variability in hospitalisation rates is explained by age, county, and their interactions.

- **F-statistic**: The F-statistic is 18.55 with a p-value < 2.2e-16, indicating that the model is statistically significant overall.

These findings are important but do not paint the entire picture. Older age groups consistently show higher hospitalisation rates, and certain counties exhibit significantly higher rates. However, the interaction between age and county is complex and requires further investigations to better understand the relationship between these two variables, and if they are truly statistically significant.

Task 4: Write the code to investigate the characteristics of patients (confirmed or suspected) who recover from COVID-19 compared to those who don't. Consider factors such as demographics (age, gender, zip code), symptoms, and timeline of diagnosis and recovery. Analyse how these factors impact the recovery outcome.

```
# Load necessary libraries
library(reader) # for reading CSV files
library(dplyr) # for data manipulation
library(ggplot2) # for data visualisation
library(tidyr) # for tidying data

# Load the CSV files
conditions <- read_csv("Documents/conditionsUG.csv")
encounters <- read_csv("Documents/encountersUG.csv")
patients <- read_csv("Documents/patientsUG.csv

# Data cleaning
# Removing any unnecessary data (column A ('...1'))
conditions <- conditions %>% select (-'...1')
encounters <- encounters %>% select (-'...1')
patients <- patients %>% select (-'...1')
```

To ensure we use only datasets which are relevant to our task, I have filtered the 'conditions' data set to only identify COVID or Suspected COVID-19 Patients, then included patient information (gender, age, etc..) into the dataset for more context.

```
# Identify COVID-19 or Suspected COVID-19 patients
covid_conditions <- conditions %>%
filter(grepl("COVID|Suspected COVID", DESCRIPTION, ignore.case = TRUE))

# Join with patients data to get demographic information and recovery status
covid_patients <- covid_conditions %>%
inner_join(patients %>% select(Id, GENDER, BIRTHDATE, DEATHDATE, ZIP, COUNTY, STATE), by
= c("PATIENT" = "Id"))

# Calculate age (reformatted the 'BIRTHDATE' to an Age, so that it is easier to analyse)
covid_patients <- covid_patients %>%
mutate(AGE = as.integer(format(Sys.Date(), "%Y")) - as.integer(substr(BIRTHDATE, 1, 4)))
```

I wasn't able to find a direct way to determine whether a patient has recovered or not, so I have opted to use death date as a metric. If a patient had covid-19 or suspected, and has no registered death date, they will be marked as "Recovered" and vice versa.

```
# Determine recovery status (this is determined by
covid_patients <- covid_patients %>%
mutate(RECOVERY_STATUS = if_else(is.na(DEATHDATE), "Recovered", "Not Recovered"))

# Join with encounters data to get encounter class information
covid_encounters <- covid_patients %>%
inner_join(encounters %>% select(PATIENT, REASONDESCRIPTION, START, STOP), by =
"PATIENT")

# Filter out patients with no encounter data
covid_encounters <- covid_encounters %>%
filter(!is.na(REASONDESCRIPTION))

# Analysis of demographics and symptoms by recovery status
demographics_analysis <- covid_encounters %>%
group_by(RECOVERY_STATUS, GENDER, AGE, ZIP, COUNTY, STATE) %>%
summarise(count = n()) %>%
ungroup()

symptoms_analysis <- covid_encounters %>%
group_by(RECOVERY_STATUS, REASONDESCRIPTION) %>%
summarise(count = n()) %>%
ungroup()

# Visualisation of Age by recovery status
ggplot(demographics_analysis, aes(x = AGE, y = count, fill = RECOVERY_STATUS)) +
geom_bar(stat = "identity", position = "dodge") +
labs(title = "Impact of Age on COVID-19 Recovery", x = "Age", y = "Count") +
theme_minimal()

# Visualisation of Gender by recovery status
ggplot(demographics_analysis, aes(x = GENDER, y = count, fill = RECOVERY_STATUS)) +
geom_bar(stat = "identity", position = "dodge") +
labs(title = "Impact of Gender on COVID-19 Recovery", x = "Gender", y = "Count") +
theme_minimal()
```

```
# Visualisation of County by recovery status
ggplot(demographics_analysis, aes(x = COUNTY, y = count, fill = RECOVERY_STATUS)) +
geom_bar(stat = "identity", position = "dodge") +
labs(title = "Impact of County on COVID-19 Recovery", x = "County", y = "Count") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Due to the large number of symptoms in the dataset, this bar graph will become unreadable if I was to include all the symptoms. Thus, I have narrowed it down to the 15 most common symptoms instead.

```
# Identify the top 15 symptoms
top_15_symptopms <- symptoms_analysis %>%
arrange(desc(count)) %>%
head(15)

# Visualise 15 most common symptoms impact on recovery outcome
ggplot(symptoms_analysis, aes(x = reorder(REASONDESCRIPTION, count), y = count, fill =
RECOVERY_STATUS)) +
geom_bar(stat = "identity", position = "dodge") +
labs(title = "Impact of Symptoms on COVID-19 Recovery", x = "Symptom", y = "Count") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

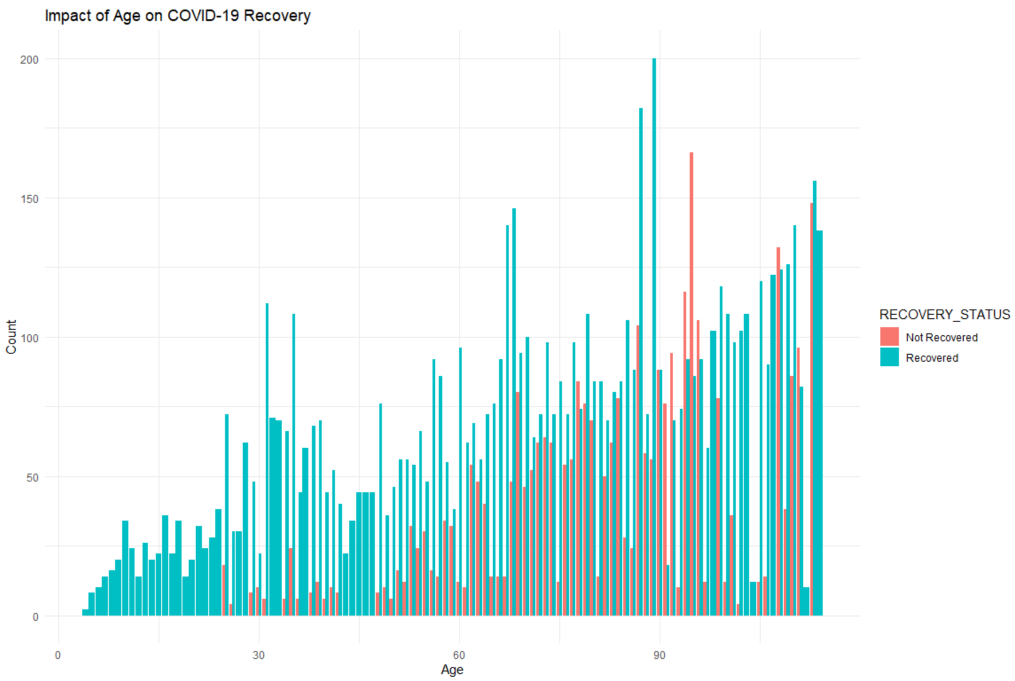

Figure 1. 6 - Impact of Age on COVID-19 Recovery

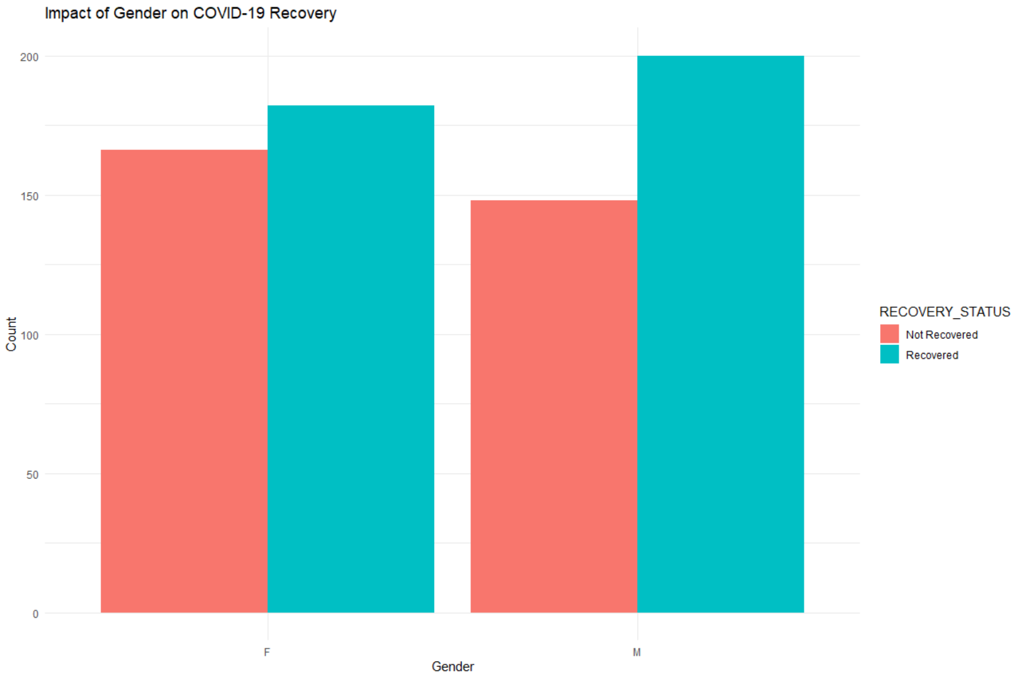

Figure 1. 7 - Impact of Gender on COVID-19 Recovery

GitHub Repository: https://github.com/Qukee/COMP1013_Assignment

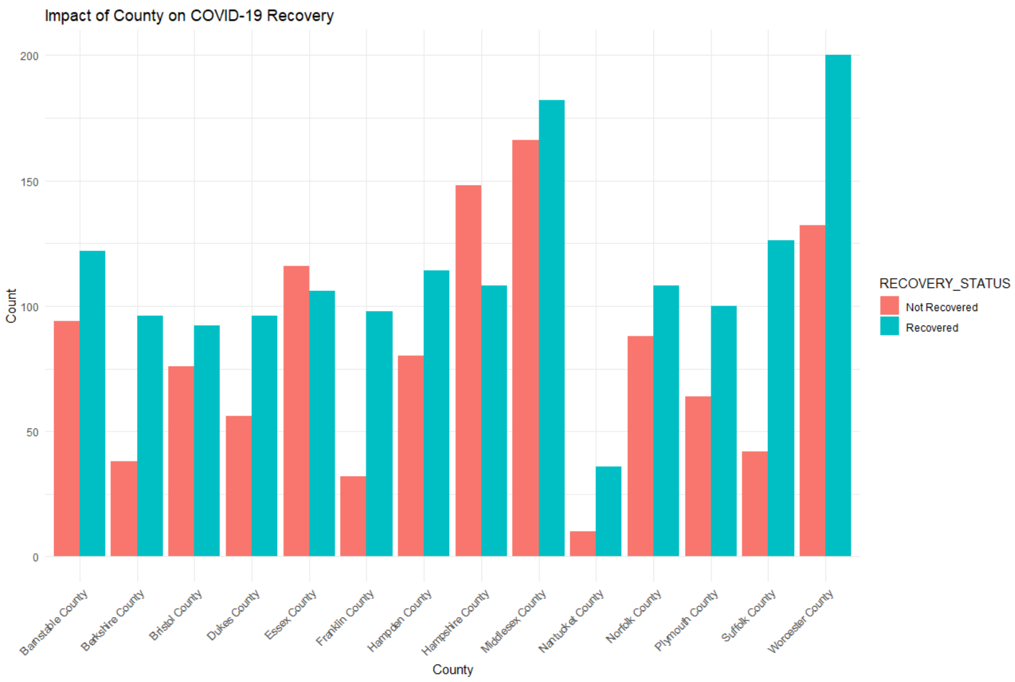

Figure 1. 8 - Impact of County on COVID-19 Recovery

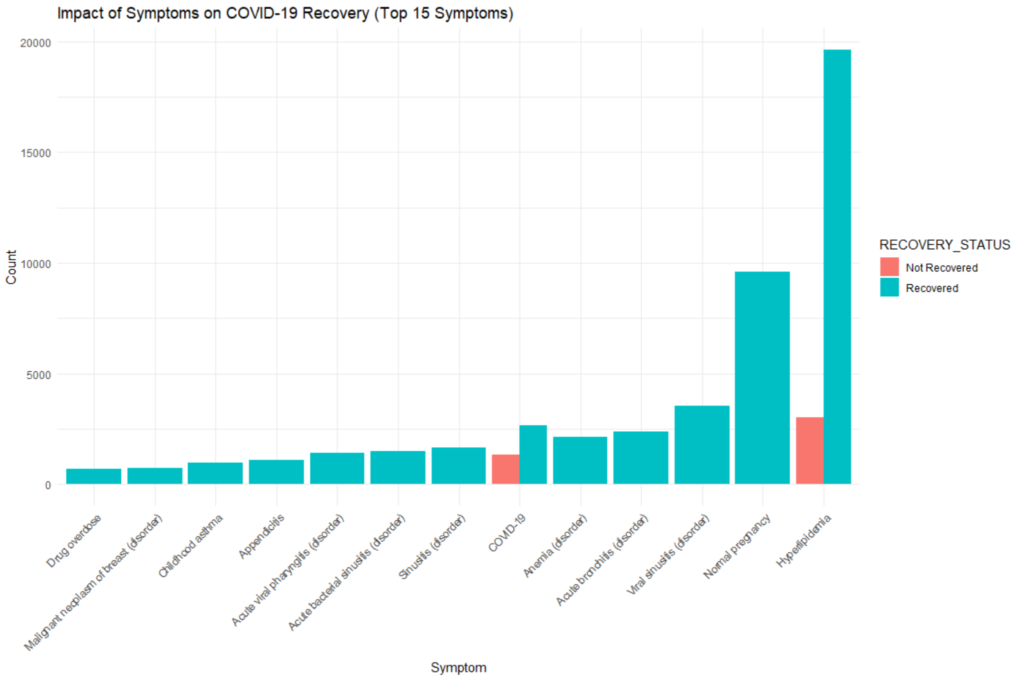

Figure 1. 9 - Impact of Symptoms on COVID-19 Recovery

GitHub Repository: *https://github.com/Qukee/COMP1013_Assignment*

Evaluation of the Characteristics of COVID-19 Patients:

The data clearly indicates that younger individuals (under 60, shown in figure 1.6) have higher recovery rates, with a large shift in *'not recovered'* patients surging at 60 years and older – with the number of recovered cases significantly outweighing the non-recovered cases below 60 years of age. As age increases, especially beyond 60, the recovery rates decline, and the number of non-recovered cases rises. This trend aligns with our previous findings that older adults are more susceptible to severe outcomes from COVID-19 due to factors like weakened immune systems and the presence of comorbidities.

Gender:

The gender-based analysis (figure 1.7) reveals that both males and females have higher recovery counts compared to non-recovery counts. However, the difference is more pronounced in males, indicating a higher recovery rate among males compared to females. This observation may warrant further investigation into the role of biological, social, and behavioural factors that might contribute to these differences.

County:

Recovery outcomes vary significantly across different counties. Some counties, such as Worcester County, show a markedly higher number of recovered cases, while others, like Hampden County, display a more balanced or even higher count of non-recovered cases. These discrepancies could be attributed to variations in healthcare infrastructure, public health interventions, socioeconomic status, population density, and other local factors that affect healthcare access and quality.

Symptoms:

The analysis of symptoms highlights that certain conditions and symptoms have a significant impact on recovery outcomes. Severe symptoms, such as those directly associated with COVID-19, show a notable number of non-recovered cases. On the other hand, common chronic conditions like hypertension and hyperlipidaemia have high recovery counts, suggesting that while these conditions are prevalent, they might not severely impact recovery if managed effectively. This emphasises the importance of symptom severity and the management of underlying health conditions in determining recovery outcomes.

Conclusion:

The combined analysis of age, gender, county, and symptoms provides a comprehensive understanding of the factors influencing COVID-19 recovery. Younger age groups and males show better recovery rates, while older individuals and certain counties exhibit higher non-recovery rates. Symptom severity plays a critical role in recovery outcomes, underscoring the need for targeted healthcare interventions and resource allocation to improve recovery rates, especially among vulnerable populations. These insights can inform public health strategies and healthcare policies aimed at mitigating the impact of COVID-19 across different demographics and regions.

GitHub Repository: https://github.com/Qukee/COMP1013_Assignment