PROGRAMMING FOR DATA SCIENCE ASSIGNMENT

Mathews Abraham

Programming for Data Science COMP 7024

DECLARATION

Name: Mathews Abraham

Student ID: 22075779

Subject Code: COMP 7024

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

Programming for Data Science

Mathews Abraham

To start with the analysis, first the necessary libraries were called using the library function in R, there are mainly three libraries used throughout our analysis, ggplot, tidyverse and dplyr.

```
#loading the libraries
library(ggplot2)

library(tidyverse)

library(dplyr)
library(knitr)

## Warning: package 'knitr' was built under R version 4.3.3
```

Question 1.Write the code to analyse the distribution of COVID patients (confirmed or suspected) across countries. 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.

```
#loading datasets
patients1 = read.csv("patientsPG1.csv") #to load the csv files
patients2 = read.csv("patientsPG2.csv")
patients = rbind(patients1,patients2)

conditions1 = read.csv("ConditionsPG1.csv")
conditions2 = read.csv("ConditionsPG2.csv")
condition = rbind(conditions1,conditions2)
conditions = subset(condition, DESCRIPTION %in% c("Suspected COVID-
19","COVID-19"), drop = TRUE)#subset dataset based on the number of covid
patients
```

As the task does not specifically mention to investigate the distribution in a particualar dataset, we have used the whole datasets here.

Here we have four datasets, two of them contains patient demographic data and others contain information about patient conditions. First, we read all the data using the 'read.csv' function in R, we used this function because the data is in a csv format. Then the dataset with demographic information was combined into one and the ones describing conditions into another single one.

First we are analyzing the distribution of covid patients across the counties, the patients data has information about the counties and the conditions have the description of the diagnosis. Since there were other information under the 'DESCRIPTION' variable but we

are only interested if the patients have suspected covid or not, we filtered the dataset with the description of diagnosis either covid-19 or not.

```
county = subset(patients, select = c("Id","COUNTY","BIRTHDATE"))
description = subset(conditions, select = c("PATIENT","DESCRIPTION"))
merged = merge(county,description, by.x = "Id",by.y = "PATIENT", all = FALSE)
#further filtering and merging
```

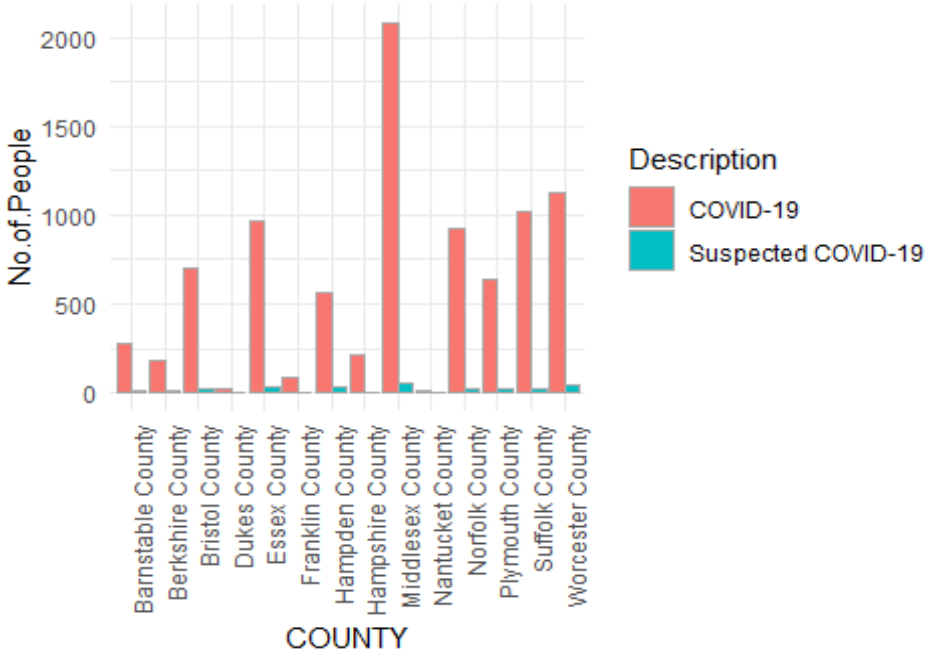
From the patients dataset, as we do not need all the information, it was subsetted with information like Id, county name and birthdate because we need birthdates later for analysis. We also made an other set with just the patient id and description from the conditions and then the description of the diagnosis and county names were combined on the basis of patient id so that we can get the patients their respective counties and if they had covid or not.

```
covid = subset(merged, DESCRIPTION == "COVID-19", drop = TRUE)#filtering
covid and suspected
s_covid = unique(merged$Id[!(merged$Id %in% covid$Id)])
sus_covid = merged[!(merged$Id %in% covid$Id),]
final = rbind(covid,sus_covid)
Q1 = subset(final,select = c("COUNTY","DESCRIPTION","BIRTHDATE"))
```

Now, we had to do some additional filtering, all the patients were suspected for covid first, but some of them did not turn covid positive. In the dataset, the ids are repeated for those who were suspected first and had covid later, so we had to remove these duplicate ids. First we took all the ids which had covid, this means these ids are repeated any way, then we filter the suspected covid ids then later further we filtered the unique suspected ids and finally binded back the unique ids together.

```
Q1_table = table(Q1$COUNTY,Q1$DESCRIPTION)
Q1.data = as.data.frame(Q1_table)
names(Q1.data) = c("COUNTY","DESCRIPTION","NO.OF.PEOPLE")
Q1.data  = Q1.data%>%
  arrange(desc(NO.OF.PEOPLE))

ggplot(Q1.data, aes(x = COUNTY, y = NO.OF.PEOPLE, fill = DESCRIPTION)) +
  geom_bar(stat = "identity",position = position_dodge(0.9),color
="darkgrey") +
  labs(title = "",x = "COUNTY",
       y = "No.of.People",
       fill = "Description") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Distribution of Covid and Suspected Covid across Counties



The above barplot is a visualisation of covid patients and suspected covid patients across different counties. Each bar represents the count of patients in each county. The number people who had just suspection is very less compared to the number of covid patients. The middlesex county with a number of 2082 had the highest number of covid confirmed covid patients whereas, the least number of cases were observed in Nantucket county with just 14 confirmed cases.

People who were suspected but did not get affected was also highest in Middlesex with a number of 53 whereas in Dukes, every person who were suspected later turned covid positive.

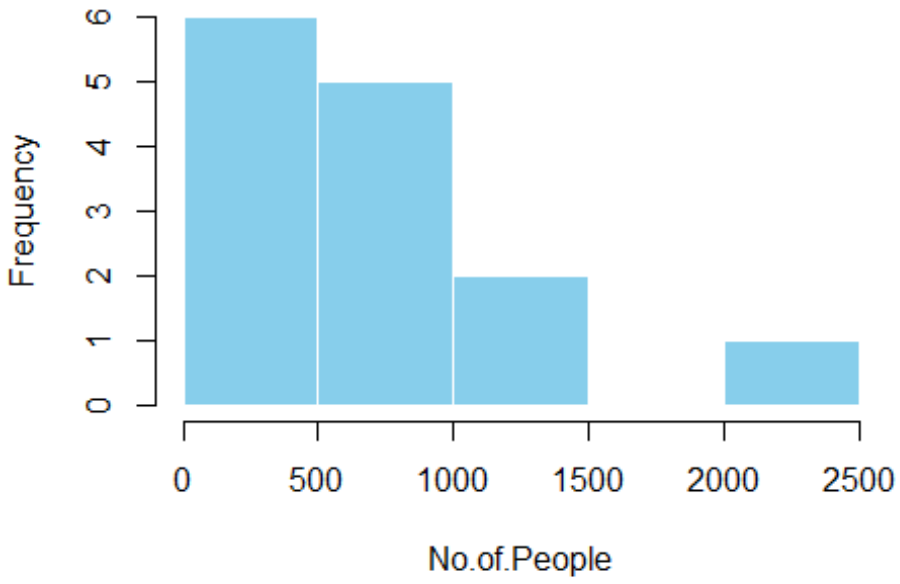
```
co.county = subset(Q1.data, DESCRIPTION == "COVID-19")#county and frquency of
covid patients
kable(co.county, caption = "COVID-19 patients across counties")#table of
covid
```

COVID-19 patients across counties

	COUNTY	DESCRIPTION	NO.OF. PEOPLE
1	Middlesex County	COVID-19	2082
2	Worcester County	COVID-19	1130
3	Suffolk County	COVID-19	1018
4	Essex County	COVID-19	967
5	Norfolk County	COVID-19	922
6	Bristol County	COVID-19	702
7	Plymouth County	COVID-19	639
8	Hampden County	COVID-19	561
9	Barnstable County	COVID-19	278
10	Hampshire County	COVID-19	216
11	Berkshire County	COVID-19	178
12	Franklin County	COVID-19	89
18	Dukes County	COVID-19	24
22	Nantucket County	COVID-19	14

```
hist(co.county$NO.OF.PEOPLE, main = "",
     xlab = "No.of.People", col = "skyblue",border = "white")#hist of
distribution
```

Distribution of covid across counties



A table with the number of covid patients was created using the 'kable' function in R and the distribution of covid patients across counties was plotted using a histogram. There was only one county with number of patients between 2000 and 2500, six counties had 0-500 number of confirmed cases and the remaining counties had numbers between 500 to 1500.

```
sus.county = subset(Q1.data, DESCRIPTION == "Suspected COVID-19")#county with
suspected case freq
kable(sus.county, caption = "Suspected COVID-19 patients across counties")
```

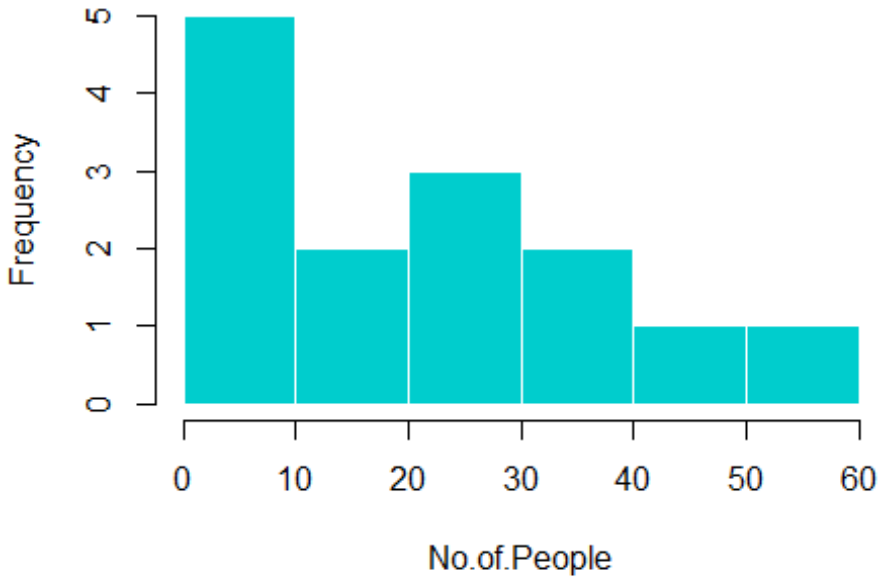
Suspected COVID-19 patients across counties

	COUNTY	DESCRIPTION	NO.OF. PEOPLE
13	Middlesex County	Suspected COVID-19	53
14	Worcester County	Suspected COVID-19	47
15	Essex County	Suspected COVID-19	32
16	Hampden County	Suspected COVID-19	32
17	Norfolk County	Suspected COVID-19	28
19	Suffolk County	Suspected COVID-19	23
20	Bristol County	Suspected COVID-19	22
21	Plymouth County	Suspected COVID-19	19
23	Barnstable County	Suspected COVID-19	11
24	Berkshire County	Suspected COVID-19	9
25	Hampshire County	Suspected COVID-19	4

	COUNTY	DESCRIPTION	NO.OF. PEOPLE
26	Franklin County	Suspected COVID-19	3
27	Nantucket County	Suspected COVID-19	3
28	Dukes County	Suspected COVID-19	0

```
hist(sus.county$NO.OF.PEOPLE, main = "",
     xlab = "No.of.People", col = "cyan3",border = "white")
```

Distribution of Suspected Covid across counties



The same process was carried out with the people who were just suspected. Most of the counties have a very lower number of people who were just suspected. As low these numbers across the counties means that most of them turned covid positive later, the higher number means there are more people who were just suspected but actually did not had covid.

Now we have to find the distribution of covid patients across different age group, for this we have to categorise the ages into four different groups 0-18, 19-35, 36-50 and 51 +. Since we do not have the direct ages of patients, instead we calculated it from their date of birth using a function.

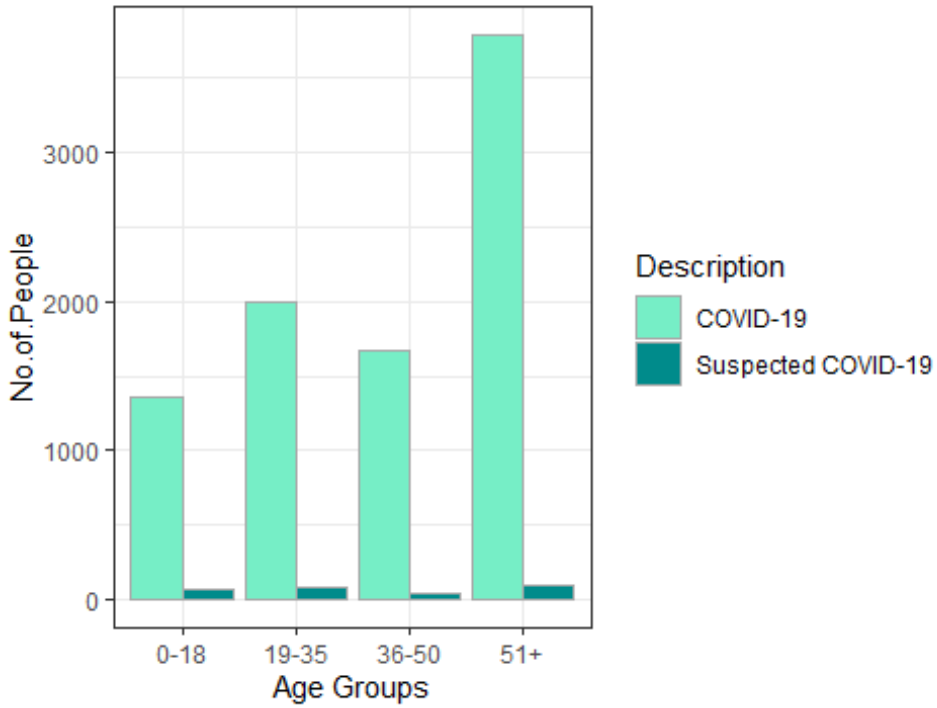
```
age = subset(Q1, select = c("BIRTHDATE","DESCRIPTION"))#filtering birthdate
for age calculation

age_calculator = function(dob) {
  today = Sys.Date()
  birth.date = as.Date(dob)
```

```
  birth.year = as.numeric(format(birth.date, "%Y"))
  current_year = as.numeric(format(today, "%Y"))

  age = current_year - birth.year

  bday = format(today, "%m-%d") < format(birth.date, "%m-%d")
  age[bday] = age[bday] - 1

  return(age)
}#function returning age from current date
```

We created a function called 'age calculator' which returns the ages of patients if they were alive today. We did not stress much into the exact age of patients because we were focusing on their age groups. Some people were above 110 years old as per the calculations but we grouped them all into one category called 51+. Based on the age groups, we had four categories 0-18, 19-35, 36-50 and above 51.

```
age$AGE = age_calculator(age$BIRTHDATE)#creating age categories
age = age %>%
  mutate(Age.group = case_when(
    AGE>=0 & AGE<=18 ~ "0-18",
    AGE>=19& AGE<=35 ~ "19-35",
    AGE>=36& AGE<=50 ~ "36-50",
    AGE>=51 ~ "51+"
  ))

#age = subset(age,select = c("Age.group"))
age.table = table(age$DESCRIPTION,age$Age.group)
age.data = as.data.frame(age.table)#making age and description as table for
ease to plot
names(age.data) = c("DESCRIPTION","AGE.GROUP","NO.OF.PEOPLE")
age.data  = age.data%>%
  arrange(desc(NO.OF.PEOPLE))

ggplot(age.data, aes(x = AGE.GROUP, y = NO.OF.PEOPLE, fill = DESCRIPTION)) +
  geom_bar(stat = "identity",position = position_dodge(0.9),colour=
"darkgrey") +
  labs(title = "",x = "Age Groups",
       y = "No.of.People",
       fill = "Description") +
  theme_bw()+
  scale_fill_manual(values = c("aquamarine2","darkcyan"))#barplot using
ggplot
```

Barplot showing Covid and Suspected covid across different age groups



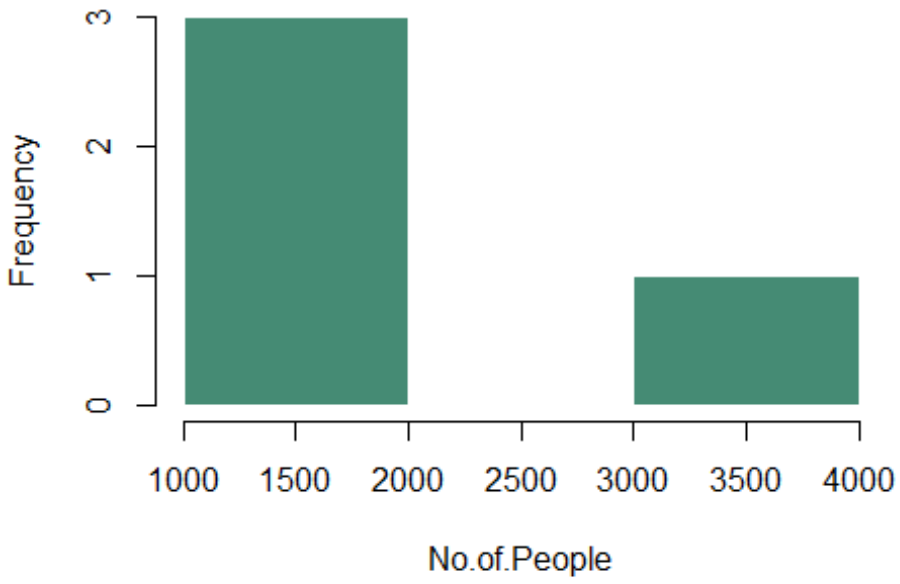
The above barplot shows the number of people belonging to two different age groups who had covid and just suspision. Most of the reported cases were above 51 years old and people who were below 18 were least affected. Even though the most affected category was 51+, the most escaped numbers are also from the same, this might have occured because of the higher number.

```
co.age = subset(age.data, DESCRIPTION == "COVID-19")
kable(co.age, caption = "")
```

DESCRIPTION	AGE.GROUP	NO.OF. PEOPLE
COVID-19	51+	3790
COVID-19	19-35	1999
COVID-19	36-50	1670
COVID-19	0-18	1361

```
hist(co.age$NO.OF.PEOPLE, main = "",
     xlab = "No.of.People", col = "aquamarine4",border = "white")#hist of
agegroup and no of people
```

Distribution of Covid Patients across different age groups



A histogram was plotted to depict the distribution of age group and people with covid. Three age groups had number of people between 1000- 2000 whereas in one age group the number of people were above 3000.

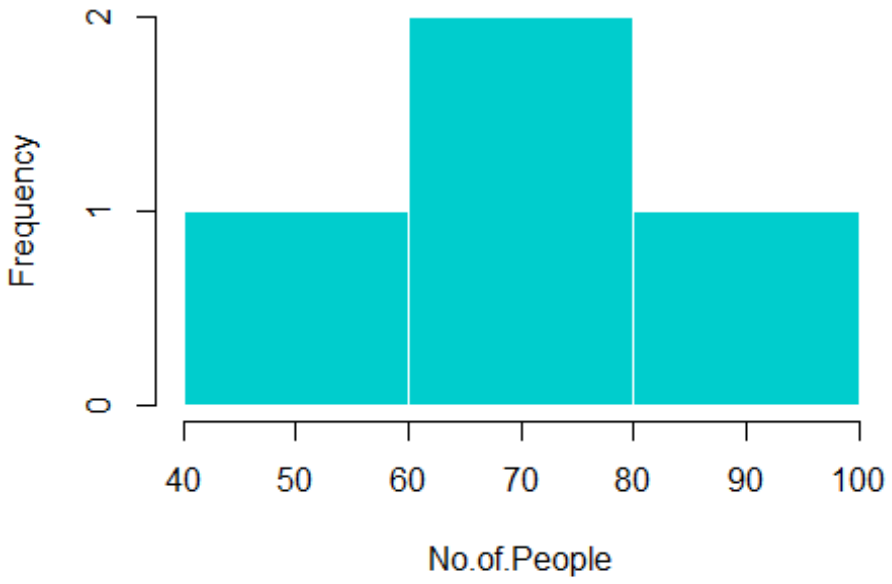
```
sus.age = subset(age.data, DESCRIPTION == "Suspected COVID-19")
kable(sus.age, caption = "Suspected COVID-19 patients across different age
groups")
```

Suspected COVID-19 patients across different age groups

	DESCRIPTION	AGE.GROUP	NO.OF. PEOPLE
5	Suspected COVID-19	51+	100
6	Suspected COVID-19	19-35	75
7	Suspected COVID-19	0-18	65
8	Suspected COVID-19	36-50	46

```
hist(sus.age$NO.OF.PEOPLE, main = "",
     xlab = "No.of.People", col = "cyan3",border = "white")#hist of age group
and suspected patients.
```

Distribution of suspected covid patients across different age groups



Another histogram for the patients who were just suspected was plotted as well. It shows the highest number of people who were just suspected falls in range from 60-80 and two of the age groups had these much number of suspected patients.

Question.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.

```
Q2 = subset(condition,select = c("PATIENT","DESCRIPTION"))#filter patient and
description
Q2= Q2%>%
  filter(PATIENT %in% final$Id)#filter ids with covid
```

To find whether the symptoms related to the patients differ by gender, first we filtered the conditions and the ids from the condition datasets. Even here, we are using both datasets condition1 and condition2 because it is not specified to analyse a specific dataset.

After taking the necessary information, it was further filtered on the basis of ids which were contracted covid or suspected.

```
symptom.count= Q2 %>%
  count(DESCRIPTION)
```

```
top.12 = symptom.count %>%
  top_n(12)

## Selecting by n

top.12 = top.12%>%
  arrange(desc(n))
top.10 = top.12%>%
  slice(-1,-2)
#top.10
```

Then we counted the number of all symptoms and took the top 12 symptoms associated with the patients. Out of the top 12, the most occuring were covid-19 and suspected covid-19 but we did not need those two, so we removed those two numbers and created a table with the top 10 common symptoms as shown above. Fever and cough are the most common symptoms followed by taste loss, fatigue and so on.

```
Q2.1 = patients%>%
  filter(Id %in% Q2$PATIENT)

gender = subset(Q2.1,select = c("Id","GENDER"))
gender.des = merge(gender,Q2, by.x = "Id",by.y = "PATIENT", all = FALSE)
```

Now we have the common symptoms but we don't know if they differ by gender, for that first we filtered the ids associated with the genders from the patients dataset and matched these ids and the ids of patients who have covid resulting in a new a data frame with Id, gender and symptoms of the covid patients.

```
symptom.gender = gender.des %>%
  count(DESCRIPTION,GENDER)#filtering gender and symptom

top.in.gender = symptom.gender%>%
  group_by(GENDER)%>%
  slice_max(order_by = n, n =12, with_ties = TRUE)#getting top 12 symtoms

reshaped = spread(top.in.gender,GENDER,n)#shaping in a table format

reshaped = reshaped%>%
  arrange(desc(c(reshaped$F)))

top.10.g = reshaped%>%
  slice(-1,-2)#removing covid 19 and suspected covid 19
```

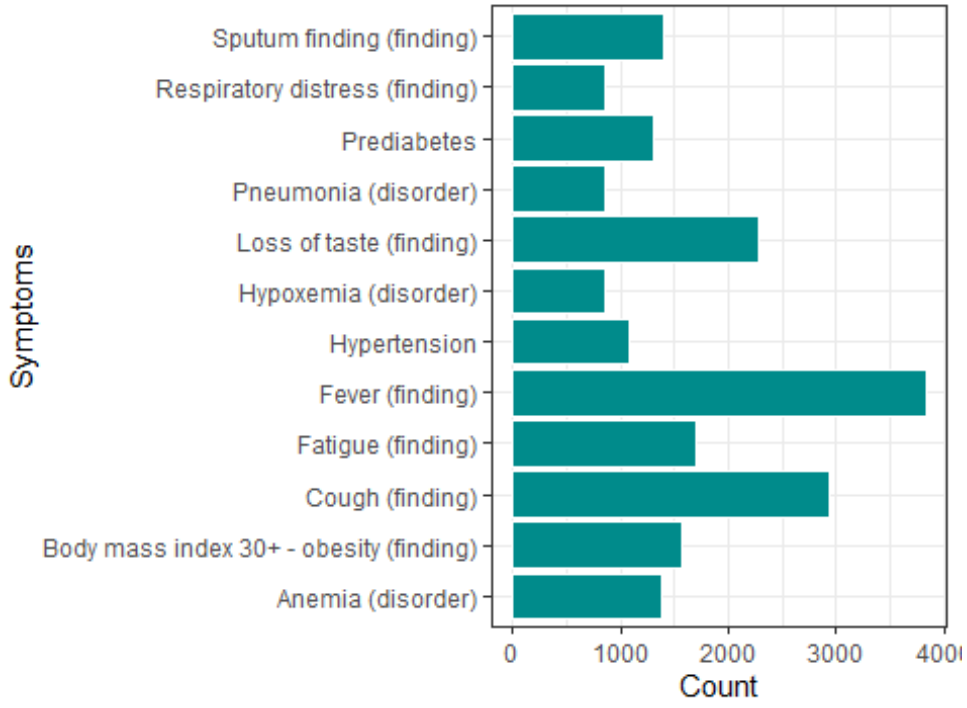
As all the symptoms were not relevant for the analysis, we took the top 10 symptoms associated with males and females.

```
male.10.symptom = subset(top.10.g,select = c("DESCRIPTION","M"))#subset male
symtoms
male.10.symptom = na.omit(male.10.symptom)
kable(male.10.symptom,caption = "")
```

DESCRIPTION	M
Fever (finding)	3833
Cough (finding)	2936
Loss of taste (finding)	2271
Body mass index 30+ - obesity (finding)	1572
Fatigue (finding)	1695
Sputum finding (finding)	1406
Prediabetes	1302
Hypertension	1084
Anemia (disorder)	1387
Hypoxemia (disorder)	865
Pneumonia (disorder)	865
Respiratory distress (finding)	865

```
ggplot(male.10.symptom, aes(x = DESCRIPTION,y = M, fill = M)) +
  geom_bar(stat = "identity",fill = "cyan4",colour= "white") +
  labs(title = "",x = "Symptoms",
       y = "Count") +
  theme_bw()+coord_flip()
```

Count of different symptoms in Male



The above table shows the top 10 symptoms in male covid patients and the barplot visualises it. As we can see the most commom symptom in males were fever followed by cough and taste loss. Symptoms such as hypoxemia, pneumonia and respiratory diseases

were the least common with 865 each in the top 10 list. Obsesity and fatigue were also some of the common symptoms in males.

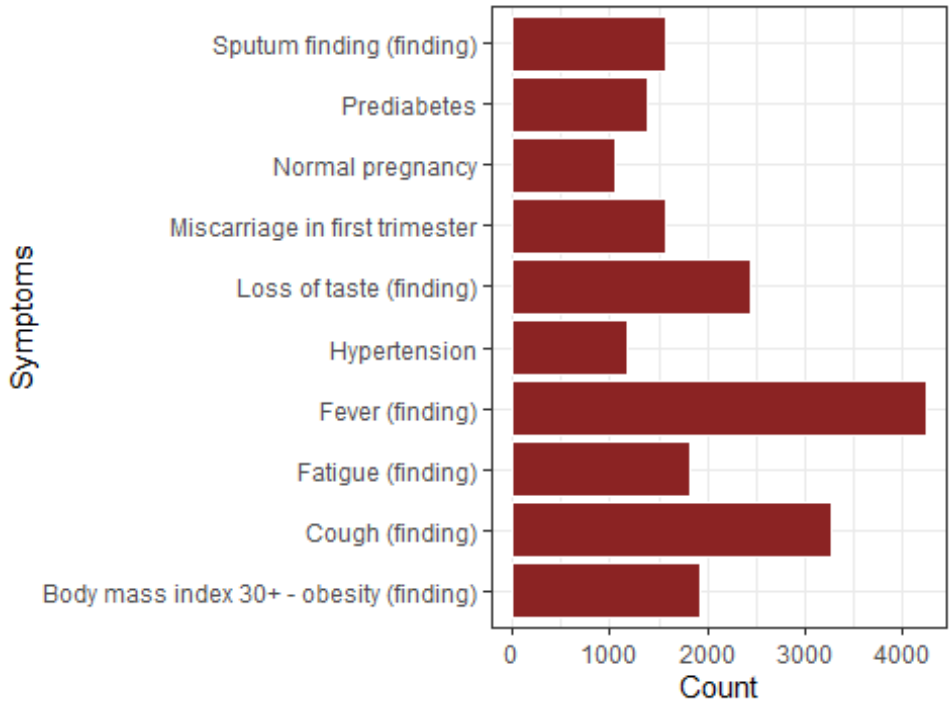
```
female.10.symptom = subset(top.10.g,select = c("DESCRIPTION","F"))#subset
female symtoms
female.10.symptom = na.omit(female.10.symptom)
kable(female.10.symptom,caption = "Top 10 symptoms in female covid patients")
```

Top 10 symptoms in female covid patients

DESCRIPTION	F
Fever (finding)	4250
Cough (finding)	3266
Loss of taste (finding)	2440
Body mass index 30+ - obesity (finding)	1918
Fatigue (finding)	1821
Miscarriage in first trimester	1580
Sputum finding (finding)	1564
Prediabetes	1377
Hypertension	1183
Normal pregnancy	1063

```
ggplot(female.10.symptom, aes(x = DESCRIPTION,y = F, fill = F)) +
  geom_bar(stat = "identity",fill = "brown4",colour= "white") +
  labs(title = "",x = "Symptoms",
       y = "Count") +
  theme_bw()+coord_flip()
```

Count of female symptoms



When it comes to females, same fever, cough and taste loss were common affecting around 4250,3266 and 2440 number of females respectively. Unlike men, women had some other symptoms such as miscarriage and normal pregnancy, not sure if they are symptoms but an evident number of women miscarriage and normal pregnancy. There are also some other symptoms such as hypertension and prediabetes in both men and women.

Question.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. Does this result vary between the 2 datasets?

The two factors that was chosen here for the analysis are gender and race. This analysis is conducted separate for both the datasets as we have to finally compare between both. The information about hospitalisation rate is in the encounters dataset and though there are six categories under the encounterclasss, we have only used four of them which are ambulatory, emergency, inpatient and urgentcare (as per the question). As the first step both the datsets, encounters1 and encounters2 were loaded using the 'read.csv' function.

```
encounters1 = read.csv("encountersPG1.csv")#loading the encounter data
encounters2 = read.csv("encountersPG2.csv")
```

```
#data1
covid.in.1 = subset(conditions1, DESCRIPTION == "COVID-19", drop = TRUE)
suspected.in.1 = subset(conditions1,DESCRIPTION == "Suspected COVID-19"&
                          !PATIENT %in% covid.in.1$PATIENT)
Q4 = rbind(covid.in.1,suspected.in.1)# filtering  covid patients in data 1

#data2
covid.in.2 = subset(conditions2, DESCRIPTION == "COVID-19", drop = TRUE)
suspected.in.2 = subset(conditions2,DESCRIPTION == "Suspected COVID-19"&
                          !PATIENT %in% covid.in.2$PATIENT)
Qn4 = rbind(covid.in.2,suspected.in.2)#filteration in data 2
```

Before starting the analysis we created two subsets for data 1 and 2 to separate the patients with covid-19 and suspected covid-19 in both datasets. Unlike what was done for first two analysis, this time the patients were taken separate. We will need this separation for the upcoming analyses as well, that is why it is done at this point.

Data 1

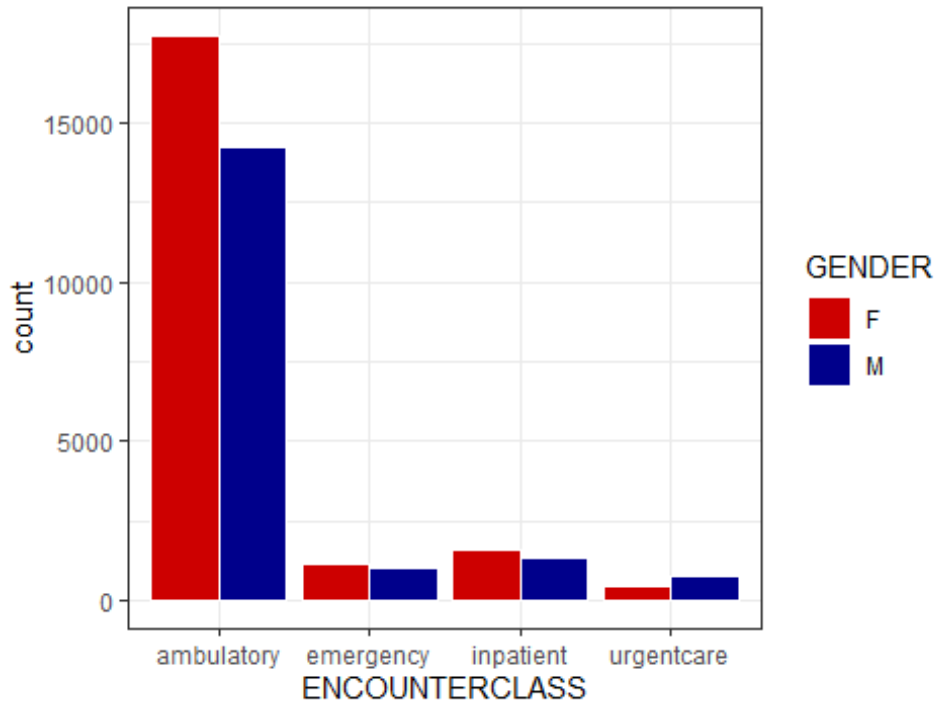
```
#DATA 1
Q3 = encounters1%>%
  filter(PATIENT %in% Q4$PATIENT)#filtering id in encounter data
Q3 = subset(Q3, select = c("PATIENT","ENCOUNTERCLASS"))#filtering encounter
class
Q3 = Q3[Q3$ENCOUNTERCLASS %in%
c("ambulatory","emergency","inpatient","urgentcare"),]
table(Q3$ENCOUNTERCLASS)

##
## ambulatory  emergency  inpatient urgentcare
##      31976       2138       2940       1162

Q3.1 = subset(patients1, select = c("Id","GENDER","RACE"))#filtering id
gender and race
Q3.2 = merge(Q3,Q3.1,by.x = "PATIENT", by.y = "Id",all = FALSE)#merging all
together
```

To start with the analysis, first we filtered the patient ids in the encounter1 dataset on the basis of the covid patients in the dataset 1. Then the four encounter categories were selected as the analysis is conducted for only four encounterclasses. Then from the patients 1 data, we filtered out the two factors that is gender and race for which we are conducting the evaluation. So, once the Ids, encounterclass, gender and race of the covid patients were obtained, everything was combined into a new dataset for the ease.

```
gender.encounter = table(Q3.2$ENCOUNTERCLASS,Q3.2$GENDER)
gender.encounter#contigency table of gender and encounter class

##
##                   F     M
##   ambulatory 17740 14236
##   emergency   1133  1005
```

```
##   inpatient   1614  1326
##   urgentcare   427   735

prop.gender.1 = prop.table(gender.encounter,margin = 1)*100
prop.gender.1#proportion of each gender under each class

##
##                      F        M
##   ambulatory 55.47911 44.52089
##   emergency  52.99345 47.00655
##   inpatient  54.89796 45.10204
##   urgentcare 36.74699 63.25301

ggplot(Q3.2, aes(x = ENCOUNTERCLASS, fill = GENDER)) +
  geom_bar(position = position_dodge(width = 0.9), color = "white") +
  labs(title = "") +
  theme_bw()+ scale_fill_manual(values = c("red3","blue4"))
```

Distribution of encounter classes among genders



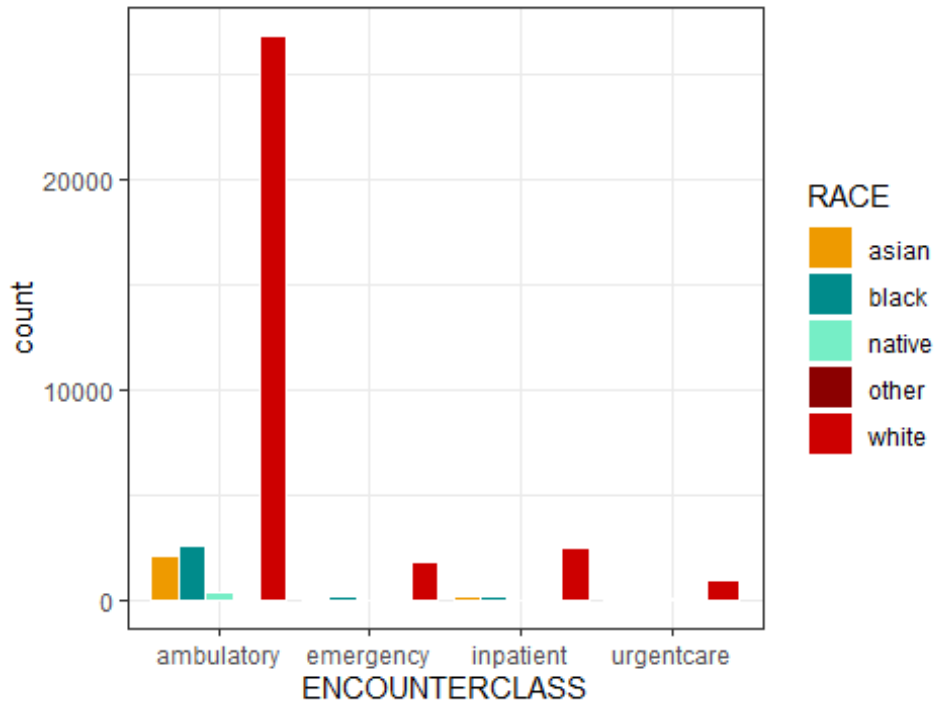
Here we are comparing the genders with different encounter classes. To understand this more easily a proportion table was created using the 'prop.table' function in R. This gives the proportion of each gender in different encounter classes.

Ambulatory services were the highest used among all, more than 15000 males and females used this services. The number of females are higher in three out of the four classes.

Ambulatory, emergency, inpatient all these services were more used by women rather than mean. When it comes to urgent care, it was more used by men than women.

```
race.encounter = table(Q3.2$ENCOUNTERCLASS,Q3.2$RACE)
race.encounter

##
##              asian black native other white
##   ambulatory  2088  2613    384    66 26825
##   emergency    139   174     16     4  1805
##   inpatient    198   193     12     2  2535
##   urgentcare   136    82     17     0   927

prop.race.1 = prop.table(race.encounter,margin = 1)*100
prop.race.1#proportion of race

##
##                    asian       black      native       other       white
##   ambulatory  6.52989742  8.17175382  1.20090068  0.20640480 83.89104328
##   emergency   6.50140318  8.13844715  0.74836296  0.18709074 84.42469598
##   inpatient   6.73469388  6.56462585  0.40816327  0.06802721 86.22448980
##   urgentcare 11.70395869  7.05679862  1.46299484  0.00000000 79.77624785

ggplot(Q3.2, aes(x = ENCOUNTERCLASS, fill = RACE)) +
  geom_bar(position = position_dodge(width = 0.9), color = "white") +
  labs(title = "") +
  theme_bw()+
  scale_fill_manual(values =
c("orange2","darkcyan","aquamarine2","darkred","red3"))
```

Distribution of encounter classes among races



Another factor race, when compared with different encounter classes, whites have the highest numbers among all. With a very high significance, in all these categories, the highest numbers are people from white race. The other numbers are not even closer to the number of white people in each category.

After whites, Asians and blacks have the highest usage of these services compared to the others. Though the numbers are small, they come in second and third positions in the utilization of these ambulatory, emergency, inpatient and urgentcare services.

Data 2

```
#DATA 2
Qn3 = encounters2%>%#same processing as data 1
  filter(PATIENT %in% Qn4$PATIENT)
Qn3 = subset(Qn3, select = c("PATIENT","ENCOUNTERCLASS"))
Qn3 = Qn3[Qn3$ENCOUNTERCLASS %in%
c("ambulatory","emergency","inpatient","urgentcare"),]
table(Qn3$ENCOUNTERCLASS)

##
## ambulatory  emergency  inpatient urgentcare
##      25897       1676       2291       1120
```

```r
Qn3.1 = subset(patients2, select = c("Id","GENDER","RACE"))

Qn3.2 = merge(Qn3,Qn3.1,by.x = "PATIENT", by.y = "Id",all = FALSE)
```

The same steps were steps and analysis were followed for the dataset 2 as well. Even in the second data, the highest users are for the ambulatory service.

```r
gender.encounter2 = table(Qn3.2$ENCOUNTERCLASS,Qn3.2$GENDER)
gender.encounter2#gende and encounter contigency
```

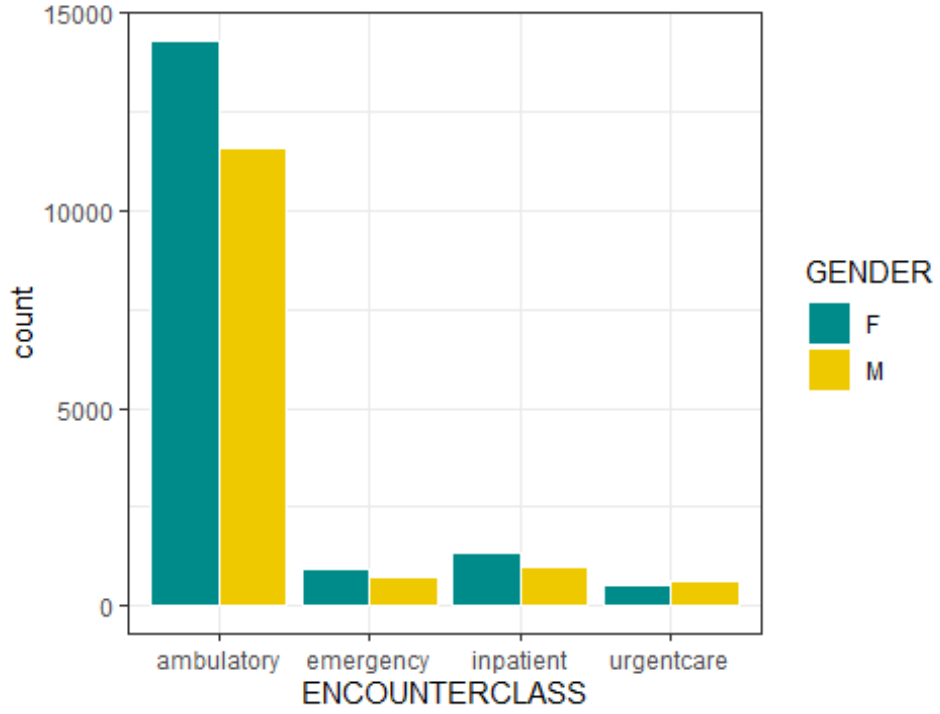
```
##
##                   F     M
##   ambulatory  14303 11594
##   emergency     941   735
##   inpatient    1332   959
##   urgentcare    497   623
```

```r
prop.gedner.2 = prop.table(gender.encounter2,margin = 1)*100
prop.gedner.2#proportion table
```

```
##
##                     F        M
##   ambulatory 55.23034 44.76966
##   emergency  56.14558 43.85442
##   inpatient  58.14055 41.85945
##   urgentcare 44.37500 55.62500
```

```r
ggplot(Qn3.2, aes(x = ENCOUNTERCLASS, fill = GENDER)) +
  geom_bar(position = position_dodge(width = 0.9), color = "white") +
  labs(title = "") +
  theme_bw()+ scale_fill_manual(values = c("cyan4","gold2"))
```

Distribution of encounter classes among genders

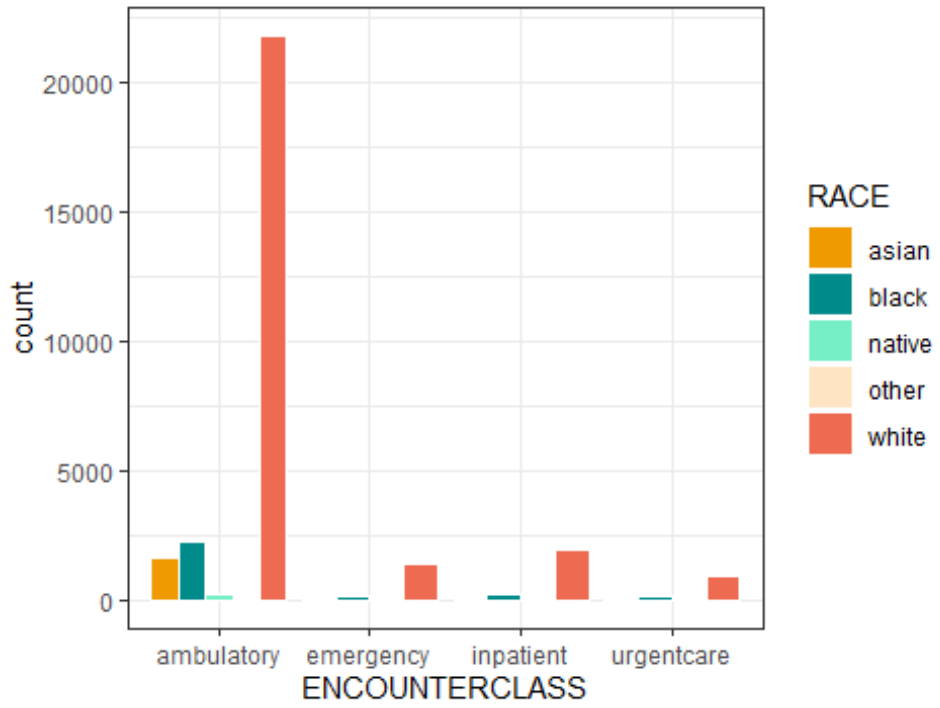


When the encounter classes were compared with the genders in the second dataset, the output result was almost similar as the first data. The proportion of women were higher than men in ambulatory, emergency and inpatient services, however the number of men who used urgent care facility is slightly higher than women.

```
race.encounter2 = table(Qn3.2$ENCOUNTERCLASS,Qn3.2$RACE)
race.encounter2#conti_table

##
##              asian black native other white
##   ambulatory  1642  2281    213     6 21755
##   emergency    103   148      3     0  1422
##   inpatient    102   213      9     0  1967
##   urgentcare    56   158      1     0   905

prop.race.2 = prop.table(race.encounter2,margin = 1)*100
prop.race.2#prop table

##
##                    asian       black      native       other       white
##   ambulatory  6.34050276  8.80797004  0.82248909  0.02316871 84.00586941
##   emergency   6.14558473  8.83054893  0.17899761  0.00000000 84.84486874
##   inpatient   4.45220428  9.29725011  0.39284155  0.00000000 85.85770406
##   urgentcare  5.00000000 14.10714286  0.08928571  0.00000000 80.80357143
```

```
ggplot(Qn3.2, aes(x = ENCOUNTERCLASS, fill = RACE)) +
  geom_bar(position = position_dodge(width = 0.9), color = "white") +
  labs(title = "") +
  theme_bw()+
  scale_fill_manual(values =
c("orange2","darkcyan","aquamarine2","bisque1","coral2"))
```

Distribution of encounter classes among races



There is not much difference among different races as well. Number of whites dominate in all the categories followed by blacks and Asians.

Overall gender and race in data 1 and 2 does affect the hospitalisation rate in a similar manner. The proportion of male and female users are almost similar in all the categories, when it comes to race, it is the same as well. The whites dominate in both the dataset whereas the number of people in other races who have utilized these services is altogether 1/4 of the overall population.

Question.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. Does this result vary between the datasets? Does this result vary between the 2 datasets?

This task analyses the factors that affected the people who recovered and did not recover from covid in both the datasets. First we are working with dataset 1. For this we had the start date when the condition was recognised and the stop date which means the date when the condition was resolved (if applicable). To find whether a person recovered or not, we assigned a new column with two values “Recovered” and “Not Recovered” for the start and stop date respectively. We also found the time difference of between these dates to get time taken for recovery.

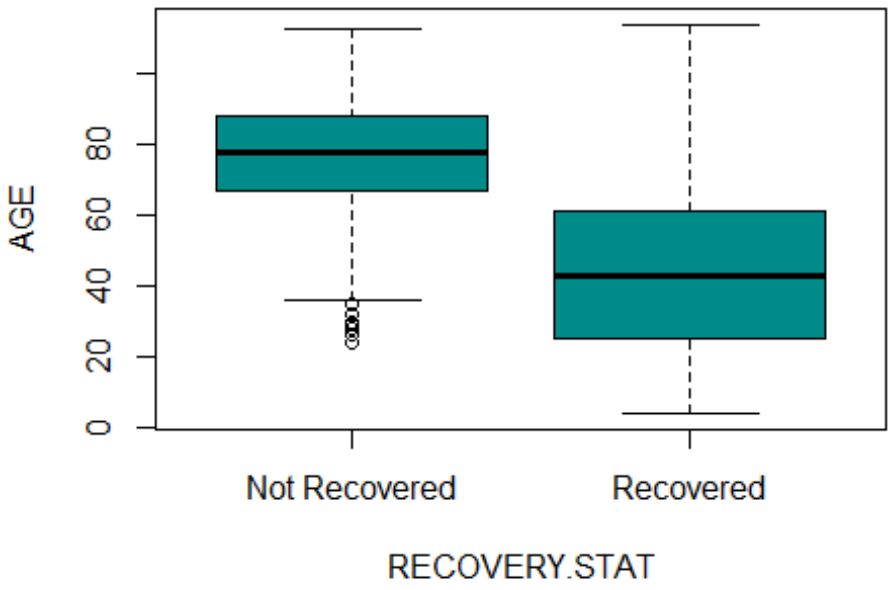
Data 1

```
#DATA 1

Q4$RECOVERY.STAT = ifelse(is.na(Q4$STOP)|Q4$STOP=="", "Not_recovered",
"Recovered")#classifying as recovered and not recovered
Q4$RECOVERY.TIME = as.Date(Q4$STOP)- as.Date(Q4$START)#calculating recovery
time

Q4.table = table(Q4$DESCRIPTION,Q4$RECOVERY.STAT)

patients1$AGE = age_calculator(patients1$BIRTHDATE)
patients1 = patients1 %>%
  mutate(AGE.GROUP = case_when(
    AGE>=0 & AGE<=18 ~ "0-18",
    AGE>=19& AGE<=35 ~ "19-35",#grouping age into categories
    AGE>=36& AGE<=50 ~ "36-50",
    AGE>=51 ~ "51+"
  ))

Q4.1 = subset(patients1,select =
c("Id","AGE","AGE.GROUP","GENDER","ZIP"))#obtaining relevant fields

Q4.2 = Q4.1%>%
  filter(Id %in% Q4$PATIENT)

Q4.3 = merge(Q4,Q4.2,by.x = "PATIENT",by.y = "Id",all = FALSE)#merging by Id

Q4.final = subset(Q4.3,select = c("PATIENT","GENDER","AGE","AGE.GROUP",

"DESCRIPTION","ZIP","RECOVERY.STAT","RECOVERY.TIME"))#final with all required
variables
```

Here we are analysing the factors such as age, gender, zip, symptom and timeline of recovery, so we calculated the age as we did for task 2 and also filtered the gender and zip code of ids who recovered or not.

```
#AGE
boxplot(AGE~RECOVERY.STAT ,data = Q4.final,col = "cyan4",main ="",names=
c("Not Recovered","Recovered"))#boxplot of age and recovery stat
```

Box plot between age and recovery stat



To check the influence of age in recovery stat, we plotted a box plot between two. The boxplot shows that the median age of people who have recovered is around 40 and who has not recovered is nearly 80. This indicates that people who did not recover were mostly older people and most of the youth survived. So age was a critical factor in recovery.

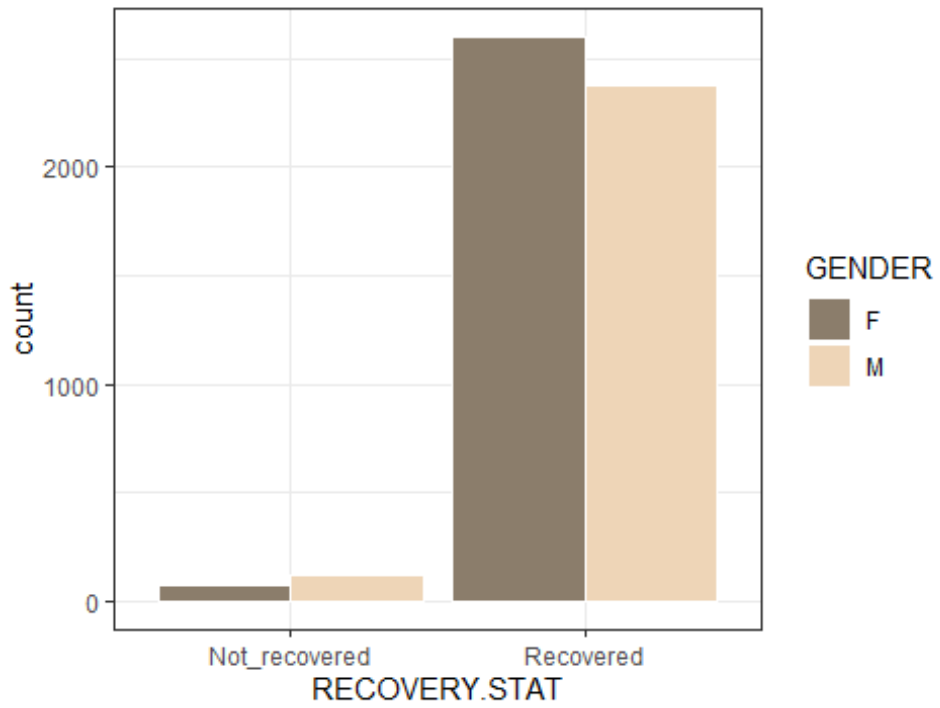
```
lm(AGE~RECOVERY.TIME,data = Q4.final)#building linear model

##
## Call:
## lm(formula = AGE ~ RECOVERY.TIME, data = Q4.final)
##
## Coefficients:
##   (Intercept)  RECOVERY.TIME
##        46.735         -0.143
```

We also fitted a linear model to relationship between the age of patients and their time of recovery. The intercept, 46.7165 depicts the expected age of a patient when the recovery time is zero. The coefficient for recovery time, -0.1428, implies that for each unit of increase in patient's recovery time, the age decreases by 0.1428 units. Which concludes younger patients recovery time is longer than older people or in other words older people do not recover, there are chances they might pass away in a short period.

```
ggplot(Q4.final, aes(x = RECOVERY.STAT, fill = GENDER)) +
  geom_bar(position = position_dodge(width = 0.9), color = "white") +
```

```
  labs(title = "") +
  theme_bw()+ scale_fill_manual(values = c("bisque4","bisque2"))
```

Gender and Recovery stat



Recovery stat when compared between genders in data 1, it shows that the number of females who recovered from covid is more than the number of males, the number of patients who did not recover are also mostly men.

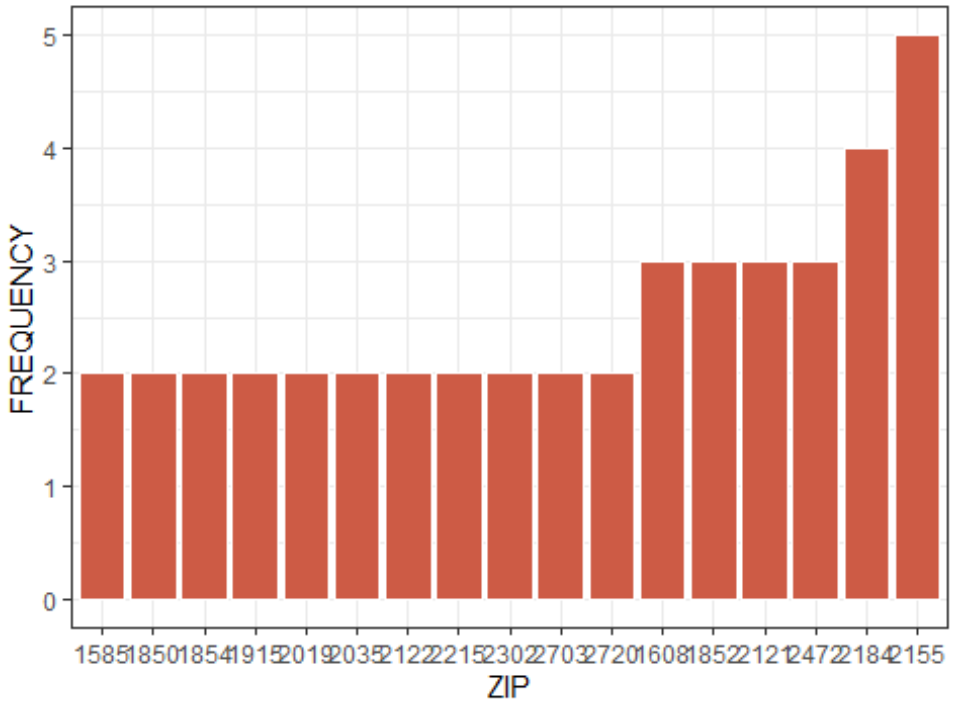
```
REC.ZIP1 = as.data.frame(table(Q4.final$RECOVERY.STAT,Q4.final$ZIP))#zip code
as dataframe
names(REC.ZIP1)= c("RECOVERY.STAT","ZIP","FREQUENCY")#giving names
NOT.REC.ZIP1 = subset(REC.ZIP1, RECOVERY.STAT=="Not_recovered")#for not
recovered
NOT.REC.ZIP1 =NOT.REC.ZIP1%>%
  slice_max(order_by = FREQUENCY, n =10, with_ties = TRUE)
REC.ZIP1 = subset(REC.ZIP1,RECOVERY.STAT=="Recovered")#for recovered
REC.ZIP1 = REC.ZIP1%>%
  slice_max(order_by = FREQUENCY, n =10, with_ties = TRUE)#obtaining top 10
```

As we have the zip codes of patients, we can check people from which place has recovered more and not.

```
ggplot(NOT.REC.ZIP1, aes(x = FREQUENCY, y = reorder(ZIP, +FREQUENCY), fill =
FREQUENCY)) +
  geom_bar(stat = "identity",fill = "coral3",colour= "white") +
```

```
  labs(title = "",y = "ZIP") +
  theme_bw()+coord_flip()
```

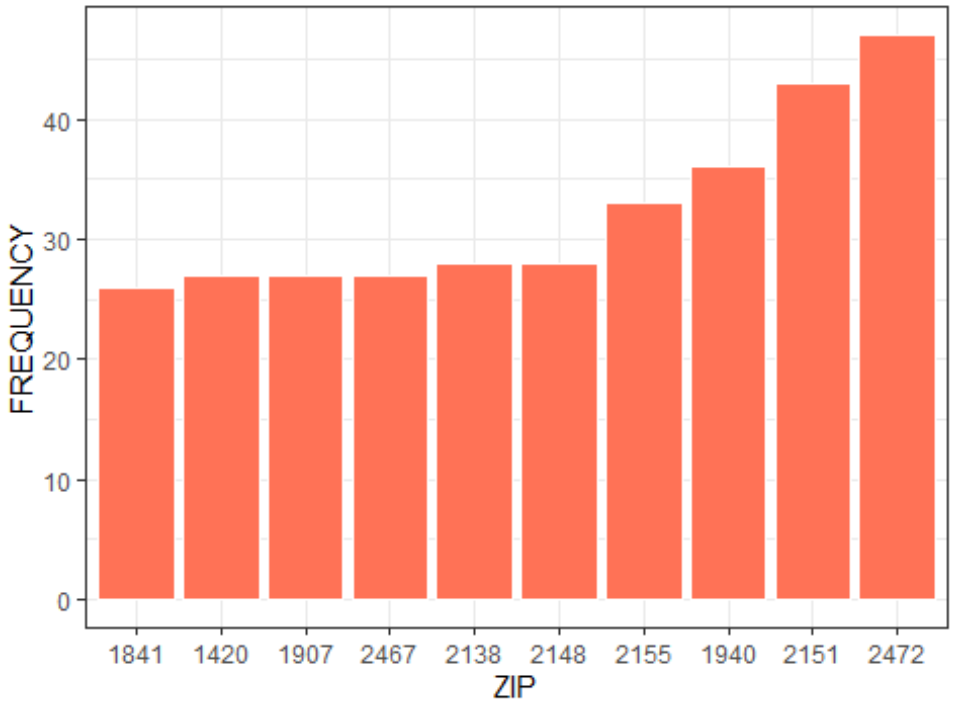
Not recovered zip code



A simple bargraph can explain it all, the highest number of people who did not recover from covid come from the area zip code 2155 with a number of 5 followed by 4 which is the area under zip code 2184.When we cross checked the zip code '2155' we found that it belongs to Middlesex County, which had the highest number of covid cases as per our analysis before. Norfolk county came second where 4 people not recovered from covid.

```
ggplot(REC.ZIP1, aes(x = FREQUENCY, y = reorder(ZIP, +FREQUENCY), fill =
FREQUENCY)) +
  geom_bar(stat = "identity",fill = "coral1",colour= "white") +
  labs(title = "",y = "ZIP") +
  theme_bw()+coord_flip()
```

Recovered Zip Code

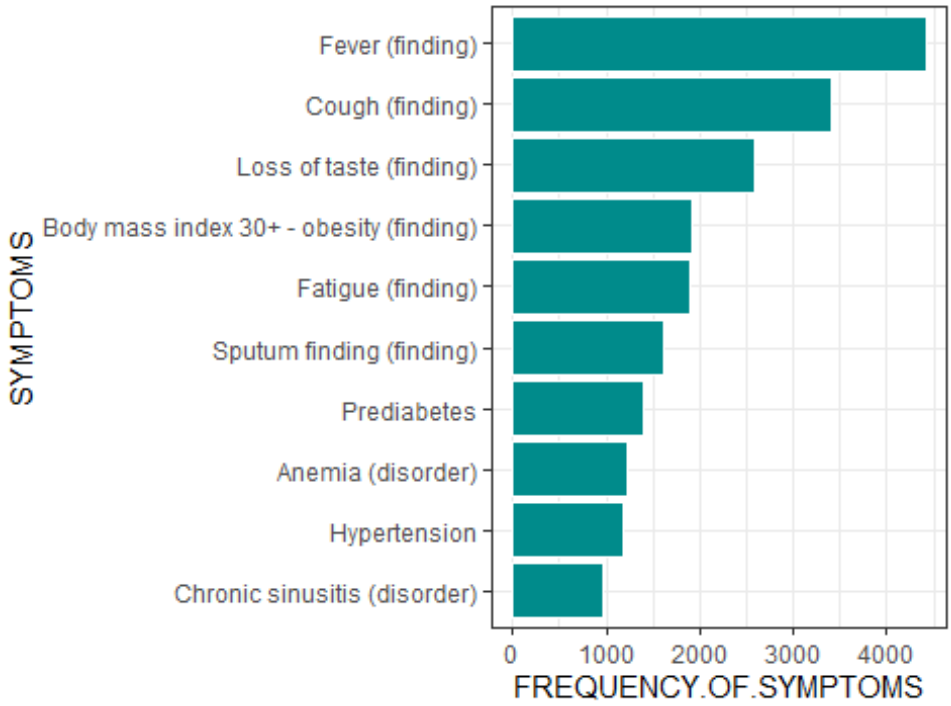


The highest number of people who recovered in data 1 comes under the ZIP code 2472 which is Middlesex County. They had the highest number diagnosis and recoveries. This was followed by Suffolk county.

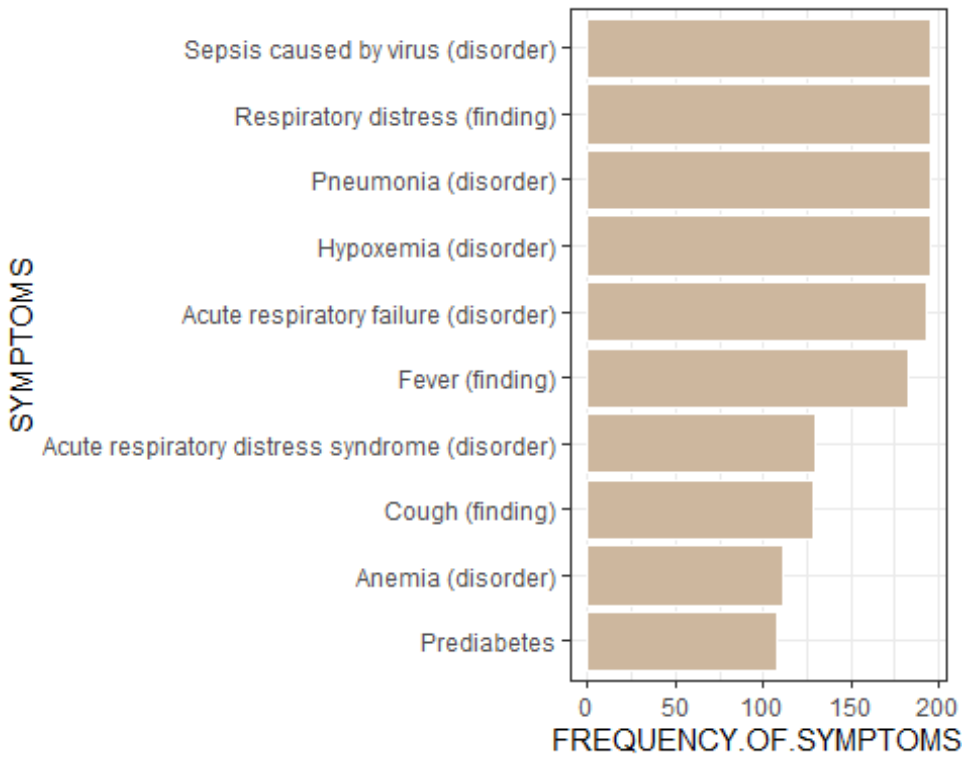
```
REC.id.1 = subset(Q4, select = c("PATIENT","RECOVERY.STAT"))
#table(REC.id.1$RECOVERY.STAT)
SYM.1 = subset(conditions1,select = c("PATIENT","DESCRIPTION"))

SYM.1 = SYM.1%>%
  filter(PATIENT %in% REC.id.1$PATIENT)#filtering id for symptoms

REC.SYM.1 = merge(REC.id.1,SYM.1,by.x = "PATIENT","PATIENT",all =
FALSE)#merging id by symptoms

REC.SYM.table.1 =
as.data.frame(table(REC.SYM.1$DESCRIPTION,REC.SYM.1$RECOVERY.STAT))

REC.SYM.table.1 = REC.SYM.table.1%>%
  arrange(desc(Freq))%>%
  slice(-1,-2)

REC.SYM.table.1 =REC.SYM.table.1%>%
  slice(-42,-43)
```

```
names(REC.SYM.table.1) =
c("SYMPTOMS","RECOVERY.STAT","FREQUENCY.OF.SYMPTOMS")

REC.SYM.FREQ.1 = subset(REC.SYM.table.1, RECOVERY.STAT == "Recovered"
)#symptoms of recovered
NOT.REC.FREQ.1 = subset(REC.SYM.table.1, RECOVERY.STAT == "Not_recovered"
)#symptos of not recovered

REC.SYM.FREQ.1 = REC.SYM.FREQ.1%>%
  slice_max(order_by = FREQUENCY.OF.SYMPTOMS, n =10, with_ties =
TRUE)#obtaining common symptoms
ggplot(REC.SYM.FREQ.1, aes(x = FREQUENCY.OF.SYMPTOMS, y = reorder(SYMPTOMS,
+FREQUENCY.OF.SYMPTOMS), fill = FREQUENCY.OF.SYMPTOMS)) +
  geom_bar(stat = "identity",fill = "cyan4",colour= "white") +
  labs(title = "",y = "SYMPTOMS") +
  theme_bw()
```

Recovered Symptom




```
NOT.REC.FREQ.1 = NOT.REC.FREQ.1%>%
  slice_max(order_by = FREQUENCY.OF.SYMPTOMS, n =10, with_ties =
TRUE)#symptoms for not recovered
ggplot(NOT.REC.FREQ.1, aes(x = FREQUENCY.OF.SYMPTOMS, y = reorder(SYMPTOMS,
+FREQUENCY.OF.SYMPTOMS), fill = FREQUENCY.OF.SYMPTOMS)) +
  geom_bar(stat = "identity",fill = "bisque3",colour= "white") +
  labs(y = "SYMPTOMS") +
  theme_bw()
```



We also filtered the top 10 symptoms associated with people who recovered and did not recover in the first dataset and visualized using bar plot. Fever, cough and taste loss were the common symtoms among people who recovered and conditions like sepsis, respiratory problems, pnemonia and so on were common amomg people who did not recover. This means that people who did not recover had more complicated symptoms than people who recovered.

Data 2

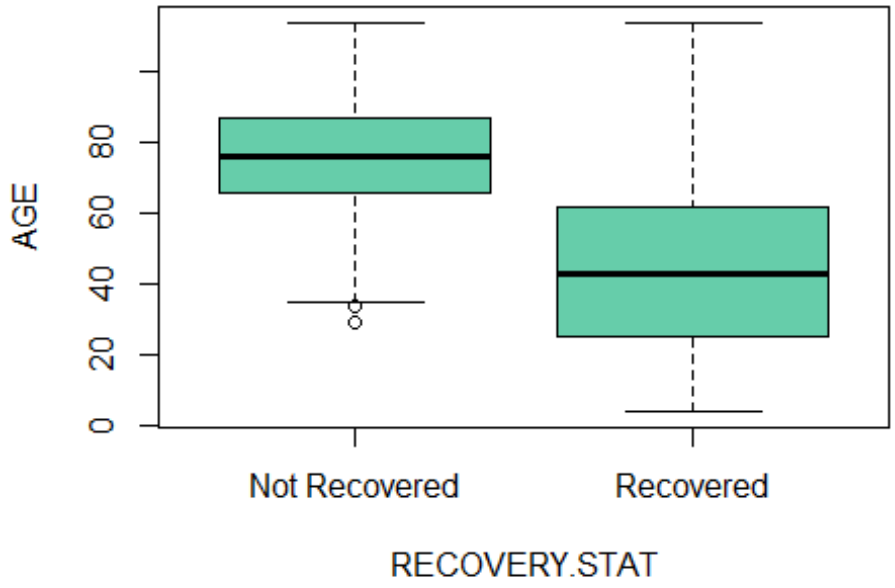
```
#DATA 2
Qn4$RECOVERY.STAT = ifelse(is.na(Qn4$STOP)|Qn4$STOP=="", "Not_recovered",
"Recovered")
Qn4$RECOVERY.TIME = as.Date(Qn4$STOP)- as.Date(Qn4$START)#same with data 2

Qn4.table = table(Qn4$DESCRIPTION,Qn4$RECOVERY.STAT)
```

The same methods of analysis were carried out data 2.

```
patients2$AGE = age_calculator(patients2$BIRTHDATE)
patients2 = patients2 %>%
  mutate(AGE.GROUP = case_when(
    AGE>=0 & AGE<=18 ~ "0-18",
    AGE>=19& AGE<=35 ~ "19-35",
    AGE>=36& AGE<=50 ~ "36-50",
    AGE>=51 ~ "51+"
```

```
  ))

Qn4.1 = subset(patients2,select = c("Id","AGE","AGE.GROUP","GENDER","ZIP"))

Qn4.2 = Qn4.1%>%
  filter(Id %in% Qn4$PATIENT)

Qn4.3 = merge(Qn4,Qn4.2,by.x = "PATIENT",by.y = "Id",all = FALSE)

Qn4.final = subset(Qn4.3,select = c("PATIENT","GENDER","AGE","AGE.GROUP",

"DESCRIPTION","ZIP","RECOVERY.STAT","RECOVERY.TIME"))

boxplot(AGE~RECOVERY.STAT ,data = Qn4.final,col = "aquamarine3",main
="",names= c("Not Recovered","Recovered"))
```

Boxplot between age and Recovery Stat



The box plot between age and recovery stat in data 2 is kind of similar as the first one. The median age of people who did not recover lies just below 80 and the age of people who survived is around 40.

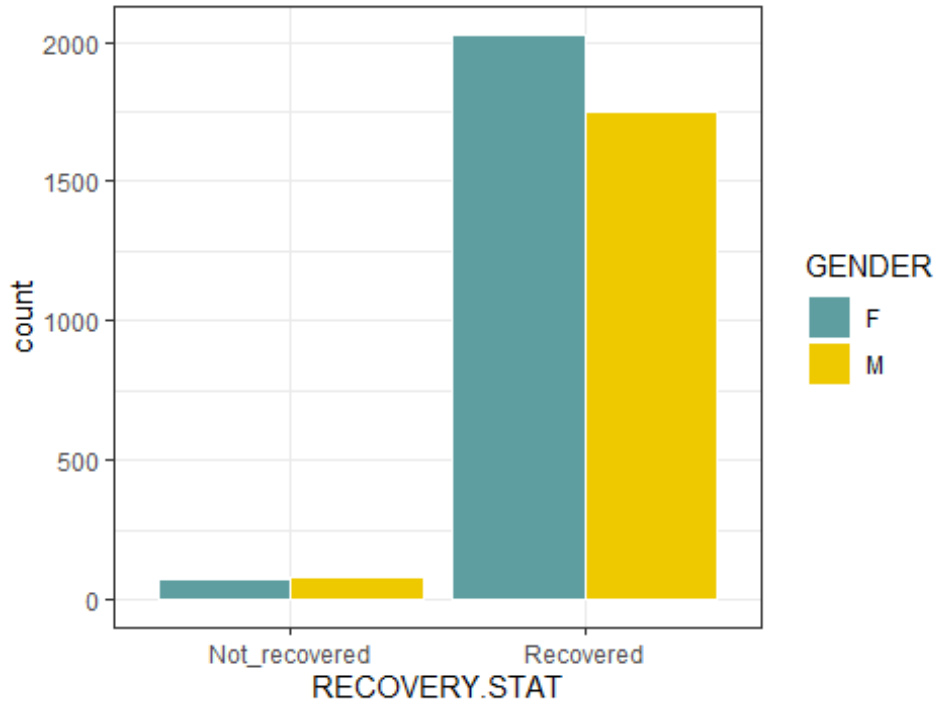
```
lm(AGE~RECOVERY.TIME,data = Qn4.final)#model for 2nd data

##
## Call:
## lm(formula = AGE ~ RECOVERY.TIME, data = Qn4.final)
```

```
##
## Coefficients:
##   (Intercept)  RECOVERY.TIME
##       49.5582        -0.2465
```

The model also states that for every one unit of increase in patient's recovery time, their age decreases by 0.2463. Though younger people need more time to recover, the chances of recovery is very high.

```
ggplot(Qn4.final, aes(x = RECOVERY.STAT, fill = GENDER)) +
  geom_bar(position = position_dodge(width = 0.9), color = "white") +
  labs(title = "") +
  theme_bw()+ scale_fill_manual(values = c("cadetblue","gold2"))
```

Recovery stat and Gender

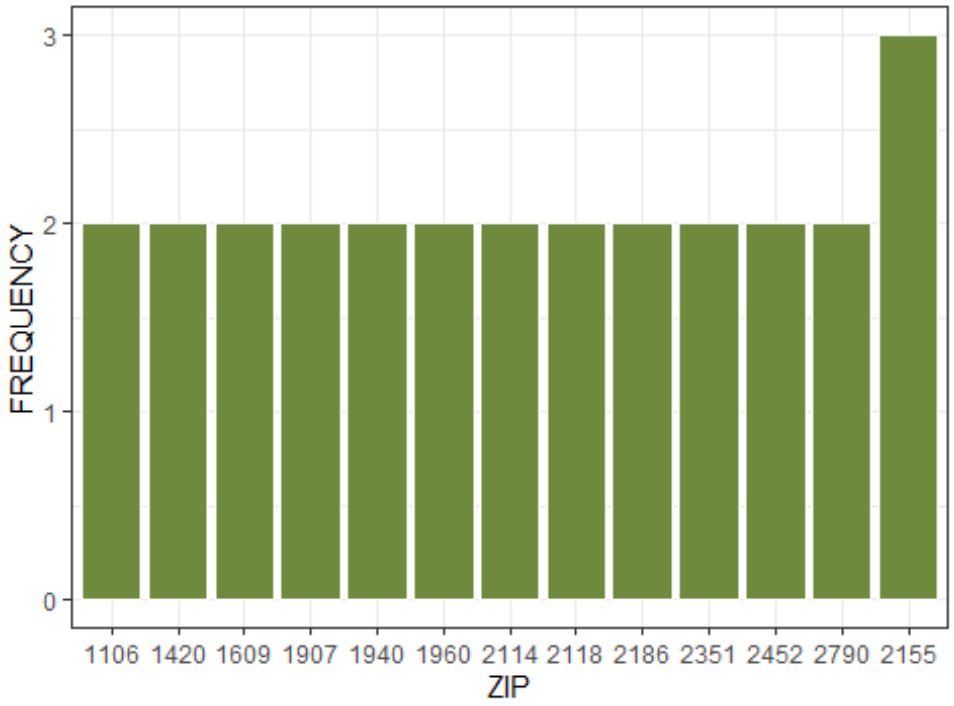


The comparison of recovery status between genders shows that, even in dataset 2, the number of females recovered more than men. Whereas the number of people who did not recover is almost the same.

```
REC.ZIP2 = as.data.frame(table(Qn4.final$RECOVERY.STAT,Qn4.final$ZIP))
names(REC.ZIP2)= c("RECOVERY.STAT","ZIP","FREQUENCY")
NOT.REC.ZIP2 = subset(REC.ZIP2, RECOVERY.STAT=="Not_recovered")
NOT.REC.ZIP2 =NOT.REC.ZIP2%>%
  slice_max(order_by = FREQUENCY, n =10, with_ties = TRUE)
REC.ZIP2 = subset(REC.ZIP2,RECOVERY.STAT=="Recovered")
```

```
REC.ZIP2 = REC.ZIP2%>%
  slice_max(order_by = FREQUENCY, n =10, with_ties = TRUE)

ggplot(NOT.REC.ZIP2, aes(x = FREQUENCY, y = reorder(ZIP, +FREQUENCY), fill =
FREQUENCY)) +
  geom_bar(stat = "identity",fill = "darkolivegreen4",colour= "white") +
  labs(title = "",y = "ZIP") +
  theme_bw()+coord_flip()
```

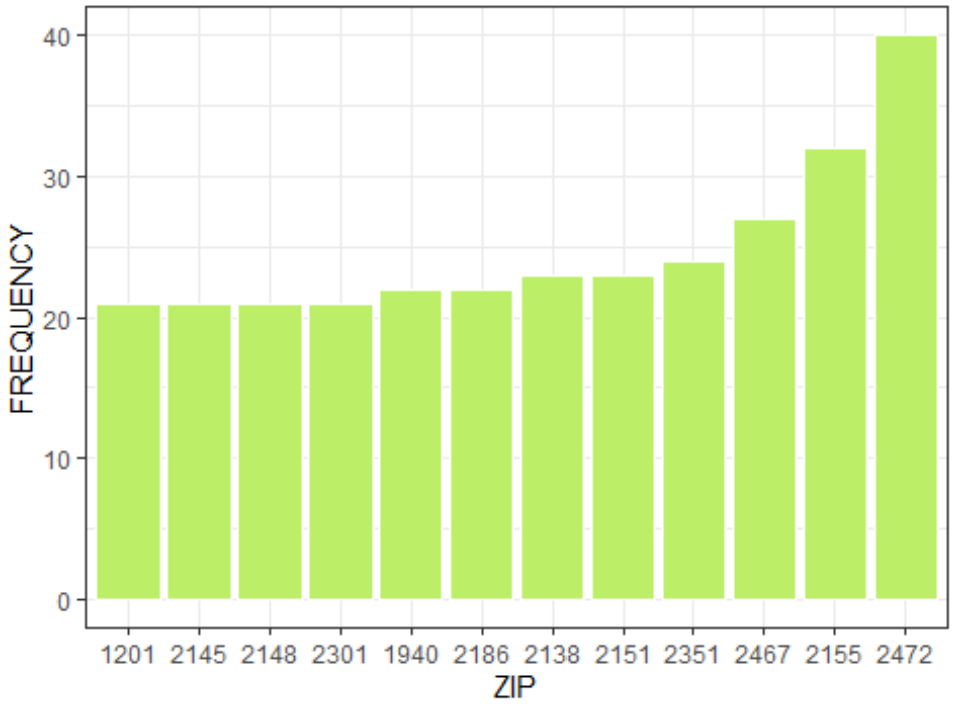
Not Recovered zip code



The highest number of patients in dataset 2 who did not recover comes under the zip code 2155 which is Middlesex County with 3 patients. All other counties reported 2 people who did not recover.

```
ggplot(REC.ZIP2, aes(x = FREQUENCY, y = reorder(ZIP, +FREQUENCY), fill =
FREQUENCY)) +
  geom_bar(stat = "identity",fill = "darkolivegreen2",colour= "white") +
  labs(title = "",y = "ZIP") +
  theme_bw()+coord_flip()
```

Recovered zip code

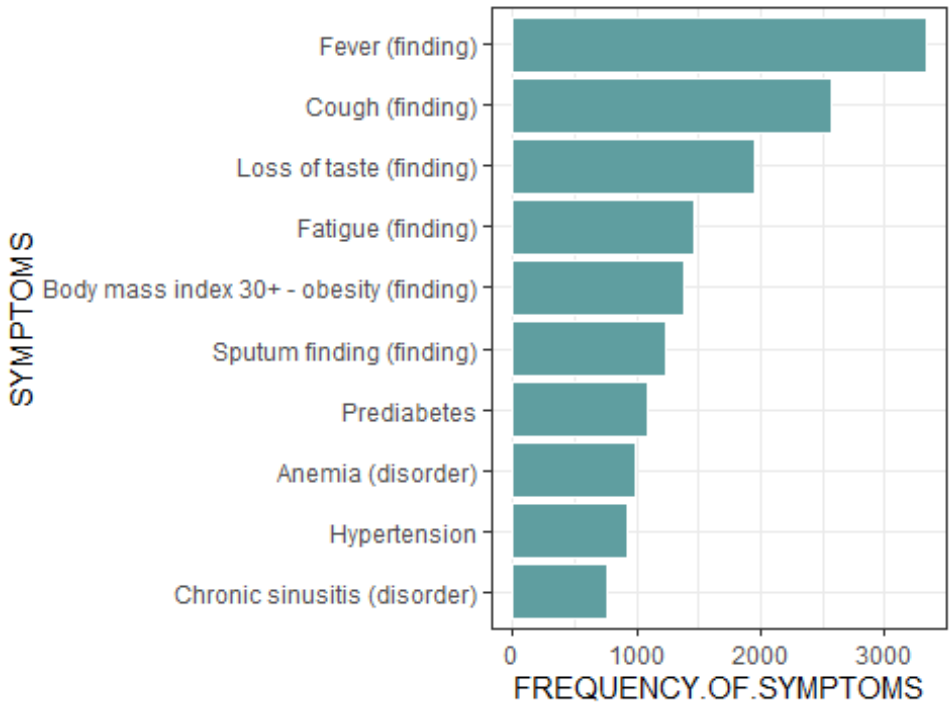


Middlesex county under zip 2472 leads highest number of people who recovered in dataset 2 followed by Norfolk and Plymouth County.

```
REC.id.2 = subset(Qn4, select = c("PATIENT","RECOVERY.STAT"))
SYM.2 = subset(conditions2,select = c("PATIENT","DESCRIPTION"))

SYM.2 = SYM.2%>%
  filter(PATIENT %in% REC.id.2$PATIENT)

REC.SYM.2 = merge(REC.id.2,SYM.2,by.x = "PATIENT","PATIENT",all = FALSE)

REC.SYM.table.2 =
as.data.frame(table(REC.SYM.2$DESCRIPTION,REC.SYM.2$RECOVERY.STAT))

REC.SYM.table.2 = REC.SYM.table.2%>%
  arrange(desc(Freq))%>%
  slice(-1,-2)

REC.SYM.table.2 =REC.SYM.table.2%>%
  slice(-42,-47)

names(REC.SYM.table.2) =
c("SYMPTOMS","RECOVERY.STAT","FREQUENCY.OF.SYMPTOMS")
```

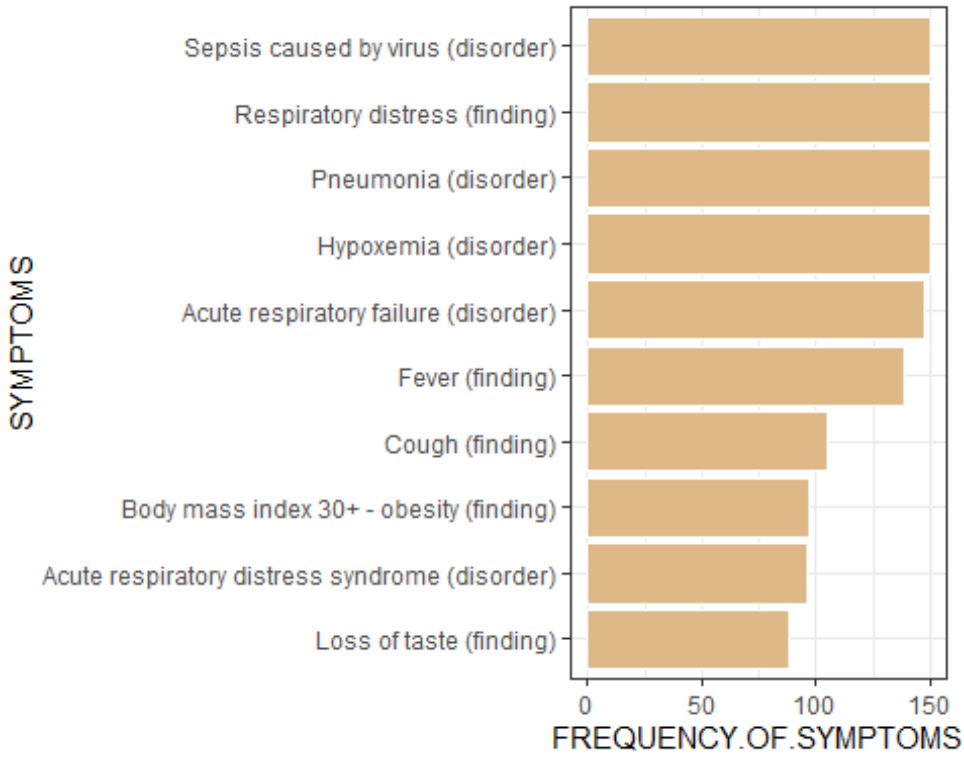
```
REC.SYM.FREQ.2 = subset(REC.SYM.table.2, RECOVERY.STAT == “Recovered” )
NOT.REC.FREQ.2 = subset(REC.SYM.table.2, RECOVERY.STAT == “Not_recovered” )

REC.SYM.FREQ.2 = REC.SYM.FREQ.2%>%
  slice_max(order_by = FREQUENCY.OF.SYMPTOMS, n =10, with_ties = TRUE)

ggplot(REC.SYM.FREQ.2, aes(x = FREQUENCY.OF.SYMPTOMS, y = reorder(SYMPTOMS,
+FREQUENCY.OF.SYMPTOMS), fill = FREQUENCY.OF.SYMPTOMS)) +
  geom_bar(stat = “identity”,fill = “cadetblue”,colour= “white”) +
  labs(title = “”,y = “SYMPTOMS”) +
  theme_bw()
```

Recovered Symptom




```
NOT.REC.FREQ.2 = NOT.REC.FREQ.2%>%
  slice_max(order_by = FREQUENCY.OF.SYMPTOMS, n =10, with_ties = TRUE)

ggplot(NOT.REC.FREQ.2, aes(x = FREQUENCY.OF.SYMPTOMS, y = reorder(SYMPTOMS,
+FREQUENCY.OF.SYMPTOMS), fill = FREQUENCY.OF.SYMPTOMS)) +
  geom_bar(stat = "identity",fill = "burlywood",colour= "white") +
  labs(y = "SYMPTOMS") +
  theme_bw()
```

Not Recovered Symptom



Symptoms associated with those who recovered in data2 are the same as 1. The most common for recovered being fever and cough and for those who did not recover, it is sepsis and respiratory distress. BMI is also a crucial factor in both data.

Overall in both the datasets, the factors affecting recovery is almost same. Though the number of observations in datset 2 is less compared to the observations in 1, the overall result appears the same.

To conclude, this analysis gave as insight into different aspects that influenced the output of covid 19 outbreak. Even the small factors that we might consider as irrelevant can be relevant in a data science or analysis perspective.