Course ID – Course Name

Student ID & Name

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.

Minh Ngoc
[Date]

R Assignment Report

Minh Ngoc

2024-05-22

```
# == Task 1 ====

# Load data of Patient who has COVID-19 or Suspected COVID demographics ====

patient_info <- read.csv("data/patientsUG.csv")

head(patient_info)
```

```
##   Record_ID                            PATIENT_ID  BIRTHDATE DEATHDATE
AGE
## 1      3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8  3/17/1988
36.16667
## 2       532 9718334c-3289-4b1c-a017-72f3df283ab3  6/13/1951
72.92778
## 3      5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2   2/6/2006
18.28056
## 4      7462 c10ee469-6182-4228-ac26-21bcf2412337 10/28/1912 11/1/2016
111.55278
## 5     10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a  6/24/1948 3/31/2020
75.89722
## 6      7818 e283d725-b355-4b86-98a5-b8274e643527   9/1/1992
31.71111
##   MARITAL  RACE GENDER        CITY         STATE         COUNTY  ZIP
## 1       M white      F    Rehoboth Massachusetts Bristol County   NA
## 2       M black      M      Boston Massachusetts Suffolk County 2113
## 3         white      M  Foxborough Massachusetts Norfolk County 2035
## 4       S black      F Springfield Massachusetts Hampden County 1020
## 5       M white      M   Braintree Massachusetts Norfolk County 2184
## 6       S black      M   Braintree Massachusetts Norfolk County 2184
```

```
conditions <- read.csv("data/conditionsUG.csv")

head(conditions)
```

```
##   Record_ID      START      STOP                            PATIENT_ID
## 1         1  2/15/2019  8/1/2019 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2         2 10/30/2019 1/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3         3   3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 4         4   3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 5         5   3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 6         6  2/12/2020 2/26/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
##                           ENCOUNTER_ID      CODE        DESCRIPTION
## 1 d5ee30a9-362f-429e-a87a-ee38d999b0a5  65363002       Otitis media
## 2 8bca6d8a-ab80-4cbf-8abb-46654235f227  65363002       Otitis media
```

```
## 3 681c380b-3c84-4c55-80a6-db3d9ea12fee 386661006    Fever (finding)
## 4 681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
## 5 681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
## 6 adedca64-700b-4fb9-82f1-9cbb658abb73  44465007    Sprain of ankle

# Filter the COVID-19 or Suspected COVID-19 conditions ====

condition_covid <- subset(conditions, (conditions$DESCRIPTION == "COVID-19")
                          | (conditions$DESCRIPTION == "Suspected COVID-19"))

head(condition_covid)

##    Record_ID     START      STOP                              PATIENT_ID
## 4          4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 5          5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 11        11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 12        12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 23        23 3/11/2020 3/11/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 24        24 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
##                            ENCOUNTER_ID      CODE        DESCRIPTION
## 4  681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
## 5  681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
## 11 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
## 12 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006           COVID-19
## 23 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840544004 Suspected COVID-19
## 24 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840539006           COVID-19

# Join the two tables: Covid Condition and Patient ====

merged_data <- merge(condition_covid, patient_info,
                     by = "PATIENT_ID", all = TRUE)

head(merged_data)

##                              PATIENT_ID Record_ID.x     START      STOP
## 1 0000b247-1def-417a-a783-41c8682be022       87721 2/18/2020 2/18/2020
## 2 0000b247-1def-417a-a783-41c8682be022       87722 2/18/2020 3/25/2020
## 3 00049ee8-5953-4edd-a277-b9c1b1a7f16b       72610  3/8/2020  3/8/2020
## 4 00049ee8-5953-4edd-a277-b9c1b1a7f16b       72611  3/8/2020 3/24/2020
## 5 00079a57-24a8-430f-b4f8-a1cf34f90060       75792 2/25/2020 2/25/2020
## 6 00079a57-24a8-430f-b4f8-a1cf34f90060       75793 2/25/2020 3/10/2020
##                            ENCOUNTER_ID      CODE        DESCRIPTION
Record_ID.y
## 1 93c3da2d-9420-49fa-94e3-7140ab9aeba1 840544004 Suspected COVID-19
9493
## 2 93c3da2d-9420-49fa-94e3-7140ab9aeba1 840539006           COVID-19
9493
## 3 dab47020-5bd0-4ce6-ae5c-e4f1ebd04627 840544004 Suspected COVID-19
7885
## 4 dab47020-5bd0-4ce6-ae5c-e4f1ebd04627 840539006           COVID-19
```

```
7885
## 5 3a23144d-0dee-4dca-90ea-0ad14c1c6909 840544004 Suspected COVID-19
8215
## 6 3a23144d-0dee-4dca-90ea-0ad14c1c6909 840539006           COVID-19
8215
##    BIRTHDATE DEATHDATE      AGE MARITAL  RACE GENDER               CITY
## 1  12/3/2007           16.45556         white      F North Attleborough
## 2  12/3/2007           16.45556         white      F North Attleborough
## 3 11/20/1984           39.49167       M white      M            Wayland
## 4 11/20/1984           39.49167       M white      M            Wayland
## 5  3/25/1991           33.14444       M white      F   Great Barrington
## 6  3/25/1991           33.14444       M white      F   Great Barrington
##           STATE           COUNTY  ZIP
## 1 Massachusetts   Bristol County   NA
## 2 Massachusetts   Bristol County   NA
## 3 Massachusetts Middlesex County   NA
## 4 Massachusetts Middlesex County   NA
## 5 Massachusetts Berkshire County 1230
## 6 Massachusetts Berkshire County 1230

# === 1.1 Distribution of patients across counties ====

# To count the occurrences of each unique value in the “county” column
# from our dataset,
# we can use the count() function from the dplyr package.

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

# Count the occurrences of each county
county_counts <- merged_data %>% count(COUNTY)

# Print the result
print(county_counts)

##               COUNTY    n
## 1  Barnstable County  529
## 2   Berkshire County  320
## 3     Bristol County 1330
## 4       Dukes County   61
## 5       Essex County 1768
```

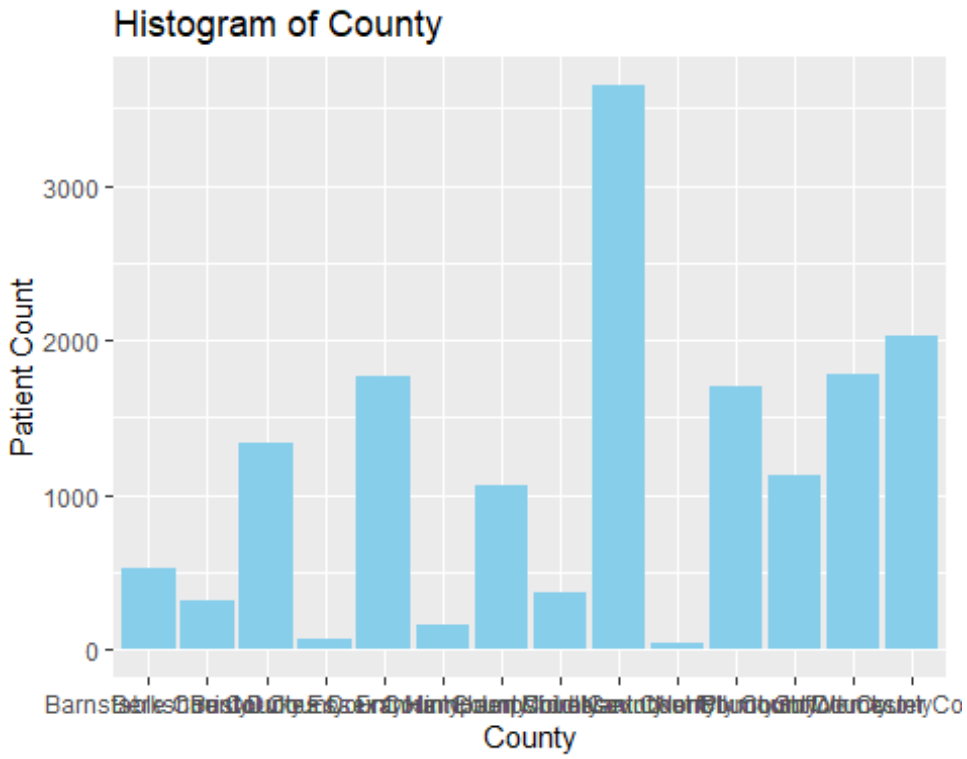
```
## 6    Franklin County  155
## 7     Hampden County 1056
## 8   Hampshire County  368
## 9   Middlesex County 3647
## 10  Nantucket County   40
## 11    Norfolk County 1704
## 12   Plymouth County 1129
## 13    Suffolk County 1775
## 14  Worcester County 2030
```

```
# See the data type
str(county_counts)
```

```
## 'data.frame':    14 obs. of  2 variables:
##  $ COUNTY: chr  "Barnstable County" "Berkshire County" "Bristol County"
"Dukes County" ...
##  $ n     : int  529 320 1330 61 1768 155 1056 368 3647 40 ...
```

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

# Create the histogram of County
ggplot(county_counts, aes(x = county_counts$COUNTY, y = county_counts$n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Histogram of County", x = "County", y = "Patient Count")
```

```
## Warning: Use of `county_counts$COUNTY` is discouraged.
## ℹ Use `COUNTY` instead.

## Warning: Use of `county_counts$n` is discouraged.
## ℹ Use `n` instead.
```




```
# Bonus: Top 10 Counties that have the most COVID Patients ====

library(dplyr)

top_counties <- county_counts %>%
  arrange(desc(n)) %>%
  head(10)  # Select the top 10 conditions

# Print the top 10 counties that have the most Covid Patients
print(top_counties)
```

```
##               COUNTY    n
## 1   Middlesex County 3647
## 2   Worcester County 2030
## 3     Suffolk County 1775
## 4       Essex County 1768
## 5     Norfolk County 1704
## 6     Bristol County 1330
## 7    Plymouth County 1129
## 8     Hampden County 1056
## 9  Barnstable County  529
## 10  Hampshire County  368
```

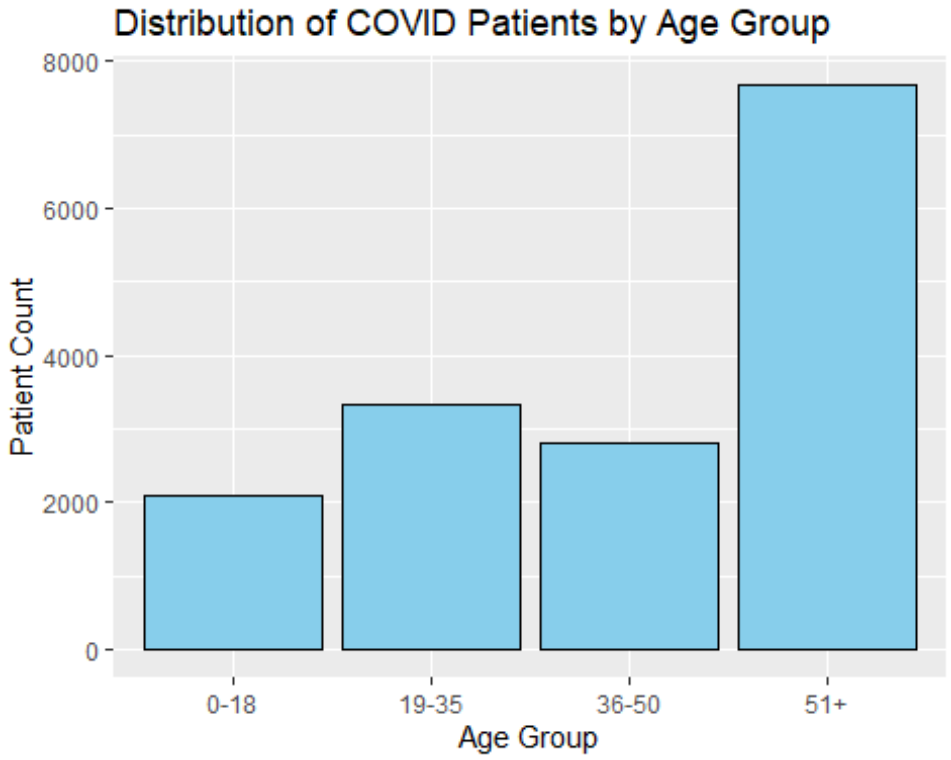
```
# =================

# 1.2. Distribution of Patients across Age Groups ====
```

```r
# Calculate the Age from Birthday column
# To quickly calculate in Excel,
# We use YEARFRAC(Birthday, TODAY()) function

head(merged_data)
```

```
##                              PATIENT_ID Record_ID.x     START      STOP
## 1 0000b247-1def-417a-a783-41c8682be022       87721 2/18/2020 2/18/2020
## 2 0000b247-1def-417a-a783-41c8682be022       87722 2/18/2020 3/25/2020
## 3 00049ee8-5953-4edd-a277-b9c1b1a7f16b       72610  3/8/2020  3/8/2020
## 4 00049ee8-5953-4edd-a277-b9c1b1a7f16b       72611  3/8/2020 3/24/2020
## 5 00079a57-24a8-430f-b4f8-a1cf34f90060       75792 2/25/2020 2/25/2020
## 6 00079a57-24a8-430f-b4f8-a1cf34f90060       75793 2/25/2020 3/10/2020
##                            ENCOUNTER_ID      CODE        DESCRIPTION
Record_ID.y
## 1 93c3da2d-9420-49fa-94e3-7140ab9aeba1 840544004 Suspected COVID-19
9493
## 2 93c3da2d-9420-49fa-94e3-7140ab9aeba1 840539006           COVID-19
9493
## 3 dab47020-5bd0-4ce6-ae5c-e4f1ebd04627 840544004 Suspected COVID-19
7885
## 4 dab47020-5bd0-4ce6-ae5c-e4f1ebd04627 840539006           COVID-19
7885
## 5 3a23144d-0dee-4dca-90ea-0ad14c1c6909 840544004 Suspected COVID-19
8215
## 6 3a23144d-0dee-4dca-90ea-0ad14c1c6909 840539006           COVID-19
8215
##    BIRTHDATE DEATHDATE      AGE MARITAL  RACE GENDER               CITY
## 1  12/3/2007           16.45556         white      F North Attleborough
## 2  12/3/2007           16.45556         white      F North Attleborough
## 3 11/20/1984           39.49167       M white      M            Wayland
## 4 11/20/1984           39.49167       M white      M            Wayland
## 5  3/25/1991           33.14444       M white      F   Great Barrington
## 6  3/25/1991           33.14444       M white      F   Great Barrington
##           STATE           COUNTY  ZIP
## 1 Massachusetts   Bristol County   NA
## 2 Massachusetts   Bristol County   NA
## 3 Massachusetts Middlesex County   NA
## 4 Massachusetts Middlesex County   NA
## 5 Massachusetts Berkshire County 1230
## 6 Massachusetts Berkshire County 1230
```

```r
# Label the Age Groups

merged_data$AGEGROUP <- cut(merged_data$AGE, breaks = c(0, 18, 35, 50, Inf),
                            labels = c("0-18", "19-35", "36-50", "51+"))

# See the number of patient by Age Group

summary(merged_data$AGEGROUP)
```

```
##  0-18 19-35 36-50   51+
##  2095  3329  2815  7673

library(ggplot2)

ggplot(merged_data, aes(x = AGEGROUP)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of COVID Patients by Age Group",
       x = "Age Group",
       y = "Patient Count")
```




```
# == Task 2 ====

# 2.1. Top 10 most common conditions (symptoms) related to the patients ====

head(conditions)

##   Record_ID      START      STOP                            PATIENT_ID
## 1         1  2/15/2019  8/1/2019 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2         2 10/30/2019 1/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3         3   3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 4         4   3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 5         5   3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 6         6  2/12/2020 2/26/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
##                            ENCOUNTER_ID      CODE        DESCRIPTION
## 1 d5ee30a9-362f-429e-a87a-ee38d999b0a5  65363002       Otitis media
## 2 8bca6d8a-ab80-4cbf-8abb-46654235f227  65363002       Otitis media
```

```
## 3 681c380b-3c84-4c55-80a6-db3d9ea12fee 386661006    Fever (finding)
## 4 681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
## 5 681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
## 6 adedca64-700b-4fb9-82f1-9cbb658abb73  44465007    Sprain of ankle
```

```
# Filtered conditions that related to COVID symptom

covid_symptoms <- subset(conditions, (conditions$DESCRIPTION == "COVID-19")
                         | (conditions$DESCRIPTION == "Suspected COVID-19")
                         | (conditions$DESCRIPTION == "Fever (finding)")
                         | (conditions$DESCRIPTION == "Cough (finding)")
                         | (conditions$DESCRIPTION == "Streptococcal sore
throat (disorder)")
                         | (conditions$DESCRIPTION == "Headache (finding)")
                         | (conditions$DESCRIPTION == "Fatigue (finding)")
                         | (conditions$DESCRIPTION == "Nausea (finding)")
                         | (conditions$DESCRIPTION == "Vomiting symptom
(finding)")
                         | (conditions$DESCRIPTION == "Loss of taste
(finding)"))

head(covid_symptoms)
```

```
##    Record_ID     START      STOP                            PATIENT_ID
## 3          3  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 4          4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 5          5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 7          7 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 10        10 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 11        11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
##                             ENCOUNTER_ID      CODE        DESCRIPTION
## 3  681c380b-3c84-4c55-80a6-db3d9ea12fee 386661006    Fever (finding)
## 4  681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
## 5  681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
## 7  1ea74a77-3ad3-4948-a9cc-3084462035d6  49727002    Cough (finding)
## 10 1ea74a77-3ad3-4948-a9cc-3084462035d6 386661006    Fever (finding)
## 11 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
```

```
library(dplyr)

condition_counts <- covid_symptoms %>% count(DESCRIPTION)

head(condition_counts)
```

```
##               DESCRIPTION    n
## 1               COVID-19 6648
## 2         Cough (finding) 4674
## 3       Fatigue (finding) 2644
## 4         Fever (finding) 6088
```

```
## 5      Headache (finding)  991
## 6 Loss of taste (finding) 3571

top_conditions <- condition_counts %>%
  arrange(desc(n)) %>%
  head(10)  # Select the top 10 conditions

# Print the top conditions
print(top_conditions)

##                            DESCRIPTION    n
## 1                   Suspected COVID-19 6863
## 2                             COVID-19 6648
## 3                Fever (finding) 6088
## 4                Cough (finding) 4674
## 5           Loss of taste (finding) 3571
## 6                 Fatigue (finding) 2644
## 7                Headache (finding)  991
## 8                  Nausea (finding)  375
## 9           Vomiting symptom (finding)  375
## 10 Streptococcal sore throat (disorder)  129

# ===
# 2.2. Top 10 conditions for genders: ====
# The Difference between Male & Female
#
# 2.2.1. Top 10 conditions for Male ====

head(patient_info)

##   Record_ID                            PATIENT_ID  BIRTHDATE DEATHDATE
AGE
## 1      3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8  3/17/1988
36.16667
## 2       532 9718334c-3289-4b1c-a017-72f3df283ab3  6/13/1951
72.92778
## 3      5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2   2/6/2006
18.28056
## 4      7462 c10ee469-6182-4228-ac26-21bcf2412337 10/28/1912 11/1/2016
111.55278
## 5     10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a  6/24/1948 3/31/2020
75.89722
## 6      7818 e283d725-b355-4b86-98a5-b8274e643527   9/1/1992
31.71111
##   MARITAL  RACE GENDER        CITY         STATE         COUNTY  ZIP
## 1       M white      F    Rehoboth Massachusetts Bristol County   NA
## 2       M black      M      Boston Massachusetts Suffolk County 2113
## 3         white      M  Foxborough Massachusetts Norfolk County 2035
## 4       S black      F Springfield Massachusetts Hampden County 1020
## 5       M white      M   Braintree Massachusetts Norfolk County 2184
## 6       S black      M   Braintree Massachusetts Norfolk County 2184
```

```r
male <- subset(patient_info, patient_info$GENDER == "M")

head(male)
```

```
##   Record_ID                            PATIENT_ID  BIRTHDATE DEATHDATE
AGE
## 2       532 9718334c-3289-4b1c-a017-72f3df283ab3  6/13/1951
72.92778
## 3      5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2   2/6/2006
18.28056
## 5     10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a  6/24/1948 3/31/2020
75.89722
## 6      7818 e283d725-b355-4b86-98a5-b8274e643527   9/1/1992
31.71111
## 7      8381 696eec8a-73f1-4aa3-a730-9609c4b6657a 10/30/2009
14.54722
## 9     12062 e586d6a9-2450-4931-9b36-91f5306acd31 10/27/1935 3/11/2020
88.55556
##   MARITAL  RACE GENDER        CITY         STATE           COUNTY  ZIP
## 2       M black      M      Boston Massachusetts   Suffolk County 2113
## 3         white      M  Foxborough Massachusetts   Norfolk County 2035
## 5       M white      M   Braintree Massachusetts   Norfolk County 2184
## 6       S black      M   Braintree Massachusetts   Norfolk County 2184
## 7         asian      M    Westwood Massachusetts   Norfolk County   NA
## 9       M white      M Northbridge Massachusetts Worcester County   NA
```

```r
library(dplyr)

male_condition <- inner_join(covid_symptoms, male,
                     by = "PATIENT_ID")

head(male_condition)
```

```
##   Record_ID.x     START      STOP                              PATIENT_ID
## 1           3  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2           4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3           5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 4          16 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 5          17 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 6          18 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
##                          ENCOUNTER_ID      CODE        DESCRIPTION
Record_ID.y
## 1 681c380b-3c84-4c55-80a6-db3d9ea12fee 386661006    Fever (finding)
1
## 2 681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
1
## 3 681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
1
## 4 eeab7c2d-71ba-4e04-af16-87a01dce7d54  25064002 Headache (finding)
3
```

```
## 5 eeab7c2d-71ba-4e04-af16-87a01dce7d54  49727002    Cough (finding)
3
## 6 eeab7c2d-71ba-4e04-af16-87a01dce7d54  84229001  Fatigue (finding)
3
##   BIRTHDATE DEATHDATE       AGE MARITAL  RACE GENDER        CITY
STATE
## 1 8/24/2017            6.730556         white      M Springfield
Massachusetts
## 2 8/24/2017            6.730556         white      M Springfield
Massachusetts
## 3 8/24/2017            6.730556         white      M Springfield
Massachusetts
## 4 6/30/1992           31.880556       S white      M    Chicopee
Massachusetts
## 5 6/30/1992           31.880556       S white      M    Chicopee
Massachusetts
## 6 6/30/1992           31.880556       S white      M    Chicopee
Massachusetts
##           COUNTY  ZIP
## 1 Hampden County 1106
## 2 Hampden County 1106
## 3 Hampden County 1106
## 4 Hampden County 1020
## 5 Hampden County 1020
## 6 Hampden County 1020

# Find the Top 10 conditions for Male

library(dplyr)

male_condition_counts <- male_condition %>% count(DESCRIPTION)

head(male_condition_counts)

##               DESCRIPTION    n
## 1               COVID-19 3161
## 2         Cough (finding) 2202
## 3       Fatigue (finding) 1271
## 4         Fever (finding) 2886
## 5      Headache (finding)  473
## 6 Loss of taste (finding) 1725

# Sort TOP 10 condition descending

male_top_conditions <- male_condition_counts %>%
  arrange(desc(n)) %>%
  head(10)  # Select the top 10 conditions

# Print the top 10 conditions for Male
print(male_top_conditions)
```

```
##                            DESCRIPTION    n
## 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                      Headache (finding)  473
## 8                        Nausea (finding)  173
## 9              Vomiting symptom (finding)  173
## 10 Streptococcal sore throat (disorder)   68

# =====
# 2.2.2. Top 10 conditions for Female ====
# =====

head(patient_info)

##   Record_ID                            PATIENT_ID  BIRTHDATE DEATHDATE
AGE
## 1      3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8  3/17/1988
36.16667
## 2       532 9718334c-3289-4b1c-a017-72f3df283ab3  6/13/1951
72.92778
## 3      5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2   2/6/2006
18.28056
## 4      7462 c10ee469-6182-4228-ac26-21bcf2412337 10/28/1912 11/1/2016
111.55278
## 5     10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a  6/24/1948 3/31/2020
75.89722
## 6      7818 e283d725-b355-4b86-98a5-b8274e643527   9/1/1992
31.71111
##   MARITAL  RACE GENDER        CITY         STATE         COUNTY  ZIP
## 1       M white      F    Rehoboth Massachusetts Bristol County   NA
## 2       M black      M      Boston Massachusetts Suffolk County 2113
## 3         white      M  Foxborough Massachusetts Norfolk County 2035
## 4       S black      F Springfield Massachusetts Hampden County 1020
## 5       M white      M   Braintree Massachusetts Norfolk County 2184
## 6       S black      M   Braintree Massachusetts Norfolk County 2184

female <- subset(patient_info, patient_info$GENDER == "F")

head(female)

##    Record_ID                            PATIENT_ID  BIRTHDATE DEATHDATE
## 1       3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8  3/17/1988
## 4       7462 c10ee469-6182-4228-ac26-21bcf2412337 10/28/1912 11/1/2016
## 8       6826 c3cad3d8-cbfa-42f3-8901-50bf814f2449  7/20/1922
## 11      6075 6b26b6e2-a7ff-4353-8b0d-48c712d4a31e 12/28/1926 2/23/1990
## 12      2356 febba7e3-0689-4b82-a0c6-f93b992e6415 12/31/1998
## 13     10598 97b6bb1a-78cd-48bb-9e03-39867c5c368e  9/12/2005
```

```
##          AGE MARITAL  RACE GENDER        CITY         STATE         COUNTY
ZIP
## 1   36.16667       M white      F    Rehoboth Massachusetts Bristol County
NA
## 4  111.55278       S black      F Springfield Massachusetts Hampden County
1020
## 8  101.82500       M white      F      Boston Massachusetts Suffolk County
2128
## 11  97.38611       M white      F  Foxborough Massachusetts Norfolk County
NA
## 12  25.38056         white      F      Ludlow Massachusetts Hampden County
NA
## 13  18.68056         white      F     Norwood Massachusetts Norfolk County
NA
```

```
library(dplyr)

female_condition <- inner_join(covid_symptoms, female,
                       by = "PATIENT_ID")

head(female_condition)
```

```
##   Record_ID.x     START      STOP                              PATIENT_ID
## 1           7 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 2          10 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 3          11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 4          12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 5          13 4/28/2020  5/8/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 6          25  3/1/2020  4/7/2020 199c586f-af16-4091-9998-ee4cfc02ee7a
##                             ENCOUNTER_ID      CODE
## 1 1ea74a77-3ad3-4948-a9cc-3084462035d6  49727002
## 2 1ea74a77-3ad3-4948-a9cc-3084462035d6 386661006
## 3 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004
## 4 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006
## 5 e03b96de-5604-4989-a2d5-03a63e041eab  43878008
## 6 8333efdf-f7bf-43bb-b73f-2b663d14c1ad  49727002
##                              DESCRIPTION Record_ID.y BIRTHDATE DEATHDATE
## 1                       Cough (finding)           2  8/1/2016
## 2                       Fever (finding)           2  8/1/2016
## 3                     Suspected COVID-19          2  8/1/2016
## 4                              COVID-19           2  8/1/2016
## 5 Streptococcal sore throat (disorder)           2  8/1/2016
## 6                       Cough (finding)           4  1/9/2004
##         AGE MARITAL  RACE GENDER     CITY         STATE          COUNTY
ZIP
## 1  7.794444         white      F  Walpole Massachusetts  Norfolk County
2081
## 2  7.794444         white      F  Walpole Massachusetts  Norfolk County
2081
## 3  7.794444         white      F  Walpole Massachusetts  Norfolk County
```

```
2081
## 4  7.794444         white      F  Walpole Massachusetts  Norfolk County
2081
## 5  7.794444         white      F  Walpole Massachusetts  Norfolk County
2081
## 6 20.355556         white      F Pembroke Massachusetts Plymouth County
NA
```

```r
# Find the Top 10 conditions for Female

library(dplyr)

female_condition_counts <- female_condition %>% count(DESCRIPTION)

head(female_condition_counts)
```

```
##               DESCRIPTION    n
## 1               COVID-19 3487
## 2         Cough (finding) 2472
## 3       Fatigue (finding) 1373
## 4         Fever (finding) 3202
## 5      Headache (finding)  518
## 6 Loss of taste (finding) 1846
```

```r
# Sort TOP 10 condition descending

female_top_conditions <- female_condition_counts %>%
  arrange(desc(n)) %>%
  head(10)  # Select the top 10 conditions

# Print the top 10 conditions for Male
print(female_top_conditions)
```

```
##                            DESCRIPTION    n
## 1                   Suspected COVID-19 3598
## 2                             COVID-19 3487
## 3                        Fever (finding) 3202
## 4                        Cough (finding) 2472
## 5               Loss of taste (finding) 1846
## 6                      Fatigue (finding) 1373
## 7                     Headache (finding)  518
## 8                       Nausea (finding)  202
## 9            Vomiting symptom (finding)  202
## 10 Streptococcal sore throat (disorder)   61
```

```r
# Compare Male and Female Top 10 Conditions: ====
cat("Male Top 10 Conditions:", "\n")
```

```
## Male Top 10 Conditions:
```

```r
print(male_top_conditions)
```

```
##                            DESCRIPTION    n
## 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                      Headache (finding)  473
## 8                        Nausea (finding)  173
## 9              Vomiting symptom (finding)  173
## 10 Streptococcal sore throat (disorder)   68
```

```
cat("Female Top 10 Conditions:", "\n")
```

```
## Female Top 10 Conditions:
```

```
print(female_top_conditions)
```

```
##                            DESCRIPTION    n
## 1                    Suspected COVID-19 3598
## 2                              COVID-19 3487
## 3                         Fever (finding) 3202
## 4                         Cough (finding) 2472
## 5                 Loss of taste (finding) 1846
## 6                        Fatigue (finding) 1373
## 7                      Headache (finding)  518
## 8                        Nausea (finding)  202
## 9              Vomiting symptom (finding)  202
## 10 Streptococcal sore throat (disorder)   61
```

```
cat("====================", "\n")
```

```
## ====================
```

```
cat("Key findings:", "\n")
```

```
## Key findings:
```

```
cat("====================", "\n")
```

```
## ====================
```

```
cat("There is a slightly difference in COVID symptoms
    between Male and Female", "\n")
```

```
## There is a slightly difference in COVID symptoms
##     between Male and Female
```

```
cat("Female has more COVID related patients than Male", "\n")
```

```
## Female has more COVID related patients than Male
```

```r
# == Task 3 ====

# Join 3 tables: Encounters, Conditions, and Patients
# To examine 2 factors of COVID Patients: Age Group and Gender

# 3.1. Join Conditions and Encounters ====

# Filter Covid conditions in dataset Conditions ====

# We already had this filtered dataset

head(condition_covid)
```

```
##    Record_ID     START      STOP                            PATIENT_ID
## 4          4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 5          5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 11        11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 12        12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 23        23 3/11/2020 3/11/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 24        24 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
##                            ENCOUNTER_ID      CODE        DESCRIPTION
## 4  681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
## 5  681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
## 11 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
## 12 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006           COVID-19
## 23 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840544004 Suspected COVID-19
## 24 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840539006           COVID-19
```

```r
# Filter Encounter Class in the dataset Encounters ====

# Read the dataset

encounter_info <- read.csv("data/encountersUG.csv")

head(encounter_info)
```

```
##   Record_ID                         ENCOUNTER_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
##                   STOP                            PATIENT_ID ENCOUNTERCLASS
## 1 2019-02-16T01:17:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     outpatient
## 2 2019-08-02T01:32:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271       wellness
## 3 2019-10-31T01:17:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     outpatient
## 4 2020-01-31T01:17:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271       wellness
## 5 2020-03-02T01:58:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     ambulatory
## 6 2019-07-08T08:17:25Z 067318a4-db8f-447f-8b6e-f2f61e9baaa5       wellness
```

```
##        CODE                        DESCRIPTION BASE_ENCOUNTER_COST
## 1 185345009             Encounter for symptom              129.16
## 2 410620009      Well child visit (procedure)              129.16
## 3 185345009             Encounter for symptom              129.16
## 4 410620009      Well child visit (procedure)              129.16
## 5 185345009 Encounter for symptom (procedure)              129.16
## 6 410620009      Well child visit (procedure)              129.16
##   TOTAL_CLAIM_COST PAYER_COVERAGE REASONCODE REASONDESCRIPTION
## 1           129.16          69.16   65363002      Otitis media
## 2           129.16         129.16         NA
## 3           129.16          69.16   65363002      Otitis media
## 4           129.16         129.16         NA
## 5           129.16          69.16         NA
## 6           129.16         129.16         NA

# Filter the dataset

encounter_filtered <- subset(encounter_info,
                             (encounter_info$ENCOUNTERCLASS == "ambulatory")
                             | (encounter_info$ENCOUNTERCLASS == "emergency")
                             | (encounter_info$ENCOUNTERCLASS == "inpatient")
                             | (encounter_info$ENCOUNTERCLASS == "urgent
care"))

# See the filtered dataset

head(encounter_filtered)

##    Record_ID                         ENCOUNTER_ID                 START
## 5          5 681c380b-3c84-4c55-80a6-db3d9ea12fee 2020-03-02T01:02:32Z
## 8          8 adedca64-700b-4fb9-82f1-9cbb658abb73 2020-02-12T08:02:25Z
## 9          9 1ea74a77-3ad3-4948-a9cc-3084462035d6 2020-03-13T08:02:25Z
## 10        10 e03b96de-5604-4989-a2d5-03a63e041eab 2020-04-28T08:02:25Z
## 12        12 11a2dfae-53d4-4d13-a74b-c540a525a1c4 2010-11-22T10:51:58Z
## 22        22 d875bd25-9aa8-464a-8021-a5a1a05a9599 2019-11-24T10:51:58Z
##                    STOP                            PATIENT_ID
ENCOUNTERCLASS
## 5  2020-03-02T01:58:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
ambulatory
## 8  2020-02-12T09:02:25Z 067318a4-db8f-447f-8b6e-f2f61e9baaa5
emergency
## 9  2020-03-13T08:52:25Z 067318a4-db8f-447f-8b6e-f2f61e9baaa5
ambulatory
## 10 2020-04-28T08:32:25Z 067318a4-db8f-447f-8b6e-f2f61e9baaa5
ambulatory
## 12 2010-11-22T11:06:58Z ae9efba3-ddc4-43f9-a781-f72019388548
ambulatory
## 22 2019-11-24T11:06:58Z ae9efba3-ddc4-43f9-a781-f72019388548
ambulatory
##         CODE                          DESCRIPTION BASE_ENCOUNTER_COST
```

```
## 5  185345009    Encounter for symptom (procedure)              129.16
## 8   50849002 Emergency room admission (procedure)              129.16
## 9  185345009    Encounter for symptom (procedure)              129.16
## 10 185345009                 Encounter for symptom             129.16
## 12 390906007     Hypertension follow-up encounter              129.16
## 22 185345009                 Encounter for symptom             129.16
##    TOTAL_CLAIM_COST PAYER_COVERAGE REASONCODE
## 5            129.16          69.16         NA
## 8            129.16          59.16         NA
## 9            129.16          59.16         NA
## 10           129.16          59.16   43878008
## 12           129.16          49.16         NA
## 22           129.16          49.16  444814009
##                       REASONDESCRIPTION
## 5
## 8
## 9
## 10 Streptococcal sore throat (disorder)
## 12
## 22           Viral sinusitis (disorder)
```

```
# Join the two datasets: conditions, encounters

library(dplyr)

merged_data_2 <- inner_join(condition_covid, encounter_filtered,
                            by = "ENCOUNTER_ID")

head(merged_data_2)
```

```
##   Record_ID.x   START.x    STOP.x                          PATIENT_ID.x
## 1           4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2           5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3          11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 4          12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 5          23 3/11/2020 3/11/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 6          24 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
##                           ENCOUNTER_ID    CODE.x      DESCRIPTION.x
Record_ID.y
## 1 681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
5
## 2 681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
5
## 3 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
9
## 4 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006           COVID-19
9
## 5 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840544004 Suspected COVID-19
23
## 6 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840539006           COVID-19
```

```
23
##                 START.y               STOP.y
## 1 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z
## 2 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z
## 3 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z
## 4 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z
## 5 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z
## 6 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z
##                           PATIENT_ID.y ENCOUNTERCLASS    CODE.y
## 1 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     ambulatory 185345009
## 2 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     ambulatory 185345009
## 3 067318a4-db8f-447f-8b6e-f2f61e9baaa5     ambulatory 185345009
## 4 067318a4-db8f-447f-8b6e-f2f61e9baaa5     ambulatory 185345009
## 5 ae9efba3-ddc4-43f9-a781-f72019388548     ambulatory 185345009
## 6 ae9efba3-ddc4-43f9-a781-f72019388548     ambulatory 185345009
##                       DESCRIPTION.y BASE_ENCOUNTER_COST TOTAL_CLAIM_COST
## 1 Encounter for symptom (procedure)              129.16           129.16
## 2 Encounter for symptom (procedure)              129.16           129.16
## 3 Encounter for symptom (procedure)              129.16           129.16
## 4 Encounter for symptom (procedure)              129.16           129.16
## 5 Encounter for symptom (procedure)              129.16           129.16
## 6 Encounter for symptom (procedure)              129.16           129.16
##   PAYER_COVERAGE REASONCODE REASONDESCRIPTION
## 1          69.16         NA
## 2          69.16         NA
## 3          59.16         NA
## 4          59.16         NA
## 5          49.16         NA
## 6          49.16         NA

# Prepare to join the 2 data frames: merged_data_2 and patients
# Join via "PATIENT_ID" column
# But in the merged_data_2 there are 2 Patient_id columns
# They are: Patient_id.x and Patient_id.y
# We need to:

# 1. Delete one column: Patient_id.y

merged_data_2$PATIENT_ID.y <- NULL

# See the dataset
head(merged_data_2)

##   Record_ID.x   START.x    STOP.x                        PATIENT_ID.x
## 1           4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2           5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3          11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 4          12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 5          23 3/11/2020 3/11/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 6          24 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
```

```
##                            ENCOUNTER_ID    CODE.x      DESCRIPTION.x
Record_ID.y
## 1 681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
5
## 2 681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
5
## 3 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
9
## 4 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006           COVID-19
9
## 5 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840544004 Suspected COVID-19
23
## 6 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840539006           COVID-19
23
##                 START.y               STOP.y ENCOUNTERCLASS    CODE.y
## 1 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z     ambulatory 185345009
## 2 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z     ambulatory 185345009
## 3 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z     ambulatory 185345009
## 4 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z     ambulatory 185345009
## 5 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z     ambulatory 185345009
## 6 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z     ambulatory 185345009
##                       DESCRIPTION.y BASE_ENCOUNTER_COST TOTAL_CLAIM_COST
## 1 Encounter for symptom (procedure)              129.16           129.16
## 2 Encounter for symptom (procedure)              129.16           129.16
## 3 Encounter for symptom (procedure)              129.16           129.16
## 4 Encounter for symptom (procedure)              129.16           129.16
## 5 Encounter for symptom (procedure)              129.16           129.16
## 6 Encounter for symptom (procedure)              129.16           129.16
##   PAYER_COVERAGE REASONCODE REASONDESCRIPTION
## 1          69.16         NA
## 2          69.16         NA
## 3          59.16         NA
## 4          59.16         NA
## 5          49.16         NA
## 6          49.16         NA
```

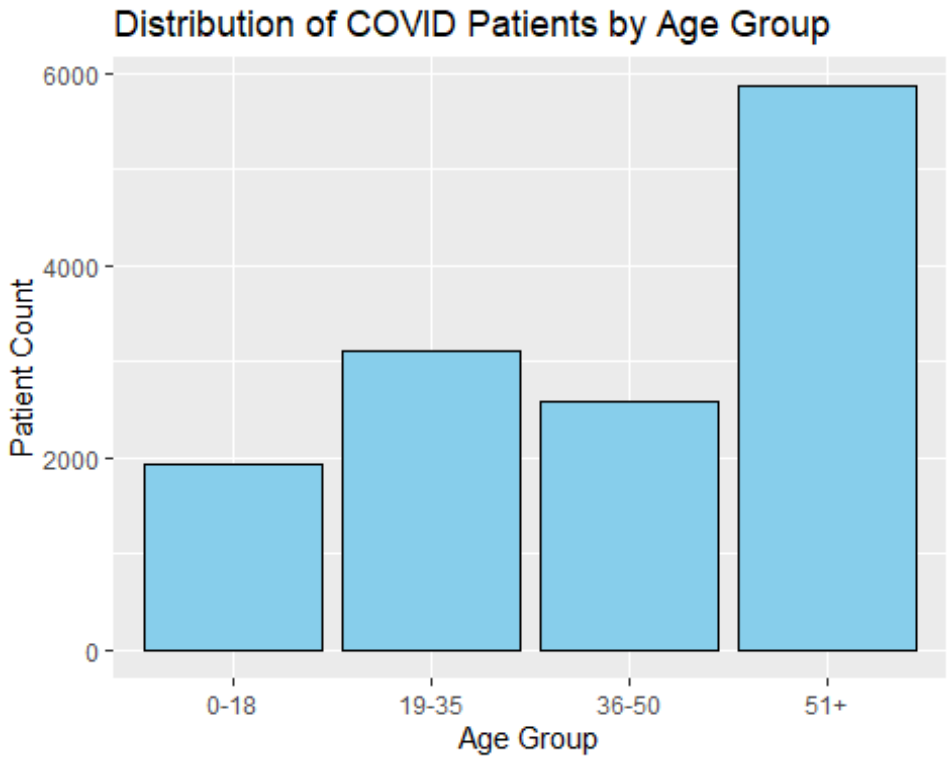
```r
# 2. Rename the column: Patient_id.x to Patient_id

colnames(merged_data_2)[colnames(merged_data_2) == "PATIENT_ID.x"] <-
"PATIENT_ID"

# See the dataset
head(merged_data_2)
```

```
##   Record_ID.x   START.x    STOP.x                            PATIENT_ID
## 1           4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2           5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3          11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 4          12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 5          23 3/11/2020 3/11/2020 ae9efba3-ddc4-43f9-a781-f72019388548
```

```
## 6          24 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
##                            ENCOUNTER_ID    CODE.x      DESCRIPTION.x
Record_ID.y
## 1 681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
5
## 2 681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
5
## 3 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
9
## 4 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006           COVID-19
9
## 5 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840544004 Suspected COVID-19
23
## 6 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840539006           COVID-19
23
##                 START.y               STOP.y ENCOUNTERCLASS    CODE.y
## 1 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z     ambulatory 185345009
## 2 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z     ambulatory 185345009
## 3 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z     ambulatory 185345009
## 4 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z     ambulatory 185345009
## 5 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z     ambulatory 185345009
## 6 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z     ambulatory 185345009
##                       DESCRIPTION.y BASE_ENCOUNTER_COST TOTAL_CLAIM_COST
## 1 Encounter for symptom (procedure)              129.16           129.16
## 2 Encounter for symptom (procedure)              129.16           129.16
## 3 Encounter for symptom (procedure)              129.16           129.16
## 4 Encounter for symptom (procedure)              129.16           129.16
## 5 Encounter for symptom (procedure)              129.16           129.16
## 6 Encounter for symptom (procedure)              129.16           129.16
##   PAYER_COVERAGE REASONCODE REASONDESCRIPTION
## 1          69.16         NA
## 2          69.16         NA
## 3          59.16         NA
## 4          59.16         NA
## 5          49.16         NA
## 6          49.16         NA

# 3. Join with dataset Patients via Patient_id column

patient <- inner_join(merged_data_2, patient_info,
                      by = "PATIENT_ID")

# See the result
head(patient)

##   Record_ID.x   START.x    STOP.x                           PATIENT_ID
## 1           4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2           5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3          11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 4          12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
```

```
## 5          23 3/11/2020 3/11/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 6          24 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
##                           ENCOUNTER_ID    CODE.x      DESCRIPTION.x
Record_ID.y
## 1 681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
5
## 2 681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
5
## 3 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
9
## 4 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006           COVID-19
9
## 5 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840544004 Suspected COVID-19
23
## 6 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840539006           COVID-19
23
##                START.y               STOP.y ENCOUNTERCLASS    CODE.y
## 1 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z     ambulatory 185345009
## 2 2020-03-02T01:02:32Z 2020-03-02T01:58:32Z     ambulatory 185345009
## 3 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z     ambulatory 185345009
## 4 2020-03-13T08:02:25Z 2020-03-13T08:52:25Z     ambulatory 185345009
## 5 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z     ambulatory 185345009
## 6 2020-03-11T10:51:58Z 2020-03-11T11:42:58Z     ambulatory 185345009
##                        DESCRIPTION.y BASE_ENCOUNTER_COST TOTAL_CLAIM_COST
## 1 Encounter for symptom (procedure)              129.16           129.16
## 2 Encounter for symptom (procedure)              129.16           129.16
## 3 Encounter for symptom (procedure)              129.16           129.16
## 4 Encounter for symptom (procedure)              129.16           129.16
## 5 Encounter for symptom (procedure)              129.16           129.16
## 6 Encounter for symptom (procedure)              129.16           129.16
##   PAYER_COVERAGE REASONCODE REASONDESCRIPTION Record_ID BIRTHDATE
DEATHDATE
## 1          69.16         NA                           1 8/24/2017
## 2          69.16         NA                           1 8/24/2017
## 3          59.16         NA                           2  8/1/2016
## 4          59.16         NA                           2  8/1/2016
## 5          49.16         NA                           3 6/30/1992
## 6          49.16         NA                           3 6/30/1992
##         AGE MARITAL  RACE GENDER        CITY         STATE         COUNTY
ZIP
## 1  6.730556         white      M Springfield Massachusetts Hampden County
1106
## 2  6.730556         white      M Springfield Massachusetts Hampden County
1106
## 3  7.794444         white      F     Walpole Massachusetts Norfolk County
2081
## 4  7.794444         white      F     Walpole Massachusetts Norfolk County
2081
## 5 31.880556       S white      M    Chicopee Massachusetts Hampden County
1020
```

```
## 6 31.880556       S white      M    Chicopee Massachusetts Hampden County
1020

# To examine 2 factors: Age Group and Gender ====

# Top Age Group  that has Covid ====

patient$AGEGROUP <- cut(patient$AGE, breaks = c(0, 18, 35, 50, Inf),
                            labels = c("0-18", "19-35", "36-50", "51+"))

# See the number of patient by Age Group

summary(patient$AGEGROUP)

##  0-18 19-35 36-50   51+
##  1930  3106  2594  5863

# Plot the histogram for Age Group

library(ggplot2)

ggplot(patient, aes(x = AGEGROUP)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of COVID Patients by Age Group",
       x = "Age Group",
       y = "Patient Count")
```



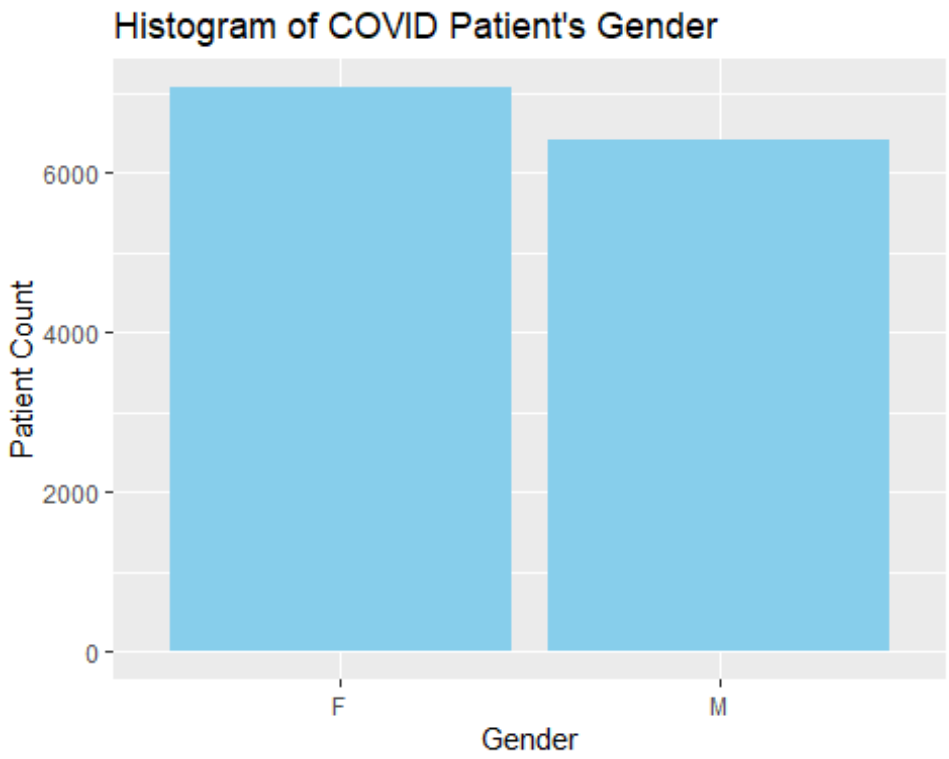
```r
# For the Gender: ====

library(dplyr)

# Count the occurrences of each Gender
gender_counts <- patient %>% count(GENDER)

# Print the result
print(gender_counts)
```

```
##   GENDER    n
## 1      F 7076
## 2      M 6417
```

```r
# See the data type
str(gender_counts)
```

```
## 'data.frame':    2 obs. of  2 variables:
##  $ GENDER: chr  "F" "M"
##  $ n     : int  7076 6417
```

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

# Create the histogram of Gender
ggplot(gender_counts, aes(x = GENDER, y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Histogram of COVID Patient's Gender",
       x = "Gender", y = "Patient Count")
```




```
cat("==================================", "\n")
```

```
## ==================================
```

```
cat("Key findings: ", "we explore that", "\n")
```

```
## Key findings:  we explore that
```

```
cat("For the Age Group:", "\n")
```

```
## For the Age Group:
```

```
summary(patient$AGEGROUP)
```

```
##  0-18 19-35 36-50   51+
##  1930  3106  2594  5863
```

```
cat("Most COVID Patients are in the Age Group from 51+ yrs old", "\n")
```

```
## Most COVID Patients are in the Age Group from 51+ yrs old
```

```
cat("Then followed by the Age Group from 19 to 35 yrs old", "\n")
```

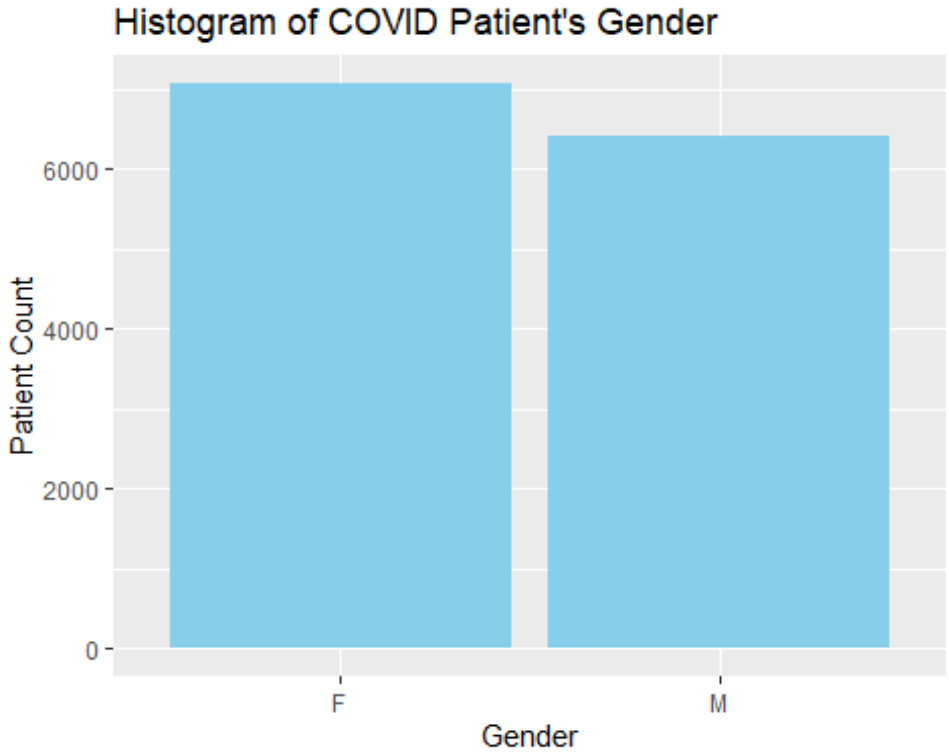
```
## Then followed by the Age Group from 19 to 35 yrs old
```

```
cat("==================================", "\n")
```

```
## ==================================
```

```
cat("For the Gender of COVID Patients:", "\n")
```

```
## For the Gender of COVID Patients:

cat("Look at the Histogram: ", "\n")

## Look at the Histogram:

ggplot(gender_counts, aes(x = GENDER, y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Histogram of COVID Patient's Gender",
       x = "Gender", y = "Patient Count")
```




```
cat("We see that Female COVID Patients are more than Male COVID Patients",
    "\n")

## We see that Female COVID Patients are more than Male COVID Patients

# ======================
# Task 4 ======
# 4.1. Filter Patients who Died because of COVID

# We will filter the dataset Encounters which description
# is "Death Certificate"

# We already had the data frame enounter_info

# See the data frame
head(encounter_info)
```

```
##   Record_ID                          ENCOUNTER_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
##                    STOP                              PATIENT_ID ENCOUNTERCLASS
## 1 2019-02-16T01:17:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     outpatient
## 2 2019-08-02T01:32:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271       wellness
## 3 2019-10-31T01:17:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     outpatient
## 4 2020-01-31T01:17:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271       wellness
## 5 2020-03-02T01:58:32Z f0f3bc8d-ef38-49ce-a2bd-dfdda982b271     ambulatory
## 6 2019-07-08T08:17:25Z 067318a4-db8f-447f-8b6e-f2f61e9baaa5       wellness
##        CODE                        DESCRIPTION BASE_ENCOUNTER_COST
## 1 185345009             Encounter for symptom              129.16
## 2 410620009      Well child visit (procedure)              129.16
## 3 185345009             Encounter for symptom              129.16
## 4 410620009      Well child visit (procedure)              129.16
## 5 185345009 Encounter for symptom (procedure)              129.16
## 6 410620009      Well child visit (procedure)              129.16
##   TOTAL_CLAIM_COST PAYER_COVERAGE REASONCODE REASONDESCRIPTION
## 1           129.16          69.16   65363002      Otitis media
## 2           129.16         129.16         NA
## 3           129.16          69.16   65363002      Otitis media
## 4           129.16         129.16         NA
## 5           129.16          69.16         NA
## 6           129.16         129.16         NA

# Filter the data
encounter_death <- subset(encounter_info,
                          encounter_info$DESCRIPTION == "Death
Certification")

# See the filtered data frame
head(encounter_death)

##      Record_ID                          ENCOUNTER_ID                START
## 86          94 a966109f-28b5-49ae-8700-02bcc2bb56d6 2016-05-14T13:29:26Z
## 521        693 66d1a1cd-9876-409a-a1cb-3f9836ac8249 2020-03-25T15:51:30Z
## 741        964 7fe393e9-35e3-4329-8c79-14d0a0f7ed7b 2010-04-23T17:07:19Z
## 947       1346 48d25632-6b54-444e-a9ba-553962f2fa04 1999-09-05T13:11:46Z
## 1032      1621 77021032-f641-4b84-9596-383aae341b45 2020-03-31T19:29:43Z
## 1171      1817 8df41367-f686-4681-9ee5-7e39613ea66b 2007-09-30T06:35:54Z
##                     STOP                              PATIENT_ID
ENCOUNTERCLASS
## 86   2016-05-14T13:44:26Z 58cac9ec-4baa-46c1-b919-0ed13572b51d
wellness
## 521  2020-03-25T16:06:30Z c70992c9-ff13-467b-9032-1901506edeef
wellness
```

```
## 741  2010-04-23T17:22:19Z cd7cdbef-ef08-48da-9302-19bf4b2dffa3
wellness
## 947  1999-09-05T13:26:46Z 419c213d-8d35-440c-bf47-a536f95df517
wellness
## 1032 2020-03-31T19:44:43Z edad31f3-5a08-4678-8d31-271a41a2aad5
wellness
## 1171 2007-09-30T06:50:54Z 944884c0-5358-479e-bdf5-c0dd87739ef7
wellness
##           CODE         DESCRIPTION BASE_ENCOUNTER_COST TOTAL_CLAIM_COST
## 86   308646001 Death Certification              129.16           129.16
## 521  308646001 Death Certification              129.16           129.16
## 741  308646001 Death Certification              129.16           129.16
## 947  308646001 Death Certification              129.16           129.16
## 1032 308646001 Death Certification              129.16           129.16
## 1171 308646001 Death Certification              129.16           129.16
##      PAYER_COVERAGE REASONCODE                       REASONDESCRIPTION
## 86                0  230690007                                  Stroke
## 521               0  840539006                               COVID-19
## 741               0  254837009 Malignant neoplasm of breast (disorder)
## 947               0   22298006                   Myocardial Infarction
## 1032              0  840539006                               COVID-19
## 1171              0  230690007                                  Stroke
```

```
# See the Covid condition
head(condition_covid)
```

```
##    Record_ID     START      STOP                              PATIENT_ID
## 4          4  3/1/2020  3/1/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 5          5  3/1/2020 3/30/2020 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 11        11 3/13/2020 3/13/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 12        12 3/13/2020 4/14/2020 067318a4-db8f-447f-8b6e-f2f61e9baaa5
## 23        23 3/11/2020 3/11/2020 ae9efba3-ddc4-43f9-a781-f72019388548
## 24        24 3/11/2020 4/15/2020 ae9efba3-ddc4-43f9-a781-f72019388548
##                            ENCOUNTER_ID      CODE        DESCRIPTION
## 4  681c380b-3c84-4c55-80a6-db3d9ea12fee 840544004 Suspected COVID-19
## 5  681c380b-3c84-4c55-80a6-db3d9ea12fee 840539006           COVID-19
## 11 1ea74a77-3ad3-4948-a9cc-3084462035d6 840544004 Suspected COVID-19
## 12 1ea74a77-3ad3-4948-a9cc-3084462035d6 840539006           COVID-19
## 23 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840544004 Suspected COVID-19
## 24 eeab7c2d-71ba-4e04-af16-87a01dce7d54 840539006           COVID-19
```

```
# Join the 2 data frames: encounter_recover and condition_covid

merged_data_3 <- merge(encounter_death, condition_covid,
                           by = "ENCOUNTER_ID", all = TRUE)

head(merged_data_3)
```

```
##                            ENCOUNTER_ID Record_ID.x              START.x
## 1 00048220-08df-43fa-97a0-045db65d789f          NA                 <NA>
## 2 00048220-08df-43fa-97a0-045db65d789f          NA                 <NA>
```

```
## 3 00171db0-2f5b-449e-947b-862440182c68          NA                  <NA>
## 4 00171db0-2f5b-449e-947b-862440182c68          NA                  <NA>
## 5 001a4cda-4939-49cc-9627-38fa6485a85a       95360 1977-09-14T03:37:33Z
## 6 001e5792-8b36-4291-99d8-23404c9226f0          NA                  <NA>
##                 STOP.x                           PATIENT_ID.x ENCOUNTERCLASS
## 1                 <NA>                                   <NA>           <NA>
## 2                 <NA>                                   <NA>           <NA>
## 3                 <NA>                                   <NA>           <NA>
## 4                 <NA>                                   <NA>           <NA>
## 5 1977-09-14T03:52:33Z c8110c1e-34fd-4c56-88ce-4b3dc0320fdc       wellness
## 6                 <NA>                                   <NA>           <NA>
##      CODE.x       DESCRIPTION.x BASE_ENCOUNTER_COST TOTAL_CLAIM_COST
## 1        NA                <NA>                  NA               NA
## 2        NA                <NA>                  NA               NA
## 3        NA                <NA>                  NA               NA
## 4        NA                <NA>                  NA               NA
## 5 308646001 Death Certification              129.16           129.16
## 6        NA                <NA>                  NA               NA
##   PAYER_COVERAGE REASONCODE                      REASONDESCRIPTION
Record_ID.y
## 1             NA         NA                                   <NA>
11260
## 2             NA         NA                                   <NA>
11261
## 3             NA         NA                                   <NA>
1058
## 4             NA         NA                                   <NA>
1059
## 5              0   94260004 Secondary malignant neoplasm of colon
NA
## 6             NA         NA                                   <NA>
24321
##    START.y    STOP.y                          PATIENT_ID.y    CODE.y
## 1 3/6/2020  3/6/2020 5b564764-d10b-4984-b3a6-e6c2fe56ae36 840544004
## 2 3/6/2020 3/22/2020 5b564764-d10b-4984-b3a6-e6c2fe56ae36 840539006
## 3 3/3/2020  3/4/2020 059b0e22-b52b-408d-b8b5-cdeb6ad9ce76 840544004
## 4 3/4/2020  4/4/2020 059b0e22-b52b-408d-b8b5-cdeb6ad9ce76 840539006
## 5     <NA>      <NA>                                <NA>        NA
## 6 3/3/2020  3/3/2020 bfec39f5-d29f-4b0c-b4f7-314142563ba4 840544004
##        DESCRIPTION.y
## 1 Suspected COVID-19
## 2           COVID-19
## 3 Suspected COVID-19
## 4           COVID-19
## 5               <NA>
## 6 Suspected COVID-19

# Join with Patients data frame
# We need to have one Patient_id in the merged_data_3
# We need to:
```

```r
# 1. Delete the column Patient_id.x

merged_data_3$PATIENT_ID.x <- NULL

# See the dataset
head(merged_data_3)
```

```
##                            ENCOUNTER_ID Record_ID.x              START.x
## 1 00048220-08df-43fa-97a0-045db65d789f          NA                 <NA>
## 2 00048220-08df-43fa-97a0-045db65d789f          NA                 <NA>
## 3 00171db0-2f5b-449e-947b-862440182c68          NA                 <NA>
## 4 00171db0-2f5b-449e-947b-862440182c68          NA                 <NA>
## 5 001a4cda-4939-49cc-9627-38fa6485a85a       95360 1977-09-14T03:37:33Z
## 6 001e5792-8b36-4291-99d8-23404c9226f0          NA                 <NA>
##                  STOP.x ENCOUNTERCLASS    CODE.x       DESCRIPTION.x
## 1                  <NA>           <NA>        NA                <NA>
## 2                  <NA>           <NA>        NA                <NA>
## 3                  <NA>           <NA>        NA                <NA>
## 4                  <NA>           <NA>        NA                <NA>
## 5 1977-09-14T03:52:33Z       wellness 308646001 Death Certification
## 6                  <NA>           <NA>        NA                <NA>
##   BASE_ENCOUNTER_COST TOTAL_CLAIM_COST PAYER_COVERAGE REASONCODE
## 1                  NA                NA             NA         NA
## 2                  NA                NA             NA         NA
## 3                  NA                NA             NA         NA
## 4                  NA                NA             NA         NA
## 5              129.16            129.16              0   94260004
## 6                  NA                NA             NA         NA
##                      REASONDESCRIPTION Record_ID.y  START.y    STOP.y
## 1                                 <NA>       11260 3/6/2020  3/6/2020
## 2                                 <NA>       11261 3/6/2020 3/22/2020
## 3                                 <NA>        1058 3/3/2020  3/4/2020
## 4                                 <NA>        1059 3/4/2020  4/4/2020
## 5 Secondary malignant neoplasm of colon          NA     <NA>       <NA>
## 6                                 <NA>       24321 3/3/2020  3/3/2020
##                           PATIENT_ID.y    CODE.y      DESCRIPTION.y
## 1 5b564764-d10b-4984-b3a6-e6c2fe56ae36 840544004 Suspected COVID-19
## 2 5b564764-d10b-4984-b3a6-e6c2fe56ae36 840539006           COVID-19
## 3 059b0e22-b52b-408d-b8b5-cdeb6ad9ce76 840544004 Suspected COVID-19
## 4 059b0e22-b52b-408d-b8b5-cdeb6ad9ce76 840539006           COVID-19
## 5                                 <NA>        NA               <NA>
## 6 bfec39f5-d29f-4b0c-b4f7-314142563ba4 840544004 Suspected COVID-19
```

```r
# 2. Rename the column: Patient_id.y to Patient_id

colnames(merged_data_3)[colnames(merged_data_3) == "PATIENT_ID.y"] <-
"PATIENT_ID"

head(merged_data_3)
```

```
##                            ENCOUNTER_ID Record_ID.x               START.x
## 1 00048220-08df-43fa-97a0-045db65d789f          NA                  <NA>
## 2 00048220-08df-43fa-97a0-045db65d789f          NA                  <NA>
## 3 00171db0-2f5b-449e-947b-862440182c68          NA                  <NA>
## 4 00171db0-2f5b-449e-947b-862440182c68          NA                  <NA>
## 5 001a4cda-4939-49cc-9627-38fa6485a85a       95360 1977-09-14T03:37:33Z
## 6 001e5792-8b36-4291-99d8-23404c9226f0          NA                  <NA>
##                  STOP.x ENCOUNTERCLASS    CODE.x       DESCRIPTION.x
## 1                  <NA>           <NA>        NA                <NA>
## 2                  <NA>           <NA>        NA                <NA>
## 3                  <NA>           <NA>        NA                <NA>
## 4                  <NA>           <NA>        NA                <NA>
## 5 1977-09-14T03:52:33Z       wellness 308646001 Death Certification
## 6                  <NA>           <NA>        NA                <NA>
##   BASE_ENCOUNTER_COST TOTAL_CLAIM_COST PAYER_COVERAGE REASONCODE
## 1                  NA                NA             NA         NA
## 2                  NA                NA             NA         NA
## 3                  NA                NA             NA         NA
## 4                  NA                NA             NA         NA
## 5              129.16            129.16              0   94260004
## 6                  NA                NA             NA         NA
##                      REASONDESCRIPTION Record_ID.y  START.y    STOP.y
## 1                                 <NA>       11260 3/6/2020  3/6/2020
## 2                                 <NA>       11261 3/6/2020 3/22/2020
## 3                                 <NA>        1058 3/3/2020  3/4/2020
## 4                                 <NA>        1059 3/4/2020  4/4/2020
## 5 Secondary malignant neoplasm of colon          NA     <NA>      <NA>
## 6                                 <NA>       24321 3/3/2020  3/3/2020
##                             PATIENT_ID    CODE.y      DESCRIPTION.y
## 1 5b564764-d10b-4984-b3a6-e6c2fe56ae36 840544004 Suspected COVID-19
## 2 5b564764-d10b-4984-b3a6-e6c2fe56ae36 840539006           COVID-19
## 3 059b0e22-b52b-408d-b8b5-cdeb6ad9ce76 840544004 Suspected COVID-19
## 4 059b0e22-b52b-408d-b8b5-cdeb6ad9ce76 840539006           COVID-19
## 5                                 <NA>        NA               <NA>
## 6 bfec39f5-d29f-4b0c-b4f7-314142563ba4 840544004 Suspected COVID-19
```

```
# Count the death patients,
# use 'dplyr' package to call the count() function

library(dplyr)

# See the number of dead people that had Death Certificate
death_counts <- merged_data_3 %>% count(DESCRIPTION.x)

print(death_counts)
```

```
##         DESCRIPTION.x     n
## 1 Death Certification  1291
## 2                <NA> 13511
```

```r
# See the number of Covid Patients:
covid_counts <- merged_data_3 %>% count(DESCRIPTION.y)

print(covid_counts)
```

```
##        DESCRIPTION.y    n
## 1           COVID-19 6648
## 2 Suspected COVID-19 6863
## 3               <NA> 1291
```

```r
# We see that dead patients are not died because of COVID

cat("Key Finding:", "\n")
```

```
## Key Finding:
```

```r
cat("Base on this below:", "\n")
```

```
## Base on this below:
```

```r
print(covid_counts)
```

```
##        DESCRIPTION.y    n
## 1           COVID-19 6648
## 2 Suspected COVID-19 6863
## 3               <NA> 1291
```

```r
cat("We see that dead patients ARE NOT died because of COVID", "\n")
```

```
## We see that dead patients ARE NOT died because of COVID
```

```r
# In the 2 datasets: Encounters and Conditions
# There is no info about "Recover" for COVID Patients
# So we cannot analyze more

# ==== END =====
```