Assignment AP

Dhyanvi Patel

2024-05-22

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. ***

Our data:

```
patients = read.csv("patientsUG.csv")
encounters = read.csv("encountersUG.csv")
conditions = read.csv("conditionsUG.csv")

head(patients)
```

```
##       X                                    Id  BIRTHDATE  DEATHDATE MARITAL
## 1  3600 6aa2e953-ad8f-48cb-909b-30fb9522ebf8 1988-03-17                   M
## 2   532 9718334c-3289-4b1c-a017-72f3df283ab3 1951-06-13                   M
## 3  5907 de9f5575-ae1c-4df5-9ef1-92a845ed99c2 2006-02-06
## 4  7462 c10ee469-6182-4228-ac26-21bcf2412337 1912-10-28 2016-11-01       S
## 5 10390 42ff8e5c-9607-490f-a256-dd6bbbd6ac2a 1948-06-24 2020-03-31       M
## 6  7818 e283d725-b355-4b86-98a5-b8274e643527 1992-09-01                   S
##    RACE GENDER        CITY         STATE         COUNTY  ZIP
## 1 white      F    Rehoboth Massachusetts Bristol County   NA
## 2 black      M      Boston Massachusetts Suffolk County 2113
## 3 white      M  Foxborough Massachusetts Norfolk County 2035
## 4 black      F Springfield Massachusetts Hampden County 1020
## 5 white      M   Braintree Massachusetts Norfolk County 2184
## 6 black      M   Braintree Massachusetts Norfolk County 2184
```

```
head(encounters)
```

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

```
head(conditions)
```

```
##   X      START       STOP                                PATIENT
## 1 1 2019-02-15 2019-08-01 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 2 2 2019-10-30 2020-01-30 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 3 3 2020-03-01 2020-03-30 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 4 4 2020-03-01 2020-03-01 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 5 5 2020-03-01 2020-03-30 f0f3bc8d-ef38-49ce-a2bd-dfdda982b271
## 6 6 2020-02-12 2020-02-26 067318a4-db8f-447f-8b6e-f2f61e9baaa5
##                              ENCOUNTER      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
```

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

```
covid_descriptions = c("COVID-19", "Suspected COVID-19")
covid_condition_indices = NULL

for (i in 1:nrow(conditions)) {
  if (conditions$DESCRIPTION[i] %in% covid_descriptions) {
    covid_condition_indices = c(covid_condition_indices, i)
  }
}

covid_conditions = conditions[covid_condition_indices, ]
```

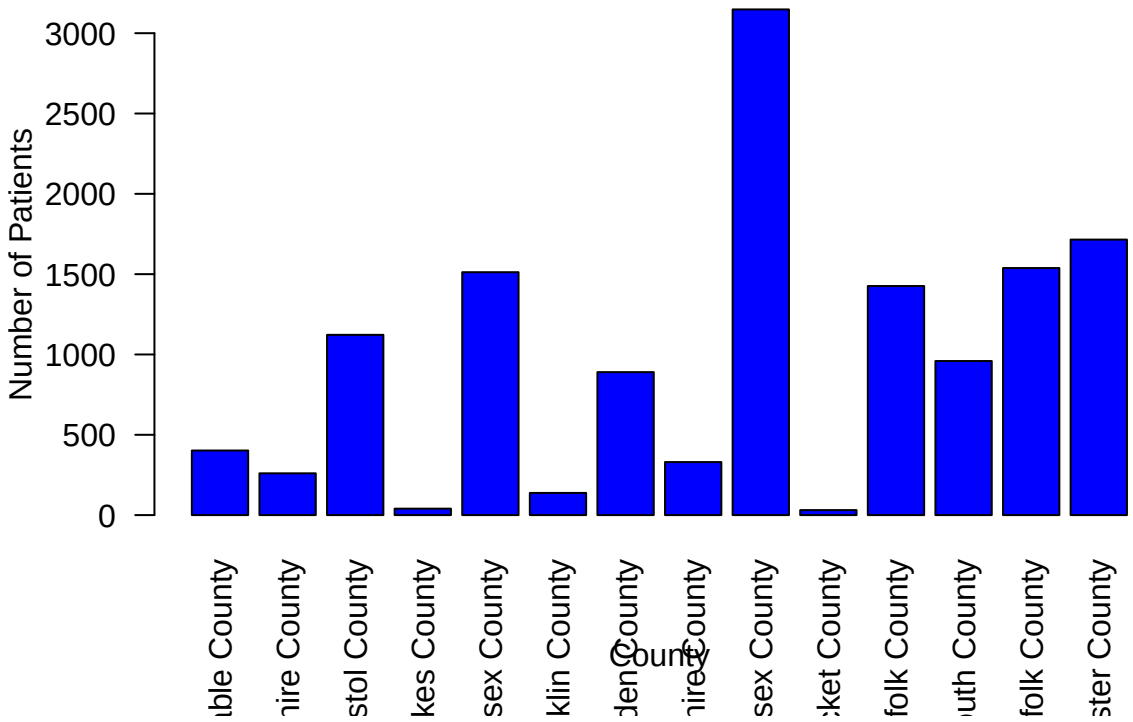
```
covid_patients = data.frame()

for (i in 1:nrow(covid_conditions)) {
  patient_id = covid_conditions$PATIENT[i]
  patient_row = patients[patients$Id == patient_id, ]
  covid_patients = rbind(covid_patients, patient_row)
}
```

```
county_distribution = table(covid_patients$COUNTY)
county_distribution
```

```
##
## Barnstable County  Berkshire County    Bristol County      Dukes County
##               402               260              1122                40
##      Essex County   Franklin County    Hampden County  Hampshire County
##              1512               138               890               330
##  Middlesex County  Nantucket County    Norfolk County   Plymouth County
##              3148                31              1426               959
##    Suffolk County  Worcester County
##              1538              1715
```

```
barplot(county_distribution, main="Distribution of COVID Patients Across Counties", xlab="County", ylab=
```



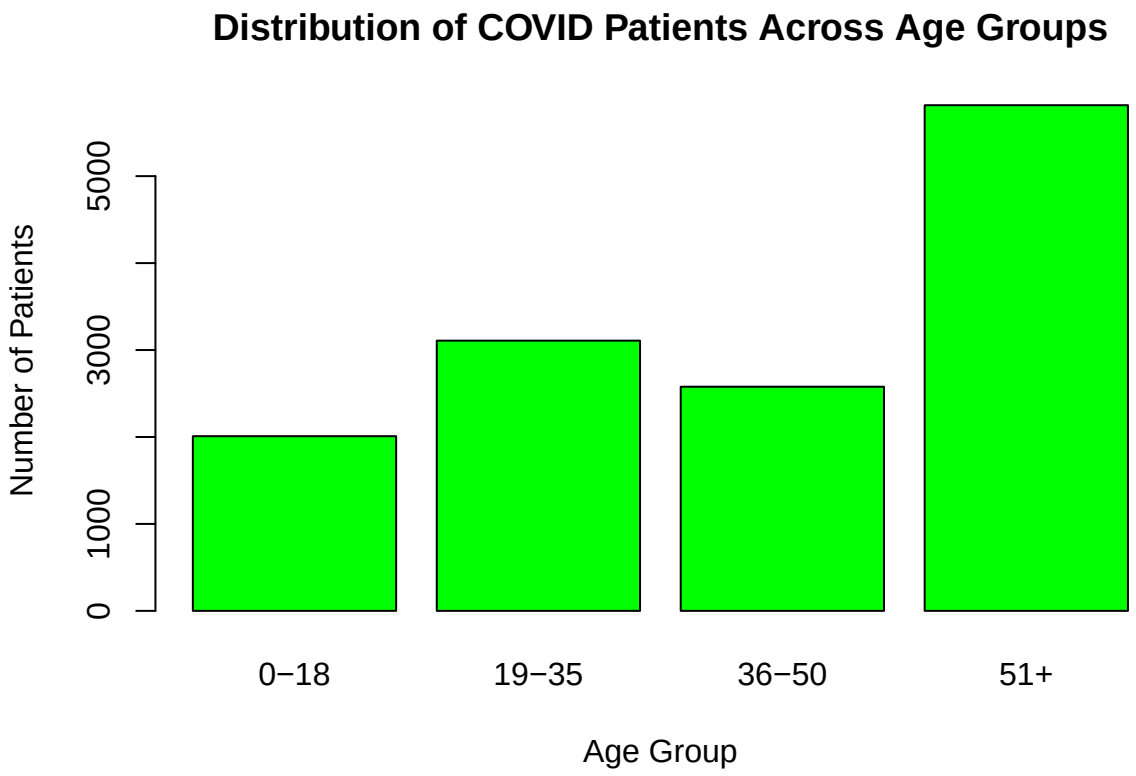

```
covid_patients$Age = as.numeric(format(Sys.Date(), "%Y")) - as.numeric(substr(covid_patients$BIRTHDATE,
```

```
covid_patients$AgeGroup = rep(NA, nrow(covid_patients))
for (i in 1:nrow(covid_patients)) {
  age = covid_patients$Age[i]
  if (age <= 18) {
    covid_patients$AgeGroup[i] <- "0-18"
  } else if (age <= 35) {
    covid_patients$AgeGroup[i] <- "19-35"
  } else if (age <= 50) {
    covid_patients$AgeGroup[i] <- "36-50"
  } else {
    covid_patients$AgeGroup[i] <- "51+"
  }
}
```

```
AgeGroups = table(covid_patients$AgeGroup)
AgeGroups
```

```
##
##  0-18 19-35 36-50   51+
##  2009  3108  2578  5816
```

```r
barplot(AgeGroups, main="Distribution of COVID Patients Across Age Groups", xlab="Age Group", ylab="Numb
```



Observations:

Distribution of COVID Patients Across Counties:

Middlesex County has the highest number of COVID patients (3148), followed by Essex County (1512) and Suffolk County (1538).
Worcester County also has a significant number of cases (1715).
Smaller counties like Dukes County (40) and Nantucket County (31) have much fewer cases.

Distribution of COVID Patients Across Age Groups:

The highest number of COVID patients is in the 51+ age group (5816).
The age groups 19-35 (3108) and 36-50 (2578) have a moderate number of patients.
The 0-18 age group has the least number of patients (2009).

Explanation:

County Distribution:

The large number of cases in counties like Middlesex, Suffolk, and Worcester may be due to higher population densities, which can facilitate faster virus transmission.
The smaller numbers in counties like Dukes and Nantucket might reflect lower population densities and possibly more effective containment measures or lower exposure rates.

Age Group Distribution:

The higher number of cases in the 51+ age group could be attributed to increased vulnerability and more frequent testing and hospital visits among older adults.

The relatively lower number of cases in the 0-18 age group could be due to lower testing rates, less exposure in school settings (due to closures), or milder symptoms leading to fewer hospital visits.

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.

```
# COVID patients in conditions dataset
covid_conditions = conditions[conditions$DESCRIPTION %in% c("COVID-19", "Suspected COVID-19"), ]

# Get the unique patient IDs of COVID patients
covid_patient_ids = unique(covid_conditions$PATIENT)

# Initialize an empty vector to store additional symptoms
symptoms = c()

# Iterate through each COVID patient to extract their additional symptoms
for (patient_id in covid_patient_ids) {
  patient_symptoms = conditions$DESCRIPTION[covid_conditions$PATIENT == patient_id]
  symptoms = c(symptoms, patient_symptoms[!patient_symptoms %in% c("COVID-19", "Suspected COVID-19")])
}

# Count occurrences of each additional symptom
other_symptoms_counts = table(symptoms)

# Get the top 10 most common additional symptoms
top_10_symptoms = head(sort(other_symptoms_counts, decreasing = TRUE), 10)
top_10_symptoms
```

```
## symptoms
##                          Fever (finding)                          Cough (finding)
##                                     6088                                     4674
## Body mass index 30+ - obesity (finding)                Loss of taste (finding)
##                                     3732                                     3571
##                                Prediabetes                      Anemia (disorder)
##                                     2952                                     2747
##                       Fatigue (finding)                           Hypertension
##                                     2644                                     2371
##                Sputum finding (finding)           Chronic sinusitis (disorder)
##                                     2260                                     1981
```

```
# Create a data frame for the top 10 symptoms
top_10_symptoms_df = data.frame(Symptom = names(top_10_symptoms), Count = as.vector(top_10_symptoms))
top_10_symptoms_df
```

```
##                                 Symptom Count
## 1                        Fever (finding)  6088
## 2                        Cough (finding)  4674
## 3  Body mass index 30+ - obesity (finding)  3732
## 4                 Loss of taste (finding)  3571
## 5                             Prediabetes  2952
## 6                       Anemia (disorder)  2747
## 7                       Fatigue (finding)  2644
## 8                            Hypertension  2371
## 9                Sputum finding (finding)  2260
## 10           Chronic sinusitis (disorder)  1981
```

These results indicate that aside from COVID-19 and suspected COVID-19, fever, cough, and loss of taste are commonly observed symptoms in COVID patients. Furthermore, conditions such as obesity, prediabetes, anemia, fatigue, hypertension, sputum finding, and chronic sinusitis are also frequently present in individuals diagnosed with COVID-19.

```r
# COVID patients in conditions dataset (excluding COVID-19 and Suspected COVID-19)
other_conditions = conditions[!conditions$DESCRIPTION %in% c("COVID-19", "Suspected COVID-19"), ]

# male covid patients
covid_male = covid_patients[covid_patients$GENDER == "M", ]

# female covid patients
covid_female = covid_patients[covid_patients$GENDER == "F", ]


# Initialize empty vectors to store additional symptoms for male and female patients
male_symptoms = c()
female_symptoms = c()

# Iterate through each male COVID patient to extract their additional symptoms
for (patient_id in covid_male$Id) {
  patient_symptoms = other_conditions$DESCRIPTION[other_conditions$PATIENT == patient_id]
  male_symptoms = c(male_symptoms, patient_symptoms)
}

# Iterate through each female COVID patient to extract their additional symptoms
for (patient_id in covid_female$Id) {
  patient_symptoms = other_conditions$DESCRIPTION[other_conditions$PATIENT == patient_id]
  female_symptoms = c(female_symptoms, patient_symptoms)
}

# Count occurrences of each additional symptom for male and female patients
male_symptom_counts = table(male_symptoms)
female_symptom_counts = table(female_symptoms)

# Get the top 10 most common additional symptoms for male and female patients
top_10_male_symptoms = head(sort(male_symptom_counts, decreasing = TRUE), 10)
top_10_female_symptoms = head(sort(female_symptom_counts, decreasing = TRUE), 10)

# Check if there are any top symptoms for male and female patients
if (length(top_10_male_symptoms) > 0) {
  top_10_male_symptoms_df = data.frame(Symptom = names(top_10_male_symptoms), Count = as.vector(top_10_m
```

```
} else {
  top_10_male_symptoms_df = data.frame(Symptom = character(), Count = integer(), Gender = character())
}

if (length(top_10_female_symptoms) > 0) {
  top_10_female_symptoms_df = data.frame(Symptom = names(top_10_female_symptoms), Count = as.vector(top_
} else {
  top_10_female_symptoms_df = data.frame(Symptom = character(), Count = integer(), Gender = character()
}

# Combine data frames for male and female symptoms
top_10_symptoms_df = rbind(top_10_male_symptoms_df, top_10_female_symptoms_df)
top_10_symptoms_df
```

```
##                                    Symptom Count Gender
## 1                          Fever (finding)  5684   Male
## 2                          Cough (finding)  4337   Male
## 3                 Loss of taste (finding)  3389   Male
## 4                        Fatigue (finding)  2504   Male
## 5  Body mass index 30+ - obesity (finding)  2338   Male
## 6                Sputum finding (finding)  2107   Male
## 7                        Anemia (disorder)  2074   Male
## 8                             Prediabetes  1958   Male
## 9                           Hypertension  1646   Male
## 10            Chronic sinusitis (disorder)  1315   Male
## 11                         Fever (finding)  6307 Female
## 12                         Cough (finding)  4874 Female
## 13                Loss of taste (finding)  3630 Female
## 14 Body mass index 30+ - obesity (finding)  2801 Female
## 15                       Fatigue (finding)  2701 Female
## 16          Miscarriage in first trimester  2363 Female
## 17               Sputum finding (finding)  2339 Female
## 18                            Prediabetes  2041 Female
## 19                          Hypertension  1764 Female
## 20                        Normal pregnancy  1636 Female
```

Similarities:

Fever and Cough are the most frequently observed additional symptoms in both male and female COVID-19 patients, with slightly higher incidences in females.
Loss of taste and Fatigue are also common in both genders, ranking high on the list of symptoms.
Body mass index 30+ - obesity and Prediabetes are present in the top 10 symptoms for both genders, highlighting a shared concern regarding obesity and prediabetic conditions among COVID-19 patients.

Differences:

Anemia (disorder) is among the top 10 symptoms for males but not for females.
Miscarriage in first trimester and Normal pregnancy are unique to the female list, emphasizing conditions related to pregnancy.
Chronic sinusitis (disorder) appears among the top symptoms for males but not for females, indicating a possible gender difference in the prevalence of sinus issues.
Sputum finding (finding) and Hypertension are common in both genders but occur with slightly different frequencies.

Conclusion: While many symptoms such as fever, cough, loss of taste, and fatigue are prevalent across both genders, there are notable gender-specific conditions. Females exhibit pregnancy-related symptoms like

miscarriage and normal pregnancy, whereas males show higher occurrences of anemia and chronic sinusitis. These differences highlight the importance of addressing gender-specific healthcare needs in managing COVID-19 patients.

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.

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.