EXECUTIVE SUMMARY REPORT 1

ALY 6010 Probability Theory and Statistics

RAHUL AVINASH JADHAV

Northeastern University

College of Professional Studies, Northeastern University, Boston, MA 02115

Contact: jadhav.ra@northeastern.edu

Submitted to Professor: Prof. Roy Wada

Date of submission: 11/07/2021

Introduction

This assignment demonstrated my capabilities in R-programming, which includes the importing csv file, cleaning the dataset, conversion of dataset and visualization. Also, we have created the data frame from CSV file, frequency tables, summary statistics, and different plots to understand the data. This report helps us to understand the details of dataset. The dataset we used for this assignment was comorbidities (any) which consisted of 96 data and 9 variables.

Analysis

a) frequency count, means, or SD

Table 1: Frequency table of Ethnicity and Race

	African-American/ Black	American Indian/ Alaska Native	Asian	Native Hawaiian/ Pacific Islander	Other	Unknown	White
Hispanic/ Latino	3	3	3	3	3	3	3
Non-Hispanic/ Latino	3	3	3	3	3	3	3
unknown	3	3	3	3	3	3	3

Frequency count helps us to know how many times a specific data is present in our dataset. As you can see from table 1. We can see that in our data set, we have 3 data of Ethnicity Hispanic/Latino of race African-American/Black

Table 2: Mean, Standard Deviation of cleaned Dataset.

	Mean	SD
comorbidity*	1	0
SEX*	2	0.82
Ethnicity*	2.33	1.26
Race*	4.29	2.39
Total_cases	4032.03	12609.08
Total_deaths	180.32	662.19
sex*	2	0.82
ethnicity*	2.33	1.26
race*	4.29	2.39
cases_in_percentage	1.59	4.96
death_in_percentage	1.59	5.83
cases_proportion	0.02	0.05

deaths_proportion	0.02	0.06

Standard deviation (SD) is a widely used measurement of variability used in statistics. It shows how much variation there is from the average (mean). A low SD indicates that the data points tend to be close to the mean, whereas a high SD indicates that the data are spread out over a large range of values. To know if the SD is high or low we can find Coefficient of Variation (CV) which is SD/mean. If CV >= 1 indicates a relatively high variation, while a CV < 1 can be considered low.

Table 3: Total Cases and Total Death in Dataset

Total Cases	Total Death
254018	11360

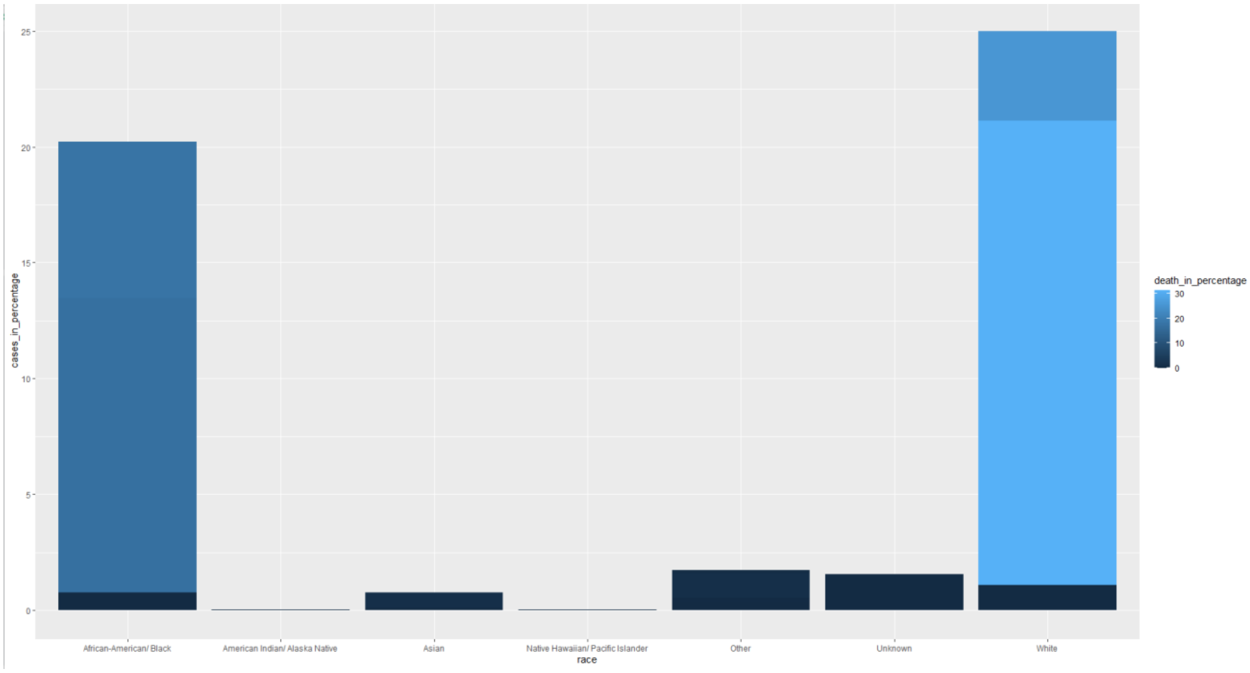
After cleaning the data and removing unnecessary data the above table shows the total cases and total deaths occurred.

Table 4: Total Death where race is African-American/ Black

Race	Total Death
African-American/ Black	3993

After creating a subset from clean dataset which contained data of only African American/ Black it was found that there were 3993 deaths.

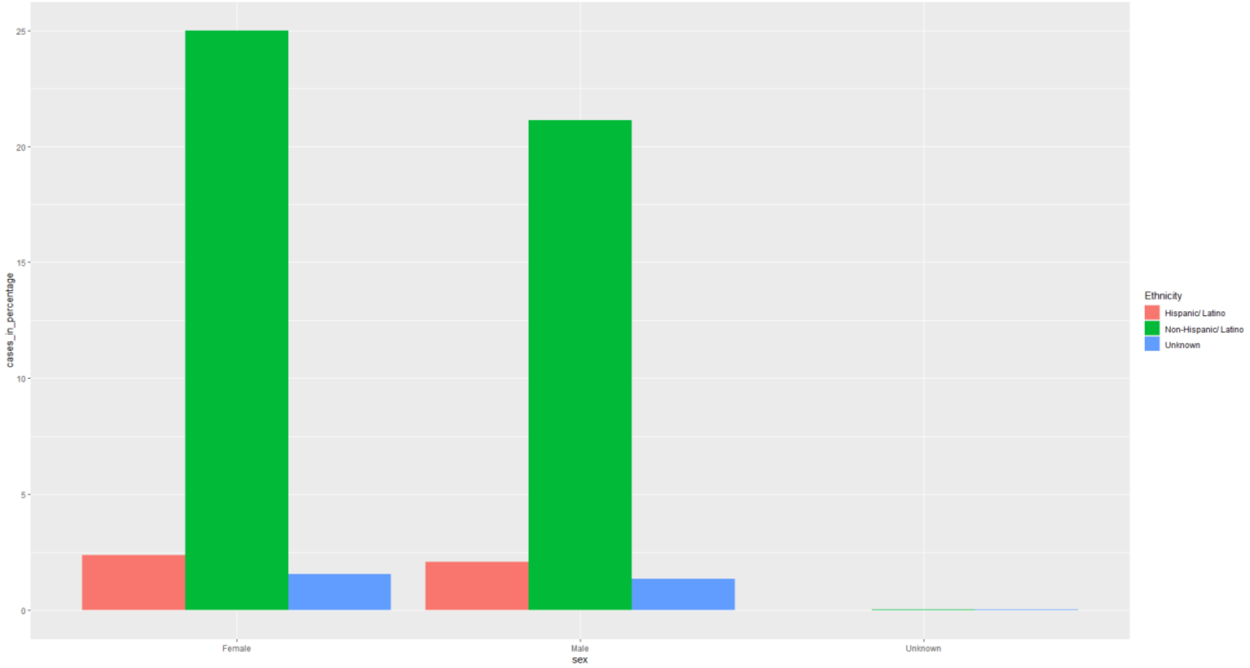
Bar plot 1: Cases and Death as per Race



From bar plot 1 we can see that people of Race white are the highest to be affected by comorbidity. As covid tends to be dangerous for people having comorbidity, we can say that the probability of people of race white being more to be affected and fatality can be more in them.

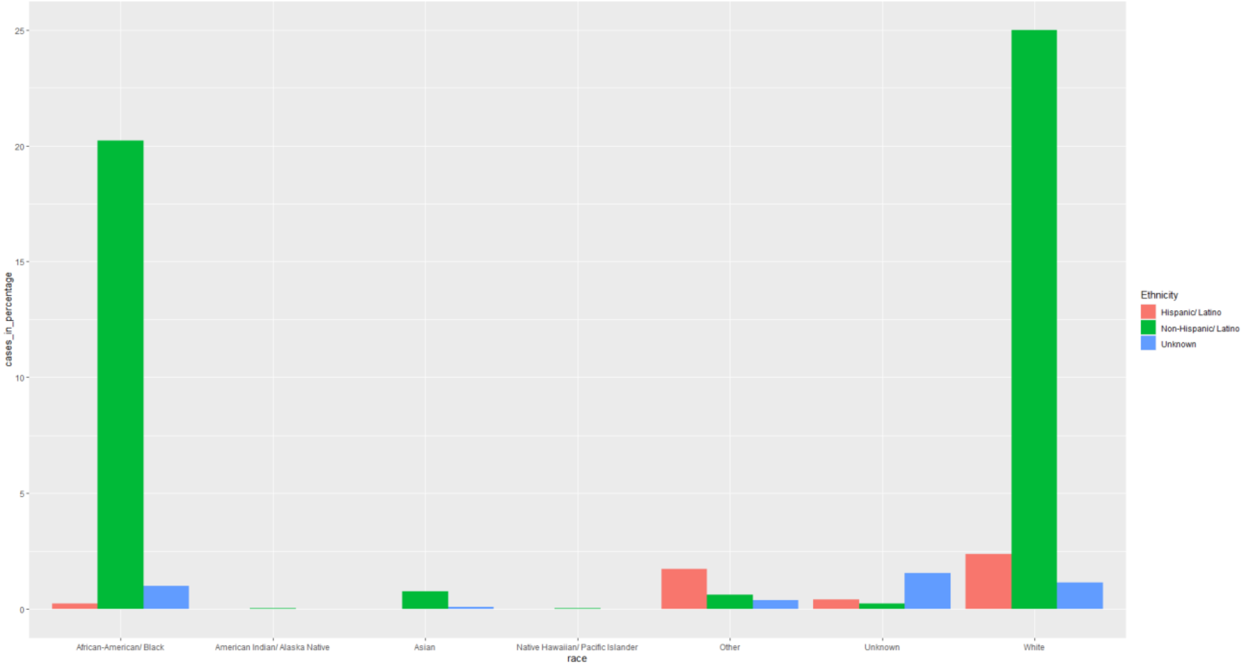
If we had more data like age group, Family Histories, Smoking habits, then it would have help us more to rule out in groups and ages how the covid affects.

Bar Plot 2: Cases in percentage as per Sex and Ethnicity



As you can see from bar plot 2, the number of Female from Non-Hispanic/ Latino are more infected by comorbidity.

Bar Plot 3: Cases in percentage as per Ethnicity and race



From bar plot 3 we can see that Non-Hispanic/ Latino of race White have the highest infected rate where as the people of race American Indian/ Alaska Native and Native Hawaiian/ Pacific islander have the lest infected rate.

b) Summary

- We can see that the Non-Hispanic/Latino people are more affected than the others
- Females are more exposed to the infection
- Total Cases and total deaths are 254018 and 11360 respectively.

Bibliography

- Kabacoff, R. (2011). *R in action: Data analysis and graphics with R*. Manning.
- *Create, modify, and delete columns — mutate*. (n.d.). Dplyr.Tidyverse.Org. https://dplyr.tidyverse.org/reference/mutate.html
- Holtz, Y. (n.d.). *Grouped, stacked and percent stacked barplot in ggplot2*. R-Graph-Gallery.Com. https://www.r-graph-gallery.com/48-grouped-barplot-with-ggplot2.html

Appendix

```
#importing libraries
packages<-(c("dplyr","psych","ggplot2"))
package.check <- lapply(
  packages,
  function(x) {
    if (!require(x, character.only = TRUE)) {
      install.packages(x, dependencies = TRUE)
      library(x, character.only = TRUE)
    }
  }
)

lapply(c("dplyr","psych","ggplot2"),require, character.only = TRUE)




#importing csv using read.csv
getwd()
setwd('C:/Users/ralph/Desktop/studies/Probability/Module 1')
main_dataset <- read.csv('comorbidities (any).csv')
headTail(main_dataset)
class(main_dataset)



#Prepare data.frame for analysis. Arrange variables (rename, drop, or keep variables), clean data (gsub or conditional assignments with ifelse() ), and apply appropriate data structure (i.e. integer, numerical, character, factors, etc.)
main_dataset <- main_dataset%>%
  mutate(
    sex = factor(sex, levels = c('Female','Male','Unknown'), labels = c('Female', 'Male','Unknown')),
    ethnicity = factor(ethnicity, levels = c('Hispanic/ Latino','Non-Hispanic/ Latino','Total','Unknown'), labels = c('Hispanic/ Latino', 'Non-Hispanic/ Latino','Total','Unknown')),
    race = factor(race, levels = c('African-American/ Black','American Indian/ Alaska Native','Asian','Native Hawaiian/ Pacific Islander','Other','Total','Unknown','White'), labels = c('African-American/ Black', 'American Indian/ Alaska Native','Asian','Native Hawaiian/ Pacific Islander','Other','Total','Unknown','White')),
    )


main_dataset <- rename(main_dataset, SEX=sex,Ethnicity= ethnicity, Race= race, Total_cases = cases, Total_deaths = deaths)

clean_dataset <- subset(main_dataset,Ethnicity != 'Total' & Race != 'Total')

headTail(clean_dataset)

attach(clean_dataset)




#Create new variables or new observations (such as proportion or total count) to aid in your analysis. Do you need to create new columns or new rows? You may want to use dplyr or tidyr commands to aggregate by groups.

clean_dataset <- clean_dataset %>%
  mutate(
    cases_in_percentage = round(Total_cases/sum(Total_cases)*100,4),
    death_in_percentage = round(Total_deaths/sum(Total_deaths)*100,4),
    cases_proportion = (proportions(Total_cases)),
    deaths_proportion = (proportions(Total_deaths))
  )

head(clean_dataset)
```

```
mytable <- function(df,outname) {
  t0 <- psych::describe(df)
  t0 <- t0 %>% select(mean, sd) %>%
    mutate(mean=round(mean,2)) %>%
    mutate(sd=round(sd,2)) %>%
    rename(Mean=mean, SD=sd)
  write.csv(t0, outname)
}
mytable(clean_dataset,"module1.csv")


table1 <- table(clean_dataset$ethnicity,clean_dataset$race)
table1
write.table(table1, file = "olstab.txt", sep = ",", quote = FALSE, row.names = F)

#data analysis and plotting

hist(Total_cases,Total_deaths)

ggplot(clean_dataset, aes(fill=death_in_percentage, y=cases_in_percentage, x=race)) +
  geom_bar(position="dodge", stat="identity")



sum(clean_dataset$Total_cases)
sum(clean_dataset$Total_deaths)

t1 <- subset(clean_dataset,race=="African-American/ Black")
sum(t1$Total_deaths)
```

```
ggplot(clean_dataset, aes(fill=Ethnicity, y=cases_in_percentage, x=sex)) +
  geom_bar(position="dodge", stat="identity",)

ggplot(clean_dataset, aes(fill=Ethnicity, y=cases_in_percentage, x=race)) +
  geom_bar(position="dodge", stat="identity",)
```