Module - 2

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: 11/14/2021

Introduction

In this assignment a dataset which consist of individual death due to covid-19 from Georgia (USA). The dataset consists of variables Age, Ethnicity, Race, Sex, County, and Chronic_Condition. For getting more insight from the data we will create a category of age as Children, youth, Adult and Senior which can help us find which age group are more susceptible to covid-19 virus.

From the given dataset it is hard to draw any meaningful conclusions, hence cleaning the dataset is necessary. For cleaning this dataset I have used the mutate() function etc.

Analysis

Table 1: Summary table of Death.csv dataset

	count	mean	sd
age	21709	72	13.91693
ethnicity*	21709	2	0.265862
race*	21709	2.9	1.407813
sex*	21709	1.5	0.499404
county*	21709	71.5	41.90576
chronic_condition*	21709	1.6	0.671959
Age_Group*	21709	2.5	0.886886

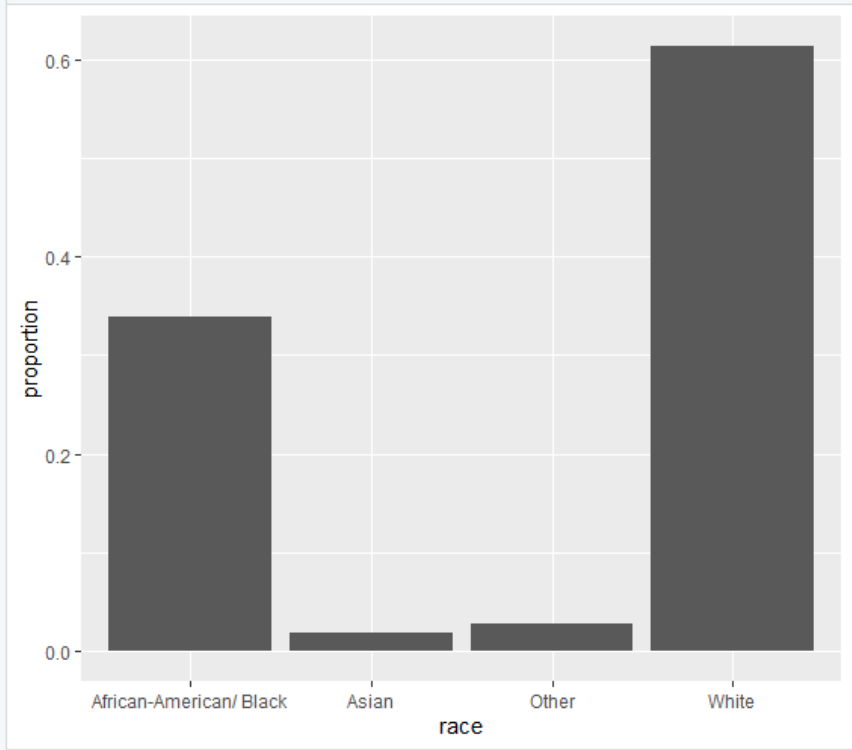
From above we can know the mean and standard deviation (SD) of the variables. Sd measurement is used to know variability in the dataset. It shows how much variation there is from average mean. A low SD means data elements are closer to the mean whereas high SD means dataset are farer from the mean. To determine if the SD is more or less, we need to calculate Coefficient Variation (CV) that is SD/mean. IF CV>=1 that means Sd is more and CV<=1 means SD is less. Mean is basically sum of all the elements divided by count of elements.

Table 2: TOP 5 country with high deaths due to covid-19

	county	count	mean(age)	sd(age)	proportion	percentage
1	Fulton	1490	73.3745	13.97078	0.069	6.863513
2	Gwinnett	1234	71.26985	14.86744	0.057	5.684278
3	Cobb	1149	73.63185	13.82964	0.053	5.292736
4	DeKalb	1087	71.92548	14.4468	0.05	5.00714
5	Non-GA Resident/Unknown State	609	70.66338	13.27889	0.028	2.805288

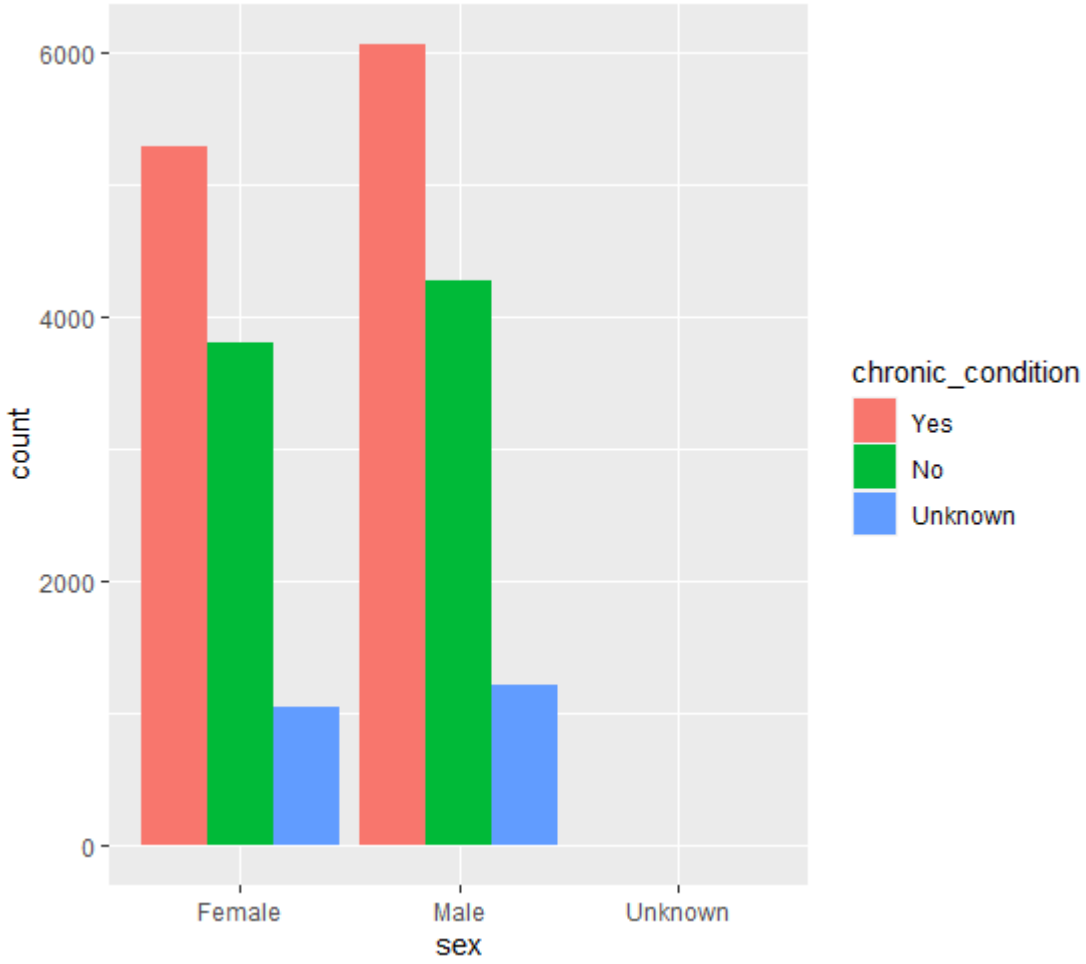
From the above table we can see that County Fulton have highest death due to covid19 that is 1490 deaths that is 6.86% of overall death followed by Gwinnett and cobb. Which consist of 5.68% and 5.29% of total deaths respectively.

Bar plot 1: Proportion of death as per race



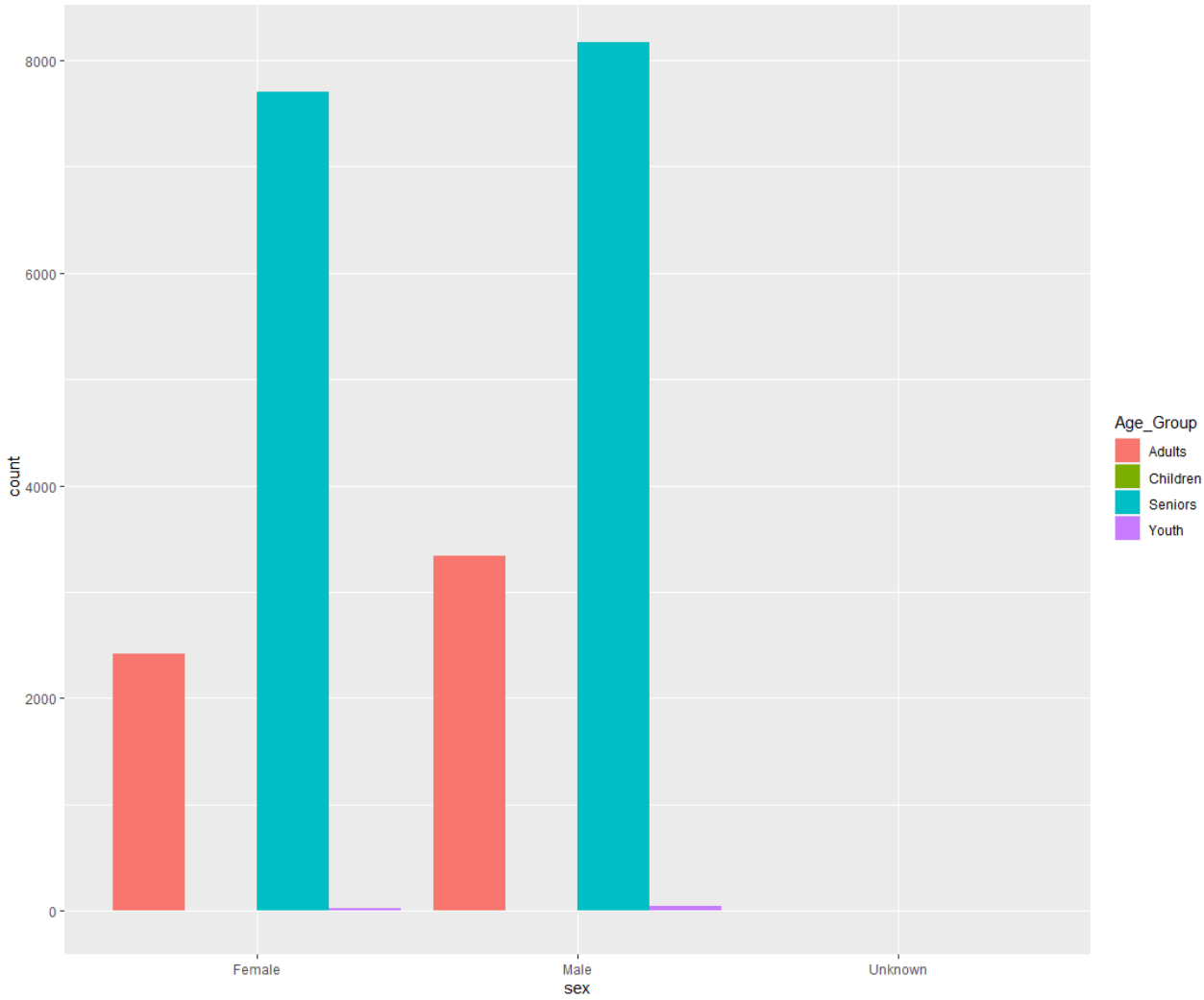
From bar plot 1 we can see proportion of death of people of Race white is highest followed by proportion of death of people of race African-American/black.

Bar Plot 2: Gender wise death of people by chronic condition



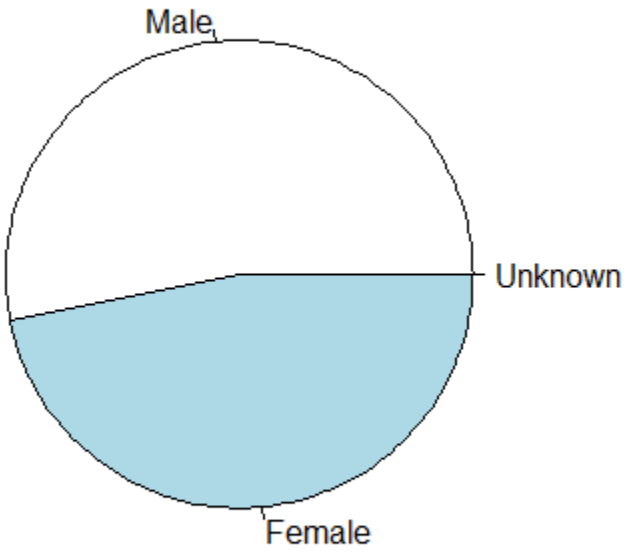
As you can see from bar plot 2, death of Male is more due to chronic condition

Bar Plot 3: Gender wise death of people as per Age group



From bar plot 3 we can see that people of age group senior(people of age above 65) have more deaths in both the sex. We can conclude that older people are more susceptible to covid_19.

PIE CHART 1 : Deaths among Sex



From the above pie chart, we can see that death of male due to covid19 is more.

Summary

- We can see that deaths of people of race white are more
- People of age above 65 are more susceptible to covid-19
- County Fulton has more deathrate
- Death of male is more due to covid19

Bibliography

- Kabacoff, R. (2011). *R in action: Data analysis and graphics with R*. Manning.
- *Age Categories, Life Cycle Groupings*. (n.d.). Statcan. https://www.statcan.gc.ca/en/concepts/definitions/age2
- *Multiple condition if-else using dplyr, custom function, or purr*. (2018, August 26). Stack Overflow. https://stackoverflow.com/questions/52028764/multiple-condition-if-else-using-dplyr-custom-function-or-purr