Prof. Roy Wada

R Practice – Module 1

ALY6010 – Probability Theory and Introductory Statistics

Abhigna Ramamurthy, NUID: 002982276

Ramamurthy.a@northeastern.edu

**Executive Summary**

The below report is a quick analysis on comorbidity data set. It has details on using R studio to clean and analyze the data to answer some interesting questions such as what proportion of a particular race has in comparison to the total number of covid deaths. The report also explores some basic data visualization tools used in R. Some recommendation on further data required to enhance the analysis has also been discussed.

**Introduction**

The data provided is a comorbidity dataset with sex, ethnicity, race, number of covid cases and number of deaths. This is a sample dataset to perform and understand basic R commands.

**Analysis**

Step 1: set working directory and read csv into R studio

Text

Description automatically generated

Views in R studio -   


Text

Description automatically generated with low confidence

Shape

Description automatically generated with medium confidence

Text, letter

Description automatically generated

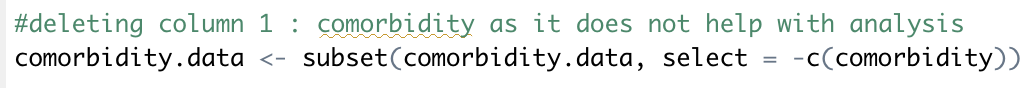
Step 2: Created a data frame for easy analysis

Text

Description automatically generated

Step 3: Cleaning of data

3.1: Deleting column 1 Comorbidity as it does not provide any analytical value



3.2: Renaming columns for easy understanding

Graphical user interface, text, application

Description automatically generated

3.3: Converting char data into factor for easy group by analysis

Columns changed include sex, ethnicity and race

Text

Description automatically generated

3.4: Checking the structure of the comorbidity data frame:

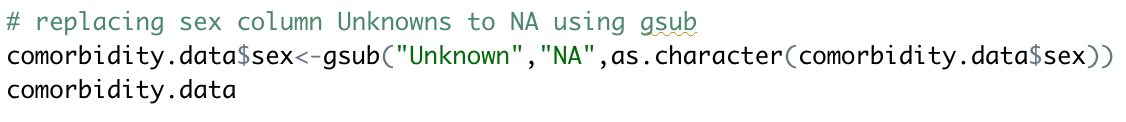
Text

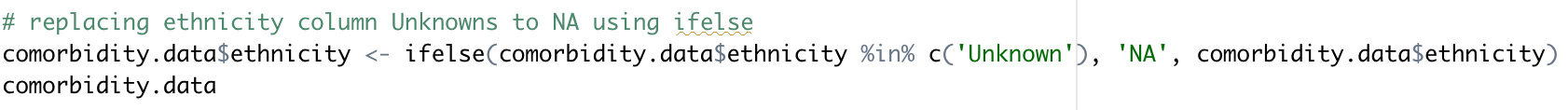
Description automatically generated

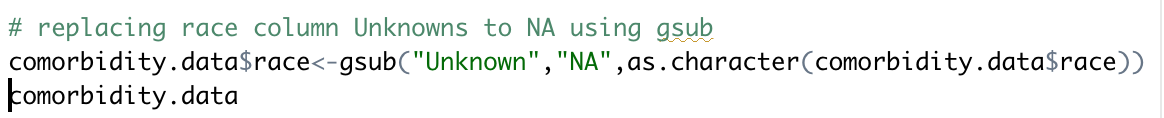
Text, letter

Description automatically generated

3.5: Using gsub and ifelse to replace Unknown values with NA in sex, ethnicity and race columns







Text, letter

Description automatically generated

3.6: Removing rows with “Total” as value in ethnicity or race columns as it shows the total of cases and deaths for that category. Keeping these rows would double up the calculations. Hence removing the rows.

Text

Description automatically generated

3.7: Using dplyr to add new columns

3.7.1: no\_of\_recovered cases   
Text

Description automatically generated

3.7.2: Adding mortality\_rate which shows the probability of death for that category of values.

Also formatted the mortality\_rate to showcase 2 decimal places only for easy analysis.

Graphical user interface, text, application, email

Description automatically generated

Step 4: Analysis

4.1: Total number of cases and total number of deaths.

Graphical user interface, text

Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

4.2: Analysis based on sex

Text

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated with medium confidence

A picture containing table

Description automatically generated

4.3: Answering question – “What proportion of deaths are among black population?”

Used group by to select based on Black race. Used filter to only select columns for Black race. Got the total sum of cases and deaths. Proportion of black population death is number of deaths in black population to total number of deaths seen across all races, which comes upto 35.14%.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

4.4: Answering: “How would you examine the possible relationship between race/ethnicity and the probability of death given that someone has been diagnosed with COVID-19?”

To answer this question, all race and ethnicity analysis was performed similar to 4.3 above and the results are combined into a data frame called race\_df and ethnicity\_df.

**Race Analysis:**

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Creating data frame race\_df

Graphical user interface, text

Description automatically generated with medium confidence

Table

Description automatically generated

The above table shows the probability of death based on the race available from the data set.

Summary of the race analysis is as follows:

Table

Description automatically generated

The analysis shows that probability of African-American/Black and White races have higher probability of death in comparison to other races.

**Ethnicity Analysis:**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text, letter

Description automatically generated

Above table shows probability of death rounded to 2 decimal places.

Summary of ethnicity analysis is as follows:

Table

Description automatically generated

The analysis shows that Non-Hispanic/Latino ethnicity has higher probability of death in comparison with other ethnicities.

4.4: Answering: “What other information or data set would you need to merge or join to enhance/complete your analysis?”

There are number of ways we can enhance the analysis:

1. If we get the exact chronic diseases suffered by the patients, we can dig deeper into exact diseases. We can also get the level of underlying diseases and determine if covid aggravated that and can find the probability of any underlying diseases, that can be prone to fatality.
2. Understanding geographical data would provide more insight as to how other aspects like availability to accessible heath care would have affected the fatality.

Step 5: Prepare and export at least one table (frequency tables, cross-tabulations or summary statistics), and at least one chart

Tables exported are as follows:

Table

Description automatically generated

Race table showing different races, number of cases, deaths and probability of death.

Table

Description automatically generated

Ethnicity table showing various ethnicities, number of cases, deaths and probability of death.

Charts:

Histogram of number of cases in the whole dataset which includes sex, ethnicity and race differentiating factors:

Graphical user interface, text

Description automatically generated

Timeline

Description automatically generated

The histogram shows that most of the categories had cases between 0-10000.

Bar chart on number of cases based on ethnicity:

Chart, bar chart

Description automatically generated

The graph shows that Non-Hispanic/Latino had the highest number of covid cases.

Bar chart for number of deaths based on ethnicity:

Chart, waterfall chart

Description automatically generated

The bar chart shows that the number of deaths from unknown ethnicities are higher.

**Conclusion**:

The assignment was a great introduction to decode how an actual manager would give a task. Every step had a certain degree of creativity to it. Learnt the basics of data cleaning very clearly. It is very important to have a pragmatic approach to data cleaning is what I understood from this assignment. Analysis is also highly driven by the quality of questions we ask. It is interesting to see how the probability of death is higher for African American/Black population in comparison with other races. Also, Non-Hispanic/Latino community with underlying chronic diseases will have more probability of death in comparison with other ethnicities. As discussed, more data with respect to the underlying chronic diseases could help greatly in triaging for primary care.