Abhigna Ramamurthy

NUID: 002982276

R Practice Module 2

ALY6010: Probability and introduction to statistics

October 3, 2021

Table of Contents

[Introduction 2](#_Toc84152872)

[Analyzing the deaths data 2](#_Toc84152873)

[Data cleaning 2](#_Toc84152874)

[Analysis 3](#_Toc84152875)

[Questions 7](#_Toc84152876)

[Conclusion 10](#_Toc84152877)

[References 10](#_Toc84152878)

# Introduction

The data under analysis in this report is deaths.csv. It has six columns with age, ethnicity, race, sex, country and chronic\_condition indicator columns. The report explains the data cleaning method employed, the descriptive analysis made, and some inferences based on that.

# Analyzing the deaths data

## Data cleaning

1. Import the file into R using read.csv( ) function and 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.)

Initial structure of the whole data without any filtration

Text, letter

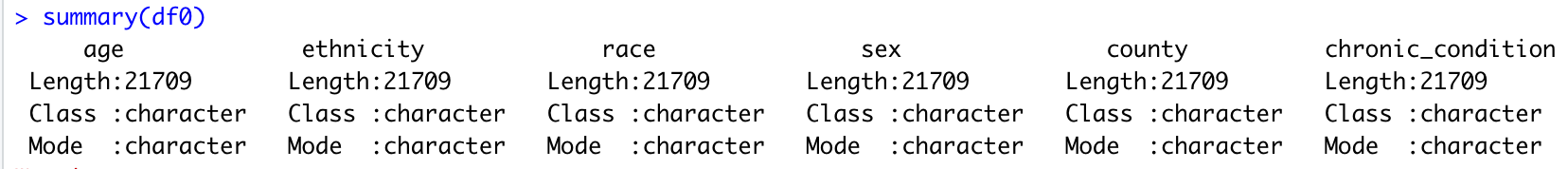
Description automatically generated

Columns of original data



Column names can be modified to reflect the data it represents in a clean format. All the columns have data that can be used for analysis. Hence keeping all the data intact.

Summary of original data



Age: Can see that age is character due to the value “90+”. It can be converted to 90 and data type to integer for analysis.

Ethnicity, Race and Sex are characters. Can be converted to factor for easier analysis by group. Unknowns on these columns can be replaced by NA.

Country is character and looks fine. It has some unknowns, which can be replaced by NA.

Chronic\_condition has Yes,No and Unknown values. Unknowns can be converted to NA.

An indicator variable based on White race or not can be created as White population has the highest number of cases.

Final structure of the data under analysis is as follows:

Text

Description automatically generated

## Analysis

1. 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?
2. You may want to use dplyr or tidyr commands to aggregate by groups.
3. You may need to create indicator variables (white indicator, black indicator, etc.)

Apart from the age variable, there is not much quantitative data to analyze. Hence new columns other than indicator columns are not necessary.

Plots created for the data analysis:

Chart, bar chart

Description automatically generated

The above plot shows the number of cases given in the dataset based on the age groups. We can see that age group of 60-89 has highest value. This means people under the age group of 61-89 are most affected (How to find the unique values in a column of an R data frame?, 2021), (length and Nuno, 2021), (ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA, 2021).

Scatter plot of the same data is below (used for comparison in the report in the section 4.4.2) (ggplot2 scatter plots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA, 2021).

Chart

Description automatically generated

Chart

Description automatically generated

This shows that the highest number of cases were reported among the White and African American/Black races (ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA, 2021).

Chart, bar chart

Description automatically generated

A grouped bar chart above shows the comparison between the total number of cases and number of cases with chronic condition grouped by race. We can conclude that almost half the total number of cases in each race is diagnosed with underlying chronic condition (ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA, 2021).

Chart, bar chart

Description automatically generated

The above grouped chart shows the number of cases categorized by sex(male/female/unknown) grouped by ethnicity. We can see that the total number of unknown gender values under each ethnicity is very less. Hence, they can be omitted from the analysis. The graph also shows that Non-Hispanic/Latino has a high number of cases compared to Hispanic/Latino ethnicity (ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA, 2021).

Chart, bar chart

Description automatically generated

The above plot shows the number of cases by countries. Only top 5 number of cases by countries are shown. We can see that Fulton has the highest number of cases of 1490 in the given dataset (ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA, 2021), (Counting the frequency of unique values - R, 2021).

Chart, histogram

Description automatically generated

The above histogram shows the distribution of age. We can see that the numbers peak after 60+ age.

## Questions

* 1. How are the data set organized differently compared to the last week?

Last week had aggregated data based on sex, ethnicity and race combinations, whereas data from this week is individual data. This can be known by looking at the amount of data in the dataset and age columns.

* 1. What proportion of deaths are among Black or Hispanic population?

38.82% of total population deaths are Black or Hispanic.

* 1. Prepare and export a descriptive summary table (for the total sample, and by group)
     1. How do you report count, proportion, means, SD, and CI in the same table?  
          
        There are many descriptive statistics libraries available in R that can be used to report all the above in one table. One such library used is psych.
     2. What do you get when you take the mean of an indicator variable?  
          
        Mean of the indicator variable represents the actual mean of number of data rows that satisfy the indicator condition. In my analysis case, White\_Indicator simply indicates rows which qualify for race as “White”. Hence the mean of this indicator shows the mean of total number of “White” race data in the dataset.
     3. By which group would you report them?

In the current dataset, there are number of ways we can group them based on the most important data we are looking to extract. I believe chronic\_condition here has more value for further analysis when combined with more data. Hence summary table is grouped by chronic\_condition value.

* 1. Make a bar chart of distribution (i.e., histogram), and a grouped bar chart  
       
     Figure 4,5 and 7 plots can be considered for this section.   
     1. When do you use a grouped bar chart?  
          
        A grouped bar chart can be used to compare count/frequencies of two or more categorical data (A Complete Guide to Grouped Bar Charts, 2021).
     2. Why is grouped bar chart preferred over a scatter chart?  
          
        When dealing with a large dataset, scatter plots become very cumbersome. Hence a grouped bar chart helps in analyzing data in a clear and methodical way (HOW TO CHOOSE WHICH TYPE OF GRAPH TO USE?, 2021).
  2. Prepare and export at least one table (summary statistics), and at least two charts (clean and with color, with appropriate titles and axis labels, etc.). You may want to look into psych::describe, hist( ), ggplot2, and plotly, among others.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **vars** | **columns** | **n** | **mean** | **sd** | **median** | **trimmed** | **mad** | **min** | **max** | **range** | **skew** | **kurtosis** | **se** |
| 1 | Age | 21709 | 71.99963149 | 13.91693023 | 74 | 73.31026542 | 14.826 | 1 | 90 | 89 | -0.84903627 | 0.640929312 | 0.094454693 |
| 2 | Ethnicity | 21709 | 2.893177945 | 0.443680918 | 3 | 3 | 0 | 1 | 3 | 2 | -3.970916814 | 13.90410667 | 0.003011278 |
| 3 | Race | 21709 | 4.854346124 | 2.822781725 | 7 | 5.067821982 | 0 | 1 | 7 | 6 | -0.597561176 | -1.611769196 | 0.019158318 |
| 4 | Sex | 21709 | 1.532682298 | 0.499403639 | 2 | 1.540560769 | 0 | 1 | 3 | 2 | -0.125452654 | -1.969574928 | 0.00338947 |
| 5 | Country | 21709 | 71.45492653 | 41.905758 | 64 | 69.22609246 | 45.9606 | 1 | 161 | 160 | 0.437087906 | -0.841204308 | 0.284415847 |
| 6 | Chronic\_Condition | 21709 | 2.419273113 | 0.671958548 | 3 | 2.524037078 | 0 | 1 | 3 | 2 | -0.732826827 | -0.581556999 | 0.004560606 |
| 7 | White\_Indicator | 21709 | 0.614491685 | 0.486726378 | 1 | 0.643099775 | 0 | 0 | 1 | 1 | -0.470434383 | -1.778773422 | 0.003303429 |

(describe function - RDocumentation, 2021), (Kabacoff, 2015)

# Conclusion

The given dataset has helped to analyze data categorically. This is especially helpful in drawing inferences based on the categories. Another important aspect least is with respect to data visualization which is the focus of this report. Showing data in a readable/understandable format is of utmost importance.

# References

2021. HOW TO CHOOSE WHICH TYPE OF GRAPH TO USE?. [online] Available at: <https://nces.ed.gov/nceskids/help/user\_guide/graph/whentouse.asp> [Accessed 3 October 2021].

Chartio. 2021. A Complete Guide to Grouped Bar Charts. [online] Available at: <https://chartio.com/learn/charts/grouped-bar-chart-complete-guide/#:~:text=A%20grouped%20bar%20chart%20is,across%20levels%20of%20the%20second.> [Accessed 3 October 2021].

Edureka Community. 2021. Counting the frequency of unique values - R. [online] Available at: <https://www.edureka.co/community/4500/counting-the-frequency-of-unique-values-r> [Accessed 3 October 2021].

Kabacoff, R., 2015. R in action. Shelter Island, NY: Manning.

length, &. and Nuno, C., 2021. 'row.names' is not a character vector of length. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/48845309/row-names-is-not-a-character-vector-of-length> [Accessed 3 October 2021].

Rdocumentation.org. 2021. describe function - RDocumentation. [online] Available at: <https://www.rdocumentation.org/packages/psych/versions/1.0-17/topics/describe> [Accessed 3 October 2021].

Sthda.com. 2021. ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. [online] Available at: <http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization> [Accessed 3 October 2021].

Sthda.com. 2021. ggplot2 scatter plots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. [online] Available at: <http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization> [Accessed 3 October 2021].

Tutorialspoint.com. 2021. How to find the unique values in a column of an R data frame?. [online] Available at: <https://www.tutorialspoint.com/how-to-find-the-unique-values-in-a-column-of-an-r-data-frame> [Accessed 3 October 2021].