Name: Frenel Cherubin  
Date: 7/28/2019  
Title: DSC 520 Statistics for Data Science Final Project

Section 1 – Week 8 – Assignment 8.1 Getting Started

• Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?

Causes of deaths have always been at the center of research worldwide. In the United States for instance, around 74% of all deaths originate as a result of either Heart disease, Cancer, Stroke and cerebrovascular diseases, and in some instance neonatal deaths. Statistical data collected have proven that for more than a decade, heart disease and cancer have claimed the first and second spots respectively as the leading causes of deaths in America. Together, the two causes are responsible for 46 percent of deaths in the United States. Third, comes chronic lower respiratory diseases. Together, they account for half deaths in the United States. Consequently, the Centers for Disease Control and Prevention (CDC) has been collecting and examining causes of death for over 30 years. This information helps researchers and doctors to take a better perspective when it comes to address growing epidemics in healthcare. Further, the data contribute to help understand how preventive measures may help people live longer and healthier lives.

• Draft 5-10 Research questions that focus on the problem statement.

What is the number 1 cause of death?

What is the leading cause of death in the US in 2018?

How can we prevent leading causes of death?

What disease has killed the most people?

Can death in connection with those diseases be prevented?

How do the diseases that lead to death originate?

Could a better diet prevent a person from getting those life-threatening diseases?

What kind of lifestyle do the victims adopt?

In the cause of neonatal deaths, did the mothers smoke/drink during pregnancy?

• Provide a concise explanation of how you plan to address this problem statement.

Causes of deaths is a serious subject that should not be taking for granted. I want to learn about the top causes of deaths not only for 1 year, but also for a couple of decades back, so I can compare how many people have died from 3 decades ago to see how research have contributed in improving growing epidemics in healthcare, so people can live healthier lives and longer.

• Discuss how your proposed approach will address (fully or partially) this problem.

My proposed approach will address this problem by chronologically analyzing the data and determining how many humans have died from leading causes of deaths, address the issue, and find solutions to help prevent growing epidemics in health care that contribute to humankind deaths and help them to live healthier and longer.

• Do some digging on a dataset that you can use to address the issue.

Original source where the data was obtained is cited and, if possible, hyperlinked.

Data

* Death rates and life expectancy at birth
* Top Five Leading Causes of Death: United States, 1990, 1950, 2000
* Infant and neonatal mortality rates: United States, 1915-2013

– I collected the three data sets from the following websites through data.gov

<https://catalog.data.gov/dataset/age-adjusted-death-rates-and-life-expectancy-at-birth-all-races-both-sexes-united-sta-1900>

<https://catalog.data.gov/dataset/top-five-leading-causes-of-death-united-states-1990-1950-2000>

<https://catalog.data.gov/dataset/infant-and-neonatal-mortality-rates-united-states-1915-2013>

I was a little concerned that I would find some datasets where I will have to do a lot of cleaning. After looking at a lot of datasets that looked very messy and complicated to work with, I collected the 3 datasets I referenced earlier. They appear to be very simple. My intent is to investigate further on the health issues that we face.

Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

All 3 of my datasets, Death rates and life expectancy at birth, Top Five Leading Causes of Death: United States, 1990, 1950, 2000, and Infant and neonatal mortality rates: United States, 1915-2013 were collected by the Centers for Disease Control and Prevention with the purpose to study death rates in America, and they are maintained by the National Center for Health Statistics (NCHS).

Death rates and life expectancy at birth was created on Metadata on March 11, 2016 and was updated February 28, 2019. It contains 5 variables.

Top Five Leading Causes of Death: United States, 1990, 1950, 2000 was created on Metadata on March 11, 2016 and was updated August 20, 2018. It contains 3 variables.

Infant and neonatal mortality rates: United States, 1915-2013 was created on Metadata on March 11, 2016 and was updated on February 28, 2019. It contains 3 variables.

At this point, the numbers of variables in each of the original datasets are mentioned above, and there are no clear indications that detail how missing values are recorded in them.

• Identify the packages that are needed for your project.

To begin, I am setting the data files as my working directory, and those data files will be imported into RStudio by clicking on Import Dataset, from Text(base), then Import. The Dataset I choose are all in Excel Comma Separated Value format. Once imported, I will use a variety of packages for the Rmarkdown file. I will call on ggplot2, so I can attempt to be more ambitious with my graphs. I will also use car, Broom, Quantpsyc and Class, so that I can try a variety of methods to evaluate the data once imported.

• What types of plots and tables will help you to illustrate the ﬁndings to your research questions?

To illustrate the finding of my research, I will begin with some histograms in order to determine the kinds of statistics I am dealing with, then I will try out some scatter plots to show the relationships that exist between either Death rates and life expectancy at birth, Top Five Leading Causes of Death: United States, 1990, 1950, 2000, certain diseases and death, or Infant and neonatal mortality rates depending on the dataset I work with to complete the rest of the project. I will probably try some that break it down by year over the course of the last couple of decades or so.

• Questions for future steps.

What kind of relationship that exist between the variables in the datasets?

If I choose the Top Five Leading Causes of Death: United States, 1990, 1950, 2000 to complete steps 2, and 3 of the projects, how do the finding show smoking for instance is among the top five leading causes of deaths?

How do I make sure the datasets are clean?

• What do you not know how to do right now that you need to learn to answer your research questions?

In order to answer my research questions, I will have to be proficient in making scatter plots where I can adjust the bars to show changes. This challenged me in past assignments, and I hope I can overcome the difficulties I faced with it.