Mini Project 2 Exercise 1 - Analysis & Visualization

David Strube
dstrube3@gatech.edu

Abstract—Having just started my second week at this new start-up company called "HealthMatters", this is an analysis and visualization for a great opportunity - a $750 Million contract with the National Center for Health Statistics (NCHS), CDC. As requested, although I am a developer, I have analyzed the data via Excel only. The NCHS team provided us with this data in the PDF form (CDC, 2014). In order to secure the bid, an analysis of the first 10 pages of data have been made understandable by the general public.

1 RAW DATA

1.1 Describe in detail how you put the data into Excel and how you cleaned it. What approaches did you take and why?

I started by copying the data from the source PDF and pasting it into a text file and then cleaning up the data manually there before cutting and pasting the data from the text file to Excel. I first copied the first two columns (Rank and Cause of death...) as it was not possible to select them separately. This led to the tedious task of removing the ellipses and numbers 1-15 from each row so that the rank and causes could be put into their own columns. Next I copied the next two columns of data (Number and Percent of total deaths), and pasted them into a text file. The values were alternating when pasted so I had to go through the task of cutting them one at a time into Excel into their respective columns. Finally I copied the last column (Rate) into a text file. This time, the problem was that all the values were on one line, and so I had to again cut and paste them one at a time, or enter a new line after each value before cutting and pasting them all into Excel.

This all seemed very inefficient, and so I got to work on writing a Python script (`ihiDataParser_MP2E1.py`) that would take in a temporary text file (*`fileIn.txt`*), and based on what data was at the beginning of the text file, parse through the data, and output to another temporary text file (*`fileOut.txt`*). I would repeat this process for 1 or 2 or 3 columns at a time on each subsequent page. Along the way, I found some complications: whereas I would copy only the two columns Number and Percent of total deaths for each page, on page 5 it was only possible to copy three columns - Number and Percent of total deaths and Rate - and so I had to write a new function to parse all three columns. I found this to be a little faster than copying just the two columns like I did on previous pages, but then on page 6, I found that this new function didn't work properly when copying all three columns. So I had to go back to copying just the two columns and then just the last one.

1.2 What Excel Functions did you use and why?

I used no Excel functions in the **Raw Data** sheet because my custom Python parser functions were better tailored to this task. However I did use functions in other sheets, which will be described later on.

2 OBSERVATION

2.1 What "story" does the data tell?

The data tells the story of what are the top 15 causes of death across all sexes, ages, and races, and then drills down into sex, specific age ranges, and races.

2.2 What data types are represented and why?

The data is represented by classifications of sex, age ranges, and races. For each classification, a list of top 15 causes of death is presented, as well as a summary of all causes and the residual causes not covered in the top 15. For each item in those lists and summaries and residuals, a set of numbers is given: the total number of deaths, percent of total deaths, and rate. These sets of numbers and lists of causes are important to determine what are the most significant causes of death for each classification as well as for humanity at large.

2.3 What are the most important characteristics of the data? Why?

The most important characteristics are Number, Percent of total deaths, and Rate. They are the most important because these characteristics can be sorted either ascending or descending to give a glimpse of what are the least or most (respectively) causes of death for each sex, age group, and race, and for all people.

2.4 What are some interesting aspects or features of the data? Why are they interesting?

When sorting the Numbers descending, *Diseases of heart (I00-I09,I11,I13,I20-I51)* and *Malignant neoplasms (C00-C97)* stood out as the top 2 sources of death among: (a) All races, both sexes, all ages, and (b) All races, male, all ages (highlighted in light red). *Diseases of heart* were the #1 killer among All races, both sexes, 85 years and over. *Malignant neoplasms* were the number one killer among both: (a) All races, both sexes, 65-74 years and (b) All races, both sexes, 75-84 years.

In terms of Percent of total deaths, *Accidents (unintentional injuries) (V01-X59,Y85-Y86)* and *Malignant neoplasms (C00-C97)* were the top 2 cause of death among these groups, ranging from 41% to 30% (highlighted in light green):

All races, both sexes, 1-4 years

All races, both sexes, 15-24 years

All races, both sexes, 25-34 years

All races, both sexes, 55-64 years

All races, both sexes, 65-74 years

All races, male, 1-4 years

All races, male, 5-14 years

All races, male, 15-24 years

All races, male, 25-34 years

All races, male, 55-64 years

All races, male, 65-74 years

When sorting descending by Rate, again we see *Diseases of heart (I00-I09,I11,I13,I20-I51)* and *Malignant neoplasms (C00-C97)* at the top, with rates ranging from 4,013.9 (All races, both sexes, 85 years and over) to 1,095.1 (All races, both sexes, 75-84 years) (highlighted in light blue). But the most interesting finding when sorting by Rate is All causes with a rate of 13,660.4 for All races, both sexes, 85 years and over, which confirms the conventional wisdom that the elderly are most at risk for death (highlighted in light yellow). This is (thankfully) further confirmed when sorting the Number column ascending and finding that the 30 lowest numbers of death are in the groups where age ranges from 1-4 years and 5-14 years (highlighted in cyan).

Note, these findings are from just the first 10 pages of data from the data source. More data may reveal different and more accurate results.

3 ANALYSIS

3.1 What types of arithmetic operations did you use? Why? What was the outcome of those?

Using the `AVERAGE` function, I obtained the average of <u>Numbers</u> for each category and put them in the row of the #1 cause of death for each category. Then, for each category, I divided the number for the #1 cause of death by the average to see how much greater the #1 cause of death was compared to the average. I found it very interesting to see that the #1 cause of death in each category ranges from about 4 to 7 times more than the average.

3.2 What types of statistical functions or equations did you use? Why? What was the outcome of those?

The mode of Numbers would not be interesting because it is only a group of 15 numbers with values differing from one another greatly and therefore it's unlikely that there would be much or any repetition of the numbers. Having calculated the mean (average) in the above steps, and since the range was clearly visible in the group of only 15 numbers in each category, that just left the median. Using the `MEDIAN` function, I calculated the median in each category and put it on the same line as the #1 cause of death, and highlighted the numbers in that row light blue so it would be easier to compare the #1 Number to the median. I then calculated how much the #1 cause of death was than the median. This varied greatly, from 9.1 (All Races, Both Sexes, ages 35-44) to 75 (All Races, Male, ages 15-24).

3.3 Describe the data displayed in your Pivot Table. Share how you chose the data elements and why they matter.

In the pivot table, I added rows for <u>Cause of death</u> and <u>Rank</u>. For columns, I chose <u>Category</u>. Upon seeing how the categories weren't displaying well in the original format, I abbreviated all of them in the source sheet so that they could be viewed better in the pivot table. (For example, "All races, both sexes, all ages" became "AR,BS,AA".) In the Filters for the pivot table, I excluded rows where the <u>Cause of death</u> was "All causes" or "All other causes (Residual)" because those are aggregate fields that would skew the data presentation.

I didn't gain much more insight from this pivot table except to learn that while some causes of death consistently topped the charts (like *Diseases of heart (I00-I09,I11,I13,I20-I51)* with ranks 1-7 and *Malignant neoplasms (C00-C97)* with ranks 1, 2, 4, & 5), other causes of death showed up in a greater number of categories. For example, *Chronic lower respiratory diseases (J40-J47)* showed up in all ranks between 3 and 14 except 11; and *Intentional self-harm (suicide) (*U03,X60-X84,Y87.0)* showed up in all ranks between 2 and 13 except 6, 9, and 12.

4 DASHBOARD

4.1 What does this data tell us?

This data tells us about the top causes of death for each category of people broken up into groups based on sex, race, and age group. The data on the Dashboard sheet narrows down data from other sheets. It shows only data for the top 3 causes of death in each category, and their respective quantities above average.

4.2 What does your analysis tell us? How does it differ from the raw data?

My analysis shows the average number of deaths in each category as well as the median death count in each category. It goes on to show how much greater the top cause of death in each category is than the average. This data is not shown in the raw data.

4.3 What do your visualizations tell us? Does it match your analysis?

My visualization shows the amount by which the top 3 causes of death are above average. It goes a little further than my analysis in that my analysis only deals with the #1 cause of death. It also illustrates that, while the top 3 causes of death are above average in most categories, this is not the case in ALL categories. Particularly, in the category "All races, both sexes, ages 55-64", the 3rd leading cause of death, "Accidents (unintentional injuries) (V01-X59,Y85-Y86)", is below average. This is a new insight that is not found in the Analysis sheet.

4.4 How does your dashboard make the raw data easier to digest?

The visualization derived from the dashboard sheet makes it easier to spot which leading causes of death are the greatest above average, and which one is below average.

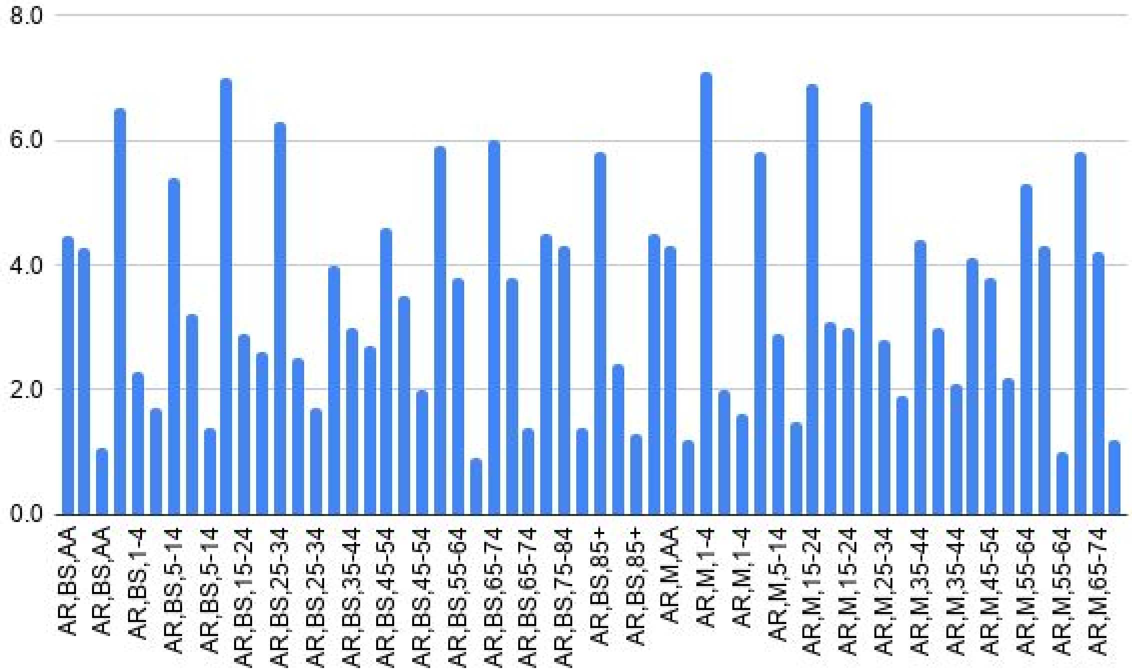

Figure 1—Visualization from Dashboard

4 REFERENCES

1. CDC. (2014). LCWK2. Deaths, percent of total deaths, and death rates for the 15 leading causes of death in 10-year age groups, by race and sex: United States, 2013. https://www.cdc.gov/nchs/data/dvs/LCWK2_2013.pdf. Published December 31, 2014. Accessed February 20, 2021.