Maggie Kahn

Business Analytics

Professor Levkoff

October 25, 2020

**Executive Summary:**

After we received instructions for this project, I immediately started looking for data that sparked some interest. I knew that sugar consumption is a huge issue in our country, so I wanted to see if it was a problem elsewhere around the world. I got my original data from one of the websites that was provided on gap minder. It showed different countries (categorical) and how much sugar in grams each day they were consuming (numerical). The data was recorded each year beginning in 1961and ending in 2013. The range of years gave me a little too much data to work with so I decided to narrow it down to just sugar consumption between the years 2006-2013. I chose this range because it held the most data for each country. After finding this data set, I looked for other variables I could compare and contrast with sugar consumption and see if they had any influence on each other. My other variables included life expectancy in years, GDP in dollars, and GDP per capital growth (%). I found these variables on the world bank data base. When I looked at all the data on sugar consumption, it was a little overwhelming to analyze so many countries, even after eliminating the years outside of 2006 to 2013. I decided the best option was to look at the countries that had the most sugar consumption, specifically the top ten. I used the same top ten countries for the remaining variables.

**Cleaning the Data:**

Once I had my data, I transferred it to excel in order to start the cleaning process. Upon opening the data, I noticed there were a lot of columns that weren’t necessary for my project but I decided to wait till I transferred it into R-studio to delete some of them. While I was in excel, I deleted the years that weren’t in the range I had previously described. I made it easier to work with in R Studio by renaming some of the columns so I would have an easier time typing them in the code. For the list of data for sugar consumption, I changed the name just to sugar, for life expectancy, I just used the word life, etc. I also created a new column and took the average sugar consumption, GDP in dollars, GDP per capita growth, and life expectancy for the years 2006 through 2013. This was another effort to cut down on the amount of data I had so I could later graph the results in a neater more concise way. Once I got into R-studio, I installed a package that has the capability to read excel files and then had it read each of the excel files that I had. I then proceeded to clean the data even further. For the sugar data, I omitted data that had no results. I then eliminated columns except for the country code and the average sugar consumption for those 10 countries. I eliminated everything besides the top ten results by writing a code that says to delete values less than the tenth value. After starting to graph my results, I realized that it was easier to read the country labels if they were shorter, so I used those instead of the full country name. I repeated this process for the GDP, GDP percent, and life expectancy.

**Exploratory Analysis:**

After I cleaned the data, I decided to plot the data in different bar plots to compare the result. First, I started with the sugar consumption data (y axis) and the ten countries that had the highest results (x axis).

Chart, bar chart

Description automatically generated

After making the graph, I noticed that the bars were at a similar level so I went ahead and labeled each bar so you could see the amount of sugar in grams. When you take a look at the graph, you can see that Cuba had the lowest sugar consumption per person per day while the United States (of course) had the highest. I also made a graph that showed the distribution for the graph above.

Chart, histogram

Description automatically generated

This graph shows a n=10 (10 total values) and a bandwidth of 3.827. It also shows that out of all the countries, most have an average of around 160gram of sugar, this is also known as the distribution.

The next graph showed the average life expectancy in years between 2006-2013. As I mention before, I took the same top ten countries and graphed them to see the results. I also added the values on top of each bar so you can clearly see the values. After looking at the graph you can tell that Switzerland had the highest life expectancy. Before I started to graph the results for this graph, I thought that the country with the lowest sugar consumption would have a higher life expectancy, but I was wrong. By looking at the graph, you can tell that Switzerland actually has a higher sugar consumption than I would think.

Chart, bar chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

This density graph shows that there was a higher distribution on the life expectancy around 78 years of age.

My next graph shows the average GDP in dollars from the years 2006-2013, using the same countries as the first graph. Again I put the values on top of each bar which was really helpful since all the values accept USA are pretty close. It was also really interesting to see which country would come in second. While sugar consumption doesn’t have a direct correlation with the GDP, I was mostly looking at if sugar consumption contributed to spending in our country. The United States has both the highest GDP and the highest sugar consumption. Chart, histogram

Description automatically generated

Chart

Description automatically generated

The density graph shows that a distribution is to the left. It also shows that the density only spikes for the Unites States and then goes back down because the rest of the values are so small.

My final graph shows the average GDP per capita (%) for the years 2006-2013. The GDP per capita shows the overall prosperity of the country. I used the same parameters as I did for the other graphs. This graph shows that Cuba has the highest GDP per capita and Barbados has the lowest. I understand why Barbados has the smallest but I was quite surprised when I saw that Cuba had the highest and not the United States or another big country. Compared to the sugar graph, I don’t believe it has a lot of correlation unfortunately.

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

The density graph shows a bandwidth of 0.9397. It also show that there was a density of around 0.2 when the x-value is around one. The distribution looks like it is skewed to the left.

**Conclusion:**

The final section should wrap up your results with a conclusion and discussion of lessons learned along the way

Overall, I do believe that there was a correlation between some of the variables like sugar consumption and GDP but I’m not quite sure if it’s a direct correlation or a coincidence because they are both pretty broad topics. For future analysis, I hope to be able to test these two variables in a more direct manner.

This project was definitely more difficult than expected. It took me a long time to figure out the right code and make sure that it works properly. Once I did get everything working, I had a good time playing with the colors of each graph and seeing what I could add to further my understanding of the topic.