Storytelling with data

I) The Dataset: “Nutrition, Physical Activity, and Obesity”.

This particular dataset was published by the *Centers for Disease Control and Prevention*, owned by the *Division of Nutrition, Physical Activity, and Obesity (DNPAO)* and includes data on the diet, physical activity, and weight status of American adults from Behavioral Risk Factor Surveillance System. The data was provided by several sources including: *Centers for Disease Control and Prevention (CDC)*, *National Center for Chronic Disease Prevention and Health Promotion*, and the *DNPAO*. The data collected is primarily used for DNPAO's Data, Trends, and Maps database, which provides national and state specific data on obesity, nutrition, physical activity, and breastfeeding. The original raw dataset published on the website contained 76.4K rows and 33 columns and thus provided a large quantity of extremely diverse data collected over the last decade from 2011 to 2019, using an overall aggregate sample size of approximately 2 million US adults.

II) **Extraction, Transformation, and Loading (ETL) Process**.

As previously mentioned, this raw dataset was vast. It contained quite a bit of information that ultimately was not useful for the analysis I wanted to do and the visualizations that I chose to create.

I deleted all the columns that I thought were unnecessary like: Data Value Footnote Symbol (which was entirely blank), Low Confidence Limit, High Confidence Limit, Geographic Location, Class ID, Topic ID, Question ID, Data Value Type and Data Value Type ID (both of which only contained the word “value” repeated thousands of times), Location ID, Data Value Unit (which was entirely blank), and the Data Source column (which only contained the name of the source “Behavioral Risk Factor Surveillance System” in all its cells). The “Year Start” and “Year End” columns simply contained the year in which the data of a particular individual was collected and thus, the two columns were identical. Therefore, I chose to delete the Year Start column and rename the Year End column as simply “Year”. Similarly, the Class and Topic Columns were separate but identical and so I deleted one of them and renamed the other as “Class and Topic”.

The Data Value Footnote Column signified all the rows in the dataset which contained no information and so I filtered out all the “Data not available because sample size is insufficient.” values from the column and removed all of those rows. Doing that left the entire column blank and so I proceeded to delete the column as well. The “Question” Column contained 6 questions that the dataset dealt with. As I had decided to focus on the topic of “Percentage of adults aged 18 years and older who have obesity”, I decided to filter out the remaining questions from that column.

I noticed that the Data Value column included only the percentage value of the sample size and so I renamed it as “Percentage of Adults”. I wanted to work with whole number values, so I inserted three additional columns in between the Percentage of Adults and Sample Size columns. In the first column I used a mathematical formula to find the numeric values using the percentage and the sample size. In the next column I used the ROUNDUP function to round up the decimal values to whole numbers (because the number of people could not be in decimals). Finally, I copied the whole number column and pasted its values into the third column, naming it “Number of Adults”, before deleting the previous 2 columns that I had used for its calculation.

This was my cleaned dataset. I then used a filter on the Stratification column and created separate sheets, categorizing the given information by the age, gender, income, race and education of the individuals as well as a sheet for the overall category. This was my final excel dataset that I proceeded to use to create my visualizations.

III) The Visualizations.

I chose to create 6 different visualizations to represent the obesity count in the diverse categories of adult US citizens, using a common colour scheme of an orange-gold gradient. The gradient portrays the highest to lowest range more efficiently. Additionally, I personally thought that the warmer orange-gold colour scheme would be ideal for this dataset as it talks about obesity which is not exactly a positive topic and thus cooler tones of blues and greens would not be as effective to the audience. The 6 Visualizations were:

1. *Top 5 US states with the highest count of obese adults.:* I created a bar chart in descending order to depict the top 5 states with the highest obesity in adults. This bar chart shows that Nebraska, Florida, Kansas, Minnesota and Maryland were the states with the highest cases of obesity; Nebraska having a count of 47,395 in total.

2. *Number of US adults with Obesity: classified by age.:* The ages of the demographic was classified into 6 different categories and thus I thought a tree map would be the most ideal way to represent this data. This Tree Map depicts that the highest population of obese adults are aged 65 and over (731,762) and the lowest population are aged 18-24 (68,897).

3. *Top 5 US states with highest count of obese adults: classified by age:* For this visualization I combined the previous two visuals to create a stacked column chart, depicting the number of people in each of the age categories in each of the top 5 states. The range of population in each age category still remains relatively similar. The count of 65+ year olds with obesity range from about 12,827 in Maryland to about 16,398 in Nebraska, whereas the count of 18-24 year olds with obesity range from 649 to 1,486.

4. *Number of US adults with Obesity: Male Vs Female:* As this data only included 2 categories (Male and Female), I thought it would be best to use a Pie Chart to represent it. The pie chart shows that on an overall, the number of adult women with obesity exceed the number of adult men, with women being 55.69% of the total (1,247,953 in number) and men being 44.31% (992,773).

5. *Trend of obesity in US adults (2011-2019) :* The most ideal way of depicting the count of obese adults by each year was by creating a line chart. This line chart shows that the trend in obesity

throughout the decade remained fairly consistent, ranging from about 230k to about 260k new cases every year. 2015 recorded the lowest cases of obesity (232,901), whereas 2016 recorded the highest (260,797).

6. *Number of obese US adults categorized by race/ethnicity:* For my final visual, I chose a heat map to represent the count of obese adults categorized by their race or ethnicity. The heat map shows that Non-Hispanic White adults have the highest cases of obesity (1,672,637) is thus represented with the darkest colour of the orange-gold spectrum. Meanwhile, Hawaiian/ Pacific Islanders have the lowest cases of obesity, totalling at only 6,194. Moreover, this visualization also shows that the number of Non-Hispanic White adults with obesity exceed the remaining races and ethnicity by a dramatically large margin, with the Non-Hispanic Black adults coming in second with a count of only 2,36,713.

IV) Summary.

In conclusion, I chose a healthcare related dataset that contained data on the diet, physical activity and weight of American adults. The dataset was extensive and intricate, but also contained a lot of unnecessary information and thus had to be cleaned by fairly complex ETL methods in order to be clearly and accurately interpreted as well as represented. I chose to focus my analysis on and create unique visualizations entirely about the problem of obesity and how many people across diverse demographics have been affected by it. I personally believe that although I tried to be as precise in analysing and portraying the data provided, the sample size was a little smaller than what such a huge country should warrant over an entire decade and thus the conclusions drawn from the data and the visualizations should be taken with a grain of salt.