Haley Mathews & Jennifer Russell
David Mimno
INFO 2950: Introduction to Data Science
May 14, 2019

Average Red Meat Consumption in the United States: Data Sheet

Meat Consumption and Agricultural Support Data

Meat Consumption Source: https://data.oecd.org/agroutput/meat-consumption.htm
Agricultural Support Source: https://data.oecd.org/agrpolicy/agricultural-support.htm

The data for meat consumption and the agricultural support farmers receive is from the Organisation for Economic Co-operation and Development (OECD). OECD collects information on many topics to help governments around the world generate prosperity and decrease poverty levels, making sure to account for the environmental implications of social and economic development[1]. This data is used for OECD's Agricultural Policy Monitoring and Evaluation publications. Overall, the OECD is a reliable source of statistics and data and is cited as a trusted source by many reputable sites, such as the U.S. Department of State. Therefore, we can be sure that there was no processes to influence the data collection.

The meat consumption data was collected to have a variable to quantitatively measure meat consumption over time. Agricultural support data was collected to try to understand the impact of the government's financial support on the amount of livestock farmers can care for and determine how farmers' resources impact consumer consumption.

To get the amount of red meat consumed by Americans, a lot of preprocessing had to be done on the data. The meat consumption data displayed kilograms per capita of beef and veal, pork, poultry, and sheep meat a person ate for the years 1990-2017. The data contained meat consumption from many countries, so first we had to select only rows where the location was the United States. Then, for each year, we added the value of the meat consumption for red meat (beef, veal, pork, and lamb) to get a total red meat consumption in the United States in kg per capita. Finally, we normalized the data by dividing by 1000, to get red meat consumption in metric tons per capita. We used this data to create the final data set, to which we added all other data.

[1] OECD, https://www.oecd.org/about/whatwedoandhow/

The agricultural support data used in our project measures the total support estimate (TSE) in millions of US dollars.To get the TSE in the proper format for the data set, we had to do some preprocessing on the data set. First, we had to iterate over every row in the data set. Then, we had to add to add the value of the column containing the value of the TSE to a list and convert those values into floats so we could run analysis on the TSE data later on. There was no data for the year 2017 available yet, so we had to add one empty string to the list so the data would be the correct size and then the TSE was added to the complete data set. In order to better analyze the data, TSE was divided by 1000 to get units of billions of US dollars.

Unemployment, GDP, and Inflation Data
Unemployment Source:
https://data.worldbank.org/indicator/SL.UEM.TOTL.NE.ZS?locations=US
GDP Source: https://data.worldbank.org/indicator/ny.gdp.pcap.cd
Inflation Source: https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?locations=US

This data comes from the World Bank, an institution that has two goals for the world: end extreme world poverty and promote prosperity for the bottom 40 percent of each country.[2] The World Bank cites that their data comes from the statistical systems of the member countries and used to monitor the effectiveness of their strategies. Because their goals are to promote development through different services, they could potentially have an interest to show unemployment or inflation rates dropping to prove the effectiveness of their actions. If true, this could give the data a conflict of interest. While this data involves people, it is unclear if they were aware of the collection or use of the data since no information on this subject is given. However, since they cite the information coming from the statistical systems of the member countries, we may assume the data from the US may come from the Census.

Unemployment, GDP, and Inflation are all key economic factors in determining how people can afford to spend their money and gage how comfortably the average American is doing financially. WE hypothesized that these factors may influence trends in red meat consumption in the United States, which is the reasoning behind why these data sets were included.

Preprocessing has to be done to get these data set in a usable form for analysis. First, all label spaces were replaced with "_". Then, the sets included data from 264 countries, and we had to remove all other countries except the United States. Next, we are only comparing the data from the years 1990 to 2017, so we removed the columns with the irrelevant years. Finally, we added the data from each set to a new list that was appended to our master data sheet. GDP was then

[2] World Bank, http://www.worldbank.org/en/about/what-we-do

normalized by being divided by 1000 to improve the clarity of the analysis, so was put into units of GDP per capita in 1,000 U.S. dollars.

Vegetarian, Vegan, and Keto Data

Vegetarian Source: https://trends.google.com/trends/explore?date=all&geo=US&q=vegetarian
Vegan Source: https://trends.google.com/trends/explore?date=all&geo=US&q=%2Fm%2F07_hy
Keto Source: https://trends.google.com/trends/explore?date=all&geo=US&q=keto

The dieting trend data sets came from Google Trends. It is a Google feature that allows you to see how frequently a query has been searched in relation to the total Google search queries over a certain period of time.[3] This data is automatically collected, and we do not believe there are any underlying processes that may have influenced what was and what was not collected because of this automatic process. Further, although people may not actively know their information is being stored about their search queries, it is free, publicly available information that can be used for any purpose.

Google trends can be a strong indication of the topics people are curious about, indicating topics that people are searching for, whether it be for academic, work, or personal motivations. We felt it was important to have qualitative statistics measuring people's interest in specific topics that may affect their personal meat consumption. To do this, we decided upon large dieting trends in the U.S. that don't eat meat (e.g. vegan, vegetarian) and one that promotes large amounts of meat (e.g. keto).

Several preprocessing steps were required to get these data sets into workable forms. First, Google Trend data only dates back to 2004, so we added an empty string in the master sheet for years 1990-2003. Next, the popularity measures were split based on month rather than years as we needed for analysis. To get just the years, we used a regular expression to just capture the years and ignore the monthly differentiation. We then took all the monthly data from the same year, added it together, then created an average for that year. With this yearly data, we appended it to our master data set.

Opinion on Climate Change

Source: https://news.gallup.com/poll/1615/environment.aspx

The data regarding American's opinions on climate change comes from Gallup, an analytics company that is known for its public opinion polling. They apply "rigorous research standards"[4] to their polls and pride themselves on transparency and integrity. However, for each data set the

[3] Google Trends, https://trends.google.com/trends/?geo=US
[4] Gallup, https://www.gallup.com/178685/methodology-center.aspx

exact sample size or demographics are not given, thus we cannot know who is being polled. Along these lines, we do not know how the data subjects were polled and if they knew the purpose of their information. Further, with a topic as divisive as climate change, the demographic of the sample is greatly important. To fully understand the influences behind the data, it would be important to see the sampled subjects. Overall, Gallup had a pristine reputation, though, and is widely cited and trusted statistical information thus we trust this information to be random and accurate.

The data concerning opinions on climate change was also included in the data set as a way to gage people's opinions on predictors directly relating to red meat consumption. As red meat is known to have detrimental environmental effects, we wanted to analyze the relationship between how much people believe they care about the environment versus actual average meat consumption.

To obtain this information, we had to scrape the Gallup webpage. This first included accessing the table where the data was stored. Then, we found all rows of years and datas. Next, we created a regular expression to capture just the years, because the dates were formatted with months and days that were irrelevant to us. For each row of data, we looped through each opinion column and added the percentages to a list. Then, based on its index, we passed each percentage to its appropriate list based on the opinion type. Next, we added the empty values to the years that were not given. Finally, we appended all of these opinion type lists to the master data set.

Obesity Rates

Obesity Prevalence Source:
https://www.statista.com/statistics/244620/us-obesity-prevalence-among-adults-aged-20-and-over/

The data concerning obesity rates in the United States among adults is from Statista.com. Statista is a well established company that consolidates data from other sources and make it available to users around the world[5]. The data was originally surveyed by the National Health Interview Survey (NCHS) and was published the Center for Disease Control and Prevention (CDC) and the NCHS. Since the CDC, a reputable organization created such widely cited data, the data can be trusted as accurate. The survey was conducted through face-to-face interviews with citizens throughout the time period of 1997 - 2017, in which time about 35,000 households were respondents. Since so many households were interviewed, we can trust that it is extremely unlikely that any processes may have influenced that data that was collected. The NCHS collects data on obesity rates, along with other health and nutritional status indicators for future use in

[5] Statista, https://www.statista.com/aboutus/

planning and evaluating public health practices. The estimates of obesity rates are found by adjusting the data to the most recent census data available. The people surveyed gave consent for their data to be used in this survey.

Red meat has many health benefits and risks. To determine if average health of an American citizen is correlated with red meat consumption, the American obesity rates over time was included in the data set. While there are many health indicators, obesity was chosen to quantify the overall health of the American population.

To get the data into the data frame, many preprocessing steps had to be taken. First of all, the obesity rates were floats with many unnecessary decimals, so we changed all of the obesity rate numbers to only have two decimal places. This way, a very small amount of information is lost and the numbers are more readable and relay information clearly. Then, since the data only went back to 1997, we had to account for the years 1990 - 1996. We added an empty string to the list of obesity rates so that the list would be the correct length to add to the complete data frame. Finally, we appended this list to the principal data set.