DATA CLEANING

Introduction

We are integrating two data sources: a data set of all tornadoes in the United States since 1950 (http://www.spc.noaa.gov/wcm/#jmc) and agricultural/crop export data from the USDA (http://www.ers.usda.gov/data-products/state-export-data.aspx). The tornado data comes from the National Weather Service. It includes information on how many tornadoes touched down in a given year, what state the tornado touched down in, what crop losses were, and how many states the tornado passed through. The agricultural export data set breaks down the agricultural exports of each state, as well as the revenue generated per export. This data set also aggregates the data so that a user can look at this information for the United States as a whole as well. From these two data sets we believe we can answer the following two questions:

- How has the tornado occurrence varied over the last 10 years?
- What is the relationship between tornado occurrences and agricultural exports in the United States?

Intellectual Property Rights

Both datasets were readily available online and posted by government sources; thus, there should be no IP or licensing concerns here. For instance, for the tornadoes dataset, the terms of use for this site specify that since the data is in the public realm, it can be used for free as long as it is not claimed as one’s own data, it is not modified, and it is not implied that NOAA is affiliated with the person using the data (http://www.weather.gov/disclaimer). Similarly, on the USDA website, their policy specifically states that ‘Most information presented on the USDA Web site is considered public domain information. Public domain information may be freely distributed or copied, but use of appropriate byline/photo/image credits is requested’ (http://www.usda.gov/wps/portal/usda/usdahome?navtype=FT&navid=POLICY_LINK).

Citations:

Storm Prediction Center WCM Page. *U.S. Tornadoes (1950-2015)* (Data File).

Retrieved November 2, 2016, from http://www.spc.noaa.gov/wcm/#jmc.

United States Department of Agriculture. (2016) *State Export Data*. (Data File). Retrieved November 2, 2016, from http://www.ers.usda.gov/data-products/state-export-data.aspx.

National Weather Service. *Disclaimer*. Retrieved November 2, 2016, from http://www.weather.gov/disclaimer.

United States Department of Agriculture. (2016) *Policies and Links.* Retrieved November 2, 2016 from

http://www.usda.gov/wps/portal/usda/usdahome?navtype=FT&navid=POLICY_LINK.

Metadata

The first dataset (US Tornadoes 1950 – 2015) had a metadata file available with it for download. This file answers the questions - who submitted the publication, who conducted the data review, who the point of contact is for additional information and when was the metadata document last updated. Furthermore, it describes what every field in the database represents and also provides a few examples. When we received the dataset, there were no column headers. We leveraged the metadata file to understand what each column represented and used this information to assign column headers in our data set. For the second dataset (State Export Data), the URL containing the Excel file for download had a data description on the webpage itself, which serves as metadata. The data set has multiple types of plant and animal product categories like *wheat, corn and beef* listed which do not require any explanation. However, it also contains other categories of plant and animal products, such as, “*Other Livestock Products”,* and *“Other Plant Products”*. The metadata provides information regarding the types of products that fall under these categories. For instance, *“Other Plant Products”* includes sugar, essential oils, planting seeds, cocoa and coffee products.

Steps to Remediate Data

The ‘State Exports’ data set has 53 spreadsheets in one Excel workbook. The first spreadsheet describes the revenue generated from agricultural exports by the United States, and the other 52 are specific to each state in the US. Since we want to determine the relationship between tornado occurrence and agricultural exports in the United States, we are considering only two states (Texas and Oklahoma), which have experienced highest tornado occurrences between 2000 and 2014, to address this research question. A state with low tornado occurrence would not help since there would not be sufficient data to back the correlation between the two. As a result, we end up with 3 spreadsheets in the second data set – One for the United States, one for Texas and one for Oklahoma.

Issues Encountered with the Data:

- The first issue encountered was missing column names in the tornadoes data set.
- After downloading the database description file (metadata) from the same link as that of the database and manually entering the column names in Microsoft Excel, there were columns that were not self-explanatory. We had to refer to information on the website as well as the database description sheet for details of attributes like State FIPS Number, F scale, County FIPS code.
- The tornado data set provides information regarding the date and time of the tornado occurrence each year, whereas the agricultural exports data set is consolidated annually. As a result, the data sets cannot be merged into one spreadsheet, and we have to analyze them on a yearly basis and not corresponding to the dates mentioned in the tornadoes data set.
- The tornadoes data set is available for a period between 1950 and 2015 and the Agricultural exports data set is available only for period the between 2000 and 2014. We had to filter data from the tornadoes data set and restrict our analysis to a period between 2000 and 2014.

Cleaning process:

1. The first step in the process of cleaning is checking the number of records in the data set and identifying the software/tool to be used. Data set 1 has 61217 tuples and data set 2 has 24 tuples. Hence, the process of cleaning can be performed in MS Excel.
2. The next step is naming the columns in the data set using the database description file downloaded from the same URL.
3. After carefully analyzing the research questions, the next step is to filter out columns in the tornadoes data set which are not being used for analysis. The columns I (State FIPS #) to AC (Wind Only) are deleted from the tornado data set.
4. The second data set (agricultural exports) has a row with a header and two rows below the attribute names which has no relevant data. These rows are deleted from the data set.
5. We deleted rows at the end of each sheet in the agricultural exports data set which provided additional details about data in the sheet.
6. The agricultural exports data set has different spreadsheets for each state. To be able to analyze the data in R, we had to move the data into 3 different files, and then import it into R.

Word Count: 1070