ETL Assignment

Andrew Johnson, Laurie Grant, Sean Evans

We as a group delved into three datasets to help examine the agricultural imports and exports of goods throughout the world. This report highlights the importance of trade and how much people rely on cooperation for their daily goods.

Extract

1 – The first was a cumulative amount of meat and livestock imported and exported to the United States. This data was pulled from the United States Department of Agriculture as an Excel document. Country consumption data was pulled from Kaggle.

2 - Our next dataset consisted of national ISO country codes. This information, obtained on the website <https://www.nationsonline.org/oneworld/country_code_list.htm>, provided an opportunity to join our separate datasets on a mutual variable. This information was extracted in *three* distinct ways. The first and easiest method involved copying directly from the website and pasting the information into an excel sheet – which was later saved onto a CSV file. The second method was utilizing Pandas and creating a Data Frame of the information. The final method of extraction utilizing web scrapping.

3 - The final dataset was pulled as a clean CSV file from the website Kaggle. Titled “Who Eats the Food That We Grow” the data is comprised of the food supply for nearly 245 countries since 1961. Provided by the Food and Agricultural Organization of the United Nations, the dataset is comprised of over 21,000 rows of data that provides Country, Country Area Code, the Foods universally utilized by each listed country since the year 1961.

Transform

- Our first dataset, comprised of the cumulative imports/exports, was initially cleaned on Excel. We sectioned the information from the original table onto separate Excel Sheets. Some of this data would not import onto MySql and required returning to Excel to remove the blank cells.

- As stated above, there were three separate ways of data cleaning the ISO data set. The initial extraction onto an excel file required no clean up. The second method of extraction required the Pandas Data Frame and manipulating

The third cleaning method of Webscrapping

-The third dataset containing the data from Kaggle was initially uploaded directly into SQL as a CSV file. Unfortunately, the final table only contained just 4,000 fields of data – a drastic loss compared to the original 21,000 rows. To solve this issue, we utilized Pandas to create a data frame which then manipulated the ISO to make it more upload friendly.

Load

Following the cleaning, each data set was properly uploaded onto a SQL Database labeled ETL. We were able to join the three separate datasets utilizing the similar variable ISO Codes. The purpose of creating this SQL - ETL Database.