Chris McGee, Jessica Schielka, Anthony Suitto

Group 8 ETL Project

In this ETL project, our group collected data from two different sources Kaggle.com and Census.gov to look at relations between avocado prices and city populations. Our avocado CSV dataset contained 18,300 rows of data and our census CSV dataset contained 29,575 rows of data both with multiple columns for data analysis.

As part of the ETL project, our group performed an Extract, Transform, and Load of our two datasets where we present the data on an HTML webpage.

Extract: Our group found two datasets in the form of CSV's. The first dataset was from Kaggle.com, which contained historical data on avocado prices and sales volume in multiple US markets. The CSV included multiple columns such as Date, Average Price, Total Volume, 4046, 4225, 4770, Total Bags, Small Bags, Large Bags, XLarge Bags, type, year, region.

The second dataset was from Census.gov which contained 2017 population data, which included multiple columns such as Country, City, Population, Housing units, Population density, Housing unit density, 2017 Median Household Income, 2017 Median Age.

Transform: In order to make the data more readable, a jupyter notebook file was created. After reading in the csv and displaying it, it was clear that both of the data sources had excess columns. To get rid of the unnecessary columns, the dataframe (.drop) function was used. Next, the column names were changed to get rid of spaces. The next thing we decided to do was to set the region as an index. We were trying to sort the data by city/region, so we performed operations on the census data using the (.groupby) function paired with (.sum) and (.mean) functions. After getting sql errors, we notice that our data was not completely clean and certain results on the census csv were causing the errors. To get rid of these errors, the (.loc) function was used to get rid of the final errors. After all the errors were caught, the data was put back into a dataframe and finally be ready to work the data into MySQL.

Load: To load the data into the SQL server we created a relational database connection from the Jupyter Notebook file. The loading gave us some issues due to exceptions involving the average price column and unknown characters in the integers in some of the other columns. To solve those issues, we had to make the average price column a 'FLOAT' data type and get rid of the dollar sign($) as SQL does not accept it. The other columns involved the census data table, where a couple rows included a dash(-) to show no data. We solved that issue using .loc as stated above. After all of those exceptions were configured, the data tables were loaded over successfully into the SQL database, where they were able to be searched from.Our schema.sql file shows the creation of the tables and their data types, as well as the selections from each table that were loaded onto our webpage.

Chris McGee, Jessica Schielka, Anthony Suitto

Data Sources: https://www.kaggle.com/neuromusic/avocado-prices/home and https://www.census.gov/