**Acovocado**

**Parker, Ariel, Joe, Matthew**

ETL Project

Are trends identifiable between organic and conventional avocados sold in California and the temperature at a specific date? In order to answer this question, we needed to find datasets that would allow us to query off of avocado sales in California as well as the weather.

To begin the extraction process we needed to find two separate datasets. Original data sources consisted of csv files pulled from Kaggle as well as National Centers for Environmental Information (NOAA). Avocado cost and volume data had been pulled from a previous study on avocado sales across multiple states in the US. Weather data was obtained via a Statewide Mapping report that provided detail regarding the temperature for 2015-2020.

In order to transform the files into a format that would allow us to take our data and load it into a database for future querying, we created a Postgres db (Avocado\_db) that would hold the cleaned csv data in two separate tables. To begin, we built a schema that first dropped all created tables if they existed. Then ensuring that we correctly identified the value types of each column on our cvs, we created our two tables, Avocados, which held our avocado sales data and Calicli, which held our climate data for California. The Postgres schema also contained a query that created a third table where we held our final table called Calisales. This table held the total volume of the two types of avocados as well as total & average price data pulled from our Avocados table.

To populate our tables, we had to load the information from the csv files into Postgres through use of a Jupyter notebook. The Jupyter notebook was connected to our Avocado database via the create\_engine() function and by using sqlalchemy, we were able to reflect the tables from our Postgres db into Jupyter and confirm our table names. After confirming the table names using the inspector, we then used the .get\_table\_names() function to print out our column names within our Avocados table and our Calicli table. By obtaining the column names, this allowed us to define our tables by creating classes so that all three of our tables would populate when the database ran. Next, we read in our cleaned csv files and created new data frames via pd.read\_csv(). We then wrote the records from the two data frames into our Avocados table and our Calicli table using the .to\_sql() function which is how we were able to take our csv data and populate our tables in our Postgres database.

Our final database in Postgres contained our query schema that allowed us to create three tables called Avocados, Calicli, and Calisales. The Avocados table which contained data from our cleaned avocado\_df, the Calicli table, which contained climate date from our cleaned calicli\_df , and the Calisales table, which consisted of total and averaged sales and price data from the Calicli table.

The last step to needed to get our data in a format for future querying, was to load the data into our database. The database produced a join via a query that pulled the total price, average price, and total volume for organic and conventional avocados for specific dates between 2015 and 2020.The query also included temperature data in the value and anomaly columns. The value column represents the temperature on the date in the record date column while the anomaly column contains temperature anomalies for that date.

Postgres was the database of choice due to our current comfort level with the program. For this project, it felt as though using a relational database such as Postgres would be more beneficial since we were looking at csv files with predefined relationships which we identified by using primary keys. We believe that by using the database we created, a peer should be able to quickly find trends between temperature and organic and conventional avocado sales in California.

Sources

Kaggle, Avocado Prices (2020), published December 2020, retrieved on December 12, 2020 from  <https://www.kaggle.com/timmate/avocado-prices-2020/>

NOAA National Centers for Environmental information, Climate at a Glance: Statewide Time Series, published December 2020, retrieved on December 12, 2020 from <https://www.ncdc.noaa.gov/cag/>