Group-8-ETL Project- Technical Report

Members: Eddie Xu, Mackenzie Baucum and Vasudha Nair

Starbucks locations vs. Weather Conditions in US

Extraction

Sources of data we will extract from:

The Open Weather Map API (http://api.openweathermap.org/data/2.5/weather?)

Kaggle- Starbucks Locations all over the world (https://www.kaggle.com/starbucks/store-locations)

The Weather Map API data is a JSON file that is converted into a CSV file during transformation and the Starbucks locations data is a CSV file (directory.csv).

Transformation of the data

The public data was then used to perform the following actions

1. To load the CSV and JSON file we employed the Pandas function in Jupyter notebook.

a) The CSV file (directory.csv) was read and converted into a dataframe. This is the Starbucks locations data all over the world (store_data). Next, we only selected the Brand name Starbucks and got down the number of rows (starbucks_df).

b) A new dataframe with only the locations in US was the next logical step. We named this US_starbucks_df.

c) The Postcode column with NAs were removed using the dropna function.

d) A city list based on Starbucks location was created to shortlist the unique cities

e) Next, looping through the list of cities, a request was performed for weather data on each. For this the weather API key was used to access the Open Weather Map API data. Cities for which the weather data was found was printed Weather found in {city_name}. Where a KeyError occurred for the data, it was printed City not found. (weather_data).

f) A weather_df data frame was created using the weather_data from above. The data was also stored as a CSV file(**weather_data_from_starbucks.csv**).

g) A cleaner dataframe for the Starbucks locations was created by dropping Brand name(new_starbucks_df). The data was also converted to a CSV file (**us_starbucks_location.csv**).

All of the above is saved in the ETL_Data.ipynb notebook.

2. Next steps in cleaning up the datasets involved dropping the variables that were not relevant to our study. For this as well as loading to Postgres we worked on a second Jupyter notebook file, ELT_Data Cleaning.ipynb. For this we used the CSV files (**weather_data_from_starbucks.csv** and **us_starbucks_location.csv)**.

a) The data frame starbucks_df was created by dropping Unnamed:0 column and renaming all the column names with lower case and introducing an underscore wherever the column names are two words.
b) We followed the same style for the weather_df dataframe and additionally removed Lat, Lng, and Country columns and renamed the cleaned data frame as new_weather_df .
c) We merged the two datasets using an inner join on city column.

Load

The last step was to transfer our data into a DataBase. We created a database in Postgres(starbucks_weather_db) and respective tables to match the columns from the Pandas dataframes using MYSQL and then connected to the database using SQLAlchemy and loaded the result. Here we were able to perform multiple queries to suit a desired criterion. For linking to Postgres, we need to enter our own username and password. From Notebook, we can find out the table names we created in Postgres after starting the engine (weather_data, starbucks_location). Next, we insert the data from new_weather_df into the table weather_data and follow the same for starbucks_df into starbucks_location table by using the .to_sql function and using the engine. Next, we inspect if we are able to read the data from Postgres using the select * from weather_data and select * from starbucks_location in notebook. We get the same data as is stored in the dataframes earlier in the notebook. We looked at two queries – one which explored the ownership type of Starbucks locations and the humidity over there and the other being the store numbers for the Starbucks locations and the cloudiness with the value of 1.

Summary

We used these datasets so that we could determine the weather conditions that were prevalent at various Starbucks locations in the US. The merged data helps us recognize various aspects of the weather like Temperature, Humidity, Cloudiness, Wind speed at these locations. The data could be used to look for trends and correlation among combination of the datasets. Looking at the weather conditions, it would be useful to find out which Starbucks locations are safe to visit frequently, and which ones have acute weather conditions as a result of which the ones to avoid.