

Project Title: Weather Conditions at Starbucks locations in US

Group Members: Eddie Xu, Mackenzie Baucum and Vasudha Nair

Project Description/Outline

Our team is trying to determine the weather conditions that were prevalent at various Starbucks locations in the US. We will be studying the various aspects of the weather like Temperature, Humidity, Cloudiness, Wind speed at these locations.

Data sets to be used

- <http://api.openweathermap.org/data/2.5/weather?> (Open Weather Map API)
- <https://www.kaggle.com/starbucks/store-locations> (All starbucks locations in the world)

The field of interest for the Weather data set are City, Latitude, Longitude, Maximum Temperature, Humidity, Cloudiness, Wind speed, Country and Date.

The field of interest for the Starbucks locations are Store Number, Store Name, Ownership Type, Street Address, City, State/Province, Postcode and Phone number.

Extract: The weather data is a JSON file that is converted into a CSV file and the Starbucks data is a CSV file.

Transform: In order to transform the public data and use it in our study we performed the following:

- Used Pandas functions in Jupyter Notebook to load the CSV and the JSON file
- Reviewed the files and transformed into data frames.
- For the Starbucks Data set we only focused on Starbucks Brand (there were other brands as well). The data has all the locations in the world. We narrowed it down to Country as US.
- Next we dropped the locations that did not have a Postcode and looked at the unique cities for these locations.
- For the Weather Dataset we focused on cities for the Starbucks locations in the US and using the API key queried it. We used the except function for KeyError if the city is not found and printed the complete Weather data
- Joined the two data sets by City (inner)
- Created queries to address our question

Load: The last step was to transfer our final output into a Database. We created a database and respective table to match the columns from the final Panda's Data Frame using Postgres database (PG admin) to store our original clean data sets. We reconnected to the database and generated additional tables for the data frames. There are some limitations to our findings due to the data available. However, we were able to address our hypothesis question in our initial project proposal listed in the ETL Project Final Technical Report