Umar Hassan & Abby Lemon

ETL – Project

March 7, 2020

Target & Starbucks (menu) / Location Data

**EXTRACT:**

Starbucks Location Data Source:

<https://www.kaggle.com/starbucks/store-locations>

Starbucks Menu Data Source:

<https://www.kaggle.com/starbucks/starbucks-menu>

Target Location Data Source:

<https://www.kaggle.com/starbucks/starbucks-menu>

This data was formatted as a CSV

**TRANSFORM:**

We loaded the data into Python and worked to identify columns of interest. For both of the locations datasets, we decided to drop international locations to have some coordination with Target location data. Furthermore, we decided to drop coordinates columns in the interest of preventing duplicate data as we already have individual addresses of the different store locations. We also decided to change the postcode data type to integer from text. The postcode (or zip code) was transformed to only keep the first five digits. This will allow for both datasets to have the same format for easy merging or joining in the future.

For the menu data, the majority of the transformation was done in the column headers. This was done by turning the column headers into a list of strings. Then some logic was applied to remove unwanted characters and spaces and replace them with underscores. The text was also edited to be all lowercase. Then the column headers were renamed. This allowed for easier SQL table creation and future querying.

We proceeded to create a SQL database with corresponding columns to stage it for connection to the Python file.

**LOAD:**

Upon the completion of the data transformation, we finalized our code and created the connection to PostgreSQL and loaded the data using Python. We choose this method because it allowed us to create a database with the data of interest for further analysis. This data was also very structured, so we thought a SQL database would suit this data type best. The relational nature of this data also lead us to use PostgreSQL.