Project Subject Area: Restaurant Businesses with Yelp and Wikipedia Datasets

Yelp is a platform customers use to find details about the business, contact and rate them. Yelp is used like Yellow Pages but is more interactive, where customers can collaborate and provide meaningful feedback of the businesses. Yelp data includes multiple information from different businesses in the United States, but I filtered and considered categories with Food and Restaurants in this project.

Data wrangling is a process of converting raw data to a clean and usable form. It's a series of steps that someone would have to take to create quality, reliable and usable datasets that can be further used in data science projects to produce insights. Here are the steps I took to complete the project milestones.

1. Data Discovery: This step of the process is when I research a subject area and select the data that I want to work with. The milestone project allowed me to look for different data sources and forms and, in this process, I was able to get hands-on extracting data from CSV, JSON, WEBSITE, API format using various Python libraries and loaded them into panda data frame. I was able to read, analyze and understand the dataset to further plan for the next steps needed to clean, transform, and organize the data.

2. Transformation: In this part of the data wrangling process is where I implemented the plan from the prior step, and there are several steps that I have to do with the data, which are as follows:

   - Structuring: This step is where I had to parse the JSON and Website (HTML) and structure them in the same format and dimension as the other datasets, which is needed so I can merge and combine data for better representation. This step also resulted in the dataset with correct columns and headings for easier reference.

   - Data Cleansing: This stage involves cleaning the raw data, where I have performed the following using Python libraries.

     - Deleting and removing unnecessary columns and data elements.
     - Identifying null columns and removing rows with several nullable fields based on a threshold.
     - Removing duplicates and outliers and handling missing data
     - Standardize columns by formatting, changing data type, and resetting indexes.

   - Data Enrichment: This phase involves looking at the cleansed data and determining how we can further enhance our information at this stage. This process is where I'm able to merge/combine data to one dataframe and also create additional elements from existing raw data such as summary count information, parsing data out of a key: value information stored in a data frame dictionary column then created that as a new column in the data frame. This stage has allowed me to make more data points that can be available and useful for data analysis.

3. Validation: This part of the data wrangling process is where I validated the dataset by looking at the shape of the dataframe (count and expected columns), viewed a subset of information, and validated the quality of the transformed data.

4. Storing Data: This part involves creating tables and inserting the dataset in the database where it can be easily accessed.

The data I used for this project is public, and I don’t see any ethical implication for privacy and security of individuals. However, data integrity can be an issue during the cleansing and transformation process, where errors can be introduced, resulting in inaccurate data. This is where the validation step in data wrangling is critical. During the data wrangling process with raw data from multiple data sources, I need to look at the right amount of data to consider so the dataset is not overweighted or less represented, that may have an impact when data is used and could result bias in machine learning models or data analysis.

Data wrangling of Yelp and Wikipedia Datasets is fun and challenging. This project has allowed me to use the knowledge and practice the skills I gained from working on different bi-weekly exercises from this class. I was able to use different Python Libraries to Read, Extract, Clean, and Transform data. This has been a great semester, and I am looking forward to continuing to learn and apply these skills to future Data Preparation projects.