**Group Project: Analysis on Yelp Business Data**

**Problem Statement:**

Yelp has an enormous size of data repository for businesses and services across the United States and worldwide. It categorizes and captures important information regarding the business, from its products and services, location, and most importantly the reviews.

Our objective is to design a program that will use existing yelp data to provide insightful analytics and help customers to find the right place for them as well as existing and future business owners to make important decisions regarding new business or business expansion.

**Data Source:**

We use the dataset provided by Yelp on their site, which consists of different datasets in.json format. Here is the link: Yelp Dataset: <https://www.yelp.com/dataset/download>

Our group mainly uses the business.json dataset for its comprehensive information that includes 61 columns. We use tip.json for analyzing user short reviews. The 2 json files which we used in the project were first converted to csv files using converter code. The json file which needs to be converted has to be manually entered in “json\_file” variable in .py file.

Converter code - <https://drive.google.com/file/d/1IWT5VIlluEPvRa-lXMNUPc5GRIxI4US-/view?usp=sharing>

Data source can be found at –

https://drive.google.com/drive/folders/1Kt6tQ\_OemQBH1tggs7WzGfpelpwWujNn?usp=sharing

**Installation instructions**

Libraries used

1. pip install seaborn
2. pip install geopy
3. pip install scikit-learn
4. conda install plotly
5. conda install geopanda

Additional Files

Following files need to be placed before execution -

1. Yelp.png 2. user\_visit\_data.csv

Execution order

Any .py can be executed first. There is only 1 dependency between GUI.py and Feature-v3.py in which GUI.py creates a file which is used in Feature-v3.py.

**Features:**

Yelp’s Dataset Summary Feature - (Summary\_GroupProject\_v2)

The summary gives the users a big picture of what is the dataset about and what are some of the features we are interested in. To summarize it, we firstly cleaned the data by dropping empty or invalid rows in the most important columns, for keeping most of the data possible. Based on the further study on locations and the restaurants, we totaled the number of cities and states.

The purpose of restaurant recommendation feature is to give users the top five restaurants in terms of rating and number of reviews, in the user’s location. The key of the location is the zip code. To rank the restaurants, the rating priority is above the number of reviews. For instance, with the same rating, the restaurants will be ranked based on the number of reviews. This is because we believe that the rating is a reference of the restaurant's quality overall, while the number of reviews is a measure of the popularity. This feature further shows how it works by giving charts after user inputting their postcode.

Yelp’s User Tip Analysis for Businesses - (feature 2)

This feature allows business owners to get a quick and useful insights from the customers’ short reviews called tips. The program basically predefined the 12 different words that would be used as a kind of a vocabulary to look up for throughout the customers’ tips and see the frequency of their occurrences. These 12 words include wait, awesome, great, fantastic, amazing, love, horrible, bad, terrible, awful, wow, and hate. Some of them represent positive comment and some negative, while the word ‘wait’ may alert business owner if there’s a concern regarding waiting time. First, the user will enter the name of their business and the program will generate all different business ids with that name and prompts the user to input which business branch they’re looking for. The program then will output a table with the 12 predefined words and the count of each along with a graph for more simplified comparison.

Yelp Restaurant Finder Feature -(GUI)

The GUI allows the user to choose three features they are looking for in a restaurant and the program will output the closest five restaurants with the user’s desired features. There is no text formatting enforced in the address text box but essentially the input should first list the address and then the state in order for the program to work, no commas are necessary (i.e. 1515 Broadway NY). Clicking the “Find Restaurants” button will run the program and clicking the “Reset” button will clear the drop-down boxes and text fields. When the program is executed, it first creates a dataset with only restaurants by filtering the “category” field within the original dataset for any text that contains “food” or “restaurant”. Next, as the user enters their desired restaurant features, the data will be filtered for any restaurants where the feature is set to “True”. There are three filters applied to the dataset for each of the user’s selected features to ultimately reach a list of restaurants that each contain the three desired features. A longitude and latitude is calculated based on the user’s address and then a “Distance” field is created in the dataset which captures how far each restaurant is from the location of the user. The dataset is then sorted in an ascending order and the top 5 closest restaurants are recommended to the user in the GUI output.

Yelp’s Notification Feature -(Feature 4 -v2)

This Feature allows Yelp to send out notification regularly to its current user and predict the business a user might like. We made a user information table called “user\_visit\_data.csv”. You need to input a user’s name (i.e. Ann / Susan / Joe) to print out recent notifications, in order to search a business that is most similar to a user’s recent visit business.

Average Restaurant Rating by state and restaurant chain -(Feature-v3)

The average user ratings of all restaurants is calculated for each US state and displayed on US map. This feature also asks users to input chain restaurant which might be in multiple states. Data is filtered for that chain restaurant and average rating is calculated for each state. Bar graph is plotted which shows states with lowest rating to highest rating.