DDR Final project Report

Does the feature data of each restaurant in Yelp affect its ranking?

Doris Pu

Krittika

Sumashree Javaji

03.20.2024
BAX 422 001

Executive Summary

Project Purpose: Our project primarily involves scraping data from Yelp for restaurants in San Francisco. Based on the data we scrape, we aim to determine whether certain characteristics of restaurants, such as ratings, number of reviews, and price, influence the ordering of search results.

Methodology: We mainly use Python for data scraping. We employ Selenium to perform a series of clicks and inputs to obtain the list of restaurants we want. Then, we use Beautiful Soup to read and save the HTML of each restaurant for later analysis. Finally, we store the data in an SQL database. In addition, we use certain regular expressions to manipulate the data. The Yelp website also has its own API, which we didn't use for this project, but we can quickly add data if we need it for subsequent analysis.

Analysis: We ensure that our data structure is suitable for business use. We convert the data into a dataframe, making sure each row represents the data for one restaurant. We encode categorical data to facilitate its use and analysis in our ML project later on. Subsequently, we perform further data cleaning, visualization, and analysis using methods like Random Forest in our ML project.

About Yelp

Yelp is an influential online platform that connects millions of users with local restaurants and other stores, making Yelp a key reference for consumers when dining out, shopping and more. The platform has a high level of user engagement, with users actively providing reviews, ratings, and photos, which continue to enrich Yelp's content and reliability. Restaurants listed on Yelp increase visibility and are seen by more users.

Against this context, our project aims to analyze Yelp data to gain insight into how restaurant characteristics such as ratings, number of reviews, and price level affect their visibility and ranking in search results. Understanding these dynamics has enormous business value, providing businesses with the strategic understanding to optimize their online presence and performance on Yelp. This analysis is not only critical for individual restaurants seeking to improve their rankings and attract more customers, but also provides broader market insights that can inform industry-wide strategies for success on the platform.

Introduction of the Data

Our dataset comprises 12 columns and 100 rows, featuring data from 100 different restaurants. The data includes the restaurant name, rating (out of a maximum of 5 points), number of reviews, cuisine type, price level, “Free Wi-Fi”, “Vegetarian”, “Street Parking”, “Takes Reservations”, “Dogs Allowed”, and “Good for Groups”. For the features from “Free Wi-Fi” to “Good for Groups”, the data is binary, with 0 indicating the absence of a particular amenity and 1 indicating its presence. The price level is categorized into $, $$, $$$, and Null, with the number of symbols($) indicating the restaurant's price level and Null signifying that the restaurant has not disclosed its prices.

Yelp displays 10 restaurants per page, and we selected data from the first nine pages, totaling 100 restaurants. This selection was made based on the practical assumption that fewer people are likely to browse beyond 10 pages of content. This consideration helps us focus our analysis on the most visible and potentially influential establishments on the platform, providing valuable insights into factors affecting restaurant visibility and customer choice on Yelp.

Detailed Methodology

STEPS:

1. **Using selenium to get to yelp page**

Go to "yelp.com"

Click on the "restaurant"

Type in "San Francisco, CA" for the region

2. **Storing html files to local**

Get the url of each restaurant

Concatenate partial url with base url to get full url

Get restaurant name as file name using regular expression

Save the html file to local

3. **Parsing and Displaying Information from Saved HTML**

Use beautifulSoup to parsing and read each file

Get all the information we needed

Converting data into a dataframe

Export csv file for further analysis

4. **Store data into a database**

Connect python and mysql

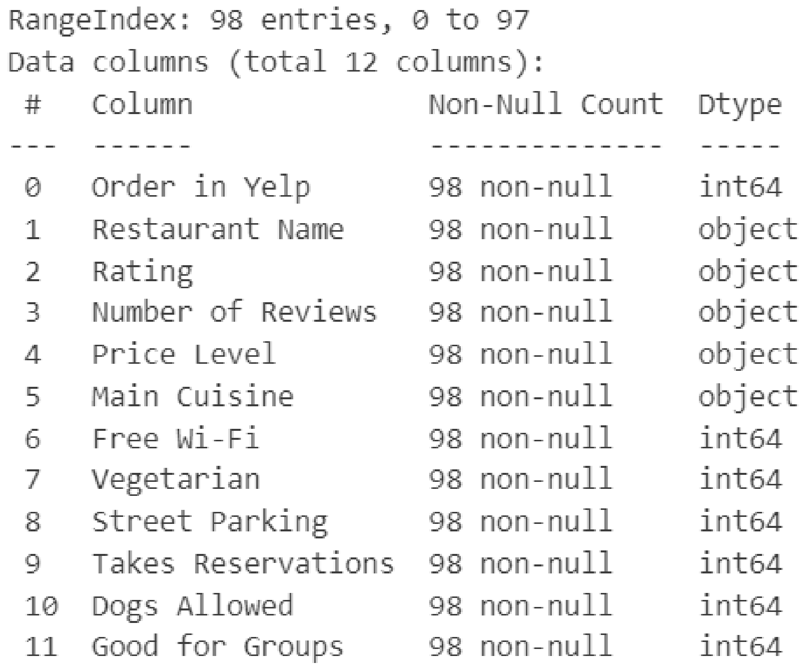
Create databases, tables, and data types

Specify the primary key

Store data in a table

Add index to specific columns

DATA Info



```
RangeIndex: 98 entries, 0 to 97
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Order in Yelp       98 non-null     int64
 1   Restaurant Name     98 non-null     object
 2   Rating              98 non-null     object
 3   Number of Reviews   98 non-null     object
 4   Price Level         98 non-null     object
 5   Main Cuisine        98 non-null     object
 6   Free Wi-Fi          98 non-null     int64
 7   Vegetarian          98 non-null     int64
 8   Street Parking      98 non-null     int64
 9   Takes Reservations  98 non-null     int64
 10  Dogs Allowed        98 non-null     int64
 11  Good for Groups     98 non-null     int64
```

Database and Business Insights

We utilize SQL as our database for storing data. Although we did not use a relational database for the data scraped in this project, we considered the potential business applications that may require adding additional information, such as specific reviews and images for each restaurant. This data is expected to grow continuously. Using a non-relational database to store such data could result in significant redundancy, making data addition, deletion, and analysis less efficient. Therefore, we opted for a relational database and employed indexing to make querying and analysis faster and more effective. This approach allows for efficient data management and scalability, ensuring our database can accommodate the evolving needs of our analysis and potential business applications.

CONCLUSION

Our project focuses on utilizing the methods learned in class to gather data needed to address our proposed business question. Moreover, we can easily obtain more or different data by adjusting the region and the number of restaurants. Our steps ensure the quality, authenticity, and up-to-dateness of the data, as well as prepare for future analysis and data modification. This approach guarantees that our project is not only tailored to our current research needs but is also adaptable for future expansions and analyses.

REFERENCES

Yelp - Company - Fast Facts, www.yelp-press.com/company/fast-facts/default.aspx. Accessed 21 Mar. 2024.

Team, Dcf. “Yelp Inc. (YELP): History, Ownership, Mission, How It Works and Makes Money.” Dcf-fm, 28 Nov. 2023, *dcf.fm*/blogs/blog/yelp-history-mission-ownership.