Data Curation Project: Preliminary observations of the Yelp Dataset

Dataset Name

Yelp Dataset (from Kaggle) - https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/data

Data Understanding and Quality Issues:

- provide descriptive statistics for numeric attributes.
- identify missing values.
- check for duplicates.
- identify any anomalies in the data (businesses with an extremely high or low number of reviews).

Data Cleaning:

- Handle missing values (impute or remove).
- Handle duplicates (remove).
- Normalize text attributes (convert all text to lowercase).

Data Integration:

- integrate two datasets (Yelp reviews) based on common attributes (**business_id**).
- Perform schema matching and mapping if required.

Data Transformation (ETL Processes):

- Extract relevant attributes for analysis.
- Transform attributes (extract the primary category from the **categories** column).
- Load into a structured form suitable for querying or analytics.

Data Analysis:

- Identify the top-rated businesses.
- Find the average rating per city or state.
- Analyze the distribution of businesses across different categories.
- Explore the relationship between the number of reviews and ratings.

Data Provenance:

- Document the source of the data.
- Track any changes or transformations made to the data.
- Store metadata or information about data processing steps.

Visualization:

- Plot the distribution of ratings.
- Visualize the number of businesses in different categories.
- Map the businesses based on their latitude and longitude to visualize their distribution.

Other/Alternative dataset - https://github.com/bookingcom/ml-dataset-mdt [Booking.com]