Yelp – What makes a highly rated restaurant?

CSPB 4502

Group 3: Eric Hanley

Project Description

This project will analyze an open dataset of restaurant reviews to understand restaurant trends, investment opportunities, and commonalities across successful restaurants. In addition to restaurants, I may evaluate users to identify trend-setters – those reviewers that highlight subsequently popular restaurants first.

- What combination of features results in an above-average rating?
  - Does this vary by region?
  - Does this vary by restaurant type?
- What type of new restaurant should an entrepreneur open with likelihood of success in a given city?
  - Are there are any undersaturated markets?
- What trends in restaurant types have occurred through time?
  - Do they differ by region?
- Which users are trend-setters? Who identifies up-and-coming restaurants?

Prior Work

- There are currently 211 notebooks on Kaggle that reference this dataset https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset
- This dataset was the basis for a student predictive modeling project for Stanford's CS229 http://cs229.stanford.edu/proj2016/report/NortzMallard-UsingYelp%20ReviewsToImproveBusinesses-report.pdf
- A quick Google search turns up multiple projects on Github based on this dataset https://www.google.com/search?q=yelp+dataset+projects&oq=yelp+dataset+projects&aqs=chrome..69i57j0i390l2.6589j0j4&sourceid=chrome&ie=UTF-8

Datasets

- The data for this project consist of three tables:
  - Business – metadata about restaurants (and other businesses)
  - Review – individual reviews of businesses
  - User – metadata about the users writing the reviews
- There are other tables in the dataset, but they will not be used in the scope of this project
- Data source: https://www.yelp.com/dataset
- These data are downloaded on my local machine

Proposed Work

- **Cleaning:** An initial review of data suggests they are mostly clean and standardized; however some fields in the Review data are free text and will likely require some cleaning to make them useful.
- **Preprocessing:** The data will have to be filtered to only include restaurants. Additionally, some characterization of the data will be necessary: categorizing restaurants by type, summarizing or extracting sentiment from review text (potentially), creating time series data from timestamped reviews
- **Integration:** The data consists of multiple tables in JSON format that will have to be read in and joined

Proposed Tools

- I haven't fully decided yet which tools I am going to use, but I am leaning toward BigQuery, Hex, or Palantir Foundry
  - Alternatively, I may spin up a "Modern Data Stack" with a combination of tools like Snowflake/BigQuery, DBT, Airflow, and some analytical + viz tooling on top of the integration/transformation layer. (this would be done for the sake of learning a new stack while doing this project)
- Initial exploration will likely occur on my local machine in Jupyter (Python)
- Production-scale analysis will be conducted in a cloud environment that offers analytical tooling, a repeatable build pipeline, and visualization.

Project Evaluation

- While there is no explicit ML/AI component scoped for this project, I will likely create a train/test split in the data to evaluate if the proposed questions can be successfully answered with the project analysis
- If I can find other analyses of the data on Kaggle that are comparable in scope, I may attempt to compare outcomes
- This project will also be peer evaluated