Market Research & Business Opportunities Based on Yelp Data

Team #105: Michael A Fronda, Hal Zhang, Vladimir Orlov, Shivani Narahari, Abhishek S Mishra, Anthony J Mazza

Introduction

Market research is crucial for the successful growth of businesses. An essential problem of market research is understanding **the relationship between supply and demand**. Our project's goal is to estimate the supply and demand of service categories in a specific geographical area using the **Yelp dataset**. These estimates will then help us uncover potential economic opportunities.

Those **who care about our project** are businesses looking to expand, local entrepreneurs and investors. Derived market research insights can aid current and future businesses and local communities in the selected area to make data-driven business decisions.

Our Data

Our project is based on the officially released **Yelp dataset** [*https://www.yelp.com/dataset*], which encompasses information on businesses, customers, customer's reviews, tips, and check-ins records. It was downloaded in the form of archived JSON files ranging **in size** from **118Mb** to **5.26Gb** and **in number of records** from **150,346** to **13,356,875** (please check the related diagram in the left for additional details). As such, our dataset is **static**, describing social activity around businesses in several North American metro areas in the time range from **February 2005** to **January 2022**.

Our Approach

While services like *Yelp*, *Google Maps*, and *OpenTable* manage and track relationships between businesses and their customers, using the same data to infer business opportunities is not a publicly available feature in the current market. **The novelty of our approach** is in **connecting the domains of social networking and market research.** We do this by extracting and measuring market demand based on customer feedback records and evaluating market supply based on existing businesses and their popularity.

Our intuition is that the number of social feedback records (reviews, tips, check-ins) submitted for a particular business effectively scores the **total demand** for this business. Part of these records with low ratings and/or implying negative sentiment represent the **unsatisfied demand** for the same business.

$$DT = \Sigma_{i=1}^{n}(1 + fr_i + cr_i + ur_i) + \Sigma_{j=1}^{m}(1 + ct_j) + chn,$$

$$DU = \Sigma_{i=1,(sr_i \le 3)}^{n}(1 + fr_i + cr_i + ur_i),$$

$$S = DT - DU;$$

DT, ***DU***, ***S*** - demand, unmet demand & supply scores; ***n***, ***m***, ***chn*** - number of reviews, tips & check-ins; ***sr*** - review star rating; ***fr***, ***cr***, ***ur*** - number of *funny*, *cool* and *useful* review tags; ***ct*** - number of tip compliments.

These scores, rolled up by business categories and geographical locations, give us a holistic view of potential business opportunities and market gaps. We also provide proper visualizations to clearly communicate the insights produced. Additionally, we leverage user review records to gauge consumer sentiment and preferences because it adds a qualitative assessment to our project and offers a more nuanced understanding of the market state.

Our Experiments & Results

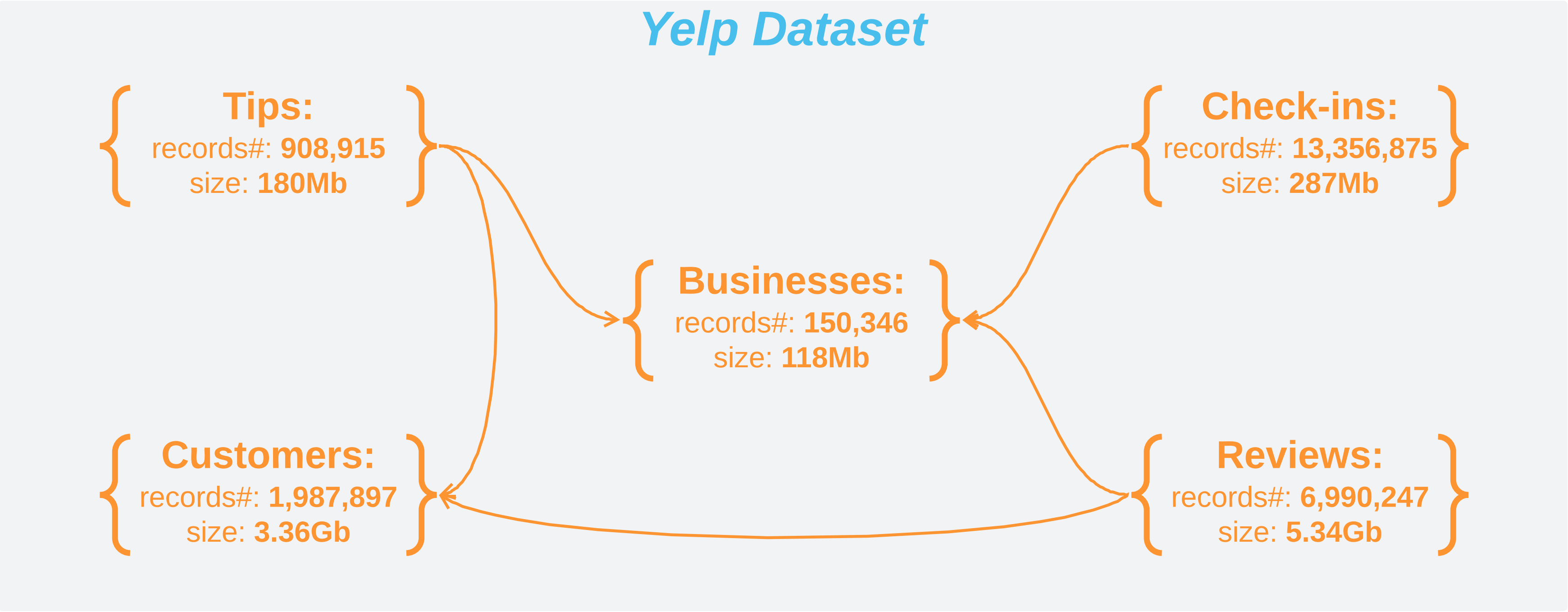
We split our dataset into businesses opened before 2020 and opened in 2020 and after. We calculated the missing demand score for businesses opened before 2020 aggregated by zip code and business category and compared the ranking of this list with the ranking of the number of businesses opened in 2020 and after also aggregated by zip and business category. This allowed us to see if a higher missing demand score in a specific zip code and business category corresponded to a higher number of new businesses opened. The Kendall Tau test statistic we calculated for **Philadelphia** businesses was **3.77e-31**, a very strong correlation. We also wanted to externally validate our missing demand score so we used the same experiment on businesses in **New Orleans** in our **Yelp dataset**. The Kendall Tau test statistic was **6.02e-10**, also a very strong correlation.

Based on the two experiments we conducted above, **the results are** that missing demand score calculations strongly correlate with the number of new businesses opened in the same zip code and business category, therefore they provide a good insight into future business opportunities. The final outcome of our project is the visualization dashboard that provides user-friendly access and navigation over the demand scores computed over geographic locations and business categories supported by sentiment analysis details.

There are no exact studies like ours that we can compare our results to but according to Camillo et. al [1], businesses that succeed open in areas with high demand. Lu et. al [2] tried to predict which businesses would stay open with an accuracy of **67.46%** and found that the restaurant being a chain helped it stay open while review count (popularity) did not. Our methods differ from these two studies which is why we think our method can generate additional insights.

Works Cited

1. **Camillo, A. A., Connolly, D. J., & Woo Gon Kim. (2008).** Success and Failure in Northern California: Critical Success Factors for Independent Restaurants. Cornell Hospitality Quarterly, 49(4), 364-380. https://doi.org/10.1177/1938965508317712
2. **Lu, Xiaopeng & Qu, Jiaming & Jiang, Yongxing & Zhao, Yanbing. (2018).** Should I Invest it?: Predicting Future Success of Yelp Restaurants. 1-6. 10.1145/3219104.3229287.



Yelp Dataset

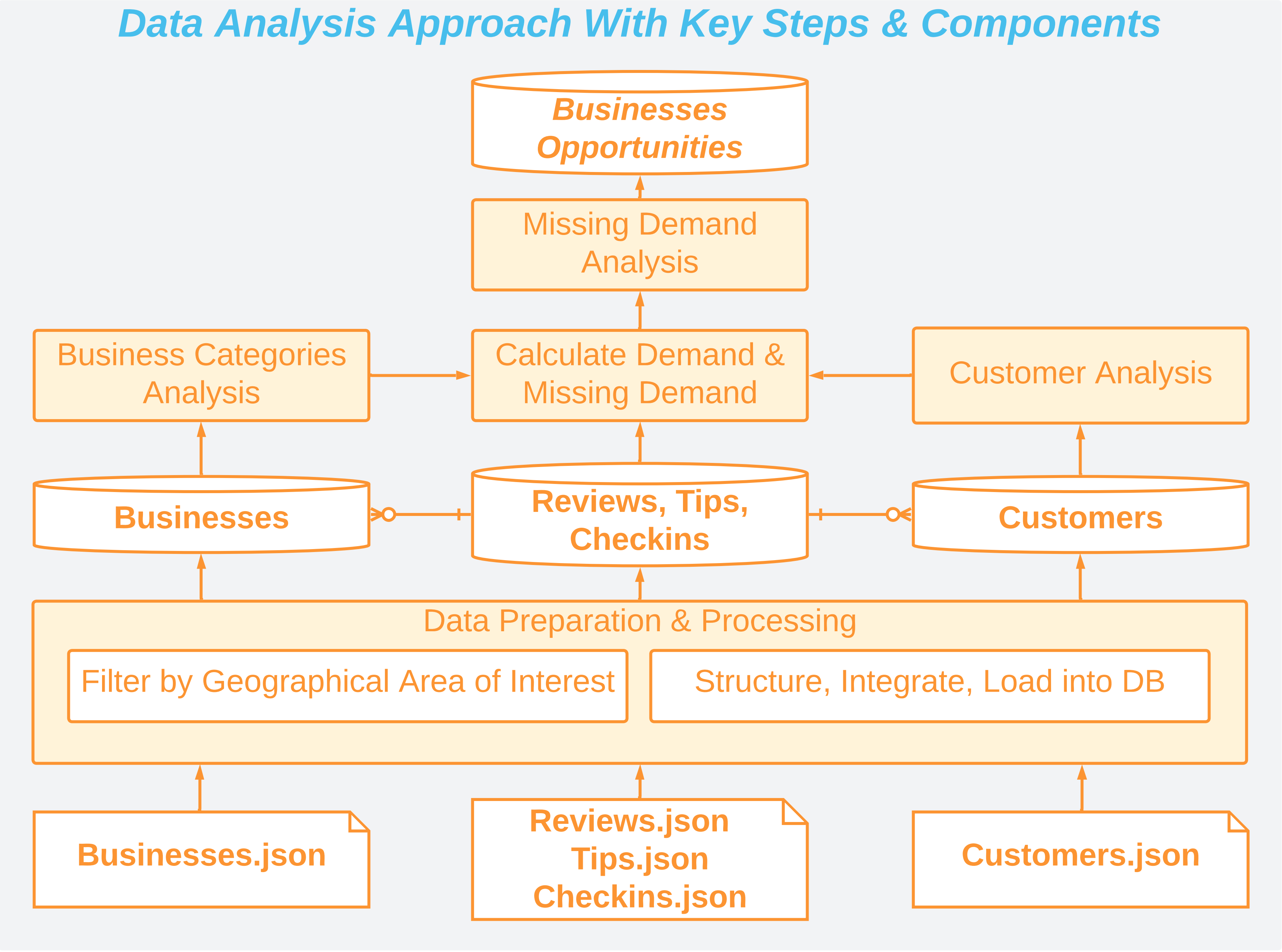
Tips: records#: **908,915** size: **180Mb**

Check-ins: records#: **13,356,875** size: **287Mb**

Businesses: records#: **150,346** size: **118Mb**

Customers: records#: **1,987,897** size: **3.36Gb**

Reviews: records#: **6,990,247** size: **5.34Gb**



Data Analysis Approach With Key Steps & Components

Businesses Opportunities

Missing Demand Analysis

Business Categories Analysis

Calculate Demand & Missing Demand

Customer Analysis

Businesses

Reviews, Tips, Checkins

Customers

Data Preparation & Processing

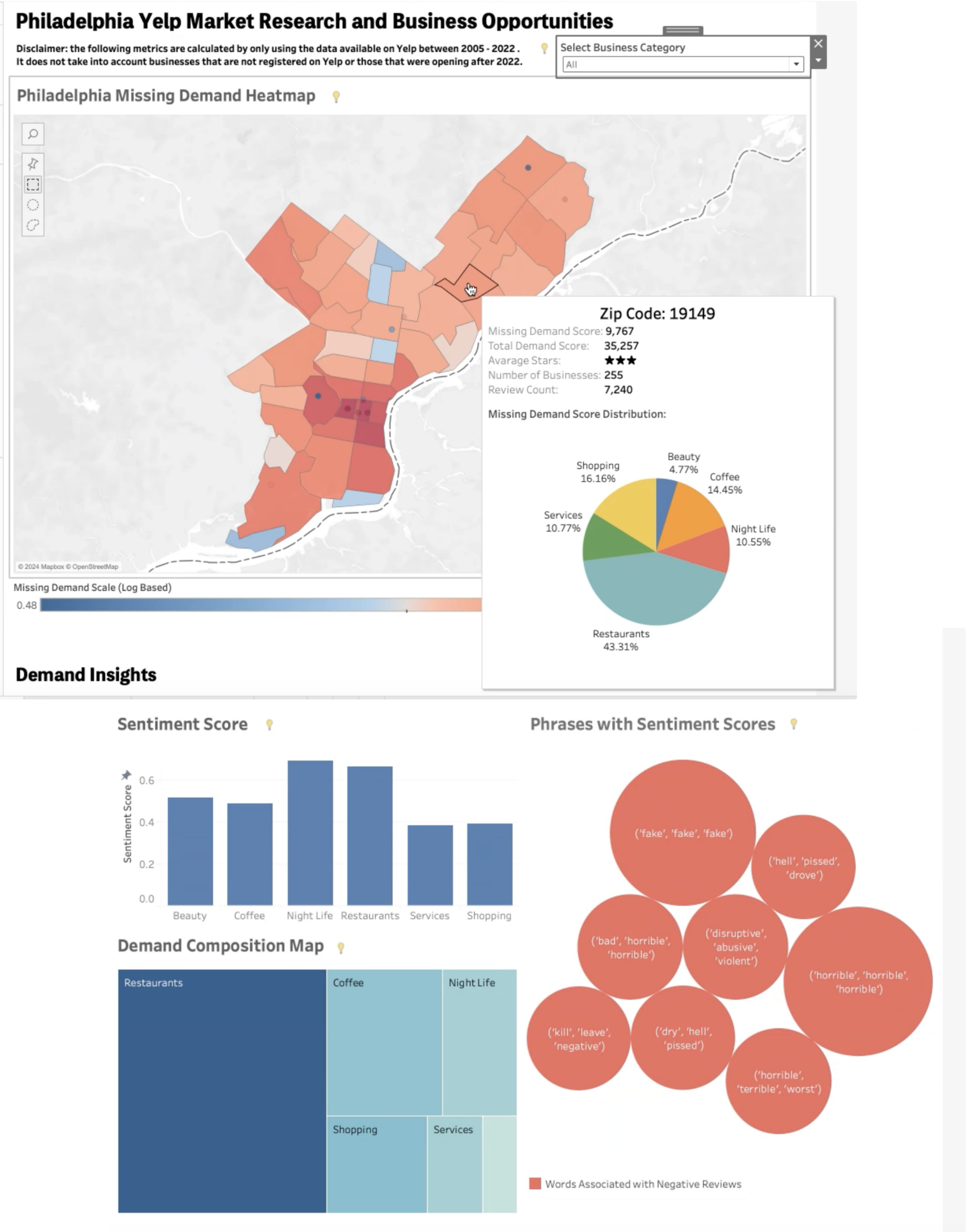
Filter by Geographical Area of Interest

Structure, Integrate, Load into DB

Businesses.json

Reviews.json Tips.json Checkins.json

Customers.json



Philadelphia Yelp Market Research and Business Opportunities

Disclaimer: the following metrics are calculated by only using the data available on Yelp between 2005 - 2022 . It does not take into account businesses that are not registered on Yelp or those that were opening after 2022.

Select Business Category: All

Philadelphia Missing Demand Heatmap

Zip Code: 19149
Missing Demand Score: 9,767
Total Demand Score: 35,257
Avarage Stars: ★★★
Number of Businesses: 255
Review Count: 7,240

Missing Demand Score Distribution:

Demand Insights

© 2024 Mapbox © OpenStreetMap
Missing Demand Scale (Log Based)
0.48
Shopping 16.16%
Beauty 4.77%
Coffee 14.45%
Services 10.77%
Night Life 10.55%
Restaurants 43.31%
Sentiment Score
Sentiment Score
0.6
0.4
0.2
0.0
Beauty
Coffee
Night Life
Restaurants
Services
Shopping
Demand Composition Map
Restaurants
Coffee
Night Life
Shopping
Services
Phrases with Sentiment Scores
('fake', 'fake', 'fake')
('hell', 'pissed', 'drove')
('bad', 'horrible', 'horrible')
('disruptive', 'abusive', 'violent')
('horrible', 'horrible', 'horrible')
('kill', 'leave', 'negative')
('dry', 'hell', 'pissed')
('horrible', 'terrible', 'worst')
Words Associated with Negative Reviews