Market Research & Business Opportunities Based on Yelp Dataset

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 component 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 [https://www.yelp.com/dataset], which encompasses information on businesses, customers, customer's reviews, tips, and check-ins records. These estimates will then help us uncover potential economic opportunities.

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 to fill this void by analytically transforming the available data into an intuitive and accessible format and enabling stakeholders to easily identify potential business opportunities and market gaps. Additionally, we leverage user review records to gauge consumer sentiment and preferences. It adds a qualitative assessment to our project and offers a more nuanced understanding of the market state.

Our methodology employs natural language processing (NLP) techniques to infer consumer preferences and satisfaction levels using user-generated content and ratings. We use clustering for extracting business categories and studying customer base. For the given dataset, we quantify services' supply and demand and render them as an interactive dashboard, highlighting areas with demand-to-supply gaps. The dashboard can help entrepreneurs, investors, and policymakers understand localized unmet demand and influence their strategic decision-making process.

Literature Survey

We drew inspiration for our methodology from various scholarly works. Cormen et al. [5] gives a mathematical foundation for graph analysis methods to use for our data analysis. Geron [6] provides a solid introduction into machine learning models, techniques, and tools to analyze our dataset. Tayeen et al.'s [11] explores the impact of location data on business review performance, helping us investigate how such data influences demand.

Singh et al.'s [1] work with Yelp data using *AzureML* and *SparkML* and Luo et al.'s [17] work with sentiment analysis suggests how to translate text analysis into insights on supply and demand. While Brath et al. [3] discuss the business implications of graph analysis, we delved deeper into the analytical aspect. Glaeser et al. [10] findings on Yelp data prompted us to focus our analysis on the city of Philadelphia, an urban area with a high business density. We consider building graphs of customers and businesses, refining McClanahan et al.'s [2] centrality analysis plus natural language processing (NLP).

We used both Louvain and Leiden community analysis on our customer and business graphs to find customer clusters and influencers [16]. Elbaghazaoui et. Al [20] suggests *PageRank* to identify influencers in addition to calculating centrality using *Networkx* in Python.

We explored various clustering methods suggested by Bishop et al. [4] and Sinaga et al. [19]. We got more meaningful results with *K-Means* when applying the algorithms to the Yelp dataset. We also considered clustering and similar techniques to derive quantitative insights from reviews, like Deho et al.'s [15] word embedding approach. We aim to enhance Jiang et al.'s [14] *word2vec* application by shifting from clustering to sentiment classification, further refining Mutinda et al.'s [13] *lexicon* and *CNN* methods to focus on Yelp reviews exclusively.

We assessed performance of various models and developed a market demand composite score formula in accordance with Li et al. [7] and Hu et al.'s [8] research on the impact of online reviews on demand, augmented by considering the volume of reviews and the influence of popular reviewers.

Evaluation of demand predictions is informed by time series analysis, drawing inspiration from Pereira et al.'s [9] work on hotel demand, but adapted to Yelp dataset.

The culmination of our project is an interactive heatmap visualization of supply and unmet demand scores, an approach that builds upon Wang et al.'s [12] *E-comp* map and extends it to include predictions of business opportunities. Ben Jones' [18] was also very instrumental in creating project visualizations.

Proposed Method

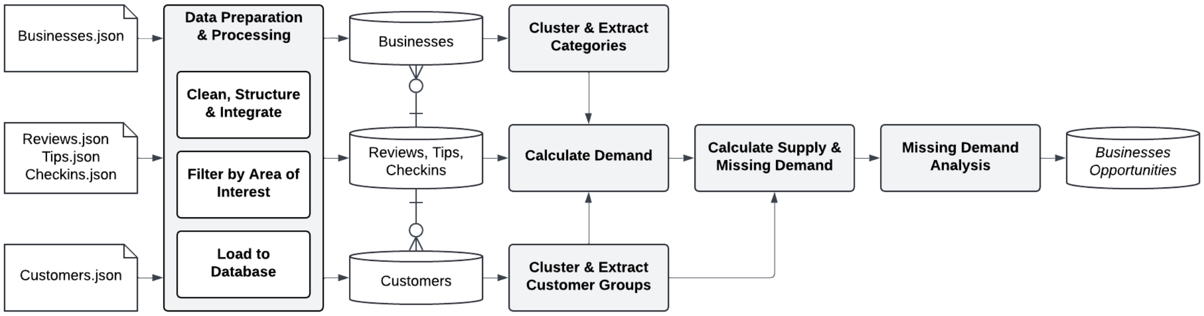


Figure 1. Analysis method proposed with all key steps and components.

Data Preparation & Processing

The Yelp dataset comes in the form of JSON files and is not convenient for subsequent analysis. At this stage, we want to restructure our data; clean any records that might have missing fields, conflicting unique ID's, broken links; integrate key dataset entities using database foreign key relationships; and finally get our data hosted in a relational database. Once our data has proper structure, field data types, indices, and relationships, we can unleash the power and fluency of SQL to manipulate it and perform subsequent analysis.

Our team decided to use PostgreSQL database as a key platform for managing the original dataset, any intermediate and final results, and analytical outcomes derived from it.

Cluster Businesses & Extract Categories

We began the analysis by obtaining a list of all business categories in Philadelphia. There were over 1000 categories, and each business can have more than one category. For the scope of the project, we needed to come up with a condensed list of overarching categories to aggregate and group businesses.

TF-IDF and *K-Means* clustering helped us to find relevant business categories. This resulted in two versions for additional analysis with aggregations up to 6 and 4 categories respectively

Cluster Customers & Extract Customer Groups

We used Louvain and Leiden community analysis to cluster customers based on businesses they reviewed. Preliminary findings show that customers tend to visit businesses around a set zip code rather than businesses that are similar in nature. We will also investigate customer and customer relationships by using centrality analysis to determine influencers and adjust our demand and unmet demand scores based on the amount of influence a customer has.

Calculate Demand, Supply, and Missing Demand

Demand for an individual business is measured based on customer reactions to this business such as reviews. We are considering adding other types of customer reactions like tips and check-ins to this formula. Each review record comes along with customer ID, his/her textual comment, star rating and *funny/cool/useful* tag counters. Reviews with star rating 3 and below inform unmet demand. Supply is represented as the difference between demand and unmet demand. These scores can then be aggregated by business category and/or zip code to calculate supply and demand across different business types and geographical areas.

$$DS = \sum_{i=1}^{n} \left(wu_i \cdot (1 + fr_i + cr_i + ur_i) \right)$$

$$DU = \sum_{i=1,\,(sr_i \le 3)}^{n} \left(wu_i \cdot (1 + fr_i + cr_i + ur_i) \right)$$

$$S = DS - DU$$

DS, *DU*, *S* - demand, unmet demand & supply scores correspondingly;
n - number of reviews;
wu - *customer* influence weight;
sr - review star rating;
fr, *cr*, *ur* - number of *funny*, *cool* and *useful* review tags correspondingly.

Missing Demand Analysis & Finding Business Opportunities

We preprocess the text by removing stop words and unwanted text formatting in the reviews. Then, we performed lemmatization to reduce inflectional forms to a common base form of each word. Visualization of text was done using *wordclouds*, separated by average star rating. We are considering further segmenting of *wordclouds* to give us an idea of text ratings.

Design of Upcoming Experiments

There are two ways that we can test if unmet demand we calculated leads to business opportunities in the future. The first way is taking advantage of the time series nature of the Yelp data containing reviews from 2005 to 2022. We can calculate the unmet demand of businesses that have reviews from 2005 to 2015 and use this to provide recommendations of business opportunities in 2015 to 2022. Then we can compare our recommendations to new businesses that have started in that period (businesses that only have reviews from 2015 on). Secondarily, we can test our analysis by calculating unmet demand with all the data we have and compare this to pattern in new businesses in the past two years (2023 and 2024) using Yelp or Google Maps. We plan on testing both approaches and choosing the most appropriate one.

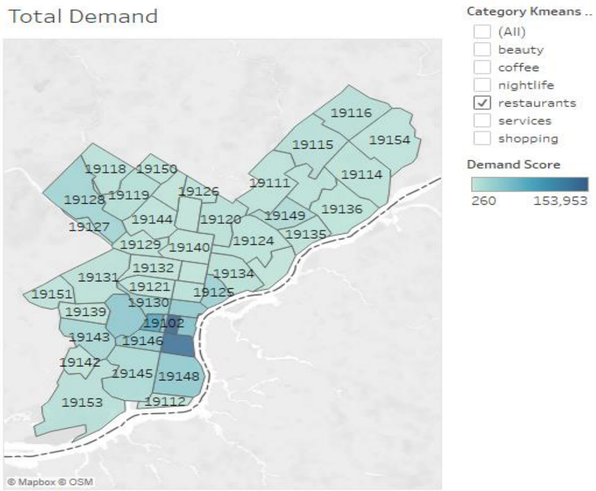
We recognize that both experiments will not provide objective accuracy metrics for the calculated unmet demand score, but they test our tool’s objective - to highlight potential business opportunities.

Results

Under construction...

Tableau interactive heatmap dashboard will show calculated total and unmet demand scores different Philadelphia zip codes. There will be the option to filter the scores by different business categories of interest.

For each Philadelphia neighborhood we will let dashboard user to drill down into details of the demand unmet there plus specific business opportunity suggestions supported by the natural language processing insights targeted for this city location plus business categories of interest.



Category Kmeans ..

- □ (All)
- □ beauty
- □ coffee
- □ nightlife
- ☑ restaurants
- □ services
- □ shopping

Figure 2. preliminary heatmap dashboard that will be updated as we finalize our analysis

Conclusion

Under construction...

Plan of Activities

All team members contributed equally to the project

Project Proposal (old)

Task	Start Date	End Date	Team Member(s)
Project Proposal	2/13/2024	2/21/2024	All
Project Literature Survey	2/21/2024	2/26/2024	All
Proposal Slides and Video	2/26/2024	2/29/2024	Anthony, Michael
Data Processing	3/01/2024	3/09/2024	Vladimir, Shivani
Modeling	3/09/2024	3/27/2024	Abhishek, Hal
Visualizations	3/27/2024	4/10/2024	Shivani, Michael, Anthony
Final Deliverables	4/10/2024	4/19/2024	All

Revised (new)

Task	Start Date	End Date	Team Member(s)
Data Preparation & Processing	3/01/2024	3/09/2024	Vladimir
Cluster & Extract Business Categories	3/13/2024	3/20/2024	Michael
Cluster & Extract Customer Groups	3/29/2024	4/07/2024	Hal
Calculate Demand	3/03/2024	3/25/2024	Vladimir, Anthony
Project Proposal & Progress Report	3/16/2024	3/30/2024	All
Calculate Supply & Missing Demand	3/23/2025	4/10/2024	Vladimir, Anthony, Michael
Results Evaluation	3/30/2024	4/10/2024	Anthony, Shivani
Interactive Dashboard(s)	4/01/2024	4/10/2024	Shivani, Hal
Final Deliverables	4/10/2024	4/19/2024	All

Citations

1. Singh, R., Woo, J., Khan, N., Kim, J., Lee, H. J., Rahman, H. A., ... & Gudigantala, N. (2019). Applications of machine learning models on yelp data. *Asia Pacific Journal of Information Systems*, *29*(1), 117-143.
2. McClanahan, B., & Gokhale, S. S. (2016). Centrality and cluster analysis of Yelp mutual customer business graph. In 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC) (pp. 592-601). https://doi.org/10.1109/COMPSAC.2016.79.
3. Brath, R., & Jonker, D. (2015). *Graph analysis and visualization: discovering business opportunity in linked data*. John Wiley & Sons.
4. Bishop, C. (2006). Pattern recognition and machine learning. Springer.
5. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to algorithms (3rd ed.). MIT Press.
6. Geron A. (2022) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 3rd Edition, O'Reilly Media.
7. Xiaolin Li, Chaojiang Wu, Feng Mai, The effect of online reviews on product sales: A joint sentiment-topic analysis, Information & Management, Volume 56, Issue 2, 2019, Pages 172-184, ISSN 0378-7206.
8. Hu, N., Liu, L. & Zhang, J.J. Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Inf Technol Manage 9, 201–214 (2008).
9. Luis Nobre Pereira & Vitor Cerqueira (2022) Forecasting hotel demand for revenue management using machine learning regression methods, Current Issues in Tourism, 25:17, 2733-2750, DOI: 10.1080/13683500.2021.1999397.
10. Glaeser, E. L., Kim, H., & Luca, M. (2017, November). Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity (Working Paper No. 24010). National Bureau of Economic Research. http://www.nber.org/papers/w24010.
11. Tayeen, A. S. M., Mtibaa, A., & Misra, S. (2020). Location, location, location! quantifying the true impact of location on business reviews using a Yelp dataset. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19) (pp. 1081–1088). Association for Computing Machinery. https://doi.org/10.1145/3341161.3345334.
12. Wang, Y., Haleem, H., Shi, C., Wu, Y., Zhao, X., Fu, S., & Qu, H. (2018). Towards easy comparison of local businesses using online reviews. Computer Graphics Forum, 37, 63-74.
13. Mutinda, James, Waweru Mwangi, and George Okeyo. 2023. "Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network" Applied Sciences 13, no. 3: 1445. https://doi.org/10.3390/app13031445.
14. Jiang, R., Liu, Y., Mentor, K., & McCann, B. (2015). A General Framework for Text Semantic Analysis and Clustering on Yelp Reviews.
15. Deho, O., Agangiba, A. W., Aryeh, L. F., & Ansah, A. J. (2018). Sentiment analysis with word embedding. In 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST) (pp. 1-4). IEEE. https://doi.org/10.1109/ICASTECH.2018.8506717.
16. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019 Mar 26;9(1):5233. doi: 10.1038/s41598-019-41695-z. PMID: 30914743; PMCID: PMC6435756.
17. Luo, Y., & Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. Sustainability (Basel, Switzerland), 11(19), 5254-. https://doi.org/10.3390/su11195254
18. Jones, B. (2014). *Communicating data with Tableau: designing, developing, and delivering data visualizations*. " O'Reilly Media, Inc.".
19. Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means clustering algorithm. IEEE Access, 8, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796
20. Elbaghazaoui, B. E., Amnai, M., & Fakhri, Y. (2022). Data profiling and machine learning to identify influencers from social media platforms. *Journal of ICT Standardization*, *10*(2), 201-218.