Analyzing the Influence of Geolocation on Business Ratings

Sathya Venugopal

1. Introduction

This study aims to analyze the impact of how the location of a business affects its ratings. In the current era of digital commerce and online reviews, a business's success is increasingly influenced by its online reputation, as reflected in user-generated ratings and reviews. While the quality of products or services offered is a primary factor in determining these ratings, this study hypothesizes that the geographical location of a business plays a significant role as well.

A business's location can affect various aspects of its operations and perception. Factors such as foot traffic, proximity to landmarks or transportation hubs, neighborhood demographics, and even local competition are likely to have a substantial impact on customer experiences and their subsequent reviews. Additionally, the socio-economic status of a neighborhood, urban versus rural settings, and cultural elements of a location might also play into how customers perceive and rate a business.

This research aims to delve deeper into the relationship between a business's geographical location and its online ratings. By leveraging a comprehensive dataset of business reviews, this study will employ predictive data analysis techniques to uncover patterns and correlations.

2. Dataset

The dataset for this study is sourced from the publicly available Google Local Reviews (2021) dataset, provided by Professor Julian McAuley at the University of California, San Diego. This comprehensive dataset provides an extensive collection of reviews, ratings, and metadata about businesses gathered from Google Local.

For the purpose of this analysis, the focus is narrowed down to the metadata about businesses located in California. This subset was chosen due to California's diverse economic landscape, encompassing a wide range of urban, suburban, and rural settings, which makes it an ideal candidate for studying the impact of location on business ratings.

California's status as a global economic powerhouse and a melting pot of cultures adds a layer of complexity to the study. The state's economy is not just driven by technology and entertainment but also encompasses significant agricultural, manufacturing, and tourism sectors. Each of these industries brings its own set of location-related factors that can influence business ratings. For instance, a tech company in Silicon Valley might be rated based on different criteria than a tourist resort in Napa Valley or a family-run vineyard in Sonoma County.

The dataset itself is a JSON file with each business being a JSON object. It contains 515,961 businesses. Here is an example of a single object from the data:


```
{
    "name": "University of California San Diego",
    "address": "University of California San
        Diego, 9500 Gilman Dr, La Jolla, CA
        92093",
    "gmap_id": "0x80dc06c4414caf4f:0
        xefb6aafc89913ea7",
    "description": null,
    "latitude": 32.8800604,
    "longitude": -117.23401349999999,
    "category": ["University"],
    "avg_rating": 4.5,
    "num_of_reviews": 936,
    "price": null,
    "hours": null,
    "MISC": {
        "Accessibility": ["Wheelchair accessible
            entrance"]
    },
    "state": null,
    "relative_results": [
        "0x80deaab8bc658bcf:0xd774e81608b43a68",
        "0x80d95686a6a04a21:0xc39f2ac6bf82f916",
        "0x80dc06c451841d31:0x940bc77e2afdb6c2",
        "0x80dc06c4414caf4f:0xe0acf21d827fd9ef",
        "0x80dc06c3689b4f99:0xdf55f97f07f34d4f"
    ],
    "url": "https://www.google.com/maps/place//
        data=!4m2!3m1!1s0x80dc06c4414caf4f:0
        xefb6aafc89913ea7?authuser=-1&hl=en&gl=us
        "
}
```


Listing 1: Example of JSON Data from the Dataset

Much of the fields in the data are not useful for the purposes of this study, so I filtered the dataset so each object only contains the `"name"`, `"address"`, `"latitude"`, `"longitude"`, `"category"`, `"avg_rating"` and `"num_of_reviews"`.

3. Dataset Properties

To better understand the nature of the dataset and what predictive tasks would be appropriate, I created multiple charts to better visualize the data. The first was a histogram, as this will allow us to see where the average ratings fall within the range and how they are distributed across the dataset. This visualization helps identify the central tendency, dispersion, and skewness of the ratings, providing insights into the ratings.

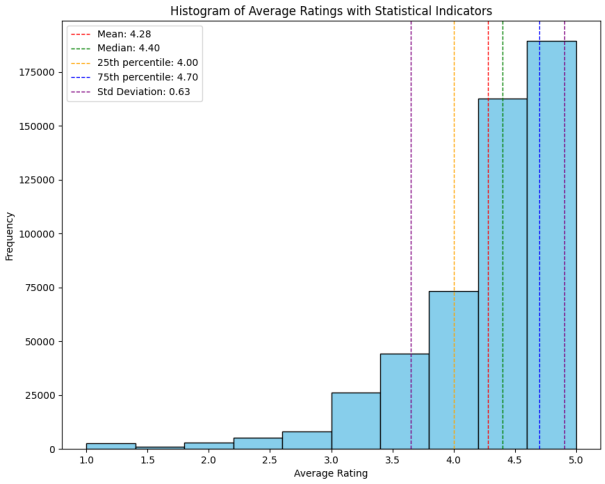

Figure 1: Histogram of average ratings showing statistical indicators

The histogram depicts the distribution of average ratings, demonstrating a significant skew towards higher ratings. The mean rating is 4.28, closely followed by a higher median of 4.40, which indicates a positive skew as more businesses are rated towards the upper end of the scale. The 25th percentile is at 4.00, and the 75th percentile is at 4.70, revealing that most businesses have ratings above 4.00, and a quarter of them score between 4.70 and 5.00. The standard deviation is relatively small at 0.63, signifying that most ratings cluster tightly around the mean. This distribution suggests that customers tend to leave positive feedback, with very few businesses receiving low ratings.

I also created a mapping of the category of business to average ratings, as this will help identify which types of businesses tend to receive higher customer ratings and may uncover industry-specific factors that influence customer satisfaction.

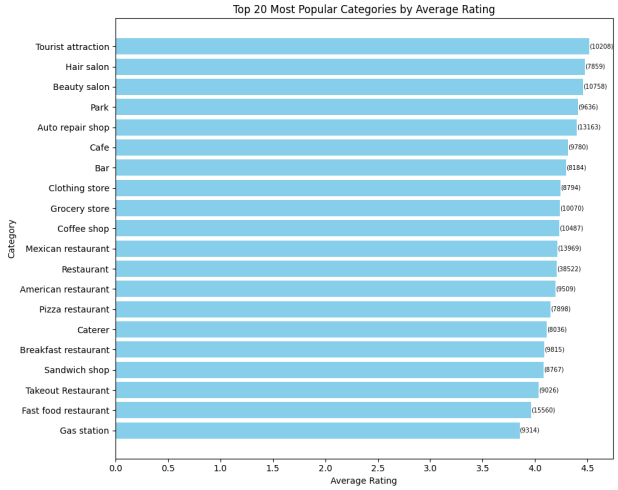

Figure 2: Bar chart showing the top 20 most popular business categories and their average ratings with their count in the dataset

The bar chart visualizes the average ratings for the top 20 most popular business categories, with the number of businesses in each category noted in parenthesis. The ratings range from 3.7 to nearly 4.5, with gas stations receiving the lowest average rating and tourist attractions receiving the highest. Notably, food-related categories, such as fast food restaurants, takeout restaurants, and pizza places, dominate the lower to middle range of the ratings spectrum, while personal care services, such as hair and beauty salons, and leisure-related categories, like parks and tourist attractions, score higher. This suggests that customers may have higher satisfaction with experiences related to personal care and leisure than with quick-service food options. The displayed counts indicate the relative popularity or frequency of reviews for each category, providing context for the average ratings.

Finally, I created a heat map showing the ratings against their latitude and longitude. This will reveal geographic trends and potential hot-spots of customer satisfaction, illustrating how ratings vary across different regions.

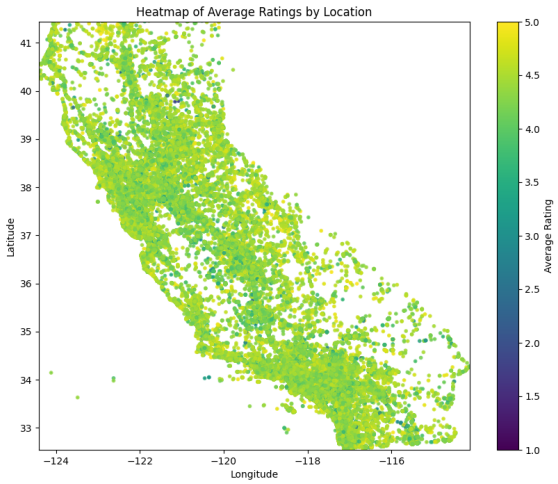

Figure 3: Heat map showing the average ratings by location

The heatmap displays the distribution of average business ratings across geographical locations in California. The latitude and longitude coordinates are plotted along the y-axis and x-axis, respectively, with the color intensity indicating the level of the average rating, as shown on the scale to the right. Warmer colors (yellow) represent higher ratings, while cooler colors (green to purple) represent lower ratings. The concentration of yellow points, particularly along the coastline and in the western part of the state, suggests that businesses in these areas tend to have higher average ratings. This could correlate with higher population densities, tourist activity, or affluence in these regions, which can influence customer satisfaction and the propensity to leave positive reviews. In contrast, the interior regions show a more varied mix of ratings with an overall tendency towards the middle of the scale (green), suggesting moderate customer satisfaction. These areas might represent a

more diverse mix of business types or different customer expectations and experiences. Overall, the heatmap suggests that geographic location is a factor in the average ratings businesses receive.

4. Predictive Task

Objective: Forecast the average rating of a business based on a combination of its geolocation (latitude and longitude), category and, number of reviews. The goal is to understand how location and business type influence customer ratings.

- **Label (Dependent Variable):**
  - Average rating of businesses (On a scale of 1-5)
- **Features (Independent Variables):**
  - **Geolocation Features:**
    - Latitude and longitude of the business.
    - Additional geospatial features like distance to significant landmarks, area density, etc.
  - **Business Category:**
    - Type of business (e.g., restaurant, retail, etc).
  - **Number of Reviews:**
    - Count of number of reviews per category

5. Baseline Models

In order to establish a initial understanding of how different factors influence business ratings, I created four simple baseline models.

5.1. Baseline Models Description

- **Model 1: Global Average Rating Predictor** – This model predicts a constant rating for all businesses. It uses the global average rating, which is a simple approach and does not differentiate between business categories or locations.
- **Model 2: Category Average Rating Predictor** – This model predicts ratings based on the business's category. It calculates the average rating for each category and uses this average to predict ratings for businesses in that category, considering the variation in ratings among different categories.
- **Model 3: Geospatial Bin Average Rating Predictor** – This model introduces a geographical element by predicting ratings based on the geospatial bin of the business. A geospatial bin refers to a specific area or zone, determined by dividing the geographical area into bins or cells based on latitude and longitude. The model calculates the average rating of all businesses within the same geospatial bin to make its predictions.
- **Model 4: Linear Regression with Combined Features** – This model applies linear regression to predict business ratings using category average ratings, latitude, and longitude. Although the latitude and longitude do not vary linearly with a businesses average rating, it still provides a solid baseline model.

5.2. Baseline Model Evaluation

The baseline models were evaluated using an 80%/20% train/test split, and the performance was measured using the Mean Squared Error (MSE) between the actual ratings and the predicted ratings. The following table presents the MSE for each model:

Model	MSE
Model 1	0.394
Model 2	0.324
Model 3	0.389
Model 4	0.320

Table 1: MSE of Baseline Models

Model 4, employing linear regression with a combination of category averages and geographical data (latitude and longitude), emerges as the most accurate with the lowest MSE of 0.320. This indicates that integrating both categorical and geographical information provides a more comprehensive understanding of what influences business ratings. The model's success suggests that the nuances captured by the combined features are crucial in accurately predicting business ratings, demonstrating the effectiveness of a multifaceted analytical approach. Model 2, focusing solely on category averages, also performs well with an MSE of 0.324, underscoring the significant impact of business categories on customer ratings. However, Model 1, a simple global average rating predictor, and Model 3, based on geospatial bin averages, show higher MSEs of 0.394 and 0.389, respectively. This contrast highlights the added value of combining category and geographical data over using either in isolation. The results collectively emphasize the importance of a holistic approach in predictive modeling, where the interplay of various features can be more accurately captured for insightful outcomes.

6. K-th Nearest Neighbor Model

I evaluated several candidate models, each with distinct advantages and disadvantages. Decision trees were considered for their excellent interpretability and ability to handle non-linear data, but they often suffer from overfitting and can produce biased results with unbalanced datasets. Support Vector Machines (SVM) were also a candidate due to their effectiveness in high-dimensional spaces and versatility with different kernel functions. However, the complexity of choosing the right kernel and parameter tuning, along with poor performance in cases of class overlap and limited scalability for large datasets, made SVMs less ideal for this analysis. Ultimately, the K-Nearest

Neighbors (KNN) algorithm was chosen for its simplicity, flexibility, and particularly for its ability to capture spatial autocorrelation - a crucial factor given the clustered nature of latitude and longitude in the dataset. While KNN excels in scenarios with strong spatial data correlation, it is computationally intensive during prediction and sensitive to irrelevant features and the choice of distance metric, considerations that were carefully managed in this study.

The model utilizes latitude, longitude, weighted category averages, and the logarithm of review counts. The logarithmic transformation moderates the impact of outliers, and the weighted category average, scaled by an arbitrarily chosen factor, adjusts each category's influence based on review volume. This approach tempers the impact of categories with fewer reviews.

Feature scaling was essential to normalize data for the KNN algorithm, which depends on distance calculations. The model employed KNN with 100 neighbors, striking a balance between overfitting and underfitting.

Model optimization was conducted using scikit-learn's GridSearchCV. This involved searching through a range of hyperparameters, including the number of neighbors and different distance metrics like Euclidean, Manhattan, and Minkowski, along with weighting schemes (uniform and distance-based). The grid search process ensured the selection of the most effective combination of parameters, further refining the model's predictive accuracy. This exhaustive approach to optimization was crucial in enhancing the KNN model's performance, leading to a more reliable and robust predictive tool. This resulted in an MSE of 0.309, outperforming baseline models.

7. Relevant Literature

7.1. Transportation and Traffic Analysis

In transportation and traffic analysis, leveraging location data has proven pivotal. Studies like those by Vlahogianni et al. (2014) utilize datasets such as GPS tracking and traffic sensor data to model traffic patterns and forecast congestion. Advanced methods, including machine learning algorithms, are employed for real-time data analysis. Companies like Uber and Lyft use these techniques to predict ride demand and adjust pricing, demonstrating practical applications in urban mobility management.

7.2. Environmental and Agricultural Studies

The integration of Geographic Information Systems (GIS) in environmental and agricultural studies, as seen in the work of Olaya et al. (2019), is transforming the field. Satellite imagery and climate data are used alongside remote sensing technologies and predictive modeling to monitor and optimize agricultural yields. This approach is reflective of broader trends in sustainable farming and environmental conservation.

7.3. Public Health and Epidemiology

Public health and epidemiology have significantly benefited from location data, as exemplified by Tatem et al. (2017). Datasets involving patient location, travel history, and infection rates aid in disease spread modeling. The use of statistical modeling and network analysis, especially evident in tracking COVID-19, aligns with broader public health strategies for managing health crises.

7.4. Crime Analysis and Public Safety

Spatial data plays a crucial role in crime analysis and public safety, with predictive analytics used for resource allocation and crime reduction, as shown in studies by Chainey et al. (2008). Techniques like spatial clustering help police departments identify and respond to crime hotspots, aligning with trends in predictive policing but also raising considerations about data privacy and ethics.

7.5. Disaster Response and Management

Location data's critical role in disaster response and management is highlighted in Cutter et al.'s (2013) research on natural disasters. Historical data, geographic information, and real-time environmental sensors are used in simulation models and risk assessment tools. These methods, as employed by agencies like FEMA, are essential for effective disaster risk assessment and response planning.

8. Conclusion

This study has comprehensively analyzed the influence of geographical location on business ratings using a large dataset from Google Local Reviews. The analysis focused on businesses in California, a region characterized by its diverse economic and geographic landscape. Through the utilization of various predictive models, including a K-Nearest Neighbors (KNN) algorithm, the research has highlighted the significant role of both location and business category in influencing customer ratings.

The results from the baseline models established that the category of a business is a strong predictor of its rating, with Model 2, which focused on category averages, outperforming other simpler models. The success of the K-Nearest Neighbors (KNN) model in this study is notably attributed to its intrinsic nature, which aligns well with the characteristics of the dataset and the research objectives. KNN thrives in scenarios where spatial relationships are key, as it predicts outcomes based on the proximity and characteristics of neighboring data points. This aspect made it exceptionally suited for analyzing business ratings in California, where geographical closeness often translates to similarity in customer preferences and experiences. The model's flexibility in handling non-linear relationships between variables without the need for assumptions about the data distribution further contributed to its superior performance, as evidenced by the lowest Mean Squared Error. KNN's ability to adaptively learn from the dataset and accurately reflect the

complex interplay between location and business category underscored its appropriateness for this nuanced analysis.

Through the visualization of data via histograms, bar charts, and heatmaps, clear trends emerged, illustrating the positive skew towards higher ratings and the variation in ratings across different business categories and geographical locations. The study's findings align with relevant literature, showing the utility of location data in various domains like traffic analysis, public health, and crime analysis, as well as in the context of business ratings.

9. References

1. Vlahogianni, E. I., et al. (2014). "Short-term traffic forecasting: Where we are and where we're going." Transportation Research Part C.
2. Olaya, Y., et al. (2019). "Spatial analysis and GIS in the study of COVID-19. A review." Science of The Total Environment.
3. Tatem, A.J., et al. (2017). "Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation." Population Health Metrics.
4. Chainey, S., et al. (2008). "The utility of hotspot mapping for predicting spatial patterns of crime." Security Journal.
5. Cutter, S.L., et al. (2013). "GI science, disasters, and emergency management." Transactions in GIS.
6. Li, J., Shang, J., McAuley, J. (2022). "UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining." Annual Meeting of the Association for Computational Linguistics (ACL).
7. Yan, A., He, Z., Li, J., Zhang, T., McAuley, J. (2023). "Personalized Showcases: Generating Multi-Modal Explanations for Recommendations." The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).