Zongrong Li
DSCI 510

Geospatial Analysis of Socioeconomic Status and Restaurant Distribution in Los Angeles

1. Motivation/Rationale for the Project:

This project aimed to explore the relationship between socioeconomic factors and the distribution of restaurants across LA. The motivation behind this analysis was to understand how variables like income, education, and employment levels influence the variety and density of restaurants. This understanding could serve urban planners and policy makers in enhancing economic development and promoting social equity through informed urban design and resource allocation.

2. Description of Data Sources:

For this project, data regarding restaurants, socioeconomic indicators, and demographic distributions was collected across Los Angeles County using several web APIs.

Data was gathered from three main sources:
Yelp API: https://docs.developer.yelp.com/docs/fusion-intro
Used for obtaining restaurant data, including locations, types, and ratings.
LA Geo Hub API: https://developers.arcgis.com/python/
Provided geographic and demographic insights across the city.
Area Deprivation Index (ADI): https://www.neighborhoodatlas.medicine.wisc.edu/
Offered socioeconomic status indicators.

The process to extract restaurant data involved the use of the yelp_downloader.py script, which automated interactions with the Yelp API. This script was configured to handle API rate limits by implementing delay mechanisms and pagination techniques to systematically collect data across different requests. This ensured the extraction of a comprehensive dataset that included approximately 1,000 entries detailing restaurant locations, types, and customer ratings.

For geographic and demographic information, the LA_CTs_downloader.py script accessed the ArcGIS API. This script fetched detailed data for over 2,000 community tracts in Los Angeles County, utilizing Python's requests library to make API calls and Python's json library to parse the responses. The script stored the geographic data in a structured format, enabling easy integration with socioeconomic data.

Socioeconomic data was obtained through the ADI_downloader.py script, which interfaced with the ADI API. This script used a similar methodology to the geographic data collection, employing loops to handle large volumes of data requests and parsing JSON formatted responses to extract key socioeconomic indicators like income levels, employment status, and educational backgrounds for matching community tracts.

Originally, I planned to analyze data at the zip code level, but limitations in data availability and API constraints necessitated a shift to community tracts. This change, while adjusting the granularity of the analysis, allowed for a more robust dataset suitable for comprehensive analysis. The project faced challenges such as API rate limitations and the management of large datasets, which were mitigated by optimizing data request sequences and employing error handling techniques to ensure data integrity and completeness.

3. Integrated Data Model:

Zongrong Li
DSCI 510

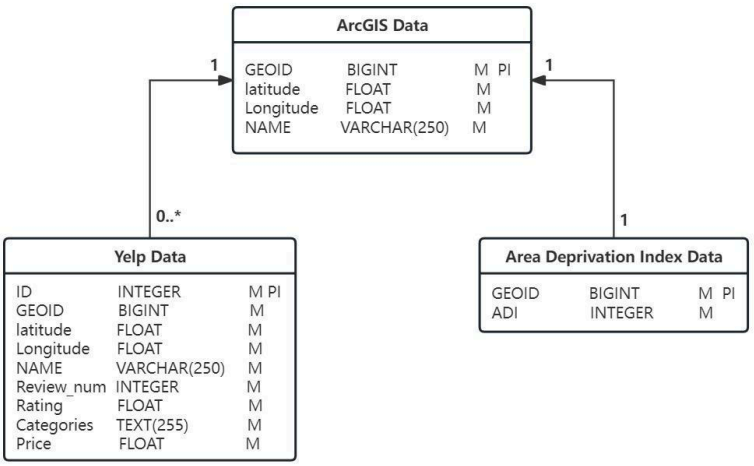

Figure 1. Diagram of Integrated Data Model

The integrated data model is structured around three principal data sets: Yelp Data, ArcGIS Data, and Area Deprivation Index (ADI) Data, with relational connections established via geographic identifiers.

The Yelp Data table is centered on individual restaurant entities, capturing essential information such as a unique identification number, a geographic identifier (GEOID), spatial coordinates (latitude and longitude), and descriptive attributes including the establishment's name, number of reviews, average rating, categories, and price level. Each restaurant entry is linked to a specific geographic location by the GEOID, which is a fundamental element in connecting to the ArcGIS dataset.

The ArcGIS Data table holds the key geographic details of community tracts within Los Angeles County. It includes the GEOID as its primary identifier, which harmonizes with the Yelp Data, and records latitude and longitude coordinates and a textual name for each region. This dataset forms the geographical backbone of the analysis, interfacing with both Yelp and ADI data.

Connecting to the ArcGIS Data is the Area Deprivation Index (ADI) Data table. It, too, employs the GEOID for consistent linkage to the geographic regions delineated in the ArcGIS table. The ADI data captures the socioeconomic status of each tract, providing a quantitative measure of deprivation that can be used for socioeconomic analysis.

In the accompanying entity-relationship diagram, the one-to-one correspondence between the ArcGIS and ADI datasets signifies that each geographic area identified by the ArcGIS data is matched with a unique socioeconomic profile from the ADI data. The Yelp Data exhibits a one-to-many relationship to the ArcGIS Data, reflecting the reality that numerous restaurants can populate a single geographic region.

This model is instrumental in enabling multidimensional analyses, such as assessing the impact of socioeconomic factors on the distribution and variety of restaurants. Through the GEOID commonality, it facilitates the integration of diverse data sets, ensuring the integrity and utility of the analysis by providing a comprehensive, linked overview of geographic, economic, and commercial datasets

4. Analyses/Visualizations:

4.1 Analytical Techniques Employed
The investigation into the relationship between socioeconomic factors and the restaurant distribution within Los Angeles County was supported by an array of analytical techniques. Geospatial mapping was central to the visualization process, overlaying restaurant data on the map of the county to assess distribution patterns. Complementing this spatial representation, regression analysis provided quantitative insight into the correlation between socioeconomic indicators and the variety of restaurants. These statistical assessments were enriched by descriptive statistics, summarizing central tendencies and variabilities such as the average restaurant ratings and prices. To discern more complex patterns, k-means clustering analysis was applied, which sorted geographic tracts into homogenous groups based on shared socioeconomic and restaurant characteristics.

Zongrong Li
DSCI 510

4.2 Visual Representation of Data

The visualizations created for this analysis serve to highlight the intersection of restaurant metrics with socioeconomic factors as categorized by the ADI National Rank.

Figure 2 shows the variance in customer engagement and perceived quality across different socioeconomic areas. Restaurants in more affluent areas (presumably indicated by a lower ADI rank) generally show higher numbers of reviews and ratings, pointing towards a greater patronage and potentially higher satisfaction levels. Meanwhile, the price categorization aligns with the economic status, with pricier establishments more frequent in wealthier neighborhoods.

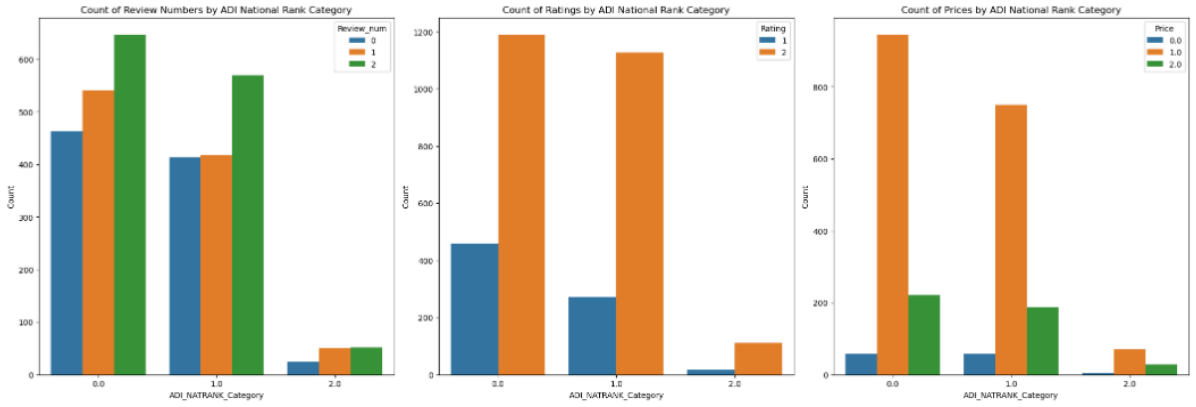

Figure 2. Bar plots for review numbers, ratings, and prices

A heatmap (Figure 3) displaying the correlation matrix for review counts, ratings, and prices elucidates the inter-relationships among these variables. A closer to 1 or -1 value indicates a stronger positive or negative correlation, respectively, while values closer to 0 suggest a lack of correlation. This heatmap aids in understanding the dynamics between how often restaurants are reviewed, how they are rated, and what price bracket they fall into.

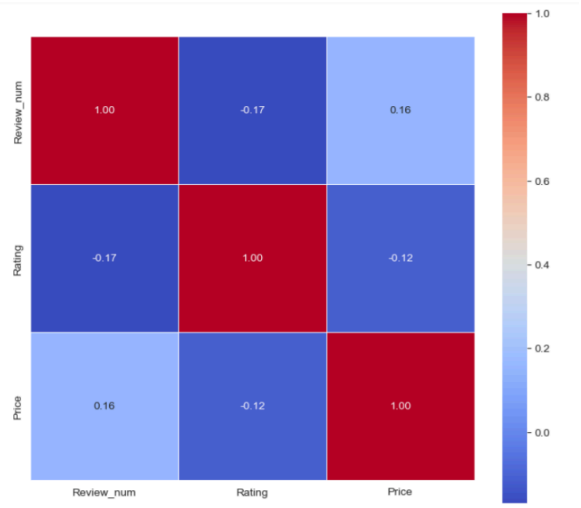

Figure 3. Heatmap

The word clouds classify restaurant types within ADI National Rank Categories (Figure 4), offering a visual representation of the most prevalent types of cuisine or restaurant styles within each socioeconomic segment. The size of each word within the cloud corresponds to the frequency of that term within the dataset, providing an at-a-glance understanding of dining options and preferences across economic divisions.

Zongrong Li
DSCI 510

Figure 4. The word clouds within ADI National Rank Categories

Lastly, figure 5 superimposes the frequency of customer reviews onto a geographical map. This type of visualization allows for the identification of hotspots where restaurants garner more engagement, potentially indicating areas of higher competition and consumer interest.

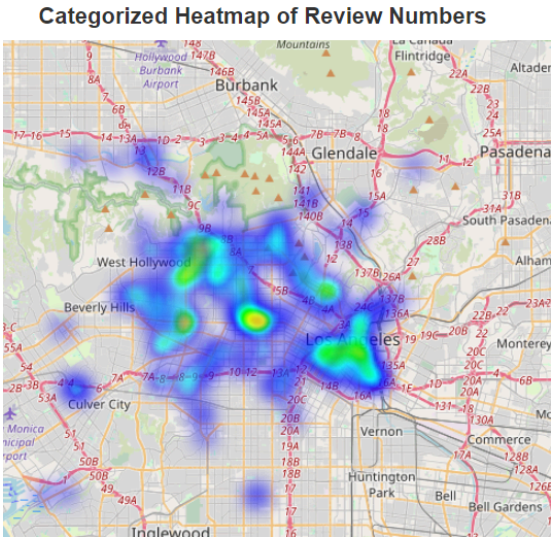

Figure 5. The categorized heatmap of review numbers

5. Conclusions:

In this study, a comprehensive examination of the restaurant industry within Los Angeles County was conducted, highlighting the integral role of socioeconomic factors in shaping the culinary landscape. The findings indicated that affluent neighborhoods tend to feature a higher frequency of restaurant reviews and more favorable ratings. This suggests a more engaged consumer base and may reflect a heightened quality of dining experiences in these locales.

Furthermore, the distribution of restaurant price points aligns with the economic affluence of an area, with higher-priced establishments predominantly situated in wealthier neighborhoods. This emphasizes the economic disparities in access to various dining experiences.

A diverse array of restaurant types was discovered in economically advantaged areas, as evidenced by the word clouds, pointing to a more eclectic and potentially experimental dining scene. This is contrasted by a less varied culinary offering in lower-income neighborhoods. Geospatial analysis via heatmaps elucidated specific districts as key gastronomic hubs, providing valuable insights for strategic planning in restaurant establishment and urban development.

The implications of this research are manifold and extend well beyond the bounds of the restaurant industry, touching upon the spheres of urban planning, economic development, and social policy. The study underscores the need for a mindful approach to fostering culinary diversity and accessibility across various socioeconomic segments. For entrepreneurs, understanding the socioeconomic dynamics at play can guide more informed decision-making in business positioning and marketing.

Zongrong Li
DSCI 510
In summary, this investigation sheds light on the intersection between socioeconomic status and dining options, providing a data-driven foundation for strategies aimed at promoting equitable growth and cultural inclusivity within the urban gastronomic fabric.

6. Future Work:

If additional time were available, several avenues could be pursued to enhance the depth and scope of this project:

1. Expanding the Dataset: By integrating more data points over a longer time frame, the analysis could capture trends and changes in consumer behavior and restaurant success metrics.
2. Incorporating Additional Variables: Factors such as cultural demographics, public transport access, and competition density could be included to refine the understanding of the restaurant industry's dynamics.
3. Advanced Statistical Modeling: More sophisticated models, like multilevel regression or machine learning algorithms, could uncover complex, non-linear relationships and predictive insights.