Value or Significance Scores of Places

MID PROGRESS PROJECT REPORT - CSE 519: DATA SCIENCE FUNDAMENTALS

***Abstract*—This mid-progress report provides an update on the ongoing data science project aimed at constructing a comprehensive dataset to assess the cultural, historical, and community significance of businesses across New York City. The primary objective of the project is to assign a significance score to each business, considering factors such as cultural relevance, historical importance, and community engagement. The dataset covers a diverse array of business types, including restaurants, churches, bookstores, clothing stores, and more, ensuring a holistic representation of New York City's commercial landscape.**

The report outlines the progress made in exploring multiple scoring methods to provide a nuanced perspective on the merit of each business. Initial findings indicate the potential for uncovering valuable insights by calculating *the rating score, "distinctiveness" of a business(or its complement- "chainness score"), walkability scores, cultural sentiment scores, and the subsequent identification of endangered independent businesses* that hold particular significance to the community being replaced by big chain restaurants.

Throughout the report, the challenges encountered during data collection, preprocessing, and analysis are discussed, along with the strategies employed to address them. Additionally, the mid-progress report outlines the next steps in the project, including further refinement of scoring methodologies, exploration of additional variables, and the integration of spatial and geographical analyses to enhance the depth of understanding.

As the project advances, it seeks to contribute meaningful insights into the intricate dynamics of businesses in New York City, offering a valuable resource for policymakers, urban planners, and community stakeholders. The ultimate goal remains to facilitate a comprehensive understanding of the multifaceted significance of businesses within the community and their broader impact on the cultural and historical fabric of New York City.

I. DATASET OVERVIEW

On collecting, scraping, generating and analyzing the data, the following is a summary of the dataset.



21985 Businesses | 59580 Reviews | 5+ Feature Scoring

1.1. Basic Information:

id	Unique identifier for each business.
alias	Business alias or nickname.
name	Official name of the business.
url	URL for the business webpage.
categories	Business categories.
coordinates	Latitude and longitude of the business.
location	Business location information.
price	Price range of the business.

1.2. Reviews and Rating Information:

rating	Average rating of the business.
review_count	Number of reviews for the business.
review_texts	Text of reviews for the business.
review_ratings	Ratings associated with the reviews.

1.3. Chain Identification:

chain_id_cosine	Cosine similarity chain id
chain_id_jaccard	Jaccard similarity chain id
chain_id_levenshtein	Levenshtein distance chain id
chain_id_fuzzy	Fuzzy matching chain id.
chain_count	Count of businesses in the same chain.
chain_type:	Independent, Local or National Chain flag
chain_avg_dist	Average distance between businesses in the same chain

1.4. Employment Information:

emp	Employment count in a zip code.
qp1	Total First Quarter Payroll info in a zip code
ap	Total Annual Payroll information in a zip code
est	Total Number of Establishments

1.5. Walkability Details:

walk_scores	Walkability scores of the business location.

1.6. Cultural Sentiment:

normalized_name	Normalized name of the business
cultural_sentiment	Cultural sentiment score of the business.

This dataset encompasses a diverse range of business and review information collected by using Yelp Fusion API[10], County Business Patterns dataset[11], web-scraping walkability scores, calculating cultural sentiment, etc. This includes business details, reviews, chain identification, score calculation, walkability scores, employment information, location details, and cultural sentiment scores. It serves as a comprehensive resource for evaluating the cultural, historical, and community significance of businesses in New York City.

II. EXPLORATORY DATA ANALYSIS

2.1 Price Range vs Frequency of Businesses:

The analysis of price range distribution across businesses reveals that the dataset has a diverse set of businesses belonging to all sections of the price range, the majority of which is in the low and mid-price ranges, providing insights into the economic diversity of the businesses in the dataset. This information is essential for understanding the affordability and consumer base of businesses in New York City.

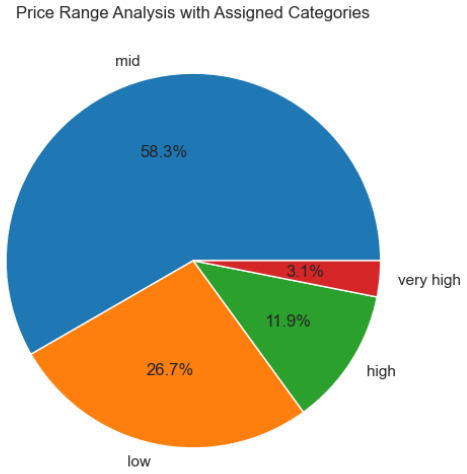


Fig. 1. Price Range of Businesses

2.2.Distribution of Review Ratings:

The histogram distribution of review ratings, ranging from 1 to 5, offers a comprehensive view of customer satisfaction levels. This analysis helps identify patterns in the sentiment of reviews and can guide businesses in improving their services based on customer feedback. We see a majority of New York City businesses tend to have good ratings online.

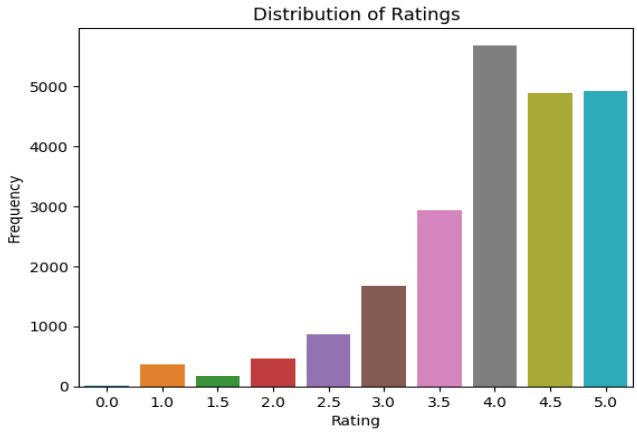


Fig. 2. Histogram Distribution of Average Ratings

2.3 Top 20 Most Popular Business Categories:

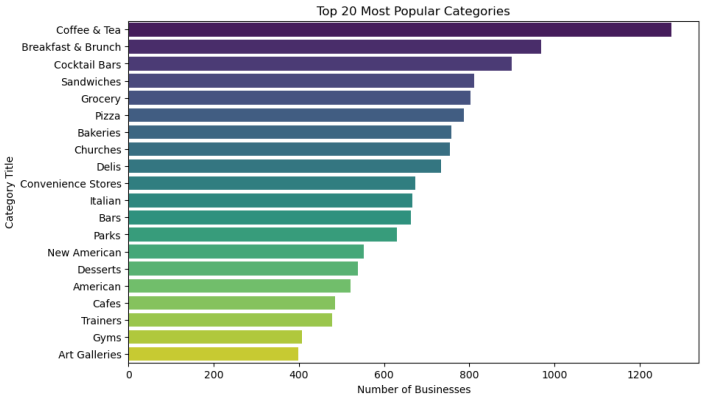


Fig. 3. Top 20 Most Popular Business Categories

The identification and visualization of the top 20 most popular business categories based on the frequency of occurrence provide valuable information about the diverse commercial landscape in New York City. We can observe there is a dominance of eateries across New York City with some churches, fitness centers, and convenience stores among others.

III. SCORING METHODOLOGIES

Scoring System Philosophy:

The scoring methodologies are designed with a **deliberate bias favoring independent businesses and businesses that were established a long time ago**. This strategic emphasis aims to recognize and elevate the unique cultural contributions of locally owned establishments within the vibrant tapestry of New York City. The scoring system serves as a tool to not only quantify but also celebrate the diverse and historically rich commercial landscape of the city.

3.1 Rating Score Calculation:

To evaluate the overall performance of businesses within the Yelp dataset, we devised a composite rating score that incorporates both the average rating and the number of ratings received. The rationale behind this approach is to give weight to both the sentiment expressed through ratings and the popularity indicated by the volume of ratings. Individual ratings were standardized using a Standard Scaler to maintain the relative differences between ratings, crucial for integrating them into a unified scoring system. Simultaneously, the number of ratings was scaled using a Min-Max Scaler. This ensured that businesses with a small number of ratings did not disproportionately influence the score, while still considering their presence in the dataset. We calculate the final rating score by multiplying the standardized rating and the scaled number of ratings for each business:

$$R_i = r_i * n_i$$

Where R_i is the Rating Score, r_i is the Standardized Rating and n_i is the Scaled_Number_of_Ratings$_i$

3.2 Chainness Score Calculation:

In the ever-evolving landscape of business diversity, the presence of chain and independent establishments influence the authenticity and uniqueness of the local culture. Building upon the insights derived from the research conducted by Liang and Andris [2], which explored the prevalence of chainness in the United States, we delve into the development of a "Chainness score" to evaluate the distinctiveness of businesses within our dataset. We assess the degree of chainness within our dataset by considering three key parameters for each business:

Chain Count: This represents the number of outlets for a specific chain within our entire dataset and provides insight into the ubiquity of a chain across various locations. To address variations in naming conventions among outlets of the same chain, we employed four methods: Cosine Similarity, Jaccard Similarity, Levenshtein Distance, and Fuzzy Matching.

Following a thorough evaluation, Fuzzy Matching proved to be most effective, providing a more accurate assessment of name similarity. Utilizing Fuzzy Matching, we refined our chain counts, ensuring accuracy by accommodating minor differences in business names

Chain Type Flag: This categorizes businesses into three distinct types. A flag value of 2 denotes a nationwide chain, identified by checking if the business belongs to the top 250 chains across various categories in the USA. A value of 1 indicates a local chain, identified by its absence in the top 250 chain list but with more than 5 chain counts in the dataset. A value of 0 suggests that it is an independent business, characterized by having chain counts less than 5.

Chain Average Distance: This measures the average distance in kilometers between all outlets of a specific chain. This was calculated by using latitude and longitude data and provides insights into the spatial distribution of a chain.

The Chainness score can be calculated as follows:

$$C_i = c_i + (t_i * d_i)$$

Where C_i is the chainness_score, c_i is the count of outlets of a chain, t_i is the chain type flag and d_i is the average distance between outlets of a chain.

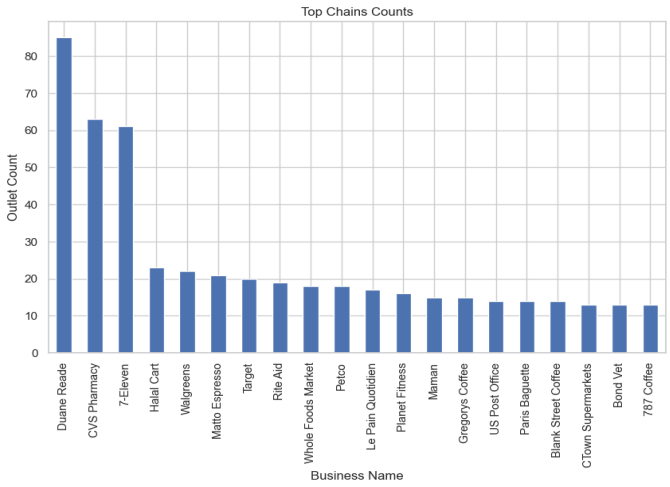


Fig. 4. Frequency of Top Chains

Higher scores are indicative of well-established, nationally recognized chains, while lower scores highlight the distinctiveness and potential cultural richness of local or independent businesses. In simple terms, a business like “7-Eleven” will have a significantly higher chainness score(hence very low distinctiveness) than a culturally unique place like “Chelsea Flea Market".

3.3 Cultural Sentiment Score Determination:

The Cultural Sentiment Score assesses a business's cultural relevance, historical importance, and community engagement by leveraging sentiment analysis and community impact metrics. This score is computed by analyzing the textual content from multiple reviews associated with each business. This can help us capture the sentiment polarity score along with the weighted sum of the presence of cultural keywords, providing a quantitative measure of the cultural significance of each business.

The mathematical equation to calculate it is as follows:

$$CS = s + \sum_{i=1}^{n} (w_i * p_i)$$

where CS is Cultural Sentiment Score, s is the sentiment polarity score obtained from TextBlob analysis of combined reviews, n is the number of cultural keywords (e.g., unique, artisanal, heritage, etc.), set of 50 in our case, w_i is the keyword weight, assigned as 0.2 for each cultural keyword, and p_i is the presence of each cultural keyword.

3.4 Economic Factor Score Calculation:

We utilized data from the [11]County Business Patterns (CBP) dataset, and incorporated key economic indicators into our analysis. The following additional columns were generated for each ZIP Code in our dataset:

Number of Employees per Establishment: Calculated as the total Mid-March Employees (EMP) divided by the Total Number of Establishments (EST).

Average Pay of Employees: Derived by dividing the Total Annual Payroll (AP) by the total Mid-March Employees (EMP).

Revenue per Establishment: Approximated as the Total Number of Establishments (EST) divided by the Total Annual Payroll (AP). The rationale behind using this metric is to gauge an establishment's revenue based on its payroll expenditure.

To combine these metrics into a single metric, we scaled each metric to ensure uniformity and comparability. Then the Economic Factor Score was computed as the sum of the scaled values of the number of employees per establishment, average pay of employees, and revenue per establishment. It can be mathematically calculated as follows:

$$E_i = a_z + n_z + r_z$$

Where E_i is the Employee Factor Score, a_z is the average annual pay per employee, n_z is the number of employees per establishment and r_z is the revenue per establishment

This composite score serves as a comprehensive indicator, capturing key dimensions of economic activity within each ZIP Code.

3.5 Walkability Score Computation:

In this study, we have employed a patented methodology [1] to calculate the walkability score for each zip code, considering various factors that contribute to the overall pedestrian experience. The study delves into pedestrian friendliness by assessing road metrics such as block length, intersection density, and population density. This multifaceted approach provides a comprehensive understanding of the walkability dynamics within each zip code.

Walk Score	Description
90–100	**Walker's Paradise** Daily errands do not require a car.
70–89	**Very Walkable** Most errands can be accomplished on foot.

50–69	**Somewhat Walkable** Some errands can be accomplished on foot.
25–49	**Car-Dependent** Most errands require a car.
0–24	**Car-Dependent** Almost all errands require a car.

Our analysis reveals a notable correlation between walkability scores and key economic indicators. Zip codes with higher walkability scores exhibit increased annual payrolls (ap), a greater number of establishments (est), and higher employment figures (emp). This underscores the pivotal role of walkability in shaping the economic landscape, serving as a reliable indicator of the employment opportunities fostered within a given region.

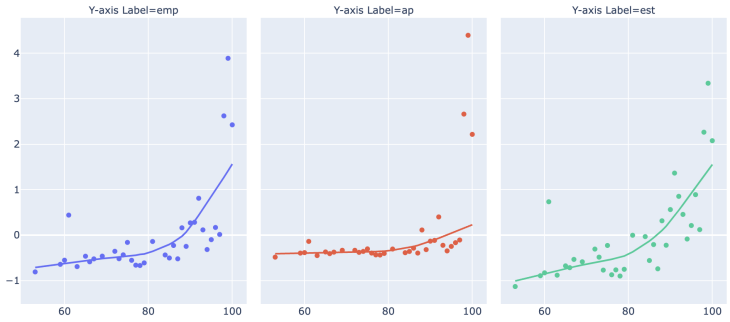


Fig. 5. Trend of Walkability score

Therefore, we determined that walkability plays a crucial role in influencing economic outcomes, our findings underscore the potential significance of pedestrian-friendly environments in driving business success and economic vitality.

3.6 Final Significance Score Synthesis (Baseline):

The culmination of our analysis results in the formation of a final Significance Score, a comprehensive metric that encapsulates various dimensions crucial to the cultural and economic fabric of New York City. This score integrates key parameters, including the Rating Score, Walkability Score, Chainness Score, Employee Factor Score, Cultural Sentiment Score, and the temporal aspect of a business represented by the scaled establishment year.

The equation governing the Final Significance Score is as follows:

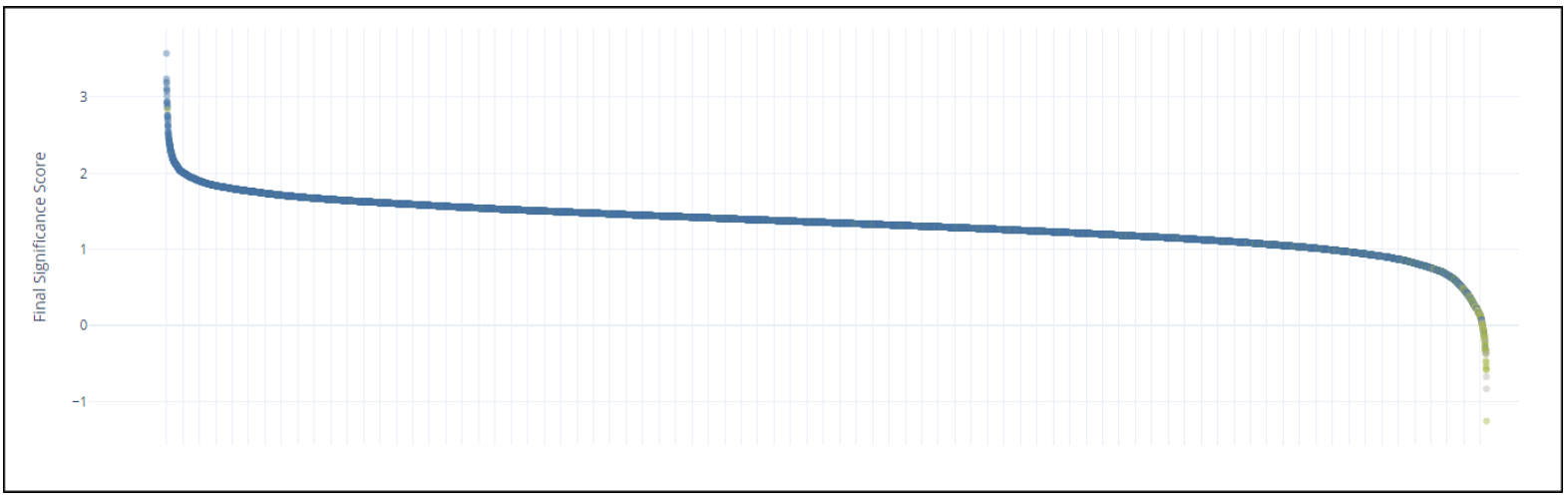
$$\textit{Final\_Significance\_Score} = \textit{Rating\_Score} + \textit{Walkability\_Score} - \textit{Chainness\_Score} + \textit{Employee\_Factor\_Score} + \textit{Cultural\_Sentiment\_Score} + (\textit{Year of Establishment} - \textit{Current Year})$$

IV. RESULT ANALYSIS

Firstly, let's plot the significant scores in descending order of their value (**Fig 6** at the end of the page) and color indicating if it's an independent, local chain, or national chain. Here, olive green indicates national chains whereas blue indicates independent businesses. We can clearly see national chains at the lower end of the graph in line with our requirements. Local chains are less in count and since we plotted 20k+ points corresponding to the business records in the dataset, it was probably overlapped by the majority of blue points.

After taking into consideration multiple factors which include cultural sentiment score, chainness score, walkability score, economic factor score, and rating score, we were successfully able to generate a continuous term "significance score", ranging from 3.57 to -1.25. This score is successfully captures how distinct and culturally significant the business is. For example, all chain businesses like "Target", "ALDI", etc. have the lowest significance scores, while places that are distinct and which have high community engagement like "The Rink at Brookfield Place" have high significance scores. Several interesting observations were found when all businesses were ranked according to the significance score. Most businesses like libraries and museums, which have a very high community engagement and cultural significance, have the highest significance score values among all other businesses. An example would be The New York Public Library, which has a significance score of 2.73 or The Poet's House, a quaint library for holding events and workshops for wordsmiths has a score of 2.86. On the other hand, commonplace businesses that are not culturally significant like "CVS Pharmacy" or "Stop & Shop" have values of -1.25 and -0.67 respectively.

Fig 6.(Below) Significant Scores of Businesses



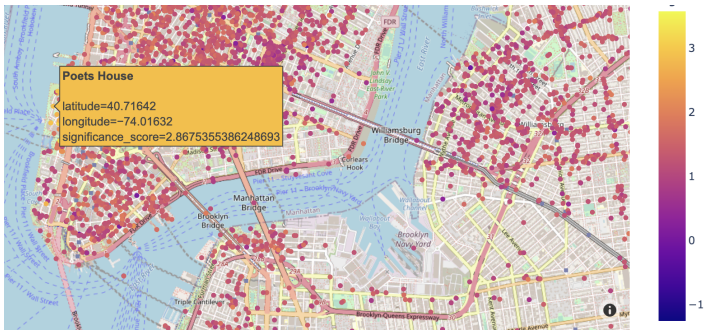


Fig 7. Distribution of Significance Scores across NYC

We plotted the significance score of each business on the map of New York City. Certain pockets of the city contain highly significant places, like Lower Manhattan. The place contains a small cluster of libraries, parks, and bars like “Poet’s House” and “Van Gogh's The Immersive Experience”

We can see that Lower Manhattan has a cluster of culturally significant places. Further analysis showed us that Midtown and Uptown Manhattan has a significant number of local chains like “2 Bros Pizza” and “The Halal Guys”.

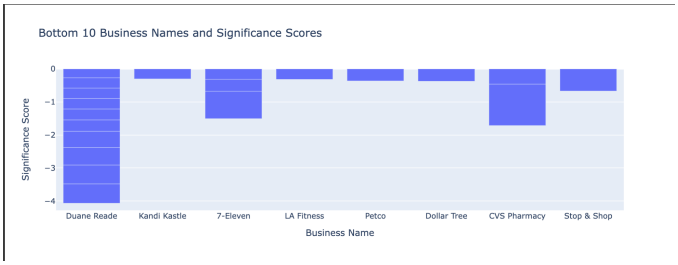


Fig 8. Bottom 10 Business and Significance Scores

We have also plotted the businesses with the least significance scores. As stated before, all of these are national-level chains and due to their high chainness scores, their significance scores are much lower. This list consists of multiple stores from “Duane Reade”, “Shop & Stop” and “Dollar Tree”.

IV. Cluster Analysis for Validation:

Given the unique nature of the developed significance scores and the absence of external validation, we employed cluster analysis as an internal validation measure. The goal was to identify patterns and groups within our dataset, acting as a validation for our calculated significance scores. The Gaussian Mixture Models (GMM) technique allowed us to categorize businesses into four clusters, revealing inherent patterns within New York City’s diverse business landscape.

The geographical distribution of these clusters is visually depicted on the map, offering insights into the spatial distribution of businesses based on the derived scores. Then we used Boxplots to analyze the distribution of various features within these clusters. This allowed us to spot trends and differences in factors like employee aspects, cultural sentiment, and walkability scores among the clusters.

Feature	F-statistic	p-value
review_count	477.96966	1.0952e-299
rating	292.7297	7.0549e-186
walk_scores	586.7369	0
chain_count	3265.2159	0
chain_avg_dist	5399.4958	0
emp	11954.4310	0
ap	10450.0506	0
est	7380.6163	0

Additionally, we used the F-statistic, a statistical measure, to assess the variance between the clusters for different features through ANOVA analysis. This statistical approach adds a level of rigor to our cluster analysis, providing a solid statistical basis for the identified groupings.

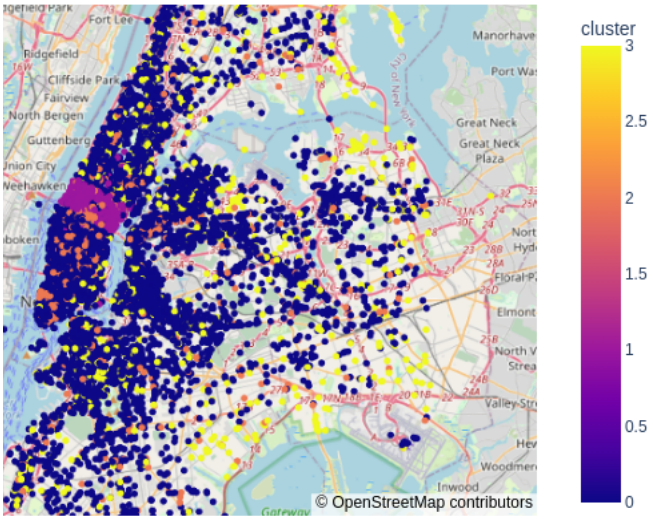


Fig 9. Distribution of clusters across NYC

References and Footnotes

1. Walk Score Methodology [Link]
2. Measuring McCities: Landscapes of chain and independent restaurants in the United States [Link]
3. Reviews, Reputation, and Revenue: The Case of Yelp.Com [Link]
4. Non‐place and placelessness as narratives of loss: Rethinking the notion of place [Link]
5. (Non‐place and placelessness as narratives of loss)
6. Total U.S. Restaurant Count Reaches 647,288, A Drop From Last Year Due to Decline in Independent Restaurant Units [Link]
7. How Fast Food Cornered the Urban Market [Link]
8. Restaurant Organizational Forms and Community in the U.S. in 2005 [Link]
9. Reviews, Reputation, and Revenue: The Case of Yelp.Com [Link]
10. Yelp Fusion API [Link]
11. County Business patterns[Link]