Value or Significance Scores of Places

FINAL PROJECT REPORT - CSE 519: DATA SCIENCE FUNDAMENTALS

ABSTRACT

Independent businesses have historically been under constant pressure due to local and national chains of businesses. We have developed a scientific way to quantify the cultural significance of a given business in 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. We have taken into account 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 how we explored 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 score of the neighborhood, cultural sentiment scores derived from reviews, and the subsequent identification of endangered independent businesses* that hold particular significance to the community being replaced by big chain businesses. After the calculation of significance scores, we have provided an in-depth analysis of the findings by providing relevant visualizations.

I. DATASET OVERVIEW

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.

Data Collection Process:

The data collection process began by utilizing the Yelp Fusion API. Initially, we used the Business Search endpoint to retrieve businesses based on specified categories and location criteria. Following this, the Business Details endpoint was employed to access detailed business information such as name, address, phone number, photos, Yelp rating, price levels, and other attributes.

Further enriching the dataset, the Reviews endpoint was invoked for each business, to retrieve up to three review excerpts per business. Once this data from Yelp was acquired, it was integrated with the additional datasets mentioned earlier, resulting in a holistic dataset.

Post-integration, the data was cleaned to ensure accuracy and consistency across all dimensions, the following is a summary of the final dataset:

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.

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.

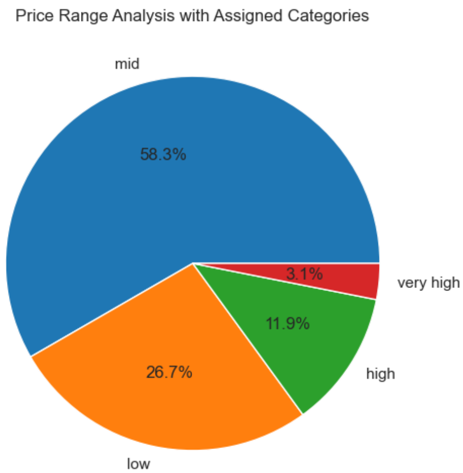

Fig. 1. Price Range of Businesses

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.

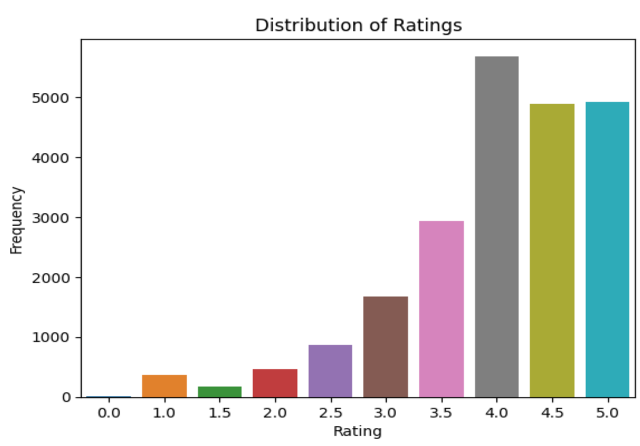

Fig. 2. Histogram Distribution of Average Ratings

2.3 Top 20 Most Popular Business Categories:

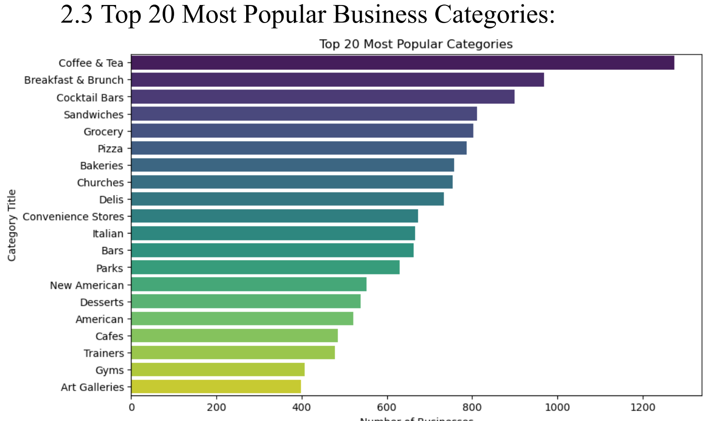

Fig. 3. Top 20 Most Popular Business Categories

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

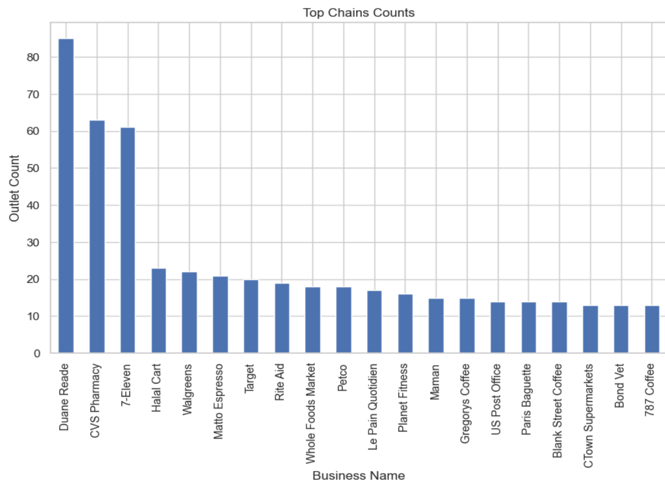

Fig. 4. Frequency of Top Chains

Chain Type: 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.

Our findings also match what Liang and Andris [2] found about water views and roads affecting where chain and independent restaurants set up businesses (refer Fig. 5). Specifically, chains exhibit a preference for locations near roads, particularly in the vicinity of highways. Conversely, independent restaurants are frequently situated near water bodies.

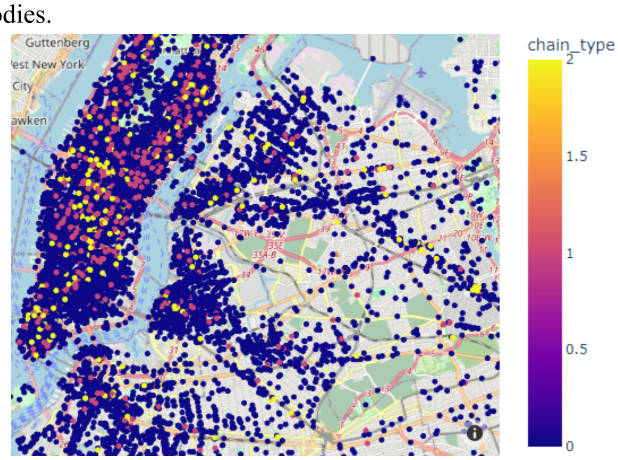

Fig. 5. Distribution of chain types across NYC

Upon generating the histogram of average distances for each chain type (refer Fig. 6), we noticed something interesting. Local chains usually have outlets close to each other, while national chains are spread out more. This may be attributed to the likelihood that they have a single owner who actively manages them, leading to a more concentrated presence.

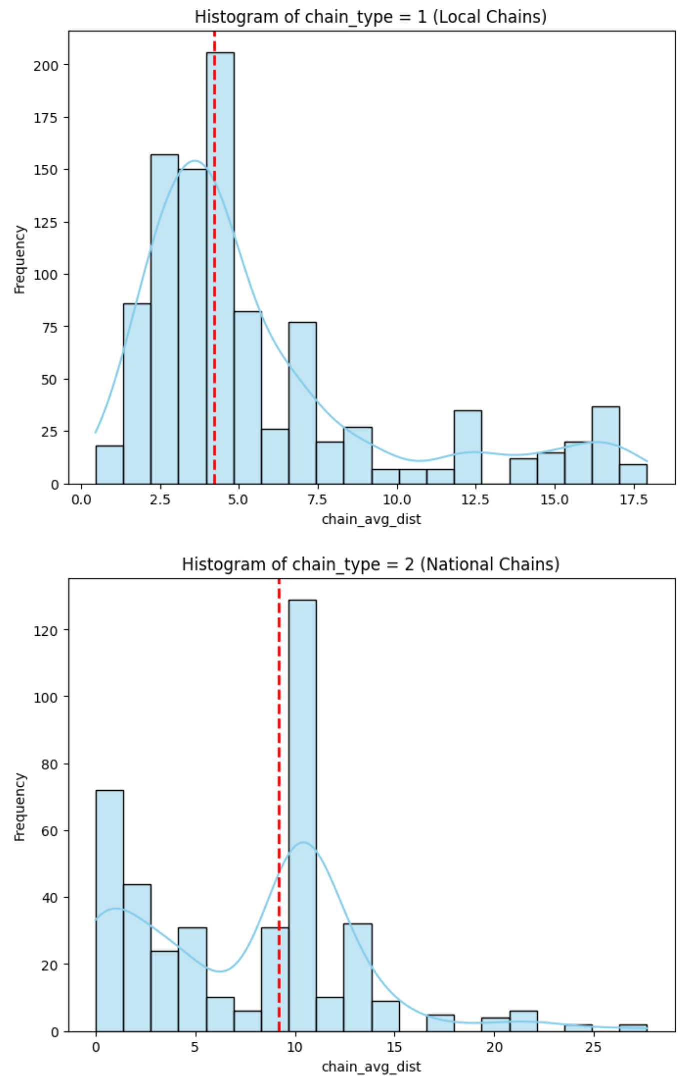

Fig. 6. Distribution of average chain distances

Finally, calculating the Chainness score after combining these features can be done 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 and d_i is the average distance between outlets of a chain.

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:

$$CS_i = s + w * \sum_{i=1}^{n} p_i$$

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 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 normalized 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. All the three values have been normalized.

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.

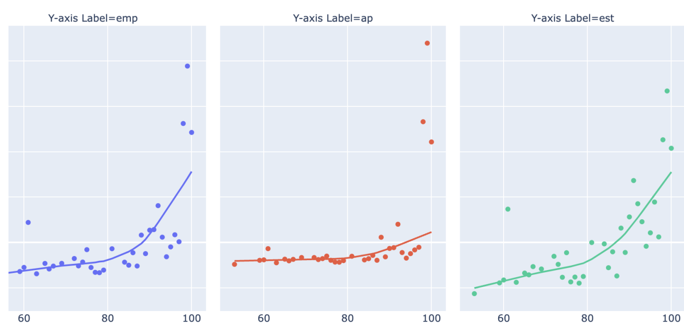

Fig. 7. Trend of Walkability score

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:

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:

$$SS_i = (w_0 * R_i) + (w_1 * W_i) - (w_2 * C_i) + (w_3 * E_i) + (w_4 * CS_i) + (w_5 * (CY - YOE_i))$$

Where SS_i is the Significance Score, R_i is the Rating Score, W_i is Walkability Score, C_i is Chainness Score, E_i is Employee Factor Score, CS_i is Cultural Sentiment Score, CY is Current Year, YOE_i is the Year of Establishment of the business. The weights for each feature is denoted by w_0-w_5.

IV. Result Analysis

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 "significance score", ranging from 0 to 100.

Baseline Model Scores:

To set a baseline, we assign equal weights to all features. Subsequently, we normalize the baseline significance score to a range of 1-100 to facilitate analysis and observation. Examining the violin plot based on chain type (independent, local, or national chain) in Fig. 9, we observe a substantial overlap among independent, local, and national chain businesses. This indicates that the baseline score is not biased toward independent chains or cultural sentiments in the reviews of places, with independent businesses generally achieving higher scores than chain businesses.

(w)eight	0	1	2	3	4	5
value	1	1	1	1	1	1

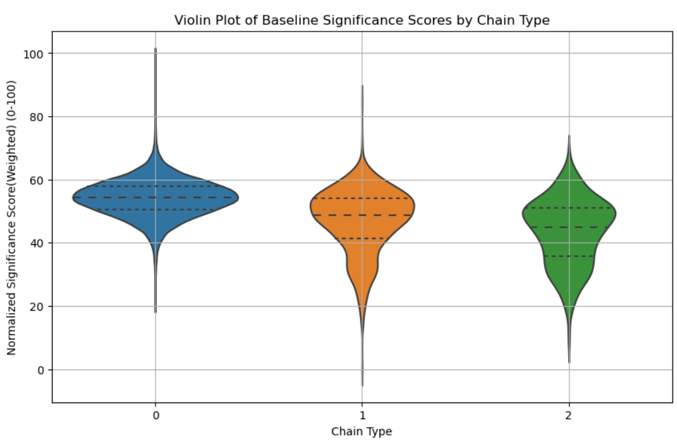

Fig. 9. Normalized Baseline Significance Score by Chain Type

Improved Model Scores:

To optimize the scores distribution focusing on cultural significance and distinguish better between the chain types, we tried trial and error to increase each feature's weight independently and observe the outcome's distribution. This led to the conclusion that increasing the weight of cultural sentiment scores and distinctiveness scoring(by incurring a larger penalty for the chainness score) we get optimized scores.

(w)eight	0	1	2	3	4	5
value	1	1	**2**	1	**4**	1

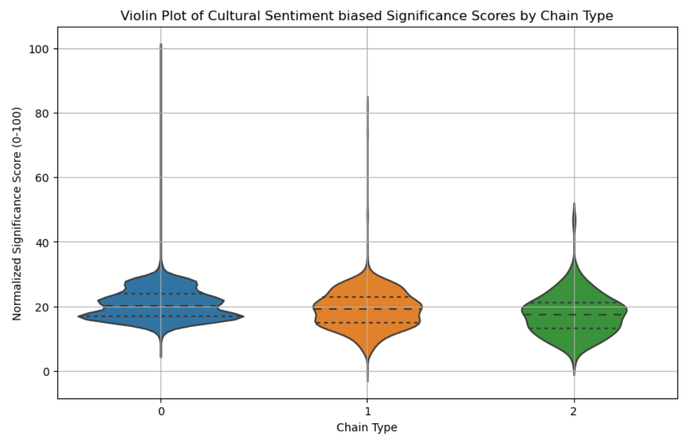

Fig. 10. Cultural Sentiment biased Score by Chain Type

The refined significance score, after optimization, presents a well-distributed range from 1 to 100. In this visualization, we've color-coded the scores to distinguish between independent businesses (blue), local chain businesses (orange), and national chain businesses (green). The analysis reveals that independent businesses consistently achieve higher scores, while national chains tend to have lower scores. This observation aligns with our initial hypothesis emphasizing the cultural significance of businesses within a specific area.

(w)eight	0	1	2	3	4	5
value	1	1	**4**	1	**2**	1

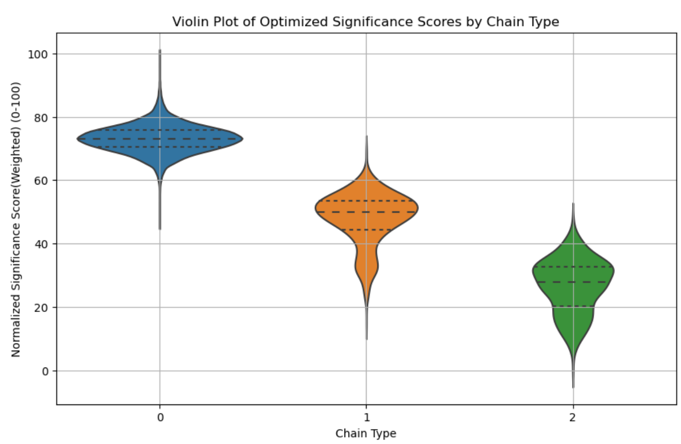

Fig. 11. Optimized Model Significance Scores- (Distinctiveness scores AND Cultural Sentiment) biased Scores by Chain Type

Our methodology allows us to quantify the impact on cultural significance when an independent business is replaced by a national chain. Notably, each occurrence of such a transition results in a discernible decrease in the average cultural significance of that location. These scores provide a quantitative measure of the decline in a place's cultural identity when such replacements occur, thereby offering valuable insights for data-driven decision-making. This approach enhances our understanding of the intricate relationship between business types and the cultural fabric of a community.

The final significance score 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 84.61 or The Poet’s House, a quaint library for holding events and workshops for wordsmiths has a score of 86.99. On the other hand, commonplace businesses that are not culturally significant like “CVS Pharmacy” or “Stop & Shop” have median values of 30.96 and 16.92 respectively.

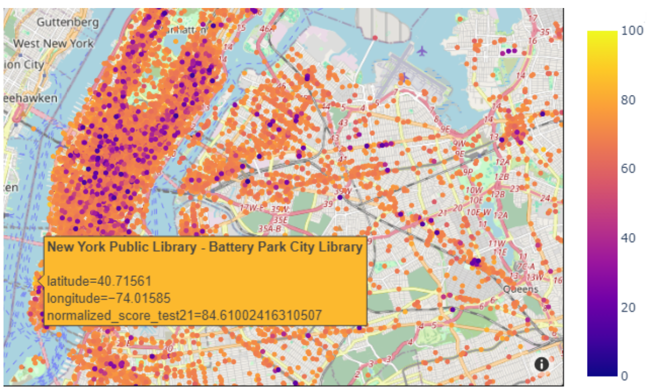

Fig 8. Distribution of Significance Scores across NYC

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 “New York Public Library”, “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 have a significant number of local chains like “2 Bros Pizza” and “The Halal Guys”

IV. Cluster Analysis

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.

To cluster similar types of businesses, we initially performed dimensionality reduction using Principal Component Analysis (PCA). To eliminate potential bias from categorical variables, we excluded the "chain_type" feature, which designates whether a business is independent or part of a chain. Including this feature resulted in clear clusters due to its three values (0 - Independent, 1 - Local Chains, 2 - National Chains). We aimed for a more robust clustering based on relevant features such as "review_counts," "chain_avg_dist," "walk_scores," "cultural_sentiment," etc. Thus, PCA was applied to eight numerical features: ['review_count', 'rating', 'price', 'walk_scores', 'chain_count', 'chain_avg_dist', 'emp_avg_pay', 'cultural_sentiment'].

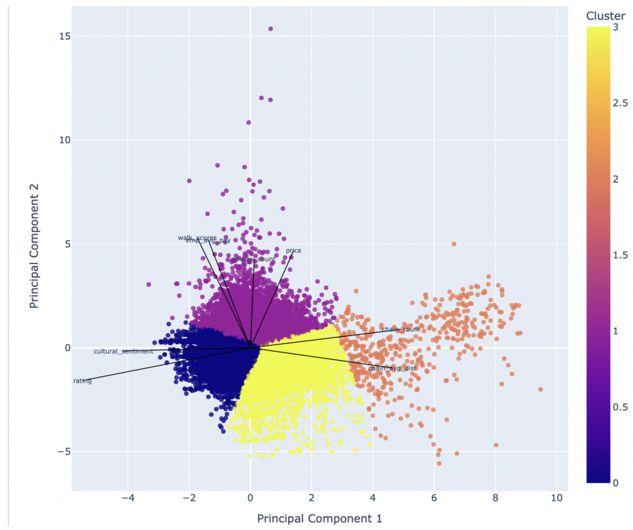

Fig. 12. PCA and Biplot

Fig. 12 illustrates the resulting biplot, enhancing the interpretation of the PCA plot. In a biplot, acute angles between features indicate a positive correlation, while opposite directions suggest a negative correlation. Notably, "walk_score" and "emp_avg_pay" are closely aligned, indicating a positive correlation, while "chain_count" opposes "cultural_sentiment," supporting our theory that chains have lower cultural sentiment and, thus, lower significance scores.

Upon analyzing each cluster, we observed that national chains, such as Target, 7-Eleven, and CVS Pharmacy, form the orange cluster on the extreme right. Businesses on the right exhibit significantly more outlets and higher chain count. Conversely, the leftmost blue cluster showcases culturally significant places, including churches, temples, mosques, synagogues, museums, cafes, libraries, and restaurants, characterized by high cultural sentiment scores and ratings.

The yellow and purple clusters contain a mix of independent and chain restaurants. Points in the purple cluster, on average, have higher significance scores, higher employee salaries, greater walkability, and higher prices. The left part of the purple cluster features independent businesses, while the right side contains businesses with multiple branches and lower significance scores. On the other hand, the yellow cluster encompasses a blend of independent and chain businesses, with lower average salaries, less walkable neighborhoods, and lower prices.

An important observation is the positive correlation between "chain_avg_dist" and "chain_count," aiding in differentiating local chains from national chains. As discussed in the chainness metric section, national chains exhibit greater distribution(average distances), validated by the biplot's alignment of the distance and chain count axes, whereas local businesses tend to open outlets in a closer radius than renowned national chains.

We took our analysis a step further using the Gaussian Mixture Model (GMM) technique to categorize businesses into four distinct clusters. This helped uncover hidden patterns in the diverse business landscape of New York City. Building on the findings from PCA and cluster analysis, GMM added a more detailed layer to our understanding. By considering the data's distribution characteristics, GMM allowed us to identify subtle groups that might not be immediately visible with other clustering methods. This approach deepened our insights into the complex relationships among businesses

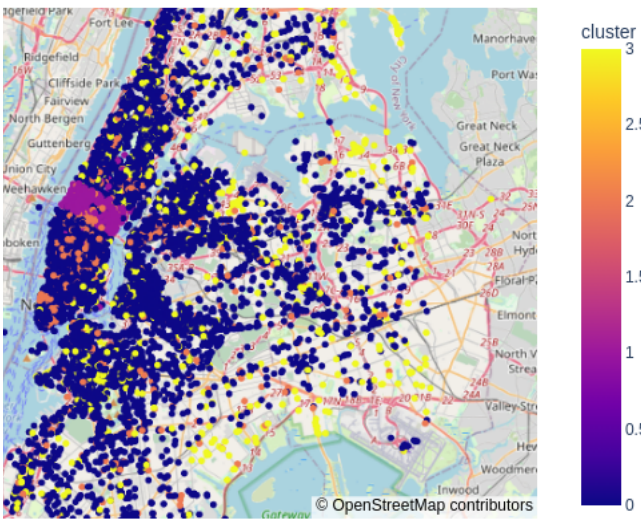

Fig 13. Distribution of clusters across NYC

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 analyzed 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.

V. External Validation

To validate our findings, we cross-referenced our significance scores with respected sources featuring lists of top businesses in various categories, such as restaurants & bars, libraries, museums, and general-purpose stores. Initially, we consulted Secret NYC's compilation of the most famous libraries in New York City [13], comparing our metric's scores for these culturally significant libraries:

Name of Library	Significance Score
New York Public Library	84.61
The Morgan Library & Museum	81.15
Brooklyn Public Library	75.46
Jefferson Market Library	76.92
New York Academy of Medicine	74.68
New York Public Library for the Performing Arts, Lincoln Square	72.50
Poets House	86.99
New York Society Library	73.62
Butler Library	74.23
Stavros Niarchos Foundation Library	75.92

Additionally, we referred to Timeout's list of the top 10 culturally significant, old, and famous restaurants in New York City [12] and compared our scoring metric against it.

Name of Restaurant	Significance Score
Katz’s Delicatessen	70.44
Peter Luger	73.87
Sylvia's	75.25
Nathan's Famous	70.03
Keens Steakhouse	78.88
Tavern on the Green	73.67
Bamonte’s	73.27
Junior’s Restaurant	75.19

Our results reveal that all the libraries and eight of the top 10 restaurants listed are present in our findings with notably high significance scores. This alignment underscores the accuracy of our scoring system, correctly assigning higher values to these businesses that are renowned, culturally significant, and deeply rooted in the community.

In a final assessment, we examined the scores assigned by our model to popular national chains that maintain multiple franchises across the country:

Name of Restaurant	Significance Score
7-Eleven	8.37
Target	16.67

Duane Reed	22.88
Stop and Shop	6.08
ALDI	21.54

It is notable that all the national chains are present in our results, aligning with our criteria of assigning lower values to businesses that are less culturally significant and lack deep community roots. The observed low significance scores for these national chains confirm the effectiveness of our scoring system in distinguishing businesses with a lower cultural and community impact.

VI. Discussion

Following an in-depth analysis of our work, we made some observations regarding the distribution of independent and non-independent businesses in New York City. Primarily, we noted that the majority of culturally significant places are concentrated in the Manhattan region. While there are notable places in other boroughs like Brooklyn or Queens, their numbers are relatively smaller, potentially due to the presence of fewer businesses and lower overall foot traffic. Notably, Battery Park City emerged as the area with the highest density of significant places, housing multiple cafes and libraries with high community engagement.

Aligning with the findings of Liang et al.[2], our observations indicate that independent businesses are predominantly situated along the shorelines, while chain businesses are more commonly found inland. This trend is evident in Fig. 8, where businesses with higher significance scores are clustered along the city's shores, while chain businesses tend to be more inland. This pattern also reinforces the significance of Battery Park City, which stands out as the area with the highest concentration of significant places per unit area, situated along the southernmost tip of the city near the Hudson River.

VII. The Road Ahead: Future Investigations

Quantifying the significance score of a business is subjective and open to diverse interpretations. What holds significance for one person might not carry the same weight for another. Consequently, addressing the nuanced nature of this problem requires careful consideration to arrive at a universally accepted metric that can precisely capture the "significance" of a place. For this study, the authors collectively defined specific metrics that were deemed effective in quantifying this aspect of a business.

A more robust approach for future endeavors could involve implementing crowdsourcing to obtain a broader perspective on public perceptions of places in their neighborhoods. Leveraging the opinions and reviews of locals who reside in proximity to a business, or in its vicinity, might offer a more accurate and community-driven understanding of independent restaurants. By tapping into the insights of those intimately familiar with the area, this approach could enhance the precision and comprehensiveness of our significance scoring system.

Another point raised during our discussions emphasizes the need to identify culturally significant places affiliated with national chains. Some national chains, like McDonald's, may have individual branches that hold cultural significance. For instance, a McDonald's branch in Times Square might carry more cultural weight than one in Queens. While we have the year of establishment as a distinguishing factor for branches of the same chain, recognizing their cultural importance demands a more detailed examination, including comprehensive reviews and a nuanced analysis. Future endeavors should explore methodologies to effectively identify and account for the cultural significance of specific branches within national chains.

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]
12. NYC’s most iconic restaurants worth visiting at least once [Link]
13. 10 Most Beautiful And Best Libraries In NYC To Spend A Day At [Link]