Preserving Nightlife Cultural Spaces:

Keeping the "City That Never Sleeps" Awake

Vivian Chen (vc1176), Ani Harnur (ash645),
Andrew Norris (ahn288), Christine Vandevoorde (cmv398)
New York University CUSP, Spring 2020

Sponsor: Office of Nightlife, NYC Mayor's Office of Media and Entertainment (MOME)
CUSP Mentor: Kim Mahler

Table of Contents

Abstract

New York City nightlife is a bustling $35.1 billion industry with world-renowned culture and community that draws people from all over the world (NYC MOME, 2019). With rising rent costs, gentrification, and most recently, Covid-19 shutdowns, the NYC Mayor's Office of Media and Entertainment (MOME) is exploring supportive policies that can help the nightlife industry recover, adapt, and thrive for years to come. By analyzing MOME's Covid-19 survey and incorporating neighborhood factors known to affect the well-being of the nightlife industry, the team will develop an interactive dashboard that displays the neighborhoods predicted to be at highest risk of experiencing permanent nightlife cultural loss. Additionally, text analysis of widely used sources describing local nightlife will reveal each neighborhood's defining cultural characteristics, thus contextualizing and humanizing potential losses.

Introduction

The unprecedented three-month shutdown resulting from Covid-19 has disproportionately affected the service and entertainment industries. Continued reports of long-standing venues permanently closing due to the extended shutdown illustrate both the sensitivity of these businesses to disruptions in the local economy, as well as expose the persistent barriers to operating these businesses face. One restaurant owner describes the pandemic as "the straw that broke the camel's back" in "an industry that has been dying for a long time" (Fortney & Dai, 2020). As dozens of culturally-rich NYC venues have already announced permanent closure in the past few months, MOME is urgently assessing the potential short- and long-term effects on NYC's nightlife industry and developing a response to minimize business closures and lasting effects on the industry (Fortney & Dai, 2020).

Problem Statement

Due to Covid-19 closures and rising rent costs, many nightlife businesses face the risk of permanent closure. The team will predict which neighborhoods are at the highest risk of permanent loss of nightlife, and characterize each neighborhood's nightlife culture through text analysis to understand what is at risk of being lost.

Literature Review

Prior to Covid-19, chronic barriers were already hindering the nightlife industry. In the 2019 NYC Nightlife Economy Report, 87% of nightlife venue owners reported rising commercial costs as an operational challenge, the most frequently cited issue, followed by regulatory red tape at 68% (NYC MOME, 2019). Linked to rising costs of space, the effects of gentrification were changing the unique nightlife footprint found in neighborhoods.

There are very few nightlife-culture-specific studies. In “Industry Dynamics and the Value of Variety in Nightlife: Evidence from Chicago,” Jacob Cosman utilizes nightlife venue entries and exits via liquor licensing datasets to show that consumers have a strong preference for nightlife areas with a high variety in offerings (2017). While this study is close in nature to this research, the team is concerned with identifying the quality of nightlife, rather than quantifying usage of nightlife spaces.

Sohrab Rahimi et al.’s study, “The Geography of Taste: Using Yelp to Study Urban Culture,” uses natural language processing techniques to predict socioeconomic geospatial boundaries from consumer’s Yelp reviews and dining habits (2018). Their model involves scraping Yelp text, filtering out English stop/filler words, and categorizing the remaining words into foods and drinks, food adjectives, and ambience adjectives. The group was successful in creating geographic clusters from their text analysis that corresponded with block-level per capita income within their study areas.

The combined impact of venues within a neighborhood is more descriptive of the overall cultural climate than the influence of individual venues. The team incorporates this precedent by focusing on neighborhoods rather than individual nightlife establishments. Furthermore, there is no established framework for measuring and profiling the cultural value of nightlife at any geographic scale, so the team instead is turning to text analysis techniques to extract cultural descriptions of nightlife from patron and owner descriptions.

The team maintains that nightlife in some neighborhoods will be more resilient to Covid-19 shutdowns than others, evidenced by differences in risk factors known to contribute to the strength of the industry: business losses due to Covid-19, staff cancellations due to Covid-19, count of venues in a neighborhood pre-Covid-19, variety of nightlife venues in a neighborhood, and cost of nightlife commercial space.

Data

The following datasets are used in analysis:

No.	Name / Source	Description	Time Range	Last Updated
0	Survey for Nightlife Businesses, Workers, and Freelancers Impacted by Covid-19 / MOME	Survey for Nightlife Businesses, Workers, and Freelancers Impacted by Covid-19. Includes data on employment changes, financial loss, home/work location, work/venue-type, etc. (.xlsx)	March 2020	March 2020
1	LiveXYZ Venue / MOME	All NYC nightlife venues open past 6pm. Main relevant features include business categorizations, location, and self-description. (.csv)	Jan - Dec 2019	Dec 2019
2	Yelp Business Listings (API Results) / Yelp	Businesses queried by the term “nightlife” within focus zip codes. Main features include business categorizations, location, hours, and 3 most recent reviews truncated to 160 characters each. (.json, .csv)	June 2020	April 2020
3	Zip Code Boundaries / NYC Open Data	Zip codes and their boundaries and population (.shp)	Up to Sept 2018	Sept 2018
4	Borough Boundaries / NYC Open Data	Boroughs and their boundaries (.shp)	May - June 2020	June 2020
5	whats-open-east-village / BetaNYC	Directory of business status - East Village, Manh (.shp)	May - June 2020	June 2020
6	whos-open-brooklyn / BetaNYC	Directory of business status - Brooklyn (.csv)	May - June 2020	June 2020
7	whos-open-queens / BetaNYC	Directory of business status - Queens (.csv)	May - June 2020	June 2020
8	whos-open-uptowngrandcentral / BetaNYC	Directory of business status - Uptown Grand Central (.csv)	May - June 2020	June 2020
9	open-business-directory / BetaNYC	Directory of business status - North Brooklyn (.csv)	May - June 2020	June 2020
10	Property Valuation and Assessment Data / NYC Open Data	Real estate assessment property data updated yearly. Data analyzed covered the past 5 years. (.csv)	FY 2014/15 - 2018/19	May 2020

Table 1: List of Data Sources

The data sources listed were chosen based on their abilities to account for both the quantitative and qualitative aspects of a neighborhood’s nightlife culture (No.’s 1-3) , and to describe the ongoing changes brought on by Covid-19 (No.’s 0, 4-9).

Exploratory data analysis of the Covid-19 survey and LiveXYZ datasets revealed certain neighborhoods were richer in the number of data points available for analysis. As expected, there is a correlation between the amount of data available and presence of nightlife. However, in order to account for a wider range of neighborhoods and reduce perpetuating the possible bias that the nightlife scenes of these neighborhoods hold greater importance than others, additional neighborhoods were chosen for analysis, as suggested by MOME representatives.

Data Limitations

While representative of a range of features pertaining to nightlife, the available data sources present two primary limitations: data quality, and data biases. In regard to data quality, the Covid-19 survey contains a dearth of both qualitative and quantitative data, and while this is a rich data source, the cleaning and enumeration required to make this usable in analysis also results in some information loss (e.g., encoding write-in responses). Additionally, data retrieved from the Yelp API is limited in temporal range (e.g., can only obtain the three most recent reviews, truncated to 160 characters) and its accuracy depends on venue owners making manual updates. The validation data is open sourced, and is limited in coverage of area and business types. With respect to data biases, these analyses utilize data from both public and private sources, and while offering much data, they are limited to venues that are commercially recorded and/or are aware of a survey such as the Covid-19 one being conducted. By aggregating data to analyze trends on a neighborhood scale, the accuracy and relevance of the findings are increased. It is important to note that the data being analyzed only reflects a limited time period of the pandemic, Phase I of NYC's shelter-in-place order. The appropriate ongoing analysis, including more in-depth validation, can take place as additional data is collected for the entirety of the pandemic.

Identifying Regions of Interest - Exploratory Data Analysis

The team filtered the LiveXYZ dataset to focus only on venues with a provided zip code. The team gleaned insight into the types of businesses that make up nightlife in NYC in three ways:

1. Total counts of venue type by zip code
2. Counts of venue type normalized by population in each zip code
3. Percentage of each venue type (in relation to total number of nightlife venues in zip code)

MOME's Covid-19 Impact of Nightlife Business Closures survey provided feedback from business owners, artists, and staff; the team gained insight on loss of income, expenses, and work opportunities. The team normalized each zip code's number of businesses in each sector by the total businesses to understand where the most nightlife venues are in New York City.

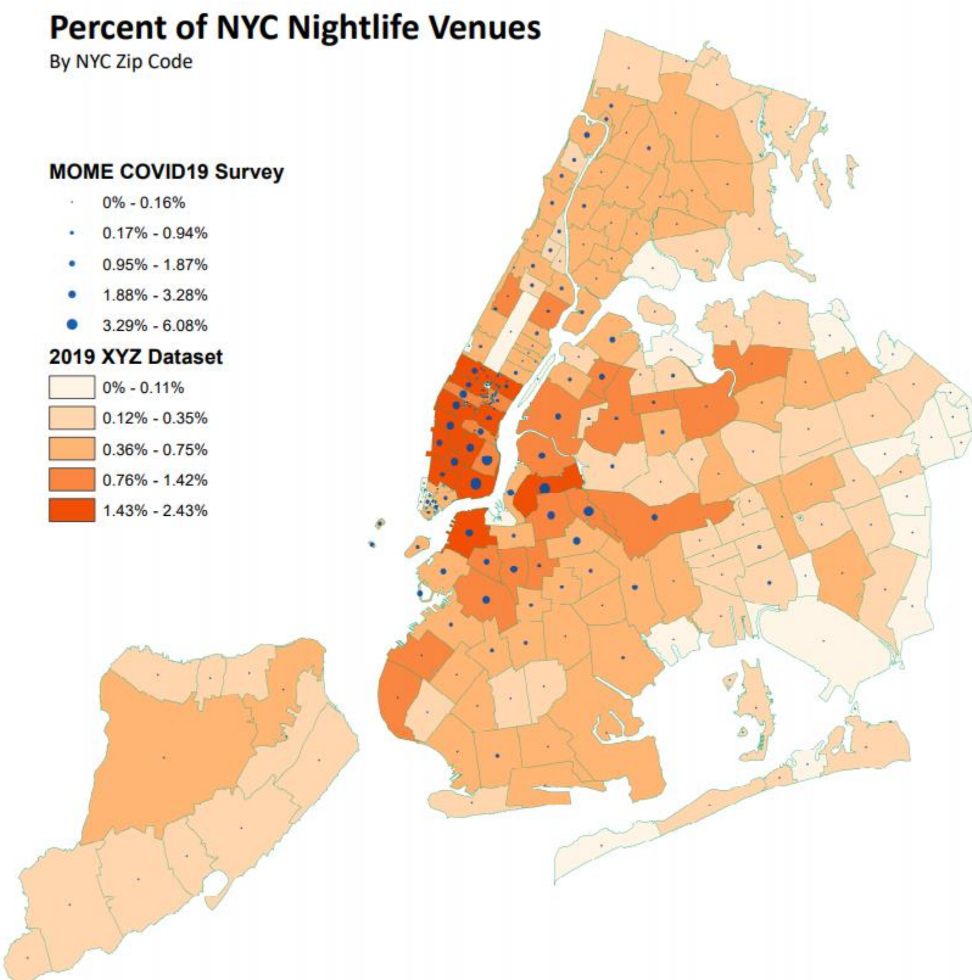

Figure 1: Percent of total NYC Nightlife Venues (Covid-19 Survey, XYZ)

The Covid-19 survey overlaid with the LiveXYZ heatmap data of nightlife business count, helped the capstone team determine the regions of interest to select zip codes that had data availability from both datasets in Manhattan, Brooklyn, and Queens. Additional neighborhoods were also included for analysis based on MOME concern. Selected zip codes, and their corresponding neighborhood, can be found in Table 4 in the Appendix.

Methodology

Determining Risk

Considering the data sources available and research findings from existing literature, the five metrics expected to impact the risk of permanent nightlife loss are listed in Table 2:

Risk Metric	*Measurement*	*Relevance*
Business loss	Measured by cumulative loss for a venue normalized by revenue, as reported by venue owners in MOME's Covid-19 impact survey, summed over all nightlife venues in a neighborhood.	Representative of both immediate and long-term economic losses on a local scale to be reflective of unique nightlife communities.
Staff cancellations	Measured by median percentage of staff cancellations over all nightlife venues in a neighborhood, as reported by venue owners in MOME's Covid-19 impact survey.	Representative of both economic loss, and the loss of human capital that is critical to the strength of the nightlife industry in a community.
Existing venue counts	Measured by the pre-Covid-19 number of nightlife venues per 1000 residents of a neighborhood.	This is indicative of the strength and presence of nightlife venues in local areas.
Cost of commercial space	Measured by the ordinary least squares coefficient of percent change in commercial property values in a neighborhood over the past 5 years.	As proprietary data on commercial rent rates is publicly unavailable, the rate of change of the values of commercial space zoned to allow nightlife venues within a neighborhood is used as a proxy to represent increases in the cost of space on a neighborhood scale.
Variety	Measured by the diversity of venue-types within a specific neighborhood, with relation to the city-wide distribution.	Neighborhoods with greater variety in venue-types are expected to be more resilient to permanent nightlife loss as mentioned in the literature review.

Table 2: Description of Risk Metrics

All 5 metrics were measured at a neighborhood level and standardized (with mean 0, standard deviation 1) so that the scales remained the same across each metric. Then, the variety and venue count metrics were multiplied by negative one so that higher percentiles would equate to higher risk, consistent with the other three metrics. To assess risk, a percentile ranking was assigned for each neighborhood-metric. The average percentile ranking across all 5 metrics was used to determine overall neighborhood risk.

Cultural Characteristics via Text Data - Exploratory Analysis

Once areas of high risk have been determined, it is important to understand exactly what type of nightlife culture is present in that neighborhood to humanize and qualitatively describe the loss. The team conducted text analysis to describe the nightlife culture of each neighborhood in as objective a manner as possible.

The team developed a program to pull and compile a large number of listings from the Yelp API using the search term "nightlife" (chosen for its encompassing range of results) for each zip code in the selected neighborhoods of interest. This retrieved the attributes: geolocation, business category, rating, price, reviews and hours of operation. Reviews and business hours were then queried by business ID and merged with the original dataset on business ID. Yelp only allows the three most recent reviews of each business to be queried and each review is truncated to 160 characters. For consistency, all description data was also truncated to 160 characters. See Tables 5 and 6 in Appendix for the LiveXYZ and Yelp data features used.

To gain an understanding of each neighborhood's nightlife cultural profile, the team performed text and sentiment analysis. Python's NLTK package was used to clean and filter LiveXYZ's Description column and Yelp's customer review column, lending insight into the most frequently used words that businesses and customers use to describe the venues.

The team performed sentiment analysis on the Yelp reviews to understand how positive/negative/neutral patrons feel toward a nightlife venue. To assess these scores on a neighborhood level, all reviews for each neighborhood were aggregated into one cell, allowing the team to determine sentiment on the neighborhood as a whole.

Gaussian Mixture Clustering

As several risk metrics are correlated, the Gaussian Mixture model was selected to account for this relationship. Via the Bayesian Information Criterion Gradient method, the optimum number of clusters was four because the gradient began leveling off at five clusters, the point at which probable overfitting occurs (Figure 7 in Appendix). Cluster assignments were calculated over ten different random states and the mode of the cluster assignments, once aligned, were evaluated to find the most stable cluster configuration, which also happened to be an actual clustering result with one of the random states individually.

Building Validation Dataset

BetaNYC, grassroot organizations, and volunteers have been working together to create an open source dataset of essential businesses and their statuses. By using the updated business statuses (whether they are open or closed), the team will be able to determine the accuracy of the risk percentile. All validation datasets were downloaded before June 22, 2020, the start date of NYC's Phase II. The business names and addresses were matched to the LiveXYZ table by using exact matching and probabilistic matching with the fuzzywuzzy Python package.

Decision Tree Regression

The team used the percent confirmed closed from the validation dataset as the target variable in the Decision Tree analysis. Due to the small sample size, leave-one-out cross validation was chosen to build as robust a model as possible. In doing so, the team hopes to account for overfitting by choosing the most appropriate depth for the tree based on averaging the results of the leave-one-out method.

Results & Discussion

Final Deliverable

The final deliverable for this project is an interactive dashboard. The dashboard shows two main concepts: risk and neighborhood profile. Risk shows each neighborhood's percentile ranking for the five risk metrics as well as their overall risk, which is the mean of all five. Neighborhood profile shows the local cultural characteristics and sentiment analysis reflecting consumer attitudes.

Risk Analysis

Figure 2 shows the results of the average percentile ranking from the five risk metrics across the 17 NYC neighborhoods of focus. The map reveals that the Manhattan neighborhoods below Central Park (Chelsea/Times Square, Lower East Side), as well as Long Island City/Astoria and the South Bronx are at less risk of seeing venues permanently close compared to the Manhattan neighborhoods above Central Park (Harlem and Inwood) as well as some popular Brooklyn neighborhoods, with Bed-Stuy/Bushwick being the most high-risk.

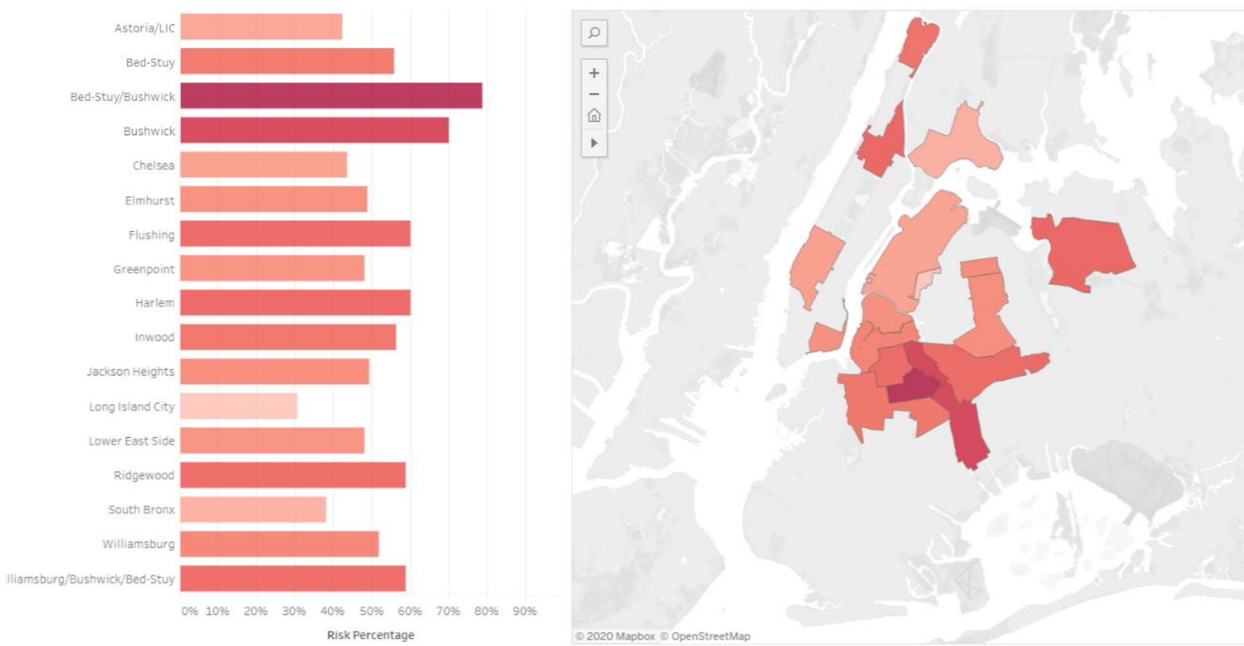

Figure 2: Heat map of Average Percentile Rankings (higher value represents greater risk)

Neighborhood Profile

Figure 3 shows the Variety metric (raw value) as it relates to neighborhood profile. Each of the five metrics are displayed individually on the dashboard. The dashboard also highlights the differences in the word usage between the LiveXYZ self-described venue descriptions versus customer word usage retrieved from Yelp in the form of two separate word clouds for each neighborhood. A noticeable trend among most neighborhoods is that LiveXYZ saw venues frequently reference the neighborhood location. Initial findings show decipherable differences between keywords that describe cultural characteristics across neighborhoods. Figures 8 and 9 in the Appendix are some examples of these word clouds. The profile dashboard also includes the results of the sentiment analysis, showing how positive, negative, or neutral customers feel towards venues in each of the 17 neighborhoods (Figure 10 in Appendix).

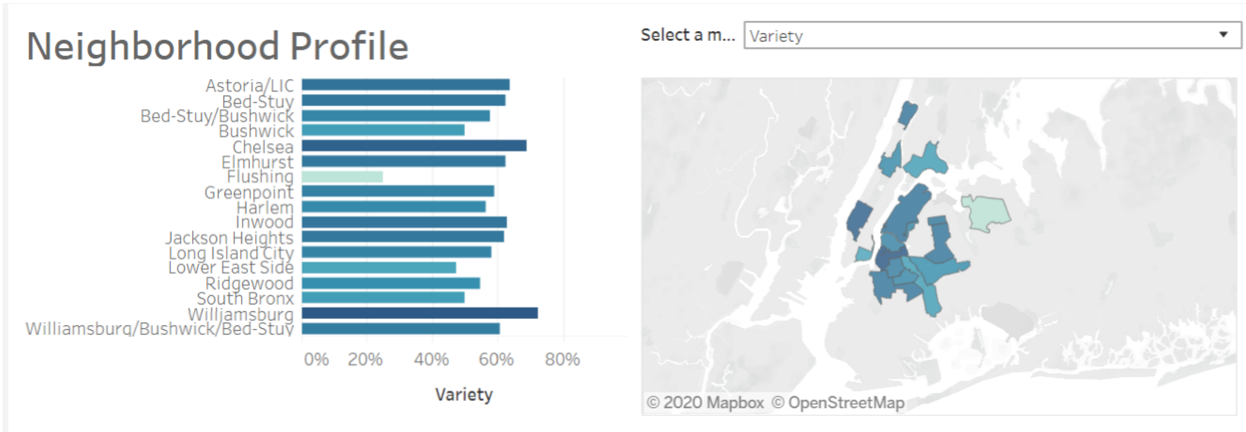

Figure 3: NYC Variety Metrics

Gaussian Mixture Clustering

As shown in Figure 4, four distinct clusters were identified. Chelsea clustered independently, likely because it had considerably more venues per capita than the other neighborhoods. The resulting clusters from GMM revealed clear patterns within each group of neighborhoods based on risk metrics, as shown in Table 3.

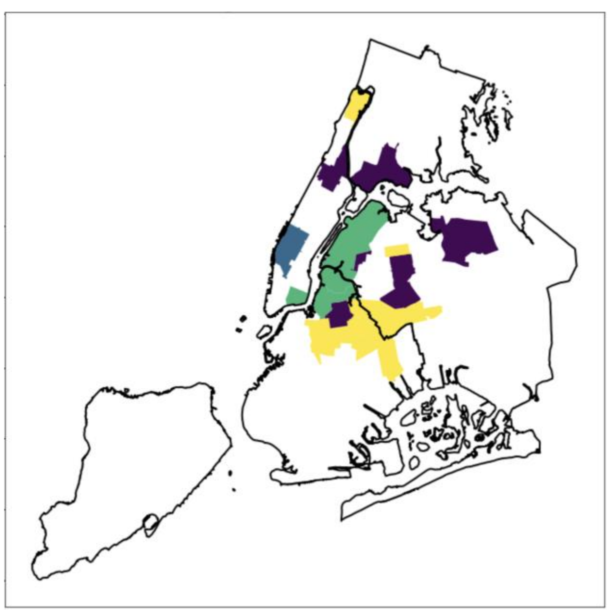

Figure 4: Cluster Assignments based on Risk Metrics

Cluster	Mean Business Loss	Mean Staff Cancellation	Mean Venues per Thousand People	Mean OLS of Percent Increase in Commercial Property	Mean Variety
Williamsburg/Bushwick/Bed-Stuy, Elmhurst, Flushing, Harlem, Long Island City, South Bronx (Purple)	1.67 (167%)	0.75 (75%)	2.50	0.03 (3%)	0.52 (52%)
Chelsea (Blue)	3.24 (324%)	1.00 (100%)	17.25	0.03 (3%)	0.69 (69%)
Astoria/Long Island City, Greenpoint, Lower East Side, Williamsburg (Green)	2.38 (238%)	1.00 (100%)	6.20	0.04 (4%)	0.61 (61%)
Bed-Stuy, Bed-Stuy/Bushwick, Bushwick, Inwood, Jackson Heights, Ridgewood (Yellow)	2.73 (273%)	1.00 (100%)	2.94	0.04 (4%)	0.58 (58%)

Table 3: Cluster Summary

Validation

Figure 5 shows the amount of confirmed businesses the team has from the open sourced tables after matching the business names and addresses. Chelsea, Inwood, and South Bronx are the only neighborhoods without validation data.

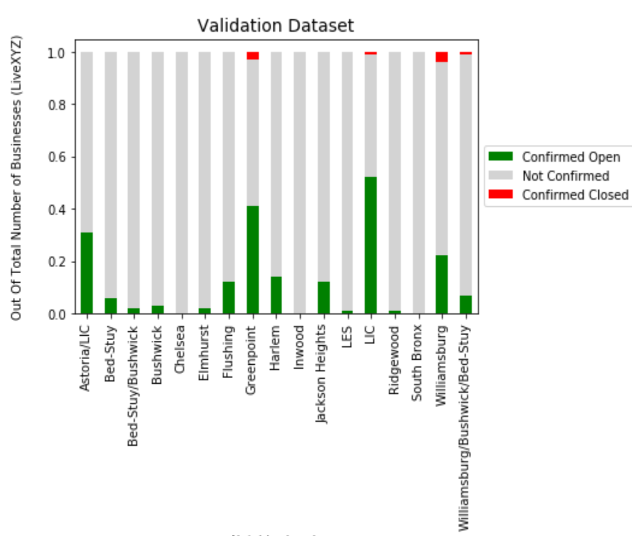

Figure 5: Operation Status as of June 19, 2020

Decision Tree Regression

Due to the small sample size of the data set and an ongoing validation data set, the decision tree should be updated as the operational status of venues continues to be confirmed. Leave-one-out cross validation showed that, on average, the best depth for the tree was two branches. The resulting tree and feature importance (Figure 6) showed that the most significant feature in determining risk of permanent closure is the change in commercial real estate property values - a problem that exists even without Covid-19 already posing challenges.

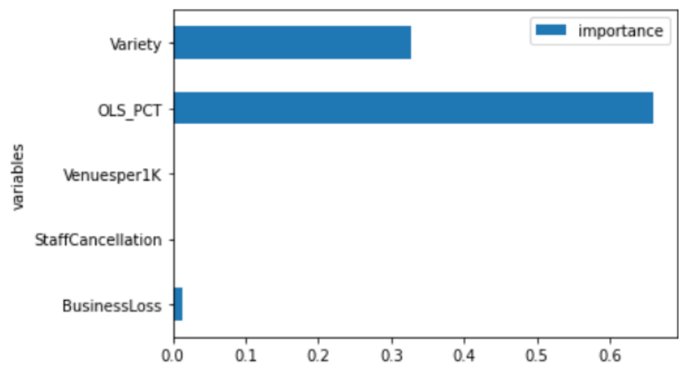

Figure 6: Decision Tree Feature Importance

Recommendations for Further Research

MOME now has the framework to continue this research as NYC enters Phases III and IV of the reopening plan. Once the validation dataset is larger in size, the decision tree model can be re-run to further understand which risk metrics are most indicative of a neighborhood experiencing permanent venue closures.

Due to the sensitive nature of businesses, the team could not gain access to granular business data (actual income, revenue, loss, attendance, etc.) that would provide an in-depth analysis of other risks that nightlife venues face in trying to maintain a physical presence in NYC.

MOME should also consider partnering with Beta NYC and other neighborhood organizations, to help in the data acquisition process and to obtain a nuanced understanding of how neighborhoods are reacting to both Covid-19 as well as other economic obstacles.

Conclusion

Demonstrating not only the economic but also the cultural contribution of the City's nightlife community is now more important than ever, to make sure they are not forgotten during the economic rebuild from the pandemic. Results up to this point indicate considerable cultural variation between neighborhoods and differences in potential risk factors. The team's interactive dashboard highlighting relative risk and neighborhood profiles gives MOME an additional resource to aid them in crafting locally-tailored policies.

Team Roles

1. Andrew Norris - Business information query; GitHub management; MOME Covid-19 survey cleaning; sentiment analysis of descriptions; word count visualizations; Property valuation OLS; Gaussian mixture clustering; website development
2. Ani Harnur - LiveXYZ venue filtering and heatmap plots; NLTK/NLP for XYZ neighborhood descriptions and word counts; Yelp vs LiveXYZ word clouds; venues per 1000 metric; standardizing risk factors and risk heat maps; dashboard input; decision tree; sentiment analysis; Github ReadMe (process maps)
3. Christine Vandevoorde - Researched alternative APIs; Yelp reviews and hours query by business ID; pulled zip code csvs; Yelp data cleaning; property valuation OLS; Gaussian mixture clustering; decision tree; website development
4. Vivian Chen - MOME Covid-19 survey cleaning; geospatial visualizations of survey and LiveXYZ data; Yelp data cleaning; word count visualizations; gathering and cleaning validation data; creating interactive dashboards; website development

Deliverables

1. Website: https://nycnightliferesilience.github.io/index.html
2. GitHub (code repository): https://github.com/nycnightliferesilience/nyu_cusp_capstone
3. Dashboards: https://public.tableau.com/profile/nycnightliferesilience#!/

References

BetaNYC. (2020). *open-business-directory*. GitHub. BetaNYA. (2020). *open-business-directory*. https://github.com/BetaNYC/open-business-directory.

BetaNYC, Alan Baglia, et al. (2020). *whos-open-queens*. GitHub. https://github.com/zhik/whos-open-queens.

BetaNYC, Brooklyn Chamber of Commerce. (2020). *whos-open-brooklyn*. GitHub. https://github.com/zhik/whos-open-brooklyn.

BetaNYC, East Village Community Coalition. (2020). *whats-open-east-village*, GitHub, https://github.com/zhik/whats-open-east-village.

BetaNYC, Uptown Grand Central. (2020). *whos-open-uptowngrandcentral*. GitHub. https://github.com/zhik/whos-open-uptowngrandcentral.

Cosman, Jacob. (2017, April 5). Industry dynamics and the value of variety in nightlife: evidence from Chicago. http://jacobcosman.ca/wp-content/uploads/2017/05/job_market_paper.pdf.

Fortney, L., Dai, S. (2020 June 18). *A Running List of NYC Restaurants That Have Permanently Closed During the COVID-19 Crisis*. NY Eater. https://ny.eater.com/2020/5/8/21248604/nyc-restaurant-closings-coronavirus.

Foster, Richard et al. (2018). *Creative Footprint New York City Music (CFP NYC): Measuring Live Music Space in NYC*. https://www.creative-footprint.org/new-york/#content.

Kensinger, Nathan. (2020, June 16). *"We Are Bleeding Out": Bar & Restaurant Owners Wonder How They'll Ever Come Back From Coronavirus Shutdown*. Gothamist. https://gothamist.com/food/bars-restaurant-nyc-rent-pandemic-coronavirus-shutdown.

Montalto, V., Jorge, C., Moura, T., Langedijk, S., Saisana, M. (2019, June). Culture counts: An empirical approach to measure the cultural and creative vitality of European cities. *Cities, 89*, 167-185. https://www.sciencedirect.com/science/article/pii/S0264275118303640#s0015.

NYC Department of City Planning. (2020 June 16). *Borough Boundaries*. NYC OpenData. https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm.

NYC Department of Information Technology & Telecommunications. (2018 September 10). *Zip Code Boundaries*. NYC OpenData. https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u/data?no_mobile=true.

New York City Mayor's Office of Media and Entertainment (2019). *NYC Nightlife Economy Impact, Assets, and Opportunities.* https://www1.nyc.gov/assets/mome/pdf/NYC_Nightlife_Economic_Impact_Report_2019_digital.pdf.

North Brooklyn Neighbors. (2020). *Essentially Open North Brooklyn*. http://northbrooklynneighbors.org/project/essentially-open/.

Rahimi, S., Mottahedi, S., & Liu, X. (2018). The Geography of Taste: Using Yelp to Study Urban Culture. *ISPRS International Journal of Geo-Information*, 7. 376. https://www.researchgate.net/publication/327642523_The_Geography_of_Taste_Using_Yelp_to_Study_Urban_Culture.

UNESCO. (2014). *Culture for Development Indicators Methodology Manual*. UNESCO. https://en.unesco.org/creativity/sites/creativity/files/cdis_methodology_manual_0_0.pdf.

UNESCO. (2001, November 2). *Universal Declaration on Cultural Diversity*. UNESCO. http://portal.unesco.org/en/ev.php-URL_ID=13179&URL_DO=DO_TOPIC&URL_SECTION=201.html.

Appendix

Neighborhood (Borough)	**Zip Codes**
Astoria (Queens)	11101,11102,11103,11105,11106
Long Island City (Queens)	11101,11102,11103,11105,11106,11104,11109,11120
Williamsburg (Brooklyn)	11206,11211,11249
Bushwick (Brooklyn)	11206,11207,11221,11237
Harlem (Manhattan)	10026,10027,10037,10030,10039
Chelsea (Manhattan)	10001,10011,10018,10019,10020,10036
Lower East Side (Manhattan)	10002
Inwood	10034, 10040
South Bronx (Hunts Point and Mott Haven)	10454, 10455, 10459, 10474
Bed-Stuy	11205, 11206, 11216, 11221, 11233,11238
Ridgewood	11385, 11386

Table 4: Regions of Interest, Neighborhoods determined by Zip Code(s)

Main Category	Highest level grouping identifier
Sub-category	Second level grouping identifier
Post Code	Location Zip Code
description	Business Description

Table 5: LiveXYZ Data Features

categories	Business-type categories
location	Full venue address and zipcode
rev1, rev2, rev3	3 most recent reviews, truncated to 160 characters apiece

Table 6: Yelp Data Features

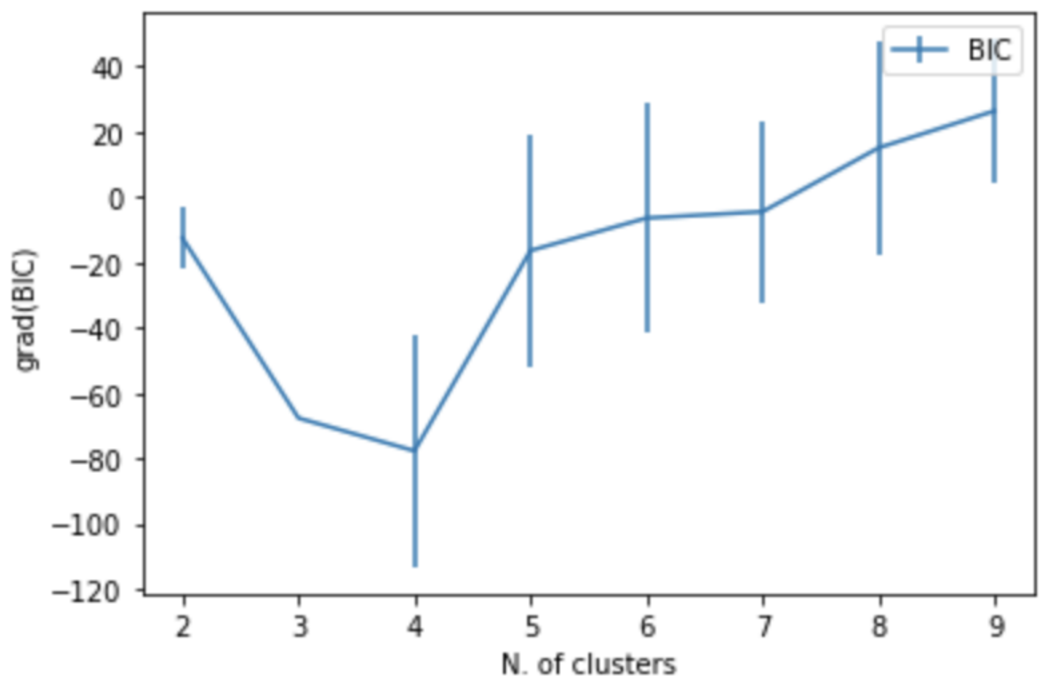

Figure 7: Gradient of BIC Scores from Gaussian Clustering

Figure 8: Lower East Side word clouds, LiveXYZ (left) vs Yelp (right)

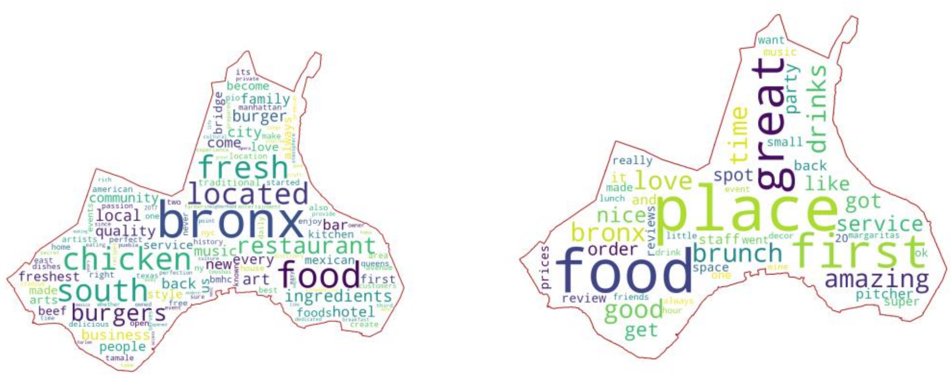

Figure 9: South Bronx word clouds, LiveXYZ (left) vs Yelp (right)

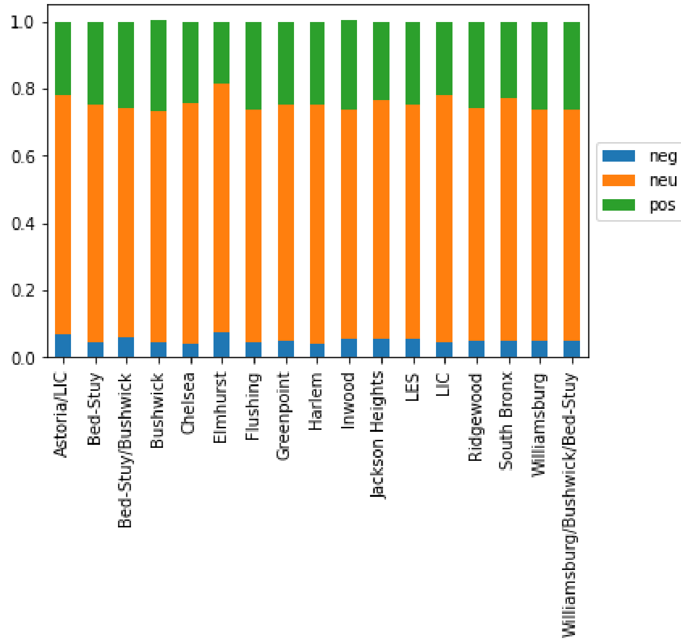

Figure 10: Yelp Sentiment Analysis