Best Locations for a Night Club in New York

# Introduction

## Background

Assume an entertainment company is choosing a location for a night club in New York City. This report will suggest the top three neighborhoods for the night club by analyzing data obtained using the Four Square API and data for New York City's neighborhoods. In addition, an appropriate genre of music will be recommended from the following: EDM (Electronic Dance Music), Trap, Hip Hop, Trance, or House based on the neighborhood's proximity to other clubs.

## Problem

Neighborhoods will be plotted and clustered based on their night life. A neighborhood which already has a medium amount of night life which is not already saturated by night clubs of the same genre would be a candidate.

Said entertainment company wants to choose an optimal location for their night club which will attract a lot of patrons. Locating the club in a popular, trendy area of the city is important. Neighborhoods with other clubs within walking distance are a good bet for oftentimes club patrons will hop from club to club, especially, if there are other genres of music close by. Thus, it will be important to locate the club in an area which is not already saturated by clubs of the same genre of music. That is, we want to avoid placing an EDM club in an area already saturated by other EDM clubs. One club of the same type might be okay for you will have curious club goers that like EDM music want to try a different venue and they are already nearby but if there are already two or three EDM clubs within walking distance or a five minute cab/uber ride that neighborhood should be avoided.

## Interest

The bar and night club industry has steady grown since the mid 1990’s reaching 23.15 billion U.S. dollars in 2015 [1]. Running a night club can cost several million dollars just for a yearly lease depending on location. A well-placed dance club can gross revenues anywhere from $5,000 to $35,000 nightly [2]. As Harold Samuel, the real estate magnate, once said “Location, Location, Location”. Location is of major influence on a brick-and-mortar business and how successful it will become.

# Data Description

## Data Sources

As mentioned above, the data used will be venue data obtained using the Four Square API and New York neighborhood data obtained from [https://geo.nyu.edu/catalog/nyu\_2451\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572%20) . This is the same data that was used in a previous assignment for this course. This data is used to identify the location of each neighborhood in New York using longitude and latitude. The Foursquare API is used to access night club venue information.

There was a third source of data that defined the boundaries of each neighborhood. It was downloaded from github at this location: <https://github.com/veltman/snd3/blob/master/data/nyc-neighborhoods.geo.json> . This data set was used to assign clubs into neighborhoods using the actual boundaries of the neighborhoods rather than their distance from some center of a neighborhood. The idea is that neighborhoods are not circles but irregular polygons, and that there is more similarity of a business within a neighborhood rather than across neighborhoods. That is, two streets could be one block over from each, exist in different neighborhoods and be significantly different economically even though they are relatively close in distance.

## Data Cleaning

The data downloaded from FourSquare was reformatted from JSON into a Pandas data frame. Only a few of the downloaded fields were used from the Four Square data set. They were:

[https://geo.nyu.edu/catalog/nyu\_2451\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572%20)

This data set was used to get the geographical centers of the neighborhoods. It was used with the FourSquare API to get all of the clubs within 3000 meters of the neighborhood center. This data set was checked for duplicates, invalid neighborhood names and bad latitude and longitude.

|  |  |
| --- | --- |
| Data Label | Description |
| Borough | Name of the borough the neighborhood is within |
| Neighborhood | Name of the neighborhood |
| Latitude | The latitude of the neighborhoods |
| Longitude | The longitude of the neighborhoods |

FourSquare API

The four square API was used to pull down the follow information. A search radius of 3000 meters was used to find all clubs within 3000 meters of the neighborhood centers defined above.

|  |  |
| --- | --- |
| Data Label | Description |
| lat | Latitude of the club |
| lng | Longitude of the club |
| distance | Distance the club was from the requested search location |
| categories | The type of venue |
| address | The address of the venue |
| labeledLatLngs | Representation of the latitude and longitude as a label |
| postalCode | The zip code of the club |
| cc | Country Code |
| city | The city the club is within |
| state | The state the club is within |
| country | The country the club is within |
| formattedAddress | The formatted address |
| crossStreet | The cross street |
| id | The FourSquare id |

The id field was NaN (not a number), this field was important for the analysis for it was used to look up more information about the venue. It was determined that many of the venues that had a NaN id were also closed temporarily due to COVID-19. This was done by choosing clubs in the candidate neighborhoods, which are the ones with the most clubs, and then checking for ratings. The clubs in the candidate neighborhoods without ratings were double checked by doing some google searches and it was discovered that these clubs without ids were indeed closed. Considering the circumstances and business closing to control the spread of COVID-19 this made sense.

<https://github.com/veltman/snd3/blob/master/data/nyc-neighborhoods.geo.json>

This data set was used to get the geographical boundaries of each neighborhood. Using the neighborhood center data was convenient to use to query the FourSquare API; however, once the clubs were obtained using the FourSquare API it was convenient to group the clubs by neighborhood using the boundaries each neighborhood. This data was converted to Polygons and the club locations were checked to see which Polygon contained the neighborhood. The data was plotted on a map to visual the neighborhoods and ensure that they were correct.

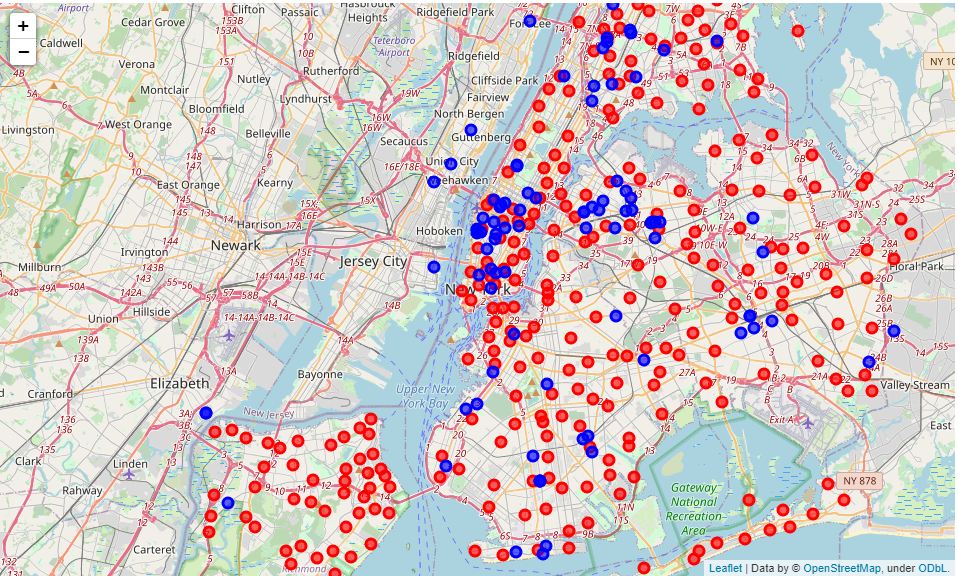
## Feature Selection

The name of the club, the neighborhood the club is contained within using the boundaries of the neighborhood, the neighborhood center that the club is closest to, the longitude of the club the latitude of the club and the id of the club were selected. The id was used to obtain the rating of the club for any club that did not have a rating a negative 1.0 rating was assigned to penalize the neighborhood. A neighborhood with a significant number of closed clubs would be heavily penalized. This was designed to avoid neighborhoods with a significant number of closed clubs. The negative was chosen to ensure that the neighborhood was sufficiently penalized. A more negative rating could also have been used.

# Methodology

## Exploratory Data Analysis

The clubs were plotted on a map of New York along with the neighborhood centers. An example of such a map is shown below. The centers of the neighborhoods are show in red. The club locations are shown in blue:



• Neighborhood centers

• Club locations

Figure 1 New York Neighborhood Centers and Club Locations

The number of clubs per neighborhood were counted and plotted on the choropleth map. The neighborhoods with the most clubs were shown in red:

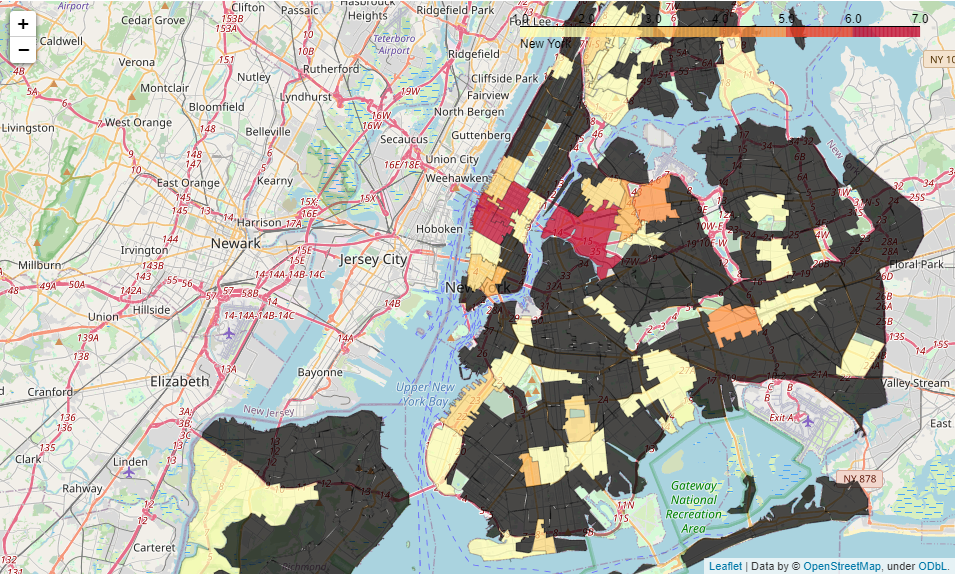


Figure 2 Choropleth Mapping showing Neighborhoods with Most Clubs

## Machine learning

# Results

# Discussion

# Conclusion

### Calculation of Target Variable

Relationship Between Improvement and Age

Relationship between improvement and minutes played

Relationship between improvement and games played

Relationship between improvement and positions

Relationship between improvement and last year’s improvement

Relationship between improvement and draft positions

Relationship between improvement and teams

# Clustering

Regression Models

Applying standard algorithms and their problems

Solution to the problems

Performance of different models

Classification Models

# Conclusions

# Future Directions

References

[1] [https://www.statista.com/topics/1752/bars-and-nightclubs](https://www.statista.com/topics/1752/bars-and-nightclubs/%23:~:text=Sales%20in%20the%20drinking%20place,alcoholic%20beverages%20for%20immediate%20consumption.)

[2] <https://www.referenceforbusiness.com/business-plans/Business-Plans-Volume-07/Nightclub.html>