Coursera Capstone: San Diego Neighborhoods and New Cafe

Nathaniel Fisher
December 11, 2020

Table of Contents

1. Introduction

1.1 Business Problem

The business problem for the capstone project involves finding an ideal neighborhood in the San Diego area to open a new cafe. With over 100 neighborhoods in the city of San Diego, determining the right neighborhood will depend on various factors related to stakeholder desires and existing businesses. The stakeholders are a group of businessmen looking to open a new family owned cafe in the San Diego area. They seek to find a location in which there is a popular breakfast/lunch dining scene, balanced competition, and centrally focused. These three factors help bring foot traffic and tourism that allow customers to discover a new cafe easily. This report will help reveal existing trends in cafe locations with the goal of recommending ideal neighborhoods to have a cafe that sets it up for future success.

2. Data

2.1 Data Sources

Data for the capstone project will be retrieved through two sources. First, a wikipedia page containing the communities/neighborhoods of San Diego will be scraped in order to have the full list of neighborhoods. This dataset will only consist of region and neighborhood names. The raw data will be clean and separated by neighborhood per row. Then, after using geocode in python to receive the coordinates for each neighborhood, location data from Foursqaure will reveal venues in each neighborhood. A Foursquare query will result in a variety of business categories in each neighborhood, and it can provide the name, coordinates, and category for each venue per row. The categorical data should include Cafe as a category, in which the analysis can determine where competitors are located.

2.2 Limitations

There are a few potential limitations with the data described above that must be acknowledged before proceeding. First, geocode in python may not be reliable for gathering coordinates for communities. Considering the size of some neighborhoods, there may be potential problems in this regard. The wikipedia data includes obscure neighborhoods that may not register for geocode and ultimately fail to retrieve coordinates. To overcome these shortfalls, the report did not include neighborhoods within neighborhoods, such as Marston Hills within Hillcrest. Wikipedia data had to be manually reviewed in this regard before and after using geocoder.

Foursquare location data is also dependent on accurate categorical data. For this reason, any restaurants corresponding with breakfast will also be reviewed. However, venue data may not show the full picture or may be missing certain venues.

2.3 Data Cleaning

Data Cleaning involved multiple steps to make wikipedia and foursquare data interpretable. Scraping neighborhood information from wikipedia required renaming and defining the columns. The data retrieved listed all neighborhoods in a few columns without being separated by commas or periods. Therefore I manually updated each row and column by adding commas to identify neighborhoods. This made it easier to split rows based on neighborhoods as separated by commas.

The geocoder sequence also presented challenges by not returning a handful of neighborhoods' coordinates and misplacing another. The missing coordinates were able to be removed because of their small sizes and proximity to larger neighborhoods that would be

included in venue retrievals. In this step, I continually produced coordinates on a map, then reviewed each one closely to ensure accuracy. For some, the map displayed neighborhoods names by default, and I was able to determine whether any were misplaced or incorrect. Through this trial and error method, I corrected the longitude and latitude manually through web searching the coordinates of the given neighborhood. Upon completion of data cleaning, the number of neighborhoods had dropped from 114 to 105.

3. Methodology

Given the maps and review of foursquare data, there is a better idea of which neighborhoods attract larger number of potential customers, tourists, and foot traffic. The methodology for this project will utilize k-mean clustering to group the neighbors together based on venue category and popular categories of the neighborhood. The dataframe consisting of the Breakfast Spot, Cafe, and Coffee shop venues will be utilized and grouped into 3 clusters to determine the areas with high competition and discover any other discrepancies. The analysis will use one hot encoding to initially rank the categories per neighborhood. Then, the kmeans function will cluster each neighborhood based on these results. Finally, the map will reflect these clusters while displaying each individual venue. The idea is to have color coded clusters that can help determine which types of competitors are located in what neighborhoods, how many are in each neighborhood, and reach conclusions as to which may be ideal for stakeholders.

4. Data Exploration and Analysis

Upon receiving coordinates for each neighborhood, the foursquare location data was able to retrieve venue results for each neighborhood. Then the venues were separated to display the most per category. Within this search, three foursquare venue categories presented potential competition to a new cafe: Cafe, Coffee Shops, and Breakfast Spots:

```
Coffee Shop           155
Mexican Restaurant    141
Italian Restaurant    105
Hotel                  94
Pizza Place            81
Café                   63
Sandwich Place         51
American Restaurant    51
Bar                    50
Breakfast Spot         45
```

By sorting by quantity, the Coffee Shop, Cafe, and Breakfast Spot venues were clear picks as competition. Their place throughout is evenly dispersed and aligned with other venue data. Meaning, the downtown areas were still the most popular:

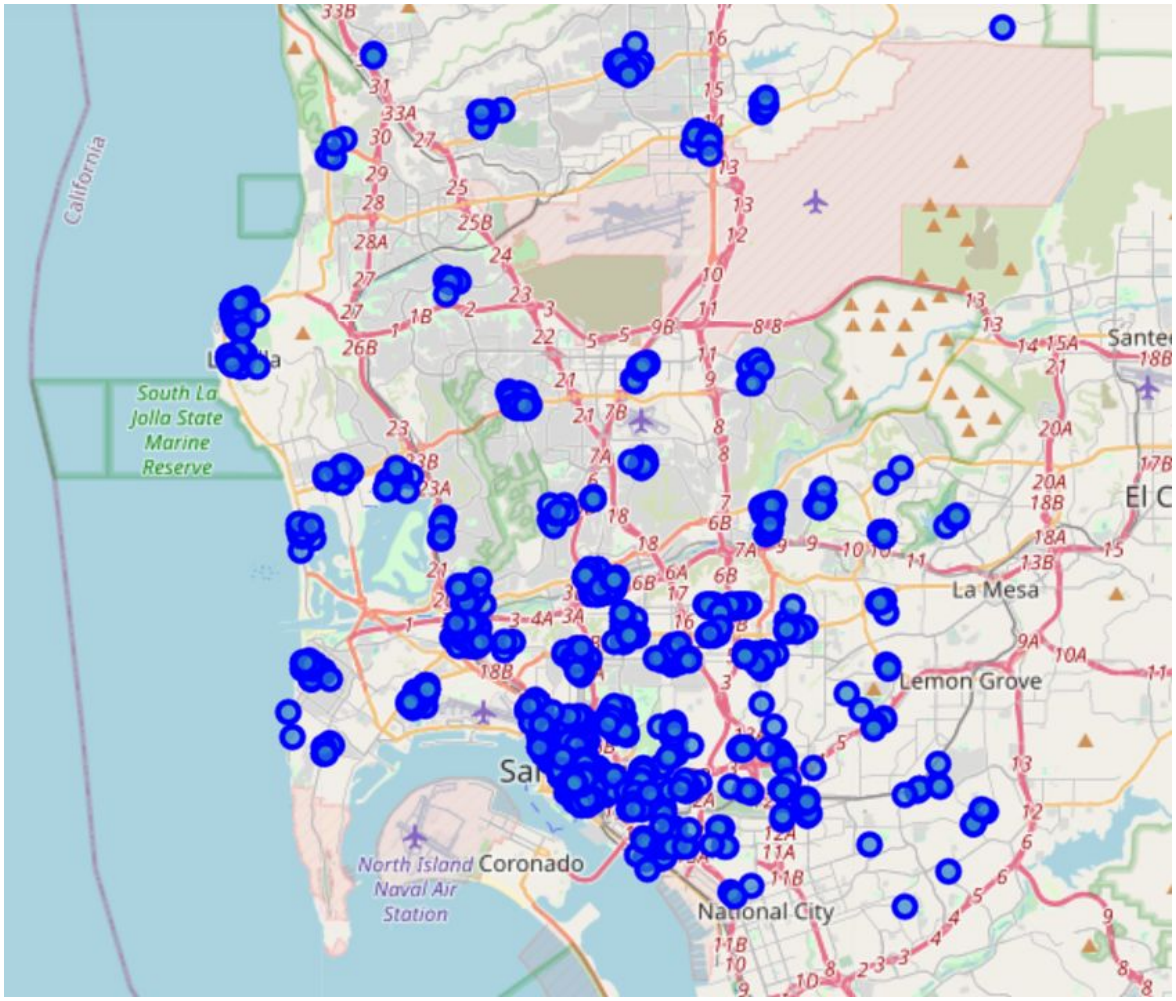


Figure 1. Related cafe venues in San Diego

Downtown San Diego is where most neighborhoods and venues are currently located. The neighborhoods with a higher number of venues suggests popular dining areas, albeit greater competition. However, the quantity is difficult to differentiate given the number of dots. To further explore this data, two additional techniques were used to better understand the relationship of competitive venues and neighborhood. A bar graph and heat map:

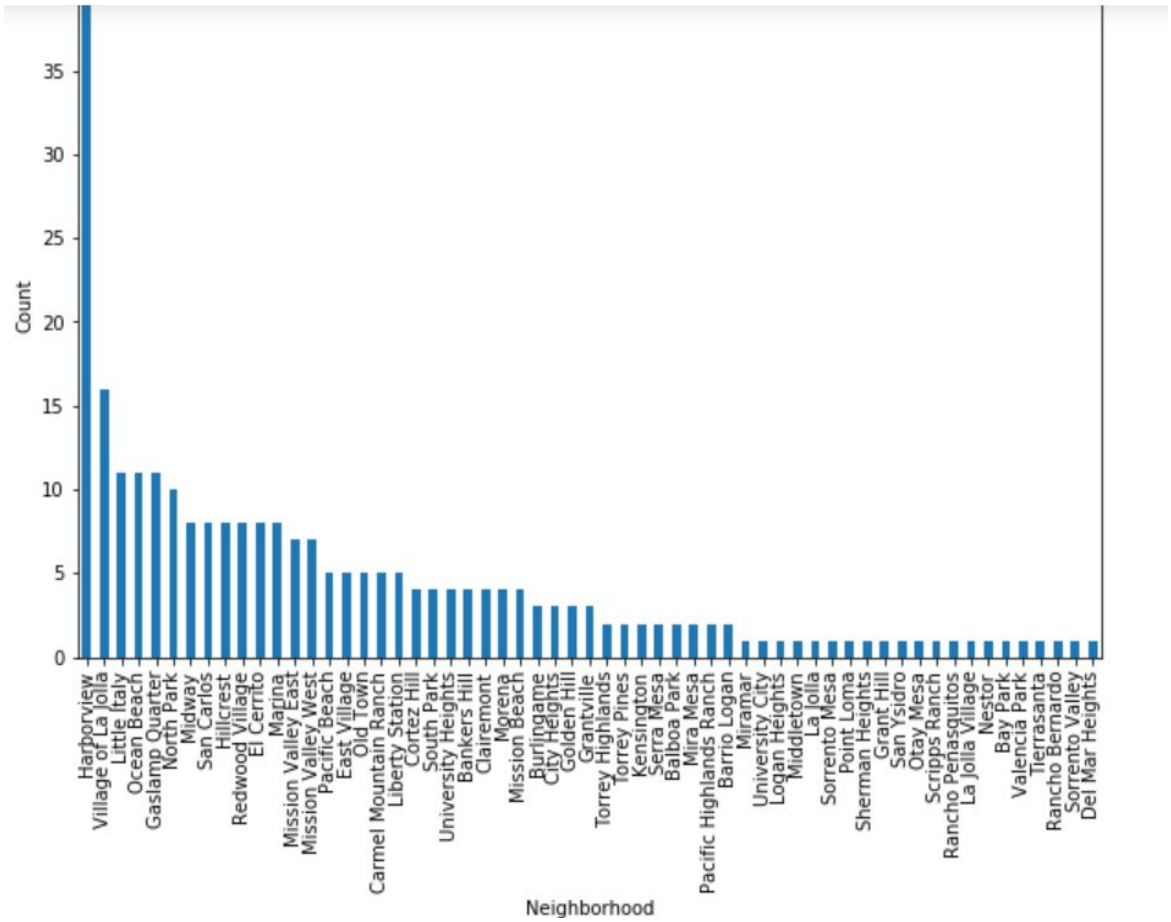


Figure 2. Bar graph of quantity of related cafe venues per neighborhood

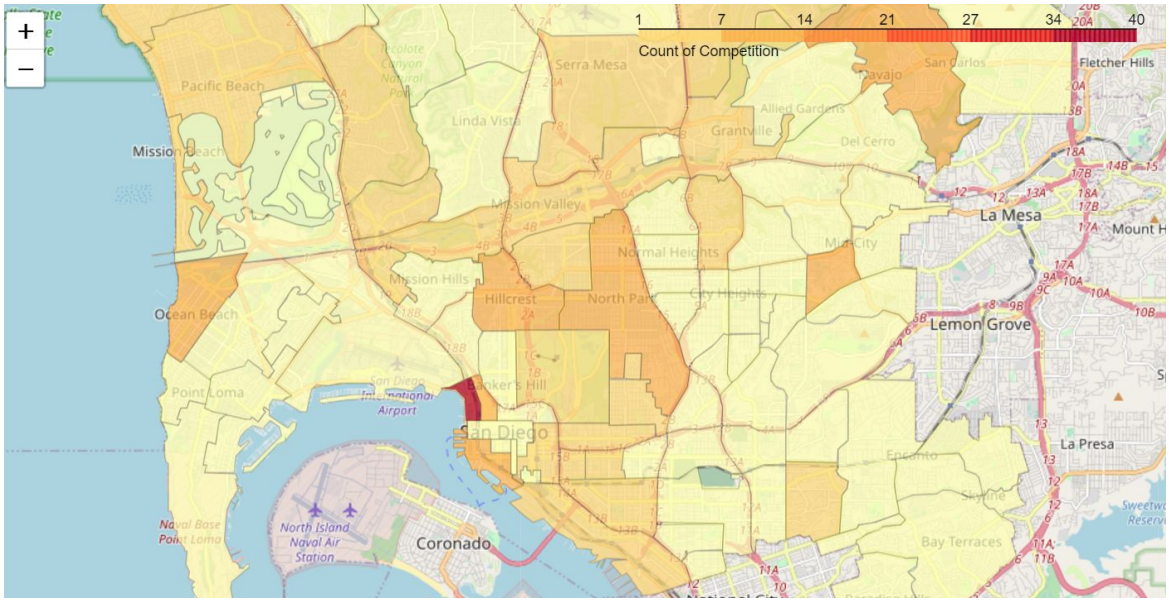
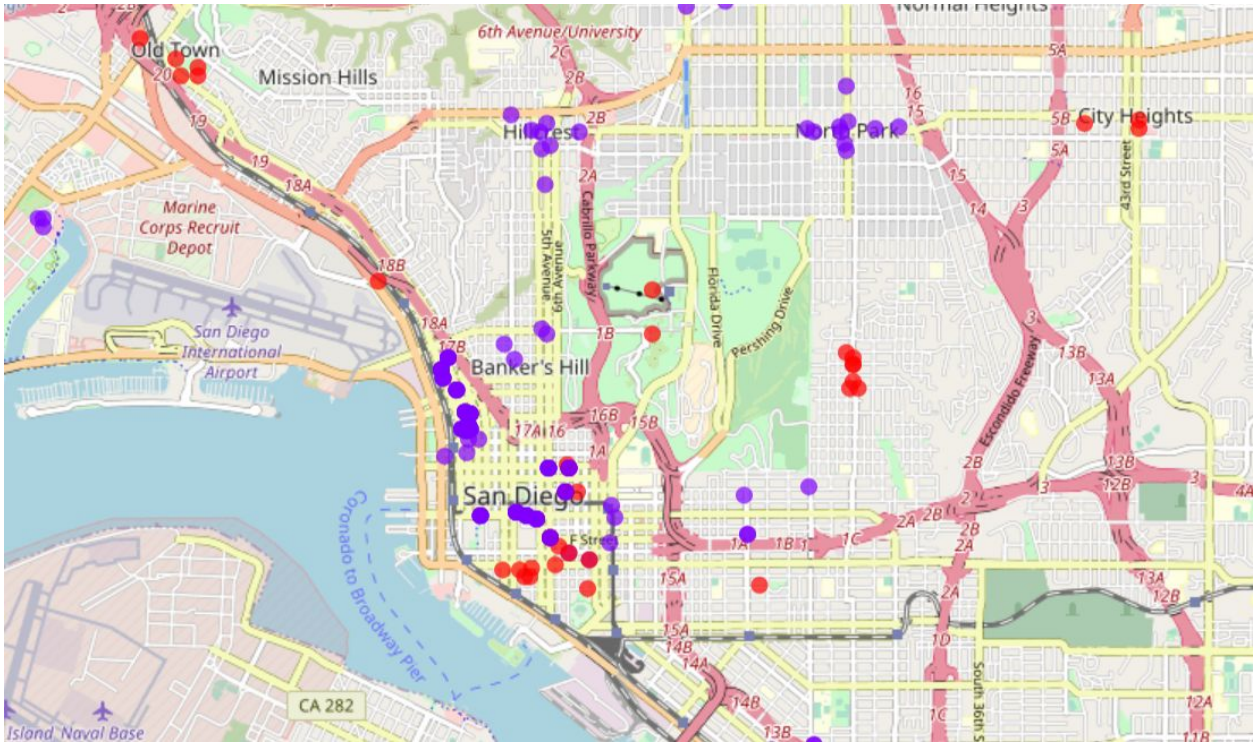


Figure 3. Heat map of quantity of cafe related venues per neighborhood

As expected the popular areas for cafes appears to be centered around western downtown, including Little Italy, Harborview, and the Gaslamp district. Additional neighborhoods with popular cafes are North Park, Hillcrest, Ocean Beach, and La Jolla.

The Kmeans clustering analysis sought to provide a better understanding of each venue's neighborhood in relation to the venue category. The Blue/Purple shade cluster represented neighborhoods that are popular with coffee shops, Red represented neighborhoods that are popular with cafes, and Cyan represented neighborhoods popular with breakfast diners (mostly northern areas of the city). The neighborhoods with a high number of venues are consistent with popular dining areas in the city and would present competition for a new cafe.



5. Results and Discussion

The analysis grouped neighborhoods based on the number of each cafe, coffee shop, and breakfast spot. Given the three categories, the choice of three clusters fit with the idea of finding an area that may be lacking. Based on the cluster analysis, downtown San Diego is filled with cafes and coffee shops. As already indicated from the heat map and value counts, the outer neighborhoods of North Park and Hillcrest are popular areas with many existing cafes. Downtown San Diego (Little Italy, Gaslamp, Harborview) is also popular, with central locations and existing cafes. Based on the stakeholder's request of locations that have a healthy balance of competition, these locations may present too many challenges to gain market share. Therefore, the outer areas of Bankers Hill and East Village appear adequate neighborhoods that are centrally

focused to the downtown area. These locations only have a few coffee shops, no visible cafes, and are near neighborhoods with an abundance of dining options. A quick search of businesses in the areas also indicate these areas are not industrial and contain some residencies. Residents in these neighborhoods may prefer a closer option to dine for breakfast/lunch.

Another option that may be less centrally focused for stakeholders is Mission Beach. As the cluster analysis indicated, Mission Beach only contained breakfast spots and may benefit a smaller cafe. Particularly this could target customers who prefer a meal or coffee later in the day. The number of breakfast spots in this area indicate a large enough population that could be beneficial for a new cafe.

6. Conclusion

The purpose of this project was to determine ideal neighborhoods to open a new cafe in San Diego, based on neighborhoods and foursquare venue data. In conclusion, the recommendation to stakeholders looking to open a new cafe in San Diego would point them towards Bankers Hill and East Village in the downtown area. These areas only have a few coffee shops, they are centrally located in the downtown region, and both are close to areas that have high popular dining. If stakeholders are open to a less central location in the city, Mission Beach offers an intriguing alternative due to its touristy, beach reputation that does not show cafes or coffee shops, and mostly contains breakfast spots. Customers may prefer a cafe in the area for an early to late afternoon meal. Further analysis is needed into these specific neighborhoods to reveal further characteristics for stakeholders to make the final decision.