

IBM Data Science Professional Specialization - Coursera

This project will try to solve a problem or question by applying data science methods on the location data gotten from FourSquare API.

San Diego research on ZIP codes

I. Description of the Project and Business Problem:

For this project we are assuming we are a company and we are trying to reach customers in San Diego. San Diego has a population estimated at 1.43 million. At the last census in 2010, San Diego's population was 1.307 million. San Diego has a population density of 4,003 people per square mile in the city proper. Starting out, it is difficult to know exactly who your target market is in San Diego, California. Even over time your ideal customers are subject to change. Therefore, it is important to evaluate your company and its products often to ensure your marketing campaigns are reaching the right people. Business Problem is to analyze and select the best location in the city for our marketing campaigns targeted advertising is a way of placing ads based on demographics, on the consumers' previous buying history or on behavior.

Many types of targeted advertising are used online, but advertisers use in other media as well.

Examples of targeted advertising include having consumers choose which ads to view and placing ads on social networking sites and even billboards that vary depending on who is viewing them.

II. Description of the data:

The main data used for this project will be from two sources:

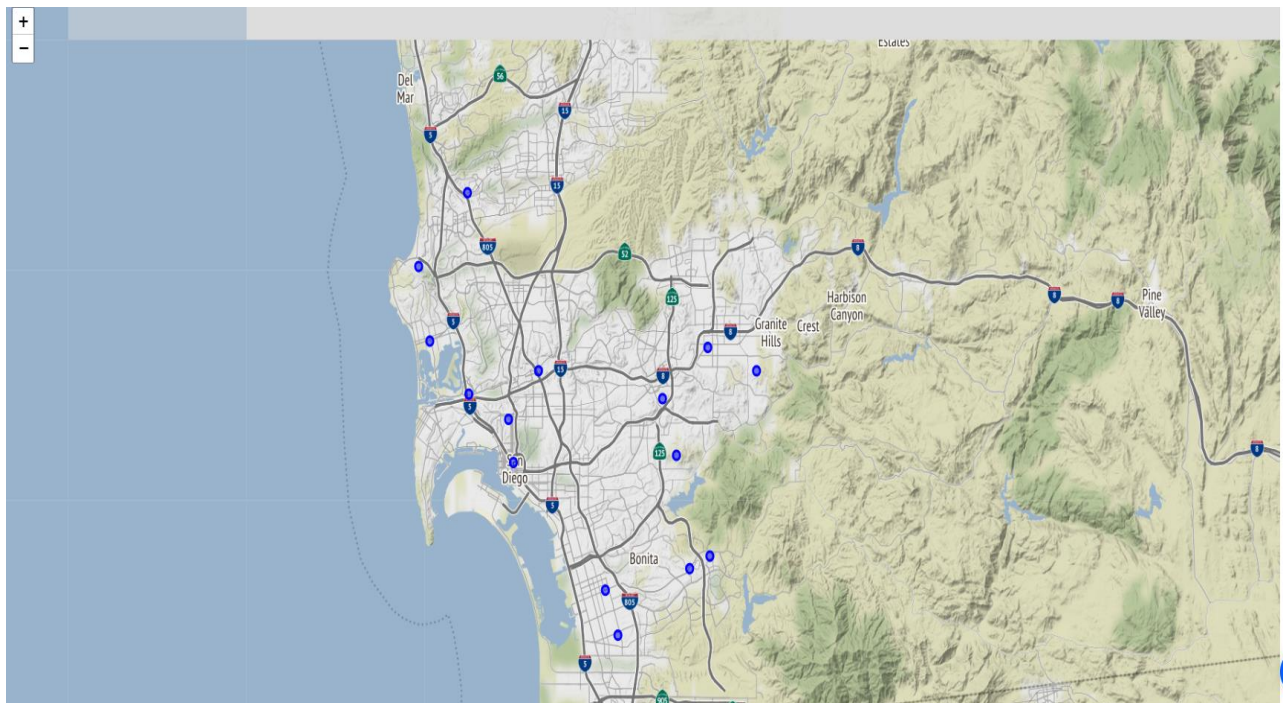
- The free of charge data we gathered from <https://www.geonames.org/postalcodes/US/CA/califonia.html>

(From there we pulled data only for San Diego's zip codes.

- The venues (FourSquare API) Byrequesting from the FourSquare API and using these geographic coordinates we obtained information for business venues near each ZIP code location.

III. Methodology

To begin, we plotted the coordinates on a folium map of San Diego. We can see that we have a good spread of point across the country.



In order to avoid overlapping, we decide to limit our search of businesses to 500 meter around each of these locations. We use the Foursquare API to create a dataframe of which includes, the business name, the type of business, and the ZIP Codes in which it resides. We create another dataframe from this which show the average amount of each business type per ZIP Code to put them into 8 groups (or kclusters). In other words, each group contains ZIP Codes with similar types of business.

	#1 Venue	#2 Venue	#3 Venue	#4 Venue	#5 Venue
0	Massage Studio	Business Service	Home Service	ATM	Music Venue
1	Business Service	Electronics Store	Gym	Home Service	Health & Beauty Service
2	Home Service	ATM	Optical Shop	Movie Theater	Music Venue
3	Bank	Ice Cream Shop	Nail Salon	Home Service	Park
4	Plaza	Gym	Pub	Park	Coffee Shop

Results :

We display the different clusters with a dataframe comprised of :

- Code : ZIP Code of the are
- Place: City where the ZIP code is found
- Latitude : North- South Coordinate
- Longitude : East Coordinate
- Cluster Labels : Group assignment based on nearby venues
- 1-6 Venue : Top 5 most common venues in the area
- We also include a color codes map displaying each location and the area it's associated business occupy.



Discussion/Conclusion :

Most of the ZIP Codes were placed into one group, cluster 5. This cluster consist mostly of various kinds of restaurants. Other notable parks include cluster 1 and 7, which consisted of mostly parks and trails/beaches, respectively. These two clusters are probably areas that can be avoided.

This data allowed us not only to see the types of business around each ZIP Code, but the larger areas in which those businesses are concentrated. Therefore we can determine the most beneficial places to advertise depending on our companies goals and in what way we should proceed in doing so.