Explore Santa Clara

A Traveller’s Guide to Best Businesses Areas in Town

1. Introduction

1.1 Background

A traveler coming to explore the town of Santa Clara would like to know which businesses are better with the purpose to either have a pleasant trip or buy a house and also know how much congested the city is. They would like to get a knowhow of the city.

With multiple tech companies like Google, Apple, LinkedIn in the vicinity, there is a huge demand

The purpose of this project is to help people in gathering an insight based on business ratings and traffic congestion in the city. It will help people making smart and efficient decision on deciding which places to focus on in the city.

1.2 Problem

Lots of people are migrating to Santa Clara as multiple tech companies like Google, Apple, LinkedIn are in the vicinity. They try to search for houses or come to explore the city or on business conventions. They try to get a knowhow of the city. This project is for those people who wish to know about the businesses in the city and also get information on the congestion in the city based on the traffic incidents in the city.

This project aim to create an analysis of features for people visiting Santa Clara to search the best businesses as a comparative analysis between zip codes. The features include traffic congestion data to give a better understanding of travel times around the city.

It will help people to get awareness of the area before moving to a new city, state, country or place for their work or to start a new fresh life.

1.3 Interest

A traveler who comes to the city of Santa Clara, will like to know the businesses in the area and how congested the city is. He will get an idea of the city based on the clusters.

2. Data Section

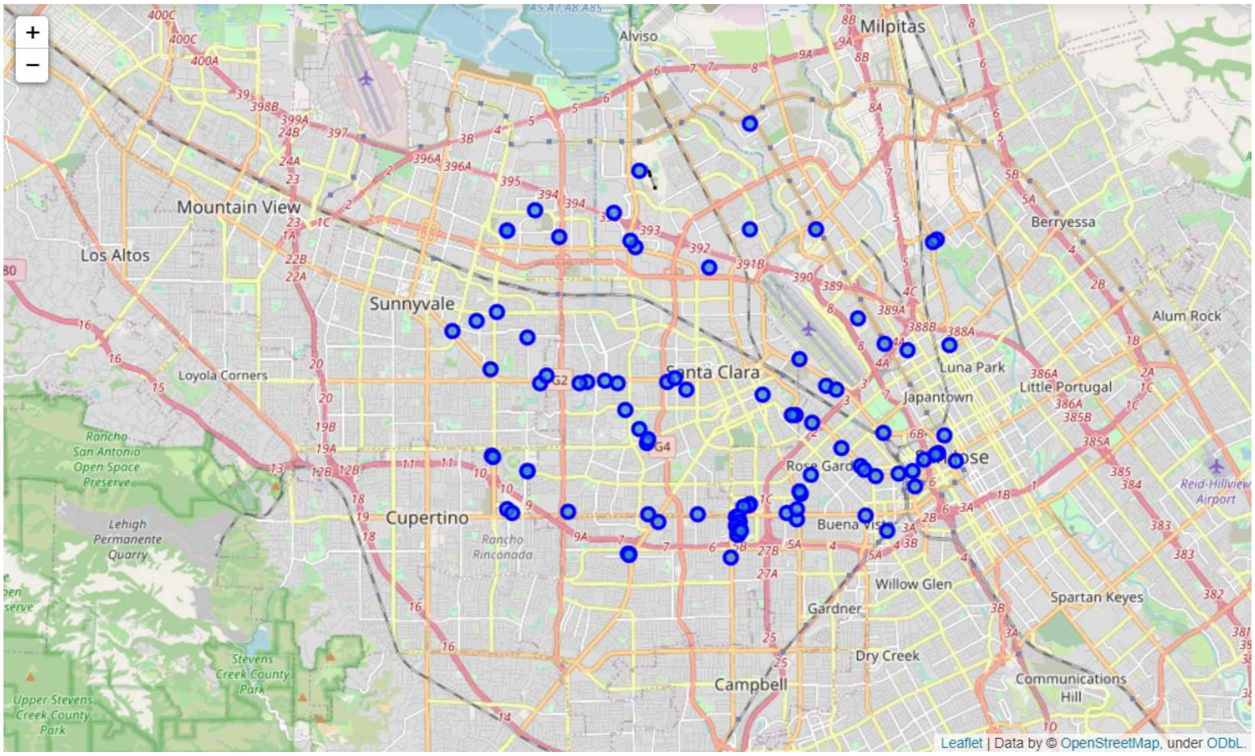
Data Link: https://api.foursquare.com/v2/venues/explore? NEAR_LOCATION = 'Santa Clara, CA'

Will use Santa Clara dataset which we scrapped from foursquare. Dataset consisting of ZipCode, Business Name, Street, Business Category, Latitude and Longitude.

Foursquare API Data:

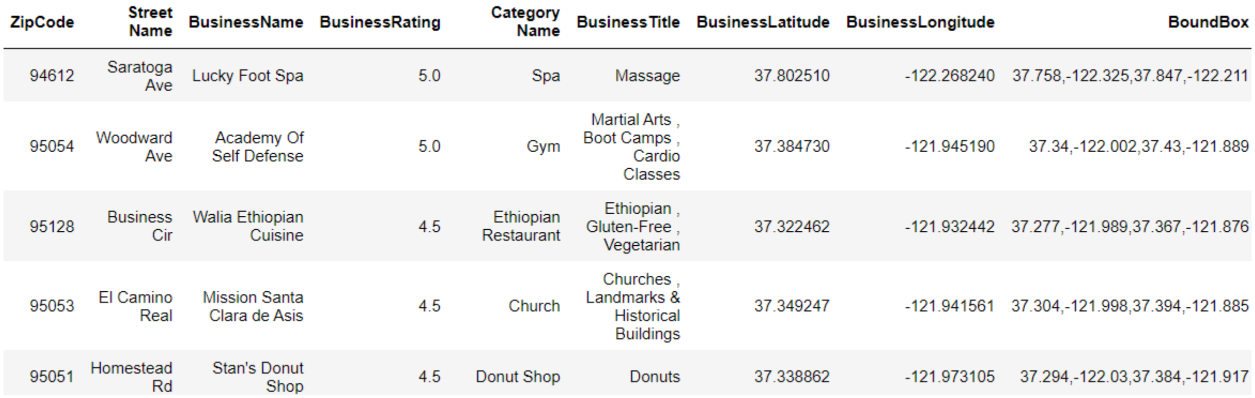
We will need data about different businesses with their categories in different neighborhoods of Santa Clara in a radius of 10km. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

We get the venues in a 10km radius.



Yelp Business Search API Data:

https://api.yelp.com/v3/businesses/search After finding the list of businesses, we then connect to the Yelp API to gather information about the business rating about the venues. We draw a bounding box of 5kms around the businesses based on their latitude and longitude.



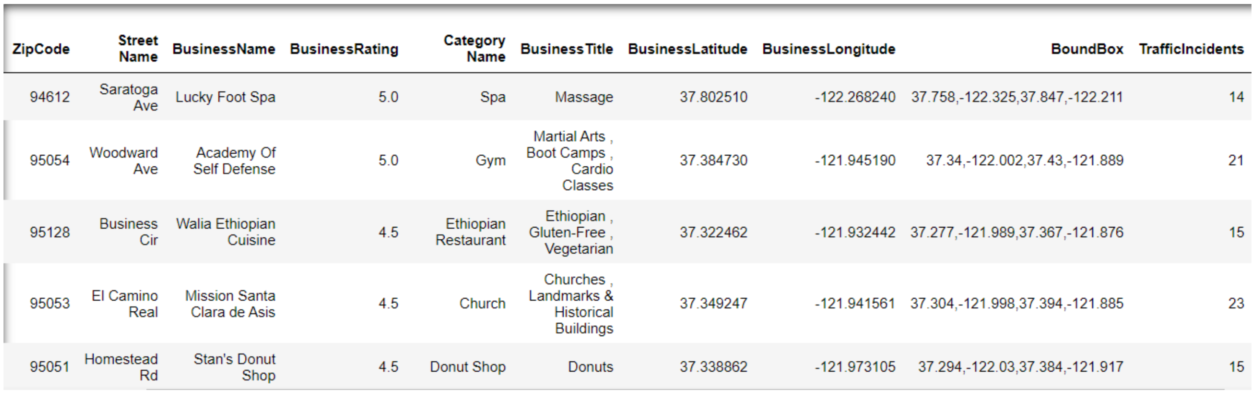
ZipCode	Street Name	BusinessName	BusinessRating	Category Name	BusinessTitle	BusinessLatitude	BusinessLongitude	BoundBox
94612	Saratoga Ave	Lucky Foot Spa	5.0	Spa	Massage	37.802510	-122.268240	37.758,-122.325,37.847,-122.211
95054	Woodward Ave	Academy Of Self Defense	5.0	Gym	Martial Arts , Boot Camps , Cardio Classes	37.384730	-121.945190	37.34,-122.002,37.43,-121.889
95128	Business Cir	Walia Ethiopian Cuisine	4.5	Ethiopian Restaurant	Ethiopian , Gluten-Free , Vegetarian	37.322462	-121.932442	37.277,-121.989,37.367,-121.876
95053	El Camino Real	Mission Santa Clara de Asis	4.5	Church	Churches , Landmarks & Historical Buildings	37.349247	-121.941561	37.304,-121.998,37.394,-121.885
95051	Homestead Rd	Stan's Donut Shop	4.5	Donut Shop	Donuts	37.338862	-121.973105	37.294,-122.03,37.384,-121.917

Bing API Data:

https://dev.virtualearth.net/REST/v1/Traffic/Incidents/

We then connect this boundingbox data to bing to get the traffic incident data. The data retrieved from a combination of FourSquare, Yelp and Bing contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

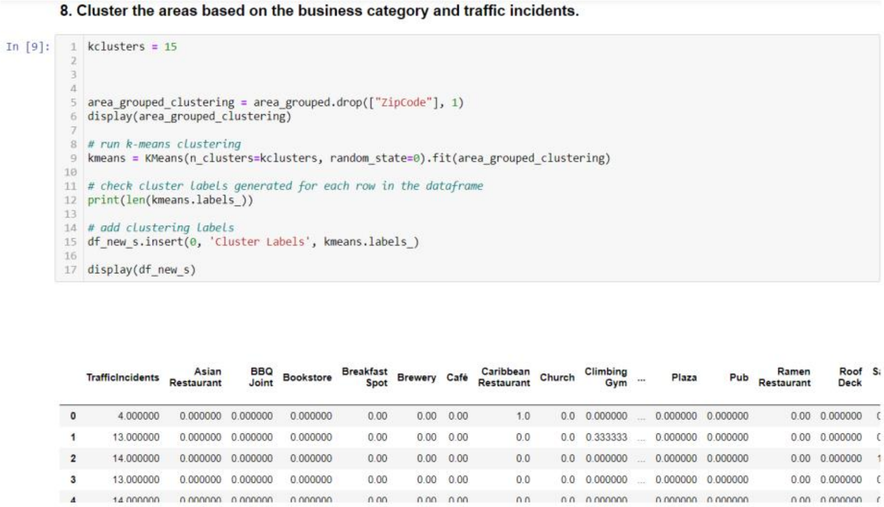
1. ZipCode
2. Street
3. Business Category
4. Business Name
5. Business Rating
6. Business Latitude and Longitude
7. BoundingBox Latitude and Longitude around the Business
8. Traffic Incident Counts in the BoundingBox



ZipCode	Street Name	BusinessName	BusinessRating	Category Name	BusinessTitle	BusinessLatitude	BusinessLongitude	BoundBox	TrafficIncidents
94612	Saratoga Ave	Lucky Foot Spa	5.0	Spa	Massage	37.802510	-122.268240	37.758,-122.325,37.847,-122.211	14
95054	Woodward Ave	Academy Of Self Defense	5.0	Gym	Martial Arts , Boot Camps , Cardio Classes	37.384730	-121.945190	37.34,-122.002,37.43,-121.889	21
95128	Business Cir	Walia Ethiopian Cuisine	4.5	Ethiopian Restaurant	Ethiopian , Gluten-Free , Vegetarian	37.322462	-121.932442	37.277,-121.989,37.367,-121.876	15
95053	El Camino Real	Mission Santa Clara de Asis	4.5	Church	Churches , Landmarks & Historical Buildings	37.349247	-121.941561	37.304,-121.998,37.394,-121.885	23
95051	Homestead Rd	Stan's Donut Shop	4.5	Donut Shop	Donuts	37.338862	-121.973105	37.294,-122.03,37.384,-121.917	15

5. Clustering Approach:

To compare the similarities of data, we decided to explore businesses, categorize them, get traffic incident data near them and group them into clusters to find similar clusters in Santa Clara. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.



8. Cluster the areas based on the business category and traffic incidents.

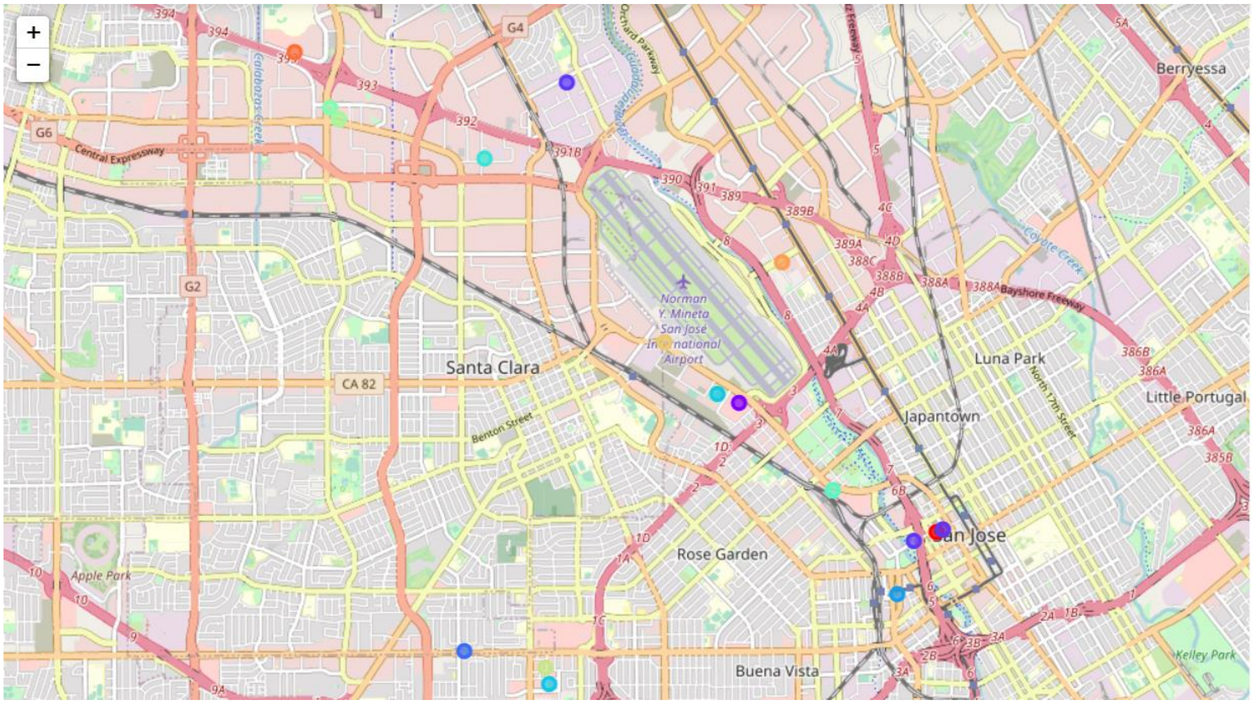
In [9]:

```
kclusters = 15



area_grouped_clustering = area_grouped.drop(["ZipCode"], 1)
display(area_grouped_clustering)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(area_grouped_clustering)

# check cluster labels generated for each row in the dataframe
print(len(kmeans.labels_))

# add clustering labels
df_new_s.insert(0, 'Cluster Labels', kmeans.labels_)

display(df_new_s)
```

	TrafficIncidents	Asian Restaurant	BBQ Joint	Bookstore	Breakfast Spot	Brewery	Café	Caribbean Restaurant	Church	Climbing Gym	...	Plaza	Pub	Ramen Restaurant	Roof Deck	S:
0	4.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	1.0	0.0	0.000000	...	0.000000	0.000000	0.00	0.000000	(
1	13.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.0	0.0	0.333333	...	0.000000	0.000000	0.00	0.000000	(
2	14.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.0	0.0	0.000000	...	0.000000	0.000000	0.00	0.000000	1
3	13.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.0	0.0	0.000000	...	0.000000	0.000000	0.00	0.000000	(
4	14.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.0	0.0	0.000000		0.000000	0.000000	0.00	0.000000	(

6. Visualizing the Clusters:

To see the clusters on the chart and then categorize into areas with businesses with higher ratings and higher congestion. To compare to other clusters where the congestion is lower.



5. Discussion Section

Problem Which Tried to Solve:

The major purpose of this project, is to suggest to people who are visiting Santa Clara, the best businesses in town and also which areas are congested.

1. Sorted list of businesses in terms of business ratings in descending order.
2. Sorted list of areas in terms of traffic congestion.

6. Conclusion Section

In this project, using k-means cluster algorithm I separated the city into 10(Ten) different clusters and for 100 different latitude and longitude from dataset, which have very-similar businesses around them. Using the charts above results presented to a particular zipcode based on higher business ratings of businesses in them and number of traffic incidents in them.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation. This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

Future Works:

This project can be continued for making it more precise in terms to find best places to stay in terms of city attractions. Best means on the basis of all required things(daily needs or things we need to live a better life) around and also in terms of cost effective.

Libraries Used to Develop the Project:

Pandas: For creating and manipulating dataframes.

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.