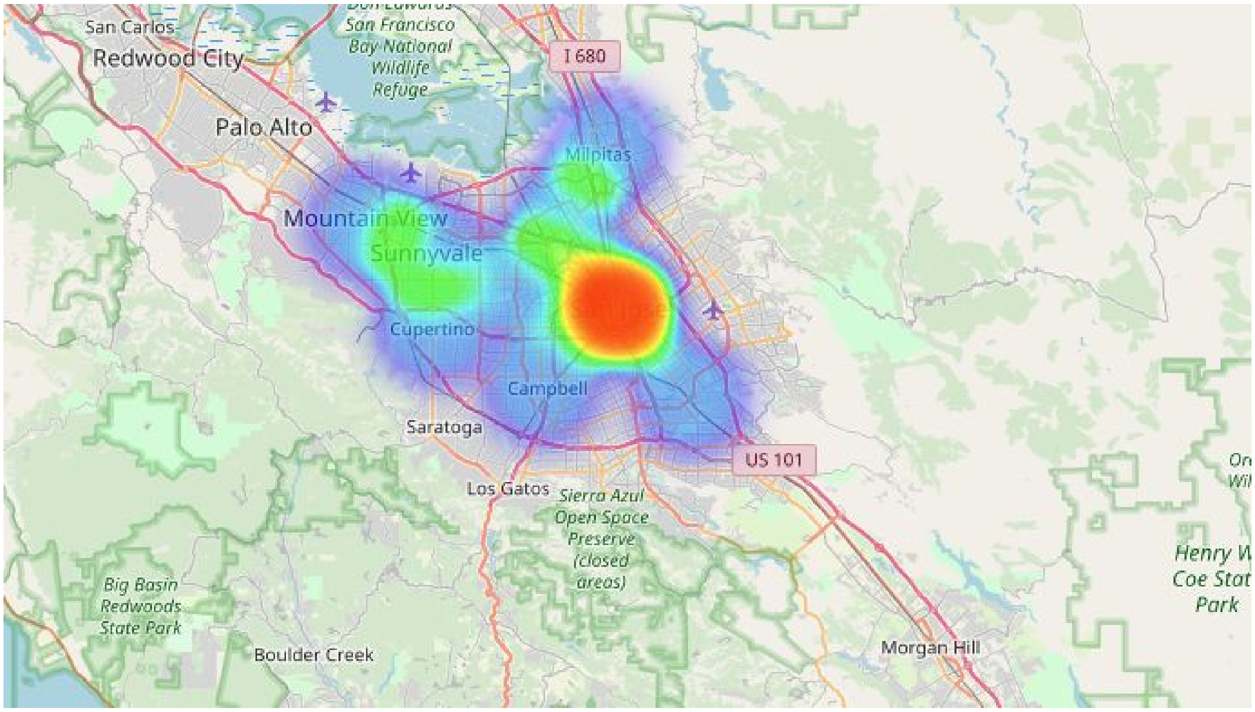
Coursera Capstone Project- IBM Data Science Specialization

Santa Clara County Rental Data Analysis



Bibhuti Nakarmi

1. INTRODUCTION

I am an aspiring data scientist, recently moved to San Jose, California. Looking for affordable good city to rent near Santa Clara County, California. The apartment in Santa Clara County must meet the following demands: apartment must be 2 or 3 bedrooms

- City should be safe with low number of violent crime
- price of rent not exceed $7,000 per month
- top amenities in the selected neighborhood shall be similar to current residence
- desirable to have venues such as coffee shops,restaurants, gym and food shops

Business Problem

The challenge is to find a suitable apartment for rent in San Jose California that complies with the demands on location, price and venues. The data required to resolve this challenge is described in the following section 2, below.

Interested Audience

I believe this is a relevant challenge with valid questions for anyone moving to other large city. The methodology is also applicable for anyone interested in exploring starting or locating a new business in any city. Lastly, it can also serve as a good practical exercise to develop Data Science skills.

2. DATA SECTION

The following data is required to answer the issues of the problem:

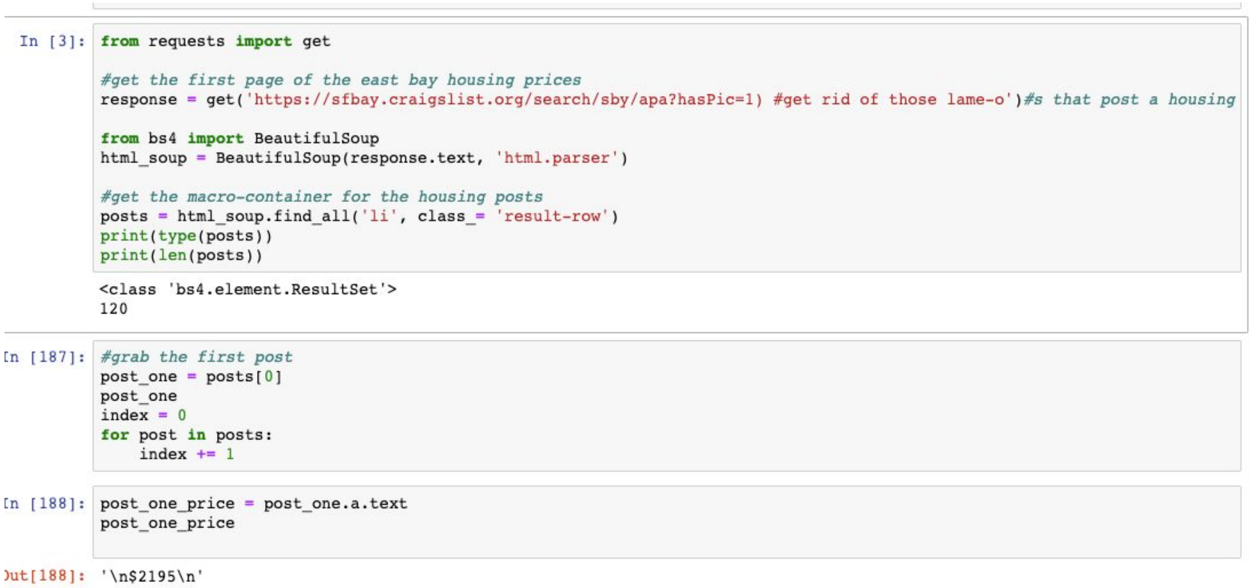
- List of Boroughs and neighborhoods of San Jose with their geodata (latitude and longitude)
- List of rail and cal stations in San Jose with their address location
- List of apartments for rent in San Jose area with their addresses and price
- Preferably, a list of apartment for rent with additional information, such as price, address, area, # of beds, etc
- Venues for each San Jose neighborhood (than can be clustered)
- Venues for rail stations, as needed
- list of crime data of each city

Data Collection

The data of this project is gathered from different sources and combined together for final exploration and analysis: Web Scraping of apartment rent data from Craigslist, Crime data of Santa Clara county, FourSquare API for gathering venues.

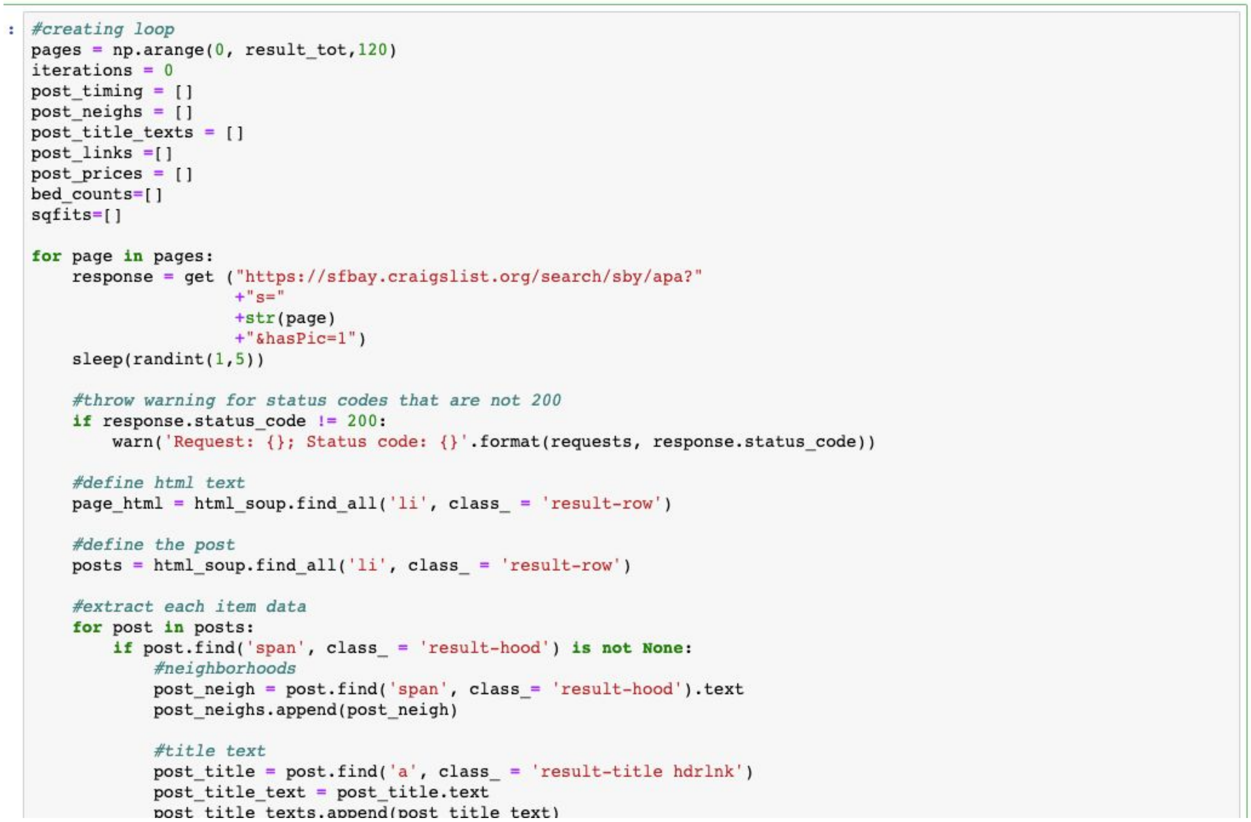
Web Scraping

I first make use of Craigslist South Bay page to scrap the rental data. I used used Beautifulsoup4 library, which is the module that can actually parse the HTML of the web page retrieved from the server.I then checked the type and length of that item to make sure it matches the number of posts on the page . You can find my import statements and setup code below:



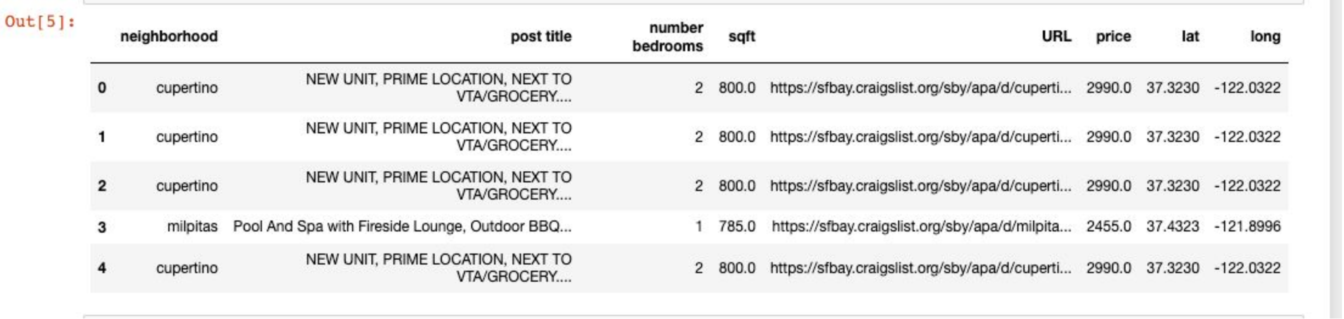
```
In [3]: from requests import get

        #get the first page of the east bay housing prices
        response = get('https://sfbay.craigslist.org/search/sby/apa?hasPic=1) #get rid of those lame-o')#s that post a housing

        from bs4 import BeautifulSoup
        html_soup = BeautifulSoup(response.text, 'html.parser')

        #get the macro-container for the housing posts
        posts = html_soup.find_all('li', class_= 'result-row')
        print(type(posts))
        print(len(posts))

        <class 'bs4.element.ResultSet'>
        120

[n [187]: #grab the first post
        post_one = posts[0]
        post_one
        index = 0
        for post in posts:
            index += 1

[n [188]: post_one_price = post_one.a.text
        post_one_price

)ut[188]: '\n$2195\n'
```

The next is to get data for all South Bay.



```
: #creating loop
  pages = np.arange(0, result_tot,120)
  iterations = 0
  post_timing = []
  post_neighs = []
  post_title_texts = []
  post_links =[]
  post_prices = []
  bed_counts=[]
  sqfits=[]

  for page in pages:
      response = get ("https://sfbay.craigslist.org/search/sby/apa?"
                      +"s="
                      +str(page)
                      +"&hasPic=1")
      sleep(randint(1,5))

      #throw warning for status codes that are not 200
      if response.status_code != 200:
          warn('Request: {}; Status code: {}'.format(requests, response.status_code))

      #define html text
      page_html = html_soup.find_all('li', class_ = 'result-row')

      #define the post
      posts = html_soup.find_all('li', class_ = 'result-row')

      #extract each item data
      for post in posts:
          if post.find('span', class_ = 'result-hood') is not None:
              #neighborhoods
              post_neigh = post.find('span', class_= 'result-hood').text
              post_neighs.append(post_neigh)

              #title text
              post_title = post.find('a', class_ = 'result-title hdrlnk')
              post_title_text = post_title.text
              post_title_texts.append(post_title_text)
```

After web scraping and cleaning the data, dataframe looks like as follows:

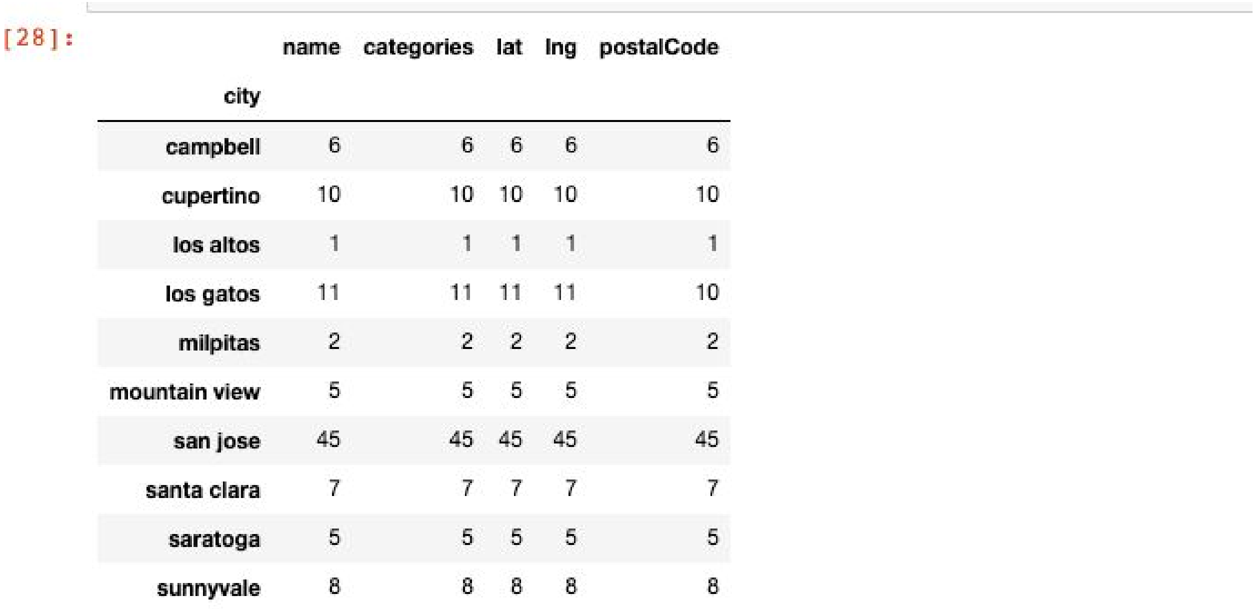


Out[5]:

	neighborhood	post title	number bedrooms	sqft	URL	price	lat	long
0	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990.0	37.3230	-122.0322
1	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990.0	37.3230	-122.0322
2	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990.0	37.3230	-122.0322
3	milpitas	Pool And Spa with Fireside Lounge, Outdoor BBQ...	1	785.0	https://sfbay.craigslist.org/sby/apa/d/milpita...	2455.0	37.4323	-121.8996
4	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990.0	37.3230	-122.0322

Using Foursqaure API to get top venue of cities in Santa Clara County:

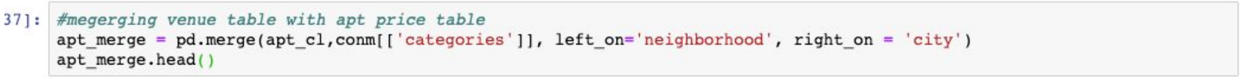
Foursquare data is very comprehensive and it powers location data for Apple, Uber etc. The popular spots returned depends on the highest foot traffic and thus it depends on the time when the call is made. So we may get different popular venues depending upon different time of the day. The call returns a JSON file and we need to turn that into a data-frame. Here I've chosen 1500 popular spots for each major districts within a radius of 14 mile. Below is the data-frame obtained from the JSON file that was returned by Foursquare —



[28]:

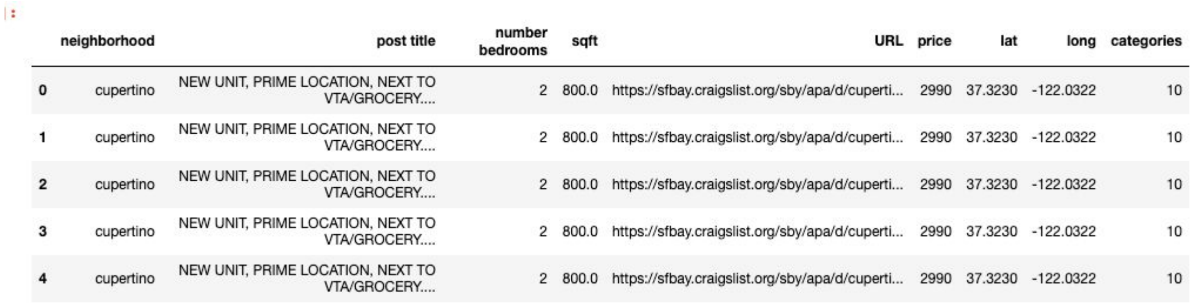
	name	categories	lat	lng	postalCode
city					
campbell	6	6	6	6	6
cupertino	10	10	10	10	10
los altos	1	1	1	1	1
los gatos	11	11	11	11	10
milpitas	2	2	2	2	2
mountain view	5	5	5	5	5
san jose	45	45	45	45	45
santa clara	7	7	7	7	7
saratoga	5	5	5	5	5
sunnyvale	8	8	8	8	8

Merging foursquare data with Craiglist dataframe.



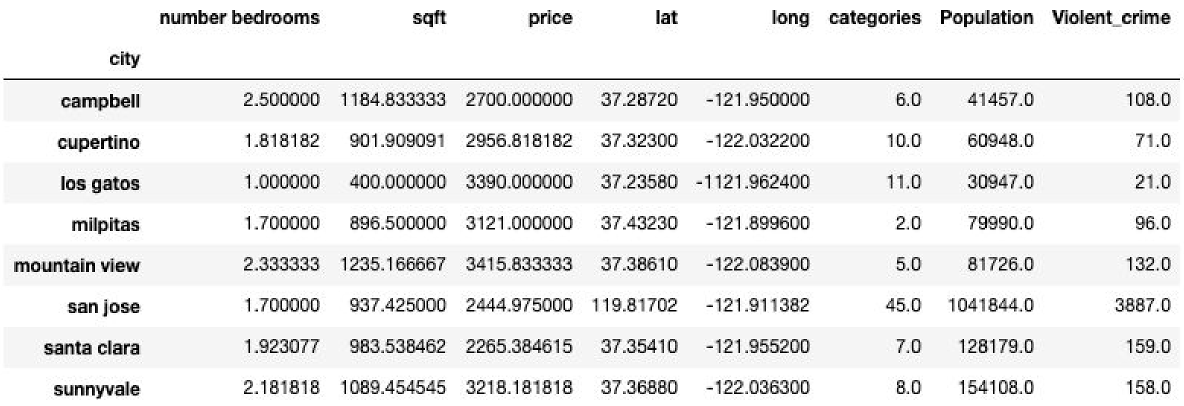
37]:

```
#megerging venue table with apt price table
apt_merge = pd.merge(apt_cl,conm[['categories']], left_on='neighborhood', right_on = 'city')
apt_merge.head()
```



:									
	neighborhood	post title	number bedrooms	sqft	URL	price	lat	long	categories
0	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990	37.3230	-122.0322	10
1	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990	37.3230	-122.0322	10
2	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990	37.3230	-122.0322	10
3	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990	37.3230	-122.0322	10
4	cupertino	NEW UNIT, PRIME LOCATION, NEXT TO VTA/GROCERY....	2	800.0	https://sfbay.craigslist.org/sby/apa/d/cuperti...	2990	37.3230	-122.0322	10

The next step in this project is to get crime data of Santa Clara County , join with df_merged and clean any data which is missing/different.The detail of steps can be seen in the code section below.



	number bedrooms	sqft	price	lat	long	categories	Population	Violent_crime
city								
campbell	2.500000	1184.833333	2700.000000	37.28720	-121.950000	6.0	41457.0	108.0
cupertino	1.818182	901.909091	2956.818182	37.32300	-122.032200	10.0	60948.0	71.0
los gatos	1.000000	400.000000	3390.000000	37.23580	-1121.962400	11.0	30947.0	21.0
milpitas	1.700000	896.500000	3121.000000	37.43230	-121.899600	2.0	79990.0	96.0
mountain view	2.333333	1235.166667	3415.833333	37.38610	-122.083900	5.0	81726.0	132.0
san jose	1.700000	937.425000	2444.975000	119.81702	-121.911382	45.0	1041844.0	3887.0
santa clara	1.923077	983.538462	2265.384615	37.35410	-121.955200	7.0	128179.0	159.0
sunnyvale	2.181818	1089.454545	3218.181818	37.36880	-122.036300	8.0	154108.0	158.0

3. METHODOLOGY SECTION

This section describes in detail the the steps performed to extract meaningful information from the data sources mentioned above.

The analysis and stragegy

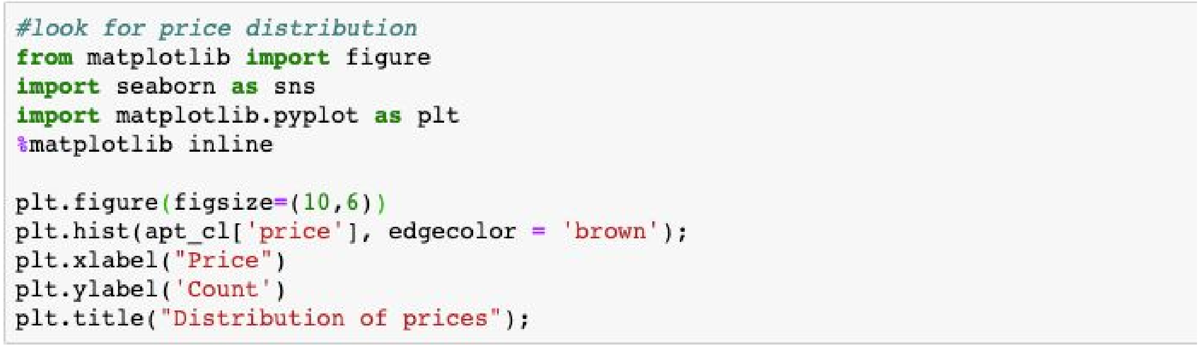
The strategy is based on mapping data obtained to facilitate choice of at least one candidate places for rent. The choices will be base on : how safe the place is (amount of violent crime, prices range , number of activities available).

The processing of these DATA and its mapping will allow to answer the key questions to make a decision:

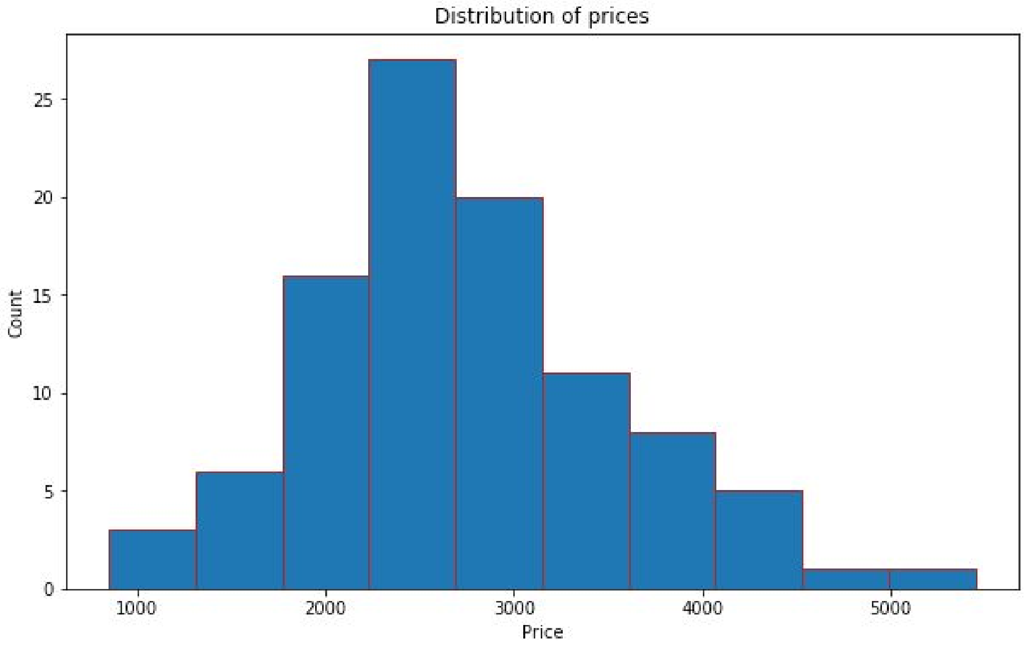
- Which is the safe city for rent?
- Which city has top number of activirities or more thins to do?
- Which city has higher distribution of range prices?

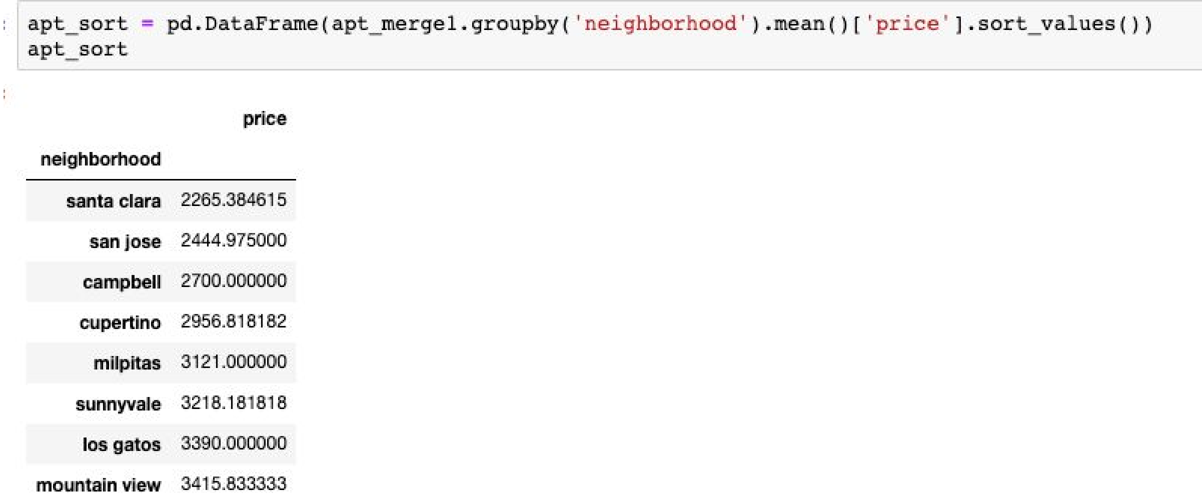
4. Exploratory Data Analysis

I wanted to see the distribution of pricing for cities in Santa Clara County.



```
#look for price distribution
from matplotlib import figure
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

plt.figure(figsize=(10,6))
plt.hist(apt_cl['price'], edgecolor = 'brown');
plt.xlabel("Price")
plt.ylabel('Count')
plt.title("Distribution of prices");
```

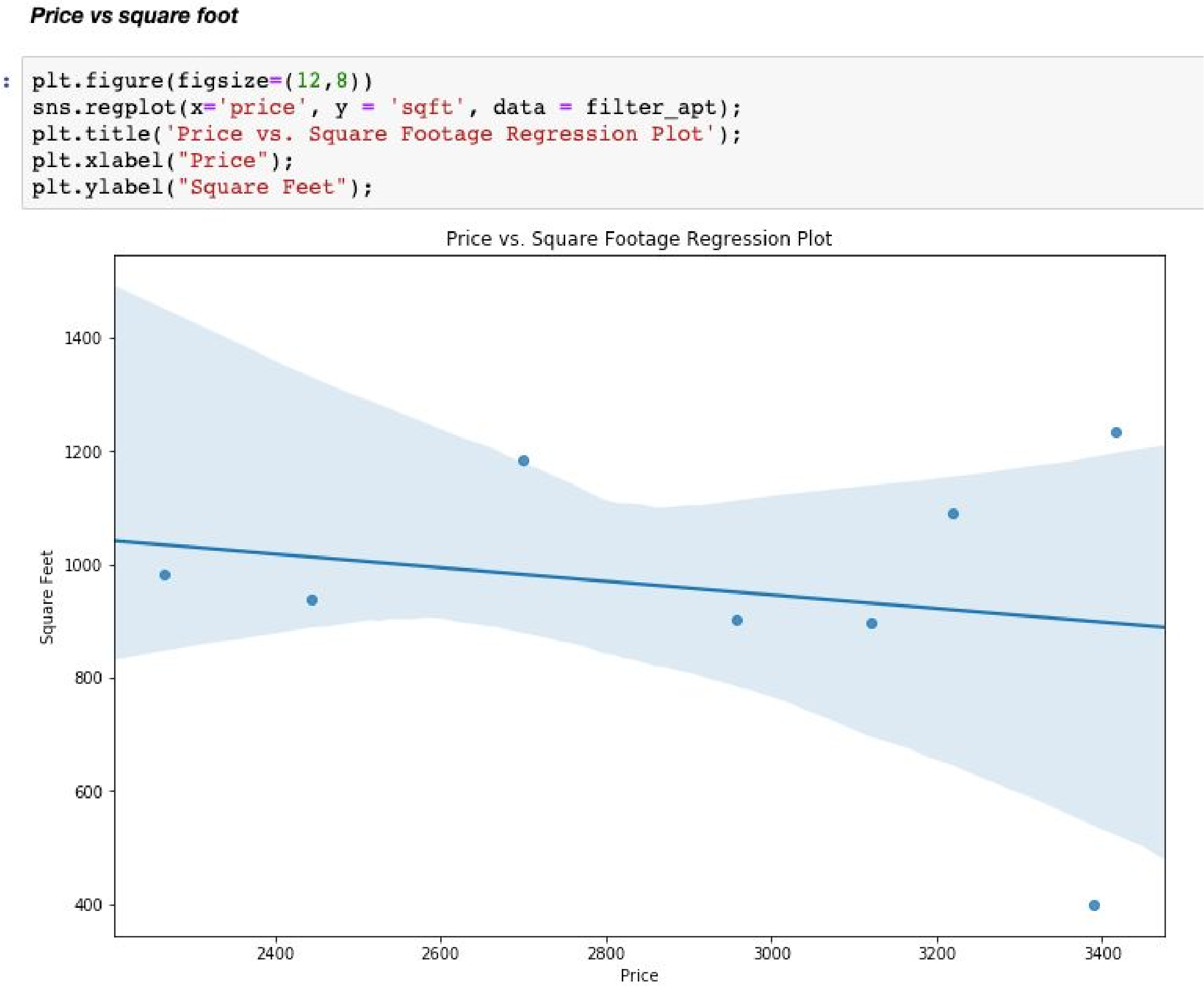




```
apt_sort = pd.DataFrame(apt_merge1.groupby('neighborhood').mean()['price'].sort_values())
apt_sort
```

	price
neighborhood	
santa clara	2265.384615
san jose	2444.975000
campbell	2700.000000
cupertino	2956.818182
milpitas	3121.000000
sunnyvale	3218.181818
los gatos	3390.000000
mountain view	3415.833333

Next is analyzing the relation between price and square footage of apartment.We can call a regplot() on these two variables to get a regression line with a bootstrap confidence interval calculated about the line and shown as a shaded region with the code below.



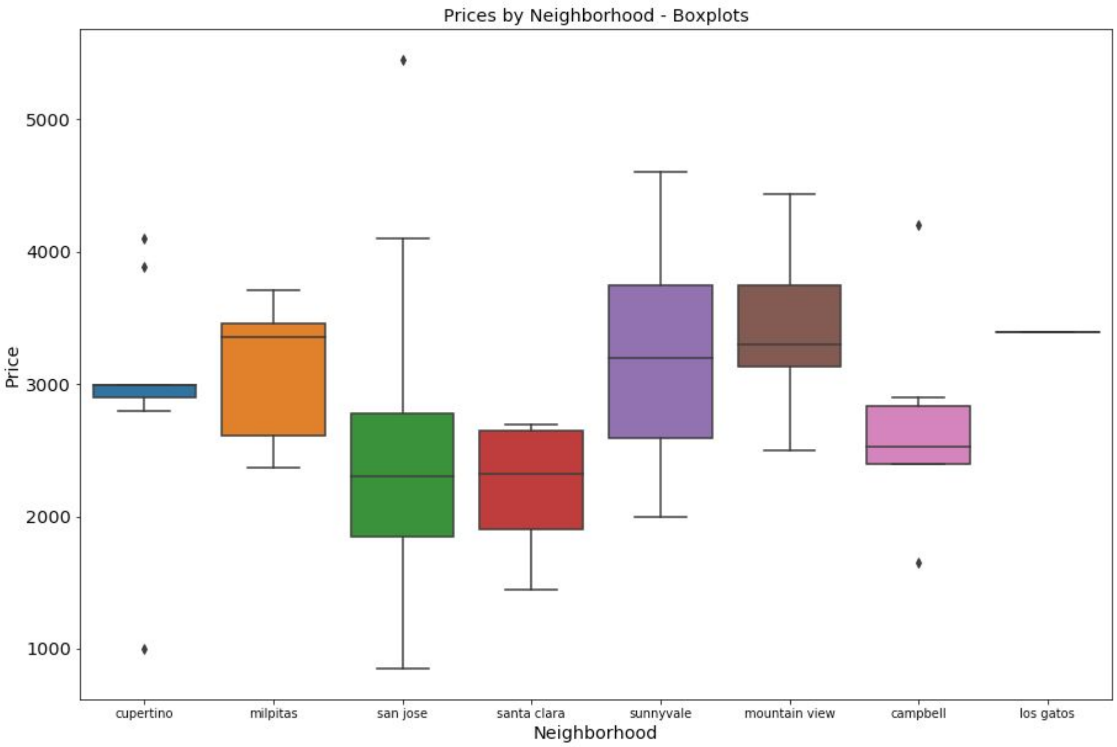
Price vs square foot

```
plt.figure(figsize=(12,8))
sns.regplot(x='price', y = 'sqft', data = filter_apt);
plt.title('Price vs. Square Footage Regression Plot');
plt.xlabel("Price");
plt.ylabel("Square Feet");
```

Figure shows square footage is linearly related to price. Lower the square feet lower the price.

Prices By Cities:

To understand the data , I looked at the spread of each neighborhood in terms of price. By doing this, I saw how prices in neighborhoods can vary, and to what degree. Here's the code that produces the plot that follows.

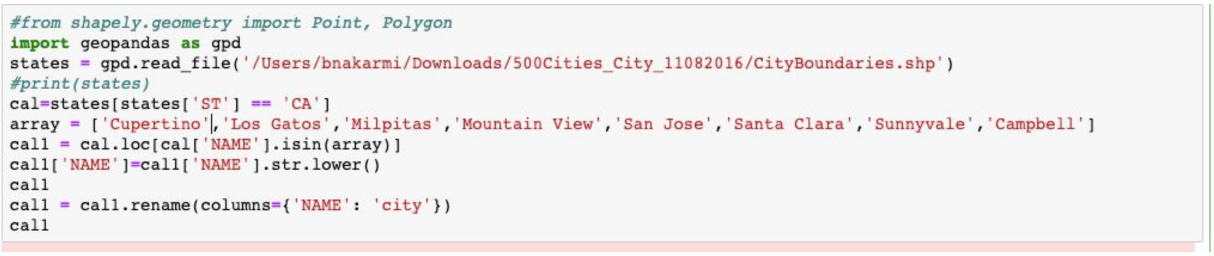


San Jose seems to have affordable price range. San Jose city has wider variety of price range distribution.

Mapping and visualizing Data

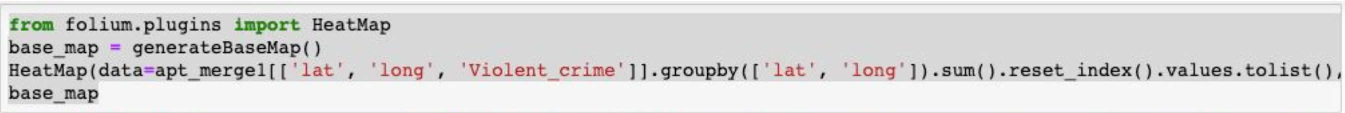
I am using Folium, the Python interface to the Leaflet JavaScript mapping library. I will explore some of the features of Folium by analyzing data I created above.

Importing shapefile downloaded from Santa Clara County website, and converting into dataframe to merge it with main dataframe.

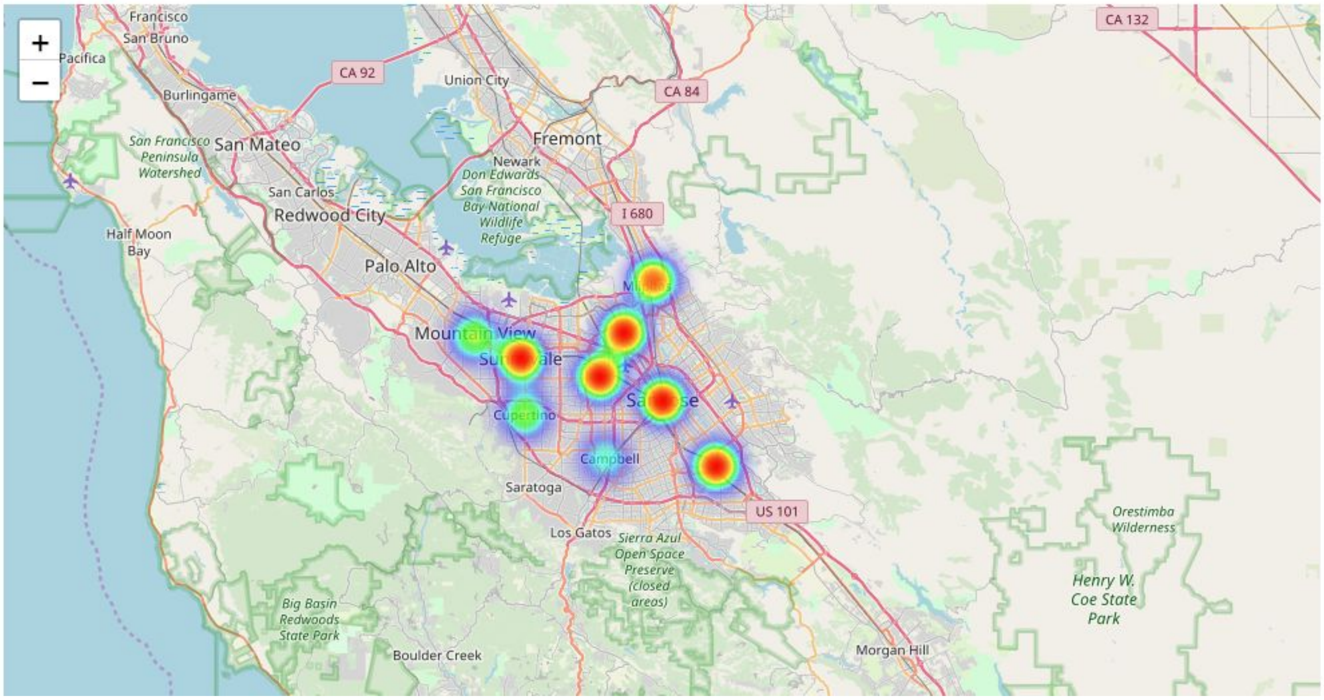


```
#from shapely.geometry import Point, Polygon
import geopandas as gpd
states = gpd.read_file('/Users/bnakarmi/Downloads/500Cities_City_11082016/CityBoundaries.shp')
#print(states)
cal=states[states['ST'] == 'CA']
array = ['Cupertino','Los Gatos','Milpitas','Mountain View','San Jose','Santa Clara','Sunnyvale','Campbell']
cal1 = cal.loc[cal['NAME'].isin(array)]
cal1['NAME']=cal1['NAME'].str.lower()
cal1
cal1 = cal1.rename(columns={'NAME': 'city'})
cal1
```

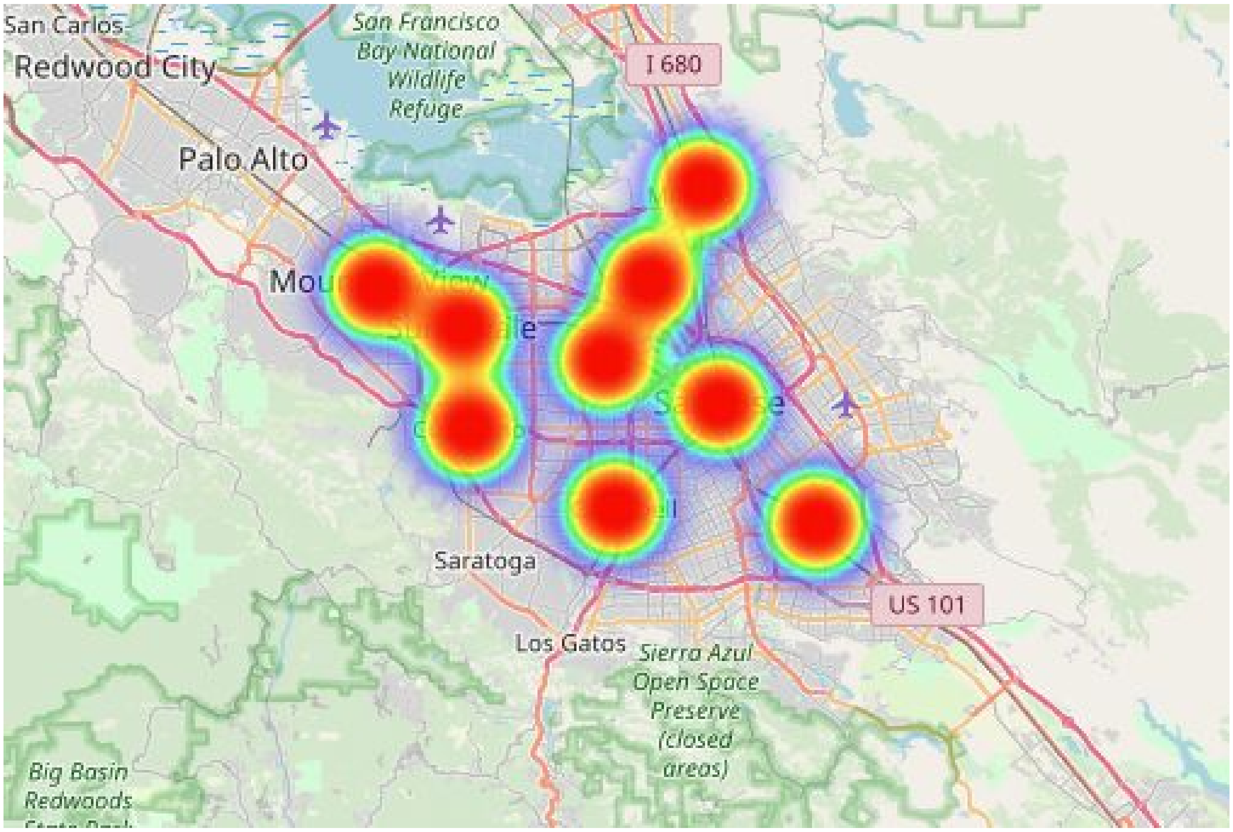
Generating Violent crime heat map



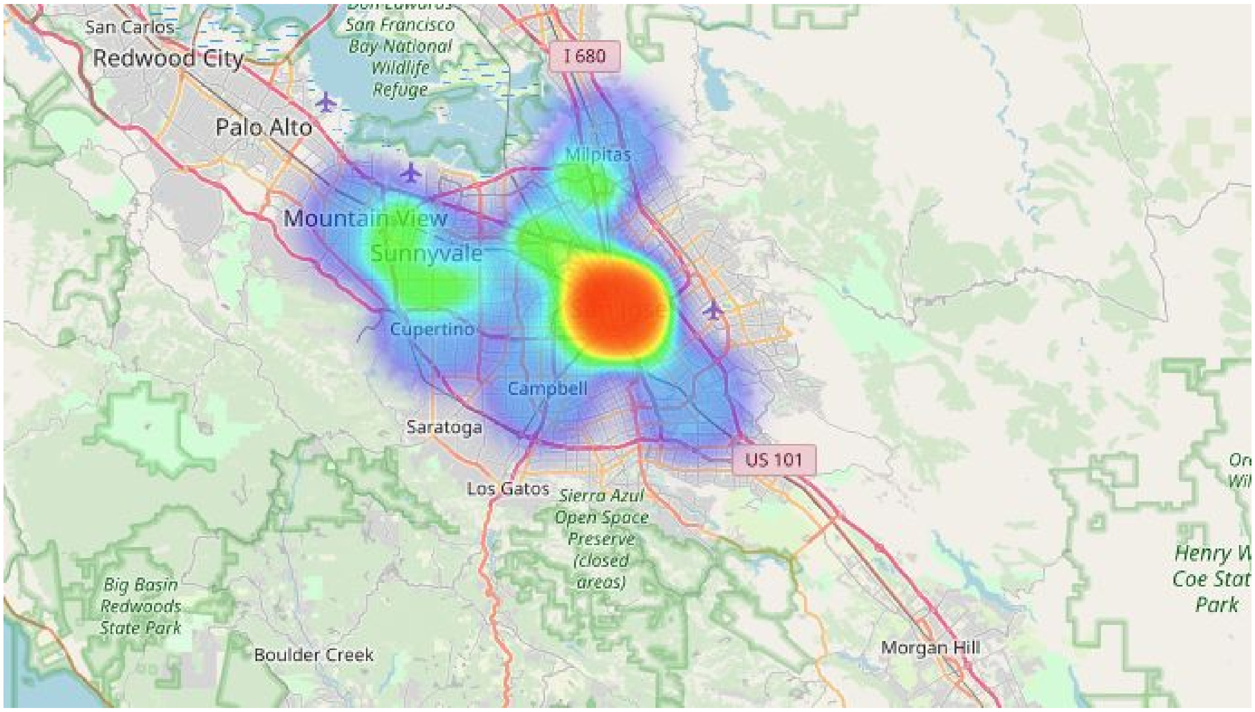
```
from folium.plugins import HeatMap
base_map = generateBaseMap()
HeatMap(data=apt_merge1[['lat', 'long', 'Violent_crime']].groupby(['lat', 'long']).sum().reset_index().values.tolist(),
base_map
```



High Price Cluster


San Carlos
Redwood City
San Francisco Bay National Wildlife Refuge
I 680
Palo Alto
Saratoga
US 101
Los Gatos
Sierra Azul Open Space Preserve (closed areas)
Big Basin Redwoods

Top Venue heat mapping



RESULTS

Based on the data analysis we can summarize that:Santa Clara County Craiglist data has following findings:

- Square footage is linearly related to price. Lower the square feet lower the price.
- San Jose seems to have affordable price range. San Jose city has wider variety of price range distribution.
- Price Range 2500 to 4000 are listed more than any other price range.
- San Jose city has highest amount of violent crime. Campbell, Mountain View and Cupertino has lowest crime occurrence.

After examining the data and maps produced above, I have chosen one location that meet the requirement: San Jose downtown has a lot of top venues but it also has a high number of violent crimes. I decided not to consider this location due to the amount of crime. Milpitas has less number of crimes and it has the second highest number of top venues. I chose Milpitas as top choice for rent.

Discussion:

I am positively impressed with the overall organization, content, and lab work. After completion of Capstone Project , I feel rewarded . I have created a good project that I can present as an example to show my potential.

CONCLUSION

Finally to conclude this project, We have got a small glimpse of how real life data-science projects look like. I've made use of some frequently used python libraries to scrape web-data, use Foursquare API to explore the Santa Clara County. There is some limitation on project data. Craigslist is not a great data source, latitude and longitude or some times address provided are incorrect. Number of bedrooms listed, price amount are misleading to attract visitor. I did analysis on City scale because of insufficient address information on Craiglist website. But the knowledge gained here will be a good starting point to perform such an analysis further.

I future I want to do more analysis on good school district and house rental price, nearness to school and price. I intend to perform some analysis related to that based on knowledge gained in this project in order to predict if such a measure will be beneficial to the community.