Capstone Project - The Battle of the Neighbourhoods (Week 2)

Case: Exploration and analysis of San Jose and its neighbourhoods

1. Description of the problem

Identifying the possibility to establish a financial advisory institution in San Jose by analysing the economic potential of the population and exploring the venues in each neighbourhood.

1.1 Background

San Jose is one of the famous cities of California, USA. It is part of the Silicon Valley, which hosts many information technology companies. These companies employ many IT professionals who are quite busy with the technological innovation and development of new software products. These professionals are fully occupied with their daily hectic routine which consumes a significant portion of their time. Consequently, it leaves them with less time to fully concentrate on their financial matters, especially studying different opportunities and making suitable investment-related decisions. Therefore, an institution which could manage their finances and provide investment advice would be of great support to them. From this perspective, San Jose offers a huge potential for establishing a financial institution to manage the finances and advise on investments for the people working in this city.

1.2 Analysis of economic indicators and venues of San Jose

Before such an institution can be set up, it is important to study the different parameters across different locations in San Jose which could indicate the investment potential. Accordingly, I have taken up the analysis of socio-economic indicators of people residing in different neighbourhoods of San Jose and the venues in each neighbourhood of San Jose. The analysis will be very helpful for institutions to potentially set up a shop and provide customized financial investment services for individuals working in San Jose along with their families.

2. Data description.

2.1 Neighbourhood data including economic indicators from a website:

The data is retrieved from the Spatial Data Repository of NYU [1] (U.S. Neighbourhoods greenness measures and social variables). The base data includes many attributes including economic indicators like average high income, percentage of people owing a house etc. The sample of the base data retrieved from Spatial Data Repository of NYU is provided below:


```
{'state': 'CA',
 'city': 'Long Beach',
 'name': 'Airport Area',
 'regionid': 272732,
 'shape_leng': 17308.1847929,
 'shape_area': 8359173.2354,
 'x': -118.154496304,
 'y': 33.8167,
 'region_id': 272732,
 'la_city': 0,
 'regionid_1': 272732,
 'dg_n': 0.132398,
 'dg_ninv': 0.867602,
 'dp_n': 0.161726,
 'dp_ninv': 0.838274,
 'pctpark_n': 0.077342,
 'meaneq_n': 0.0,
 'dg_mean': 321.37,
 'dp_mean': 116.38,
 'pct_park': 7.728,
 'mean_eq': 0.0,
 'youngfolks': 0.226277,
 'popdensity': 5741.323529,
 'diversity': 67.864706,
 'pc_income': 33412,
 'avg_hinc': 86903,
 'avg_hval': 437100,
 'pct_own': 0.607968,
 'pct_rent': 0.357973,
 'pct_white': 0.654763,
 'pct_hispan': 0.238231,
 'pct_black': 0.052221,
 'medage_cy': 39.370588,
 'unemprt_cy': 8.505882,
 'medhinc_cy': 73346,
```


Based on the objective which is to establish a financial institution, the following attributes have been identified as a basis for analysis:

- Neighbourhood name
- Latitude
- Longitude
- Population density
- Average high income
- Percentage owning houses
- Percentage renting
- Median age
- Median high income
- City parks

The .json file from the website contains data related to many other cities in the USA, in addition to San Jose. Therefore, base data has been first filtered to retrieve only the San Jose relevant indicators.

[1]: https://geo.nyu.edu/catalog/stanford-xq082nw3443

2.1.1 **Data samples for economic indicators**

2.1.1.1 Sample data for all the cities

	City	Neighborhood	Latitude	Longitude	popdensity	avg_hinc	pct_own	pct_rent	medage_cy	medhinc_cy	city_parks
0	Long Beach	Airport Area	33.8167	-118.154496	5741.323529	86903	0.607968	0.357973	39.370588	73346	0.129
1	Long Beach	Alamitos Heights	33.7738	-118.125871	7060.266667	110908	0.570474	0.380876	43.066667	83046	0.129
2	Long Beach	Belmont Heights	33.7639	-118.151191	15536.411111	84302	0.318873	0.623607	40.588889	66667	0.129
3	Long Beach	Belmont Shore	33.7589	-118.137396	13146.320000	104479	0.353005	0.584101	40.273333	83425	0.129
4	Long Beach	Bixby Area	33.8405	-118.176421	9901.688000	74772	0.514261	0.443579	36.672000	62332	0.129

2.1.1.2 San Jose data after clean-up

	City	Neighborhood	Latitude	Longitude	popdensity	avg_hinc	pct_own	pct_rent	medage_cy	medhinc_cy	city_parks
448	San Jose	Almaden Valley	37.2211	-121.869849	5612.600000	176918	0.734866	0.237213	41.210811	140164	0.068
449	San Jose	Alum Rock-East Foothills	37.3772	-121.825214	10173.289474	113307	0.674909	0.300745	35.323684	89187	0.068
450	San Jose	Berryessa	37.4016	-121.856383	9173.466667	130216	0.744864	0.236287	37.942424	108157	0.068
452	San Jose	Blossom Valley	37.2549	-121.843005	9977.140984	115350	0.657751	0.317510	36.403279	98611	0.068
453	San Jose	Buena Vista	37.3212	-121.916699	13386.700000	79466	0.275472	0.667481	32.340000	58664	0.068
454	San Jose	Burbank	37.3213	-121.930542	10040.090909	91458	0.370584	0.587205	35.945455	69339	0.068
455	San Jose	Cambrian Park	37.2598	-121.913999	7439.178261	132315	0.674165	0.302950	40.337681	103343	0.068
457	San Jose	Downtown	37.3405	-121.890340	12526.985714	86554	0.343653	0.603385	32.780000	63735	0.068
458	San Jose	East San Jose	37.3347	-121.825193	13127.814474	102843	0.676735	0.302644	32.343421	85678	0.068
459	San Jose	Edenvale-Seven Trees	37.2797	-121.817703	12941.859375	114389	0.674291	0.299457	32.309375	92459	0.068
460	San Jose	Rose Garden	37.3296	-121.931553	9117.513333	104442	0.447511	0.517272	38.103333	78196	0.068
462	San Jose	Evergreen	37.2981	-121.770502	8227.902326	140063	0.799158	0.178919	36.944186	116423	0.068
463	San Jose	Fairgrounds	37.3020	-121.858670	13640.700000	83621	0.480870	0.485012	31.651163	65232	0.068
464	San Jose	North San Jose	37.3831	-121.931127	5048.100000	107008	0.454169	0.474309	33.514286	84430	0.068
465	San Jose	North Valley	37.3763	-121.874664	12035.338000	101165	0.557130	0.406239	33.468000	84418	0.068
466	San Jose	Santa Teresa	37.2366	-121.793081	7900.364706	136960	0.728919	0.244759	37.592157	112249	0.068
467	San Jose	West San Jose	37.3010	-121.983818	9151.487097	129873	0.580150	0.392746	40.151613	105168	0.068
468	San Jose	Willow Glen	37.2963	-121.901595	8251.063889	117085	0.571972	0.395732	39.108333	90779	0.068

2.1.2 Data utilization:

The economic indicators like the ‘average high income’, ‘percentage of own’, ‘’percentage of rent’, ‘median high income of city’ etc. are analysed to identify their potential in enabling a potential investor to decide upon establishing a financial advisory institution.

2.2 **Data from neighbourhood and venues of San Jose – Explored using Foursquare**:

Using Foursquare, the San Jose neighbourhood is explored to identify the venues for each neighbourhood. From the perspective of establishing a financial advisory institution, the venues which generate economic activity do play a significant role. Hence, the different venues which are in the neighbourhood of San Jose are good economic indicators.

2.2.1 Data samples from Four Square:

2.2.1.1. Sample neighbourhood data of San Jose – obtained using Four Square


```
{'meta': {'code': 200, 'requestId': '5ca61b0c9fb6b714159b85af'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': '$-$$$$', 'key': 'price'},
    {'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Almaden Valley',
  'headerFullLocation': 'Almaden Valley, San Jose',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 50,
  'suggestedBounds': {'ne': {'lat': 37.239100018000016,
    'lng': -121.847287177019l},
   'sw': {'lat': 37.20309998199998, 'lng': -121.8924114949809}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4b79c338f964a52051102fe3',
       'name': 'Tacos Al Pastor',
       'location': {'address': '6469 Almaden Expy',
        'lat': 37.220333202570416,
        'lng': -121.86239385301369,
        'labeledLatLngs': [{'label': 'display',
          'lat': 37.220333202570416,
          'lng': -121.86239385301369}],
        'distance': 666,
        'postalCode': '95120',
        'cc': 'US',
        'city': 'San Jose',
        'state': 'CA',
        'country': 'United States',
        'formattedAddress': ['6469 Almaden Expy',
         'San Jose, CA 95120',
         'United States']],
```


2.2.1.2 Different neighbourhoods of San Jose

```
Almaden Valley
Alum Rock-East Foothills
Berryessa
Blossom Valley
Buena Vista
Burbank
Cambrian Park
Downtown
East San Jose
Edenvale-Seven Trees
Rose Garden
Evergreen
Fairgrounds
North San Jose
North Valley
Santa Teresa
West San Jose
Willow Glen
```

2.2.2 Sample venues in the neighbourhoods of San Jose

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Almaden Valley	37.2211	-121.869849	Almaden Community Center	37.221499	-121.869231	Gym / Fitness Center
1	Almaden Valley	37.2211	-121.869849	Parma Park	37.221672	-121.871146	Playground
2	Almaden Valley	37.2211	-121.869849	Jakes Playlot	37.221509	-121.870844	Playground
3	Almaden Valley	37.2211	-121.869849	Boulder Ridge Golf Club Grill	37.224674	-121.866432	Restaurant
4	Alum Rock-East Foothills	37.3772	-121.825214	Antipastos By De Rose	37.380269	-121.827502	Deli / Bodega

2.2.3 Data utilization: The data related to venues in different neighbourhoods of San Jose provides an indicator of economic activity within the neighbourhoods. Basically, the venues trigger a lot of economic activity.

2.3 Data cluster: Using the K-means clustering algorithm, the neighbourhoods are divided into clusters and a map was provided for visualization.

3. Methodology

From a technical perspective, I used Python data frames, Folium, matploplib and the relevant libraries for doing this analysis. The San Jose map was created using latitudes and longitudes, which were derived using geolocator.

The economic indicators of San Jose were derived from the base data upon filtering. Python data frame was used to store and manipulate the resultant data. Accordingly, I have created a data frame and populated the attributes which are economic indicators of San Jose and neighbourhoods as shown below.

	City	Neighborhood	Latitude	Longitude	popdensity	avg_hinc	pct_own	pct_rent	medage_cy	medhinc_cy	city_parks
448	San Jose	Almaden Valley	37.2211	-121.869849	5612.600000	176918	0.734866	0.237213	41.210811	140164	0.068
449	San Jose	Alum Rock-East Foothills	37.3772	-121.825214	10173.289474	113307	0.674909	0.300745	35.323684	89187	0.068
450	San Jose	Berryessa	37.4016	-121.856383	9173.466667	130216	0.744864	0.236287	37.942424	108157	0.068
452	San Jose	Blossom Valley	37.2549	-121.843005	9977.140984	115350	0.657751	0.317510	36.403279	98611	0.068
453	San Jose	Buena Vista	37.3212	-121.916699	13386.700000	79466	0.275472	0.667481	32.340000	58664	0.068
454	San Jose	Burbank	37.3213	-121.930542	10040.090909	91458	0.370584	0.587205	35.945455	69339	0.068
455	San Jose	Cambrian Park	37.2598	-121.913999	7439.178261	132315	0.674165	0.302950	40.337681	103343	0.068
457	San Jose	Downtown	37.3405	-121.890340	12526.985714	86554	0.343653	0.603385	32.780000	63735	0.068
458	San Jose	East San Jose	37.3347	-121.825193	13127.814474	102843	0.676735	0.302644	32.343421	85678	0.068
459	San Jose	Edenvale-Seven Trees	37.2797	-121.817703	12941.859375	114389	0.674291	0.299457	32.309375	92459	0.068
460	San Jose	Rose Garden	37.3296	-121.931553	9117.513333	104442	0.447511	0.517272	38.103333	78196	0.068
462	San Jose	Evergreen	37.2981	-121.770502	8227.902326	140063	0.799158	0.178919	36.944186	116423	0.068
463	San Jose	Fairgrounds	37.3020	-121.858670	13640.700000	83621	0.480870	0.485012	31.651163	65232	0.068
464	San Jose	North San Jose	37.3831	-121.931127	5048.100000	107008	0.454169	0.474309	33.514286	84430	0.068
465	San Jose	North Valley	37.3763	-121.874664	12035.338000	101165	0.557130	0.406239	33.468000	84418	0.068
466	San Jose	Santa Teresa	37.2366	-121.793081	7900.364706	136960	0.728919	0.244759	37.592157	112249	0.068
467	San Jose	West San Jose	37.3010	-121.983818	9151.487097	129873	0.580150	0.392746	40.151613	105168	0.068
468	San Jose	Willow Glen	37.2963	-121.901595	8251.063889	117085	0.571972	0.395732	39.108333	90779	0.068

Each of the economic indicators ‘average high income’, ‘the percentage of ownership’, ‘percentage of rent’, ‘median high income in the city’ and ‘median age‘ across different neighbourhoods were considered for analysis.

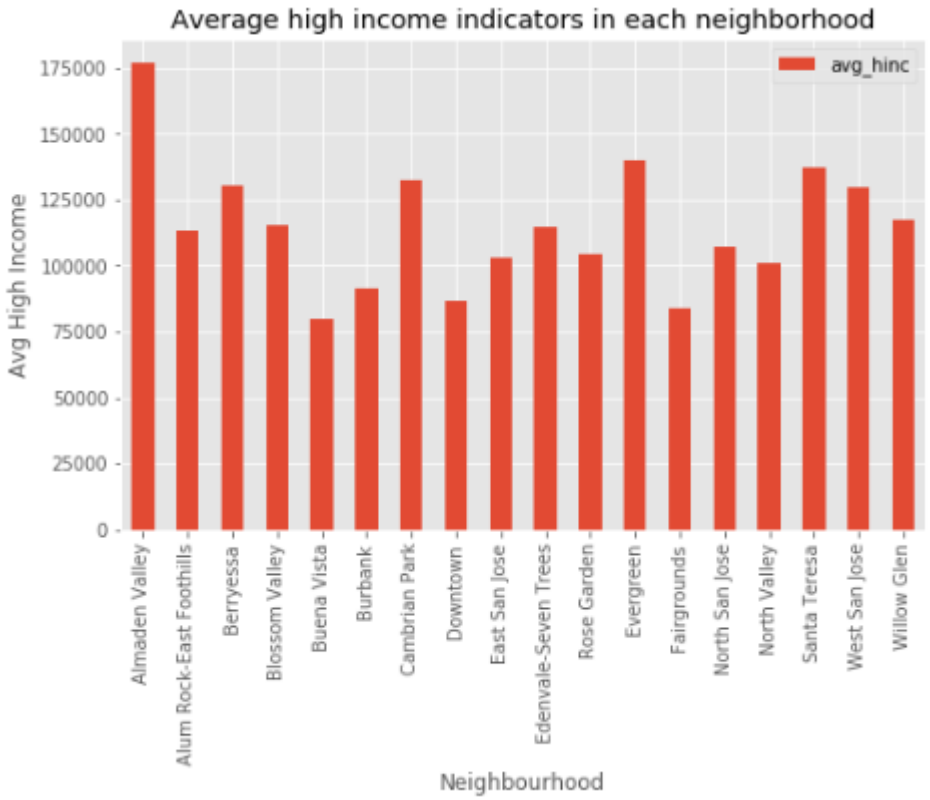
The average high income of 100’000 USD was considered as a threshold to determine the potential for a financial advice. In the case of median high income, it was fixed as 90’000 USD.

In order to enable an investor to make a decision on establishing a financial advisory institution, the following factors were considered as indicators:

- The proportion of neighbourhoods which have more than average high income.
- The proportion of neighbourhoods having more than median high income
- The distribution of percentage of owners and percentage of renting houses amongst the neighbourhoods.

For the purpose of illustration, one of the important attributes, namely 'avg_hinc' (average high income), was used to create a bar chart along with the different neighbourhoods of San Jose. For this purpose, a separate data frame was created with only Neighbourhood and avg_hinc as attributes.

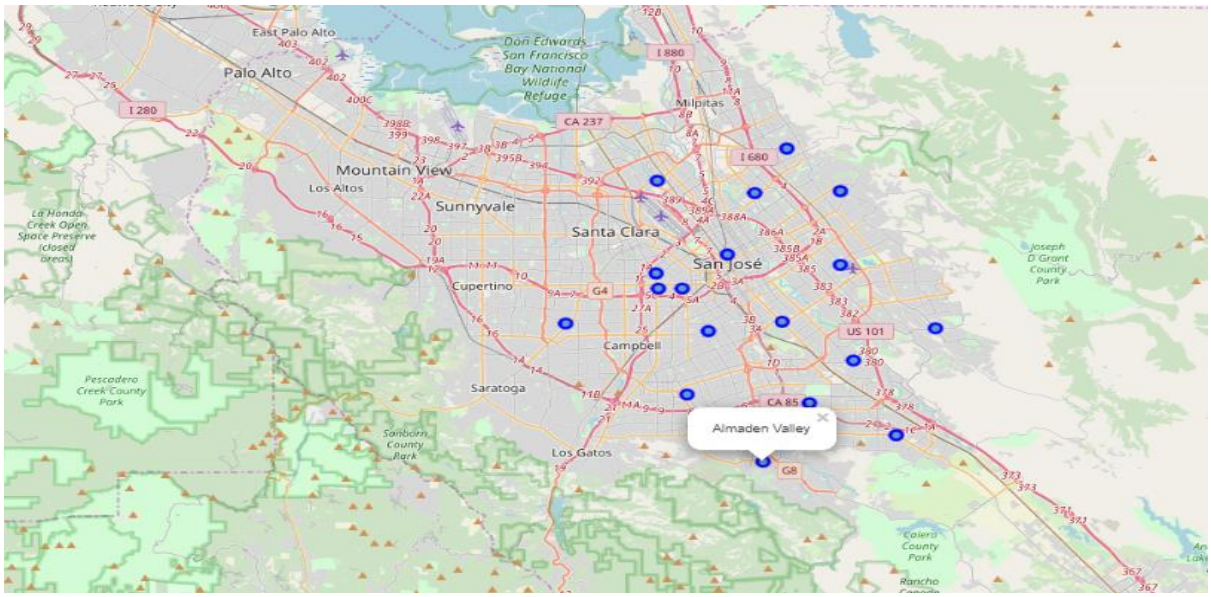
Neighborhood	avg_hinc
Almaden Valley	176918
Alum Rock-East Foothills	113307
Berryessa	130216
Blossom Valley	115350
Buena Vista	79466

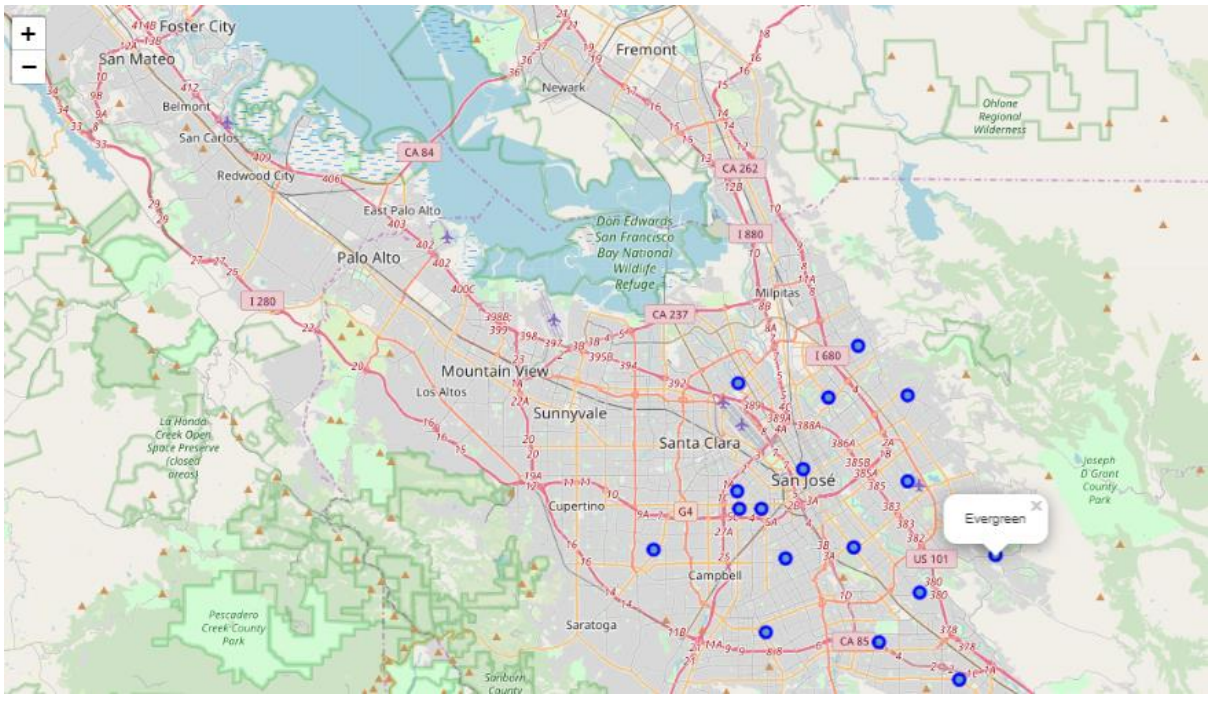


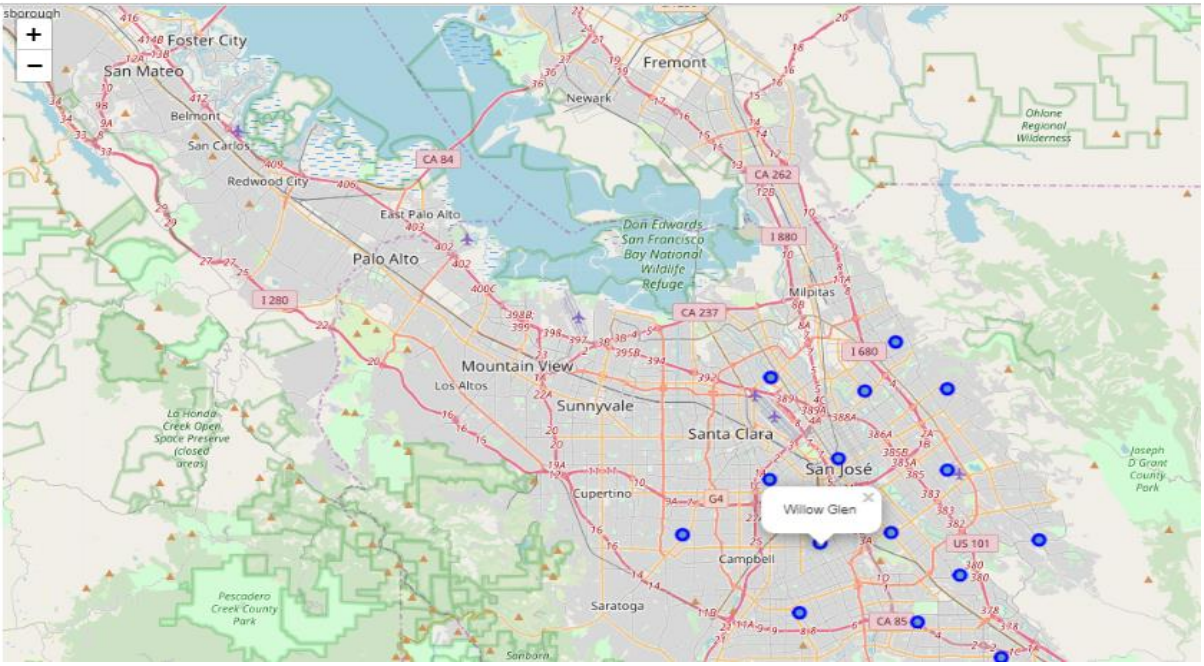
A graphical view of San Jose was provided:

The latitudes and longitudes of San Jose were identified based on address using an agent. According to the latitudes and longitudes, a map of San Jose was created using Folium. The neighbourhoods were superimposed on top of the map. In addition to the statistical data, this map provides a supplementary graphical view of the city and the neighbourhoods for evaluation by a potential investor.

Sample visuals with two the highest average income neighbourhood (Almaden Valley) and two others with average high income more than 100'000 USD s is provided below:







In addition to the analysis based on the Neighbourhood and economic indicators, the venues of the neighbourhoods are also explored and analysed using Foursquare. For that purpose, the latitudes and longitudes of each neighbourhood were used. The top 200 venues within a range of 2000 metres were explored for each neighbourhood.

Each neighbourhood was analysed and the data was grouped by neighbourhood by taking the mean of the frequency of occurrence of each category as shown in the below sample:

	Neighborhood	Airport Terminal	American Restaurant	Arts & Crafts Store	Asian Restaurant	Auto Garage	BBQ Joint	Bakery	Bank	Bar	Beer Garden	Big Box Store	Bookstore	Breakfast Spot	Bubble Tea Shop
0	Almaden Valley	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000
1	Alum Rock-East Foothills	0.000000	0.000000	0.000000	0.000000	0.166667	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000
2	Berryessa	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000
3	Blossom Valley	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000
4	Buena Vista	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.02439	0.000000	0.00000	0.024390	0.000000	0.02439	0.000000
5	Burbank	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.055556	0.00000	0.000000	0.055556	0.00000	0.000000
6	Cambrian Park	0.000000	0.000000	0.142857	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.142857	0.000000	0.00000	0.000000
7	Downtown	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.023810	0.00000	0.000000	0.02381	0.000000	0.000000	0.00000	0.000000
8	East San Jose	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000
9	Edenvale-Seven Trees	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000
10	Evergreen	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.058824	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000

Subsequently, for each neighbourhood, the venues were sorted to display the top 10 venues as shown in the below sample:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Almaden Valley	Playground	Gym / Fitness Center	Restaurant	Yoga Studio	Farmers Market	Cosmetics Shop	Dance Studio	Deli / Bodega	Department Store	Dessert Shop
1	Alum Rock-East Foothills	Martial Arts Dojo	Pizza Place	Video Store	Deli / Bodega	Auto Garage	Fast Food Restaurant	Dog Run	Farmers Market	Falafel Restaurant	Eye Doctor
2	Berryessa	Construction & Landscaping	Fish & Chips Shop	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Discount Store	Dive Bar	Dog Run	Ethiopian Restaurant
3	Blossom Valley	Food Truck	Mexican Restaurant	Convenience Store	Farmers Market	Optical Shop	Japanese Restaurant	Dance Studio	Deli / Bodega	Department Store	Dessert Shop
4	Buena Vista	Motorcycle Shop	Dessert Shop	Grocery Store	Pharmacy	Video Store	Thrift / Vintage Store	Fast Food Restaurant	Smoke Shop	Men's Store	Other Repair Shop

As a next step, the data was clustered using K-means as it is the most common cluster method and is fit for clustering the neighbourhoods in San Jose

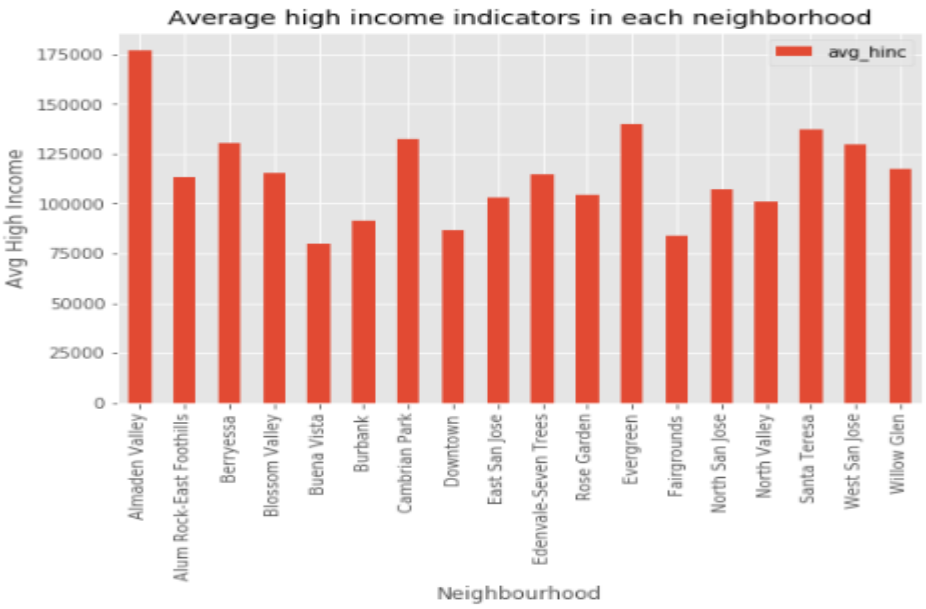
On the one hand the average high income levels provide an indicator for economic potential of the population to setup an advisory financial institution. On the other hand, the venues data would indicate the economic activity in the city.

Hence, both the income related data and venue related data are merged to provide a comprehensive view of San Jose and the neighbourhoods. The merged data would provide the basis to potential investors to evaluate and decide to establish a financial advisory institution.

4. Results

Upon analysis of the economic indicators, the results are

- 14 out of 18 neighbourhoods which means nearly 78 percent of neighbourhoods have 'avg_hinc' (average high income) value of more than 100'000 USD.
  Result: It is a very high figure. Hence the result is positive which indicates the availability of disposable income in large number of neighbourhoods. Hence it presents a good market potential for advises on financial matters especially investments.



- 9 out of 18 neighbourhoods which is 50 percent of the neighbourhoods have a 'medhinc_cy' (median high income of above 90000 USD.

  Result: I this case also, the figure is quite good presenting a good market potential for financial advice.

- In 12 out of the 18 neighbourhoods, which is nearly 67 percent of neighbourhoods, the percentage of owners is more than 50 percent in the population. In the rest of the neighbourhoods, the percentage of rent is high. Basically, it is a balanced distribution of the percentage of owners and percentage of people renting houses across the neighbourhoods.

Result: A very positive result. Those who already own a house can be offered better mortgage/interest rate related advices. Those who are on rent can be provided with advises regarding purchase of property. Hence, there is a good potential for financial advice.

- The attribute 'medage_cy' shows that San Jose is a city of middle age group combined with average high income.
  Result: Even though it is not a significant factor, the population with above 30 years age group are quite active in terms of financial investment. For eg. The population among this age where people are still renting property can aspire for an investment in property.

Basically, the attributes in the data frame respectively the economic indicators provide a positive potential as a market which would enable a potential investor to decide upon establishing a financial advisory institution.

In addition, both the economic indicators data and venue related data retrieved from Foursquare are merged to provide a comprehensive view of San Jose and the neighbourhoods. The merged data would provide the basis to potential investors to evaluate and decide to establish a financial advisory institution.

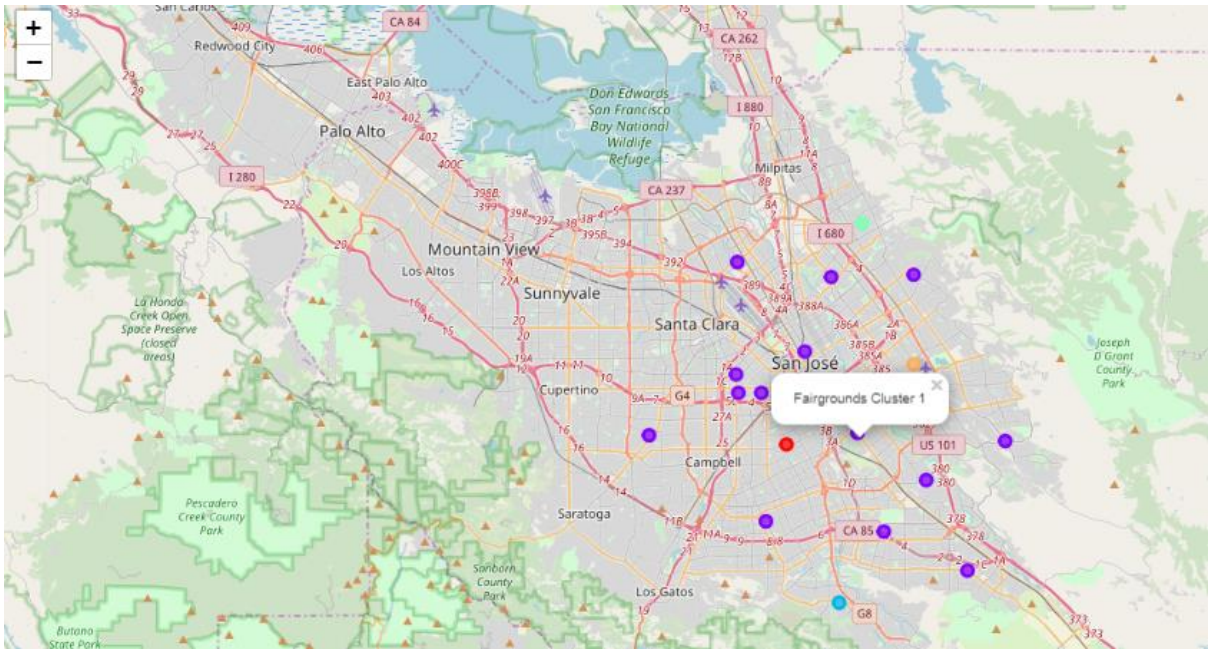
Sample is provided below:

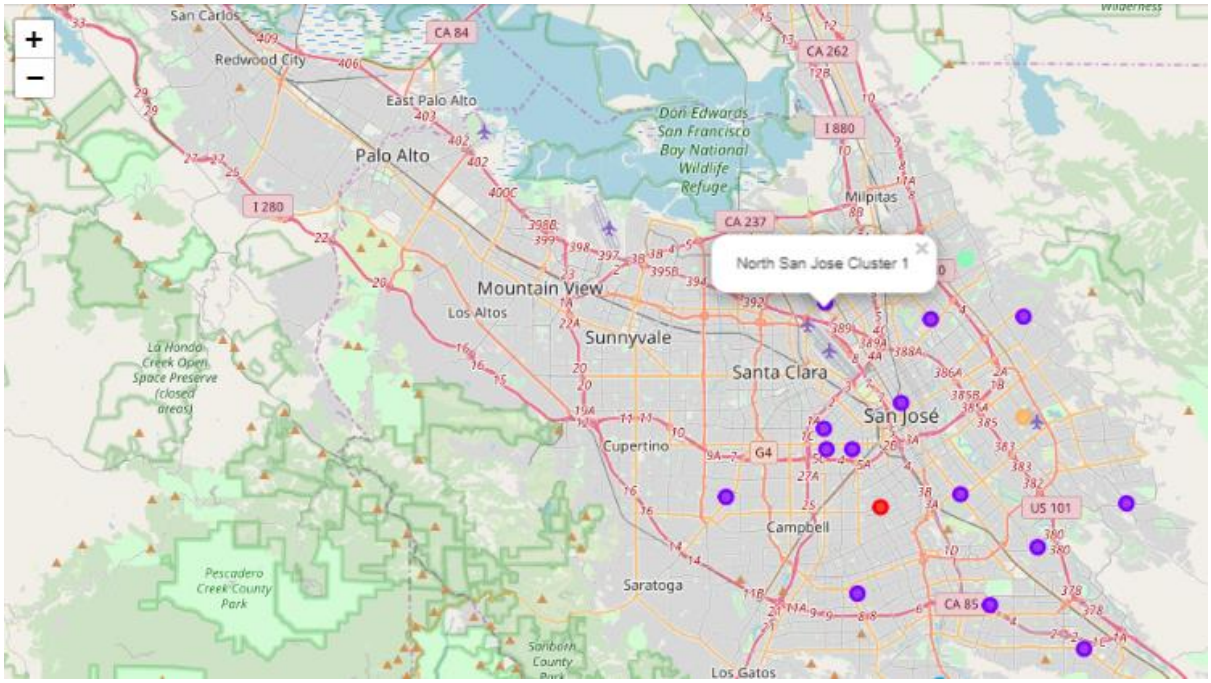
	City	Neighborhood	Latitude	Longitude	popdensity	avg_hinc	pct_own	pct_rent	medage_cy	medhinc_cy	city_parks	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	San Jose	Almaden Valley	37.2211	-121.869849	5612.600000	176918	0.734866	0.237213	41.210811	140164	0.068	2	Playground	Gym / Fitness Center	Restaurant	Yoga Studio
1	San Jose	Alum Rock-East Foothills	37.3772	-121.825214	10173.289474	113307	0.674909	0.300745	35.323684	89187	0.068	1	Martial Arts Dojo	Pizza Place	Video Store	Deli / Bodega
2	San Jose	Berryessa	37.4016	-121.856383	9173.466667	130216	0.744864	0.236287	37.942424	108157	0.068	3	Construction & Landscaping	Fish & Chips Shop	Dance Studio	Deli / Bodega
3	San Jose	Blossom Valley	37.2549	-121.843005	9977.140984	115350	0.657751	0.317510	36.403279	98611	0.068	1	Food Truck	Mexican Restaurant	Convenience Store	Farmers Market
4	San Jose	Buena Vista	37.3212	-121.916699	13386.700000	79466	0.275472	0.667481	32.340000	58664	0.068	1	Motorcycle Shop	Dessert Shop	Grocery Store	Pharmacy
5	San Jose	Burbank	37.3213	-121.930542	10040.090909	91458	0.370584	0.587205	35.945455	69339	0.068	1	Convenience Store	Discount Store	Mexican Restaurant	Karaoke Bar

The merged data shows a lot of venues (shops, construction firms, restaurants, Pharmacies) which also trigger a lot of economic activity and their employees also are a potential market. So in this case also, the result is positive.

Additionally, there is no financial advisory institution yet. So no competition is foreseen.

The potential investor can view the geography and clusters using the visual maps provided below with two neighbourhoods of Cluster 1:





Overall, the economic indicators, venue data and the visual maps provide a good source of information indicating the market potential which will enable an investor to establish a financial advisory institution.

5. Discussion

San Jose offers a huge potential for establishing a financial institution to manage the finances and advise on investments for the people working in this city.

Considering the diversity of the data and complexity involved, I have used the most important economic indicators of San Jose and neighbourhoods, their venues for analysis.

Such analysis can be extended to other cities also for identifying their economic potential. In addition, the base data contains additional attributes which can be retrieved for different purposes. For eg. Analysis of such data could also help to setup institutions for other purposes, for example education institutions, child care centres, hospitals.

I used the K-means algorithm as part of this clustering study. For more accurate guidance, the exploration and analysis can be extended to streets in the neighbourhoods to even indicate a potential address for establishment of new institutions.

6. Conclusion

In general for any city and specifically as in the case of San Jose, the economic indicators provide a good view of the activity and potential to set up a financial institution.

The case study shows that essentially, the data plays a very significant role to identify the current situation of a location in terms of activity (for eg. economic) and also for making predictions using statistics and machine learning.