A description of the problem and a discussion of the background

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.

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 study different opportunities and make 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.

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 setup a shop and provide customized financial investment services for individuals working in San Jose along with their families.

A description of the data and how it will be used to solve the problem.

The data is retrieved from the Spatial Data Repository of NYU [2] (U.S. Neighbourhoods greenness measures and social variables). The data includes many attributes and the following attributes would be mainly used for analysis:

- Population density
- Average high income
- Percentage owning houses
- Percentage renting
- Median age
- Median high income
- City parks.

In addition, the Neighbourhood name and the coordinates (latitude and longitude) are the attributes.

The .json file from the website contains data related to many other cities in the USA, in addition to San Jose. Therefore, data needs to be pre-processed and filtered to retain only the San Jose relevant indicators.

Data samples for data retrieved from Spatial Data Repository of NYU:

1. Base containing the data for all the cities


```
{'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,
 'medfinc_cy': 72176
```


2. Head of dataframe containing the data for all 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

3. 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

In order to obtain the data related to venues, I would be using Four Square to explore and analyse each Neighbourhood of San Jose to identify the venues in the neighbourhood. Both the income related data and venue related data would be merged to provide a comprehensive view of San Jose and the neighbourhoods. I would be using the K-means algorithm to cluster the neighbourhoods.

Data samples from Four Square:

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': -121.8472871770191},
   '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. 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
```

3. Sample venues in the neighbourhoods of San Jose

3]:

	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

On the one hand the income levels provide an indicator for economic potential of the population to setup an advisory financial institution. On the other hand, the data related to venues in different neighbourhoods of San Jose provides an indicator of economic activity within the neighbourhoods. These two when merged together would provide the basis to potential investors and enable them to decide and setup an investment advisory institution.