1. Introduction

1.1 Business Problem:
The city of Navi Mumbai is one of the largest planned cities of India. It is located on the west coast of the Indian subcontinent, in the state of Maharashtra. The city is well within the reach of the main city of Mumbai and considering the high rate of property in the main city, Navi Mumbai has seen a lot of growth in residential complexes in the last 10 - 20 years.

Considering the above factors, we will be solving the business problem of identifying the best location or the pincode within Navi Mumbai where one could open a restaurant. The proximity of Navi Mumbai to the main city of Mumbai and the connectivity between the two cities will help maximize the business opportunity if one decides to start a restaurant in the city of Navi Mumbai.

1.2 Beneficiaries:
The beneficiaries of solving this business problem would range from organized food giants like McDonalds, Jubilant Foods, Pizza Hut, etc to small and medium businesses looking to benefit from managing a single restaurant.

2. The Data

2.1 Data Source:
The data is sourced from two locations. The first is the All India Pincode directory. Click here to check the website. Here we will find the pin codes for all the locations within Navi Mumbai.

The second data source is the GeoNames postal code files for all countries. Click here to check the website. On this website we will find latitudes and longitudes against all the pincodes in India.

2.2 Data Description:
First we will import both datasets to check the respective data in each file:
All India Pincode directory -

```
[56]: df_Pin = pd.read_csv('Pincode_30052019.csv')
      df_Pin.head()
```

[56]:

	CircleName	RegionName	DivisionName	OfficeName	Pincode	OfficeType	Delivery	District	StateName
0	Andhra Pradesh Circle	Kurnool Region	Anantapur Division	A Narayanapuram B.O	515004	BO	Delivery	ANANTHAPUR	Andhra Pradesh
1	Andhra Pradesh Circle	Kurnool Region	Anantapur Division	Akuledu B.O	515731	BO	Delivery	ANANTHAPUR	Andhra Pradesh
2	Andhra Pradesh Circle	Kurnool Region	Anantapur Division	Alamuru B.O	515002	BO	Delivery	ANANTHAPUR	Andhra Pradesh
3	Andhra Pradesh Circle	Kurnool Region	Anantapur Division	Allapuram B.O	515766	BO	Delivery	ANANTHAPUR	Andhra Pradesh
4	Andhra Pradesh Circle	Kurnool Region	Anantapur Division	Aluru B.O	515415	BO	Delivery	ANANTHAPUR	Andhra Pradesh

Latitudes and longitudes from GeoNames -

```
[54]: df_IN = pd.read_csv('IN.csv')
      df_IN.head()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/IPython/core/interactiveshell.py:3072: DtypeWarning: Columns (4,6) have mixed types.Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

[54]:

	countrycode	postalcode	placename	adminname1	admincode1	adminname2	admincode2	adminname3	admincode3	latitude	longitude	accuracy
0	IN	744301	Sawai	Andaman & Nicobar Islands	1	Nicobar	638	Carnicobar	NaN	7.5166	93.6031	4.0
1	IN	744301	Carnicobar	Andaman & Nicobar Islands	1	Nicobar	638	Carnicobar	NaN	9.1833	92.7667	3.0
2	IN	744301	Mus	Andaman & Nicobar Islands	1	Nicobar	638	Carnicobar	NaN	9.2333	92.7833	4.0
3	IN	744301	Lapathy	Andaman & Nicobar Islands	1	Nicobar	638	Carnicobar	NaN	9.1833	92.7667	3.0
4	IN	744301	Kakana	Andaman & Nicobar Islands	1	Nicobar	638	Carnicobar	NaN	9.1167	92.8000	4.0

Next, from the All India Pincode directory, we will filter the data for Navi Mumbai and using the pin codes which is common in both the files, merge it with the data of latitudes and longitudes

```
[58]: is_Navi =  ['Navi Mumbai Region']
      df_Nmum = df_Pin[df_Pin.RegionName.isin(is_Navi)]
      df_Nmum.head()
```

[58]:

	CircleName	RegionName	DivisionName	OfficeName	Pincode	OfficeType	Delivery	District	StateName
78787	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Abhona S.O	423502	SO	Delivery	Jalgaon	Maharashtra
78788	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Adgaon B.O	423101	BO	Delivery	Jalgaon	Maharashtra
78789	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Aghar BK B.O	423201	BO	Delivery	Malegaon	Maharashtra
78790	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Aghar KH B.O	423208	BO	Delivery	Malegaon	Maharashtra
78791	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Ahergaon B.O	422209	BO	Delivery	Malegaon	Maharashtra

```
[86]: df_final5=pd.DataFrame(df_final4)
      df_final5.head()
```

[86]:

	Pincode	CircleName	RegionName	DivisionName	OfficeName	OfficeType	Delivery	District	StateName	latitude	longitude
0	423502	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Abhona S.O	SO	Delivery	Jalgaon	Maharashtra	20.0947	73.9282
12	423101	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Adgaon B.O	BO	Delivery	Jalgaon	Maharashtra	20.3237	74.2071
32	423201	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Aghar BK B.O	BO	Delivery	Malegaon	Maharashtra	20.5498	74.4557
37	423208	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Aghar KH B.O	BO	Delivery	Malegaon	Maharashtra	20.2592	74.0714
47	422209	Maharashtra Circle	Navi Mumbai Region	Malegaon Division	Ahergaon B.O	BO	Delivery	Malegaon	Maharashtra	20.1704	73.9923

We can use the above final dataframe to identify clusters that are best suited to run a restaurant business.

Now that our data is ready, we can move on to the analysis part.

3. Methodology

We will apply the K-means cluster analysis on the location data to identify the pin codes where we can suggest to start a restaurant.

What is K-means clustering?
Clustering is an exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different.

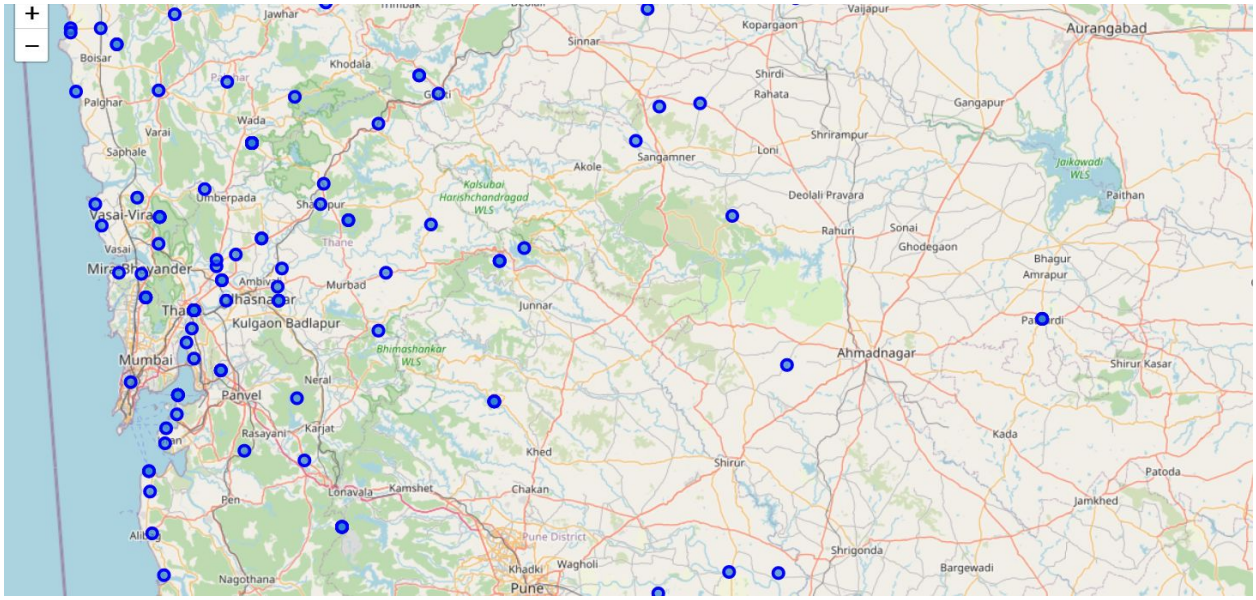
Unlike supervised learning, clustering is considered an unsupervised learning method since we don't have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

The K-means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Using the foursquare API, we will fetch the top 10 venues for all Navi Mumbai pin codes and use the K-means clustering algorithm to find the best locations to start a restaurant business.

What is foursquare API?
The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag.

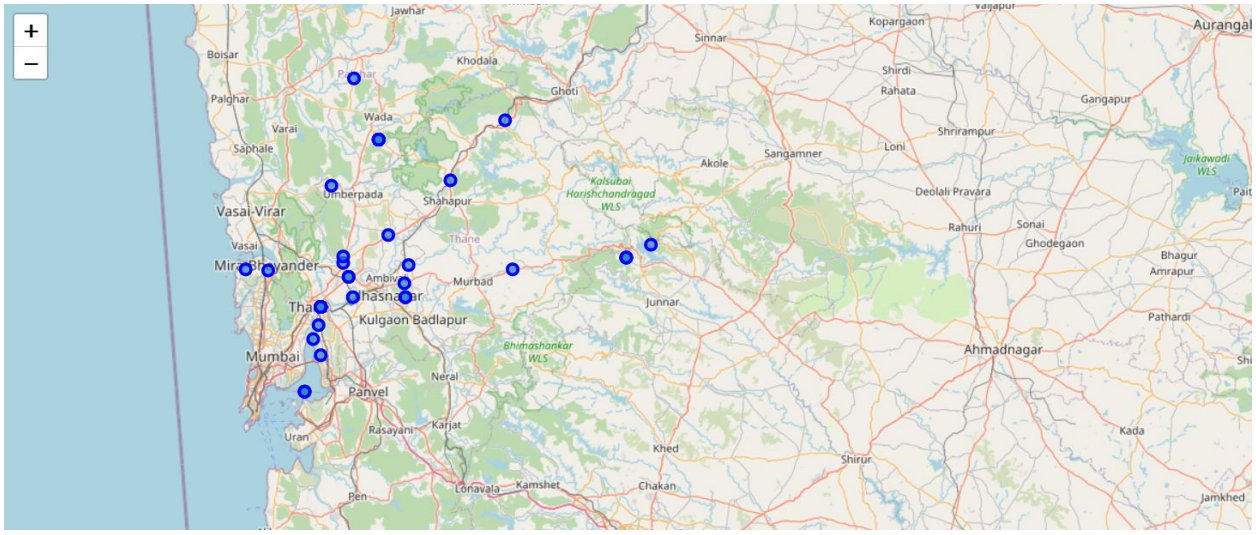
First, let us view our data set using folium maps:



As we can see from the above data, the initial dataset for Navi Mumbai consists of pin codes that are spread all across Maharashtra. This will not help in our analysis as we are looking for locations within Navi Mumbai that are closer to the main city of Mumbai.

Hence, we shall filter the data further upto all the pin codes within the district of Thane, located in the city of Navi Mumbai.

The below folium map shows the location of the pincodes in the district of Thane.



Looking at the above folium map, we can conclude that the pin codes in the district of Thane are located closer to the main city of Mumbai and hence we can proceed with our K-means cluster analysis.

Next, we will use the unique client credentials shared by the Foursquare API to fetch the top 100 venue categories from all the neighborhoods, within a radius of 500 meters, in the Thane district and put it in a Pandas dataframe.

[52]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Airoli B.O	19.151	72.9962	Domino's Pizza	19.148078	72.995161	Pizza Place
1	Airoli B.O	19.151	72.9962	Hotel Vaibhav Sip N Dine	19.147927	72.999466	Hotel Bar
2	Airoli B.O	19.151	72.9962	Café Coffee Day	19.148130	72.995247	Café
3	Airoli B.O	19.151	72.9962	McDonald's	19.147545	72.995163	Fast Food Restaurant
4	Airoli B.O	19.151	72.9962	Sector-9 Bus Stop	19.148233	72.994297	Bus Station

From the resulting dataframe, we can conclude that we have been able to fetch 24 unique categories for all the neighborhoods.

Using one hot coding, we shall fetch the different categories of venues for all the neighborhoods in the Thane district.

[55]:

	Neighborhood	ATM	Asian Restaurant	Burger Joint	Bus Station	Café	Chinese Restaurant	Convenience Store	Fast Food Restaurant	Gym	Hotel	Hotel Bar	Ice Cream Shop	Indian Restaurant	Lake	Multiplex	Nature Preserve	Pizza Place	Plaza	Restaurant
0	Airoli B.O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1	Airoli B.O	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2	Airoli B.O	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Airoli B.O	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4	Airoli B.O	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

According to the frequency, we can check the top 10 venues for all the neighborhoods and put them in a single dataframe. This will allow us to run K-means cluster analysis on a single dataset.

:61]:

	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	Aghai B.O	Lake	Train Station	Hotel Bar	Asian Restaurant	Burger Joint	Bus Station	Café	Chinese Restaurant	Convenience Store	Fast Food Restaurant
1	Airoli B.O	Hotel Bar	Asian Restaurant	Bus Station	Café	Pizza Place	Fast Food Restaurant	Gym	Toy / Game Store	Train Station	Burger Joint
2	Chamble B.O	Restaurant	Train Station	Hotel Bar	Asian Restaurant	Burger Joint	Bus Station	Café	Chinese Restaurant	Convenience Store	Fast Food Restaurant
3	Dombivali I.A. S.O	Hotel	Train Station	Toy / Game Store	Asian Restaurant	Burger Joint	Bus Station	Café	Chinese Restaurant	Convenience Store	Fast Food Restaurant
4	Dombivali S.O	Snack Place	Café	Pizza Place	Fast Food Restaurant	Indian Restaurant	Gym	Train Station	Hotel	Asian Restaurant	Burger Joint

We can then check the clusters one by one to identify the exact set of locations where we can suggest to investors for opening a restaurant.

[63]:

	Pincode	CircleName	RegionName	DivisionName	OfficeName	OfficeType	Delivery	District	StateName	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	400708	Maharashtra Circle	Navi Mumbai Region	Navi Mumbai Division	Airoli B.O	BO	Non Delivery	THANE	Maharashtra	19.1510	72.9962	0.0	Hotel Bar	Asian Restaurant	Bus Station	Café	Pizza Place
1	400614	Maharashtra Circle	Navi Mumbai Region	Navi Mumbai Division	Belapur Node III S.O	SO	Non Delivery	THANE	Maharashtra	19.1941	73.0002	NaN	NaN	NaN	NaN	NaN	NaN
2	400706	Maharashtra Circle	Navi Mumbai Region	Navi Mumbai Division	Darave B.O	BO	Delivery	THANE	Maharashtra	18.9894	72.9610	NaN	NaN	NaN	NaN	NaN	NaN
3	400701	Maharashtra Circle	Navi Mumbai Region	Navi Mumbai Division	Ghansoli S.O	SO	Delivery	THANE	Maharashtra	19.1167	72.9833	NaN	NaN	NaN	NaN	NaN	NaN
4	400703	Maharashtra Circle	Navi Mumbai Region	Navi Mumbai Division	K.U.Bazar S.O	SO	Non Delivery	THANE	Maharashtra	19.0787	73.0005	0.0	Theater	Bus Station	Café	Hotel	Train Station

4. Results & Observations:

We will look at each cluster one by one and identify the one that helps us best in resolving our business problem.

The first cluster has four pin codes and out of the 40 most common venues in this cluster, 11 venues are not related to restaurants. 72.50% of the venues in these four pin codes belong to the restaurant category.

The second cluster has 2 pin codes and out of the 20 most common venues in this cluster, 8 venues are not related to the restaurants category. 60% of the venues in these 2 pincodes belong to the restaurant category.

The third cluster has 1 pin code and out of the 10 most common venues in this cluster, 4 venues are not related to the restaurants category. 60% of the venues in this pincode belong to the restaurant category.

The fourth cluster has 2 pin codes and out of the 20 most common venues in this cluster, 9 venues are not related to the restaurants category. 55% of the venues in these 2 pincodes belong to the restaurant category.

The fifth and the final cluster has 2 pin codes and out of the 20 most common venues in this cluster, 6 venues are not related to the restaurants category. 70% of the venues in these 2 pincodes belong to the restaurant category.

5. Conclusion

From the above results and observations, we can conclude that the first cluster with 72.50% on the top 10 most common venues and four pin codes, is our best chance of success if we want to start a restaurant business.