**Setting franchises for mid-range car in Mumbai**

**Priya Yogendra Rana**

**15th Feb 2021**

1. **Introduction**
   1. **Business Background & Problem**
   2. **Concerns**
   3. **Assumptions**
   4. **Advantage**
   5. **Target Audience**
2. **Data Acquisition and Cleaning**
   1. **Data Sources**
   2. **Data Cleaning**
   3. **Feature Selection**
3. **Methodology**
   1. **Exploratory Data Analysis**
      1. **Calculating the average real estate prices**
      2. **Relationship between real estate prices and rental prices**
4. **Clustering Modelling**
   1. **Determine the type of clustering model**
   2. **Determine the number of clusters required**
   3. **Calculate the cluster center**
   4. **Visualise the clusters**
   5. **Map and define each cluster**
5. **Results**
   1. **Determine the location coordinates**
   2. **Explore the areas around the coordinates**
6. **Observations & Recommendations**
7. **Conclusion & Future Directions**
8. **Introduction**
   1. **Background & Problem**

Mumbai being called a city of dreams, is hugh (approx. 233 sq mi), densely populated (approx. 1.84 crores) and busy. It provides with an opportunity to promote and sell shopping goods like mid-range or premium brands which have more viable market in Mumbai rather than in any small city in India. However, the geographical expanse of the city also poses a dilemma as to where to set shop. There are plethora of areas and localities - some more better off while others aren’t. Some are easily approachable and others aren’t. Also, there are localities which are posh but lack in footfalls.

So, the business problem revolves around how to identify and locate an approximate location for Car showroom ranging approximately around Rs. 7-8 lacs. As there is no demographics available for Mumbai - area or locality wise, it poses a question of how to identify profitable areas within Mumbai and narrow them down to few. The localities have to be such that they not only provide an exposure the brand is looking for but also attract the right set of customers in good numbers. Basically, the objective is to make the company profitable by locating places which can generate maximum revenue for them.

* 1. **Concerns**

The first concern is to how to identify most appropriate areas for the company, which are popular and can bring the buyer a complete shopping experience. Reason being, people tend to combine couple of activities together even while shopping a product like car. Just to step out for checking a car brand is not supposed to be a chore. Choosing a brand can be a practical decision but stepping out to check the car combined with a complete shopping experience will lead to a positive decision. So, the areas/localities for the car showroom position has to be such that the buyers can plan other activities as well. The second and the final concern is that there is no up-to-date information on the way the localities are spelled and no particular dataset where the property rates of all localities are mentioned. In Mumbai, lot of localities are clubbed under one locality where real estate rates are concerned.

* 1. **Assumption**

Unfortunately, I have to make an assumption in this project as there is no demographics data available locality wise. To narrow down or eliminate localities, I will remove localities where minimum property rates are not above Rs.15000 per sq. ft. Reason being, the purchasing capacity can somewhat be judged by the real estate rates. In Mumbai, it is very difficult to live or own a property which is highly priced if the concern person is not well off and have good purchasing power.

* 1. **Advantage**

Fortunately, I know the ins and outs of Mumbai in detail – having been living for more than two decades there. The outcome of the model will give an idea how on the mark or vice-versa the model is.

* 1. **Target Audience**

Finally, the target audience is the company who is new on the block and has a mid-range car as a product to be sold in Mumbai. However, the model that will be prepared will be such that given a new set of criteria and little bit of tweaking in the script, we can prepare a customisable solution for another company as well.

1. **Data Acquisition and Cleaning**
   1. **Data Sources**

There will be 3 dataset that will help resolve the business problem in question.

The **first dataset** will be from Foursquare API. This dataset will contain many columns but the ones which will solve my purpose are the following:

**venue id, name, address, geographical location (latitude and longitude), postal codes, category**.

This dataset will form the base of the project. This main purpose of this dataset will be to provide me with the list of postal code and geographical locations of all popular places in Mumbai. It will be very difficult to extract the locality from the address column in the dataset as there are various ways to write a locality name in Mumbai. Hence to proceed further, I will be using the postal codes to determine the localities with the help of subsequent dataset.

The **second dataset** will be from a website (https://mumbai7.com/postal-codes-in-mumbai/) having the most up-to-date list of all **postal codes** along with their respective **locality name** in Mumbai. I will be scrapping the dataset directly from the website. It will help to standardize the locality names and further help in merging the above dataset with the third dataset. In other words, it will help connect the first and third dataset on a common column i.e. locality and subsequently lead the formation of final dataset.

However, there are issues that needs to be resolved here. There will be some postal codes and localities that need to be revisited. Reason being, real estate prices are not as per the postal codes in Mumbai. Couple of localities are lugged under one locality in real estate world. The other reason to revisit them is few postal codes are incorrect or the fact that there are plenty of ways to write a locality name, which doesn’t help with the normalization. These few records need to be corrected one by one.

The **third dataset** will be from another website (www.99acres.com). This website will provide with the real estate prices of all localities in and around Mumbai. For all the localities, we will get following:

**Locality name, min retail price per sq ft, max retail price per sq ft, quarter change, rental rates for 1, 2, & 3 BHK**.

Out of all, the first 3 will be required as of now and merged with the first dataset. Now the dataset is complete for further processing. Then another column will be introduced, which will be the mean of Minimum Retail Price per square feet and Maximum Retail Price per square feet. This column along with the latitude and longitude of the popular venues will be used in clustering model to find the apt coordinates of centre of each cluster, the model defines.

* 1. **Data Cleaning**

I was aware that data cleaning will take considerable time and it did. As stated above, normalizing the three datasets on one column was a cumbersome process. The first dataset generated from Foursquare API came with considerable number of blanks postal code. Some of the respective addresses contained area details, on the basis of whom I populated the blank postal codes. However, there were few records which were still Null and 1 postal code which was incorrect as its length exceeded the normal length. This was corrected by writing separate scripts for them. Upon further analysis, there were 3 postal codes which were invalid and no more in use. To correct them, another script was written for individual cases as there were not many records to make changes to.

Upon mapping these locations, there was another realisation that there will be certain postal codes against whom real estate prices will not be available. In Mumbai, there are many localities who are clubbed under another locality with different postal code for real estate prices. Here at this stage, I was able to locate one such postal code and changed its postal code with the one which lists the real estate prices of that postal code.

To have a standard locality name, I took the help of Mumbai municipality website that provides a comprehensive list of postal code and locality name. But here, there were 6 cases where I had to correct the locality name as they were misspelled before finally merging this dataset with the previous name.

Then I extracted the real estate prices from a website but I took only records of those localities who existed in my previous dataset. There was only 1 record which didn’t contained any real estate details. For that particular locality, I did some google search and updated the record separately. Like this finally my dataset was ready for further analysis and processing.

* 1. **Feature Selection**

As and when extracted data either from Foursquare API or tables/data extracted from website, I eliminated columns which would not solve any purpose with their existence in the dataset.

The following table shows the columns deleted/ignored and the reasons:

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Kept Features | Ignored/deleted Features | Reasons for dropping |
| FourSquare API set | venue id, name, address, geographical location (latitude and longitude), postal codes, category | referralId, reasons.count, reasons.items, crossStreet, labelled latlang, country, state, city, photos, group, pageid, neighborhood | These fields were not solving any purpose. I didn’t need them for any analysis nor for clustering. |
| Mumbai Postal Code | Postal Code, Locality | None | Not Applicable |
| Data from 99acres.com | Locality Name, Minimum Retail per sqft, Maximum Retail per sqft, Minimum 1BHK Rental | Quarter Change, Maximum IBHK Rental, Minimum IIBHK Rental, Maximum IIBHK Rental, Minimum IIIBHK Rental, Maximum IIIBHK Rental | They didn’t solve any purpose in any kind of analysis nor in the model. However, I didn’t drop these columns to show the rates simultaneously along with the cluster group for comparisons. |

1. **Methodology**
   1. **Exploratory Data Analysis**

This project required limited use of exploratory data analysis. There were 2 places I needed to perform the exploratory analysis – once to calculate the mean and another to calculate the correlation between two columns in the final dataset.

* + 1. **Calculating the average retail estate prices**

The dataset scraped from the retail estate website gives the minimum and maximum real estate prices. For simplicity sake and further comparisons there was a need to calculate the average. This also it easier to eliminate the areas which are below the mark. If the average retail price is below Rs. 15000 per sqft and the minimum retail price is also below Rs. 15000 per sqft, then these localities need to be excluded from the model. The localities deleted were not exactly where people would be interested to buy a mid-range car.

* + 1. **Relationship between** **real estate prices and rental prices**

For the model, there was a need to decide which columns will be used. There were columns about real estate price - both if the person owned the property or rented a property. Correlation was the best way to decide whether rental rate needed to be included or not. Correlation was performed so that it becomes easier to decide if rental prices should be included in the model and whether its inclusion will make any significant difference to the model ahead. As the correlation was highly positive with the score of .88, I decided not to consider the rental rates as it would not make significant difference to the model.

1. **Clustering Modelling**
   1. **Determine the type of clustering model**

There are 7 types of clustering models. But I narrowed down to k-means because there was a need to develop clusters and find the clusters centers – which will be the locations (coordinates) to get an idea where the company can set up franchise.

* 1. **Determine the number of clusters required**

Out of the total localities, the model helped in identifying the most feasible number of clusters. For this I used the Elbow method. In this method, where there is a bend which is significant, that is the number which we should consider as the number of clusters needed. The graph below helped in reaching to the conclusion that 3 clusters will be best.

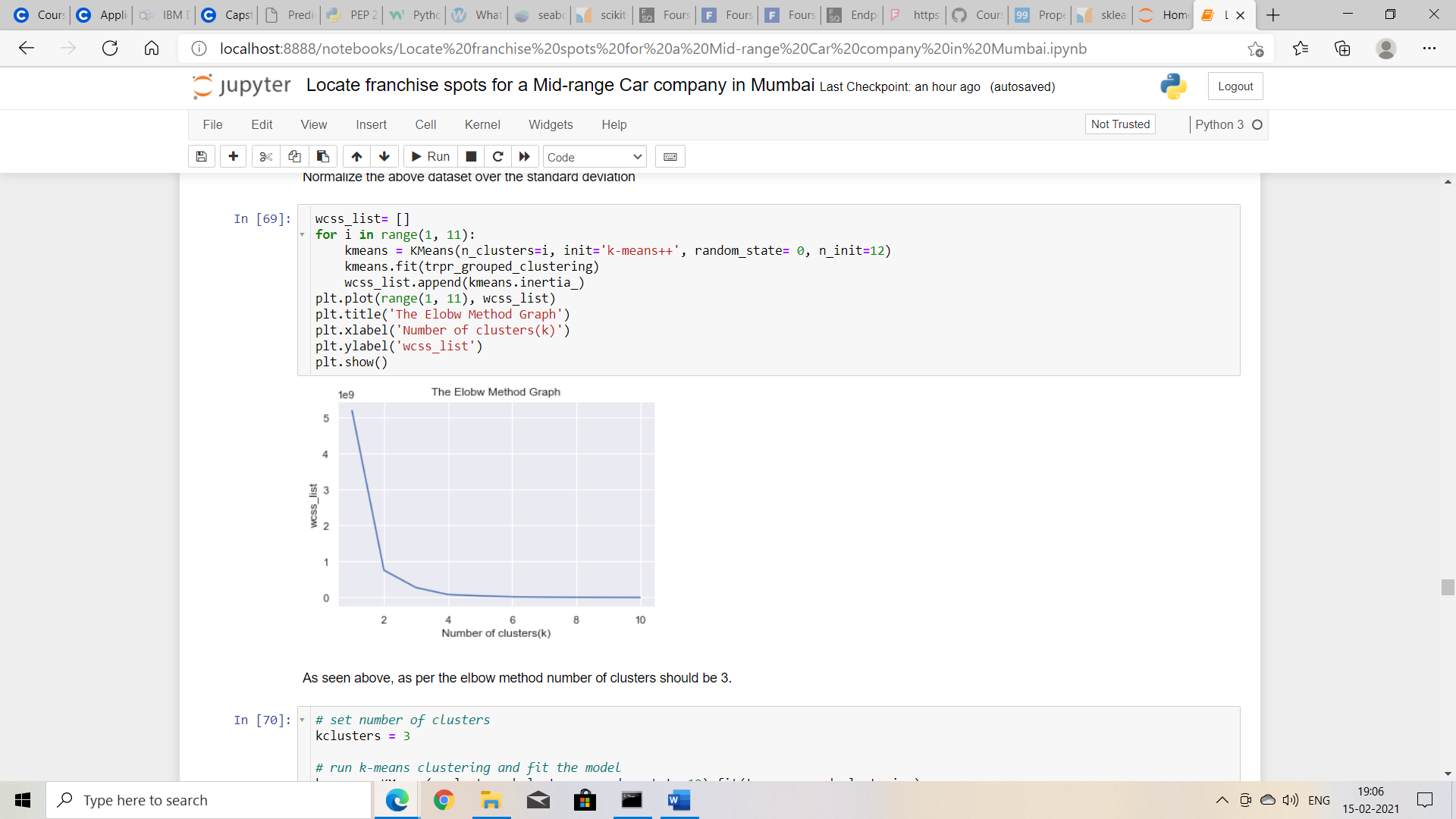


Figure 1: Graph to decipher how many clusters should be there.

* 1. **Calculate the cluster center**

The cluster centers provided by the model can be plotted on the graph to get an estimation, alignment and placement of the clusters. Here’s is what the model plotted:

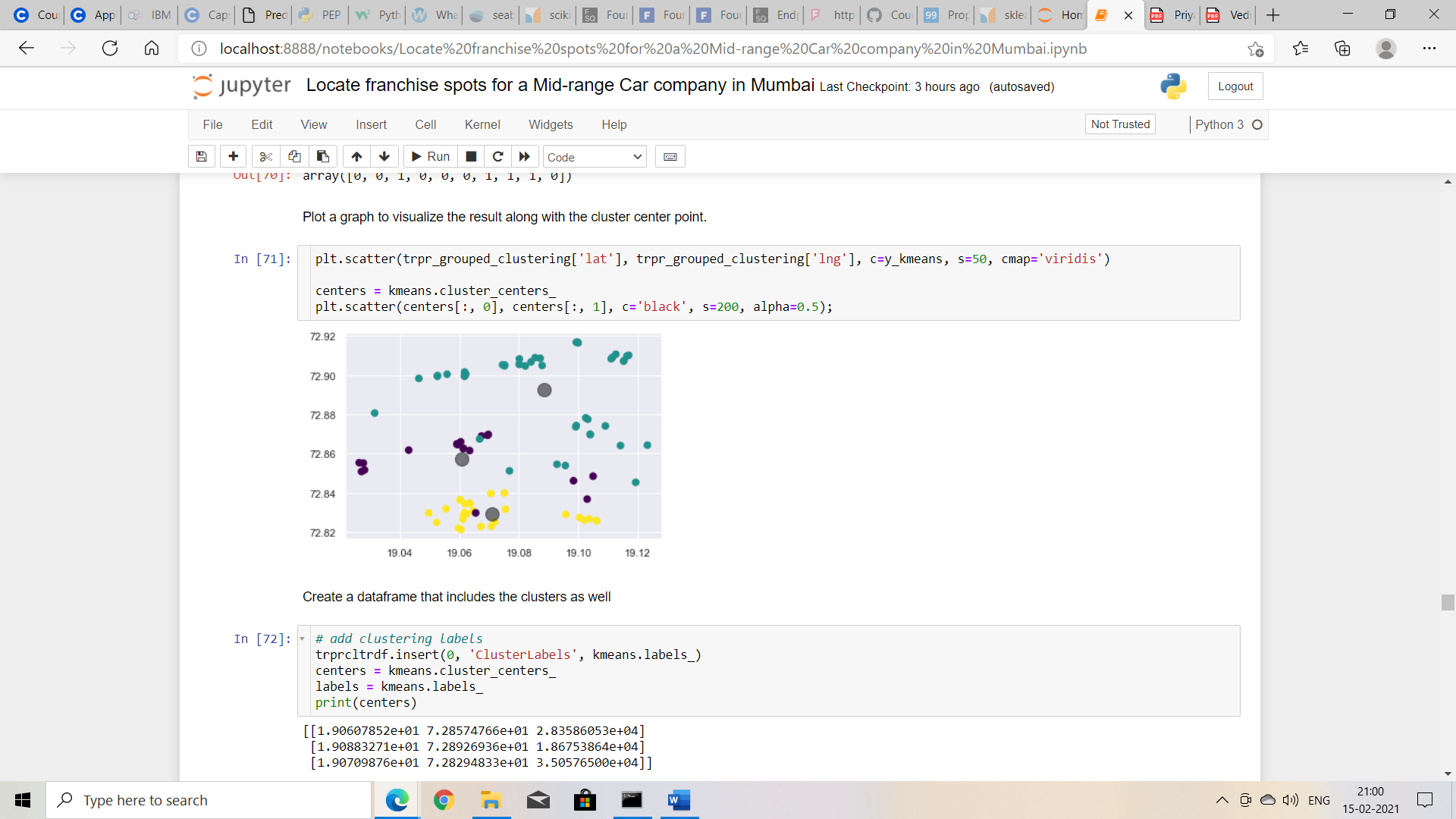


Figure 2: Cluster Graph

* 1. **Visualise the clusters**

When we check the shape of the cluster center, the results show 3 clusters in 3 dimensions. The cluster center themselves are 3 dimensional points and can be represented. The model also allows us to visualize what the centers look like. And after using the appropriate script, here’s what it looks like:

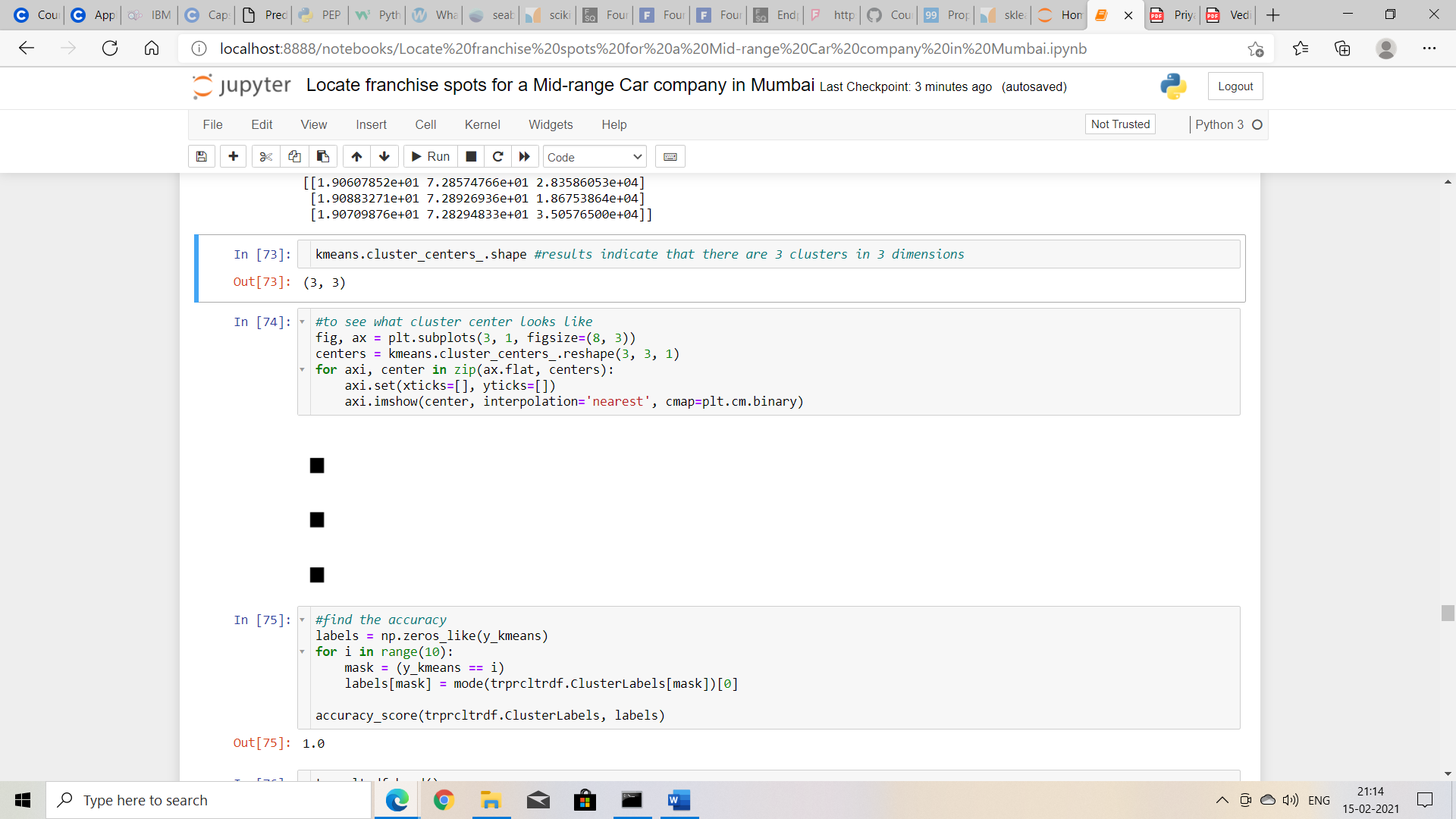


Figure 3: Visualize how cluster center would look like

* 1. **Accuracy Score and Confusion Matrix**

We can check the accuracy score and cross check it with the confusion matrix. The model generated an accuracy score of 1 and when checked against the confusion matrix, we get the following output:

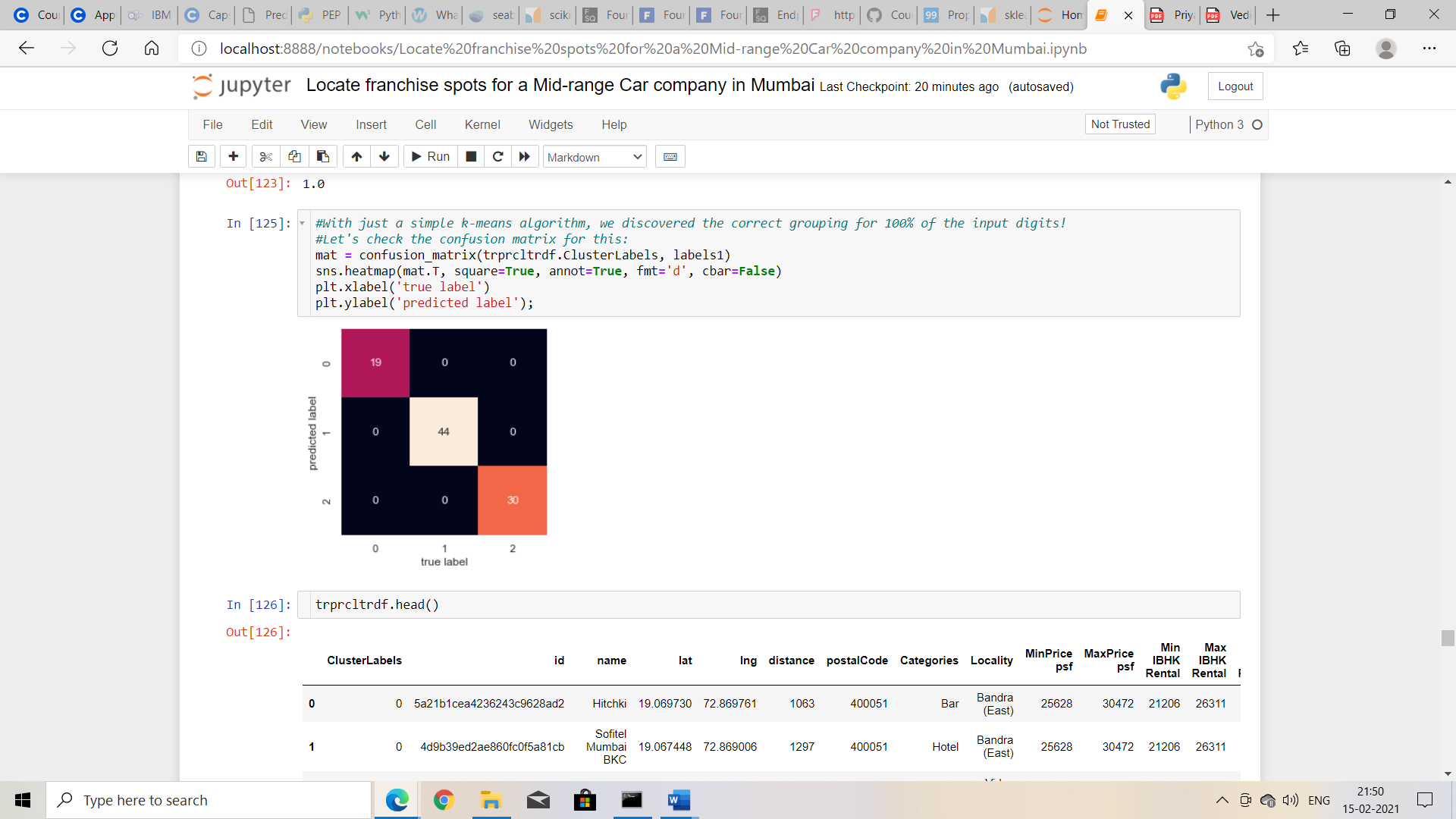


Figure 4: Confusion Matrix

* 1. **Map and define each cluster**

Plotting the clusters on the map will give fair idea of how the venues are placed and under which cluster. It won’t be divided on purely geographical basis as real estate price also plays an important role.

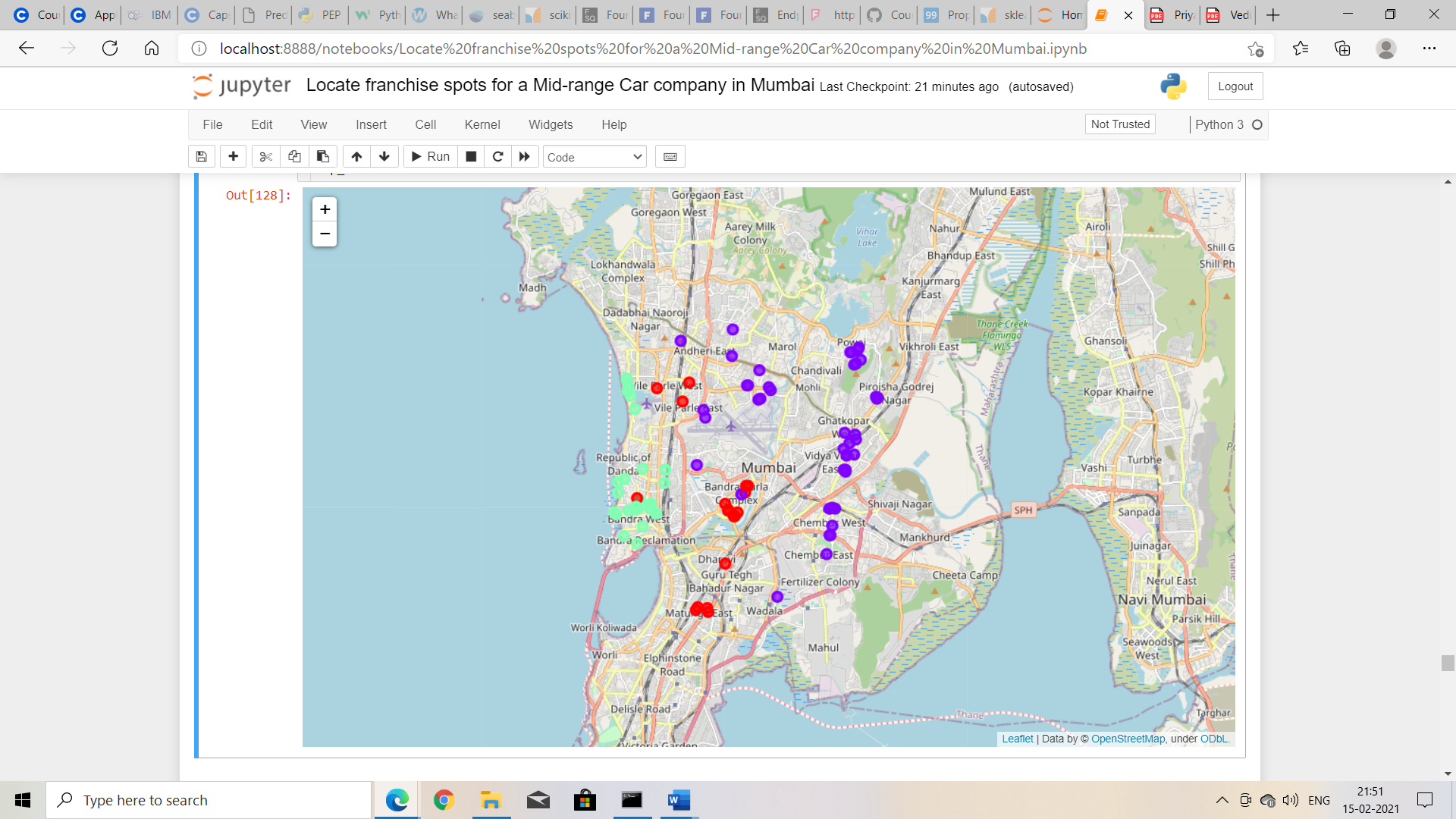


Figure 5: Mapping the cluster groups on the map.

1. **Results**

The model generated 3 coordinates and central point of real estate rates between the areas that fall in that particular cluster.

* 1. **Determine the location coordinates**

The three coordinates which are the center of each cluster when plotted on google map lies in Khar (West), Bandra Kurla Complex and Ghatkopar (West).

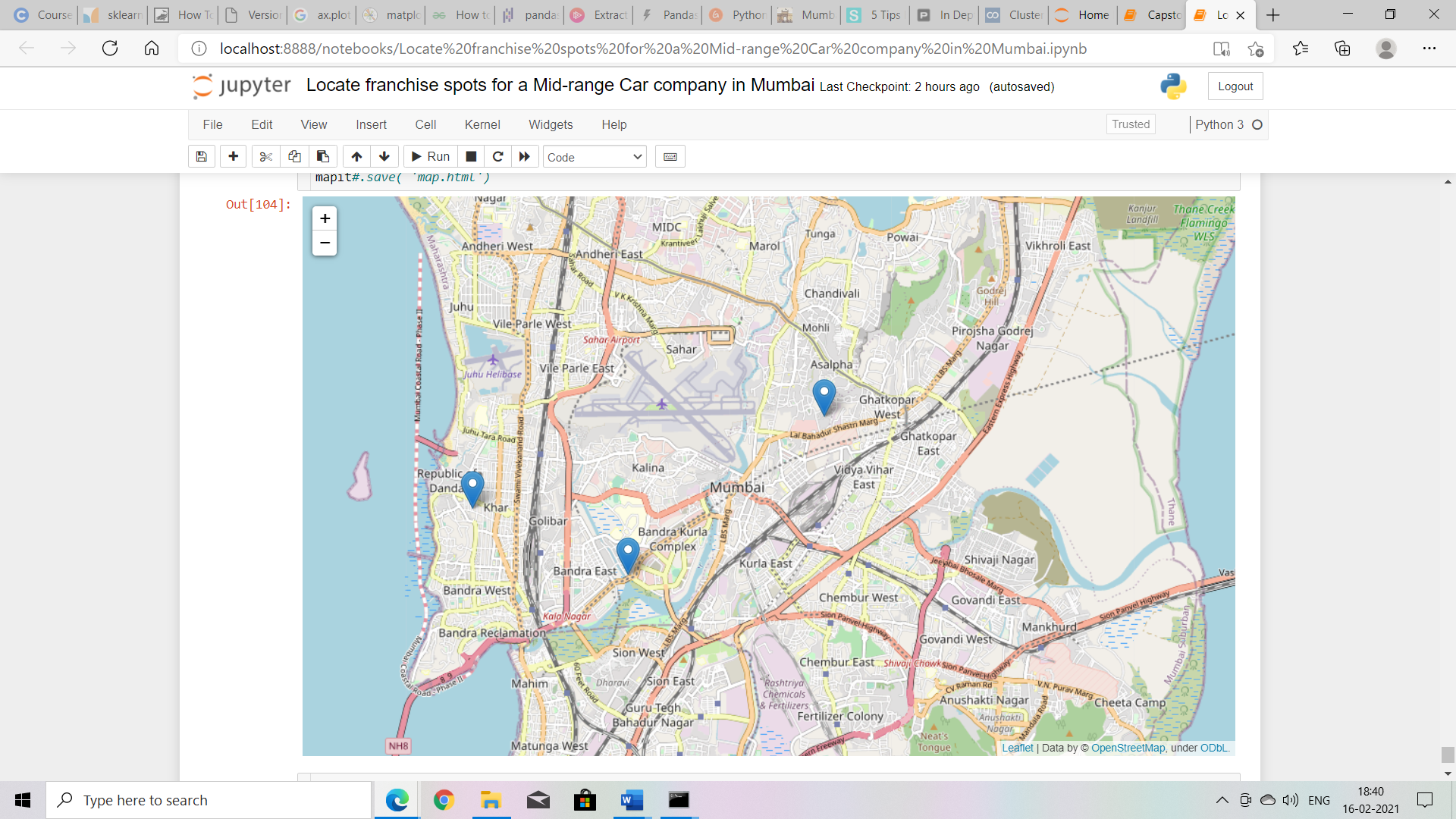


Figure 6: Mapping the cluster center on a map

* 1. **Explore the areas around the coordinates**

The three coordinates are further passed to Foursquare API to explore the area around with a radius of 500m. For each of the coordinates I can find venues that are popular, the category these venues fall under. This is to judge whether the area and nearby places can lure the prospective buyers.

1. **Observations & Recommendations**

My observations are that the three areas identified are possibly one of the best locations to set franchises for a mid-range car. I can corroborate the results with my practical experience that I have garnered been living there since decades. All these 3 locations experience maximum footfall from the section of the society who can conveniently afford a mid-range car. Franchises towards South Mumbai or North Mumbai would not have been a good choice. Reason being, large section of South Mumbai opts for expensive or high-priced car. Similarly, majority in North Mumbai prefers low budget cars.

So, my recommendation is that the organisation can search for an affordable commercial place around these locations to attract the right set of customers.

1. **Conclusions and Future Directions**

To conclude, reaching to a decision for the most viable spots for setting up franchises and unavailability of demographics data can be covered up with real estate prices of the city. Second, if the rental rates are in highly positive correlation with real estate rates then we can simply consider the property rates into consideration only. The features mentioned above can be enough to reach to a conclusion.

If in other times, there is a need to identify sites for another product or services, then a simple change in the criteria and assumptions - which makes the first part of the project - can be enough to reach to a feasible conclusion. There is no need to make any significant changes in the model. As long as the end result is to identify venues, with a little tweaking this project will take care of the stakeholders’ interests easily.