Project Report

Final Project Report for the IBM Professional Data Science certification Capstone Project

On Coursera

Rajvardhan Dixit

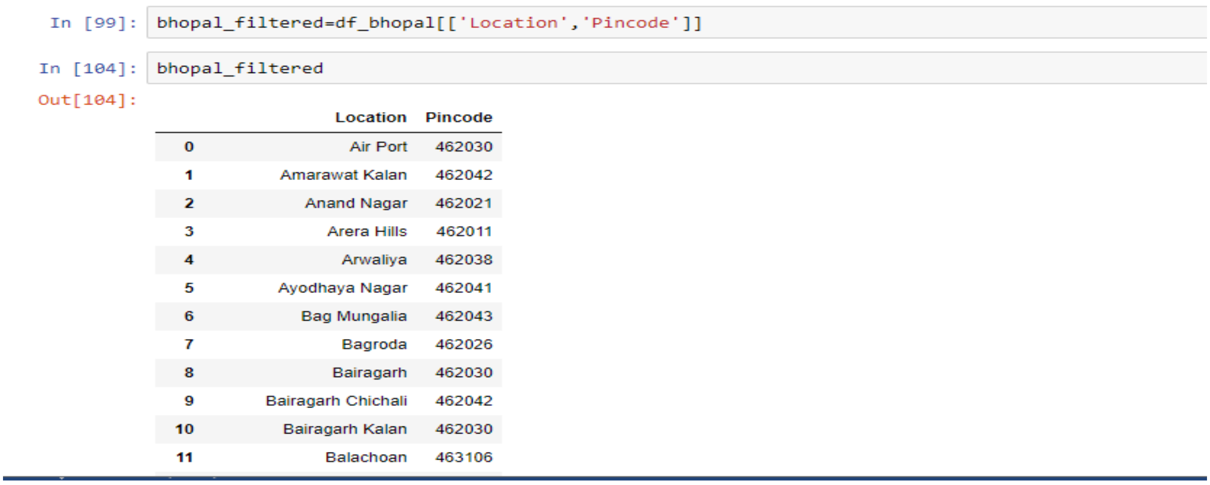
rdixit810@gmail.com

Introduction/ Business Problem: In this project I am going to create a dataset by using various websites, of Bhopal- Capital city of Madhya Pradesh and try to suggest best locations to invest in real estate as development of any part of city or land prices depends on the various utilities in the area like medical facilities, Public transportation, Education and food outlets. These things make life in the city easy so if a place have more utilities then the property price at that area will also become high.

Data Used: Dataset that can be used in this project is not available completely on internet so I created it using various websites and APIs. I took the pin code and Location names of Bhopal from Indian Post websites then I used geopy library to find out the Coordinates (Latitude and Longitude) of all the pin codes.

After that I used Foursquare API to create one more data frame about the three queries (medical, transport and food) one can add more number of queries to get more accurate results. Then add the two data frame and remove duplicate columns to get the perfect data frame for further use.

Indian Post Dataset:

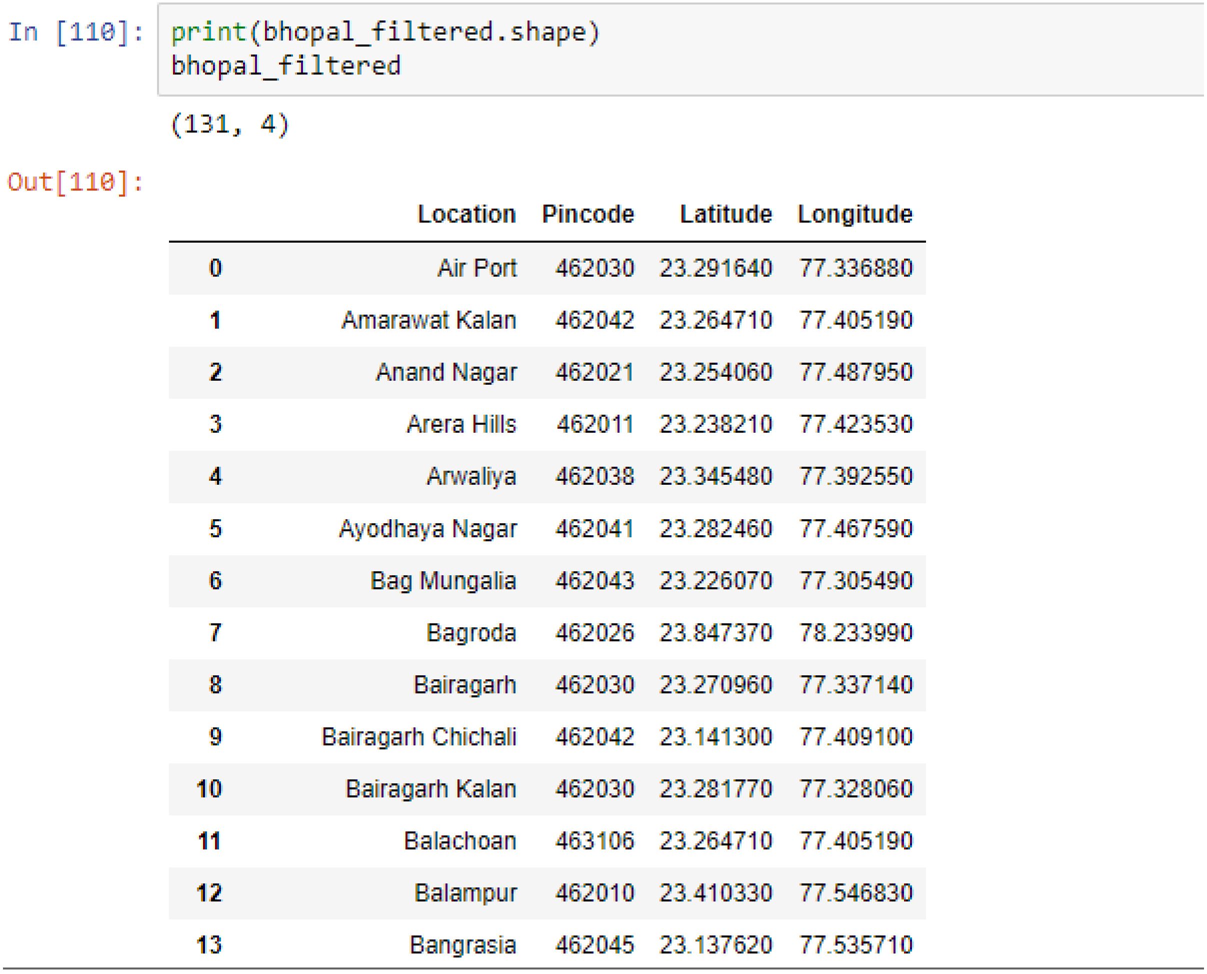


```
In [99]: bhopal_filtered=df_bhopal[['Location','Pincode']]

In [104]: bhopal_filtered
Out[104]:
```

	Location	Pincode
0	Air Port	462030
1	Amarawat Kalan	462042
2	Anand Nagar	462021
3	Arera Hills	462011
4	Arwaliya	462038
5	Ayodhaya Nagar	462041
6	Bag Mungalia	462043
7	Bagroda	462026
8	Bairagarh	462030
9	Bairagarh Chichali	462042
10	Bairagarh Kalan	462030
11	Balachoan	463106

Coordinates Dataset:



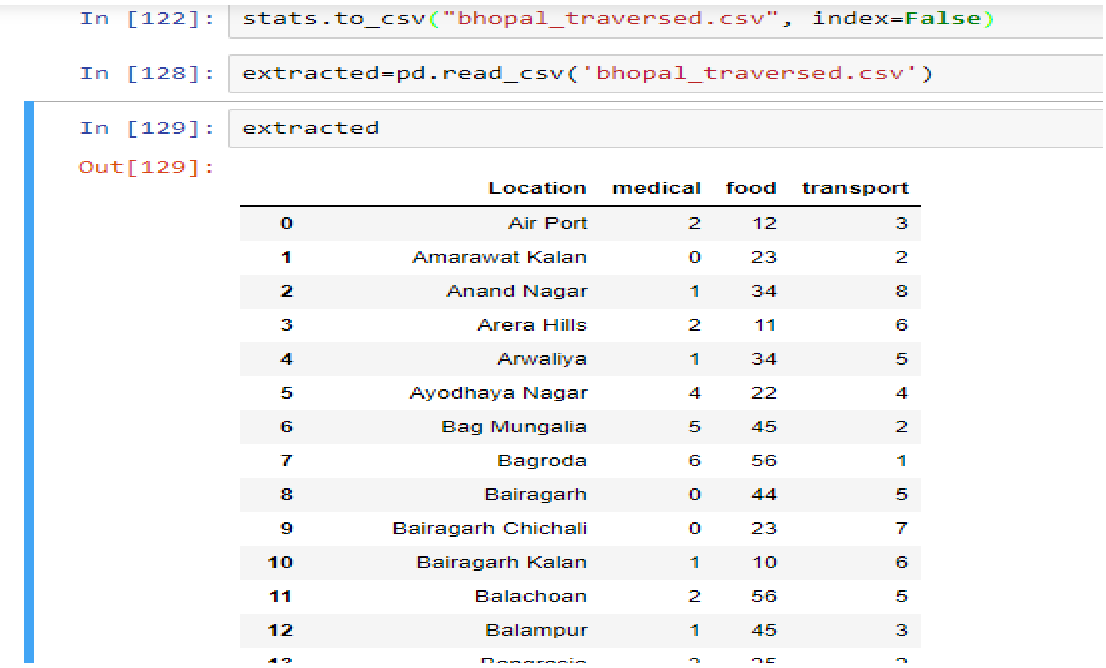
In [110]:
```python
print(bhopal_filtered.shape)
bhopal_filtered
```

```
(131, 4)
```

Out[110]:

	Location	Pincode	Latitude	Longitude
0	Air Port	462030	23.291640	77.336880
1	Amarawat Kalan	462042	23.264710	77.405190
2	Anand Nagar	462021	23.254060	77.487950
3	Arera Hills	462011	23.238210	77.423530
4	Arwaliya	462038	23.345480	77.392550
5	Ayodhaya Nagar	462041	23.282460	77.467590
6	Bag Mungalia	462043	23.226070	77.305490
7	Bagroda	462026	23.847370	78.233990
8	Bairagarh	462030	23.270960	77.337140
9	Bairagarh Chichali	462042	23.141300	77.409100
10	Bairagarh Kalan	462030	23.281770	77.328060
11	Balachoan	463106	23.264710	77.405190
12	Balampur	462010	23.410330	77.546830
13	Bangrasia	462045	23.137620	77.535710

Services dataset:



In [122]:
```python
stats.to_csv("bhopal_traversed.csv", index=False)
```

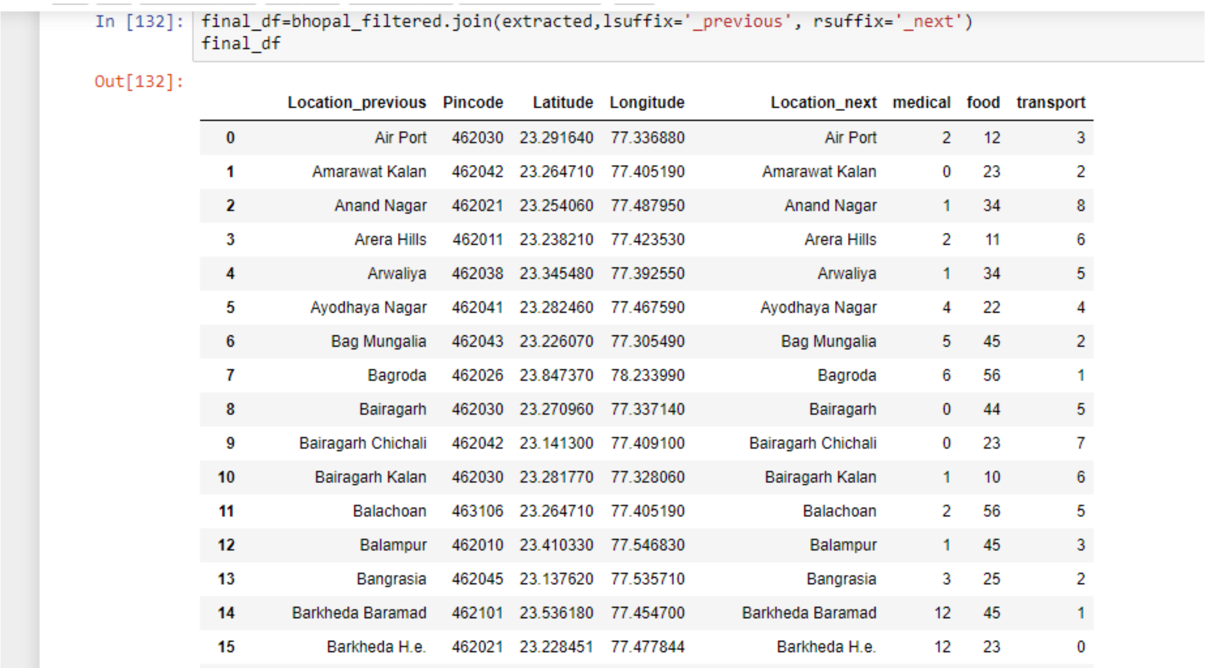
In [128]:
```python
extracted=pd.read_csv('bhopal_traversed.csv')
```

In [129]:
```python
extracted
```

Out[129]:

	Location	medical	food	transport
0	Air Port	2	12	3
1	Amarawat Kalan	0	23	2
2	Anand Nagar	1	34	8
3	Arera Hills	2	11	6
4	Arwaliya	1	34	5
5	Ayodhaya Nagar	4	22	4
6	Bag Mungalia	5	45	2
7	Bagroda	6	56	1
8	Bairagarh	0	44	5
9	Bairagarh Chichali	0	23	7
10	Bairagarh Kalan	1	10	6
11	Balachoan	2	56	5
12	Balampur	1	45	3
13	Bangrasia	[illegible]	[illegible]	[illegible]

Final dataset created for analysis:



```
final_df=bhopal_filtered.join(extracted,lsuffix='_previous', rsuffix='_next')
final_df
```

	Location_previous	Pincode	Latitude	Longitude	Location_next	medical	food	transport
0	Air Port	462030	23.291640	77.336880	Air Port	2	12	3
1	Amarawat Kalan	462042	23.264710	77.405190	Amarawat Kalan	0	23	2
2	Anand Nagar	462021	23.254060	77.487950	Anand Nagar	1	34	8
3	Arera Hills	462011	23.238210	77.423530	Arera Hills	2	11	6
4	Arwaliya	462038	23.345480	77.392550	Arwaliya	1	34	5
5	Ayodhaya Nagar	462041	23.282460	77.467590	Ayodhaya Nagar	4	22	4
6	Bag Mungalia	462043	23.226070	77.305490	Bag Mungalia	5	45	2
7	Bagroda	462026	23.847370	78.233990	Bagroda	6	56	1
8	Bairagarh	462030	23.270960	77.337140	Bairagarh	0	44	5
9	Bairagarh Chichali	462042	23.141300	77.409100	Bairagarh Chichali	0	23	7
10	Bairagarh Kalan	462030	23.281770	77.328060	Bairagarh Kalan	1	10	6
11	Balachoan	463106	23.264710	77.405190	Balachoan	2	56	5
12	Balampur	462010	23.410330	77.546830	Balampur	1	45	3
13	Bangrasia	462045	23.137620	77.535710	Bangrasia	3	25	2
14	Barkheda Baramad	462101	23.536180	77.454700	Barkheda Baramad	12	45	1
15	Barkheda H.e.	462021	23.228451	77.477844	Barkheda H.e.	12	23	0

Methodology: In this project I used Location data of various localities of Bhopal City from Indian post website and created a data frame using pandas then by using the pin codes of the localities I extracted the coordinates (Longitude and latitude) of all those coordinates using geopy library and make a separate data frame and then I joined the two data frame by using the join function.

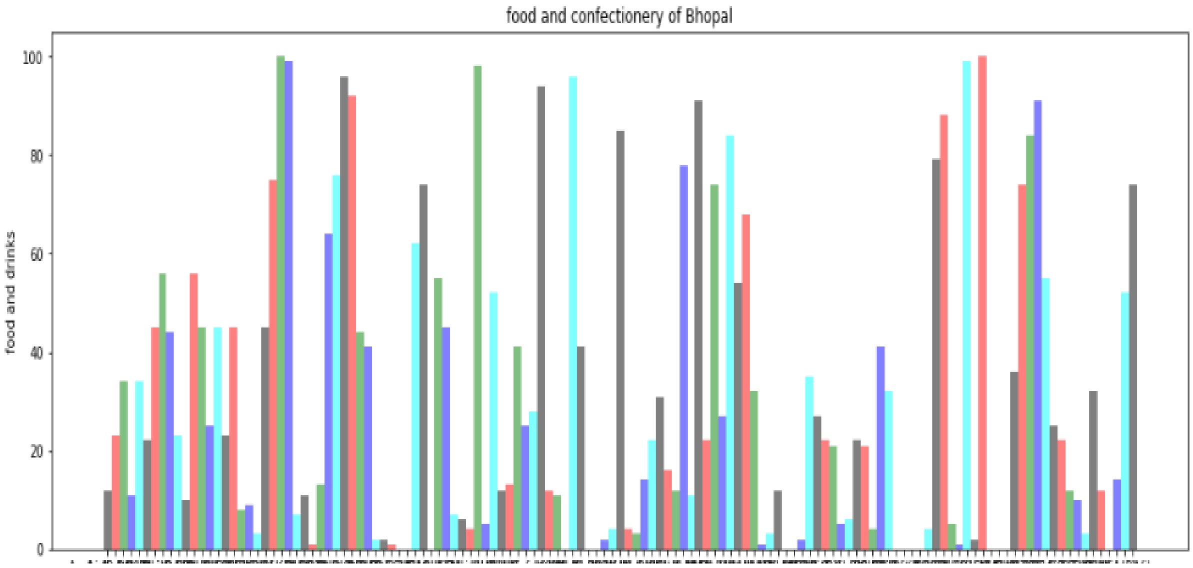
Now, by using the coordinates and Foursquare API, I extracted the number of medical centres, public transport and food and drinks outlet in the 5000ft radius of all the pin codes. I plot them as bar plots to easily understand the dispersion of data on various pin codes and also plot all the pin codes on the Bhopal city map using Folium Library.

This suggest that the pin codes at the centres and at some of the posh locations of the city have more facilities then the others.

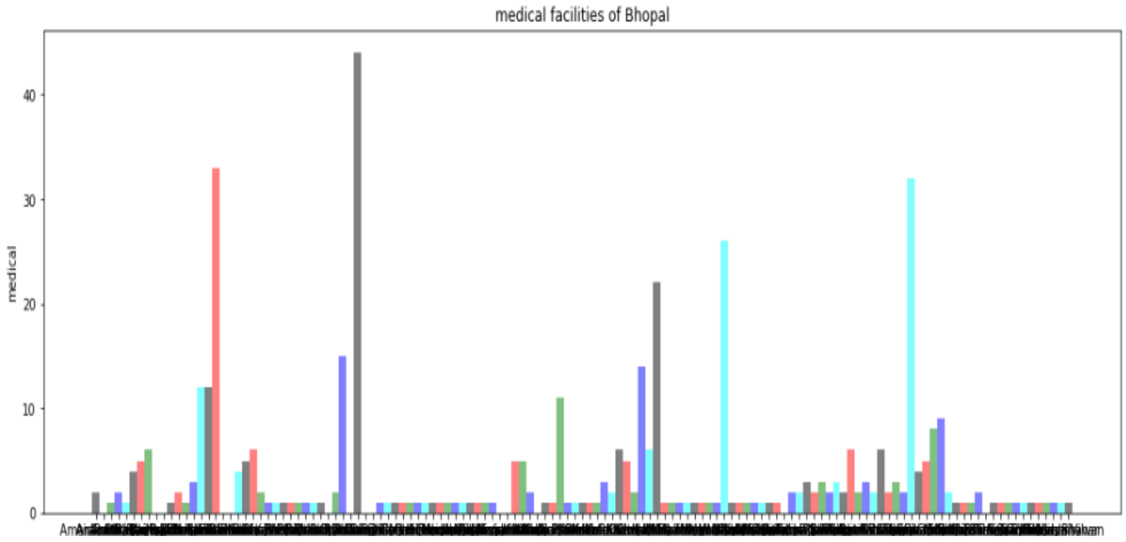
Then I used K means clustering algorithms and then make 4 clusters or categories from the dataset that I created and then place each pin code in any one out of four categories and then plot them on the Bhopal city map to check the property prices in the city. Category 1 Localities have least price while Category 4 in Cyan Colour have most expensive properties because of high level of public utilities availability.

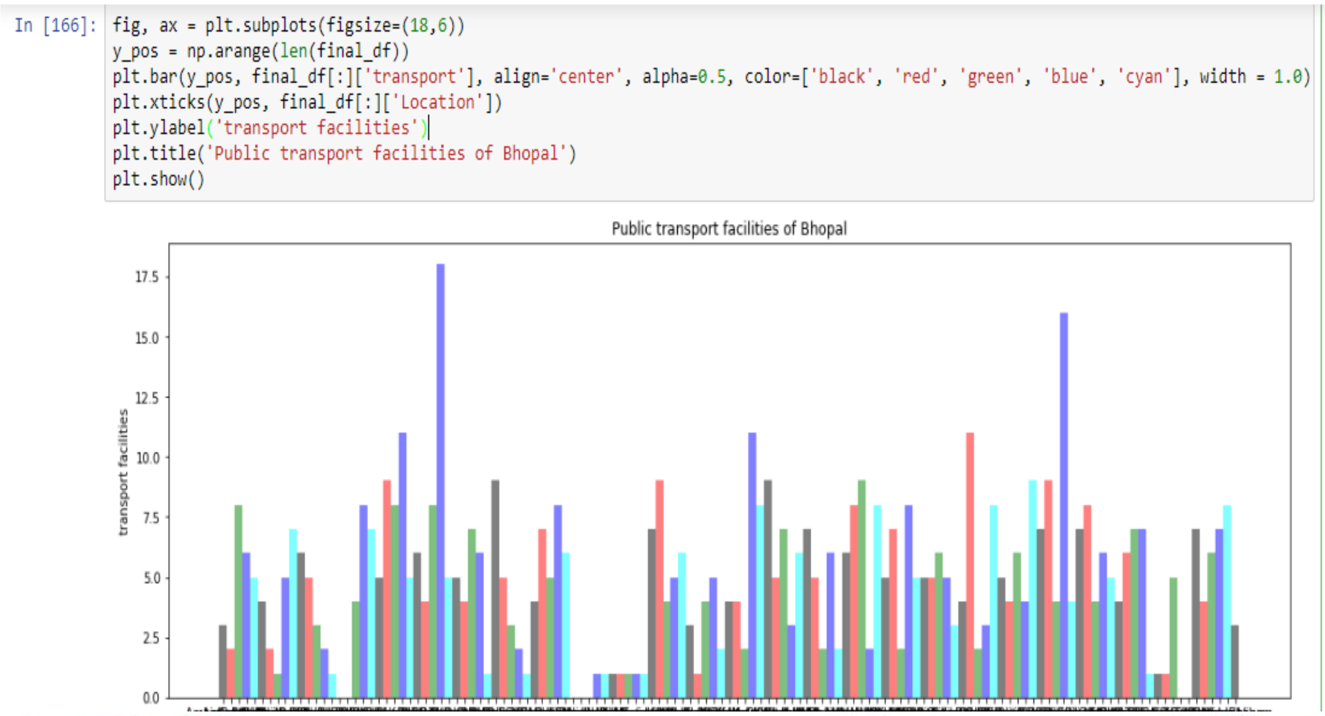
Some Visuals that was plotted:

```
In [168]: fig, ax = plt.subplots(figsize=(18,6))
          y_pos = np.arange(len(final_df))
          plt.bar(y_pos, final_df[:]['food'], align='center', alpha=0.5, color=['black', 'red', 'green', 'blue', 'cyan'], width = 1.0)
          plt.xticks(y_pos, final_df[:]['Location'])
          plt.ylabel('food and drinks')
          plt.title('food and confectionery of Bhopal')
          plt.show()
```



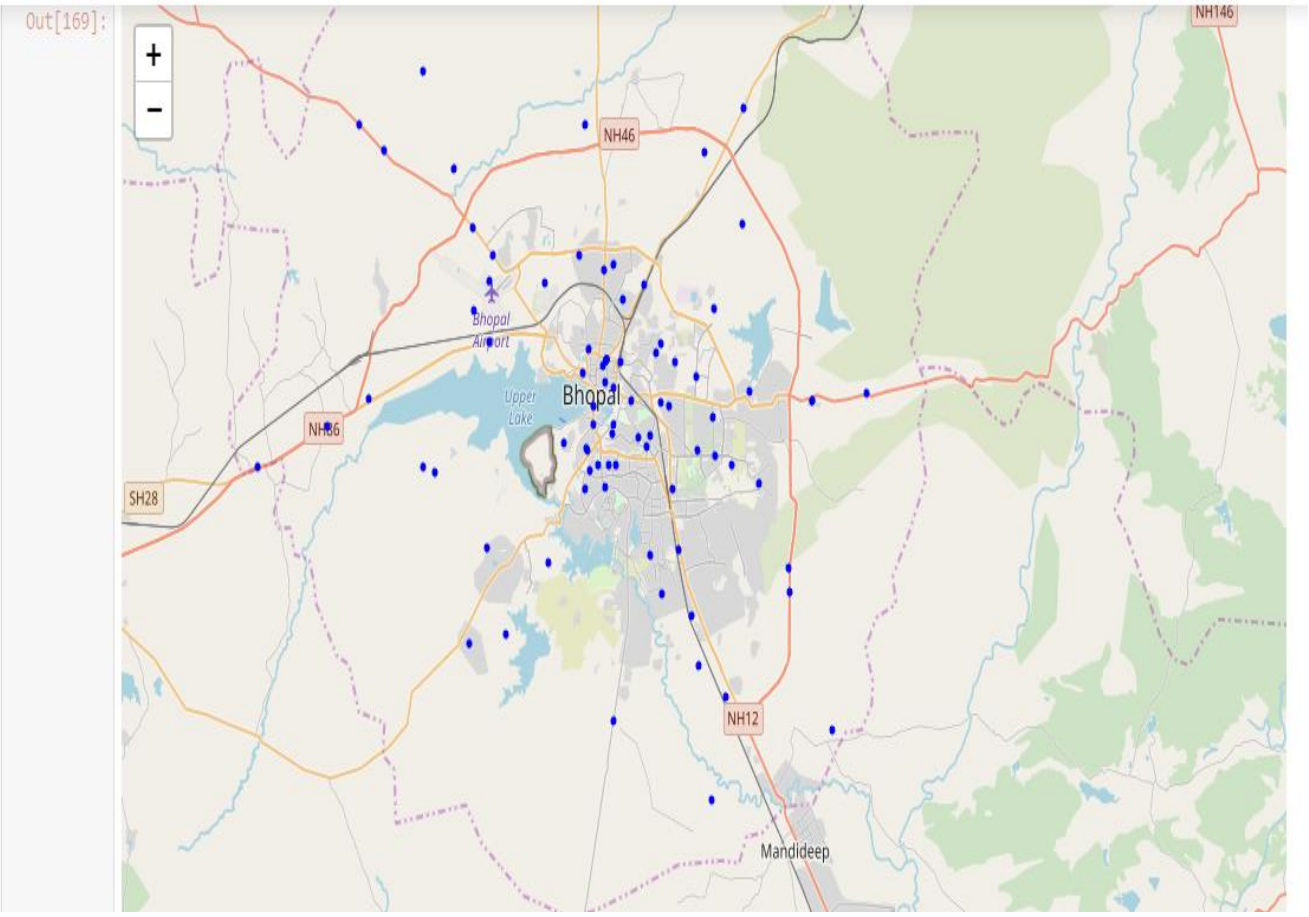
```
In [167]: fig, ax = plt.subplots(figsize=(18,6))
          y_pos = np.arange(len(final_df))
          plt.bar(y_pos, final_df[:]['medical'], align='center', alpha=0.5, color=['black', 'red', 'green', 'blue', 'cyan'], width = 1.0)
          plt.xticks(y_pos, final_df[:]['Location'])
          plt.ylabel('medical')
          plt.title('medical facilities of Bhopal')
          plt.show()
```





In [166]:

```
fig, ax = plt.subplots(figsize=(18,6))
y_pos = np.arange(len(final_df))
plt.bar(y_pos, final_df[:]['transport'], align='center', alpha=0.5, color=['black', 'red', 'green', 'blue', 'cyan'], width = 1.0)
plt.xticks(y_pos, final_df[:]['Location'])
plt.ylabel('transport facilities')
plt.title('Public transport facilities of Bhopal')
plt.show()
```



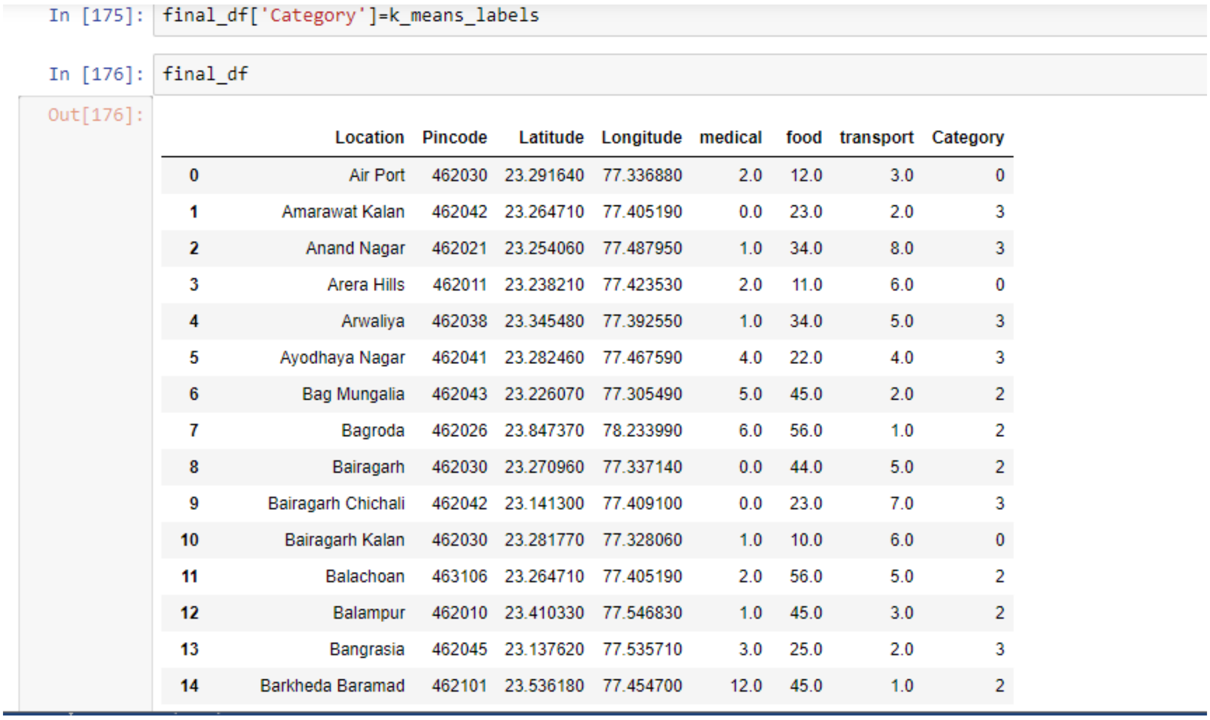
Out[169]:

Results:

At last we got a map and a dataset clustered in four categories and depicted on City map that shows the property prices using four different colours.

Here Category 1 have least property price and Category 4 have highest property price.

Categorised Dataset:



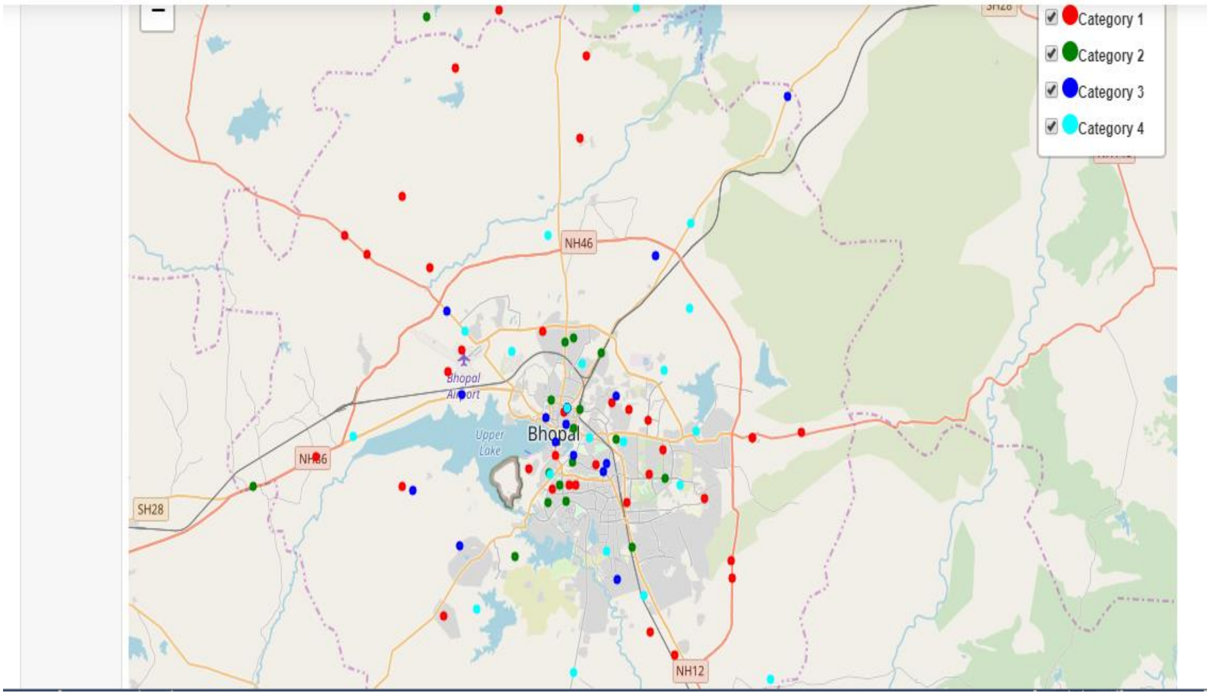
```
In [175]: final_df['Category']=k_means_labels
```

```
In [176]: final_df
```

Out[176]:

	Location	Pincode	Latitude	Longitude	medical	food	transport	Category
0	Air Port	462030	23.291640	77.336880	2.0	12.0	3.0	0
1	Amarawat Kalan	462042	23.264710	77.405190	0.0	23.0	2.0	3
2	Anand Nagar	462021	23.254060	77.487950	1.0	34.0	8.0	3
3	Arera Hills	462011	23.238210	77.423530	2.0	11.0	6.0	0
4	Arwaliya	462038	23.345480	77.392550	1.0	34.0	5.0	3
5	Ayodhaya Nagar	462041	23.282460	77.467590	4.0	22.0	4.0	3
6	Bag Mungalia	462043	23.226070	77.305490	5.0	45.0	2.0	2
7	Bagroda	462026	23.847370	78.233990	6.0	56.0	1.0	2
8	Bairagarh	462030	23.270960	77.337140	0.0	44.0	5.0	2
9	Bairagarh Chichali	462042	23.141300	77.409100	0.0	23.0	7.0	3
10	Bairagarh Kalan	462030	23.281770	77.328060	1.0	10.0	6.0	0
11	Balachoan	463106	23.264710	77.405190	2.0	56.0	5.0	2
12	Balampur	462010	23.410330	77.546830	1.0	45.0	3.0	2
13	Bangrasia	462045	23.137620	77.535710	3.0	25.0	2.0	3
14	Barkheda Baramad	462101	23.536180	77.454700	12.0	45.0	1.0	2

Depiction: on map:



Discussion:

The results are accurate to a certain extent but as we know the prediction is very complicated process and needs to consider a lot of things then we consider in the application, however the red dots shows less price of properties and cyan dots shows high value of land and property in the area.

Conclusion:

IN this project I created a application for the real estate developers, city planners, students of architecture and planning to understand the development of city. The goals of the project can be satisfied with this but to develop the application to the market standards needs a little more efforts as the prices cannot be predicted just by three queries some more queries about educational institutes, amusement parks, city heritage locations and mall and museum will make the application more accurate. By increasing the radius will also increase the number of facilities in a area and by doing this the application will be more better.