Applied Data Science Capstone Project
Exploring Sports Venues in Navi Mumbai, India

Vijayandra Mahadik

Table of Contents

Introduction

The aim of the project is to explore different variety of options to play different sports within Navi Mumbai and then group places based on similar sport options they provide.

The target audience is someone who is planning to shift to Navi Mumbai or planning to stay for long time period where they want their kids or themselves to engage in different sport activities for example if a person likes to be a member of sports club or is a professional player who plays Soccer, Golf, Volleyball or Basketball and wants to live in a place which is nearby saving travel time.

This project can also help business investors to locate potential areas where they can start a new sport venue based on the existing common venues within Navi Mumbai. For instance, If a given sport venue like Sports club or Tennis court is only situated in very few areas and is mostly booked all the time having players wait for their slots to come, there can pe a potential business opportunity to open a similar venue in a nearby area. Another potential business opportunity would be to create new venues near corporate companies to host their sport events.

Data

The railway stations from Navi Mumbai were considered as Neighborhood. The list of station names was taken from below link:

https://en.wikipedia.org/wiki/Trans-Harbour_line_(Mumbai_Suburban_Railway)

The coordinates of each stations were then derived using geopy library.

Methodology

The first step was to find coordinates of neighborhood in Navi Mumbai. The neighborhood dataset considered here consists of all the stations from Trans-harbour railway line of Navi Mumbai (excluding Thane and Panvel as they are separate cities). There are in total 13 stations on this line which are part of Navi Mumbai.

As the coordinates of all these stations were not readily available at one site or readily available for download, the dataset was created by iteratively passing station names to geopy library. As there are multiple places with name Mansarovar hence 'Mansarovar, Navi Mumbai' was used to find its coordinates.

	Neighborhood	Latitude	Longitude
0	Airoli	19.158272	72.996709
1	Rabale	19.136637	73.002781
2	Ghansoli	19.119331	72.999510
3	Kopar Khairane	19.102852	73.003075
4	Turbhe	19.076165	73.017662
5	Sanpada	19.065977	73.009533
6	Vashi	19.075713	73.000354
7	Juinagar	19.056169	73.018245
8	Nerul	19.033612	73.018140
9	Seawoods-Darave	19.022192	73.018738
10	CBD Belapur	19.018987	73.039095
11	Kharghar	19.025773	73.059185
12	Mansarovar	19.016434	73.080655

Figure 1: Neighborhood dataset

In the next step, Foursquare API was used to get sport venues using “Athletics & Sports” search category ID. 250 Venues were searched within a radius of 1.5 km of each station.

Some data clean-up was done after API call.

- As the stations are close to each other, some venues got picked for two different stations creating duplicates. The first match was kept, and others were dropped. This way duplicates were removed.
- One of the stations “Khandeshwar” had no sport venue hence removed it.
- Many Gym venues with different names such as 'Gym', 'Gym / Fitness Center', 'Gym Pool', 'Boxing Gym' were grouped into a single venue ‘Gym’.

Final data set had 13 stations with 16 unique venue categories.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
91	Vashi	19.075713	73.000354	Sector 9 Ground	19.080126	72.993422	Garden
92	Vashi	19.075713	73.000354	Talwalkars Fitness Centre, Vashi	19.073968	72.991247	Gym
93	Vashi	19.075713	73.000354	NMSA Gymnasium	19.072205	72.992036	Gym
97	Vashi	19.075713	73.000354	Fr.Agnels Basketball Court	19.077486	72.993571	Basketball Court
98	Vashi	19.075713	73.000354	father agnel Football Ground	19.077786	72.993859	Soccer Field
99	Vashi	19.075713	73.000354	nmsa basketball court	19.071135	72.992470	Basketball Court

Figure 2: Sample data of Vashi station which has different venues

For visual understanding, using Folium library a map of Navi Mumbai with neighborhood stations superimposed on top was created.

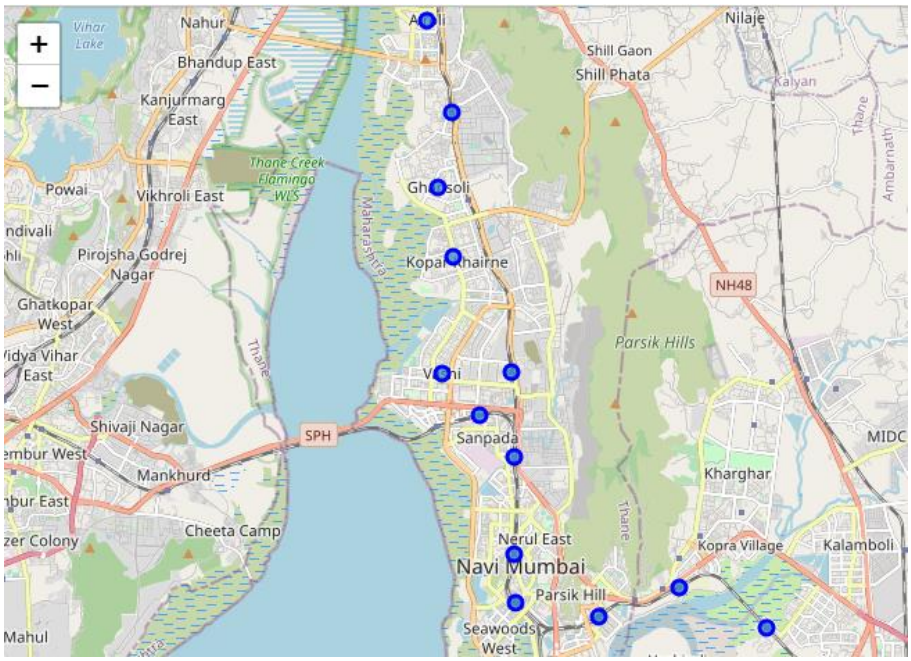

Figure 3: Navi Mumbai map

All the categorical values were converted into values using one-hot encoding and then their mean of frequency of occurrence of each category was taken.

Neighborhood	Athletics & Sports	Baseball Field	Basketball Court	Garden	Golf Course	Gym	Hotel	Martial Arts Dojo	Residential Building (Apartment / Condo)	Soccer Field	Sporting Goods Shop	Sports Club	Tennis Court
Airoli	0.133333	0.000000	0.000000	0.00	0.000000	0.666667	0.000000	0.000000	0.000000	0.000000	0.000	0.200000	0.000000
CBD Belapur	0.250000	0.000000	0.000000	0.00	0.000000	0.625000	0.000000	0.000000	0.000000	0.000000	0.125	0.000000	0.000000
Ghansoli	0.000000	0.000000	0.000000	0.00	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000
Juinagar	0.250000	0.000000	0.000000	0.00	0.000000	0.500000	0.000000	0.000000	0.000000	0.250000	0.000	0.000000	0.000000
Kharghar	0.111111	0.000000	0.000000	0.00	0.000000	0.777778	0.000000	0.000000	0.000000	0.111111	0.000	0.000000	0.000000
Kopar Khairane	0.000000	0.000000	0.000000	0.00	0.000000	0.833333	0.000000	0.166667	0.000000	0.000000	0.000	0.000000	0.000000
Mansarovar	0.000000	0.000000	0.250000	0.00	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000
Nerul	0.173913	0.000000	0.000000	0.00	0.000000	0.695652	0.000000	0.000000	0.000000	0.000000	0.000	0.086957	0.000000
Rabale	0.000000	0.000000	0.333333	0.00	0.000000	0.666667	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000
Sanpada	0.173913	0.043478	0.043478	0.00	0.043478	0.521739	0.043478	0.000000	0.043478	0.086957	0.000	0.000000	0.000000
Seawoods-Darave	0.000000	0.000000	0.000000	0.00	0.000000	0.777778	0.000000	0.000000	0.000000	0.000000	0.000	0.111111	0.111111
Turbhe	0.117647	0.058824	0.058824	0.00	0.058824	0.529412	0.058824	0.000000	0.058824	0.058824	0.000	0.000000	0.000000
Vashi	0.160000	0.000000	0.120000	0.04	0.040000	0.440000	0.000000	0.000000	0.000000	0.120000	0.000	0.040000	0.000000

Figure 4: Mean of frequency of occurrence

Top 5 most common venues across each neighborhood were plotted to get a fair idea of how the clustering would look like.

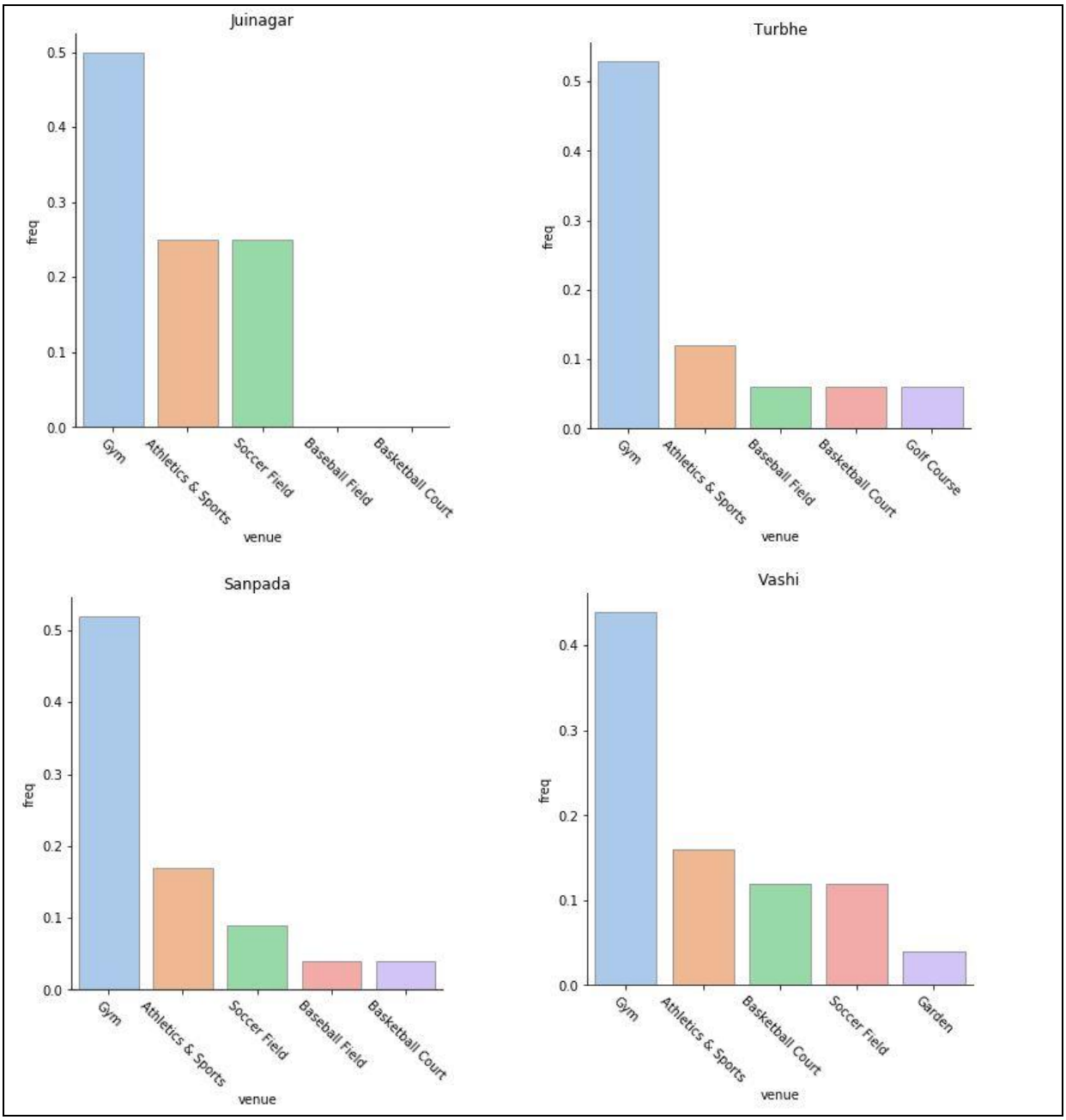

Figure 5: Top 5 common venues in Juinagar, Turbhe, Sanpada and Vashi

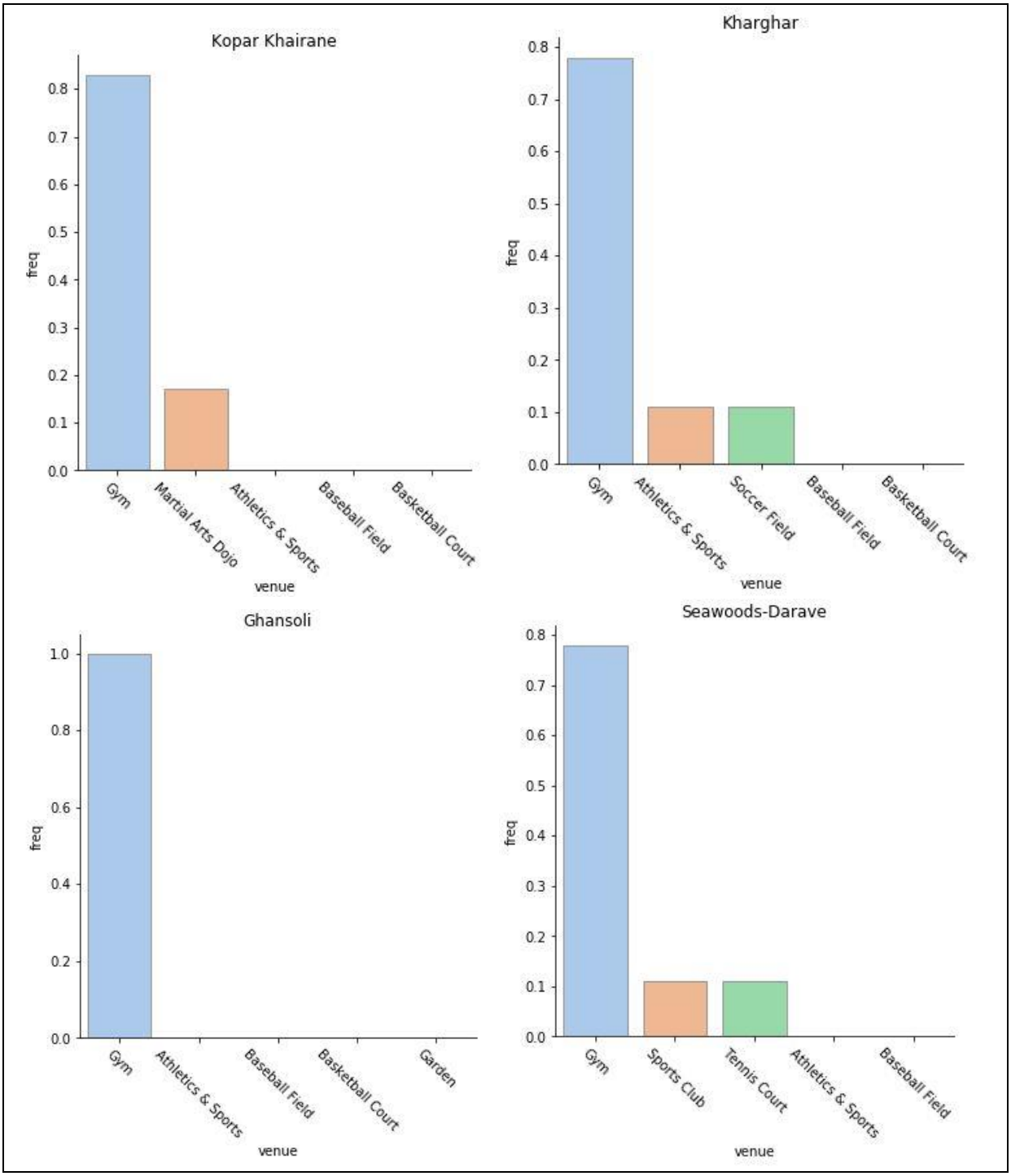

Figure 6: Top 5 common venues in Kopar Khairane, Khargar, Ghansoli and Seawoods

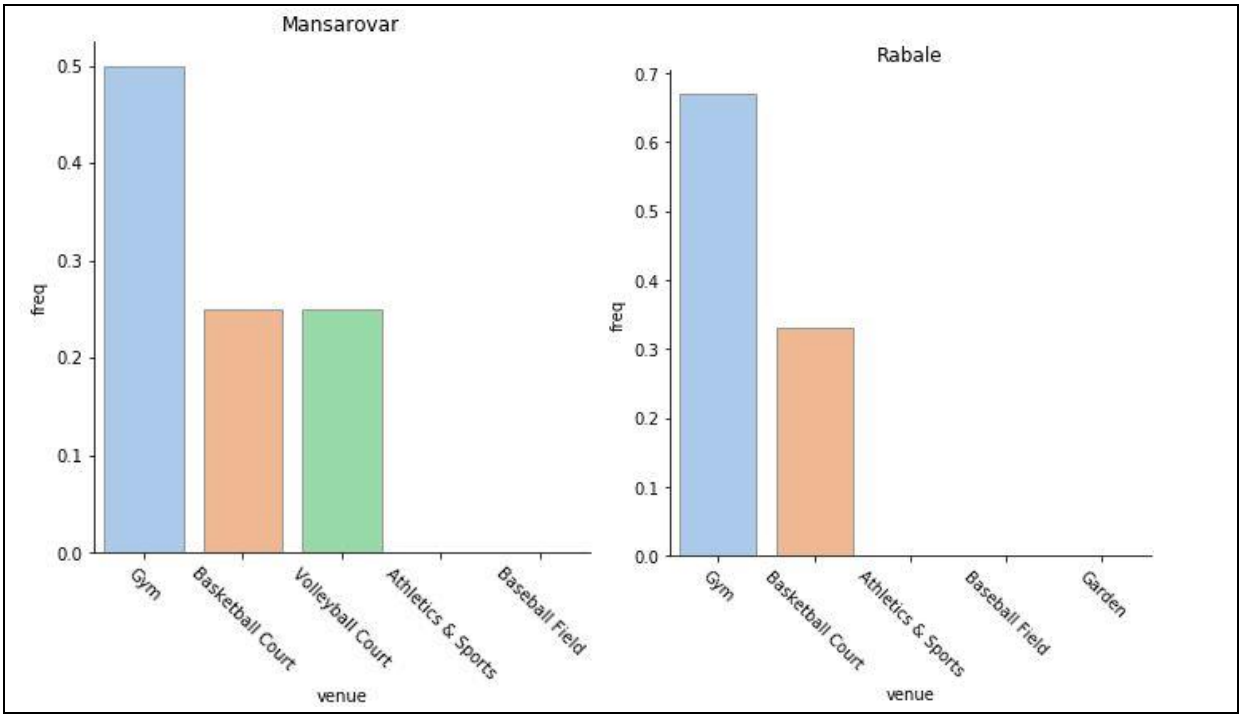

Figure 7: Top 5 common venues in Mansarovar and Rabale

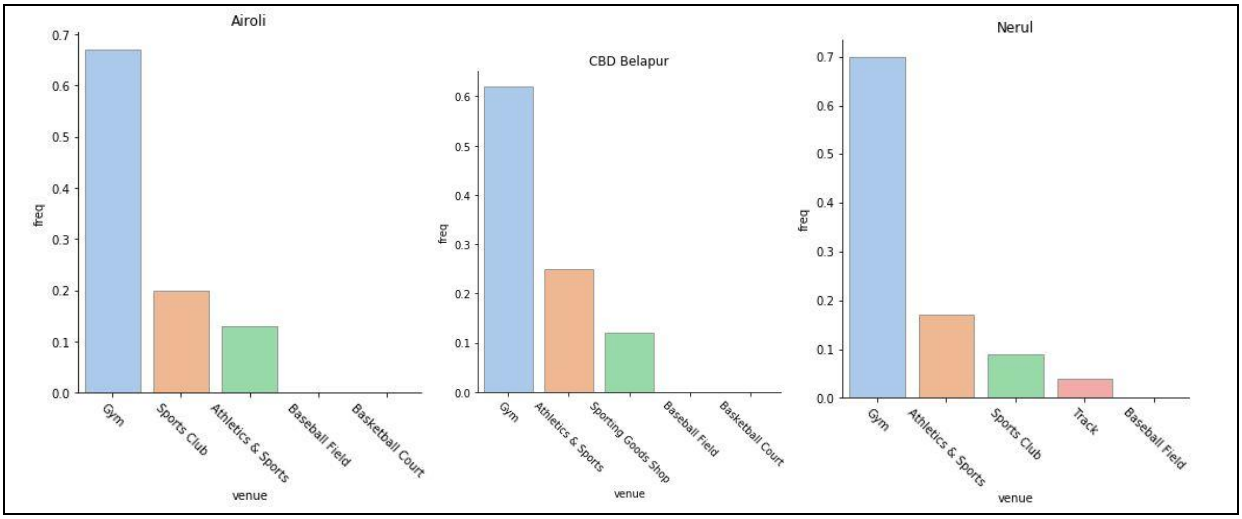

Figure 8: Top 5 common venues in Airoli, CBD Belapur and Nerul

	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	Airoli	Gym	Sports Club	Athletics & Sports	Yoga Studio	Volleyball Court	Track	Tennis Court	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)
1	CBD Belapur	Gym	Athletics & Sports	Sporting Goods Shop	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Soccer Field	Residential Building (Apartment / Condo)
2	Ghansoli	Gym	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)	Martial Arts Dojo
3	Juinagar	Gym	Soccer Field	Athletics & Sports	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Residential Building (Apartment / Condo)
4	Kharghar	Gym	Soccer Field	Athletics & Sports	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Residential Building (Apartment / Condo)

Figure 9: Top 10 most common venues for each neighborhood

K-means clustering was used on top of this data to create K similar clusters. As there were only 13 neigborhoods, lower K value of the range 2-5 was tried. On an iterative run with different K value, K=4 gave better results. After running K-means clustering with K=4, cluster labels ranging from 0-3 were generated which was then merged with top 10 most common venue data.

Neighborhood	Latitude	Longitude	Cluster Labels	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
Airoli	19.158272	72.996709	0	Gym	Sports Club	Athletics & Sports	Yoga Studio	Volleyball Court	Track	Tennis Court	Sporting Goods Shop	Soccer Field
Rabale	19.136637	73.002781	3	Gym	Basketball Court	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field
Ghansoli	19.119331	72.999510	2	Gym	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)
Kopar Khairane	19.102852	73.003075	2	Gym	Martial Arts Dojo	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field
Turbhe	19.076165	73.017662	1	Gym	Athletics & Sports	Soccer Field	Residential Building (Apartment / Condo)	Hotel	Golf Course	Basketball Court	Baseball Field	Yoga Studio

Figure 10: Top 10 most common venues for each neighborhood with cluster labels (0-3)

Results

Below map shows clusters of similar stations created from K-means clustering:

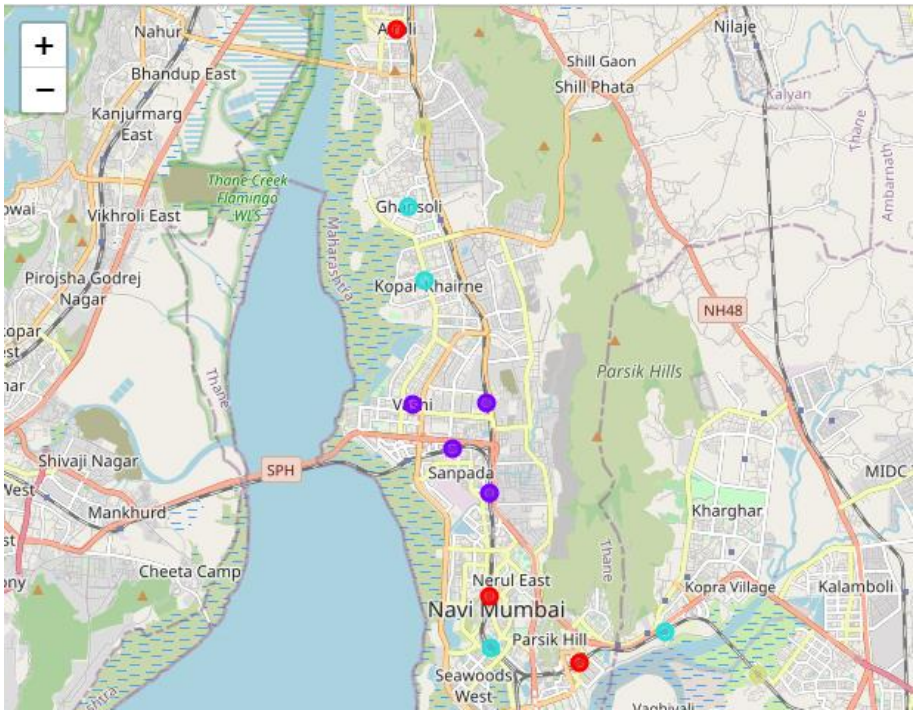

Figure 11: Result of Clustering Algorithm

4 clusters were created with below details:

- Cluster 1 (Red) – Consists 3 stations namely Airoli, Nerul, CBD Belapur
- Cluster 2 (Purple) - Consists 4 stations namely Turbhe, Sanpada, Vashi, Juinagar
- Cluster 3 (Cyan) - Consists 4 stations namely Ghansoli, Kopar Khairane, Seawoods-Darave,Khargar
- Cluster 4 (Pale Yellow) - Consists 2 stations namely Rabale, Mansarovar

Each cluster data was plotted and analysed. Right side of the plot created a list of all venues sorted with most occurrence at the top.

	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	Airoli	Gym	Sports Club	Athletics & Sports	Yoga Studio	Volleyball Court	Track	Tennis Court	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)
8	Nerul	Gym	Athletics & Sports	Sports Club	Track	Yoga Studio	Volleyball Court	Tennis Court	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)
10	CBD Belapur	Gym	Athletics & Sports	Sporting Goods Shop	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Soccer Field	Residential Building (Apartment / Condo)

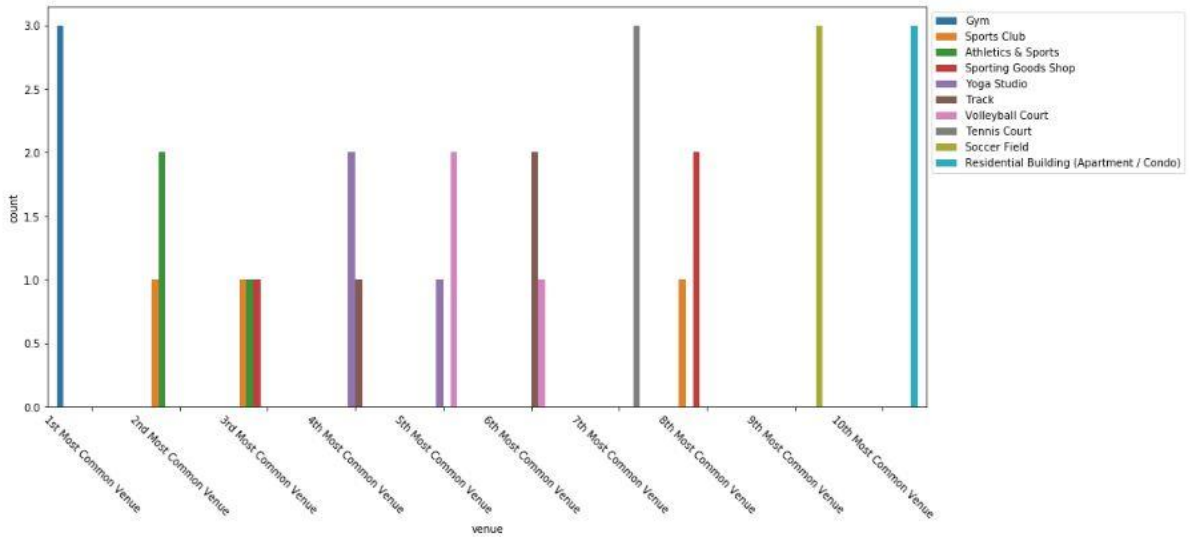

Figure 12: Cluster 1 – Gym and Sports Club

	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
4	Turbhe	Gym	Athletics & Sports	Soccer Field	Residential Building (Apartment / Condo)	Hotel	Golf Course	Basketball Court	Baseball Field	Yoga Studio	Volleyball Court
5	Sanpada	Gym	Athletics & Sports	Soccer Field	Residential Building (Apartment / Condo)	Hotel	Golf Course	Basketball Court	Baseball Field	Yoga Studio	Volleyball Court
6	Vashi	Gym	Athletics & Sports	Soccer Field	Basketball Court	Yoga Studio	Sports Club	Golf Course	Garden	Volleyball Court	Track
7	Juinagar	Gym	Soccer Field	Athletics & Sports	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Residential Building (Apartment / Condo)

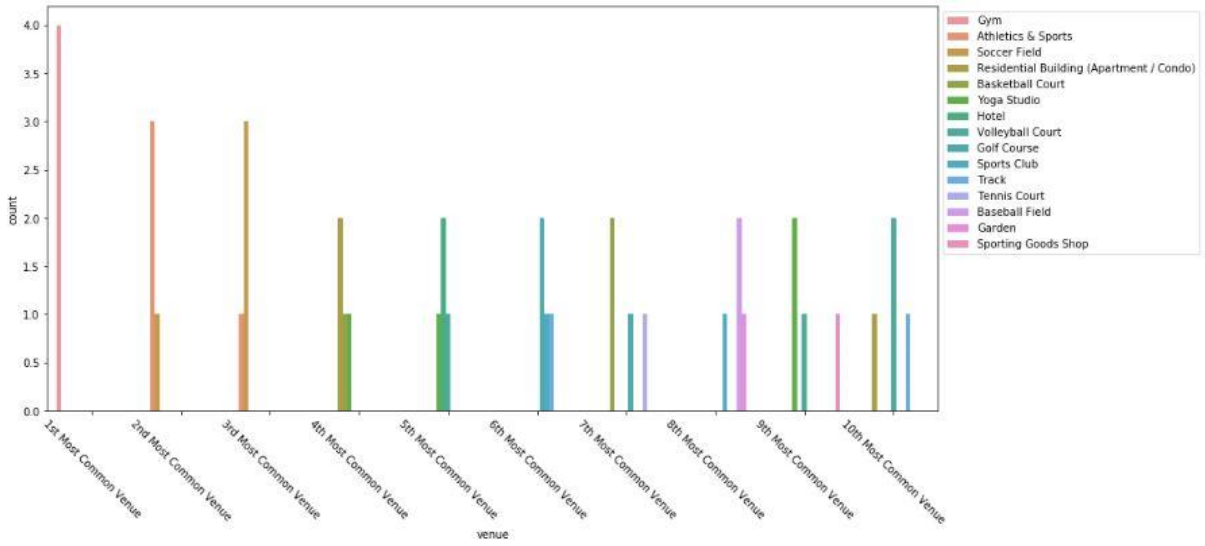

Figure 13: Cluster 2 - Soccer, Golf, Baseball

	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
2	Ghansoli	Gym	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)	Martial Arts Dojo
3	Kopar Khairane	Gym	Martial Arts Dojo	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)
9	Seawoods-Darave	Gym	Tennis Court	Sports Club	Yoga Studio	Volleyball Court	Track	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)	Martial Arts Dojo
11	Kharghar	Gym	Soccer Field	Athletics & Sports	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Residential Building (Apartment / Condo)

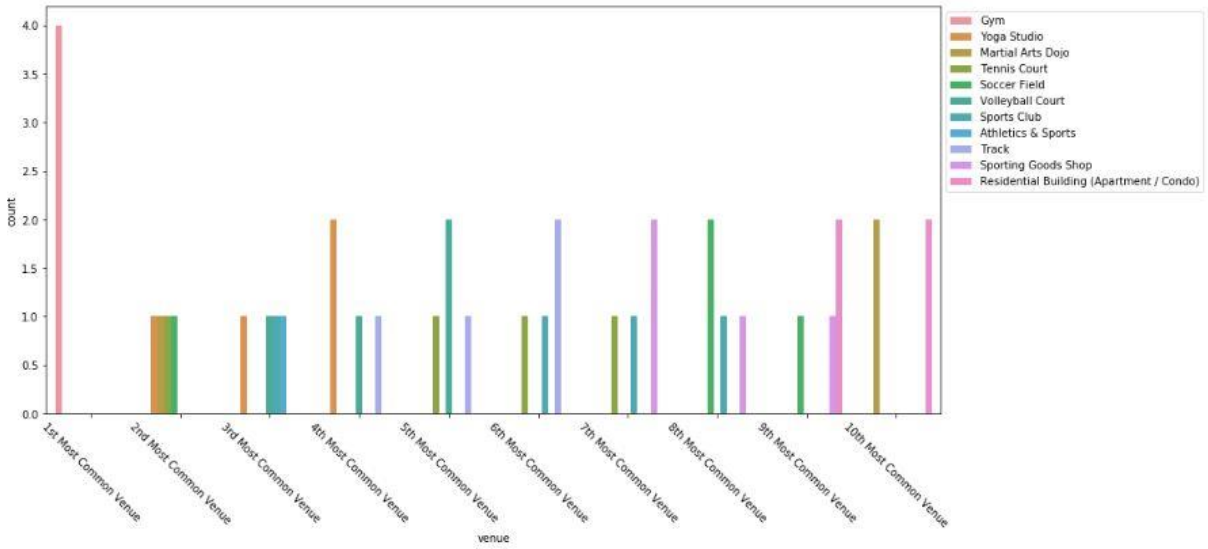

Figure 14: Cluster 3 - Yoga, Martial Arts, Tennis

	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
1	Rabale	Gym	Basketball Court	Yoga Studio	Volleyball Court	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)
12	Mansarovar	Gym	Volleyball Court	Basketball Court	Yoga Studio	Track	Tennis Court	Sports Club	Sporting Goods Shop	Soccer Field	Residential Building (Apartment / Condo)

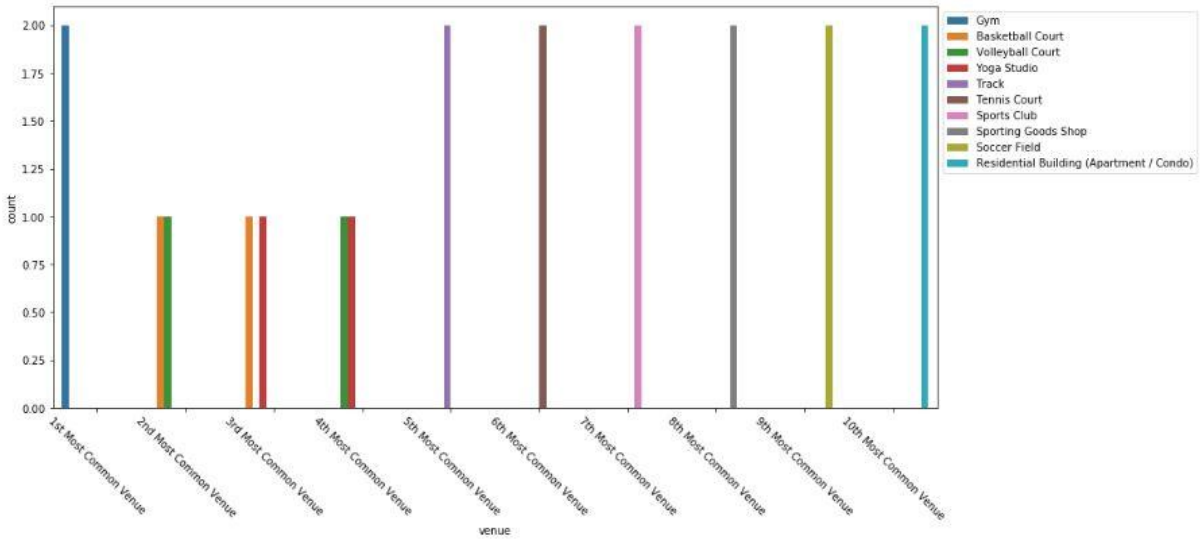

Figure 15: Cluster 4 - Basketball, Volleyball

From the above clustering below conclusion was made:

- All the stations had Gym as their first most common venue. No surprises, it is very common these days.
- Cluster 1 (Red) - 3 stations Airoli, Nerul, CBD Belapur; Sports Club
- Cluster 2 (Purple) - 4 stations Turbhe, Sanpada, Vashi, Juinagar; Open field games such as Soccer, Golf, Baseball
- Cluster 3 (Cyan) - 4 stations Ghansoli, Kopar Khairane, Seawoods-Darave,Khargar; Yoga, Martial Arts, Tennis
- Cluster 4 (Pale Yellow) - 2 stations Rabale, Mansarovar; Games played on courts such as Basketball, Volleyball

Discussion

Clearly the clustering provides a great categorization of the different sports being played. Some areas (Purple dots) have lot of open space as compared to others hence host open field games. Some areas (Red dots) has venue to play multiple sports at the same location which may include indoor games as well. If your kids or your partner plays different game than you then this is a great place for entire family to hang around and enjoy under one roof. Looking at this map a Sportsperson can decide where to stay depending on what sports he plays. This cannot be the sole reason to decide which place you live but can help make decisions.

Also, this would be a handy insight for a Businessman to start new venue. For instance, looking at our clusters, people living in Airoli don't have Golf course and the next nearest they have is in Turbhe region which is around 11km away. This shows a potential opportunity to open a Golf course in Airoli. But other factors should also need to be considered such as cost of setup, land availability and so on.

Conclusion

The clustering of sport venues can give you a fair idea of what options you got to play any game. This can help distinguish or find similar places which can help you decide which place you want to move or live or identify new business opportunities. However, the above insights will not be sufficient alone to take decision but if we combine additional data such as property cost, population of the area, average salary, ratings and cost of membership to these venues, how old these venues are, travel options etc then it will serve in better decision making.