STUDYING THE COVERAGE OF FOURSQUARE CROWDSOURCING DATA

for restaurants in the city of Mumbai

Introduction

- Foursquare is a service that provides crowdsourced data for various venues nearby your location. However, as a native of Mumbai, I found that Foursquare data for the area was lacking.
- This study aims to qualitatively understand the lack of data in this region. It will limit itself to the scope of restaurants and other eateries, and compare the Foursquare data with a relatively complete data obtained from Zomato.
- The results of this study can help quantify the coverage difference between Foursquare, a US based service and Zomato, an India based service.

Data Sourcing

For this study, we made use of the following data:

- Mumbai Data
  - Post offices and Pin Codes: This Data encompasses a list of pin codes, local post office names and the governing districts. It will provide a way to segregate localities on basis of a definite locations
    We will limit our scope to Mumbai only.
    Source: https://mumbai7.com/postal-codes-in-mumbai/
- Restaurant Data
  - Foursquare API:
    API requests to Foursquare provided the currently documented restaurants and eateries in the vicinity of the post offices.
  - Zomato API:
    API requests to Zomato provided the currently documented restaurants and eateries. This data was be considered more accurate and used to compare the Foursquare data

Locality

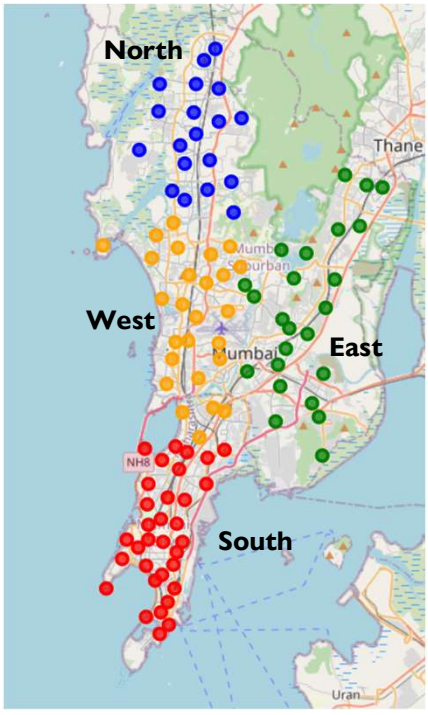
- In order to efficiently capture our data, we use the different localities of Mumbai as starting points for API calling.
- These localities are not however, formal. There boundaries are blurred, with multiple names for the same localities.
- As a starting point, we used the different Pin Code areas as a reference for a list of localities since the area they cover is appropriate for our purpose. Wherever a Pin Code area was too large, we used appropriately chosen additional localities to supplement our list.
- The Geo Coordinates were then obtained using Nominatim. Any Locality name that did not return any value or an incorrect location was replaced by an alternate name.

Sectors

Mumbai is often informally divided into four major areas:

1. **South Mumbai**, majority of which falls in Mumbai City District. It is equivalent to the downtown region of a major city. It is rich in commercial, industrial, and tourist areas as compared to the suburbs.
2. **North Mumbai**, which was historically a quiet suburb far away from the bustling of the Mumbai City. Recently it has seen a rising growth, with more people coming in as the city slowly expands.
3. **Eastern Mumbai**, an industrial and semi-suburban region that developed along the Central and Harbour Lines of the Mumbai Suburban Railway. Continuing up these railway lines we reach into Thane and Navi Mumbai.
4. **Western Mumbai**, a semi-suburban region between South Mumbai and North Mumbai. It is now growing rapidly, as the commercial and industrial centres move outward.

Thus we can subdivide our localities into these sectors. These are indicative and the boundaries are subjective.

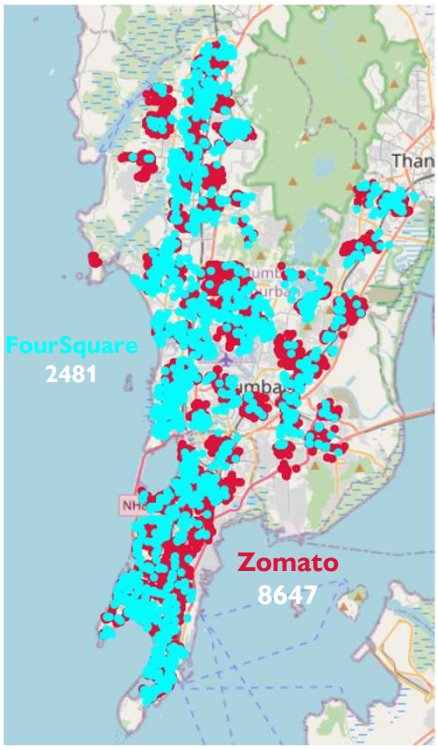


Methodology

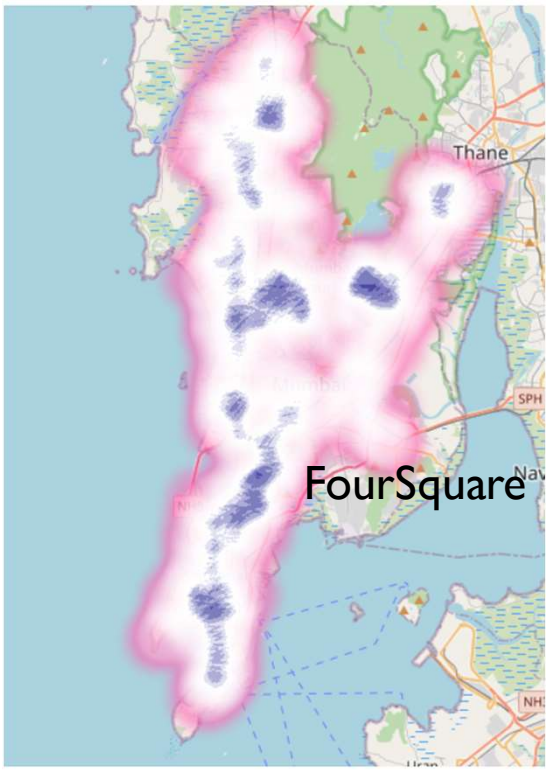
- In this project, we wish to see how many venues are being returned for a particular locality, as opposed to how many should be returned.
- In order to do this, we used restaurant data from Zomato as a baseline. The Zomato data accounts for >90% of all venues. However, this has some back draws:
  - Venue Coordinates are not accurate; there is a error margin of ~25m.
  - API returns only 100 restaurants for a query; the max value for a Locality is 100 venues
  - API search query has a bug w.r.t 'radius'; the venues are not filtered by the radius argument
- The limitation that derives from this is that we cannot effectively match the two datasets, as the names of venues also do not match. Since we are only concerned with the coverage, it can be obtained by alternate means
- For the first part, we collected the required data: Venues from FourSquare and Zomato around each Locality in radius of 1 kilometer.
- In the second part, we visualized the obtained data by exploring different visual tools:
  - We used Heat Maps to visualize the venue distribution for both FourSquare and Zomato venues.
  - We used superimposed Bubble Maps to:
    - visualize the contrast in venue counts for each locality.
    - visualize the top 20 localities w.r.t FourSquare venues
    - visualize the contrast in venue counts for each sector.
  - We used Cluster Maps to visualize the venue count clusters for FourSquare venues.
- This lead to a qualitative understanding between the contrast of Zomato and FourSquare data coverage.
- In the third and final part, we attempted to create a measure of the contrast of the data coverage for the two providers. This value, called *Coverage*, is the percentage of FourSquare Venues to Zomato Venues for a given Locality. Using it, we can guess the relative coverage across Localities.
- Finally we presented a Heat Map for the *Coverage* to visualize the Foursquare data for restaurants in Mumbai.

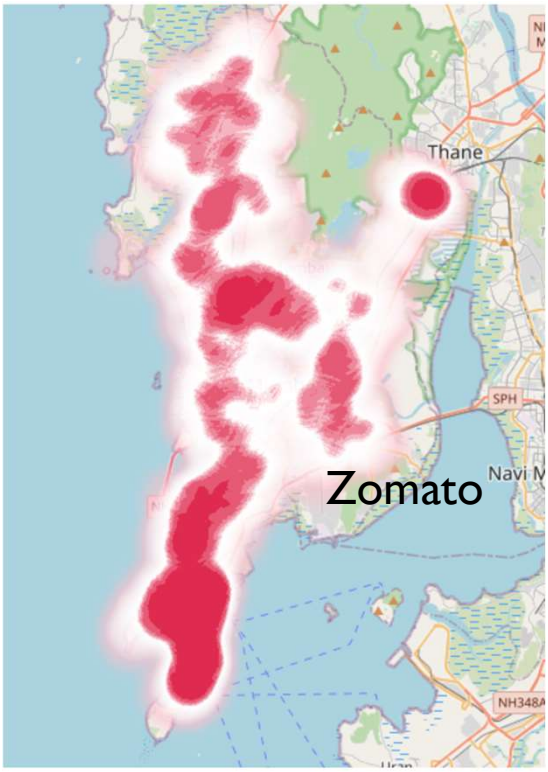
Venue Data

- Venue data was obtained using API calls to Zomato and FourSquare
- Parameters were set to get all food related venues in a radius of 1000 meters
- Maximum venues returned for a locality was set at 100
- Obtained data was filtered for any duplicate values by using unique Venue ID / Restaurant ID in the dataset
- Venue Coordinates obtained were then plotted as seen on the right.
- A total of **8647** unique venues were found from Zomato.
- A total of **2481** unique venues were found from FourSquare

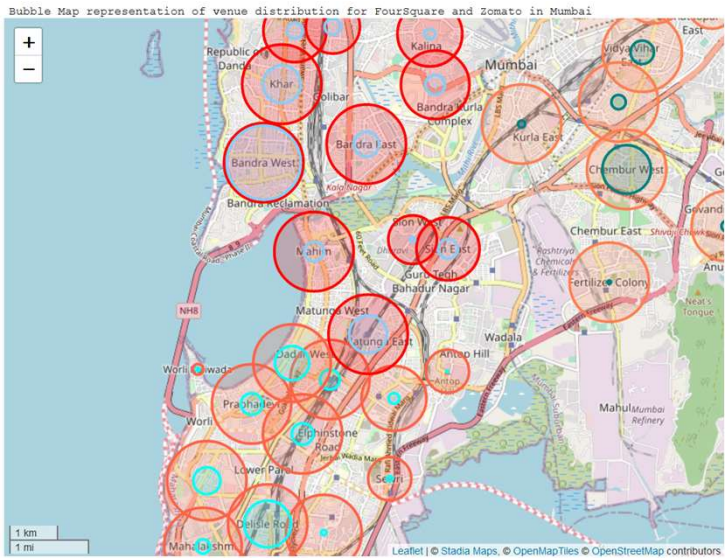


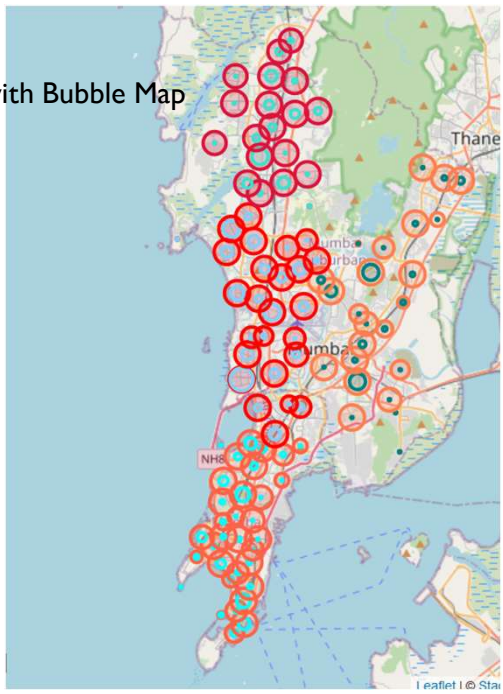
Venue Distribution with Heat Maps





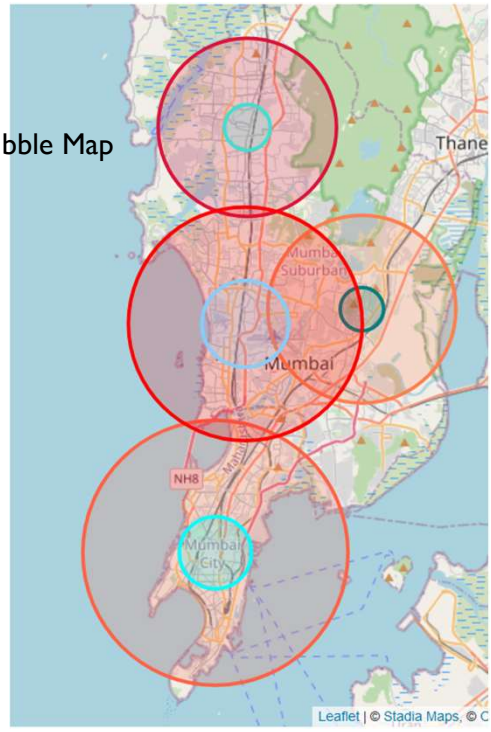
Data Contrast by Locality with Bubble Map





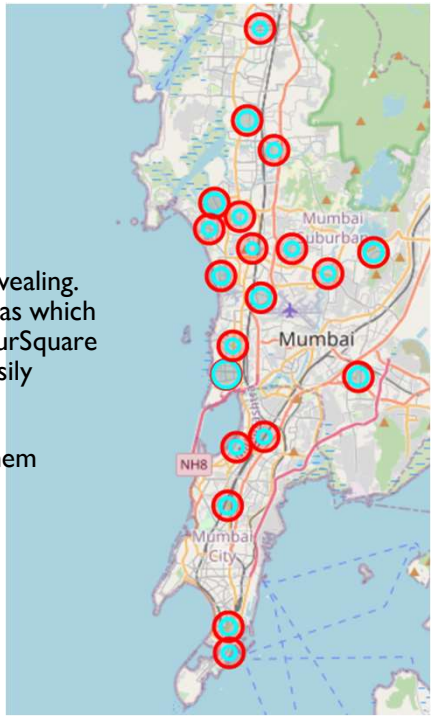
Data Contrast by Sector with Bubble Map

Sector	Latitude	Longitude	FourSquare Venues	Zomato Venues
North	19.2027	72.8481	459	1773
South	18.9735	72.8299	725	2652
East	19.1048	72.9133	431	1887
West	19.0969	72.8469	866	2335



Top FourSquare venues

Locality	Sector	Local Latitude	Local Longitude	Zomato Venues	FourSquare Venues
Bandra (West)	West	19.058336	72.830267	100	93
Lokhandwala	West	19.143189	72.824081	100	62
Vile Parle (East)	West	19.096288	72.848380	100	61
Chembur West	East	19.057110	72.899388	100	61
Malad (West)	North	19.184013	72.841216	100	60
Powai	East	19.118720	72.907348	100	59
Delisle Road	South	18.993260	72.831012	100	58
Hutatma Chowk	South	18.933307	72.831377	100	57
Juhu	West	19.107021	72.827528	100	53
Henry Road	South	18.920362	72.831837	96	49
Khar	West	19.072458	72.833707	100	48
Matunga	West	19.027436	72.850147	100	48
Saki Naka	East	19.108221	72.883582	100	47
Versova	West	19.130252	72.821377	100	47
Borivali West	North	19.229456	72.847991	100	43
Bhavanishankar Road	South	19.022237	72.835657	97	42
Jogeshwari (West)	West	19.136394	72.837382	100	42
Goregaon (East)	North	19.169262	72.855255	100	41
MIDC	West	19.120512	72.864637	100	41
Cama Road	West	19.120651	72.843866	100	39



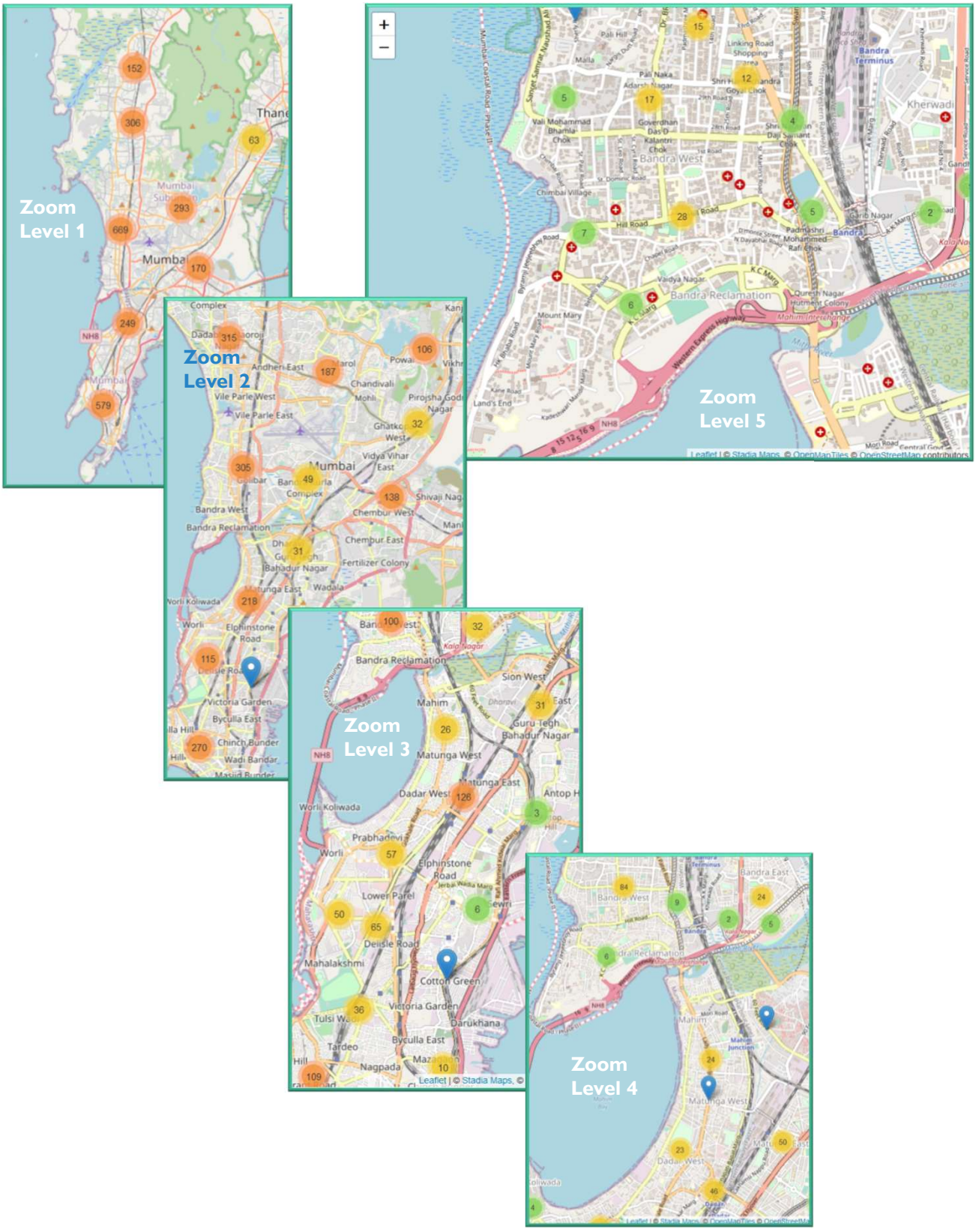
- This map is revealing. The major areas which have more FourSquare Venues are easily identifiable.
- We can call them ***Hotspots***

Hotspots

These Hotspots are

- **Bandra, Khar** : Upscale Coastal semi-suburb with multiple major tourist attractions
- **Lokhandwala, Jogeshwari** : Upmarket and affluent residential and commercial neighbourhood
- **Vile Parle, MIDC, Saki Naka** : Areas surrounding CSMI Airport and nearby hotels
- **Juhu, Versova** : Coastal areas with beaches and upscale neighbourhoods
- **Powai** : Upscale residential neighbourhood on banks of Powai Lake
- **Henry Road, Hutatma Chowk, Delisle Road** : Old neighbourhoods in South Mumbai with heritage tourist attractions
- **Malad, Goregaon, Chembur** : Upcoming upmarket residential localities in North-Western and Eastern Mumbai

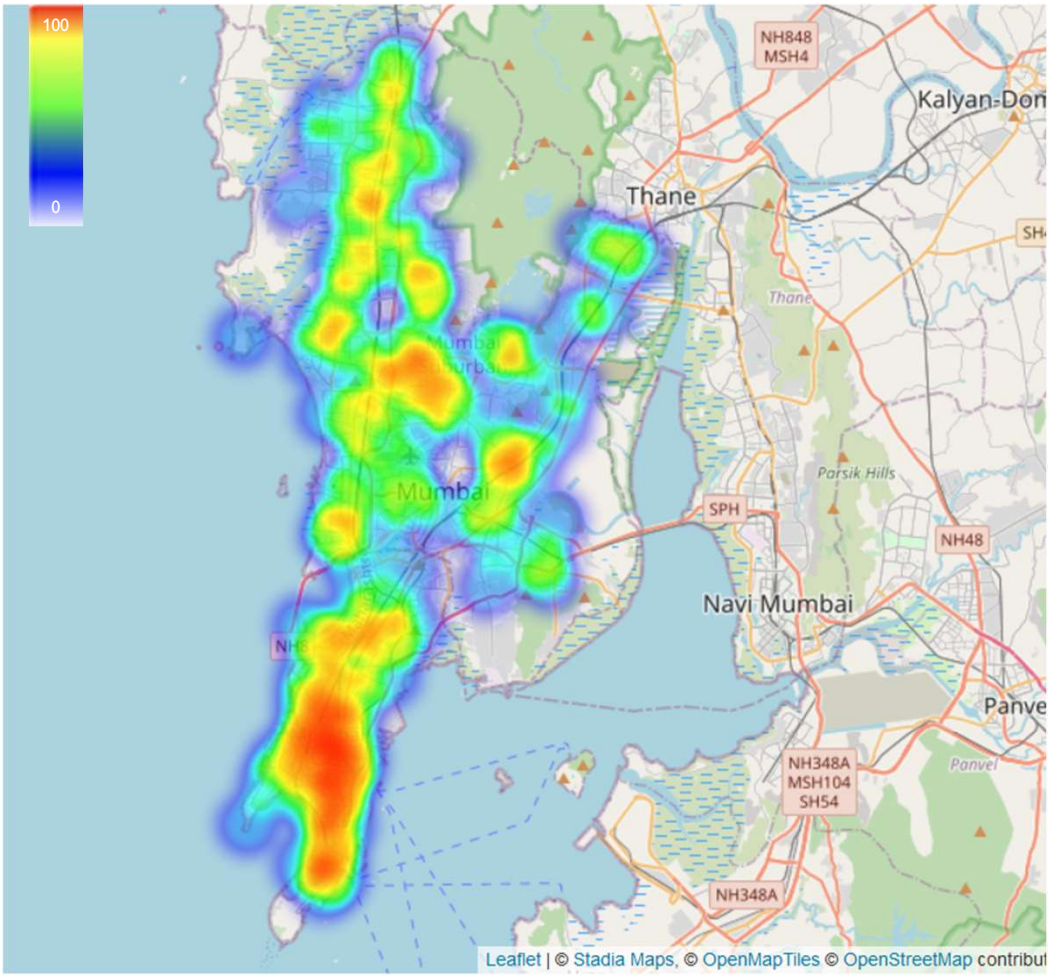
Cluster Map for FourSquare



Coverage of FourSquare with Heat Map

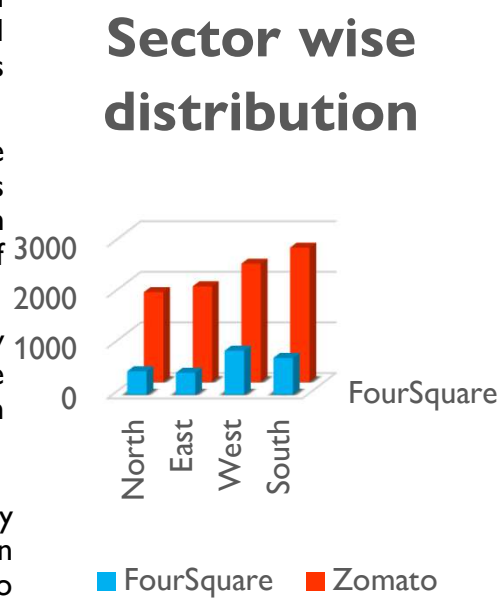
- We define *Coverage* as
- $Coverage = 100 \times \frac{Number\ of\ FourSquare\ Venues\ in\ the\ Locality}{Number\ of\ Zomato\ Venues\ in\ the\ Locality}$

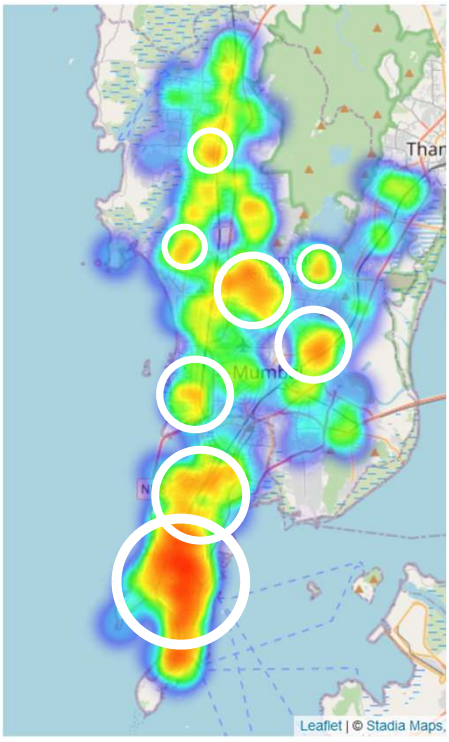
- Since we are measuring *Coverage*, any Localities with FourSquare Venues exceeding Zomato got their *coverage* normalized to 100.
- *Coverage* values were then taken as weights and a Heat Map was generated for all the localities as shown below



Inferences

- Our analysis shows that compared to data obtained from Zomato, which was relatively uniformly spread across all Localities in Mumbai, FourSquare Data was prominent only in certain areas.
- This was inferred from generated HeatMaps of the Venues. A Sector wise count showed that the bias is towards Western and South Mumbai, while in depth analysis using bubble maps gave the concentration of venues in the ***Hotspots***
- These Hotspots are places on the most-visited by tourists list in Mumbai. These are places of affluence and culture, where neighboring restaurants draw in clientale from foriegn visitors.
- Since FourSquare is a crowdsources its data, it is very likely that a large amount of it is submitted by foreign tourists (many of which use FourSquare regularly) who visit these Hotspots as compared to local residents (where FourSquare has low popularity).





- Thus, there is an automatic bias generated since it is concentrated only in the Hotspots. Zomato, a local (Indian) restaurant aggregator has no such bias in its data.
- The Cluster Map generated supports this, as we see the clusters coagulating on the hotspots.
- In order to better quantify this, we defined a custom measure called Coverage, which is a percentage of FourSquare venues to Zomato Venues. This was then used as a measure of weight to generate a Heat Map.
- This map clearly visualises our goal. We see the Coverage intensity highest in South Mumbai, and other areas of medium-high intensity being all the other Hotspots.
- The quantitative accuracy of Coverage is debatable. We are treating the Zomato data as a baseline; however, since the API does not return more than 100 restaurants, the value may exceed 100 in certain localities. There, the value of Coverage will be lowered. But this does not discount the qualitative accuracy. Qualitatively, we can certainly infer that the Hotspots are regions where the value of Coverage is higher.

Conclusion

- The purpose of this project was to study and understand the FourSquare data coverage. By obtaining restaurants from both FourSquare and Zomato, we got venues for each Locality. A Visual Exploratory Data Analysis was performed to see the characteristics of venue distribution, which led to pattern similar to well known Hotspots in Mumbai; primarily regions in South Mumbai, Western Coast, and near the Airport. A custom measure called Coverage was defined and corresponding HeatMap was generated. This allows a concise measure to visualise and study the FourSquare Data Coverage.
- Relevant Stakeholders-primarily FourSquare-can use this model to understand the nature of Crowdsourced data generation. The biases generated by tourist generated data can possibly cause non-uniformity in the coverage of such a service.