

Prediction of New Opportunities for Bakery Businesses in Pune

Divya Vyas

27 May 2020

2. Data acquisition and cleaning

2.1 Data sources

To get further details about the potential areas first I needed to get names and coordinates with the Postal Codes of all the different areas of Pune District.

1. The same details were taken from the below mention website :-

<https://geographic.org/streetview/india/maharashtra/pune/pune.html>

2. . After acquiring the data from above website, it was found that the data of some of the possibly potential areas like “Koregaon Park, Boat Club Road, Fatima Nagar etc” were not available. The data for the same was acquired from google searches.
3. After getting all area locations, we needed to understand food and lifestyle preferences of the city. Hence, we got all venues details corresponding to the respective areas from FourSquare API.

2.2 Data cleaning

Pune district consists of a large geographical area. Many such locations fall in the small town or rural categories. The data set acquired consisted of all such areas as well.

As the category of our bakery does not quite appeal to the audiences of such locations, I needed to focus mainly on areas that fall in the main Pune city area.

The Postal Code of all the main Pune city areas begin with ‘411’. Hence, I filtered the dataset accordingly.

In order to easily take a glimpse at any particular area I sorted the list in Ascending order. After acquiring venues details from FourSquare I divided data into 2 datasets as follows

1. **All the areas wise Venues details (across all categories)** : This dataset would help us to understand location wise food and lifestyle preferences of the audiences.
2. **Area wise Bakery details:** this would help us to understand the distribution of bakery location in the main city areas so that we could identify locations that do not have sufficient bakeries but fall into the same cluster as the areas which have.
3. **Zone wise data:** A other way to understand data would be to divide the main city areas in to 5 zones. Mainly, “South, North, East, West and central”.
4. **Zone wise Bakery distribution:** The data was then further processed to see the bakery distribution in the above mention’s zones.

3.Exploratory Data Analysis

3.1 Zonal Data

Upon exploring data, according the zones for the bakery category. It was found that the bakeries in Pune were largely evenly distributed. However there seems to be more opportunities in the East and North zone compared to the others. This is evident in the graph below.

Central and South zone seems to be having a higher competition in bakeries.

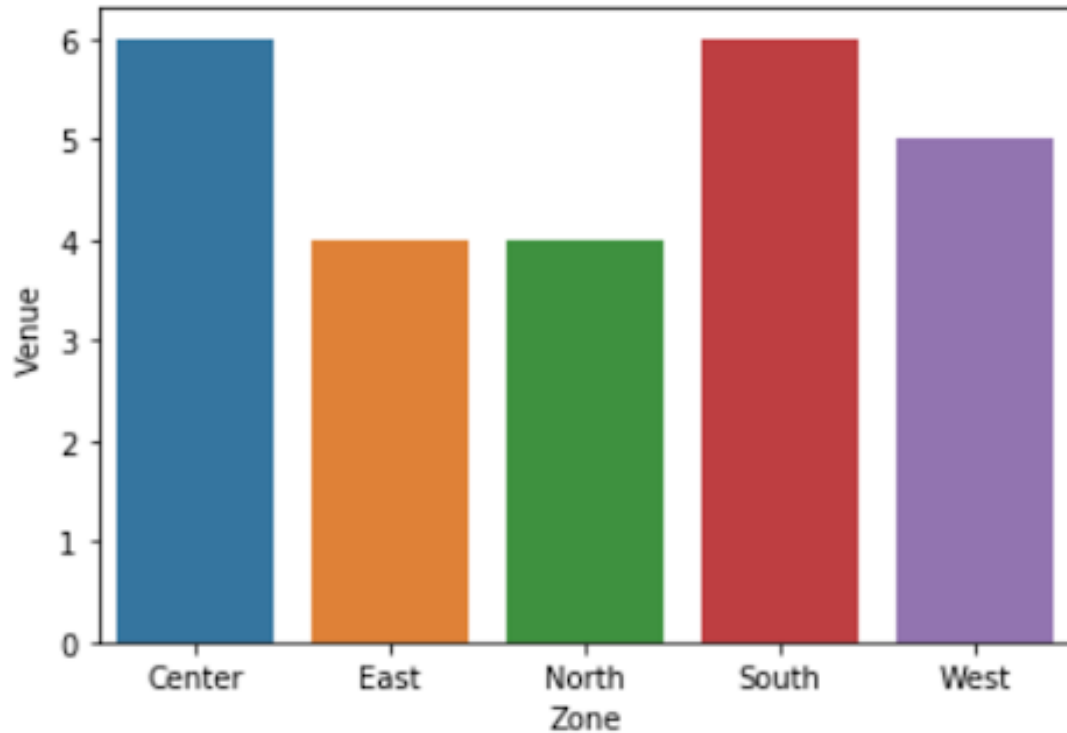


Figure.1 Zone wise venues distribution

3.2 Neighbourhood-wise Venue Distribution

This was to understand the category preferences of all the different neighbourhoods.

I took a count to check how many venues were distributed across different neighborhoods. I got a total of 143 unique categories.

I used “Onehot” encoding after which I took the “groupbymeans” to view weightages of categories across neighbourhood.

I found out the top 5 venues in each of the neighborhoods. This would help me understand the neighborhoods better and also form a strong basis for input to a machine learning algorithm that would follow post data exploration

Finally, I put all this data into a data frame. The data now appears in a 2- dimensional format and the columns are arranged in the order of most preferred locations for each of the respective neighborhoods.

3.3 Heatmap for key categories and neighbourhoods

The data set of Pune City consist of large number of data. Further the data acquired for FourSquare also consisted of a large number of venues. In order to better understand the cultural and lifestyle preferences of people the large data set would certainly give us good insights about the audiences.

However, to explore and comprehend the data better, I decided to filter the data into a few key categories which primarily belong food industries. I also filtered only some of the key neighbourhoods in the city of Pune.

Further I decided to plot this on the heat map to draw some key inferences

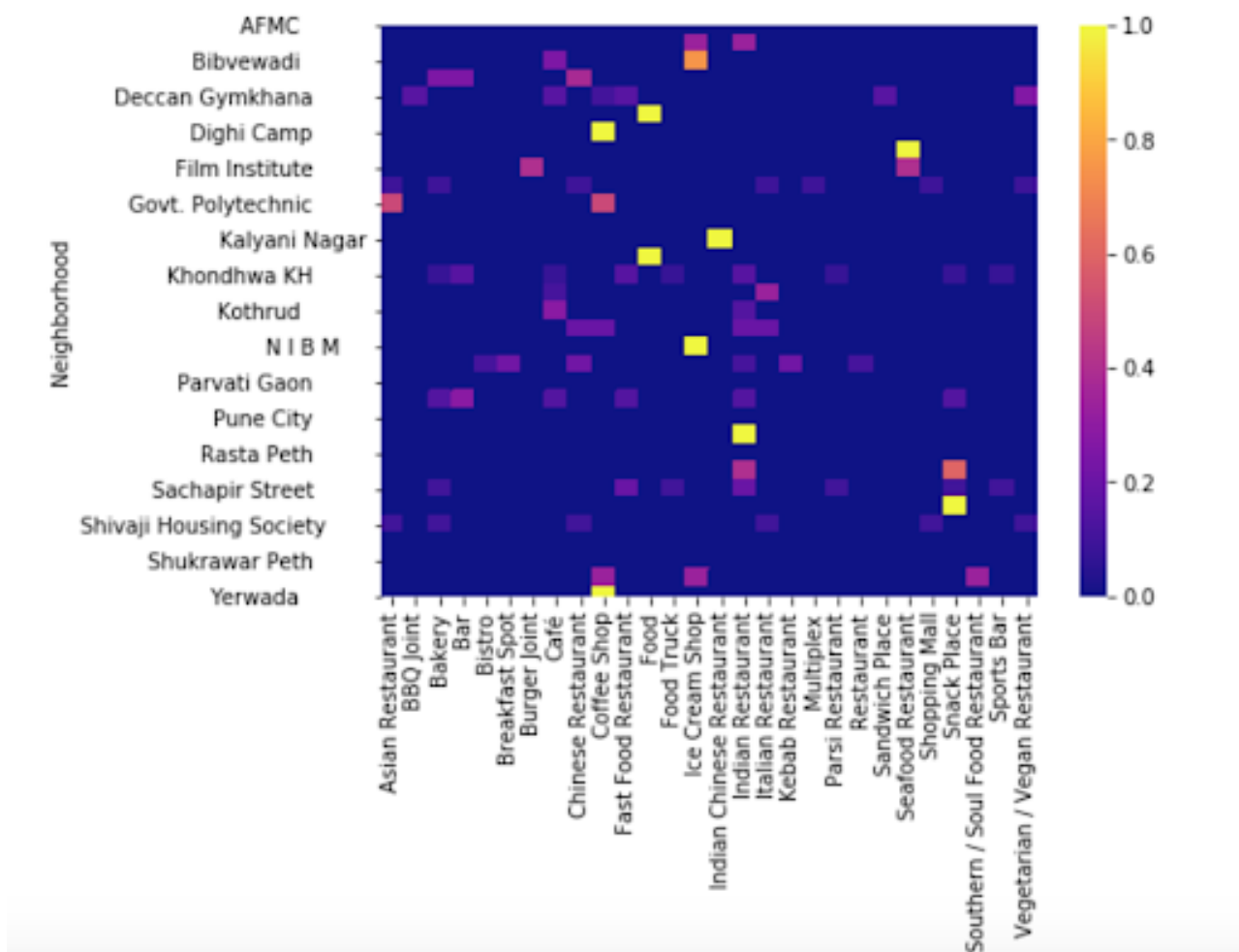


Figure.2 Heatmap for Venue Category and Neighborhoods

Inferences from Heatmap:

- Areas like Kalyani Nagar, Sadashiv Peth, Yerwada and NIBM seem to be bustling with food trucks, restaurants, bakeries, snack joints and cafes.
- Areas like Pune Cent East, Kondhwa and Boat Club Road seem to be having decent number of bakeries around their respective vicinities.