

The background is a dark blue gradient. It is decorated with various geometric elements: thin white vertical lines of varying lengths, small squares in teal, orange, and pink, and larger squares in teal and orange. Some of these shapes are solid, while others are just outlines. They are scattered across the slide, creating a modern, tech-like aesthetic.

Clustering in Mumbai

An overview

Introduction

Several small squares in various colors (cyan, pink, orange) are scattered in the top right corner of the slide.

We will be looking at neighbourhood data in Mumbai to ascertain spots or clusters in the city that are ideal for a new restaurant.

This data should help one understand the various neighbourhoods in Mumbai and choose a good neighbourhood to start a restaurant in.


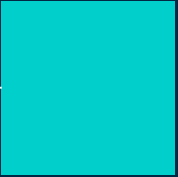
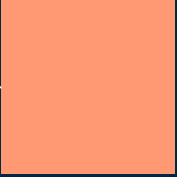
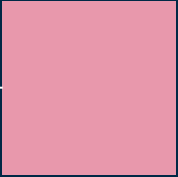
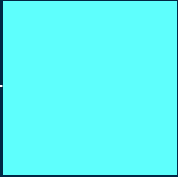
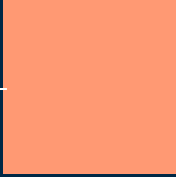
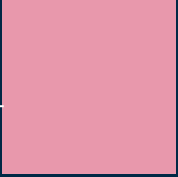
Two small squares, one orange and one cyan, are located in the bottom left corner of the slide.

TABLE OF CONTENTS

| | | | | | |
|--|--|---|--|---|--|
|  01 Business Problem |  02 Data |  03 Methodology |  04 Analysis |  05 Results & Discussion |  06 Conclusion |
|--|--|---|--|---|--|

Business Problem

In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to restaurateurs interested in opening a restaurant in Mumbai. Here we will try finding if someone wants to open a new restaurant in the city which location is best suited for it keeping in mind the competitors and which income group of people will be attracted most to it based on the population of the neighbourhood.

Since there are lots of restaurants in Mumbai, we will try to detect locations that are not already crowded with restaurants but still have some competition. We would also prefer locations as close to city center as possible, assuming that first two conditions are met.

We will use KMeans clustering after the initial cleaning of Data to generate a few most promising neighbourhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decision are:

- All existing restaurants in the neighborhood (any type of restaurant)
- Age group of people with their income

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods. Following data sources will be needed to extract/generate the required information:

- Neighbourhood data in reference with its coordinates and zipcode/postal code
- Number of restaurants and their type and location in every neighborhood using Foursquare API

Methodolgy

To solve the problem I am going to use "K-Means Clustering Algorithm ". K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are: The centroids of the K clusters, which can be used to label new data Labels for the training data (each data point is assigned to a single cluster)

Analysis

The data for Mumbai was hard to find however after successfully organising the data I imported it to the notebook and checked for missing values. 228 Neighbourhoods in Mumbai were plotted on a map to better visualise the data and for an outsider to get a glance at the city and the several neighbourhoods within it.

After this I used foursquare API to explore and find top venues in each location. I defined a function such that it would input the co-ordinates of each neighbourhood in Mumbai and provide us with a dataframe of useful information. This gave us a total of 2815 venues with their co-ordinates and categories so that we can better understand each neighbourhood.

We apply one hot encoding where the integer encoded variable is removed and a new binary variable is added for each unique integer value. This will help the machine better fit the model and output useful clusters. We then sort the data into the top ten most visited venues in each neighbourhood. We then apply KMeans clustering to find form the neighbourhoods into clusters.

Finally, I have mapped the clusters after assigning each one a different color for better understanding. The map would help in understanding which cluster one should choose.

Results

We can see that the six clusters have varying sizes. We can see that each cluster has different most visited venues for example Cluster 5 includes neighbourhoods that have more gyms and parks in them. cluster 1 has however provided us with a group of 64 neighbourhoods whose first two most common venues are mostly restaurants. This implies that these neighbourhoods have a great variety of various restaurants and could imply a more competitive environment for any restaurateur which could also translate to higher profits.

For more 'bull-ish' businessman Cluster 6 would be a better fit as these neighbourhoods do see a high number of restaurants and eateries however it is not as concentrated as Cluster 2. This would mean less risks and would be good for any new businessmen/restaurateur who wants to enter the market.

Conclusion

In conclusion, The goal of this project was to ascertain optimal locations for businesses to bloom. The total number of neighbourhoods in Mumbai is 228. Through KMeans clustering and analysis we can remove more than half the neighbourhoods if the restaurateur selects Cluster 6. However if the restaurateur selects a more 'bear-ish' approach and selects Cluster 2 they would subtract approximately 72% of the total neighbourhoods.

