**Capstone Project – The Battle of Neighbourhoods**

**Report**

**By**

**Gaurav Tiwari**

1. **Introduction**

India is one of the major economies of world, with a large consumer population and a wide variety of products and services in market. This makes up for a high amount of daily financial transactions.

An Automated teller machine (ATM) is electronic telecommunication device that provide customers to perform financial transaction such as cash withdrawal, deposits, fund transfer and other bank services at any time without a need for directly going to bank.

While India is moving towards digital cashless transaction still a very large percentage of transactions are with cash. This large cash demand makes ATM a key service/product for a bank’s customer, and right location to set up an ATM can be a good strategy for any bank.

**1.1 Business problem**

For a bank to increase its reach to customer in terms of providing services relies on ATM as they are the easiest and most convenient way to acquire cash. It’s important for a bank to find best location to reach large customers and this will be the focus here. We will be tackling one of the cities in India and find out best location for ATM.

Here for this project I have chosen Ahmedabad city located in Gujarat, India.

1. **Data Gathering and Processing**

**2.1 Data Gathering**

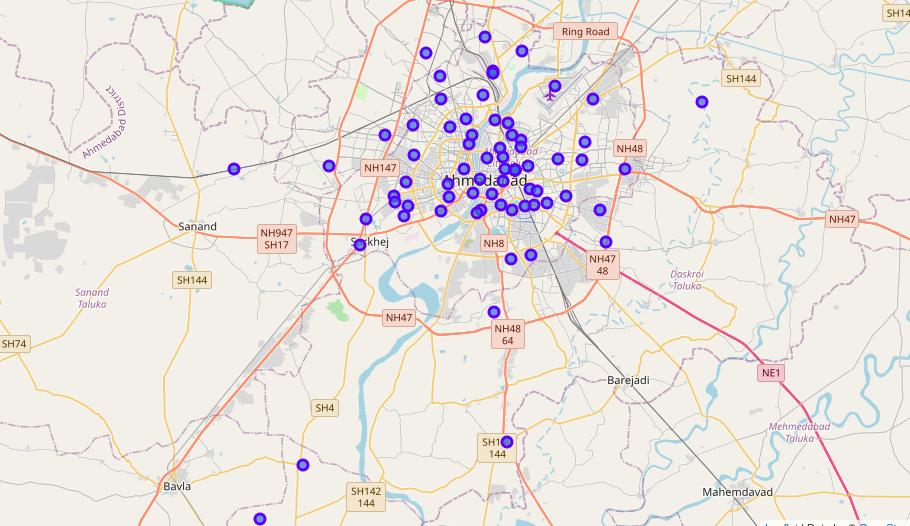
* Data required here involved neighbourhoods in Ahmedabad, geographical coordinates of neighbourhoods, and venues in these neighbourhoods.
* Neighbourhoods of Ahmedabad were gathered by web scraping performed using BeautifulSoup from Wikipedia. A total of 81 neighbourhood in the city.
* Geographical coordinates were acquired using geocoder. ArcGIS. This gave us Latitude and Longitude of the neighbourhoods. 

Figure 2.1 Visualization of Ahmedabad

* Venues data was obtained from Foursquare through an API. An API call was made with personal credential. This provided the information of venue, venue category, neighbourhood location, venue Latitude and Longitude.

**2.2 Data Processing**

* First step of data processing was merging Neighbourhood DataFrame with geographical coordinates DataFrame, which resulted in a DataFrame with 3 columns of Neighbourhood, Latitude and Longitude.
* Next utilizing Foursquare API, a new DataFrame was created with columns of Venue, Venue Latitude, Venue Longitude and Venue Category, in addition to the 3 columns from previous DataFrame i.e. Neighbourhood, Latitude and Longitude. These were merged according to Neighbourhood. 

Figure 2.2 Ahmedabad Venues

1. **Methodology**

The methodology in this project consists of two parts:

**3.1 Exploratory Data Analysis**

Visualise and extract the neighbourhoods in that borough to find the 10 most common venues in each neighbourhood.



Figure 3.1 Top ten venues for each neighbourhood

**3.2 Modelling**

To find similar neighbourhoods we will be clustering similar neighbourhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. To find out best K we will be finding out distortion/inertia and also if needed Silhouette score. K-mean will group similar neighbourhoods into clusters, these clusters should have a lot of common venues and this will help us understand which clusters and in turn which neighbourhoods are best.

K-means clustering: It is a type of unsupervised learning, which is used when you have unlabelled data. The main objective of this algorithm is to find groups in data, which is represented by value K. The Algorithm works iteratively and the data points are clustered on basis of similar feature.

For finding optimal K we have used two methods:

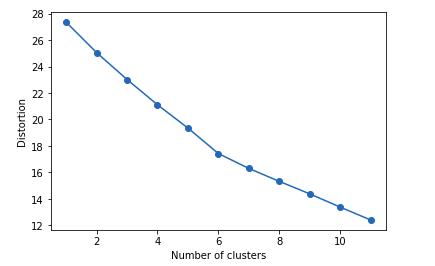
1. Using Inertia/Distortion: This is what is called The Elbow method. Inertia is sum of squared distances of sample from their closest cluster. Distortion is the average of squared distances from the cluster centre. 

Figure 3.2 Distortion vs Number of clusters

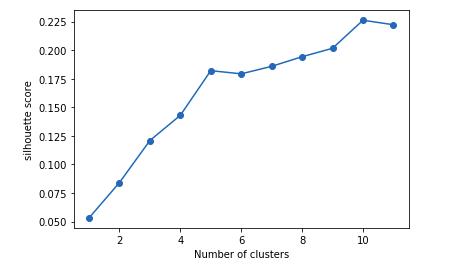
1. Silhouette score: It is a measure of how similar an object is to its own cluster compared to other clusters. The Silhouette ranges from -1 to +1 where a high value shows a well match between objects in cluster and poor match to objects in another cluster. 

Figure 3.3 Silhouette Score

Once the best K is found we cluster neighbourhoods in those and observe each cluster individually. Ideally clusters will have neighbourhoods with very similar venues. Here we will also visualize map with different neighbourhoods in different cluster labelled.

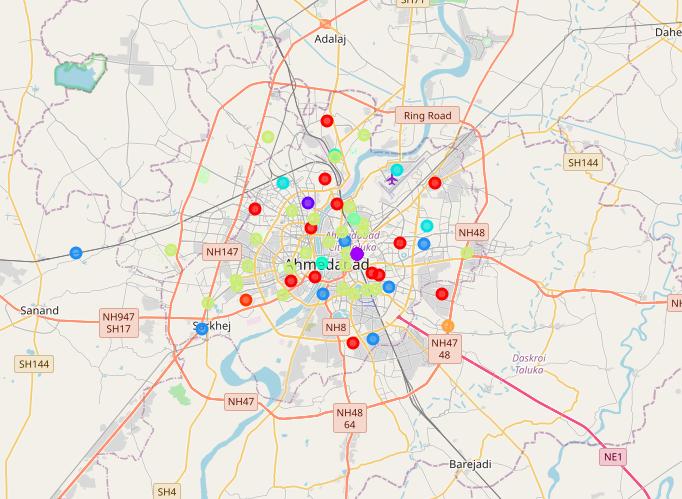


Figure 3.4 Visualization of clusters of neighbourhoods in Ahmedabad

1. **Result**

As per Silhouette score, we came up with optimal number of clusters k = 10 Following which we visualized cluster and observed each cluster individually.

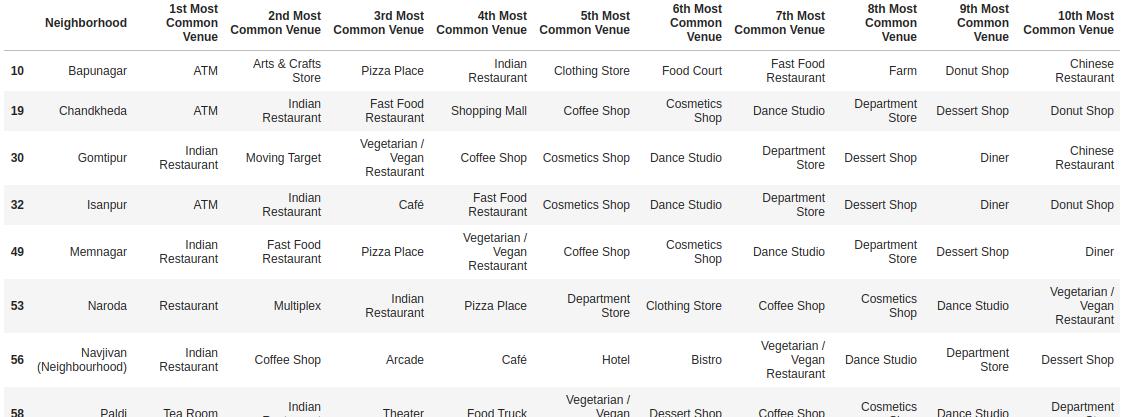
Cluster 1 is the 2nd biggest cluster observed and consisting of 13 neighbourhoods with Restaurant, coffee shops and other outside eateries as most common venues. We also see that 3 neighbourhood have ATM as the most common venue. 

Figure 4.1 Cluster 1

Cluster 2 is 3rd biggest cluster with 8 neighbourhood with completely similar common venues with Motel, Restaurant and Men’s store being top 3 common venue. 

Figure 4.2 Cluster 2

Cluster 3 consist of 2 neighbourhood with Bus station as most common venue. 

Figure 4.3 Cluster 3

Cluster 4 consist of 7 neighbourhood with ATM being most common venue and other venues show similar characters with Restaurant and coffee shops being few of common venues. 

Figure 4.4 Cluster 4

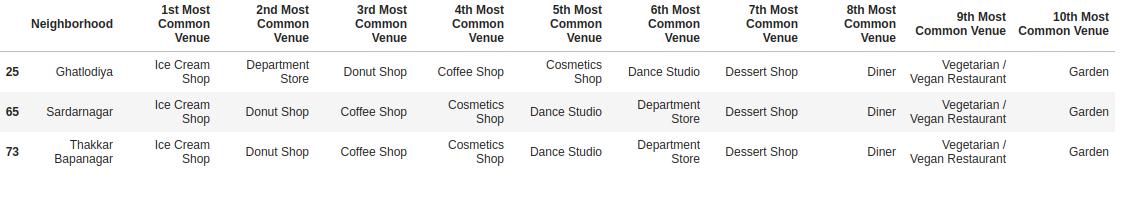
Cluster 5 consist of only 3 neighbourhood with almost completely same common venues. 

Figure 4.5 Cluster 5

Cluster 6 consist of 2 neighbourhoods with completely same common venues. 

Figure 4.6 Cluster 6

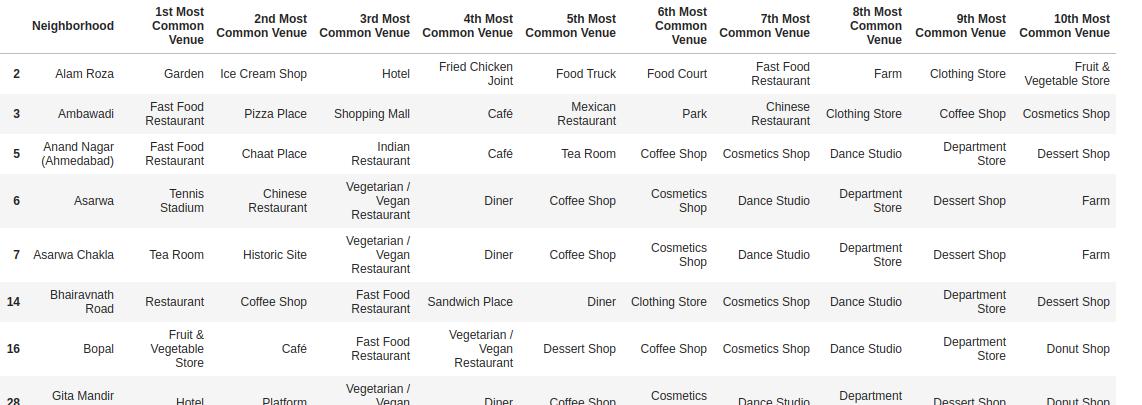
Cluster 8 is biggest cluster with 27 neighbourhoods, this has a lot of neighbourhoods with Restaurants as one of most common venues. 

Figure 4.7 Cluster 8

Cluster 7, 9 and 10 consist of only 1 neighbourhood each.

1. **Discussion**

Once we observe each cluster individually, we see a lot of important insights into how clusters were formed. To begin with cluster 4, neighbourhoods in this clusters have ATM as first most common venue. Clearly this location has a large number of ATMs already present and clearly shouldn’t be focus for further increase in ATM number as per this analysis. But this also shows that high frequency of ATMs is also accompanied with high frequency of restaurants and shopping outlets.

* Cluster 1 is 2nd largest cluster and it consist of 3 neighbourhoods which show high number of ATMs as well as high number of restaurants and shopping outlets. But none of the other neighbourhoods have ATMs in top ten venue. As all neighbourhoods in same cluster have similar characters these neighbourhoods in cluster 1 should be a prime location for new ATMs
* Cluster 2 is 3rd largest cluster and it consist of neighbourhood that have completely same venues, most common venues do consist of restaurants and shopping outlets but this cluster also doesn’t have any ATMs in top ten. This could be a good place for new ATMs.
* Cluster 8 is largest cluster and doesn’t show a particularly strong similarity which can make neighbourhoods in this cluster a prime spot for new ATMs, but some of the neighbourhoods have potential to be a good spot based on significant number of restaurant and shopping outlets.
* Other clusters are too small to be useful here but could be utilized in a further study involving daily traffic into picture to see if these locations are involved with high number of people visiting. As places like bus and train station can be a good spot for ATMs as well.

1. **Conclusion**

This project shows us the patterns associated with ATMs location when looking from the point of most common venues. Neighbourhoods with high number of ATMs show us the other most common venues around a particular ATM. This information can be utilized by a bank to pick best spot for their new ATMs to provide a better reach to its customers. This analysis could also be accompanied with daily road traffic to get much better location.