Exploring Neighborhood of Colleges under Delhi University

Saransh Bajpai
17th July 2020

1. Introduction

1.1. Background

Every year thousands of students across India enroll in various programs in different colleges in the capital city of Delhi. Since its inception in 1922 University of delhi has grown to be the most prestigious university with an increasing number of colleges and enrollments year after year. Students look forward to joining the college of their choice based on their merit and interest. A student gets a choice to choose between a number of colleges ranging from one to as high as fifteen based on his merit.

1.2. Problem

A student before joining any college wants to make an informed choice and wishes to weigh his preference on the basis of concrete facts. When on one hand the rank of the college matters to him, other factors such as the cost of living in the vicinity of the college and the accessibility of services in and around the college(knowledge of neighborhood) is also a factor. Google search can help him/her explore colleges individually, but doing this for 70+ colleges and them comparing them altogether is a tedious task. Moreover the data which a student seeks to study is neither readily available nor clean enough or in the same metrics for all the college locations. Therefore a comprehensive study of college location needs to be done.

1.3. Interest

Students looking forward to enroll in any college in the capital city of Delhi would be interested in this study. Also since this study categorizes college in terms of rental price and venues in vicinity, this report shall interest business people to set up affordable renting space and other marketplaces.

2. Data acquisition and Cleaning

2.1. Data Source

The list of colleges under Delhi University was scrapped from their **official website**: www.du.ac.in/du/index.php?page=list-of-colleges.

The rental prices for various localities was scrapped from the website - https://www.makaan.com/price-trends/property-rates-for-rent-in-delhi
Makaan.com is a company that lists real estate pricing and their details all across India.

The coordinates of college as well as localities of Delhi was retrieved using **Nominatim geocode**r. Since this package couldn't find coordinates for each and every college,the leftover entries were typed in manually.

The FourSquare api was used thereafter to get the nearby venue details. Foursquare helps find trending, nearby or specific categorical places in and around a location based on its geographical coordinates.

2.2. Data Cleaning

- Some of the colleges which run evening classes in the same college premise were repeated in the list. The duplicate entries were removed from the list.
- The scraped data from makaan.com gave a table containing the avg rent prices for 1,2 and 3 BHK housing as well as a range of price for the same for 700+ localities. Many cells of this table were missing. Before dropping, if the data for 1BHK was missing and the price for either 2 or 3 BHK was given, it was used to calculate and fill in the missing value(1 BHK).
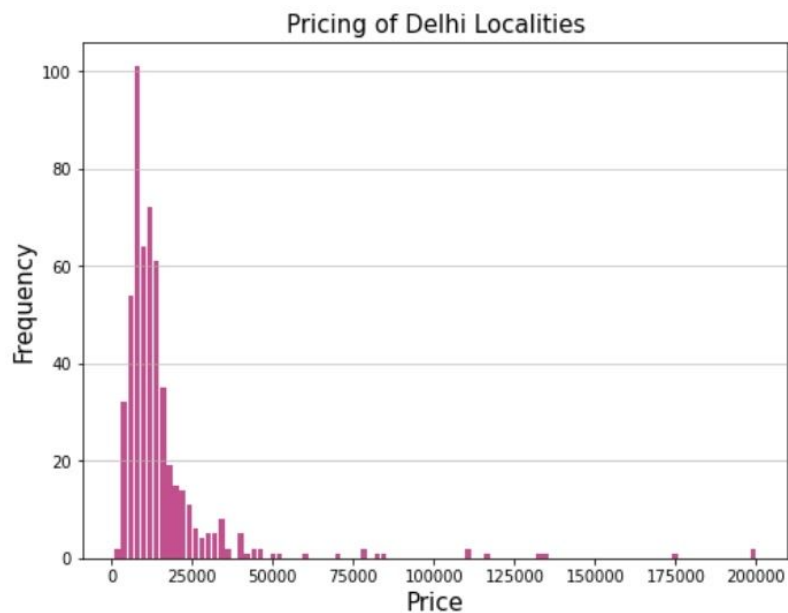
2.3. Feature Selection

For the rent pricing data **1 BHK avg rent price** was chosen as the primary metrics and hence all other columns : 2 BHK or 3 BHK rent price as well as the bracket range were only considered to formulate the primary metrics column.

3. Methodology

3.1. Exploratory Data Analysis

3.1.1. Colleges and Avg Rent relationship


Pricing of Delhi Localities

The locality average pricing data scraped from makaan.com contains many outliers in high range values. These prices are of the expensive housing which aren't availed by students and therefore need to be removed from our analysis.
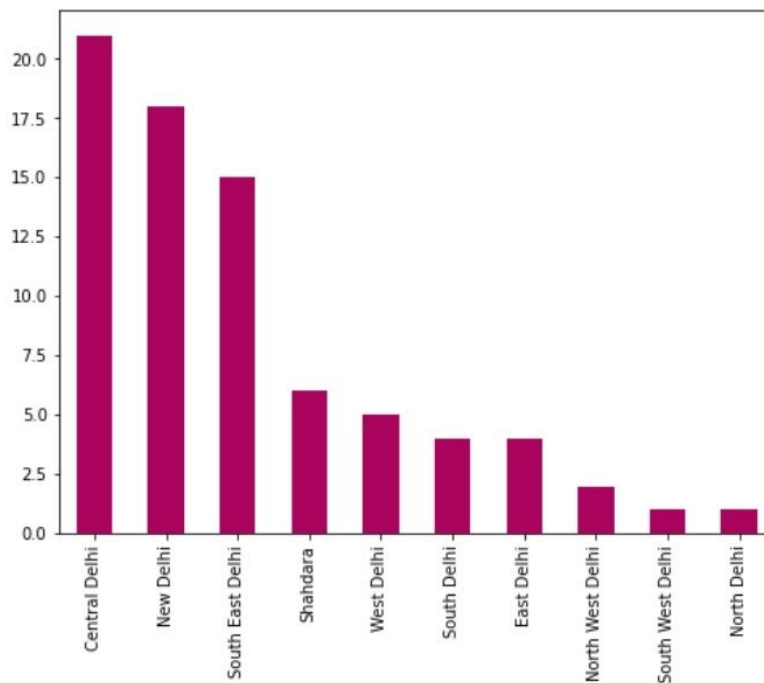
3.1.2. Locality and Avg Rent
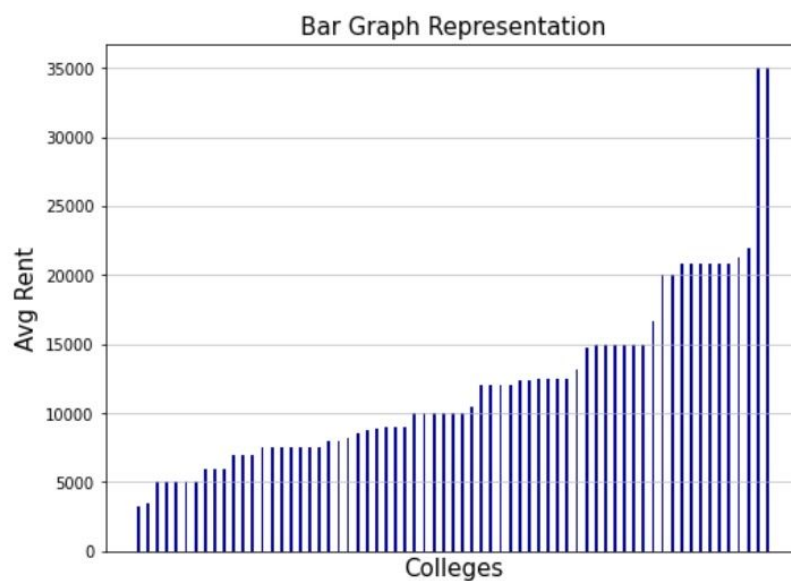
Bar Graph Representation

The average rent of places is mostly between 5000-20,000 thousand with outliers as high as 35000.

Pricing Histogram
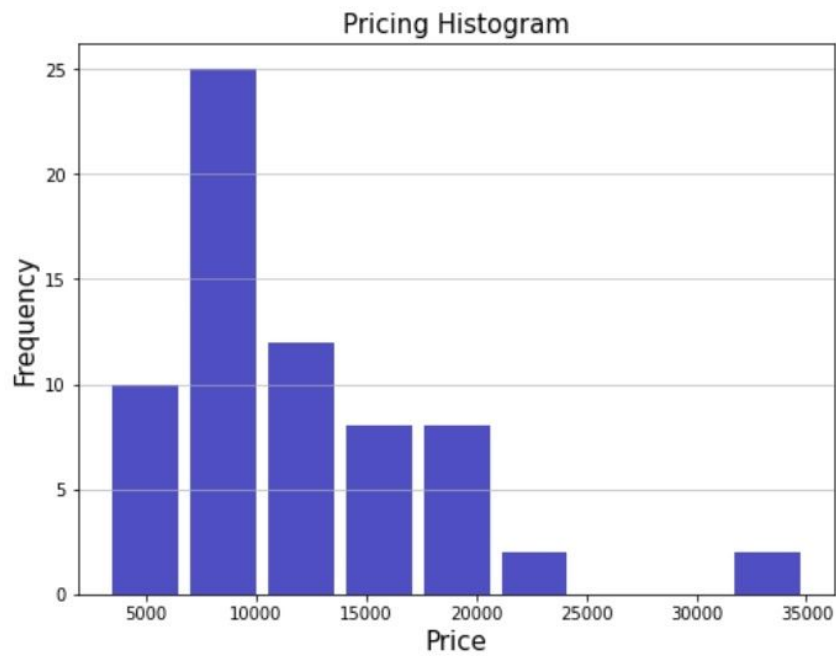
The average rent of places is mostly between 5000-20,000 thousand with outliers as high as 35000.

3.1.3.   Colleges Zone wise

Most of the colleges are in Central,New or South East Delhi region. West, South, East Delhi have mediocre numbers of colleges while North West, South West and North Delhi have the least number of College campuses.

3.2.   Clustering Model

3.2.1.   Standardization

The data contains venue details in the form of one-hot encoding and rent pricing in the range of inr 5,000-35,000. Therefore data standardization is needed before plugging the values into the cluster model.
Min-max Scaler is used for the standardization.

3.2.2.   Optimal 'k' for K-means

Elbow method can be used for finding the optimal k, but doesn't guarantee an optimum result always. The elbow graph did not show a unique value of k (as expected), since the data is very homogenous. At k = 4 the drop is substantial, therefore acting as the best value for the clustering.

3.2.3.   Machine Learning Model : K- Means

K-means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters). In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure

4.   Result and Discussion

4.1.   Clustering and Visualization

The colleges are clustered into 4 clusters represented as a scatter plot.
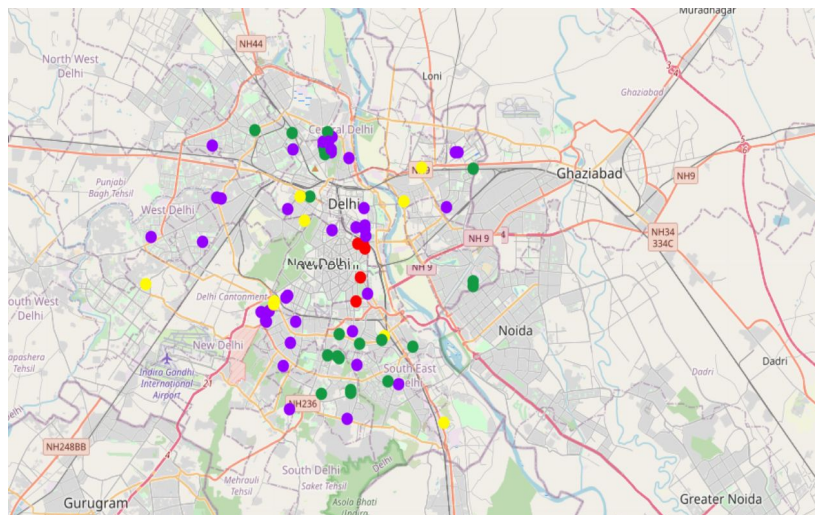
4.1.1. Cluster 1

This cluster contains colleges of particularly **Central Delhi zone** which have a **mediocre** living expense. The neighborhood of such colleges consists of places like *restaurants,hostels and parks.*

4.1.2. Cluster 2

This cluster contains colleges of mostly **all zones** which have a **least** living expense. The neighborhood of such colleges consists of places like *department stores and shops for all needs.*

4.1.3. Cluster 3

This cluster contains colleges of **New Delhi zone** which have a **very high** living expense. The neighborhood of such colleges consists of places like *cafes, plaza and bistro.*

4.1.4. Cluster 4

This cluster contains colleges present in the **New Delhi zone** that have a **considerably expensive** rent price. The neighborhood of such colleges consists of lavish places such as Theatre, Art Gallery & Museum, Arcade.

5. Conclusion

The colleges of Delhi can be clustered in four different clusters based on their pricing and neighborhood data. The division of clusters shows a similarity with the zones in which the capital city of Delhi is already divided. The New Delhi Zone has the highest housing price as visible from cluster 3 and 4. This is true since the New Delhi region consists of places of national importance thus adding value to its region. The Central Delhi zone has mediocre housing options in the range of 12-15 thousand.

We can also conclude that the average rent price of a neighborhood has a relation with the venues surrounding it. Colleges in neighborhoods with considerably high housing prices have places like cafes, bistros, art museums and theatres. On the other hand, colleges in neighborhoods with considerably decent/low housing prices have places like department stores and shops small and big of all necessities.This is a classic example of how venues decide the pricing of the neighborhood.

Therefore using this study a student can have the knowledge of colleges within a common cluster and can use it while deciding college of his choice backed by his financial status and expected quality of living.