

“Exploratory Data Analysis”

Mohammed Salmanuddin¹, Rushikesh Kulkarni², Atharva Mohite³, Prof. Mahendra Patil⁴

^{[1],[2],[3]} Students, Department of Computer Engineering, Atharva College of Engineering, Mumbai, Maharashtra, India ^[4] Professor, Department of Computer Engineering, Atharva College of Engineering, Mumbai, Maharashtra, India

Date of Submission: 15-04-2023

Date of Acceptance: 25-04-2023

ABSTRACT – This project aims to help incoming students find suitable accommodation by using K-Means and DBSCAN clustering algorithms. The analysis is based on students' preferences for amenities, budget, and proximity to the location. The data consists of accommodation details in various neighborhoods of the city.

The study utilized exploratory data analysis techniques, such as descriptive statistics, univariate visualization, and multivariate visualization, to gain insights into the dataset. K-Means and DBSCAN clustering algorithms were applied to classify the accommodation into different clusters based on the preferences of the students. The results showed that both algorithms successfully classified the accommodation into clusters, with K-Means providing a more structured clustering, and DBSCAN being more flexible and able to detect outliers and noise.

In conclusion, the project successfully applied K-Means and DBSCAN clustering algorithms to assist students in finding the best accommodation in a new city. The study provided valuable insights into the preferences of students and how they influence the choice of accommodation. The findings of the study can assist incoming students in finding the most suitable accommodation based on their preferences.

Keywords: Machine Learning, Data Visualization, Data Cleaning, Student accommodation, Geolocation, Geographic Information Systems, Evaluation.

I. INTRODUCTION

Exploratory Data Analysis (EDA) is an approach for data analysis that utilizes a variety of techniques to summarize main characteristics of the data set, often with visual methods. EDA is useful for a range of purposes such as: Maximizing insights into a data set, mapping out underlying structure of the data, identifying useful variables, detecting outliers and anomalies and Testing a hypothesis. EDA is about getting to know and

understanding your data before making any assumptions about it. Different techniques used for analysis of the data are outlined below:

- 1) Clustering and dimension reduction: Creates graphical displays of high-dimensional data with many variables.
- 2) Univariate visualization: Method of looking at a one variable of interest.
- 3) Multivariate visualizations: Analysis of multiple variables at the same time.
- 4) K-Means clustering.
- 5) Predictive Models.

II. K-MEANS CLUSTERING

We are given a data set of items, with certain features, and values for these features (like a vector). The task is to categorize those items into groups. To achieve this, we will use the kMeans algorithm; an unsupervised learning algorithm. ‘K’ in the name of the algorithm represents the number of groups/clusters we want to classify our items into.

The algorithm will categorize the items into k groups or clusters of similarity. To calculate that similarity, we will use the euclidean distance as measurement. The algorithm works as follows:

1. First, we initialize k points, called means or cluster centroids, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that cluster so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters. This project involves the use of K-Means Clustering to find the best accommodation for students in a city by classifying accommodation for incoming students on the basis of their preferences on amenities, budget and proximity to the location.

Implementing the project will take you through the daily life of a data science engineer - from data preparation on real-life datasets to visualizing the data and running machine learning

algorithms, to presenting the results.

III. DBSCAN CLUSTERING

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm used in machine learning to group similar data points based on their spatial proximity and density. Unlike other clustering algorithms that rely on a predetermined number of clusters, DBSCAN is capable of finding clusters of arbitrary shapes and sizes, making it a flexible and versatile tool for clustering data.

The DBSCAN algorithm starts by selecting an unvisited data point and examining its neighborhood defined by the eps parameter. If there are at least minPts points in the neighborhood, the point is considered a core point and a cluster is formed around it. The algorithm then expands the cluster by recursively adding all neighboring points that also have at least minPts neighbors in their own neighborhood.

The result of DBSCAN is a set of clusters, each containing a group of data points that are closely packed together and separated from other clusters by areas of lower density. The algorithm is capable of detecting clusters of arbitrary shapes and sizes, and it can handle noisy and sparse datasets.

By analyzing the data using DBSCAN, it is possible to identify clusters of students who have similar preferences and needs. This information can be used to make better decisions about the design and location of student accommodation facilities.

IV. OBJECTIVE

While people migrate to a new city for various purposes, like education, job location, etc., one needs to handle the issues like a house or a place to stay, food necessities in that location, environment, and many others.

To avoid searching for a rental house manually by visiting place to place if there is properly analyzed data regarding the rental house, and food preferences with preferred location then the difficulties of an immigrant can be reduced as it is a basic necessity while migrating to a new city. This need led us to think of an idea to provide such properly analyzed clustered data for a given location which can be helpful while looking for a place to stay.

We have thought of using a specific means of clustering method to cluster this unanalysed data properly and present it to the client. In this analysis, the main problem is the proper clustering of the available data and using that clustered data to plot the data on the geolocational map according to the clusters for a better understanding.

The objective is to use the K- means and DBSCAN algorithm as it is an unsupervised learning method of Machine Learning technique.

It is relatively simple to implement and understand, guarantees convergence and mainly generalizes to clusters of different shapes and sizes.

V. PROPOSED SOLUTION

The existing system contains hostels and apartments for rent, and it has bought and sold options. It doesn't recommend accommodation in our budget. It has rare cases of rental houses on our preferences. It also doesn't recommend restaurants, gyms etc., based on users' preferences previous research lacks the accuracy of true recommendations.

The Proposed system recommends hostels, apartments as well as houses and it also displays the details of those houses, apartments and hostels. It recommends accommodation within our budget and based on preferences given. It has large cases of houses on our budget. It also recommends restaurants, gyms etc., based on users' budgets. It provides true recommendations without much lacking. We are using the K-means algorithm in this project, but it has a drawback when two circular clusters centered at the same mean have different radii. K-Means uses median values to define the cluster center and doesn't differentiate between the two clusters. It also fails when the sets are noncircular. To overcome this drawback, we use the DBSCAN Algorithm along with K-means. By using both K-means and DBSCAN, we can take advantage of the strengths of both algorithms. K-means can be used to identify initial clusters, which can then be refined using DBSCAN. This hybrid approach can help to overcome the limitations of K-means while still maintaining its efficiency, as K-means can be computationally faster than DBSCAN.

Overall, combining K-means and DBSCAN can lead to more accurate and robust clustering results, especially when dealing with complex and non-circular clusters.



1. Get Datasets from the pertinent locations (Data Collection)
2. Clean the Datasets to prepare them for analysis. (Data Cleaning via Pandas)
3. Visualize the data using boxplots. (Using Matplotlib /Seaborn /Pandas)

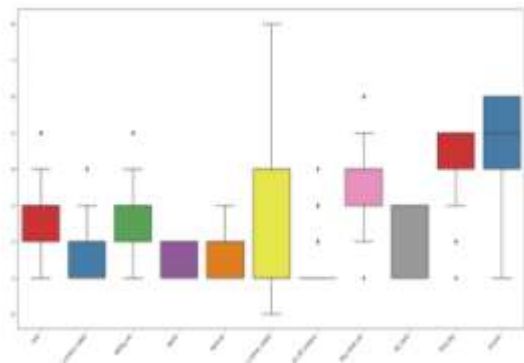
4. Fetch Geo-locational Data ((Foursquare API REST APIs)
5. Use K-Means Clustering to cluster the locations.
6. Discover the locations on the map. (Using Folium/Seaborn)

VI. TOOLS / TECHNOLOGIES USED

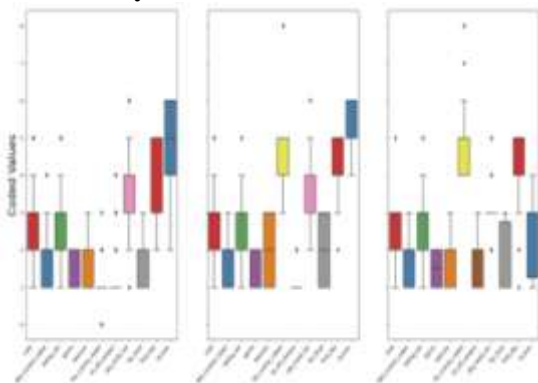
1. **Python** :- Programming Language using for the Code Implementation of Exploratory Analysis of Data.
2. **Vscode** :- An Integrated Development Environment used for implementing the entire project.
3. **FourSquare API** :- An API used to fetch geolocational data.
4. **Seaborn** :- It is used to visualize the data using boxplots.
5. **Folium** :- Used for plotting locations on the map.
6. **Pandas** :- It is used for data cleaning to prepare the data for further analysis

VII. RESULT AND ANALYSIS

1. BoxPlot Cleaned Data



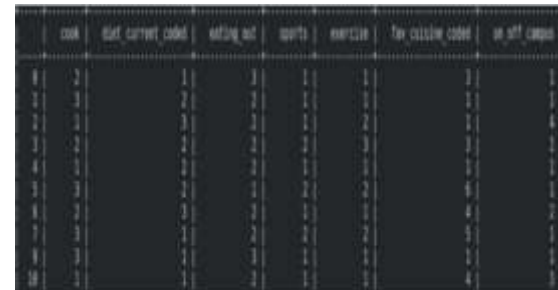
2. BoxPlot By K-Means



3. Cleaned API Data

| | position_lat | position_lng | Cafes | Stores | Gyms |
|----|--------------|--------------|-------|--------|------|
| 0 | 19.1235 | 72.9131 | 20 | 8 | 20 |
| 1 | 19.121 | 72.8854 | 5 | 8 | 20 |
| 2 | 19.1177 | 72.8837 | 6 | 11 | 20 |
| 3 | 19.1627 | 72.9329 | 6 | 8 | 9 |
| 4 | 19.096 | 72.9136 | 15 | 9 | 14 |
| 5 | 19.1665 | 72.9558 | 6 | 11 | 20 |
| 6 | 19.0781 | 72.9117 | 12 | 10 | 20 |
| 7 | 19.1719 | 72.9559 | 8 | 16 | 20 |
| 8 | 19.1628 | 72.8579 | 13 | 6 | 16 |
| 9 | 19.1237 | 72.8466 | 15 | 8 | 20 |
| 10 | 19.1462 | 72.8439 | 6 | 3 | 14 |
| 11 | 19.0963 | 72.8542 | 18 | 2 | 16 |
| 12 | 19.1019 | 72.8455 | 18 | 9 | 20 |
| 13 | 19.1409 | 72.8393 | 19 | 8 | 20 |
| 14 | 19.1409 | 72.839 | 20 | 8 | 20 |
| 15 | 19.125 | 72.9896 | 1 | 7 | 2 |
| 16 | 19.1251 | 72.9898 | 1 | 7 | 2 |
| 17 | 19.1256 | 72.9905 | 1 | 8 | 3 |
| 18 | 19.1259 | 72.9907 | 1 | 8 | 3 |
| 19 | 19.1234 | 72.9906 | 1 | 10 | 5 |

4. Cleaned Data



A screenshot of a terminal window displaying the cleaned data in a tabular format, similar to the table in section 3.

5. Clustered API Data (K-Means)

| | position_lat | position_lng | Cafes | Stores | Gyms | Cluster |
|----|--------------|--------------|-------|--------|------|---------|
| 0 | 19.1235 | 72.9131 | 20 | 8 | 20 | 2 |
| 1 | 19.121 | 72.8854 | 5 | 8 | 20 | 0 |
| 2 | 19.1177 | 72.8837 | 6 | 11 | 20 | 0 |
| 3 | 19.1627 | 72.9329 | 6 | 8 | 9 | 1 |
| 4 | 19.096 | 72.9136 | 15 | 9 | 14 | 2 |
| 5 | 19.1665 | 72.9558 | 6 | 11 | 20 | 0 |
| 6 | 19.0781 | 72.9117 | 12 | 10 | 20 | 0 |
| 7 | 19.1719 | 72.9559 | 8 | 16 | 20 | 0 |
| 8 | 19.1628 | 72.8579 | 13 | 6 | 16 | 2 |
| 9 | 19.1237 | 72.8466 | 15 | 8 | 20 | 2 |
| 10 | 19.1462 | 72.8439 | 6 | 3 | 14 | 0 |
| 11 | 19.0963 | 72.8542 | 18 | 2 | 16 | 2 |
| 12 | 19.1019 | 72.8455 | 18 | 9 | 20 | 2 |
| 13 | 19.1409 | 72.8393 | 19 | 8 | 20 | 2 |
| 14 | 19.1409 | 72.839 | 20 | 8 | 20 | 2 |
| 15 | 19.125 | 72.9896 | 1 | 7 | 2 | 1 |
| 16 | 19.1251 | 72.9898 | 1 | 7 | 2 | 1 |
| 17 | 19.1256 | 72.9905 | 1 | 8 | 3 | 1 |
| 18 | 19.1259 | 72.9907 | 1 | 8 | 3 | 1 |
| 19 | 19.1234 | 72.9906 | 1 | 10 | 5 | 1 |

6. Clustered Locations of Student Accommodations (K-Means)



7. Clustered API Data (DBSCAN)

| | position_lat | position_lng | Cafes | Stores | Spots | Cluster_dbSCAN | Cluster |
|----|--------------|--------------|-------|--------|-------|----------------|---------|
| 0 | 19.1225 | 72.9112 | 20 | 8 | 20 | 0 | 2 |
| 1 | 19.121 | 72.8954 | 5 | 8 | 20 | -1 | 0 |
| 2 | 19.1177 | 72.8817 | 0 | 11 | 20 | -1 | 0 |
| 3 | 19.1627 | 72.9138 | 6 | 8 | 19 | -1 | 1 |
| 4 | 19.096 | 72.9138 | 13 | 9 | 20 | -1 | 2 |
| 5 | 19.1683 | 72.8958 | 6 | 11 | 20 | -1 | 0 |
| 6 | 19.0701 | 72.9117 | 12 | 10 | 20 | 0 | 0 |
| 7 | 19.1729 | 72.9559 | 8 | 10 | 20 | -1 | 0 |
| 8 | 19.1628 | 72.9079 | 13 | 9 | 10 | 0 | 2 |
| 9 | 19.1217 | 72.8864 | 15 | 8 | 20 | 0 | 2 |
| 10 | 19.1462 | 72.8859 | 6 | 3 | 14 | 0 | 0 |
| 11 | 19.0961 | 72.8942 | 18 | 2 | 16 | -1 | 2 |
| 12 | 19.1619 | 72.8853 | 18 | 9 | 20 | 0 | 2 |
| 13 | 19.1489 | 72.8963 | 19 | 8 | 20 | 0 | 2 |
| 14 | 19.1499 | 72.879 | 20 | 8 | 20 | 0 | 2 |
| 15 | 19.125 | 72.8896 | 1 | 7 | 2 | -1 | 1 |
| 16 | 19.1261 | 72.8988 | 1 | 7 | 2 | -1 | 1 |
| 17 | 19.1254 | 72.8985 | 1 | 8 | 2 | -1 | 1 |
| 18 | 19.1259 | 72.8987 | 1 | 8 | 5 | -1 | 1 |
| 19 | 19.1214 | 72.9086 | 1 | 10 | 5 | -1 | 1 |

8. Clustered Locations of Student Accommodations (DBSCAN)



VIII. APPLICATIONS

The project model could help students and workers identify areas with a high concentration of accommodations that fit their budget and preferences, allowing them to make more informed decisions about where to live.

The project model can be used to analyze and predict the demand for accommodation in a specific location, which can be useful for businesses in the hospitality industry.

The clustering algorithms used in the model can also be applied to other datasets with similar features, such as restaurant or retail store locations.

IX. FUTURE SCOPE

The project model can be further refined and expanded by incorporating additional features, such as pricing data or customer reviews.

The model can be integrated with existing booking platforms to provide real-time recommendations for users based on their preferences and location.

The project can be extended to include predictive analytics for seasonal fluctuations in demand, which can help businesses optimize pricing and inventory management.

X. CONCLUSION

In conclusion, the project model aimed to develop a clustered map model that would assist immigrant students and workers in finding suitable accommodations in a new place. The project utilized several techniques and methodologies such as data mining, clustering algorithms, and Gantt charts to implement the solution effectively. The results showed that the application was successful in clustering similar accommodations based on location, price, and amenities, and it provided accurate recommendations to the users.

The project has great potential for future applications and improvements. The use of machine learning algorithms could enhance the accuracy of recommendations, and the inclusion of a feedback system could further improve the user experience. Additionally, the project could be expanded to cover more places and provide information on other aspects such as transportation and local culture. Overall, the project has the potential to greatly benefit international students and workers by providing them with a user-friendly platform to find accommodations and settle into a new environment with ease.

ACKNOWLEDGEMENT

We owe sincere thanks to our college Atharva College of Engineering for giving us a platform to prepare a project on the topic "Exploratory Analysis on Data" and would like to thank our Principal Dr. Ramesh Kulkarni for instigating within us the need for this research and giving us the opportunities and time to conduct and present research on the topic.

We are sincerely grateful for having Prof. Mahendra Patil as our guide and Prof. Suvarna Pansambal, Head of the Computer Engineering Department, for their encouragement, constant support and valuable suggestions. Moreover, the completion of this research would have been impossible without the cooperation, suggestions and help of our friends and family.

REFERENCES

- [1]. Exploratory Data Analysis Using Dimension Reduction [Tejas Nanaware , Prashant Mahajan , Ravi Chandak, Pratik Deshpande, Prof. Mahendra Patil]
- [2]. Automating Exploratory Data Analysis via Machine Learning [Tova Milo, Amit Somech]
- [3]. Visualization Methods for Exploratory Data Analysis [IEEE A.Nasser , D.Hamad , C.Sar]

- [4]. Exploratory Analysis of Geo-Locational Data - Accommodation Recommendation [M. Sumithra, A.Sai Pavithra, L.Sowmiya]
- [5]. Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means [Akhilesh Kumar Singh;Shantanu Mittal;Prashant Malhotra]
- [6]. Exploratory Data Analysis using Artificial Neural Networks by Sriram D , Kalaivani K , Ulaga Priya K , Saritha A , Sajeevram A
- [7]. Exploratory analysis of the fire statistics using automatic time series decomposition [M.M. Tatur;A.G. Ivanitskiy]