# Student Accommodation With Preference Analysis and Geo-location Data

1st Alli Khadga Jyoth
*Data Science and Engineering*
*IISER Bhopal*
*khadga19@iiserb.ac.in*
*19024*

***Abstract*—Finding suitable housing is a major issue when relocating to a new location. The place must be to their liking in order for them to have suitable lodging. In this project, we investigate the preferences of approximately 100 students and, using machine learning techniques such as k-Means clustering, attempt to find them the best possible accommodation on campus. The procedure for preference-based accommodation is as follows: • Data Collection • Data Cleaning and Visualization • Collecting Geolocation Data • K-Means Clustering • Geo visualization of the results With the data cleaned, we were able to use the Elbow method to find the optimal number of Clusters in our data, which is 8. Then we gathered the geolocation data and divided the accommodations into eight appropriate groups for our students. This project makes good use of the algorithm by grouping students based on their preferences.**

***Index Terms*—K-Means, Preference Analysis, Elbow Method, Student Preference, Exploratory Analysis, Geolocation Data, Machine Learning, ML**

## 1. Introduction

In this day and age, migrating from one place to another for study, employment, or whatever the reason might be is ordinary. When someone moves to a new place, accommodation is a real problem one has to face. One might find an excellent place to live but is not happy because it is not his liking. For one to have suitable accommodation, the place must be to his/her liking. In this project, we explore the preferences of around 100 students and, with the help of some machine learning techniques like k-Means clustering, try to find them an optimal accommodation around the Campus.

## 2. Background

The study by Mansoor and Hussain Ali [1] has shown that there is statistical evidence that student satisfaction in hostels has an impact on their academic performance. The authors also mention the five essential factors that have a significant influence on students' academic performance. These factors are accommodation, food facilities, inmate cooperation, library facilities, and safety & security. Though factors such as inmate cooperation, library facilities cannot be controlled by the student. The factors such as accommodation, food facilities, and safety & security are in the student's hands since they can choose to stay in that place or not. This poses the problem of suggesting student accommodation outside the campus based on their preferences. This project aims to solve that by taking students' preferences and applying some machine learning algorithms to find the best accommodation spots.

## 3. Materials and Methodology

The process of Preference-based accommodation is as follows:

- Data Collection
- Data Cleaning and Visualization
- Collecting Geo-location Data
- K-Means Clustering
- Geo visualization of the results

### 3.1. Data Collection

The student preference data is collected from kaggle [2]. There are 61 columns in the data, including the student's GPA and preferences.

TABLE 1. CLEANED DATA SHOWING FIRST 5 OBSERVATIONS

| cook | eating_out | employment | ethnic_food | exercise | fruit_day | income | on_off_campus |
|---|---|---|---|---|---|---|---|
| 2 | 3 | 3 | 1 | 1 | 5 | 5 | 1 |
| 3 | 2 | 2 | 4 | 1 | 4 | 4 | 1 |
| 1 | 2 | 3 | 5 | 2 | 5 | 6 | 2 |
| 2 | 2 | 3 | 5 | 3 | 4 | 6 | 1 |
| 1 | 2 | 2 | 4 | 1 | 4 | 6 | 1 |

TABLE 2. CLEANED DATA CONTINUED

| pay_meal_out | sports | veggies_day | diet_current_coded | fav_cuisine_coded | pay_meal_out | fav_food |
|---|---|---|---|---|---|---|
| 2 | 1 | 5 | 1 | 3 | 2 | 1 |
| 4 | 1 | 4 | 2 | 1 | 4 | 1 |
| 3 | 2 | 5 | 3 | 1 | 3 | 3 |
| 2 | 2 | 3 | 2 | 3 | 2 | 1 |
| 4 | 1 | 4 | 2 | 1 | 4 | 3 |

## 3.2. Data Cleaning and Visualization

Data cleaning and visualization are critical phases of data analysis. Since no data that we obtain is ever perfect and requires some cleaning before it can be used. Furthermore, Data Visualization is the process of visualizing the obtained data in plots, which helps us see the fuller picture by eliminating the slight imperfections in the data and showing the trends in the data and helps us to focus on data as a whole rather than focusing on a single data point.

**3.2.1. Data Cleaning.** No data is ever perfect, So we had to clean our data since it contains Nan values which would eventually throw up errors during our classification. Therefore, during the cleaning phase, we removed all the rows that had Nan values. This made our data uniform, and we were left with around 100 observations. see table 1,2.

**3.2.2. Visualization.** After Cleaning the data and choosing the set of preferences that we want to include in our model, we make a box plot 1 showing the stretch of our variables.

With all the cleaned data and box plots plotted, we then divided the student preferences into clusters to find the optimal accommodation based on their clusters effectively. For this, we are going to use Elbow Method on K-Means Cluster. But before that, some background on K-Means Algorithm

K-means Clustering. It is an Unsupervised learning algorithm [3] which works by an iterative process. In this method, the data set is divided/clustered into k number of predefined non-overlapping clusters or subgroups. It is making the inner points of the cluster as similar as possible while attempting to keep the clusters in distant space. It allocates data points to a cluster so that the sum of the squared distance between the cluster centroid and the data point is at a minimum. At this position, the cluster's centroid is the arithmetic mean of the data points in the clusters.

So, the classical $k$-means can be written as an optimization problem[4]:

$$\mathcal{U}^* = \underset{U \in \mathcal{P}}{\arg\min} \max_{c_k \in \chi} \sum_{k=1}^{C} \sum_{i=1}^{n} U_{k,i} ||c_k - \vec{x}_i||_d^2$$

for some distance $|| \cdot ||_d$, membership matrix U (which assigns clusters to each $\vec{x}$, $c_k$ is the kth cluster center, and $\vec{x}_j$ is the jth data point. Basically, $U_{i,k}$ is 1 if$\vec{x}_i$ is in the kth cluster and 0 otherwise. The summations basically compute the total distance of each data point to its cluster center.

Elbow Method. A curve is drawn between "within the sum of squares" (WSS) and the number of clusters in this method. The plotted curve resembles a human arm. The elbow method is so named because the point of the elbow in the curve yields the greatest number of clusters. The value of WSS changes very slowly after the elbow point in the graph or curve, so the elbow point must be considered to give the final value of the number of clusters[3].

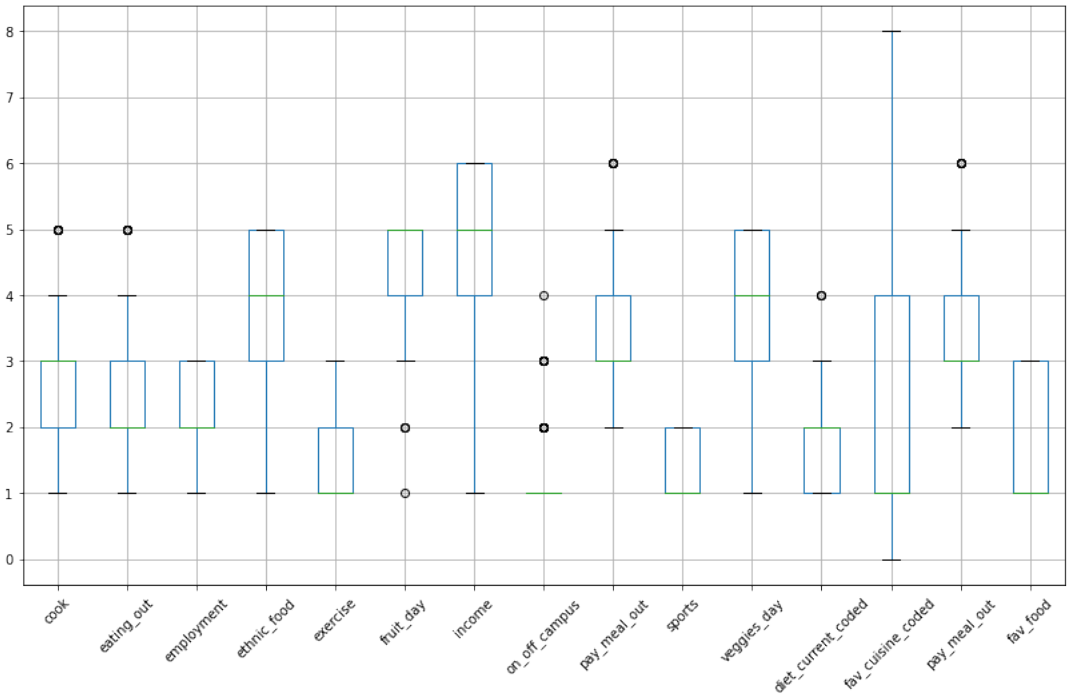


Figure 1. Box Plot of the Cleaned Data

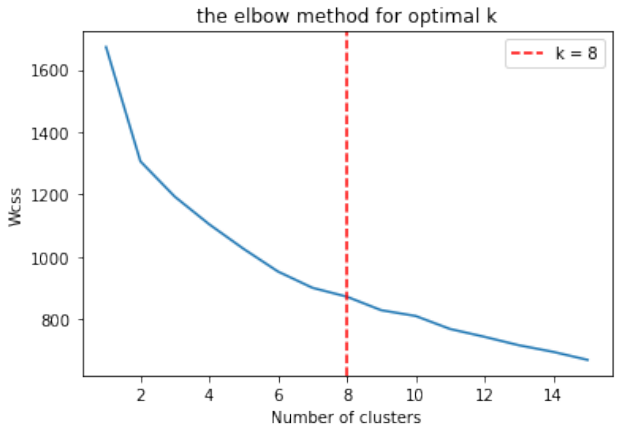


Figure 2. Elbow Method for cleaned Data

With the data cleaned, we were able to use the Elbow method to find the optimal number of Clusters in our data which is 8 see fig 2. Now, we had to extract and clean the Geo-location data to try to accommodate the students.

### 3.3. Fetching Geo-location Data

Collecting Geolocation data meant that we needed to get the latitude and longitude of the place where the campus is located, for which we used an online website [5]. After which, we used Foursquare API[6] to collect all the information regarding the place. The process of extracting the valuable information from the Geodata involved

Data Extraction and Cleaning. In this step, we search the locality of the campus for any accommodation facilities for a radius of 10 KM. This gives us locations of all the nearby apartments, hostels around the campus. These accommodation facilities are later clustered based on student preferences.

Preferences analysis. The location of these apartments is then used as a center point to find the ease of access to the student preferred locations like restaurants, gyms, stores, etc. The ease of access is measured through the number of such places available around the apartment within a certain radius.

### 3.4. Clustering the Geo-locations

Using the cleaned and processed Geo-location data, we then cluster the geo-locations of apartments into eight groups. That is the number of different groups of people from the initial analysis of student preferences.

### 3.5. Geo plotting the Results

After successfully clustering the geo-location data into 8 groups, the final step in this project is to visualize/plot the results on the map see fig 3. For this task, we used an open-source python library Folium [7].

## 4. Summary

In summary, this project involved collecting and cleaning the Student preference data. Then we extracted 15 valuable variables from the data on which we did a KMeans clustering. However, since we did not know the optimal K value, we used the Elbow method to find the optimal K for us, which turned out to be 8. After which, we had to collect the geolocation data. We first found the campus coordinates, and then we gathered details of all the nearby residential apartments. We used its geolocation data to count the number of nearby restaurants, gyms, and stores within a certain radius of the apartment. Then, we used the KMeans cluster to divide all the apartments into eight groups since that is the number of groups we got from student preferences data. Then these eight groups are plotted on a map with different colors representing different groups.

## 5. Conclusion

This project concludes that machine learning algorithms can be used in vastly different applications where the user knows what they want, i.e., has the labeled data, and when the user does not know what he wants, i.e., when they do not have any labeling for the data. The KMeans clustering is one of the unsupervised learning algorithms where it automatically groups into clusters of similar properties without the user having to look for similarities manually. This project utilizes the algorithm effectively and applies to group the students based on their preferences and also gives us a method to suggest them the accommodation based on their preferences automatically. In this project, we were successfully able to divide students into eight groups and suggest the preferred type of accommodation.

## 6. Discussion

This project only used a few variables available to cluster the students into groups. This means we are not utilizing the full potential of the data. The reason for only choosing a few of the variables was that most of the variables in the data were subjective and non-numeric types. So to incorporate the subjective preferences into our classification, we need to convert them into a numeric format. This is done to some extent in some of the variables, but we lost much of the semantic information that the data possessed. Furthermore, using K-means meant that we were only looking at the similarities of the data and not at the semantic information contained within it. So, one way to vastly improve the model is to use NLP techniques to extract the essential features of the data, code it into a numeric form, and then do an ontological clustering of the data that preserves the semantic relationship between the variables.

## References


[1] U. Mansoor and M. Hussain Ali, "Impact of hostel students' satisfaction on their academic performance in sri lankan universities," 2015.

[2] BoraPajo, *Food choices*, kaggle.com, 2017. [Online]. Available: https://www.kaggle.com/borapajo/food-choices/version/5.

[3] *K- means clustering algorithm — how it works — analysis implementation*, EDUCBA, May 2019. [Online]. Available: https://www.educba.com/k-means-clustering-algorithm/.

[4] user3658307, *Kernel k-means formula notation*, Mathematics Stack Exchange, URL:https://math.stackexchange.com/q/2319242 (version: 2017-06-12). eprint: https://math.stackexchange.com/q/2319242. [Online]. Available: https://math.stackexchange.com/q/2319242.

[5] *Gps coordinates - latitude and longitude finder*, Gps-coordinates.org, 2019. [Online]. Available: https://gps-coordinates.org/.

[6] *Find nearby places*, Foursquare Developer Documentation. [Online]. Available: https://developer.foursquare.com/reference/places-nearby.

[7] *Folium — folium 0.12.1 documentation*, Github.io, 2013. [Online]. Available: https://python-visualization.github.io/folium/ (visited on 12/05/2021).

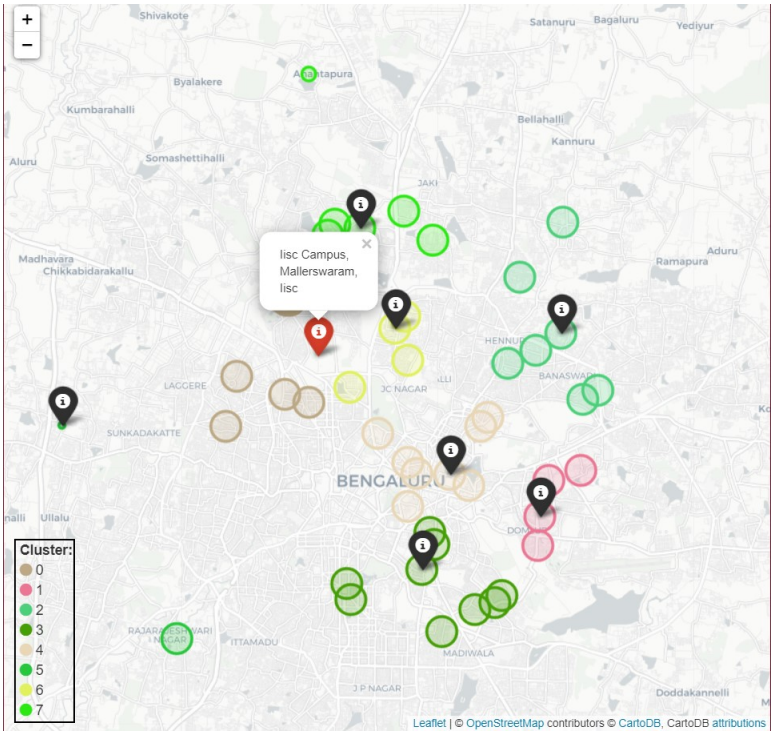


Figure 3. Plotted Result on Map