ISSN No:-2456-2165

Exploratory Analysis and Geolocation of Data to Help Student Find Housing Facilities

Sudhanshu Kadam[1], Tejas Desai[2], Rohan Dengle[3], Viraj Gawade[4]
Student, Computer Engineering, S.I.E.S Graduate School of Technology, Nerul, India [1]
Student, Computer Engineering, S.I.E.S Graduate School of Technology, Nerul, India [2]
Student, Computer Engineering, S.I.E.S Graduate School of Technology, Nerul, India [3]
Student, Computer Engineering, S.I.E.S Graduate School of Technology, Nerul, India [4]

Abstract:- Geolocation of data is a important topic in academic research. Geolocation and the use of geographic information systems have become fundamental tools in many disciplines Because they can link databases and display geographic data, geographic information systems have become essential tools in many disciplines. This project is basically to help students who relocate from various places for study purposes. This app use geolocation of data to try to provide the students best location for there stay. In this app many factors are used to try to provide the with the best location.

I. INTRODUCTION

Geolocation technology collects real-time information about various people and clusters to find the best location. This information is typically used for location monitoring as well as for analysis of location. From an operational point of view, the geolocation simply identifies the type of data and uses it to cluster to find the ideal location [13]. Traditionally students that are new to the city face a lot of problems while relocating for college or for job purposes. We have seen this in person when some of our friends relocated and all the problems that they faced. So this geolocation will help all the students to familarize themselves with the area and get the best roommates for the specific period of time. Finding a ideal place during your education is a important thing. as you may not find all the things that you need. We here are going to use kmeans clustering and try to make it easier. Many times the student don't get all the necessary things around them so we here are going to try and consider all the factors and try to provide them the best results.

II. LITERATURE SURVEY

To better understand the project, we studied different papers, videos and websites. Some of the papers that were studied are given below also there findings are stated.

A. Related Work

Implementing vector analysis in geological data to figure out exact position and better placement of geological data. The vector analysis approach works when only three points are available. There may be occasions where more measurements are available, we wish to use all the measurements at once [1].

To prove how statistical analysis can be proved very useful for geolocation of data, this paper presents a technique for evaluating the similarity between various variables [2]._In this paper we combine the largest minimum distance algorithm and the traditional K-Means algorithm to propose an improved K-Means clustering algorithm. The improved K-Means not only_keeps the high efficiency of standard K-Means but also raises the speed of convergence effectively by_improving the way of selecting initial cluster focal point [3]._Study of this paper describes the behavior of K-means algorithm. Through this paper we have try to overcome the limitations of K-means algorithm by proposed algorithm in this paper we presented an algorithm for performing K-means clustering. The experimental result demonstrated that our scheme wants to improve the direct K-means algorithm [4].

B. Survey of the Existing Application

- NoBroker [14]
- Magicbricks [15]
- 99acres [16]
- Flatchat [17]
- Nestaway [18]

All these applications just provide housing by using minimal features that are of no interest to students.

C. Components In The System

➢ *Hardware*

Hardware used is a computer system running on Windows 10 or above. RAM 4GB or above. CPU 1.8GHz processor and above.

➢ *Software*

Software requirements are Android Studio, XML, Android Phone/Emulator, Google collab, Folium.

➢ *Libraries Used*

- Pandas
- Seaborn
- Folium
- Numpy
- Matplotlib
- Geopy
- Sklearn
- Scipy

- Minisom
- Pandas.io.json
- Geopy.geocoders
- Json_normalize
- Nominatim

III. METHODOLOGY

A. Downloading the Dataset

We searched for the dataset that would be appropriate for our project as we required various fields in our project which can be further used for geolocation and appropriate plotting of area, so after downloading such dataset we will now go towards the next step.

B. Cleaning and Visualisation of Data

➢ *Data cleaning*

Cleaning the data is very important. Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include: Removal of errors when multiple sources of data are at play. Fewer errors make for happier clients and less-frustrated employees. So the actually dataset was extremely vast and consisted of data which was not required. So in data cleaning process we deleted the unwanted columns and made the dataset ready for data visualisation.[19]

➢ *Data visualisation*

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion. Data visualisation can prove to be very helpful as we can see the data beforehand and how it is and all the trends can be observed.[20]

- *Pairplot*

A pairplot plot a pairwise relationships in a dataset. The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column. That creates plots as shown below.[21]

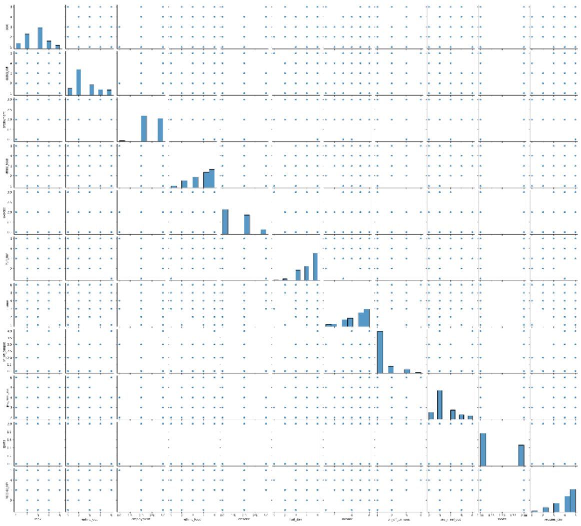


Fig 1 Pairplot

- *Boxplot*

In descriptive statistics, a **box plot** or **boxplot** is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.[22]

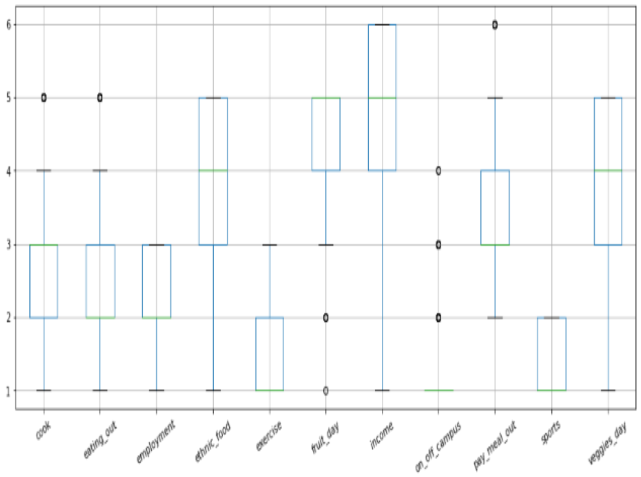


Fig 2 Boxplot

- *Implement K-means clustering on data*

K-Means Unsupervised learning process called clustering divides the unlabelled dataset into many clusters. Here, K specifies how many pre-defined clusters must be produced as part of the process. For example, if K=2, there will be two clusters; if K=10, there will be three clusters; and so on. It provides a straightforward method for categorising the groups in the unlabelled dataset on our own, without the requirement for any training. It also enables us to cluster the data into several groups. Each cluster has a centroid assigned to it because the algorithm is centroid-based. This algorithm's primary goal is to reduce the total distances between each data point and its corresponding clusters. K-means clustering is utilised in this project. Here we will be setting the max value of K to be 10 it can be anything between that.

For the first part the algorithm takes value to be 7 and plots the elbow method graph.

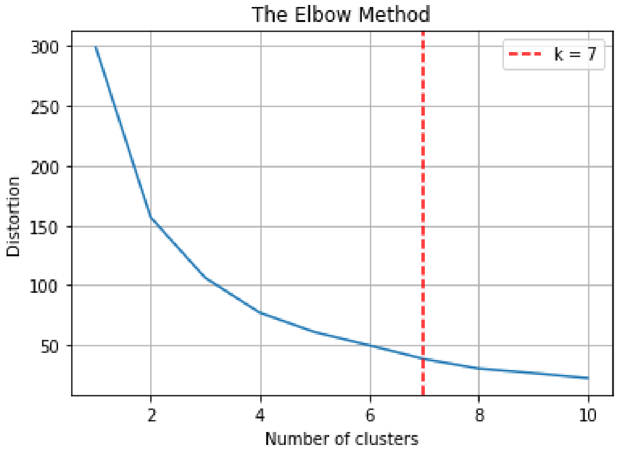


Fig 3 Elbow method when k=7

We will be adding two more columns to the dataset so that the data is more diverse and students can get the maximum benefit. After we will again run the k-means algorithm on the dataset using the optimal value of K by elbow method.

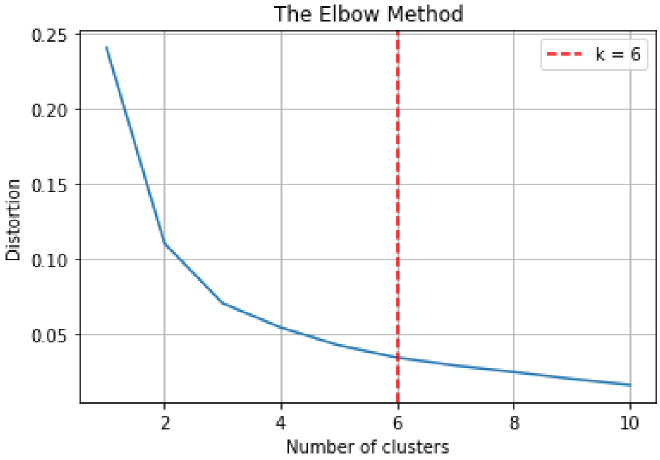


Fig 4 Elbow method k=6

- *Data cleaning*

Data cleaning process for extracting necessary data columns from the dataset. As we saw above cleaning the data is very important. Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include: Removal of errors when multiple sources of data are at play. Fewer errors make for happier clients and less-frustrated employees. So the actually dataset was extremely vast and consisted of data which was not required. So in data cleaning process we deleted the unwanted columns and made the dataset ready for data visualisation. As we have a big dataset all the factors won’t be necessary to us so what we do is we take only the necessary data columns so that we have a more accurate and precise data going into the further part of the process.

- *Extract geolocation data*

Any type of information that makes it possible to pinpoint the position of a person or an object on Earth with some degree of accuracy is referred to as geolocation data. Typically, a signal from an electronic device, such a mobile phone, connected automobile, or smart watch, is used to create this data. In general, geolocation data serves one or more of three purposes. The most typical example is identifying an object's position on Earth using longitude and latitude coordinates. A digital artefact like a picture or social media post can also get geographical information added to it. Finally, it can provide additional context for a well-known location, such as what is there and when it is accessible to the public. So once we done with data cleaning, visualization and k-means clustering we will go on to extracting the geolocation data that will be ready to be plotted on the map. In this we have used the foursquare api for extracting the geolocation data.

- *Plotting the results on the map*

Once we are done with all the processes we come down to the final part that is plotting the results on the map. Plotting accurate data on the map is very important as it proves to be valuable the more accurate the plotting is more helpful it will be. Map plotting is the process of importing data from CRM, ERP or Excel spreadsheets into a web mapping software solution. This enables you to visualize the data geographically instead of simply looking at a pile of numbers, which can provide an enhanced level of insight into your operation. Map plotting solutions allow you to create maps based on your data and supplemented with zip codes, counties, states and key demographic data. These maps can be shared as printed wall maps, image files, or as map URL’s, with other key members of your team. In the situation the maps will give the students approximate idea of what is the ideal location for there stay accordingly they can checkout housing in the given location.[23]

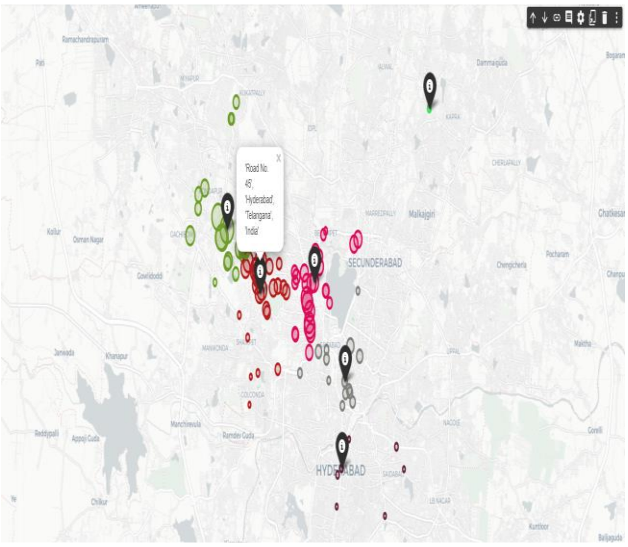


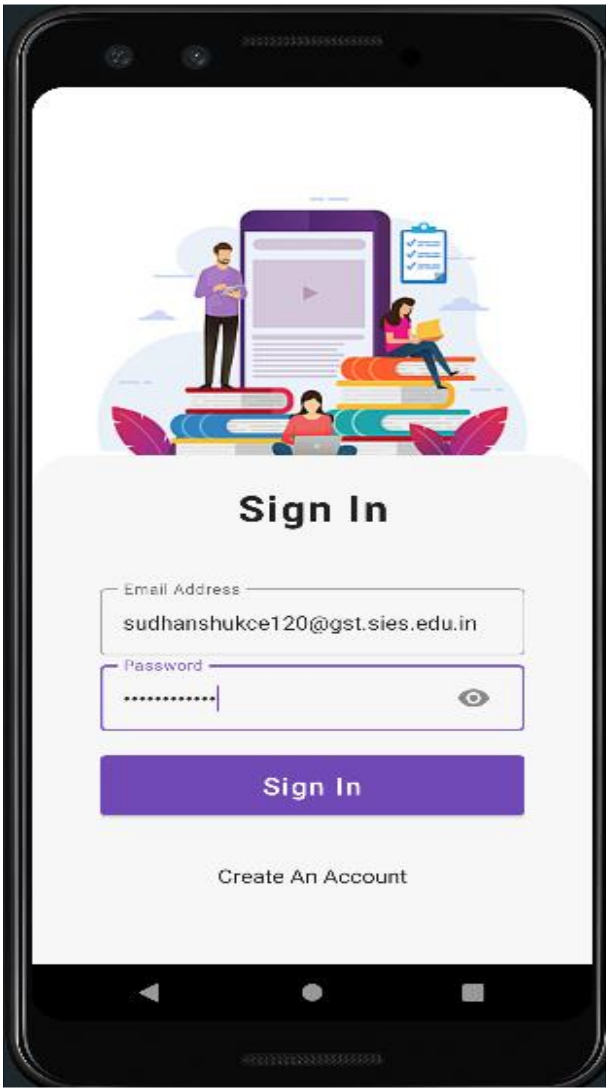
Fig 5 Plotted Map

IV. RESULTS AND CONCLUSION

In order to give this system a frontend in which people can enter the details of the kind of area they want to live in we developed a app. So in this app the person can all the details that are required for a student to find a student a good locality. The fields include GPA, Gender, Employment status, Preferred cuisine, Income/Allowance, Preferred recreational activities. Once the user enters al this details and click on "FIND ME A LOCALITY" he gets a locality found according to the choices that he entered and it will be perfect for him to live in.

Here are some of the screenshots of the app.

➢ *Login Page*

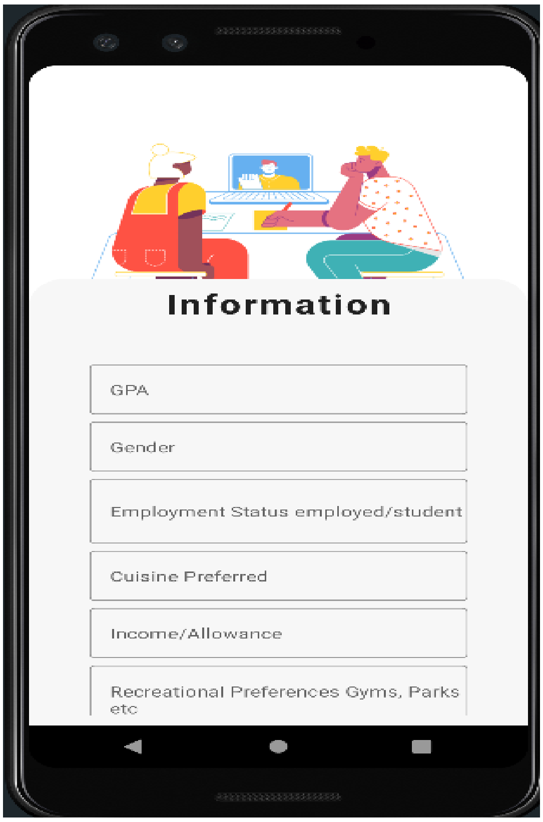

Sign In

Email Address: sudhanshukce120@gst.sies.edu.in

Password: ••••••••••••

Fig 6 Login page

➢ *Info Page*


Information

GPA
Gender
Employment Status employed/student
Cuisine Preferred
Income/Allowance
Recreational Preferences Gyms, Parks etc

Fig 7 Info page

➢ *Final Result Page*

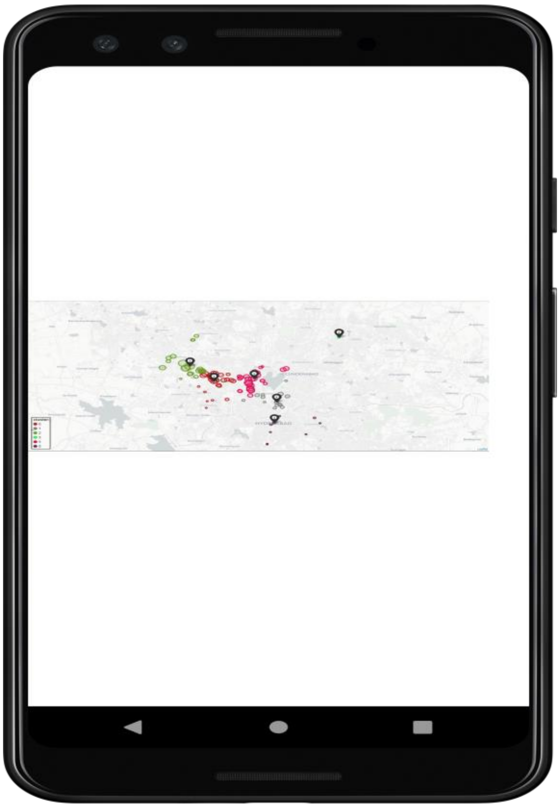


Fig 8 Result page

This paper gives a conclusive idea about how geolocation can be used for making a app that can used for helping students to find ideal location for there stay during there whole education phase. The results give us the idea of how the app works and how it can used by students for there well being. It also gives a idea of how data cleaning, data visualisation and k-means clustering works. Also how geolocation of data works. So in a few words this paper gives a brief idea about the entire project.

V. FUTURE SCOPE

Currently this system can only give you a locality according to your needs but if we collect more data we also suggest flats, rooms etc. Also this can be added to any particular college website so that students can direct take advantage of it and find there housing even before coming to the new city.

REFERENCES

[1]. Vector Analysis of Geological Data, McIntyre, D. Vector Analysis of Geologic Data. Science (New York, N.Y.).

[2]. Statistical Analysis of Geological Data, Griffiths, John. (Statistical Analysis of Geological Data. Journal of Geology.

[3]. A Clustering Method Based on K-Means Algorithm, Youguo Li, Haiyan Wu Department of Computer Science, Xinyang Agriculture College.

[4]. Analysis and Study Of K-Means Clustering Algorithm, **Sudhir Singh and Nasib Singh Gill Dept of Computer Science & Application M. D. University, Rohtak, Haryana**

[5]. Aftabizadeh, A.R., Wiener, J., Xu, J.M., 1987. Oscillatory and periodic solutions of delay differential equations with piecewise constant argument. Proceedings of the American Mathematical Society 99, 673–679. Andrews, W.K., 1984. Non-strong mixing autoregressive processes.

[6]. J. Application. Prob. 21, 930–934. Apel, J.R., 1987. Principles of Ocean Physics. Academic Press, London. Box, G.E.P.,

[7]. Jenkins, G.M., 1976. Time Series Analysis Forecasting and Control. Revised edition, San Francisco

[8]. Holden Day. Casdagli, M., 1989. Nonlinear prediction of chaotic time series. Physica D 35, 335–356. Casdagli, M., 1991. Chaos and deterministic versus stochastic non-linear modeling.

[9]. J. R. Stat. Soc. B 54, 303–328. Caviedes, C.N., 1975.

[10]. El Nino 1972: its climatic, ecological, ˜ human and economic implications. Geographical Review 65, 493–509.

[11]. Cooke, K.L., Wiener, J., 1990. A Survey of Differential Equations with Piecewise Constant Argument, Delay Differential Equations and Dynamical Systems.

[12]. Springer-Verlag, Berlin. Farmer, J.D., Sidorowich, J.J., 1987. Predicting chaotic time series. Phys. Rev. Lett. 59, 845–848

[13]. APILayer https://blog.apilayer.com/what-is-geolocation-data-where-to-get-it-and-examples/#:~:text=Free%20Geolocation%20APIs%3F-,What%20Exactly%20is%20Geolocation%20Data%3F,generated%20by%20an%20electronic%20device.

[14]. NoBroker https://www.nobroker.in/?utm_source=google&utm_medium=cpc&utm_campaign=Search_Brand_Mumbai&adgroup=Nobroker&gclid=Cj0KCQjw_r6hBhDdARIsAMIDhV8pB3yIkgioiTQAbO0vDvod6H4MW0b55qoArKsri-UaZSI5olcU-ikaAuaXEALw_wcB

[15]. Magicbricks https://www.magicbricks.com/

[16]. 99acres https://www.99acres.com/

[17]. Flatchat https://flatchat.app/

[18]. Nestaway https://www.nestaway.com/

[19]. Wikipedia https://en.wikipedia.org/wiki/Data_cleansing

[20]. Tableau https://www.tableau.com/learn/articles/data-visualization

[21]. pythonbasicsorg https://pythonbasics.org/seaborn-pairplot

[22]. Wikipedia https://en.wikipedia.org/wiki/Box_plot

[23]. Mapbusinessonline https://www.mapbusinessonline.com/