A Major Project Report On

Explanatory Analysis of Geolocational Data For Accomodation

Submitted to partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS AND COMMUNICATION ENGINEERING

By

Mr. MALVIYA DEEPAK (20911A04K8)

Mr. KEERTHIPATI JATIN VARMA (20911A04L6)

Mr. ARUNKONDA PRASHANTH (20911A04J8)

Under the Guidance of

Mrs. MOHANA LEELA

Assistant Professor

Department of Electronics and Communication Engineering

VIDYA JYOTHI INSTITUTE OF TECHNOLOGY

(An Autonomous Institution)

(Approved by AICTE, accredited by NAAC, NBA & permanently Affiliated to JNTUH, Hyderabad) Aziz Nagar Gate, C.B. Post, Hyderabad-500075

2023-2024

VIDYA JYOTHI
INSTITUTE OF TECHNOLOGY
AN AUTONOMOUS INSTITUTION

(Approved by AICTE, Accredited by NAAC, NBA & permanently Affiliated to JNTUH, Hyderabad)

Aziz Nagar Gate, C.B. Post, Hyderabad-500075

DEPARTMENT OF ELECTRONICS AD COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that the project report titled "**EXPLANATORY ANALYSIS OF GEOLOCATIONAL DATA FOR ACCOMODATION**" is being submittedby **Malviya Deepak (20911A04K8), Keerthipati Jatin Varma (20911A04L6),Arunkonda Prashanth (20911A04J8)** in partial fulfillment for the award of the Degreeof Bachelor of Technology in **Electronics and Communication Engineering,** is a record of bonafide work carried out by them under my guidance and supervision. These results embodied in this project report have not been submitted to any other University or Institute for the award of any degree or diploma.

Internal Guide

Mrs. Mohana Leela

Assistant Professor /

Head of Department

Dr. M.Rajendra Prasad

Professor

External Examiner

VIDYA JYOTHI
INSTITUTE OF TECHNOLOGY
AN AUTONOMOUS INSTITUTION

(Approved by AICTE, Accredited by NAAC, NBA & permanently Affiliated to JNTUH, Hyderabad)

Aziz Nagar Gate, C.B. Post, Hyderabad-500075

DECLARATION

We Mr. MALVIYA DEEPAK(20911A04K8)Mr. KEERTHIPATI JATIN VARMA (20911A04L6)Mr. ARUNKONDA PRASHANTH (20911A04J8) Hereby declare that the project entitled, "EXPLANATORY ANALYSIS OF GEOLOCATIONAL DATA FOR ACCOMODATION" submitted for the degree of Bachelor of Technology in Electronics and Communication Engineering is original and has beendone by us and this work is not copied and submitted anywhere for the award of any degree.

Mr. MALVIYA DEEPAK (20911A04K8)

Mr. KEERTHIPATI JATIN VARMA (20911A04L6)

Mr. ARUNKONDA PRASHANTH (20911A04J8)

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to the project guide, **Mrs. Mohan Leela – Assistant Professor**,), Vidya Jyothi Institute of Technology, Hyderabad for her timely cooperation and valuable suggestions while carrying out this work. It is her kindness that made us learn more from her.

We are grateful to **Dr GANESAN P M.E., Ph.D.,** Project Coordinator and Professor ECE Dept for his time to time, much needed valuable guidance throughout my study.

We are really grateful to **Dr. M. RAJENDRA PRASAD M.E., Ph.D.,** Professor & HOD ECE Dept,Vidya Jyothi Institute of Technology for his time to time, much needed valuable guidance throughout our study.

We express my hearted gratitude to **Dr. A. PADMAJA M.Tech., Ph.D.,** Dean (Accreditation & Ranking), Vidya Jyothi Institute of Technology for giving us spontaneous encourage for completing the project.

We thank **Dr. E. SAIBABA REDDY, M.E., Ph.D.,** Principal, Vidya Jyothi Institute of Technology for encouraging us in the completion of our project.

It is our privilege to express our gratitude and indebtedness to **Dr. PALLA RAJESHWAR REDDY, M.Sc., Ph.D.,** Secretary& Correspondent of Vidya Jyothi Institute of Technology for his moral support.

We express our heartful thanks to **Staff of Electronics and Communication Department,** Vidya Jyothi Institute of Technology helping us in carrying out our project successfully.

Mr. MALVIYA DEEPAK (20911A04K8)

Mr. KEERTHIPATI JATIN VARMA (20911A04L6)

Mr. ARUNKONDA PRASHANTH (20911A04J8)

ABSTRACT

The proliferation of geolocational data has created vast opportunities for extracting valuable insights and understanding human behaviour in relation to location. This project aims to perform an explanatory analysis of geolocational data using the K-means clustering algorithm to uncover hidden spatial patterns, identify distinct groups, and gain a deeper understanding of the underlying structure within the data.

The study will leverage a diverse range of geolocational data sources, including GPS traces, check-ins, and geotagged social media posts, along with relevant contextual information. The K-means clustering algorithm will be applied to partition the data into distinct clusters based on spatial proximity, enabling the identification of groups with similar location-based characteristics.

Keywords: Geolocational data, explanatory analysis, K-means clustering, spatial patterns, cluster identification, spatial distribution analysis, cluster profiling, temporal analysis, human behavior, insights.

INDEX

LIST OF FIGURES

CHAPTER-1

INTRODUCTION

- Analyzing geo-locational data enables research into places and regional human behaviour. Those who travel frequently may find it difficult to locate the suitable area to reside. In 2021, India accounted for 1.57% of total international tourist visits. India welcomed 17.9 million more international visitors in 2020 compared to 2019, an increase of 3.5%. India is the eighth most visited country in the Asia-Pacific region and now holds the 22nd place worldwide. It would be challenging for them to find a location to stay and enjoy their vacation because India is attracting a lot of attention from tourists. And also the people migrating to different place may find difficulties to locate an ideal place with their priorities and preferences.
- We thus recommend which would be ideal for them based on the place they choose as well as their preferences for the area. Individuals who move to a new place will likely have particular preferences and considerations, therefore analysis of geo-location is used to pinpoint the optimal locations. The situation would be hassle-free and time-saving if the consumers lived close to their preferred locations.
- The methodology can be applied to any location of one's choosing, and can vary according to user preference. To make the data points in each group more comparable to one another than those in the other groups, the data points are only separated into a number of groups. In other words, the goal is to group data items based on the characteristics they have in common.
- K-Means clustering is the best clustering technique for grouping things depending on how similar they are. K-Means clustering is an algorithm for Unsupervised Learning, which clusters an unlabeled dataset into distinct groups. The variable K represents the number of clusters that the algorithm will create. For instance, if K is set to 2, then the algorithm will create two distinct groups. This process is used to group unlabeled data into different categories without any prior training. The algorithm takes in the unlabeled dataset as input, divides it into K clusters, and repeats this process until it finds the optimal clusters.
- to go out for a good meal every once in a while, for social purposes. Either way, the food one eats is an important aspect regardless of where one lives.
- If a person moves to a new place, they already have some preferences and taste. It would save both user and the food providers a lot of benefits if they liveclose to their preferred outlets. It is convenient

for the owners and provide better sales and saves time for the user.

- This data science project explores a data engineer's daily life, covering data preparation, visualization, machine learning, and result presentation. It focuses on the importance of food preferences and location for someone new to an area, aiming to find nearby food outlets aligning with their tastes.
- The project aids convenience, customer satisfaction, and sales. Real-life datasets are collected, preprocessed, and visualized to understand outlet distributions and preferences.
- Machine learning algorithms recommend nearby outlets based on personal tastes. Results are presented in a user-friendly format, applicable to restaurant managers and hotel owners for strategic location decisions. This beginner-friendly project demonstrates data-driven decision-making's relevance in lifestyle optimization.

CHAPTER-2

LITERATURE SURVEY

- This project involves recommending hotels, gyms and other needs for the user who has accommodated to an area newly. It is difficult for a user to find all the places in a newly accommodated area. So, it is easy if we recommend nearby places. One is too tired to fix oneself a home cooked meal frequently. Even if a person gets home cooked meal every day, it is not unusual to want to go out for a good meal every once in a while for social purposes. If a person moves to a new place. They already have some preferences and taste. It would save both user and the food providers a lot of benefits if they live close to their preferred outlets. It is convenient for the owners and provide better sales and saves time for the user.
- Throughout the last few decades, researchers have examined the use of geo-location data to detect travel-related events and reasons. These research look at recurring GPS trajectories and use rules, models, and machine learning to identify personal movement patterns like home, work, shopping, and leisure. To get knowledge into the ongoing business processes, we are seeking to assess travel events as activities of various businesses in the current job.
- In recent years, there has been a noticeable surge in immigration. When these individuals arrive in the target nation, the majority of them are students who require longterm lodging. However, because he is new to the area and does not know many relevant locations, this creates a difficulty. Therefore, this research article defines an effective methodology for Accommodation Recommendation through the use of K Nearest Neighbor Clustering along with Artificial Neural Networks and Decision Making. The experimental evaluation has been performed which has proved the superiority of the presented technique.
- Exploratory analysis of geo-location data is an important aspect of gaining insights into the patterns and trends within the data. Several publications provide guidance on how to conduct exploratory analysis of geo-location data. "Exploratory Data Analysis with R" by Roger D. Peng is a widely cited book that provides an overview of exploratory data analysis techniques and their applications using the R programming language. The book includes examples of exploratory analysis on geo-location data. This book introduces the key concepts and techniques of spatial data analysis, including exploratory analysis of geo-location data. It provides a thorough explanation of how to prepare and analyze spatial data, and how to identify patterns and trends within the data.
- In order to accurately identify different groups of consumers based on behaviour, demographic, and

other factors, customer segmentation studies make use of a significant quantity of customer data. To find the ideal number of clusters, several strategies are used in customer segmentation, however each method has limitations, such as the DBSCAN algorithm failing in the case of changing density clusters. The K-means approach, on the other hand, guarantees convergence, warm-starts the centroid positions, and quickly adjusts to changes and form ideal cluster sizes. With the help of this project, marketers will be able to more effectively tailor their promotional, marketing, and product development strategies to various audience segments and encourage people to buy the product.

- The "City Tour Traveler" system, is an app for the locations URL and is based on GPS and the Internet, was recommended as a straightforward way to provide travel data and details to mobile applicants. The best way to see the city will be made possible by the Travel App's effective design, which includes precise instructions, locations, and models. The application may be used to schedule a journey for a specific period of time. It would also be helpful for people who are new with that particular location and desire to see the city swiftly.
- According to the results, by knowing the number of clusters K-Means clustering does provide a useful solution. According to our readings, the popular location-based social networks FOUR Square API and Here Geocoding & Search API are used frequently to retrieve data. As a result, we used the Here Geocoding & Search API, which assists in determining the preferred locations for the provided latitude and longitude and typically produces accurate results.

CHAPTER-3
FEASIBILITY STUDY

In order to evaluate if the project can be done in the given time frame, we are using the TEL-evaluation methods, where we cover the feasibility of the project from a technological, economical, and legal perspective. Those perspectives would help us have a broad vision of the requirements and implications related to the project. We also discuss in this section the methodology used in conducting the project.

3.1 Technological Sides

In terms of technology, conducting an explanatory analysis of geolocational data project would require suitable tools and technologies for data collection, preprocessing, spatial analysis, and visualization. This may include geographic information system (GIS) software, programming languages like Python.

3.2 Economical Side

From an economic perspective, the feasibility study would consider the costs associated with acquiring and maintaining the necessary technology and infrastructure. It would also assess the potential return on investment (ROI) or cost savings that could be achieved through the analysis. Additionally, factors such as data availability, data acquisition costs, and potential revenue generation from the project would be evaluated.

By considering both the technological and economic aspects in the feasibility study, you can determine the viability and potential benefits of the geolocational data analysis project.

CHAPTER-4

SYSTEM REQUIREMENTS

4.1 EXISTING SYSTEM

- The existing geolocational data exploration system involves data collection from GPS, smartphones, and social media.
- Preprocessing handles errors and missing values, while geospatial visualization reveals patterns on interactive maps. Spatial analysis uncovers relationships, and GIS tools aid management and analysis.
- Location-Based Services offer real-time navigation and personalized recommendations. Data privacy and security protect sensitive information. This system empowers urban planning, transportation, marketing, and disaster management, supporting data-driven decisions.
- Advancements in technology and analysis methods are expected to expand its applications further.

4.2 PROPOSED SYSTEM

- A proposed system for explanatory analysis of geolocational data might incorporate advances in machine learning and artificial intelligence. By leveraging these technologies, the system could offer more sophisticated predictive and explanatory capabilities.
- For example, it might use K-means clustering algorithms to identify hotspots or anomaly detection models to reveal unusual patterns in the data. Additionally, the system could integrate with data from various sources, such as social media or IoT devices, to provide a comprehensive understanding of the geospatial context.
- Keep in mind that the field of geospatial analysis is continually evolving, and there might have been further advancements and new systems developed beyond my last update.
- For the most up-to-date information, I recommend looking into recent research papers, academic journals, consulting experts in the field.

4.3 SYSTEM REQUIREMENTS

4.3.1 SOFTWARE REQUIREMENTS

- OS : Windows 11 and above (with any web browser)
- Technology : Python, HTML
- Modules : Pandas, Seaborn,Matplotlib, ScikitLearn,Folium
- API : REST API’s

4.3.2 HARDWARE REQUIREMENTS

- **RAM: 4GB or higher**
- Processor: Intel i5 or above
- Hard Disk: 4GB or above

4.4 REQUIREMENTS DEFINITION

After the severe continuous analysis of the problems that arose in the existing system, we are now familiar with the requirement required by the current system. The requirements that the system needs are categorized into functional and non-functional requirements. These requirements are listed below:

4.4.1 FUNCTIONAL REQUIREMENTS

Functional requirements specify which features or functions should be included in a system in order to satisfy client expectations and meet business needs. The functional requirements define the connection between the inputs and outputs, based on the premise. To obtain the output, all the operations that must be carried out on the input data must be specified. This includes defining the input and output data validity checks, the parameters impacted by the operations, and any additional operations required to convert the inputs into outputs. Functional requirements specify the behavior of the system for valid input and output.

4.4.2 NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements describe features, characteristics, and capacity of the system and they may constraints the boundaries of the proposed system.

The following are the non-functional requirements that are essential depending on the performance, cost, and control and give security efficiency and services.

Based on the above-explained non-functional prerequisites, they are as follows:

- User friendly
- The system should provide better accuracy.
- To perform efficiently with better throughput and response time

CHAPTER-5

SYSTEM DESIGN

5.1 UML Diagrams

UML diagram is designed to let developers and customers view a software system from a different perspective and in varying degrees of abstraction. UML diagrams are commonly created in visual modeling tools include. In its simplest form, a use case can be described as a specific way of using the system from a User's (actor's) perspective. A more detailed description might characterize a use case as:

- a pattern of behavior the system exhibits
- a sequence of related transactions performed by an actor and the system
- delivering something of value to the actor

Use cases provide a means to

- capture system requirements
- communicate with the end users and domain experts
- Test the system

Use cases are best discovered by examining the actors and defining what the actor will be able to do with the system. Since all the needs of a system typically cannot be covered in one use case, it is usual to have a collection of use cases. Together this use case collection specifies all the ways of using the system.

A UML system is represented using five different views that describe the system from a distinctly different perspective. Each view is defined by a set of diagrams, which is as follows.

User Model View

- This view represents the system from the user's perspective.
- The analysis representation describes a usage scenario from the end user's perspective.

Structural model view

- In this model, the data and functionality come from inside the system.
- This model view models the static structures.

Behavioral Model View

- It represents the dynamic of behavior as parts of the system, depicting the interactions of collection between various structural elements described in the user model and structural model view.

Implementation Model View

- In this, the structural and behavioral parts of the system are represented as they are to be built.

Environmental Model View

- In this, the structural and behavioral aspects of the environment in which the system is to be implemented are represented.

UML is specifically constructed through two different domains they are:

- UML Analysis modeling, which focuses on the user model and structural model views of the system.
- UML design modeling, which focuses on behavioral.

5.1.1 USE CASE DIAGRAM:

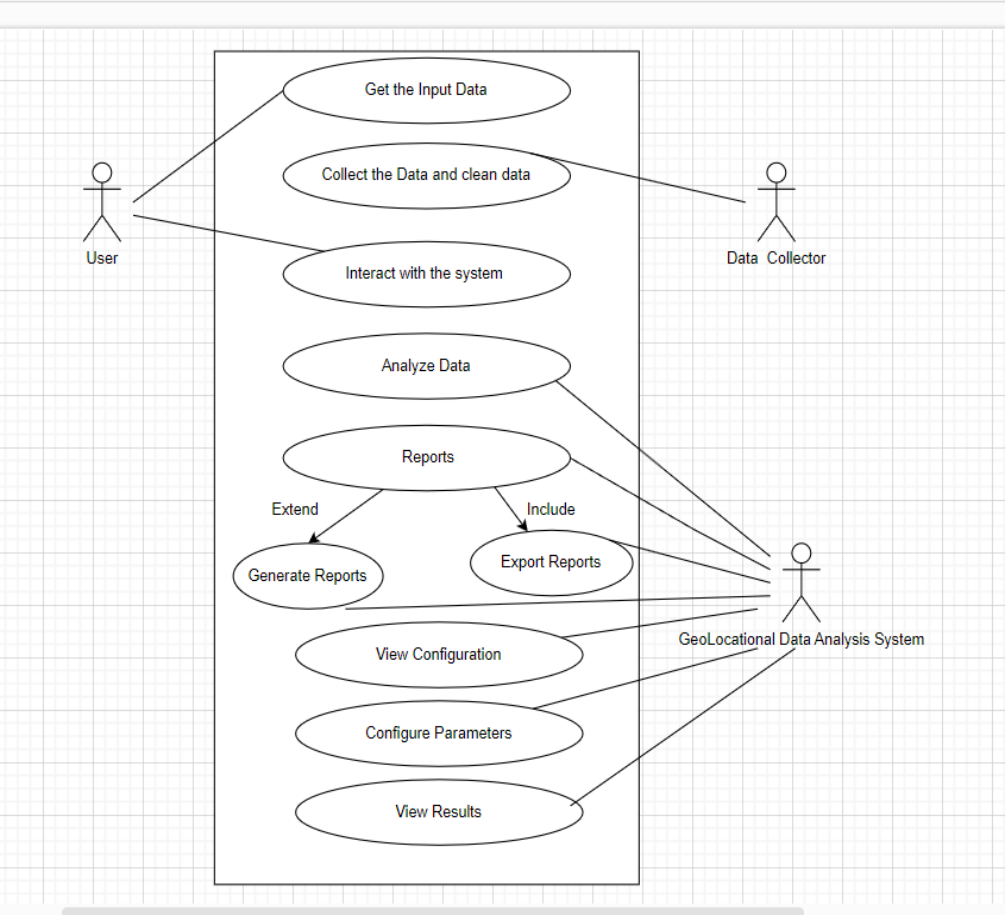


Figure. 5.1.1: Use Case Diagram

The above use case diagram represents the actors and the process in the application.

5.1.2 COMPONENT DIAGRAM:

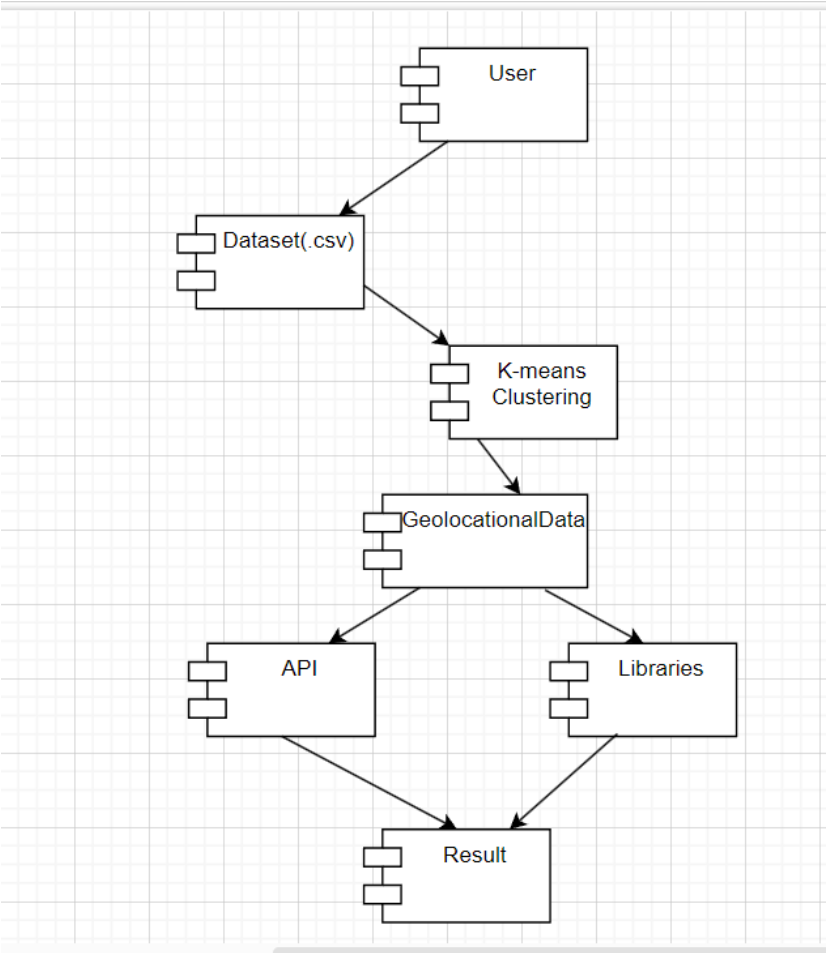


Figure. 5.1.2: Component diagram

The above component diagram describes the organization and wiring of the physical components in a system. It contains user, Dataset, Geolocational data as the main components and it shows the relationship between result, API and libraries components.

5.1.3 ACTIVITY DIAGRAM

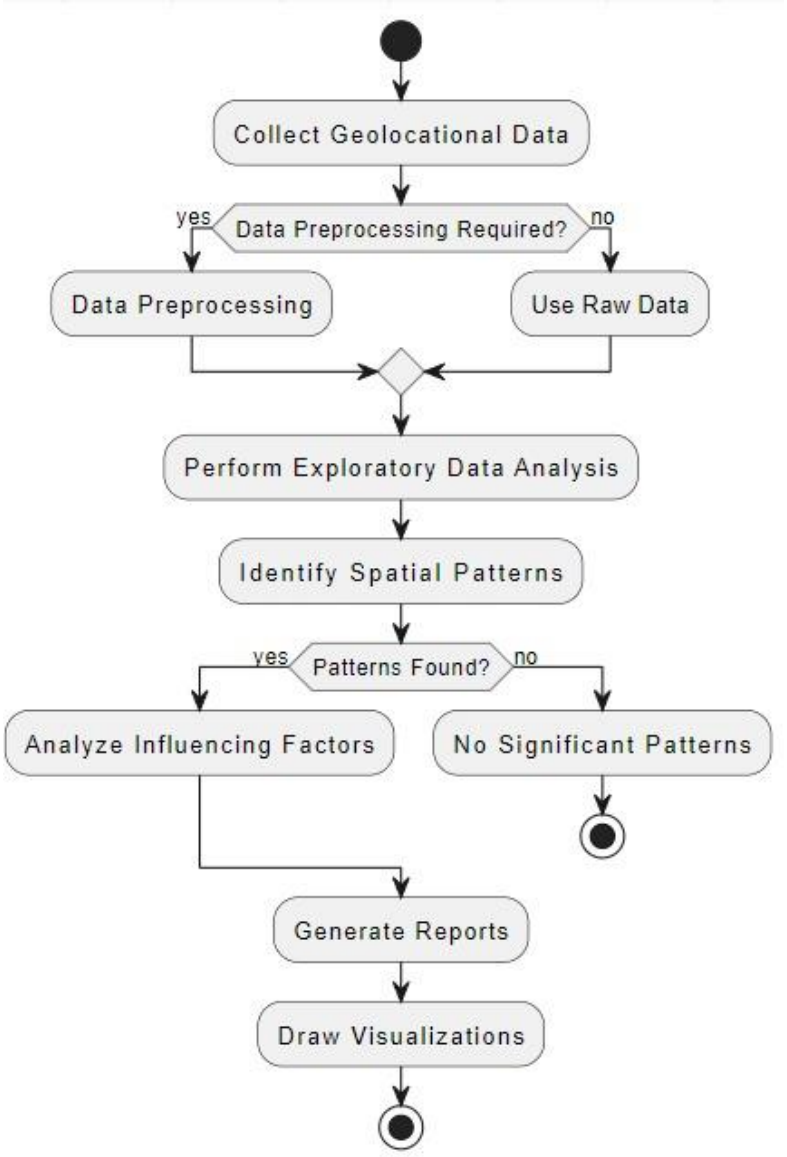


Figure. 5.1.3: Activity Diagram

The activity diagram depicts the flow from one activity to another activity. In the above activity diagram, it depicts the flow of steps involved in analyzing and interpreting the data. It includes activities like data collection,preprocessing,spatial analysis and visualization.

5.1.4 CLASS DIAGRAM:

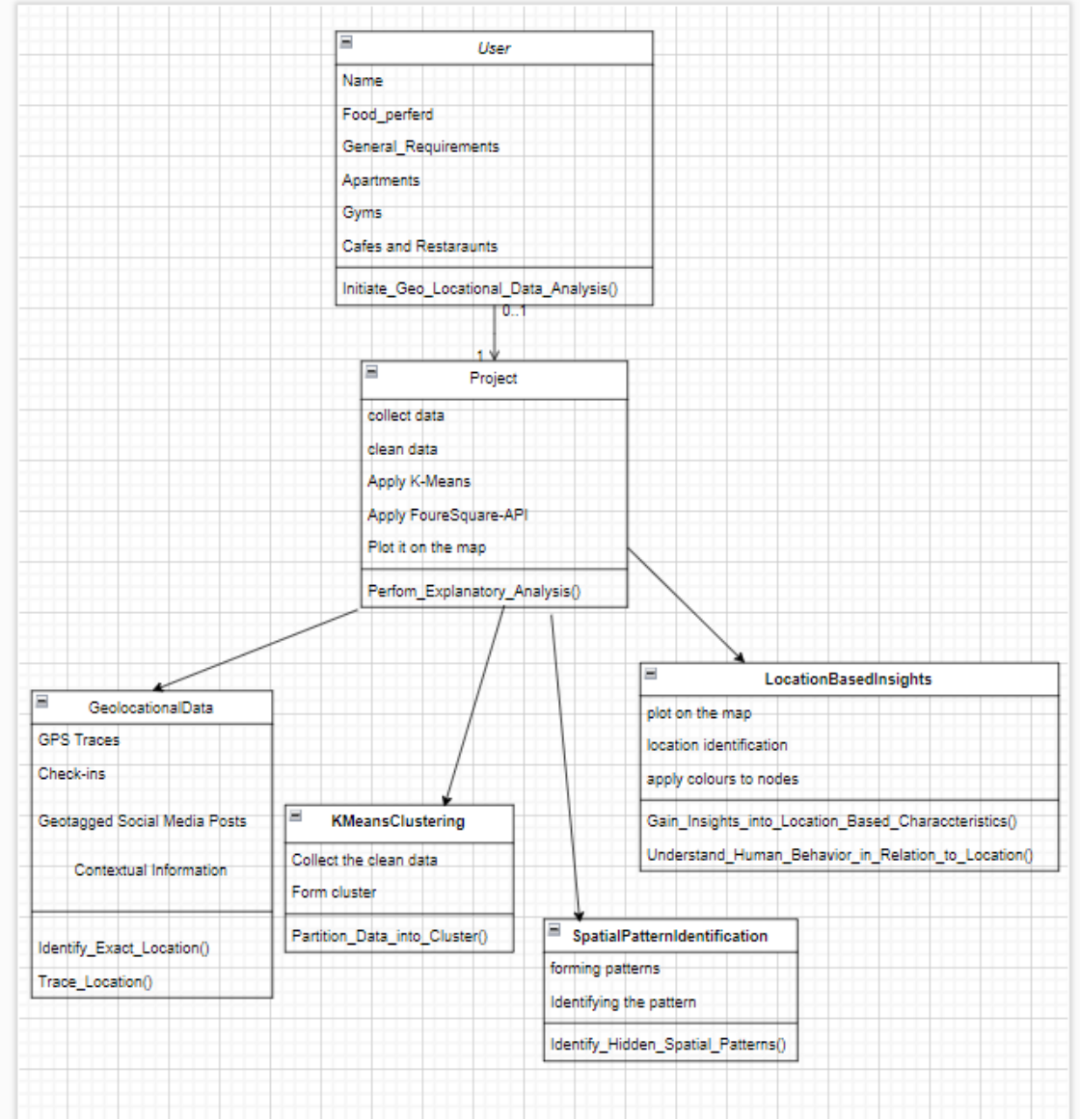


Figure. 5.1.4: Class Diagram

The class diagram represents the class, relations between the classes in the activity. Class diagram consists of classes, attributes, relations.

5.1.5 SEQUENCE DIAGRAM:

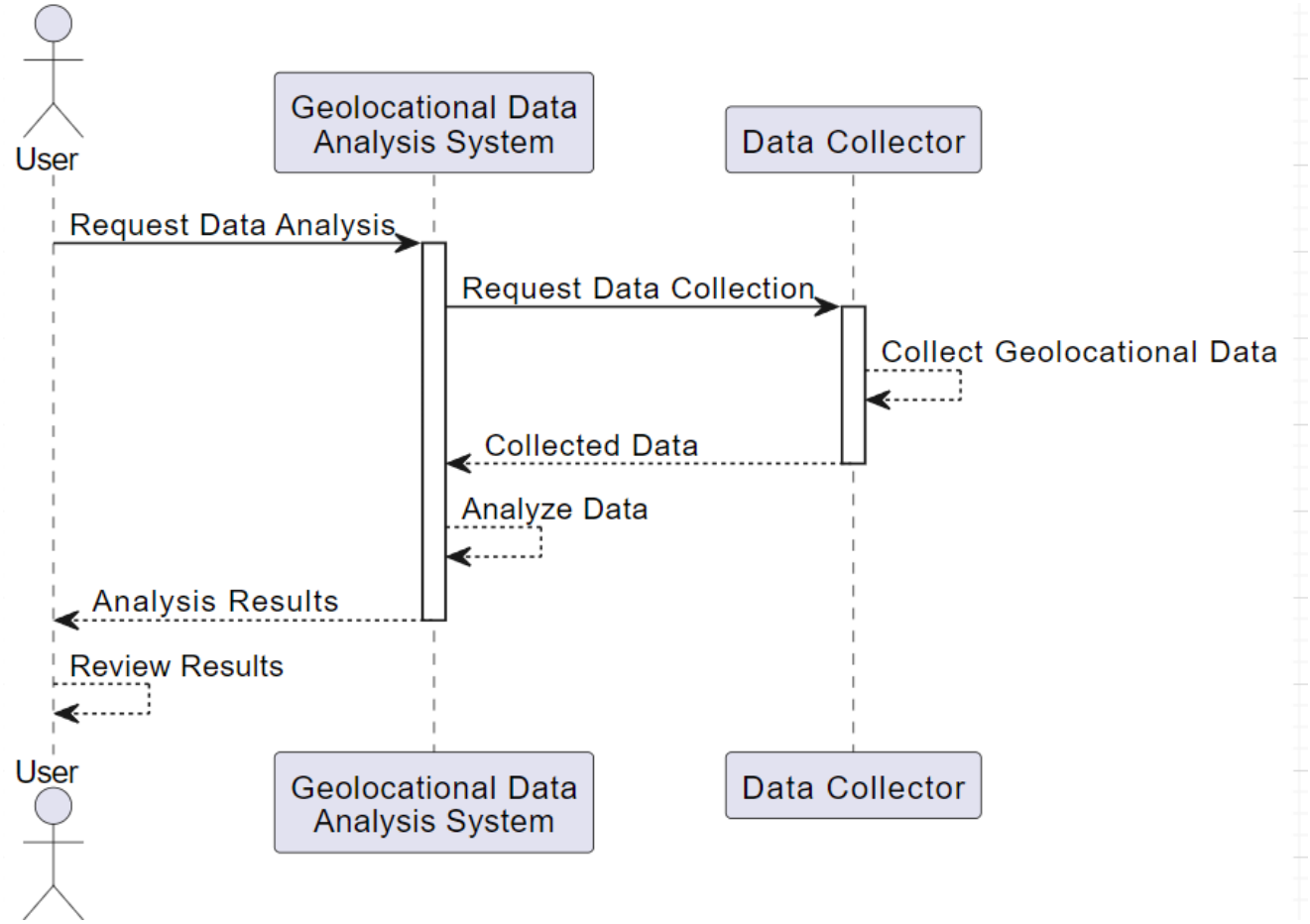


Figure. 5.1.5: Sequence Diagram

A sequence diagram is a Unified Modeling Language (UML) diagram that illustrates the sequence of messages between objects in an interaction. It consists of User,Geolocational Analysis Data System, Data Collector.

5.1.6 DATA FLOW DIAGRAM:

Figure 5.1.6: Data Flow Diagram

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the vus subprocesses the data moves through.

CHAPTER 6

SOFTWARE IMPLEMENTATION

6.1 Machine Learning

We probably know that computers don't communicate with each other the way that people do. Instead, computers require codes or directions. These binary codes and commands allow computers to process needed information. Every second, billions upon billions of ones and zeros are processed in order to provide you with the information you need.

So, what does that have to do with your ability to post your latest pictures online? Everything.

The methods by which computers communicate with each other through the use of markup languages and multimedia packages is known as **web technology**. In the past few decades, web technology has undergone a dramatic transition, from a few marked-up web pages to the ability to do very specific work on a network without interruption. Let's look at some examples of web technology.

6.1.1 Technologies used in the application

- HTML

  HTML or Hyper Text Mark-up Language is the standard mark-up language used to create web pages. HTML was created in 1991 by Tim Berners-Lee at CERN in Switzerland. It was designed to allow scientists to display and share their research.

  HTML is written in the form of HTML elements consisting of tags enclosed in angle brackets (like <html>). HTML tags most commonly come in pairs like <h1> and </h1>, although some tags represent empty elements and so are unpaired, for example <imp>. The first tag in a pair is the start tag, and the second tag is the end tag (they are also called opening tags and closing tags).

- PYTHON

  Python has a rich ecosystem of libraries and frameworks specifically designed for machine learning and artificial intelligence. Some of the most popular ones include TensorFlow, PyTorch, scikit-learn, Keras, and pandas. These libraries provide pre-built functions and tools for various machine learning tasks. Python excels at data manipulation and analysis.Libraries like NumPy, pandas, and Matplotlib are essential tools for data preprocessing, exploration,and visualization , which are

critical steps in any machine learning project.Python is widely adopted in the industry for machine learning applications. Many tech companies and research organizations use Python for developing cutting-edge AI and ML solutions.

Interface Used:

- **API**

  Being a scripting language, **JavaScript cannot run on its own. In fact, the browser is responsible for running JavaScript code**. When a user requests an HTML page with JavaScript in it, the script is sent to the browser, and it is up to the browser to execute it. The main advantage of JavaScript is that **all modern web browsers support** JavaScript. So, you do not have to worry about whether your site visitor uses Internet Explorer, Google Chrome, Firefox, or any other browser. JavaScript will be supported. Also, JavaScript **runs on any operating system** including Windows or Mac. Thus, JavaScript overcomes the main disadvantages of JavaScript (Now deprecated) which is limited to just IE and Windows.

- **FoureSquare API**

  PHP is one of the most widely used server-side scripting languages for web development. Popular websites like Facebook, Yahoo, Wikipedia and so on, and our very own application is developed using PHP.

  PHP is so popular because it's very simple to learn, code and deploy on the server, hence it has been the first choice for beginners for decades.

  In this tutorial series, we will be covering all the important concepts of PHP language from basics to advanced and will also share some ready-to-use, useful code snippets for beginners to kickstart their web development project.

6.1.2 Tools Used

Colab:

Google Colab is free to use and provides access to a GPU (Graphics Processing Unit) for faster computation. It also offers TPUs (Tensor Processing Units) for even more accelerated machine learning tasks . Colab integrates with Google Drive, allowing you to store and access your notebooks directly from your Google Drive account. It also supports importing datasets and files from your drive.Colab comes with many pre-installed Python libraries and packages commonly used in data science and machine learning, such as NumPy, pandas, Matplotlib, TensorFlow, and PyTorch.

6.1.3 Learning Machine Learning is easy

Learning machine learning can be relatively easy for those with a strong foundation in mathematics and programming. With abundant online resources, tutorials, and courses, you can grasp the basics of algorithms, data preprocessing, and model building. Libraries like TensorFlow and scikit-learn simplify implementation. However, mastering ML involves continuous practice, staying updated with evolving techniques, and understanding real-world applications. It can become challenging due to complex math, hyperparameter tuning, and data nuances. So, while the initial learning curve is manageable, achieving expertise in machine learning demands dedication and ongoing learning.

6.1.4 It's Performance

It is typically assessed using various metrics, such as accuracy, precision, recall, F1 score, and others, depending on the nature of the problem. Good machine learning performance means that the model can make accurate predictions or classifications on new, unseen data. Achieving high performance often involves optimizing the model's parameters, selecting appropriate features, and handling issues like overfitting and bias.

6.1.5 It is an Open Source

Technically, the point is that it is an open-source project, and they release patches often.

It has interfaces to a large variety of library files

PYTHON supports a large variety of library files.

Support available

Online support is available for using API.

6.2 DATASETS(.csv files)

Datasets are collections of data organized for various purposes, such as analysis, research, or machine learning.They can contain structured or unstructured data and as the foundation come in various formats, including spreadsheets, databases, text, images, and more. Datasets serve as the foundation for training and evaluating machine learning models, conducting statistical analyses, and generating insights in various domains. High-quality datasets are

essential for accurate and reliable results in data-driven applications.

6.3 FoureSquare API

The Foursquare API (Application Programming Interface) is a set of tools and services provided by Foursquare, a location-based social networking platform. It allows developers to integrate location and venue data into their applications, websites, or services. Here are some key aspects of the Foursquare API:

1. Venue Data: The API provides access to a vast database of location-specific information, including details about venues, such as restaurants, shops, and points of interest.

2. Location Data: Developers can access geospatial data, including location coordinates, place names, and user check-in information.

3. Search and Recommendations: The API enables developers to implement location-based search and recommendation features, helping users discover nearby places of interest.

4. User Interaction: Developers can create applications that allow users to check in at locations, leave reviews, and interact with Foursquare's social features.

5. Analytics: Foursquare's API also provides analytics and insights into user behavior and location trends, which can be valuable for businesses and research.

6. Licensing: Access to the Foursquare API may require developers to adhere to specific licensing terms and usage restrictions, depending on the type of application and the volume of requests.

Foursquare's API has been used in various applications, such as location-based recommendations, social check-ins, and geospatial analysis. Developers can use it to enhance their applications with location-based features and data.

6.4 Colab

Colab, short for Google Colaboratory, is a cloud-based, free-to-use platform provided by Google for machine learning and data science. It offers the following key features:

1. Jupyter Notebooks: Colab allows users to create and run Jupyter notebooks, making it easy to write and execute code in a collaborative and interactive environment.

2. Free GPU/TPU: Colab provides access to free GPU (Graphics Processing Unit) and TPU (Tensor Processing Unit) resources, which are valuable for accelerating deep learning and data analysis tasks.

3. Cloud Storage: It integrates with Google Drive, allowing users to store and access data and notebooks directly from their Google accounts.

4. Pre-installed Libraries: Colab comes with popular Python libraries like TensorFlow, PyTorch, and scikit-learn pre-installed, simplifying the setup for machine learning projects.

5. Collaboration: Multiple users can collaborate in real-time on the same notebook, making it ideal for team projects.

6. Internet Access: Colab notebooks have internet access, enabling data retrieval, web scraping, an integration with online resources.

7. Code Sharing: Users can easily share their Colab notebooks with others, fostering collaboration and knowledge sharing.

Colab is a convenient and accessible platform for machine learning and data analysis, particularly for those who don't have access to high-end hardware or want to collaborate on projects with others.

6.5 Data Analysis

Data analysis is the process of inspecting, cleaning, transforming, and interpreting data to extract valuable insights and make informed decisions. Here are key aspects of data analysis:

1. Data Collection: The process begins with gathering relevant data from various sources, which can include databases, spreadsheets, sensors, surveys, and more.

2. Data Cleaning: Data is often messy, containing errors, missing values, and inconsistencies. Data cleaning involves preprocessing to ensure data accuracy and completeness.

3. Data Exploration: Analysts explore the data to understand its structure, distribution, and relationships. This often involves using descriptive statistics, data visualization, and exploratory data analysis (EDA) techniques.

4. Data Transformation: Data may need to be transformed to make it suitable for analysis. This can include normalization, scaling, feature engineering, and creating new variables.

5. Data Analysis Techniques: Analysts apply statistical, mathematical, and computational techniques to gain insights from the data. Common techniques include regression analysis, hypothesis testing, clustering, and machine learning.

6. Visualization: Data visualization plays a crucial role in data analysis. It helps convey findings and patterns through charts, graphs, and visual representations.

7. Interpretation: Analysts draw conclusions from the data analysis, identifying trends, patterns, and relationships. This interpretation leads to actionable insights.

8. Decision Making: The insights derived from data analysis are used to inform business decisions, research, policy-making, and other areas.

8. Reporting: Results are often presented in reports, dashboards, or presentations to communicate findings to stakeholders.

Data analysis is a fundamental process in various fields, including business, science, healthcare, finance, and more. It helps organizations and individuals make data-driven decisions and solve complex problems.

6.6 Sample Code

Exploratory Analysis of Geolocational Data

Data Collection

[ ]
```
import pandas as pd
data=pd.read_csv("food_coded.csv")
```

[ ]
```
data
```
output

[ ]

Data Cleaning

The process of Extracting the features, (and dealing with different kinds of values as well as NaN values) is known as Data Cleaning.

[ ]
```
data.columns
```
output
```
Index(['GPA', 'Gender', 'breakfast', 'calories_chicken', 'calories_day',
    'calories_scone', 'coffee', 'comfort_food', 'comfort_food_reasons',
    'comfort_food_reasons_coded', 'cook', 'comfort_food_reasons_coded.1',
    'cuisine', 'diet_current', 'diet_current_coded', 'drink',
    'eating_changes', 'eating_changes_coded', 'eating_changes_coded1',
    'eating_out', 'employment', 'ethnic_food', 'exercise',
    'father_education', 'father_profession', 'fav_cuisine',
    'fav_cuisine_coded', 'fav_food', 'food_childhood', 'fries', 'fruit_day',
    'grade_level', 'greek_food', 'healthy_feeling', 'healthy_meal',
    'ideal_diet', 'ideal_diet_coded', 'income', 'indian_food',
    'italian_food', 'life_rewarding', 'marital_status',
    'meals_dinner_friend', 'mother_education', 'mother_profession',
    'nutritional_check', 'on_off_campus', 'parents_cook', 'pay_meal_out',
    'persian_food', 'self_perception_weight', 'soup', 'sports', 'thai_food',
    'tortilla_calories', 'turkey_calories', 'type_sports', 'veggies_day',
```

'vitamins', 'waffle_calories', 'weight'],
dtype='object')

[ ]
```
column=['cook','eating_out','employment','ethnic_food', 'exercise','fruit_day','income','on_off_campus','pay_meal_out','sports','veggies_day']
```

[ ]
```
d=data[column]
```

[ ]
```
d
```
output

Data Exploration and Visualisation

[ ]
```
import seaborn as sns
sns.pairplot(d)
```
output

[ ]
```
pip install matplotlib
```
output
```
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.1.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.12.0)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (4.43.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.4.5)
Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.23.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (23.2)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotlib)
(1.16.0)
```

Boxplot of Dataset

[ ]
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

```
ax=d.boxplot(figsize=(16,6))
ax.set_xticklabels(ax.get_xticklabels(),rotation=30)
```
output

[ ]
```
d.shape
```
output
(126, 11)

[ ]
```
s=d.dropna()
```

[ ]
```
pip install minisom
```
output
```
Collecting minisom
  Downloading MiniSom-2.3.1.tar.gz (10 kB)
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: minisom
  Building wheel for minisom (setup.py) ... done
  Created wheel for minisom: filename=MiniSom-2.3.1-py3-none-any.whl size=10588
sha256=60ce0400d9c3c02a55fdd2ca97bc5071290f1d76fe151dcd62e5b9f326c04ec3
  Stored in directory: /root/.cache/pip/wheels/c7/92/d2/33bbda5f86fd8830510b16aa98c8dd420129b5cb24248fd6db
Successfully built minisom
Installing collected packages: minisom
Successfully installed minisom-2.3.1
```

Run KMeans Clustering on the data

[ ]
```
## for data
import numpy as np
import pandas as pd
## for plotting
import matplotlib.pyplot as plt
import seaborn as sns
## for geospatial
import folium
import geopy
## for machine learning
from sklearn import preprocessing, cluster
import scipy
## for deep learning
import minisom
```

[ ]

```
f=['cook','income']
X = s[f]
max_k = 10
## iterations
distortions = []
for i in range(1, max_k+1):
    if len(X) >= i:
        model = cluster.KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
        model.fit(X)
        distortions.append(model.inertia_)
## best k: the lowest derivative
k = [i*100 for i in np.diff(distortions,2)].index(min([i*100 for i
    in np.diff(distortions,2)]))
## plot
fig, ax = plt.subplots()
ax.plot(range(1, len(distortions)+1), distortions)
ax.axvline(k, ls='--', color="red", label="k = "+str(k))
ax.set(title='The Elbow Method', xlabel='Number of clusters',
       ylabel="Distortion")
ax.legend()
ax.grid(True)
plt.show()
```

output

[ ]

Get Geolocational Data

[ ]

```
from pandas.io.json import json_normalize
import folium
from geopy.geocoders import Nominatim
import requests
CLIENT_ID = "KTCJJ2YZ2143QHEZ2JAQS4FJIO5DLSDO0YN4YBXPMI5NKTEF" # your Foursquare ID
CLIENT_SECRET = "KNG2LO22BPLHN1E3OAHWLYQ5PQBN14XYZMEMAS0CPJEJKOTR" # your Foursquare Secret
VERSION = '20200316'
LIMIT = 10000
```

[ ]

```
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    17.448372, 78.526957,
```

```
    30000,
    LIMIT)
```

```
[ ]
 results = requests.get(url).json()
```

```
[ ]
 results
output
{'meta': {'code': 200, 'requestId': '651f9be9670b2730dc39dcc4'},
 'response': {'queryRefinements': {'target': {'type': 'path',
    'url': '/venue/explore',
    'params': {'ll': '17.448372,78.526957', 'radius': '30000'}},
   'refinements': [{'query': 'Food'},
    {'query': 'Nightlife'},
    {'query': 'Coffee'},
    {'query': 'Shops'},
    {'query': 'Arts'},
    {'query': 'Outdoors'}]},
  'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Hyderabad',
  'headerFullLocation': 'Hyderabad',
  'headerLocationGranularity': 'city',
  'totalResults': 115,
  'suggestedBounds': {'ne': {'lat': 17.71837227000027,
    'lng': 78.80945141641452},
   'sw': {'lat': 17.17837172999973, 'lng': 78.24446258358547}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4c1f7229b306c928046b68b7',
       'name': 'Fifth Avenue Bakers',
       'location': {'address': 'Sainikpuri',
        'lat': 17.487673346331192,
        'lng': 78.54279287522861,
        'labeledLatLngs': [{'label': 'display',
          'lat': 17.487673346331192,
          'lng': 78.54279287522861}],
        'distance': 4687,
        'cc': 'IN',
        'state': 'Andhra Pradesh',
        'country': 'India',
        'formattedAddress': ['Sainikpuri', 'Andhra Pradesh', 'India']},
       'categories': [{'id': '4bf58dd8d48988d16a941735',
         'name': 'Bakery',
         'pluralName': 'Bakeries',
         'shortName': 'Bakery',
         'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/bakery_',
          'suffix': '.png'},
         'primary': True}],
```

'photos': {'count': 0, 'groups': []},
'createdAt': 1277129257},
'photo': {'id': '5410601c11d2a80070a4d902',
'createdAt': 1410359324,
'prefix': 'https://fastly.4sqi.net/img/general/',
'suffix': '/58807053_Az0izB4nYIHej7aiec35wqyHdb71ssAD5yHmESElVBk.jpg',
'width': 540,
'height': 960,
'visibility': 'public'},
'referralId': 'e-0-4c1f7229b306c928046b68b7-0'},
{'reasons': {'count': 0,
'items': [{'summary': 'This spot is popular',
'type': 'general',
'reasonName': 'globalInteractionReason'}]},
'venue': {'id': '55e9d8dc498e8a5c51f30331',
'name': 'Cinepolis CCPL',
'location': {'address': 'Malkajgiri',
'lat': 17.45728215392065,
'lng': 78.53682348691787,
'labeledLatLngs': [{'label': 'display',
'lat': 17.45728215392065,
'lng': 78.53682348691787}],
'distance': 1442,
'cc': 'IN',
'city': 'Hyderabad',
'state': 'Telangana',
'country': 'India',
'formattedAddress': ['Malkajgiri', 'Hyderabad', 'Telangana', 'India']},
'categories': [{'id': '4bf58dd8d48988d17f941735',
'name': 'Movie Theater',
'pluralName': 'Movie Theaters',
'shortName': 'Movie Theater',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/movietheater_',
'suffix': '.png'},
'primary': True}],
'photos': {'count': 0, 'groups': []},
'createdAt': 1441388764},
'photo': {'id': '5cc280766fd626002c9ffaa9',
'createdAt': 1556250742,
'prefix': 'https://fastly.4sqi.net/img/general/',
'suffix': '/65453630_Ei2r1Z2hekZIgq6qRE9IShh5NV-q_cZFhdoy8_sxnOs.jpg',
'width': 1900,
'height': 897,
'visibility': 'public'},
'referralId': 'e-0-55e9d8dc498e8a5c51f30331-1'},
{'reasons': {'count': 0,
'items': [{'summary': 'This spot is popular',
'type': 'general',
'reasonName': 'globalInteractionReason'}]},
'venue': {'id': '4e5b27778877c8a76dfc0f0c',
'name': 'Vivanta by Taj',
'location': {'address': 'Vivanta By Taj - Begumpet, Hyderabad',
'crossStreet': '1-10-147 & 148, Mayuri Marg',
'lat': 17.44349840742226,
'lng': 78.4605560466047,
'labeledLatLngs': [{'label': 'display',
'lat': 17.44349840742226,

'shortName': 'Afghan',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/afghan_',
'suffix': '.png'},
'primary': True}],
'photos': {'count': 0, 'groups': []},
'createdAt': 1365782335},
'photo': {'id': '5adc5f8695a7227f2ca4821c',
'createdAt': 1524391814,
'prefix': 'https://fastly.4sqi.net/img/general/',
'suffix': '/26232825_ZQoPKWHXojunPbMvgLaHcPZ0v9U4hfRIbH0hi9KDu4w.jpg',
'width': 988,
'height': 1900,
'visibility': 'public'},
'referralId': 'e-0-51682f3fe4b0c86be4c2d508-97'},
{'reasons': {'count': 0,
'items': [{'summary': 'This spot is popular',
'type': 'general',
'reasonName': 'globalInteractionReason'}]},
'venue': {'id': '4ff37a1ee4b01d081ec95995',
'name': 'Karachi Bakery',
'location': {'address': 'Domestic Depatures',
'crossStreet': 'Shamshabad International Airport',
'lat': 17.235138042504662,
'lng': 78.43028778701525,
'labeledLatLngs': [{'label': 'display',
'lat': 17.235138042504662,
'lng': 78.43028778701525}],
'distance': 25864,
'cc': 'IN',
'city': 'Hyderabad',
'state': 'Telangana',
'country': 'India',
'formattedAddress': ['Domestic Depatures (Shamshabad International Airport)',
'Hyderabad',
'Telangana',
'India']},
'categories': [{'id': '4bf58dd8d48988d16a941735',
'name': 'Bakery',
'pluralName': 'Bakeries',
'shortName': 'Bakery',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/bakery_',
'suffix': '.png'},
'primary': True}],
'photos': {'count': 0, 'groups': []},
'createdAt': 1341356574},
'photo': {'id': '514dad27e4b02f9bf8a7d0e2',
'createdAt': 1364045095,
'prefix': 'https://fastly.4sqi.net/img/general/',
'suffix': '/812503_W2o7ZvYwqKg4KVq-gzwmbhNStiehsztZfwpuOEra1Ig.jpg',
'width': 717,
'height': 959,
'visibility': 'public'},
'referralId': 'e-0-4ff37a1ee4b01d081ec95995-98'},
{'reasons': {'count': 0,
'items': [{'summary': 'This spot is popular',
'type': 'general',
'reasonName': 'globalInteractionReason'}]},

'venue': {'id': '51f2a57b498ebcf535947605',
'name': 'Yum Yum Tree Arabian',
'location': {'lat': 17.309767,
'lng': 78.47579,
'labeledLatLngs': [{'label': 'display',
'lat': 17.309767,
'lng': 78.47579}],
'distance': 16358,
'cc': 'IN',
'country': 'India',
'formattedAddress': ['India']},
'categories': [{'id': '4bf58dd8d48988d1c0941735',
'name': 'Mediterranean Restaurant',
'pluralName': 'Mediterranean Restaurants',
'shortName': 'Mediterranean',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/mediterranean_',
'suffix': '.png'},
'primary': True}],
'photos': {'count': 0, 'groups': []},
'createdAt': 1374856571},
'photo': {'id': '555b9c84498eacdb455758c1',
'createdAt': 1432067204,
'prefix': 'https://fastly.4sqi.net/img/general/',
'suffix': '/65453630_NzxmwKe6Sx-jadef8b6RS8f8Z7kc1Caa7VAC_IHDA8M.jpg',
'width': 720,
'height': 960,
'visibility': 'public'},
'referralId': 'e-0-51f2a57b498ebcf535947605-99'}]}]}}

[ ]

```
venues = results['response']['groups'][0]['items']
nearby_venues = json_normalize(venues)
```

output

```
<ipython-input-24-745f8c248222>:2: FutureWarning: pandas.io.json.json_normalize is deprecated, use pandas.json_normalize instead.
nearby_venues = json_normalize(venues)
```

[ ]

```
nearby_venues
```

output

[ ]

[ ]

[ ]

```
nearby_venues
```

output

Adding two more Columns Restaurant and Others

1. Restaurant: Number of Restaurant in the radius of 20 km
2. others:Number of Gyms, Parks,etc in the radius of 20 km

[ ]

```
resta=[]
oth=[]
for lat,long in zip(nearby_venues['venue.location.lat'],nearby_venues['venue.location.lng']):
    url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
     CLIENT_ID,
     CLIENT_SECRET,
     VERSION,
     lat,long,
     1000,
     100)
    res = requests.get(url).json()
    venue = res['response']['groups'][0]['items']
    nearby_venue = json_normalize(venue)
    df=nearby_venue['venue.categories']

    g=[]
    for i in range(0,df.size):
     g.append(df[i][0]['icon']['prefix'].find('food'))
    co=0
    for i in g:
     if i>1:
       co+=1
    resta.append(co)
    oth.append(len(g)-co)

nearby_venues['restaurant']=resta
nearby_venues['others']=oth
nearby_venues
```

output

[ ]

Changing the Column Name

[ ]

```
lat=nearby_venues['venue.location.lat']
long=nearby_venues['venue.location.lng']
```

Install the minisom library using pip

MiniSom is a minimalistic and Numpy based implementation of the Self Organizing Maps (SOM). SOM is a type of Artificial Neural Network able to convert complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. Minisom is designed to allow researchers to easily build on top of it and to give students the ability to quickly grasp its details.

[ ]

```
 pip install minisom
```

```
account_circle
Requirement already satisfied: minisom in /usr/local/lib/python3.10/dist-packages (2.3.1)
```

Run K Means clustering on the dataset, with the optimal K value using Elbow Method

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

[ ]

```
 f=['venue.location.lat','venue.location.lng']
 X = nearby_venues[f]
 max_k = 10
 ## iterations
 distortions = []
 for i in range(1, max_k+1):
    if len(X) >= i:
       model = cluster.KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
       model.fit(X)
       distortions.append(model.inertia_)
 ## best k: the lowest derivative
 k = [i*100 for i in np.diff(distortions,2)].index(min([i*100 for i
     in np.diff(distortions,2)]))
 ## plot
 fig, ax = plt.subplots()
 ax.plot(range(1, len(distortions)+1), distortions)
 ax.axvline(k, ls='--', color="red", label="k = "+str(k))
 ax.set(title='The Elbow Method', xlabel='Number of clusters',
       ylabel="Distortion")
 ax.legend()
 ax.grid(True)
 plt.show()
```

```
account_circle
```

[ ]

```
city = "Hyderabad"
## get location
locator = geopy.geocoders.Nominatim(user_agent="MyCoder")
location = locator.geocode(city)
print(location)
## keep latitude and longitude only
location = [location.latitude, location.longitude]
print("[lat, long]:", location)
```
```
account_circle
Hyderabad, Bahadurpura mandal, Hyderabad District, Telangana, India
[lat, long]: [17.360589, 78.4740613]
```

[ ]
```
nearby_venues.head()
```
```
account_circle
```

[ ]
```
nearby_venues.columns
```
```
account_circle
Index(['referralId', 'reasons.count', 'reasons.items', 'venue.id',
       'venue.name', 'venue.location.address', 'venue.location.lat',
       'venue.location.lng', 'venue.location.labeledLatLngs',
       'venue.location.distance', 'venue.location.cc', 'venue.location.state',
       'venue.location.country', 'venue.location.formattedAddress',
       'venue.categories', 'venue.photos.count', 'venue.photos.groups',
       'venue.createdAt', 'venue.location.crossStreet', 'venue.location.city',
       'venue.location.postalCode', 'venue.venuePage.id',
       'venue.location.neighborhood', 'restaurant', 'others'],
      dtype='object')
```

Data Cleaning Process for Extracting Necessary Columns in the Dataset

---

[ ]
```
n=nearby_venues.drop(['referralId', 'reasons.count', 'reasons.items', 'venue.id',
       'venue.name',
       'venue.location.labeledLatLngs', 'venue.location.distance',
       'venue.location.cc',
       'venue.categories', 'venue.photos.count', 'venue.photos.groups',
       'venue.location.crossStreet', 'venue.location.address','venue.location.city',
       'venue.location.state', 'venue.location.crossStreet',
       'venue.location.neighborhood',   'venue.venuePage.id',
       'venue.location.postalCode','venue.location.country'],axis=1)
```

[ ]
```
n.columns
```
```
account_circle
```

```
Index(['venue.location.lat', 'venue.location.lng',
       'venue.location.formattedAddress', 'venue.createdAt', 'restaurant',
       'others'],
      dtype='object')
```

New Dataset

```
[ ]
 n
account_circle
```

```
[ ]
```

Dropping Nan Values from Dataset

```
[ ]
 n=n.dropna()
 n = n.rename(columns={'venue.location.lat': 'lat', 'venue.location.lng': 'long'})
 n
account_circle
```

Convert Every Row of Column ***'venue.location.formattedAddress'*** from List to String

---

```
[ ]
 n['venue.location.formattedAddress']
account_circle
0                [Sainikpuri, Andhra Pradesh, India]
1     [Ground Floor, Lifestyle Building, Greenlands ...
2     [Opposite Hussain Sagar Lake (Tank Bund Road),...
3     [Vivanta By Taj - Begumpet, Hyderabad (1-10-14...
4           [Malkajgiri, Hyderabad, Telangana, India]
                      ...
95                            [Barakas, India]
96    [Domestic Depatures (Shamshabad International ...
97                                  [India]
98    [Rajiv Gandhi International Airport, Hyderabad...
99    [RGIA Hyderabad, Hyderabad 500409, Telangana, ...
Name: venue.location.formattedAddress, Length: 100, dtype: object
```

```
[ ]
 spec_chars = ["[","]"]
 for char in spec_chars:
   n['venue.location.formattedAddress'] = n['venue.location.formattedAddress'].astype(str).str.replace(char, ' ')
```

account_circle

```
<ipython-input-41-804ee3019677>:3: FutureWarning: The default value of regex will change from True to False in a future version.
In addition, single character regular expressions will *not* be treated as literal strings when regex=True.
  n['venue.location.formattedAddress'] = n['venue.location.formattedAddress'].astype(str).str.replace(char, ' ')
```

[ ]

```
n
```

account_circle

Plot the clustered locations on a map

[ ]

```
X = n[["lat","long"]]
max_k = 10
## iterations
distortions = []
for i in range(1, max_k+1):
    if len(X) >= i:
        model = cluster.KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
        model.fit(X)
        distortions.append(model.inertia_)
## best k: the lowest derivative
k = [i*100 for i in np.diff(distortions,2)].index(min([i*100 for i in np.diff(distortions,2)]))
## plot
fig, ax = plt.subplots()
ax.plot(range(1, len(distortions)+1), distortions)
ax.axvline(k, ls='--', color="red", label="k = "+str(k))
ax.set(title='The Elbow Method', xlabel='Number of clusters',
       ylabel="Distortion")
ax.legend()
ax.grid(True)
plt.show()
```

account_circle

[ ]

```
k = 6
model = cluster.KMeans(n_clusters=k, init='k-means++')
X = n[["lat","long"]]
## clustering
dtf_X = X.copy()
dtf_X["cluster"] = model.fit_predict(X)
## find real centroids
closest, distances = scipy.cluster.vq.vq(model.cluster_centers_,
                     dtf_X.drop("cluster", axis=1).values)
```

```
 dtf_X["centroids"] = 0
 for i in closest:
    dtf_X["centroids"].iloc[i] = 1
 ## add clustering info to the original dataset
 n[["cluster","centroids"]] = dtf_X[["cluster","centroids"]]
 n
account_circle
```

```
[ ]
 ## plot
 fig, ax = plt.subplots()
 sns.scatterplot(x="lat", y="long", data=n,
                 palette=sns.color_palette("bright",k),
                 hue='cluster', size="centroids", size_order=[1,0],
                 legend="brief", ax=ax).set_title('Clustering (k='+str(k)+')')
 th_centroids = model.cluster_centers_
 ax.scatter(th_centroids[:,0], th_centroids[:,1], s=50, c='black',
           marker="x")
account_circle
```

```
[ ]
 model = cluster.AffinityPropagation()
```

```
[ ]
 k = n["cluster"].nunique()
 sns.scatterplot(x="lat", y="long", data=n,
                 palette=sns.color_palette("bright",k),
                 hue='cluster', size="centroids", size_order=[1,0],
                 legend="brief").set_title('Clustering (k='+str(k)+')')
account_circle
```

```
[ ]
x, y = "lat", "long"
color = "cluster"
size = "restaurant"
popup = "venue.location.formattedAddress"
marker = "centroids"
data = n.copy()
## create color column
lst_elements = sorted(list(n[color].unique()))
lst_colors = ['#%06X' % np.random.randint(0, 0xFFFFFF) for i in
              range(len(lst_elements))]
```

```
data["color"] = data[color].apply(lambda x:
                lst_colors[lst_elements.index(x)])
## create size column (scaled)
scaler = preprocessing.MinMaxScaler(feature_range=(3,15))
data["size"] = scaler.fit_transform(
               data[size].values.reshape(-1,1)).reshape(-1)
## initialize the map with the starting location
map_ = folium.Map(location=location, tiles="cartodbpositron",
                  zoom_start=11)
## add points
data.apply(lambda row: folium.CircleMarker(
           location=[row[x],row[y]],
           color=row["color"], fill=True,popup=row[popup],
           radius=row["size"]).add_to(map_), axis=1)
## add html legend
legend_html = """<div style="position:fixed; bottom:10px; left:10px; border:2px solid blac
k; z-index:9999; font-size:14px;"> <b>"""+color+""":</b><br>"""
for i in lst_elements:
     legend_html = legend_html+""" <i class="fa fa-circle
     fa-1x" style="color:"""+lst_colors[lst_elements.index(i)]+""""">
     </i> """+str(i)+"""<br>"""
legend_html = legend_html+"""</div>"""
map_.get_root().html.add_child(folium.Element(legend_html))
## add centroids marker
lst_elements = sorted(list(n[marker].unique()))
data[data[marker]==1].apply(lambda row:
           folium.Marker(location=[row[x],row[y]],
           draggable=False,  popup=row[popup] ,
           icon=folium.Icon(color="black")).add_to(map_), axis=1)
## plot the map
map_
```

CHAPTER-7

SYSTEM ANALYSIS

The purpose of system analysis is to study and understand an existing or proposed system to identify its problems, needs, and opportunities, with the goal of designing and implementing effective and efficient solutions. It involves evaluating the current state of a system, defining requirements, and planning improvements or new system development. System analysis helps organizations optimize processes, enhance productivity, and meet their objectives.

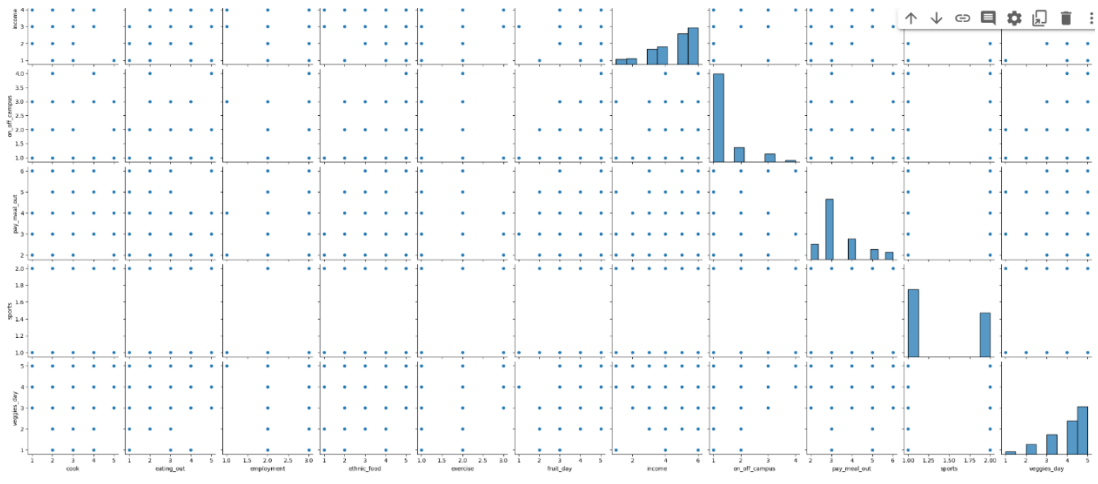
7.1 TYPES OF ANALYSIS

7.1.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a data analysis approach that involves visually and statistically examining a dataset to understand its characteristics, uncover patterns, relationships, and anomalies, and gain initial insights. EDA helps in data preprocessing, feature selection, and informing the design of machine learning models. It is a crucial first step in the data analysis process.

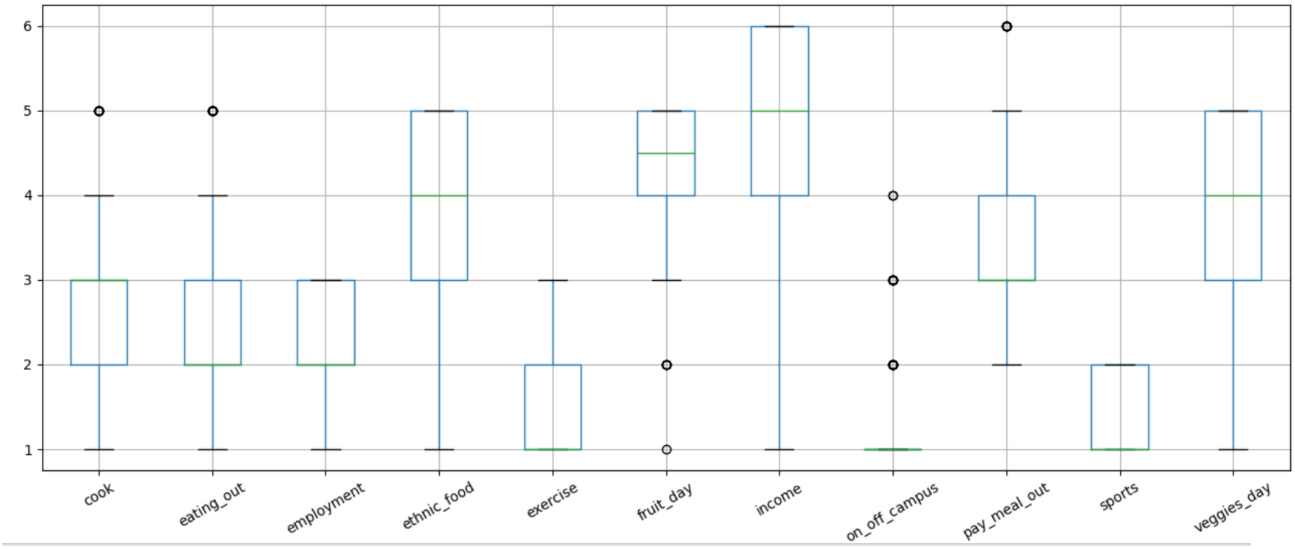
7.1.2 Pairplot Analysis

A pair plot analysis is a data visualization technique that displays pairwise relationships between variables in a dataset. It generates a grid of scatterplots, histograms, or other visual representations to help explore correlations, patterns, and distributions between different pairs of variables simultaneously. It is a valuable tool for understanding multivariate data and identifying potential relationships or trends.



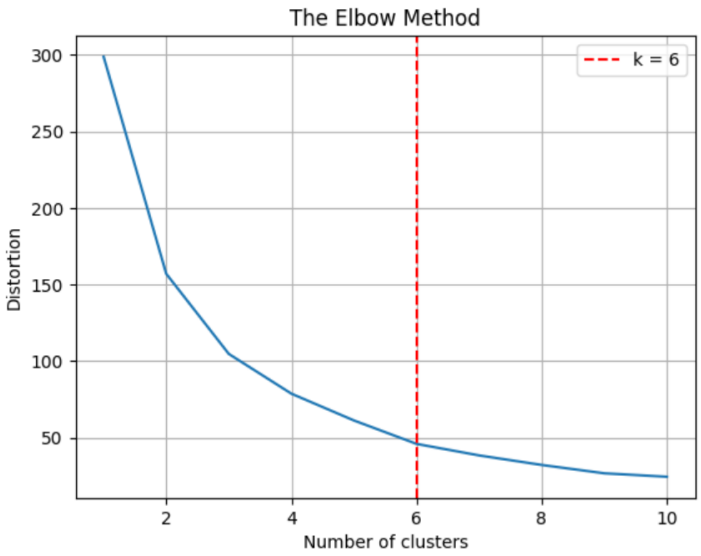
7.1.3 Box Plot Analysis

A box plot, also known as a box-and-whisker plot, is a data visualization technique that provides a summary of the distribution of a dataset. It displays the median, quartiles, and potential outliers in the data using a rectangular box and "whiskers." Box plots help visualize the central tendency and spread of data and are useful for identifying skewness, variability, and outliers in a dataset.



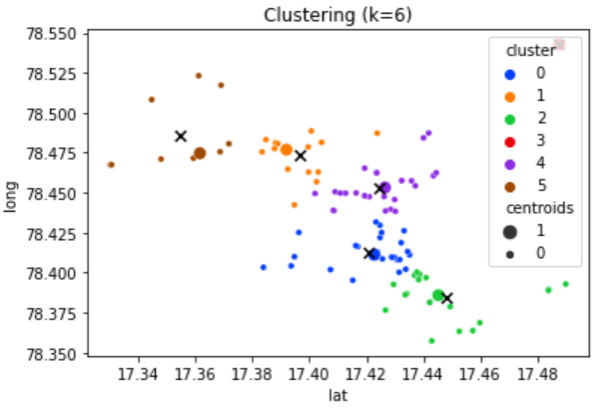
7.2 Elbow Method

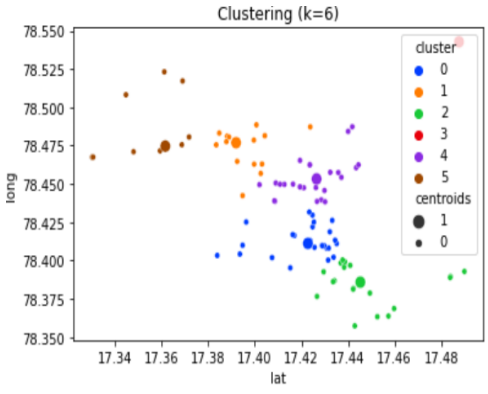
The elbow method is a technique used in clustering analysis, particularly in K-means Clustering. It helps determine the optimal number of clusters for a dataset by plotting the within-cluster sum of squares (WCSS) as a function of the number of clusters and identifying the "elbow" point in the plot. The "elbow" represents the point where adding more clusters no longer significantly reduces WCSS, indicating an appropriate number of clusters for the data.



7.3 K-means Clustering

K-means clustering is a machine learning algorithm used to partition a dataset into a specified number of distinct groups or clusters. It works by iteratively assigning data points to the nearest cluster center and recalculating cluster centers to minimize the within-cluster sum of squares. K-means is widely used for data segmentation, pattern recognition, and feature engineering in various applications.





7.4 ANALYSIS METHODOLOGIES

The following are the testing methodologies:

- **Exploratory Data Analysis (EDA).**
- **Pairplot Analysis.**
- **Box Plot Analysis.**
- **Elbow Method.**
- **K-means Clustering.**

CHAPTER-8

RESULTS AND OUTPUT SCREENS

8 . RESULT:

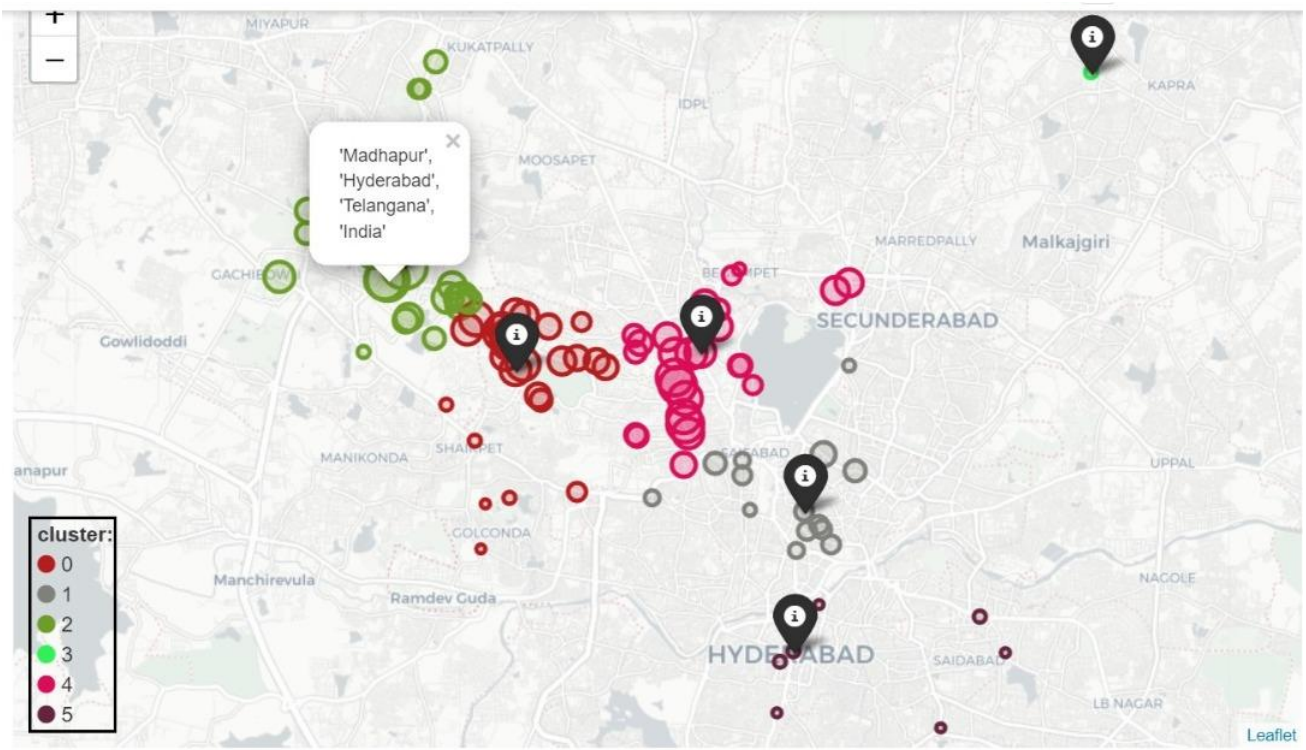


Figure. 8:

Depicting Clusters on Geo-location map.

The outcomes are shown as, Restaurants, Departmental stores, and Gyms are most abundant in cluster 0 (Green), whereas they are least prevalent in cluster 1 (Gray), Cluster 2(Red) which contains more restaurants but fewer departmental stores and gym, these are depicted in Fig 8.

8.1 UPLOADING THE DATASET

We need data to do data analysis! Fetch the data we need and set up your environment before you move on to data analysis.

Uploading the dataset in explanatory analysis of geolocational data involves importing spatial information into a tool or platform for analysis. This typically includes latitude and longitude coordinates, allowing you to explore and visualize geographical patterns. This step is crucial for conducting spatial analysis, such as mapping, clustering, or identifying trends based on location, in order to derive insights from the geolocational dataset.

8.2 CLEANING THE DATASET

Cleaning a geo-locational dataset involves addressing issues that may compromise data quality and analysis accuracy. Common steps include:

1. Handling Missing Values:
   - Identify and address missing values in latitude, longitude, or other relevant fields.
   - Consider imputation methods or removal of incomplete records.

2. Outlier Detection:
   - Identify and handle outliers that may distort spatial patterns.
   - Utilize statistical methods or visualization tools to spot anomalies.

3. Duplicate Removal:
   - Remove duplicate entries to avoid redundancy and ensure accurate analysis.
   - Check for identical or very similar records based on location and time.

4. Coordinate Validation:
   - Ensure latitude and longitude values fall within valid ranges for the given geographic area.
   - Eliminate records with incorrect or unrealistic coordinates.

5. Consistency Checks:
   - Verify consistency in data formats, units, and scales for geo attributes.
   - Standardize units and formats to ensure uniformity.

6. Address Standardization:
   - Standardize address formats to enhance geocoding accuracy.
   - Use geocoding services to convert textual addresses to precise coordinates.

7. Temporal Alignment:
   - Ensure consistency and coherence in temporal data.
   - Check for time zone discrepancies and standardize timestamps.

8.3 RUN ON K-MEANS ALGORITHM :

Running the K-Means algorithm in geo-locational data analysis involves:

1. Data Preprocessing: Prepare geo-coordinates for clustering.

2.Selecting K: Determine the optimal number of clusters (K) based on the nature of the data.

3.Algorithm Execution: Apply the K-Means algorithm to group data points into K clusters.

4.Cluster Interpretation: Analyze clusters to understand spatial patterns or groupings in the geo-locational data.

5.Visualization: Present results through maps or other visualizations for clear interpretation.

8.4 MAP FROM FOURESQUARE API :

In explanatory analysis of geolocational data using the Foursquare API, a map is a visual representation that displays relevant information about locations. Foursquare provides a mapping interface that allows users to plot and analyze geolocational data, such as venues, check-ins, or user-generated content. By leveraging the Foursquare API's mapping capabilities, analysts can create interactive maps to illustrate spatial patterns, trends, or relationships within the data. These maps serve as a powerful tool for conveying insights about the distribution and characteristics of points of interest, helping users better understand and interpret geolocational information.

8.5 PLOTTING ON MAP :

In explanatory analysis of geolocational data, plotting on a map involves visually representing data points, such as locations of interest or user check-ins, on a geographical map. This method allows analysts to explore spatial patterns, trends, and relationships within the data. By leveraging tools like Foursquare API or other mapping libraries, users can create informative visualizations that enhance the understanding of geolocational information, making it easier to communicate insights about the distribution and characteristics of specific points on the map.

8.6 ANALYSIS :

In explanatory analysis of geolocational data, analysis involves uncovering patterns, trends, and relationships within the spatial information. Analysts use various techniques to explore and interpret the data, such as clustering to identify spatial groupings, heat mapping to visualize concentration levels, and statistical methods to uncover correlations. The goal is to derive meaningful insights about the distribution and characteristics of locations, helping to explain and communicate the geospatial aspects of the data.

8.6.1 BOX-PLOT ANALYSIS :

In geolocational data analysis, a box plot provides a concise visual summary of the distribution and variability of a dataset. It displays key statistics such as median, quartiles, and potential outliers. In the context of geolocational data, a box plot can be used to compare and contrast location-based metrics, such as the number of check-ins or ratings, across different geographic

regions. This graphical representation aids in identifying patterns, variations, and potential outliers in the distribution of geospatial data, contributing valuable insights to the explanatory analysis process.

8.6.2 Clustering :

In geolocational data analysis, clustering involves grouping spatial data points based on their proximity or similarity. This technique helps identify patterns and spatial structures within the data. In explanatory analysis, clustering can reveal insights into the distribution of locations or venues, highlighting regions with similar characteristics. For example, it can be used to identify clusters of popular venues in a city. By visually grouping similar data points on a map, clustering facilitates the exploration and interpretation of spatial patterns in geolocational datasets.

8.7 Requirement DataSet :

In explanatory analysis of geolocational data, the requirement dataset refers to the specific data needed to address the analytical objectives. This dataset typically includes relevant geospatial information such as coordinates, venue types, user check-ins, or other location-based attributes. The choice of variables in the requirement dataset depends on the research questions and goals of the analysis, guiding the selection of data necessary for exploring patterns, trends, and relationships within the geographical context.

8.8 Food related csv file :

In explanatory analysis of geolocational data related to food, a CSV file may contain information about various aspects such as restaurant locations, cuisines, ratings, and user reviews. Analysts can use this data to create visualizations, conduct statistical analyses, and derive insights into spatial patterns of popular cuisines, concentration of high-rated restaurants, or trends in user preferences across different geographic locations. The CSV file serves as a structured dataset that facilitates the exploration and communication of key findings regarding the geolocational aspects of the food-related data.

8.9 Food_coded.csv file :

The "Food_coded.csv" file in explanatory analysis of geolocational data likely contains information related to food, possibly sourced from a geolocational dataset. It may include details about different food items, their characteristics, or associated attributes. Analysts can use this file to explore and visualize geospatial patterns related to food consumption, preferences, or trends. The file could be a valuable resource for uncovering insights about the intersection of location and food-related behaviors in the context of the geolocational dataset being analyzed.

GPA	Gender	breakfast	calories_c	calories_d	calories_s	coffee	comfort_f	comfort_f	comfort_f	cook	comfort_f	cuisine	diet_curre	diet_curre	drink	eating_cha	eating_cha	eating_cha
2.4	2	1	430	nan	315	1	none	we dont h	9	2	9	nan	eat good a	1	1	eat faster	1	1
3.654	1	1	610	3	420	2	chocolate,	Stress, bor	1	3	1	1	I eat abou	2	2	I eat out m	1	2
3.3	1	1	720	4	420	2	frozen yog	stress, sad	1	1	1	3	toast and	3	1	sometime	1	3
3.2	1	1	430	3	420	2	Pizza, Mac	Boredom	2	2	2	2	College die	2	2	Accepting	1	3
3.5	1	1	720	2	420	2	Ice cream,	Stress, bor	1	1	1	2	I try to eat	2	2	I have eat	3	4
2.25	1	1	610	3	980	2	Candy, bro	None, i do	4	3	4	nan	My curren	2	2	Eating rice	1	3
3.8	2	1	610	3	420	2	Chocolate	stress, bor	1	2	1	1	I eat a lot	3	1	I started e	2	5
3.3	1	1	720	3	420	1	Ice cream,	I eat comf	1	3	1	1	I eat a ver	1	2	Freshmen	2	5
3.3	1	1	430	nan	420	1	Donuts, ic	Boredom	2	3	2	1	I eat what	1	1	I snack les	2	8
3.3	1	1	430	3	315	2	Mac and c	Stress, ang	1	3	1	1	I eat healt	1	1	I cook a lo	1	3
3.5	1	1	610	3	980	2	Pasta, gra	Boredom	2	1	2	1	i eat very l	1	2	Nun	3	4
3.904	1	1	720	4	420	2	chocolate,	sadness, s	3	3	3	1	I am very l	1	1	Less meat	4	5
3.4	2	1	430	3	420	2	Cookies, p	Sadness, b	3	5	3	1	I focus mo	1	2	I have bee	2	5
3.6	1	1	610	3	420	2	ice cream,	stress, bo	1	2	1	1	Not as hea	2	2	not as hea	1	3
3.1	2	1	610	3	420	2	Pizza, fruit	Friends, er	2	3	2	1	Making su	1	2	I knew I w	3	4
nan	2	2	430	nan	980	2	cookies, d	boredom	2	4	2	1	I like a lot	2	2	none	3	4
4	1	1	265	3	420	1	Saltfish, C	Stress	1	3	1	nan	I eat very l	1	1	More Wat	2	5
3.6	2	1	430	3	980	2	chips, coo	I usually o	2	3	2	1	My curren	3	2	I would sa	1	2
3.4	1	1	720	3	980	1	Chocolate	Sadness, s	3	3	3	1	I eat lots o	1	1	I ate at the	4	6

indian_foo	italian_foo	life_rewar	marital_st	meals_din	mother_e	mother_p	nutritiona	on_off_ca	parents_c	pay_meal	persian_fo	self_perce	soup	sports	thai_food	tortilla_ca	turkey_cal	type_spor
5	5	1	1	rice, chick	1	unemploy	5	1	1	2	5	3	1	1	1	1165	345	car racing
4	4	1	2	Pasta, stea	4	Nurse RN	4	1	1	4	4	3	1	1	2	725	690	Basketball
5	5	7	2	chicken ar	2	owns busi	4	2	1	3	5	6	1	2	5	1165	500	none
5	5	2	2	Grilled chi	4	Special Ed	2	1	1	2	5	5	1	2	5	725	690	nan
2	5	1	1	Chicken Pa	5	Substance	3	1	1	4	2	4	1	1	4	940	500	Softball
5	5	4	2	Anything t	1	Hair Braid	1	1	2	5	5	5	1	2	4	940	345	None.
5	5	8	1	Grilled chi	4	Journalist	4	2	2	2	5	4	1	1	5	940	690	soccer
1	3	3	1	chicken, st	2	cook	4	1	1	5	1	3	1	2	1	725	500	none
5	5	8	2	Pasta, Fish	5	Elementar	2	1	2	3	5	4	2	2	5	725	345	none
4	5	3	2	pasta sala	5	Pharmace	5	1	3	3	4	3	1	1	4	580	345	field hocke
1	5	8	1	chicken al	4	Chidos Cle	2	3	1	2	2	1	1	1	2	940	345	soccer
5	5	1	2	Chicken pa	4	Court Rep	5	1	1	3	5	2	1	1	5	940	500	Running
3	4	9	2	Cereal, piz	4	Child care	2	1	2	2	3	5	2	1	3	940	500	Soccer and
3	5	10	2	pasta, chic	4	business c	2	2	2	3	3	3	1	1	5	1165	850	intramura
2	3	1	1	Pizza, chic	4	Charity wo	2	2	2	3	1	3	1	1	4	940	500	Hockey
1	5	nan	2	pizza buffa	4	Librarian	1	1	2	3	1	2	2	1	1	940	345	Hockey
5	3	9	2	Curry goat	2	Police	4	1	2	2	2	3	1	2	1	580	345	nan
3	5	7	2	Grilled chi	2	stay at hor	4	1	1	6	2	2	1	1	3	940	500	hockey

eating_ou	employme	ethnic_foo	exercise	father_ed	father_pr	fav_cuisin	fav_cuisin	fav_food	food_chilc	fries	fruit_day	grade_lev	greek_foo	healthy_fe	healthy_m	ideal_diet	ideal_diet	income
3	3	1	1	5	profesor	Arabic cuis	3	1	rice and c	2	5	2	5	2	looks not	being heal	8	5
2	2	4	1	2	Self emplc	Italian	1	1	chicken ar	1	4	4	4	5	Grains, Ve	Try to eat	3	4
2	3	5	2	2	owns busi	italian	1	3	mac and c	1	5	3	5	6	usually inc	i would sa	6	6
2	3	5	3	2	Mechanic	Turkish	3	1	Beef strog	2	4	4	5	7	Fresh fruit	Healthy, fr	2	6
2	2	4	1	4	IT	Italian	1	3	Pasta, chic	1	4	4	4	6	A lean pro	Ideally I w	2	6
1	3	4	2	1	Taxi Drive	African	6	3	Fries, plair	1	2	2	2	4	Requires v	My ideal d	2	1
2	3	5	1	4	Assemble	Thai	4	1	grilled chic	1	4	4	5	4	Protein, ve	I would id	2	4
2	2	2	2	3	Business g	Anything a	5	1	chicken, ch	1	5	2	3	3	A healthy	My ideal d	2	5
5	2	5	nan	5	High Scho	Seafood	1	3	Shrimp, sp	1	4	1	5	7	Colorful	The same	6	5
3	3	5	1	5	commissic	Italian	1	1	Pasta, Egg	1	5	1	5	3	Chicken ar	Lots of pro	2	4
2	1	5	1	2	Idk	Orange ch	4	1	Chicken	1	5	3	1	9	Chicken, a	Chicken, fi	7	3
1	2	5	1	3	Home Ma	Italian	1	1	Scalloped	1	5	2	5	1	lean prote	More heal	2	5
1	3	4	3	3	Shirt desig	Chinese	4	3	Pizza, chic	1	4	1	3	9	A salad wi	My ideal d	1	5
4	2	5	2	2	business c	italian	1	1	steak, lasa	1	5	3	4	8	Lots of veg	organic/he	2	5
2	3	4	2	4	Commodit	Chinese	4	3	Pizza	1	5	3	4	2	Green and	Eat a little	1	5
4	3	1	1	5	Hockey Co	italian foo	1	2	pizza	1	3	1	1	6	chicken, ve	I wish I ate	2	4
1	2	2	2	1	Constructi	Jamaican	7	1	Curry, Stev	1	5	1	2	7	Not too m	Vegetables	2	1
2	3	4	1	2	self emplc	American	5	3	chicken fir	1	3	1	3	8	for me usu	I like how	1	6
3	2	1	3	5	Engineer	Chicken	5	3	Chicken	1	5	2	3	6	Everything	Lots of fru	2	5

CHAPTER:9

CONCLUSION

K-Means clustering technique has been used to cluster the data which is taken as a geo-locational data from the Here Geocoding and Search API based on the uses location. We have built a simplistic website that accepts the user location and then outputs a map that has been filled with geographic data clusters. It is user friendly, making it simple to use. This website aids in resolving a problem that affects migrants frequently. With this service, we may find housing options fast that fit our preferences and are useful for migrants. We also included the nearby locations to our website options for restaurants, banks and schools.

CHAPTER-10
FUTURE ENHANCEMENT

Analyzing geolocational data offers valuable insights into spatial patterns. Enhancements like real-time integration, predictive modeling, and interactive visualizations can provide dynamic, user-friendly experiences. Ensuring data privacy, scalability, and mobile compatibility is crucial. Incorporating feedback mechanisms and IoT integration can keep the project relevant and effective in evolving geospatial contexts. Additionally, integrating real-time data sources and implementing interactive visualizations can provide a more dynamic and engaging analysis experience. Enhancements like geospatial clustering, network analysis, and geofencing can further enrich the insights derived from the data. It's important to continuously explore new techniques and technologies to enhance the project's capabilities and stay at the forefront of geolocational data analysis.

CHAPTER 11

REFERENCES

[1] M. Sumithra, Sai Pavithra, Sowmya, Swetha, Srinithi. Exploratory Analysis of Geo-Locational Data - Accommodation Recommendation. International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 07 | July 2022.

[2] Gaurang Ajmera, Alok Singh 2018. Hierarchical Data Analysis on Geo-locational Data using Machine Learning.

[3] S. G. K. Patro et al., "A Hybrid Action-Related K-Nearest Neighbor (HAR-KNN) Approach for Recommendation Systems," in IEEE Access, vol. 8, pp. 90978-90991, 2020, DOI: 10.1109/ACCESS.2020.2994056.

[4] Al-Dabooni, S. and Wunsch, D., 2018. Model order reduction based on agglomerative hierarchical clustering. IEEE transactions on neural networks and learning systems, 30(6), pp.1881-1895.

[5] Roger D Peng. Exploratory Data Analysis with R, A book published by Lulu.com in 2016 April 20.

[6] Srinivas Chellaboina, Maneesh Gembali, Sathya Priya. Product Recommendation based on Customer Segmentation Engine. Published in 2022 2nd International Conference on Intelligent Technologies (CONIT) Date of Conference: 24-26 June 2022.

[7] Nemani, Y.M., Yadav, R., Patki, M., Padave, O. and Bhelande, M.M., 2018. City Tour Traveler: Based on Foursquare API. City, 5(04).

[8] Psyllids, A., Yang, J. and Bozzons, A., 2018. Regionalization of social interactions and points-of-interest location prediction with geosocial data. IEEE Access, 6, pp.34334-34353.

[9] Patel, P., Sivaiah, B. and Patel, R., 2022, July. Approaches for finding Optimal Number of Clusters using K-Means and Agglomerative Hierarchical Clustering Techniques. In 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) (pp. 1-6). IEEE.

[10] Wang, P., Ding, C., Tan, W., Gong, M., Jia, K. and Tao, D., 2022. Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. IEEE Transactions on Multimedia.

[11] Daraio, E., Cagliero, L., Chiusano, S. and Garza, P., 2022. Complementing Location-Based Social Network Data with Mobility Data: A Pattern-Based Approach. IEEE Transactions on Intelligent Transportation Systems.

[12] Sharma, S. and Batra, N., 2019, February. Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 568- 573). IEEE.

[13] Sumithra, M., Shruthi, S., Ram, S., Swathi, S., Deepika, T., "MRI image classification of brain tumor using deep neural network and deployment using web framework", Advances in Parallel Computing, 2021, 38, pp. 614–617.