**Lab Report**

Title: *<*Prospectus\_Final Project *>*

Notice: Dr. Bryan Runck

Author: Rongxuan Zhu

Date: 9/29/2021

**Project Repository:** *< >*

**Google Drive Link:** *<N/A.>*

**Time Spent:** *<10 hours >*

**Abstract**

The unsupervised machine learning method k mean clustering algorithm will be used to group sets of latitude and longitude coordinate points from Travel Behavior Inventory (TBI) Household Survey Interview Data for 2010 in Minnesota to determine the clustering pattern of household and trip origins and destination location. Elbows curve will be used to determine the optimal number of clustering groups k. The latitude and longitude coordinate points will be grouped into k groups using the unsupervised k means algorithm. This project will combine spatial analysis with the power of machine learning to determine and obtain the clustering pattern of TBI data to understand how the household and trip locations are segregated spatially.

**Problem Statement**

This project will apply the unsupervised machine learning method k mean clustering algorithm in Jupyter notebook python environment to find the clustering pattern and a better way to label the Travel behavior Inventory Household Survey Interview Data for 2010 and to figure out how the household and trip locations are segregated spatially in Minnesota.

*Table 1. <Problem Statement Elements >*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Requirement** | **Defined As** | **(Spatial) Data** | **Attribute Data** | **Dataset** | **Preparation** |
| 1 | Travel Behavior Inventory Household Interview Survey | Raw input dataset for k mean clustering algorithm from MNDOT | Household survey data | Travel Behavior Inventory Household Interview Survey | [Mn GeoSpatial Commons](https://gisdata.mn.gov/dataset/us-mn-state-metc-society-tbi-home-interview2010) | Downloaded from website and import into Jupyter Notebook |
| 2 | K Mean Clustering Algorithm | Unsupervised Machine Learning Method |  |  |  |  |

**Input Data**

The Travel Behavior Inventory survey was conducted from fall 2010 to spring 2012. The respondents are people from different cities in Minnesota. The TBI data contains three datasets, household, person, and trips. The latitude and longitude coordinate data of household and trip origins and destinations will be used to run the k mean clustering algorithm to find the clustering pattern of respondents’ household and trip locations.

Table 2. <Input Data>

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Title** | **Purpose in Analysis** | **Link to Source** |
| 1 | Travel Behavior Inventory Household Interview Survey | Raw input dataset for k mean clustering from MNDOT | [Mn GeoSpatial Commons](https://gisdata.mn.gov/dataset/us-mn-state-metc-society-tbi-home-interview2010) |

**Methods**

First is to use the elbow curve to determine the optimal number of clusters. The elbow curve run k mean clustering on the dataset for a range of values of k (for example 1 to 10) and for each k value, it will calculate the Sum of Squared Errors (SSE). The centroids will get closer to the centroids of the clusters as K increases. The Elbow curve first increase and then decline after it reaches the optimal k value. After the k value is determined, the k means clustering algorithm will be implemented in Python using Scikit-Learn (Scikit Learn,2007-2021) and Pandas Library. This algorithm identifies a centroid and the data points that are clustered into their groups by the centroid they are the closest to.

Diagram

Description automatically generated

Figure 1. Data flow diagram.

**Results**

*Diagram

Description automatically generated*

Figure 2. K Means Results (Sharma, 2021)

The k means the algorithm will group similar and unlabeled items in the form of clusters as Figure 2 shows. The GPS points will be grouped as clusters with the nearest mean. The results will be visualized by plotting the data colored by each label.

**Results Verification**

The data will be imported and mapped in ArcGIS Pro and a clustering analysis tool such as Hot Spot analysis will be run to generate a clustering pattern which will be compared with the results of the k mean clustering algorithm in python.

**Discussion and Conclusion**

After the k means algorithm, the latitude, and longitude coordinates of household, trip origins and destination will have a new clustering label that helps to describe the data and find the spatial clustering pattern. The results will help to analyze the respondent's travel behavior as to which locations are more likely to have interactions between households as their origins and destination located. Having clustered the household and trip locations, we can have a different perspective of the TBI data.

**References**

Scikit Learn (2007-2021). Retrieved September 29, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

Sharma, P. (2021). *K means Clustering simplified in Python: K Means Algorithm*. Analytics Vidhya. Retrieved September 29, 2021, from https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/.

**Self-score**

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Description** | **Points Possible** | **Score** |
| **Structural Elements** | All elements of a lab report are included **(2 points each)**:  Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score | 28 | **28** |
| **Clarity of Content** | Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**. | 24 | **24** |
| **Reproducibility** | Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified. | 28 | **28** |
| **Verification** | Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**. | 20 | **20** |
|  |  | 100 | **100** |