K- Means Clustering
Uber Pickup Data NYC

Project Report

Devananth V

EP20BTECH11004

Nikhil Krishna A R

ME20BTECH11031

Data Science and Analysis
Course Code:EP4130

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Supervisor: Shantanu Desai

May 2023

Contents

Abstract

In this project we analyse the Uber pickup data NYC of april 2014. We apply k means clustering on the dataset and divide the area into clusters. We also analyse the time vs Uber pickup frequency in weekdays and weekends.

The dataset used in this analysis is the Uber Pickups in New York City dataset from Kaggle, which contains information on the frequency of Uber pickups in New York City over a period of six months. We use only the month of data of April 2014 for our project.

In this report, we will explore how K-means clustering can be used to analyze the frequency of Uber pickups in different areas.

Introduction

1.1 Clustering

Cluster analysis aims at grouping data objects based only on information found in the data that de- scribes the objects and their relationships. The objective of clustering is to produce groupings of things that are similar to (or connected to) one another and distinct from (or unrelated to) one another. The better, or more distinct the clustering, the greater the homogeneity inside a group and the greater the difference across groups. Cluster analysis has always been crucial in a wide range of fields, including statistics, biology, pattern recognition, information retrieval, machine learning, and data mining, it is sometimes only a starting point for other objectives like data summarization, whether for understanding or utility.

There are various kinds of clustering methods depending on the need and application at hand. In this work, we present one of the oldest and most used clustering technique - The K-means.

1.2 K-means Clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled data set into different clusters. Based on a similarity measure such as euclidean distance between points, it tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different(far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the

cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

1.2.1 K-Means Algorithm

Given a set of n observations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and a desired number of clusters k, the K-means clustering algorithm proceeds as follows:

1. Initialize k cluster centers $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$ randomly or using a predefined method.

2. Repeat until convergence:

   - Assign each observation $\mathbf{x}_i$ to the cluster with the closest center $\mathbf{c}_j$ based on the Euclidean distance $d(\mathbf{x}_i, \mathbf{c}_j) = \sqrt{\sum_{p=1}^{d}(x_{ip} - c_{jp})^2}$, where d is the dimensionality of the data.

   - Recalculate the cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$ as the mean of the observations assigned to each cluster:

     $$\mathbf{c}_j = \frac{1}{n_j} \sum_{i=1}^{n} \mathbf{x}_i \cdot [j = \text{argmin}_l d(\mathbf{x}_i, \mathbf{c}_l)], \tag{1.1}$$

     where n_j is the number of observations assigned to cluster j.

   - The algorithm terminates when the cluster assignments no longer change or a maximum number of iterations is reached.

   - The objective function of K-means clustering is the sum of squared distances between each observation and its closest cluster center:

     $$J = \sum_{i=1}^{n} \min_{j=1}^{k} d(\mathbf{x}_i, \mathbf{c}_j)^2. \tag{1.2}$$

     The algorithm aims to minimize this objective function by finding optimal cluster centers that minimize the inertia.

Analyzing the Dataset

In this project we will analyse the dataset : Uber trip data from 2014 (April)

There are six files of raw data on Uber pickups in New York City from April to September 2014. We only consider the data for April, 2024. The file has has the following columns:

1. Date/Time : The date and time of the Uber pickup
2. Lat : The latitude of the Uber pickup
3. Lon : The longitude of the Uber pickup
4. Base : The TLC base company code affiliated with the Uber pickup

2.1 Plotting the Data

2.1.1 Ride Volume by Date

First we will make a bar plot of uber rides per day in April 2014.

From the plot, it seems that the daily ride volumes follow a cyclical pattern.

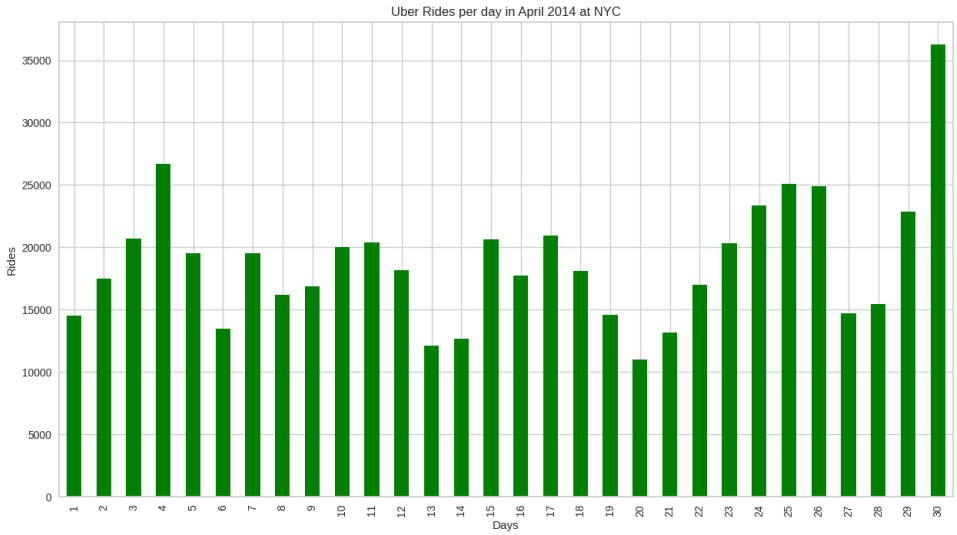

Figure 2.1: Total Rides Vs Date

2.1.2 Ride Volume by Days in Week

Based on the previous plot, it seems like there might also be a pattern throughout the weeks.

To confirm this, we can create another plot based on the days of the week.

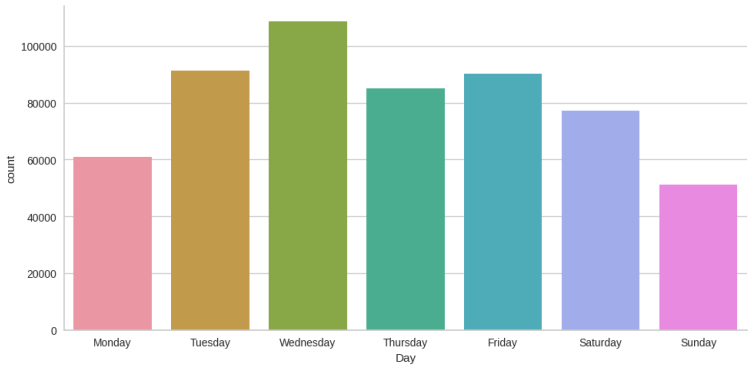

Figure 2.2: Total Rides Vs Day

2.1.3 Ride Volume by Hours

We can also plot the ride occurrences by hours of a day.

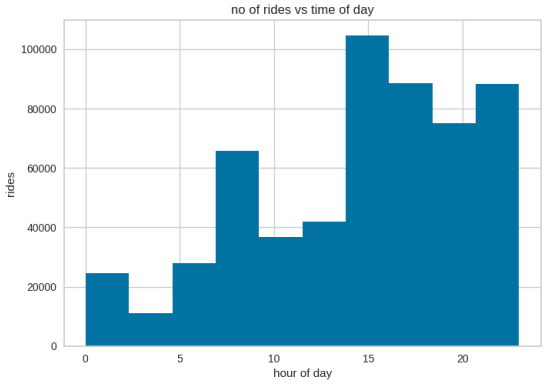

Figure 2.3: Total Rides Vs Hours

2.1.4 Observation

It appears that Monday and Sunday have the least counts, while Tuesday and Wednesday have the most rides.

The plot indicates that approximately 60% of rides occurred between 14:00 - 21:00.

Possible Reason:

- Commuting patterns: Tuesdays and Wednesdays are typically busy weekdays when people commute to work and go about their regular activities. This could explain the higher ride counts on these days. In contrast, Mondays and Sundays may see fewer rides as people may be off work or have more flexible schedules.

- Business travel: Tuesdays and Wednesdays are often popular days for business travel, as professionals may travel to different locations for meetings, conferences, or other work-related events. This could contribute to the higher ride counts on these days compared to Mondays and Sundays.

- Weekend leisure activities: Mondays and Sundays may see fewer rides as people may prefer to stay at home or engage in leisure activities closer to their residences during weekends, resulting in lower ride counts.

2.2 Clustering the Data

2.2.1 Elbow Method

We use the elbow method to find the number of ideal clusters(k) to be used.

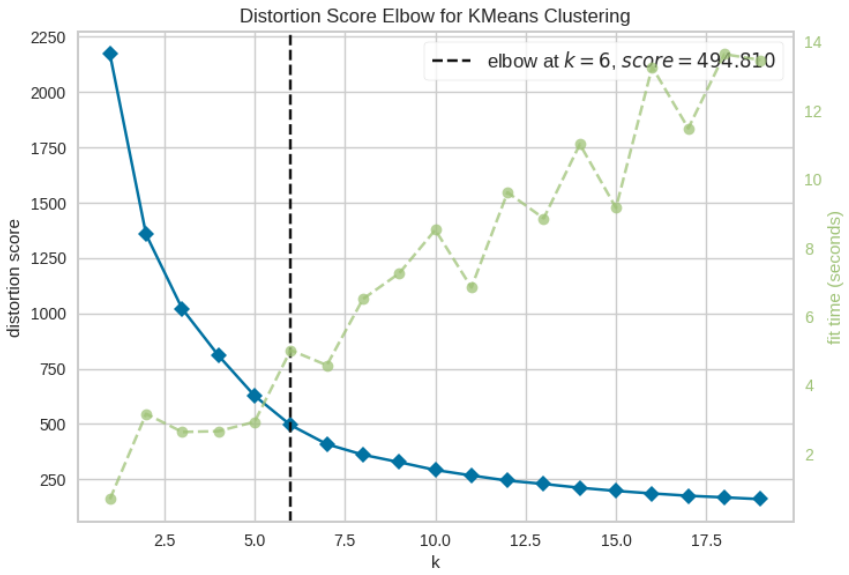

Figure 2.4: Elbow Method

2.2.2 Finding the Centroids

We find the K-Means clustering to find k number of clusters and their respective centroids.

We plot the centroid using Lattitude and Longitude.

Now, we use the "folium" package available in python to plot the centroids on a map.

The Map shows the areas with the most probability of getting a pickup. It is helpful for a uber driver to

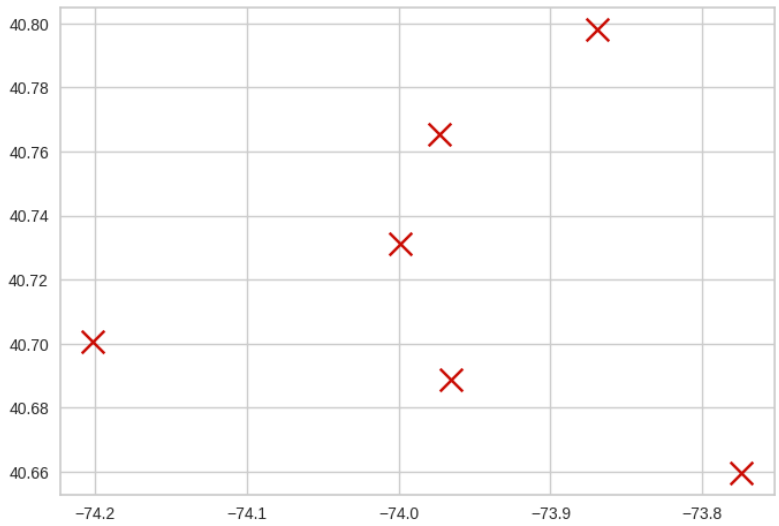

Figure 2.5: Elbow Method

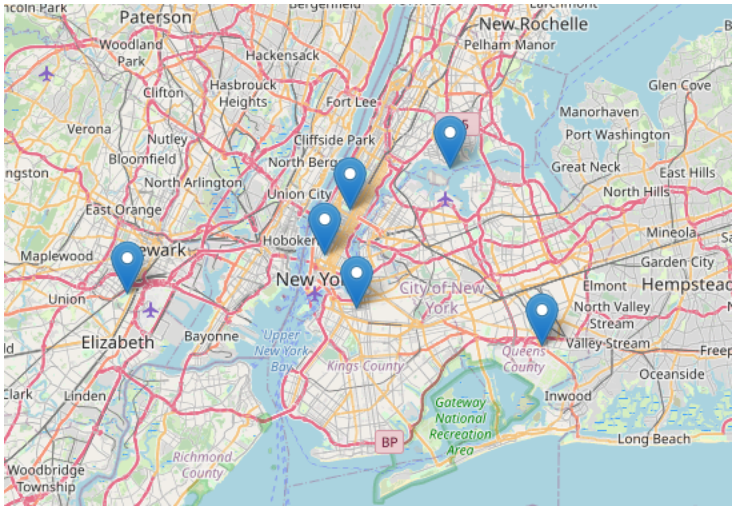

Figure 2.6: Hotspots

wait in this area, so as to get the more number of pickup request. Also the the pickup point on an average will be shortest from them.

2.3 Clusters

The image given below gives the scatter plot of the clusters we got by K-means method.

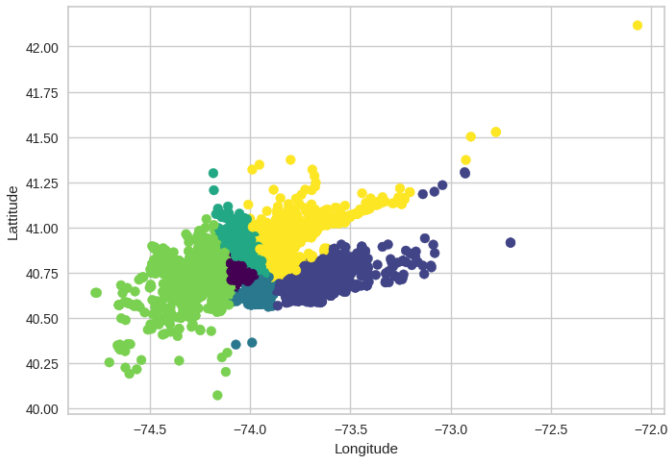

Figure 2.7: Clusters

The plot given below shows the number of pickups vs cluster. The plots given below shows number of rides

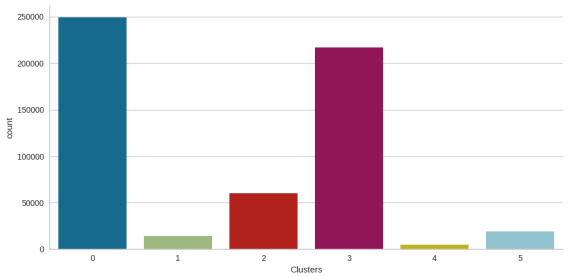

Figure 2.8: No of Pickups per Cluster

per day vs time for each of the cluster.

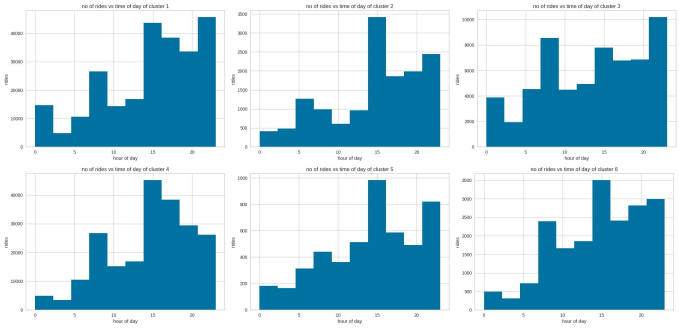

Figure 2.9: No of Rides Vs Time per Cluster

2.4 Weekdays and Weekends

Now we compare the data for weekdays and weekends.

2.4.1 Weekdays

The plot shows number of pickups Vs day. Then we find the hotspots for this data.

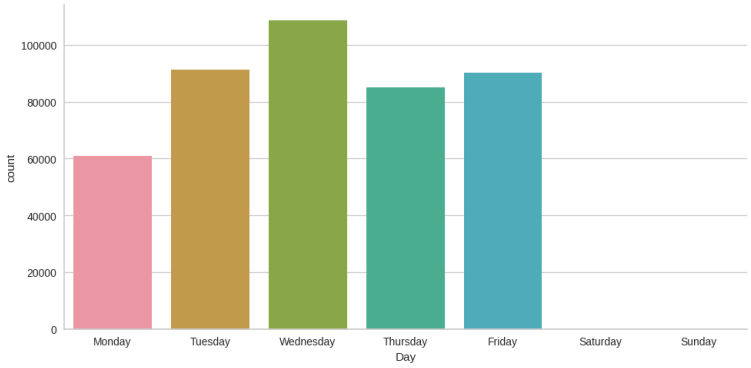

Figure 2.10: No of Pickups Vs Day

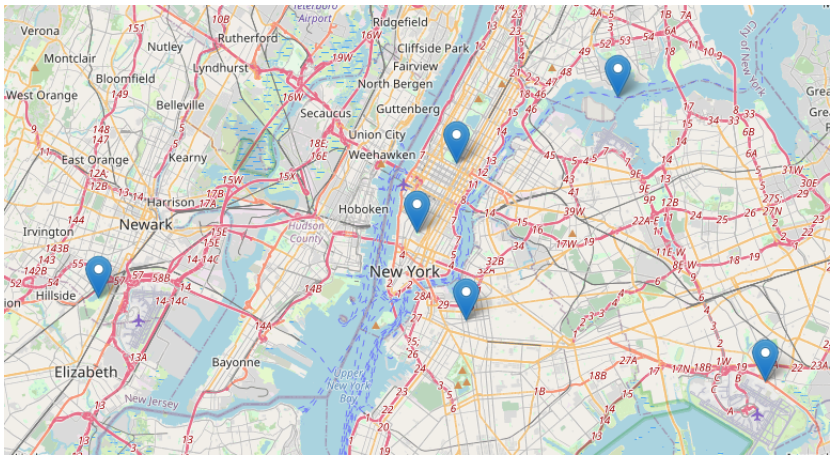

Figure 2.11: No of Pickups Vs Day

2.4.2 Weekends

The plot shows number of pickups Vs day. Then we find the hotspots for this data.

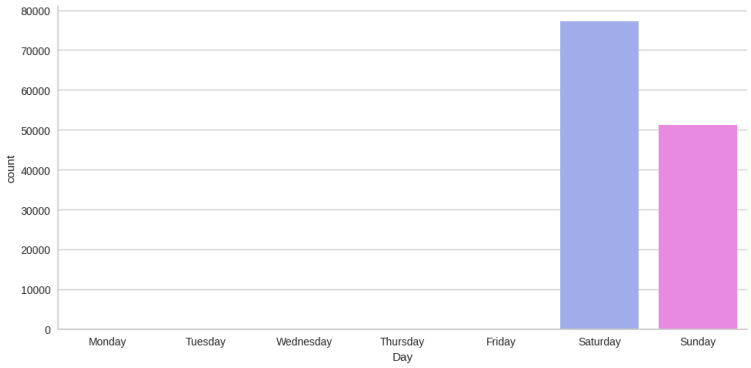

Figure 2.12: No of Pickups Vs Day

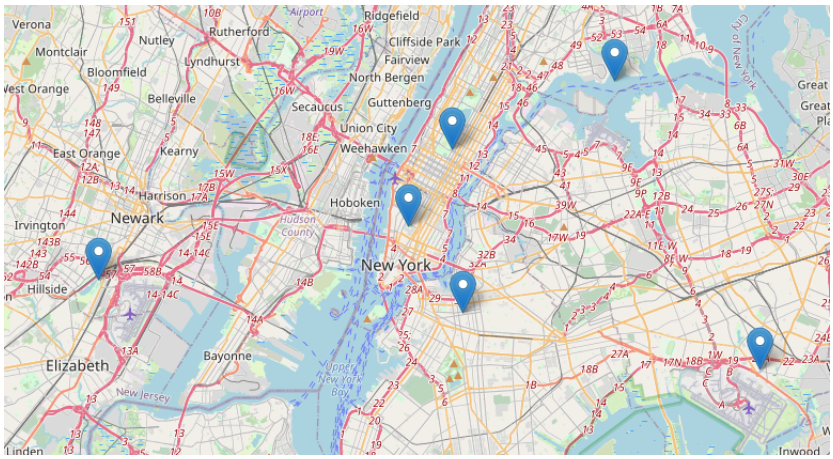

Figure 2.13: No of Pickups Vs Day

2.5 Acknowledgement

We would like to thank Prof. Shantanu Desai for giving us this wonderful opportunity to credit the project work we have done and for teaching the concepts in class.

2.6 Code Repository

https://github.com/ep20btech11004/projects/tree/main/DSA_Project/Data

2.7 References

- K-means Clustering : Wikipedia