CSE-564 VISUALIZATION

LAB REPORT – 2B

Seattle Airbnb Open Data

By: Mayuresh Pingale

SBU ID: 114910589

Aim:

The project aims to visualize a dataset by using various charts developed using the D3.JS library.

1. To perform Multi-Dimension Scaling (MDS) on data and variables and visualize it using Scatter Plot.
2. To visualize all data (categorical and numerical) using Parallel Co-ordinates Plot
3. To find good PCP axes ordering from user interactions and show only the selected variables

Dataset:

The Seattle Airbnb open dataset was taken from Kaggle and publicly made available by Airbnb. This dataset describes the listing activity of homestays in Seattle, Washington. It includes full descriptions, average scores, reviews, availability, etc. The dataset contains Seattle Airbnb listings, ratings, and related data. The dataset is a good mix of numerical and categorical variables with 3818 properties and 92 attributes describing them.

Link to Dataset: https://www.kaggle.com/datasets/airbnb/seattle

Feature Selection:

The following Numerical are the definitions of the attributes in the sampled data.

1. Accommodates - The number of people/guests that can stay on the property
2. Bathrooms - The number of restrooms available in the house
3. Bedrooms - The number of bedrooms in the house
4. Price - Price of the property in dollars per day
5. Minimum Nights - Minimum number of days the property needs to be booked.
6. Number of Reviews - Number of reviews available on the property

7. Review Score Ratings - The ratings of the property given by the previous customers.
8. Review Score Cleanliness- The cleanliness rating of the property
9. Review Score Location - Ratings of the location of the property

Importance/ Value of Dataset:

This dataset was selected because it gave the real-world picture of renting and listings properties on Airbnb in Seattle. The dataset had a good mixture of categorical and numerical variables that was useful in the visualization of different charts. The data is helpful to find the vibe of each Seattle neighbourhood. Also, it can be used to analyse what factor affects the pricing of the property.

Deployment:

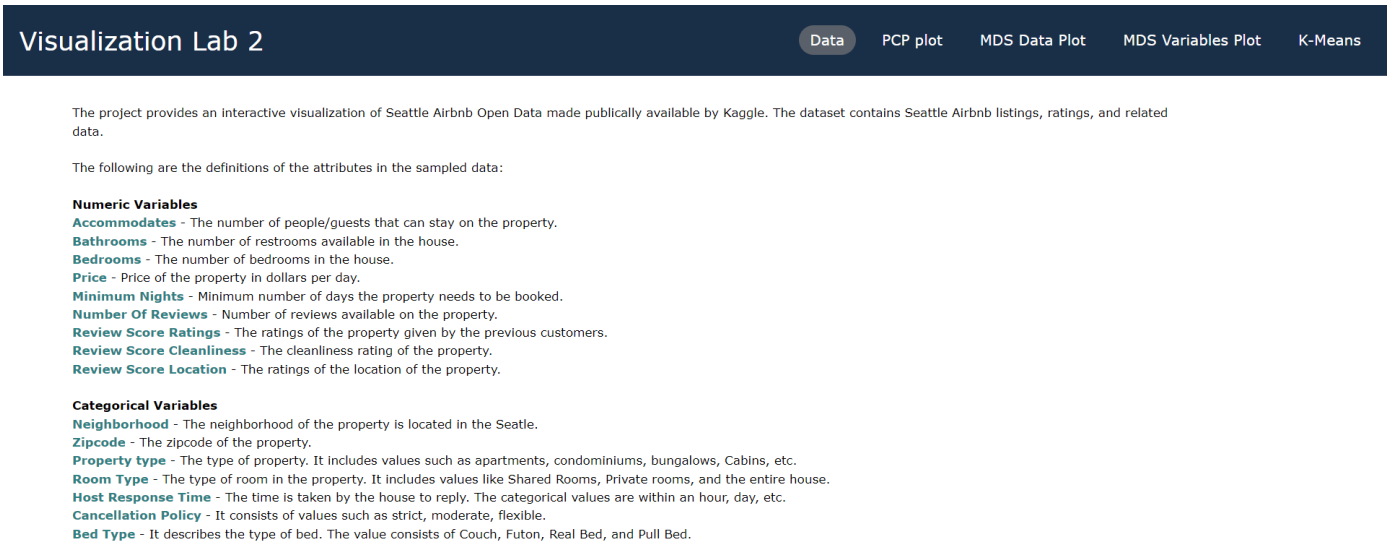
To run this project, just install flask. Please make sure you have installed Machine Learning Libraries like numpy, pandas, scikit, etc

Run: python backend.py

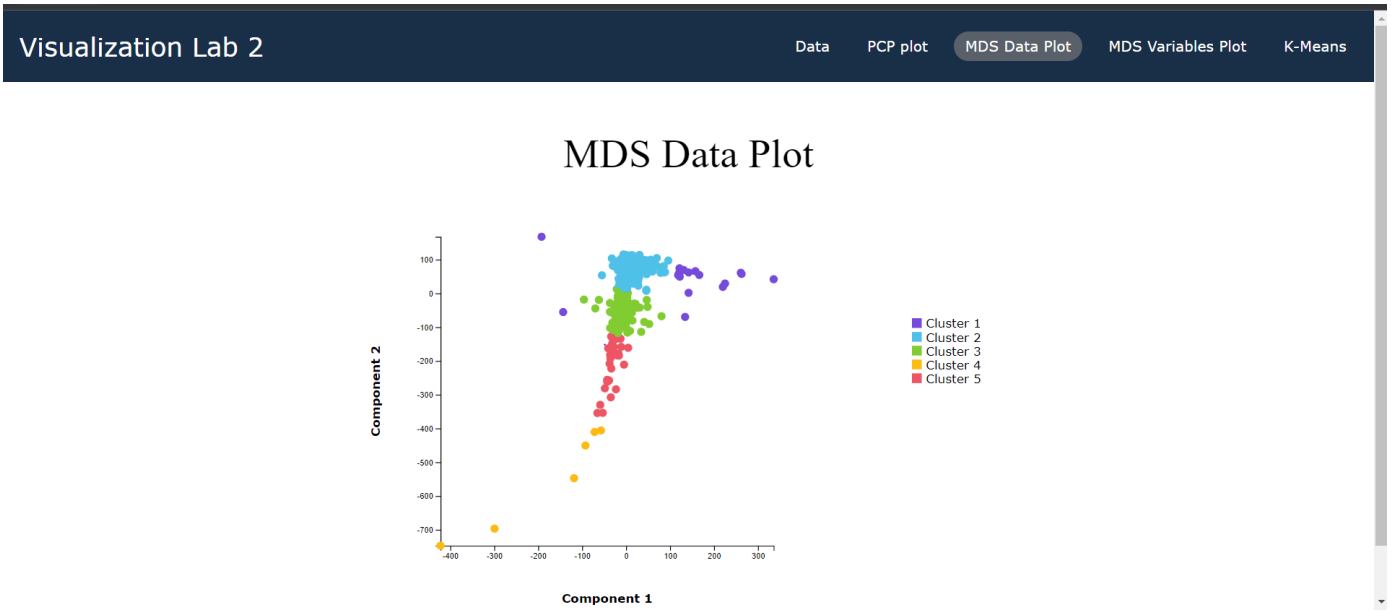
Please make sure you have an active Internet.

Features of Application:

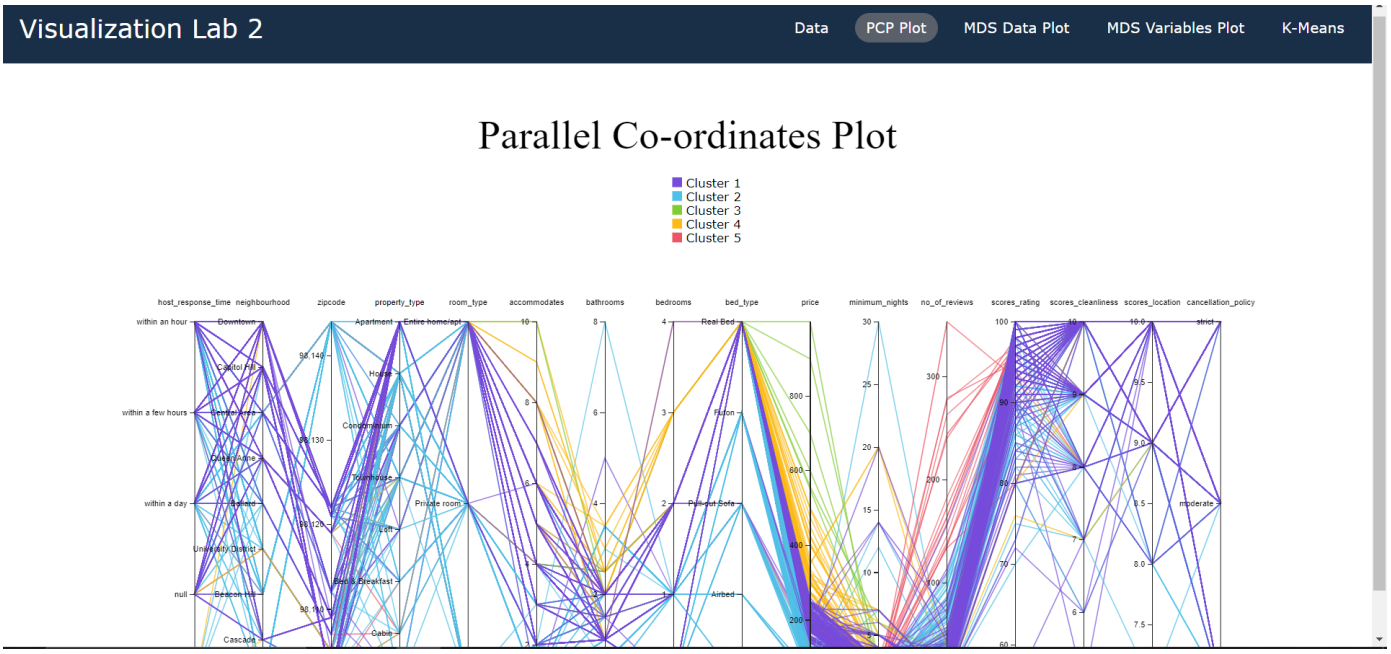
The application opens with the data page that provides information about the dataset. The navigation bar provided ease to switch between tabs. The user can select the type of graph, and the corresponding page will be displayed. The project has 5 HTML files, each corresponding to a different tab.



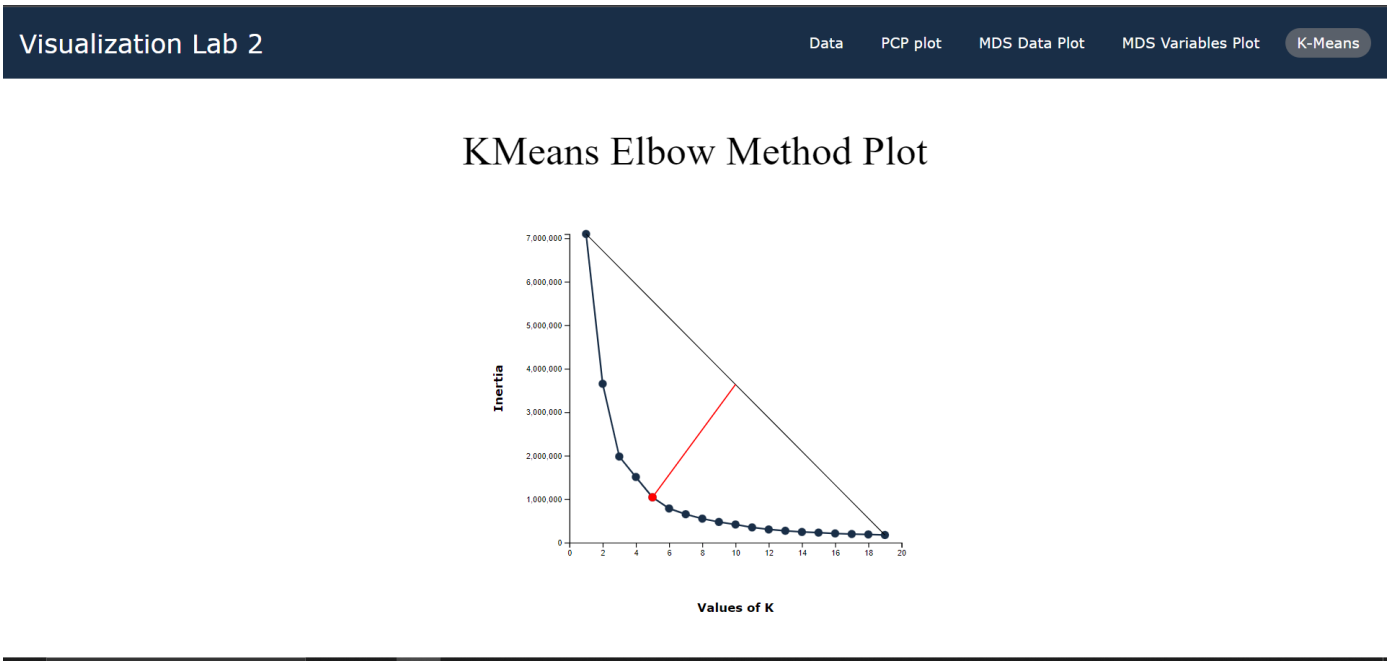
When the user selects a MDS Data plot, MDS operation is performed at the backend. The data will be visualized using Scatter Plot Matrix. The dots are colored according to the cluster they belong. A legend is also displayed for easy reference.



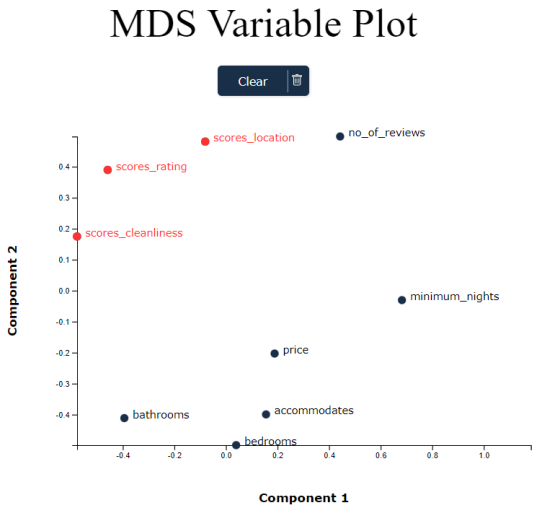
The PCP tab displays parallel co-ordinates plot for all the data dimensions. The dimensions are moveable. So, users can order the PCP plot. They can come up with any meaningful order. After reordering plot looks like this. But user can order any way(n!). The lines are colored according to cluster they belong. A color legend is also displayed at top. User can also filter the values based on values. This will help them in analysis.

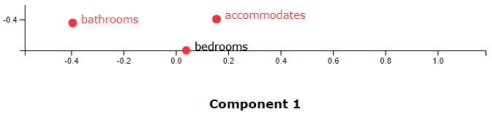


The k-Means tab displays the elbow method used for selecting the optimal value of K. The Y-axis displays inertia while the X-axis the value of K. According to the elbow method, the optimal value of **K is 5.** The highlighted bar shows the selected value of K.

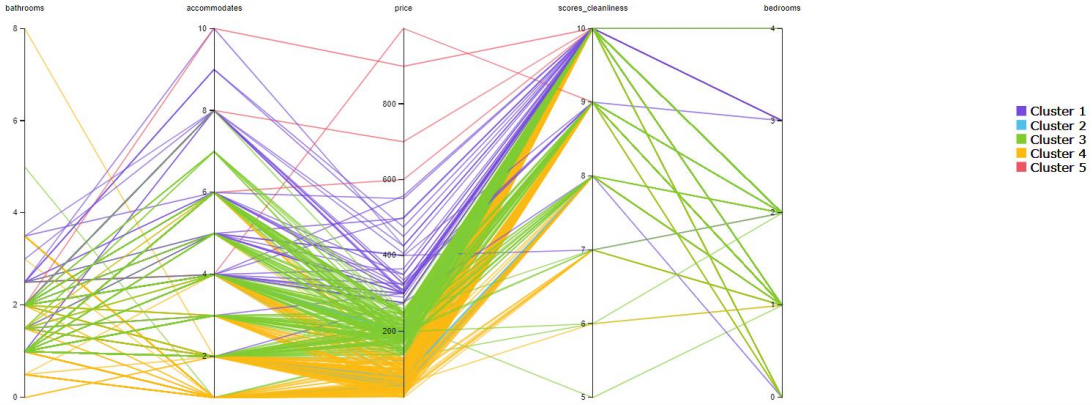


MDS Variable Plot Tab displays two things. The first part displays a scatterplot of Numerical Variables. Based on the variables selected by the user, The Parallel Plot gets updated. The orders of selection are preserved. The Parallel co-ordinate Plot is color-coded according to the clusters. A color legend is displayed for clusters. A clear button is provided at the top that clear user selection.





Parallel Plot



Interesting Observations:

1. From MDS Variables Plot, the top co-related features are number_of_reviews, scores_location, scores_cleanliness, and score_ratings.
2. The number of bathrooms and bedrooms is also related to each other.
3. By analyzing Parallel Co-ordinates Plot, the price of the room is high when cleanliness and locations score is high. Also price is low when the bed_type is Futon.

Advantages of Different Visualizations:

1. Multidimensional scales can assess how closely related different values are. It uses distances to find out co-relations.
2. A parallel co-ordinate plot makes it easy to perceive trend. The axes order can be changed to check out the trend between two different axes.

Disadvantages of Different Visualizations:

1. MDS does not deal in real numbers. The algorithm is very slow.
2. In parallel co-ordinates plot, the overlaying of data lines for common data values among data entries is issue.