CSE-564 VISUALIZATION

LAB REPORT – 2

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 dimension reduction and data visualization using PCA using Interactive Scree-Plot and Bi Plot
2. To use the PCA components less than the dimensionality index and visualize the data using Scree-Plot Matrix
3. To use k-means to find clusters and color the points

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:

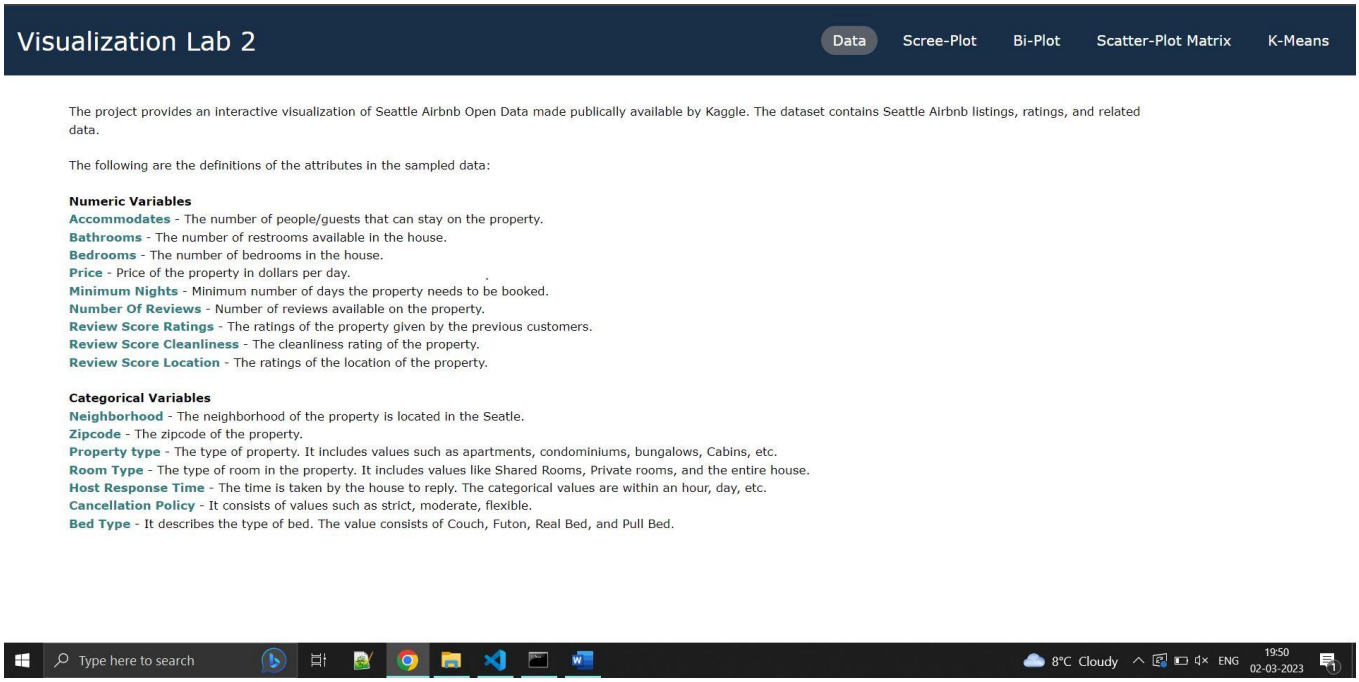
To run this project, just install flask. Please make sure you have installed Machine Learning Libraries like numpy, pandas, scikit, etc

Run: python backend.py

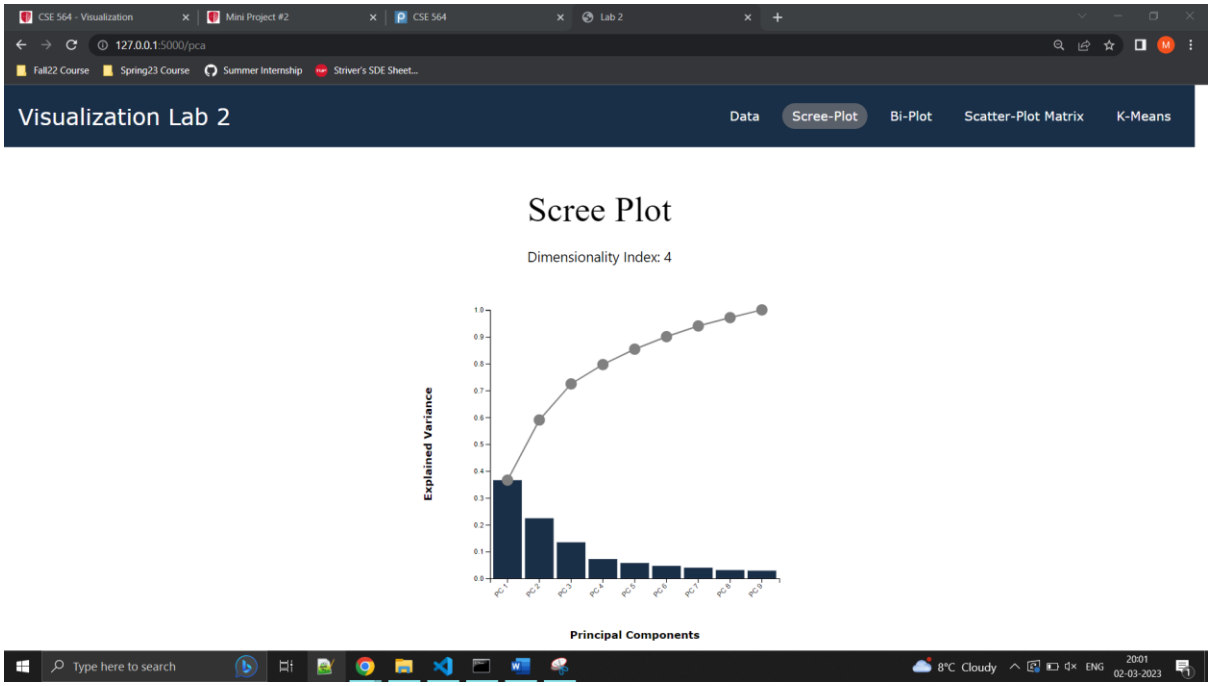
Please make sure you have an active Internet.

Features of Application:

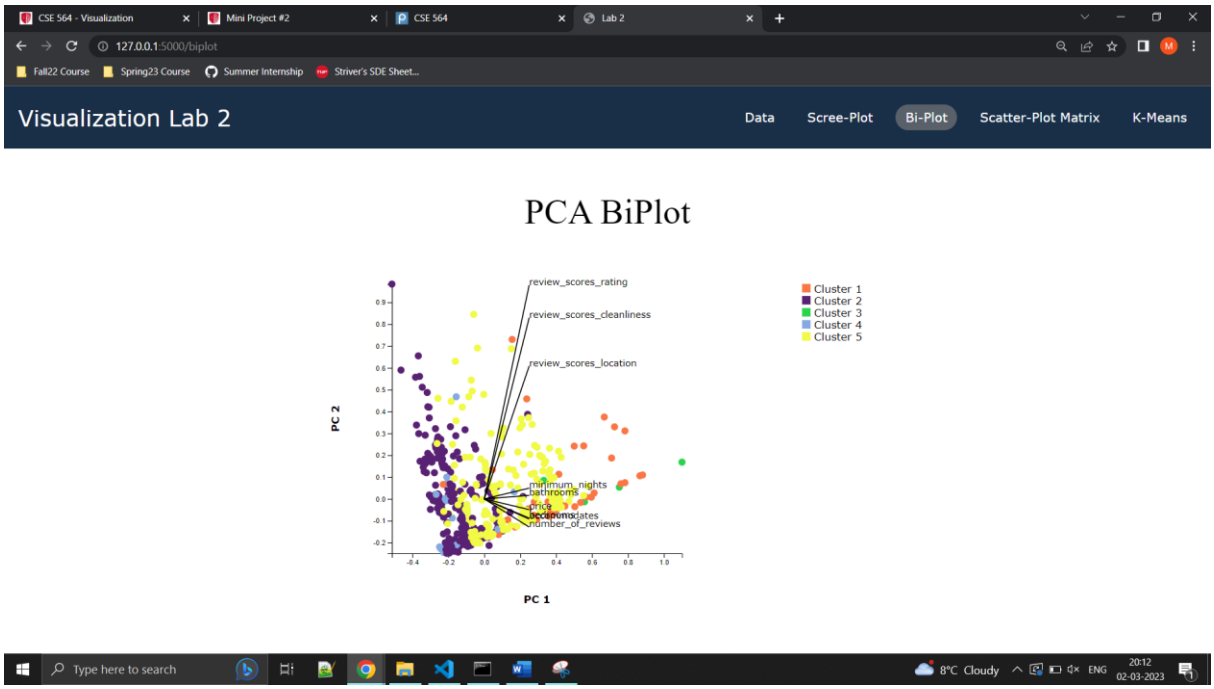
The application opens with the data page that provides information about the dataset. The navigation bar provided ease to switch between tabs. The user can select the type of graph, and the corresponding page will be displayed. The project has 5 HTML files, each corresponding to a different tab.



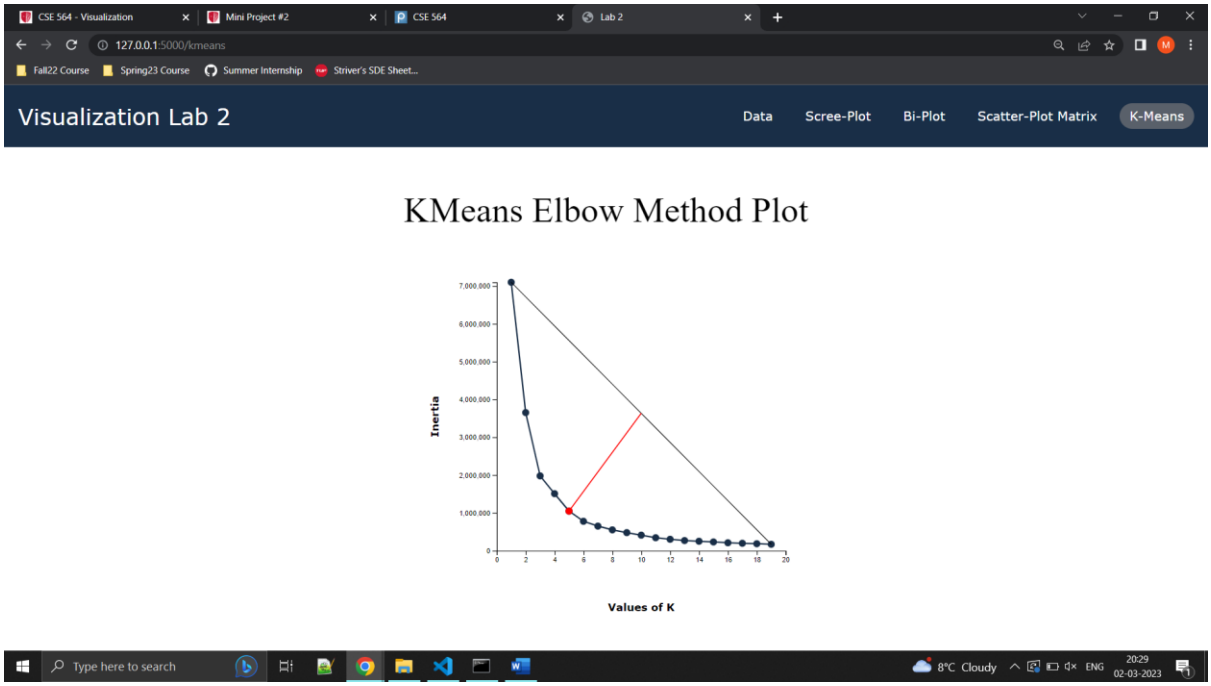
When the user selects a scree plot, A Scree Plot will be displayed where the user has the option of selecting the dimensionality index. The dimensionality index is used to find out the top 4 features. On hovering over a particular value, a tooltip is displayed that shows values. According to the value chosen by the user, the scatter-plot matrix page will get updated.



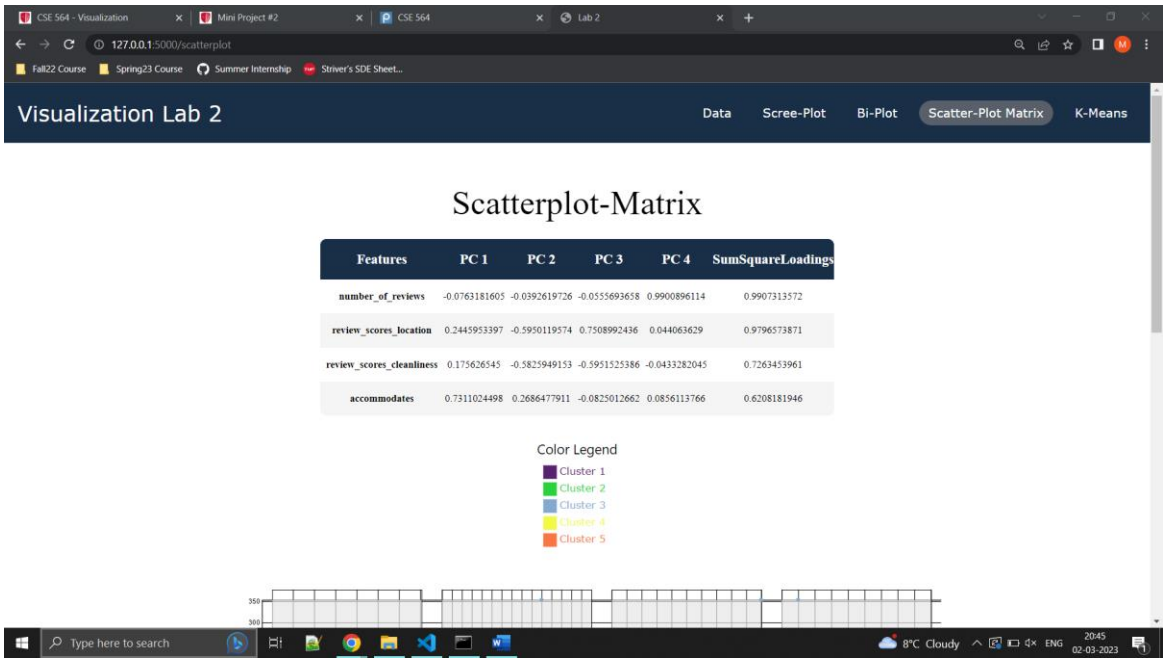
The bi-plot tab displays the top two principal coordinates and shows an arrow for loadings of each feature. The dots are colored according to the cluster they belong. A legend is also displayed for easy reference. Furthermore, a tooltip is also provided for feature names.

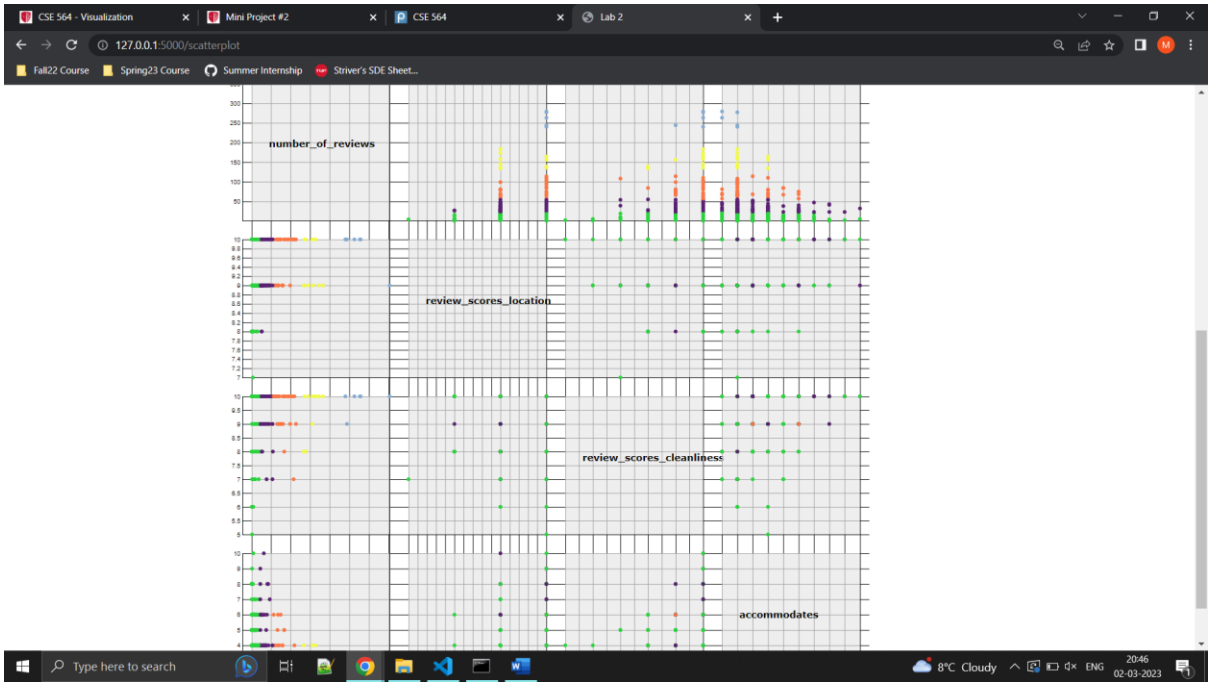


The k-Means tab displays the elbow method used for selecting the optimal value of K. The Y-axis displays inertia while the X-axis the value of K. According to the elbow method, the optimal value of **K is 5.** The highlighted bar shows the selected value of K.



Scatterplot Matrix Tab displays two things. The first part displays a table that shows the top 4 features according to their contribution to each Principal Component. The dimensionality index is used to select top PCA components and values are displayed in sorted order according to the sum of square loadings. The second part shows all possible bivariate scatterplots arranged into a matrix. Each scatterplot is color-coded according to the clusters. A color legend is displayed for clusters.





Interesting Observations:

When we select the dimensionality index as 4, 80% variance is conserved. According to the sum of the square of loadings, the top 4 features are number_of_reviews, review_scores_location, review_scores_cleanliness, and accommodates. These features play an important role in a property.