Visualization assignment 2

Part 1

Dataset:

https://archive.ics.uci.edu/ml/datasets/wine

Video Link:

https://youtu.be/Os54aSR78BQ

About Data:

This dataset is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Originally the dataset was supposed to be used for testing classification models. But in this assignment I also tried to observe if we get clusters which are similar to the already given classification labels.

Attributes:

The original dataset had some possible spelling mistakes. I have added corrected spelling if someone finds the given attribute names difficult to understand. All attributes are numerical.

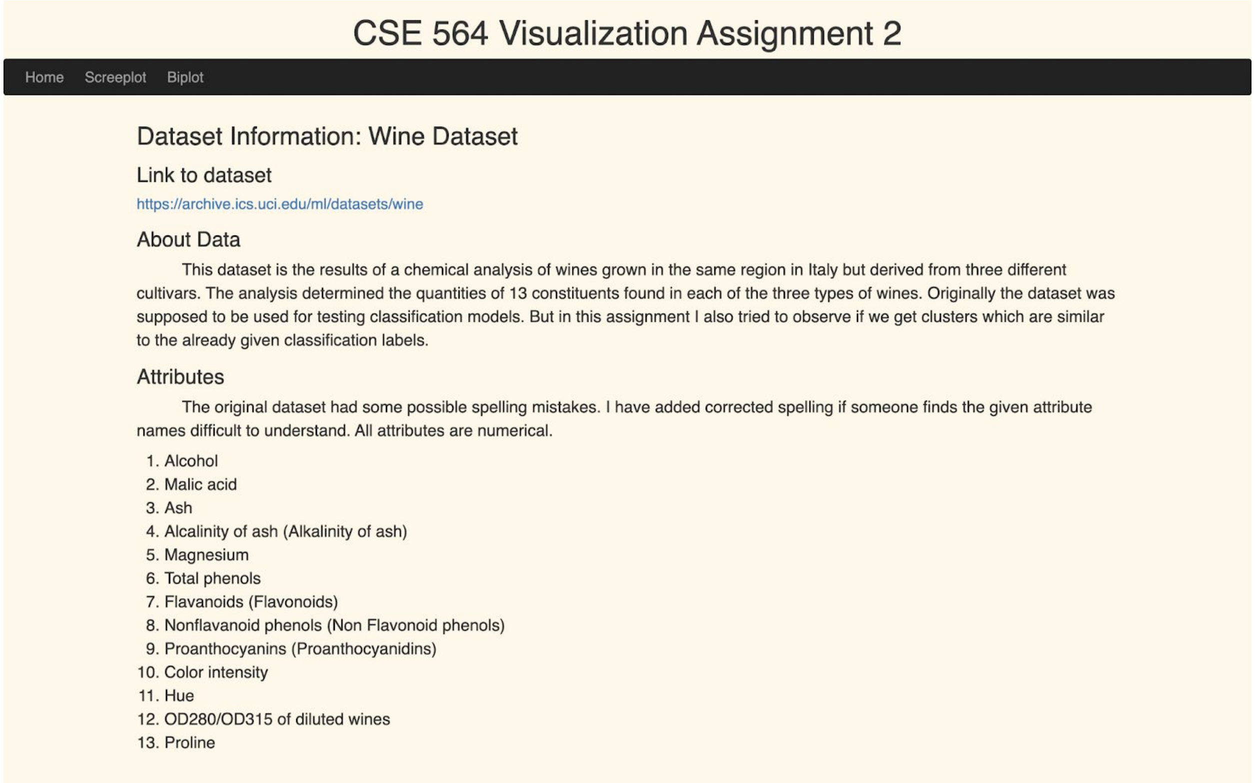
1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash (Alkalinity of ash)
5. Magnesium
6. Total phenols
7. Flavanoids (Flavonoids)
8. Nonflavanoid phenols (Non Flavonoid phenols)
9. Proanthocyanins (Proanthocyanidins)
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

Execution instructions:

Run app.py file to start the execution.

Implementation:

Homepage:



CSE 564 Visualization Assignment 2

Dataset Information: Wine Dataset

Link to dataset

https://archive.ics.uci.edu/ml/datasets/wine

About Data

This dataset is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Originally the dataset was supposed to be used for testing classification models. But in this assignment I also tried to observe if we get clusters which are similar to the already given classification labels.

Attributes

The original dataset had some possible spelling mistakes. I have added corrected spelling if someone finds the given attribute names difficult to understand. All attributes are numerical.

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash (Alkalinity of ash)
5. Magnesium
6. Total phenols
7. Flavanoids (Flavonoids)
8. Nonflavanoid phenols (Non Flavonoid phenols)
9. Proanthocyanins (Proanthocyanidins)
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

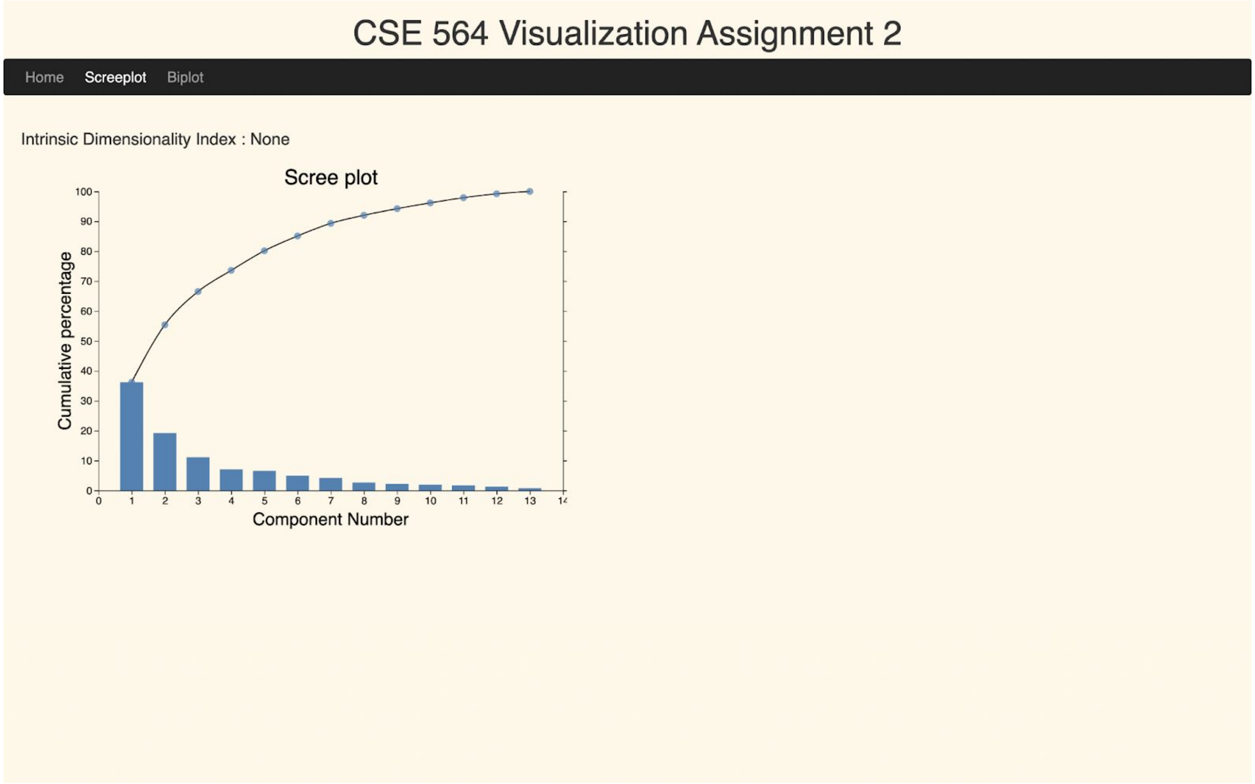
First homepage gives information about the dataset and its attributes.

Task 1:

1. Scree Plot:

After clicking on the ‘Screeplot’ in the navbar base scree plot will be displayed (Image 1). Initially intrinsic dimensionality index isn’t selected.

Image 1



CSE 564 Visualization Assignment 2

Intrinsic Dimensionality Index : None

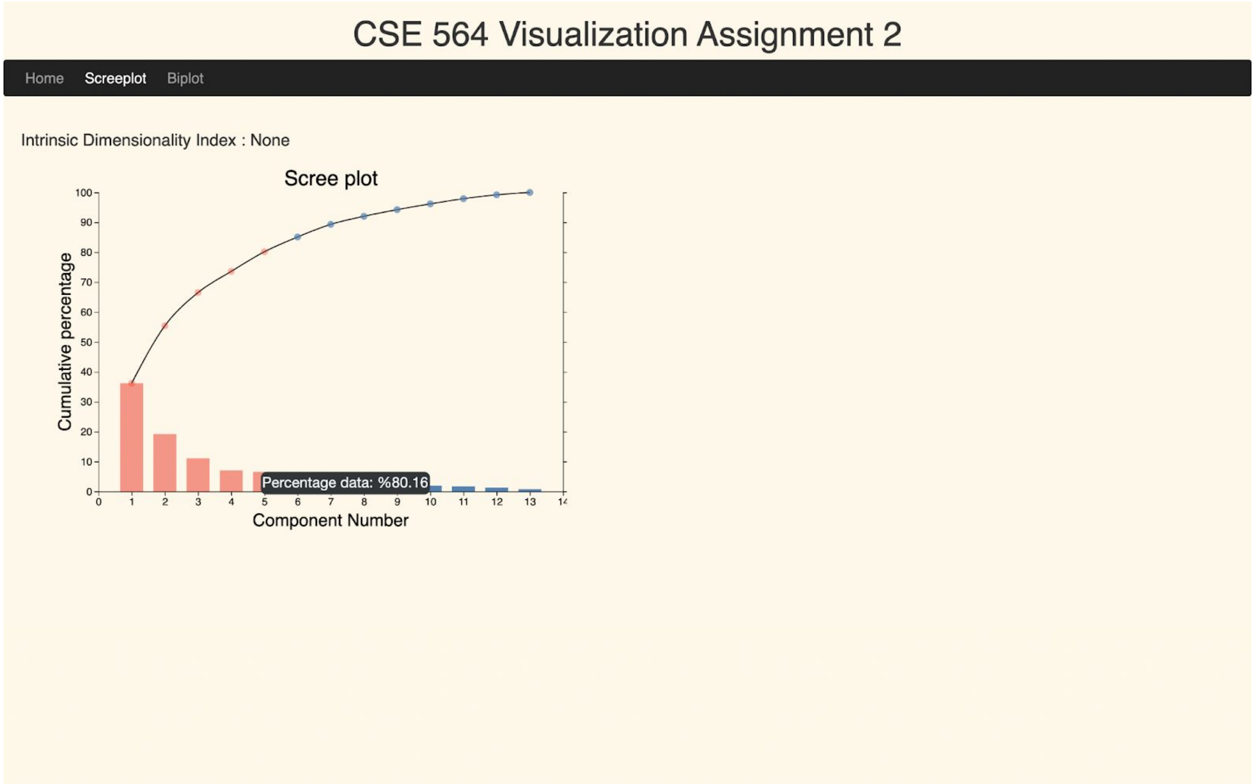
When a user hovers on any of the bars or circles their corresponding percentage value will be displaced in the right side of the pointer (Image 2). This percentage represents the amount of information present in the PCA components until that bar, from left to right. While hovering over the bar, all previous bars and circles also get highlighted to give users a visual understanding of the amount of information they are selecting.

Task 1:

2. Intrinsic dimensionality index(di):

After deciding the intrinsic dimensionality index (di) we can click on either bar or the point above it to set ‘di’. The reason behind keeping bars and circles is that, as the information present in the components decreases, the height of bars also decreases and it’s difficult to select them (Image 3). In that case it’s convenient to select from the point above the bar. After selecting ‘di’ the opacity of the selected bar changes to 1 and color also changes to red (Image 4).

Image 2

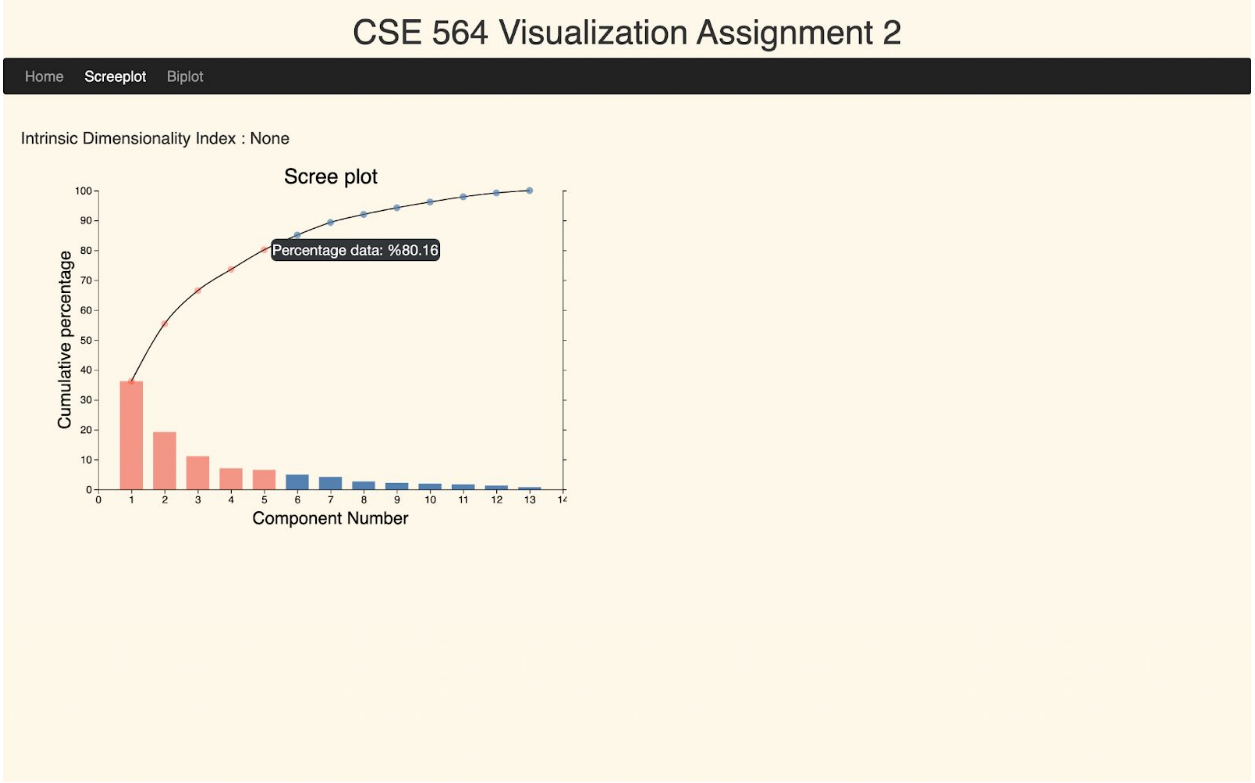


CSE 564 Visualization Assignment 2

Intrinsic Dimensionality Index : None

Percentage data: %80.16

Image 3



CSE 564 Visualization Assignment 2

Intrinsic Dimensionality Index : None

Task 2:

1. Display the table with four important features;

After clicking on the bar or circle ‘di’ will be selected and sent to the backend. On the backend part based on the sum of squared loadings best four features will be decided. These four attributes will be sent back to the frontend with their respective sum of squared loadings and it will be displayed in a table (Image 4).

Task 2:

2. Scatterplot matrix:

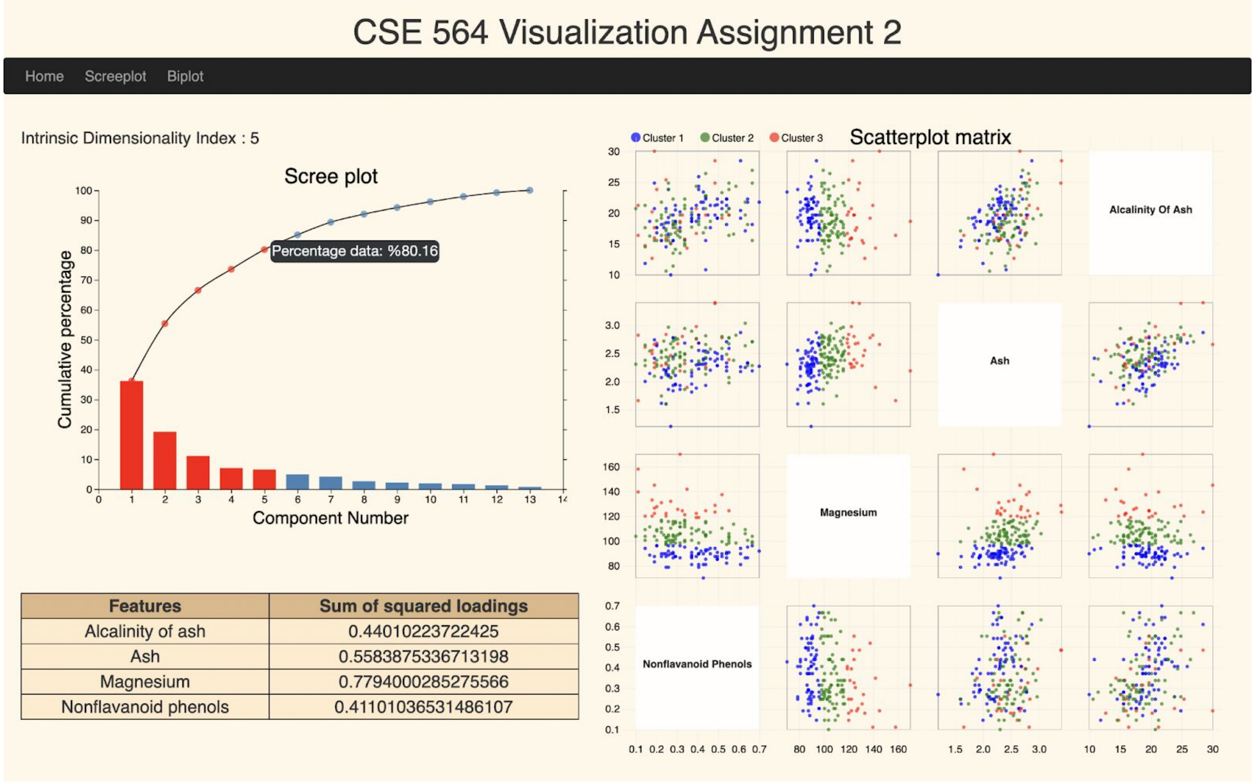
Just like the above table from Task 2.2 the values for those attributes will also be sent to the frontend and a scatterplot matrix of those 4 important features will be displayed (Image 4).

Task 2:

3. K-means:

Along with the scatterplot matrix data labels assigned by K-means algorithm will be sent to the frontend and clusters will be colored in the scatterplot matrix based on these labels. The default cluster size has been set to three. As the original dataset was for classification purposes with 3 output labels, setting K-means clusters to 3 might give interesting results.

Image 4



CSE 564 Visualization Assignment 2

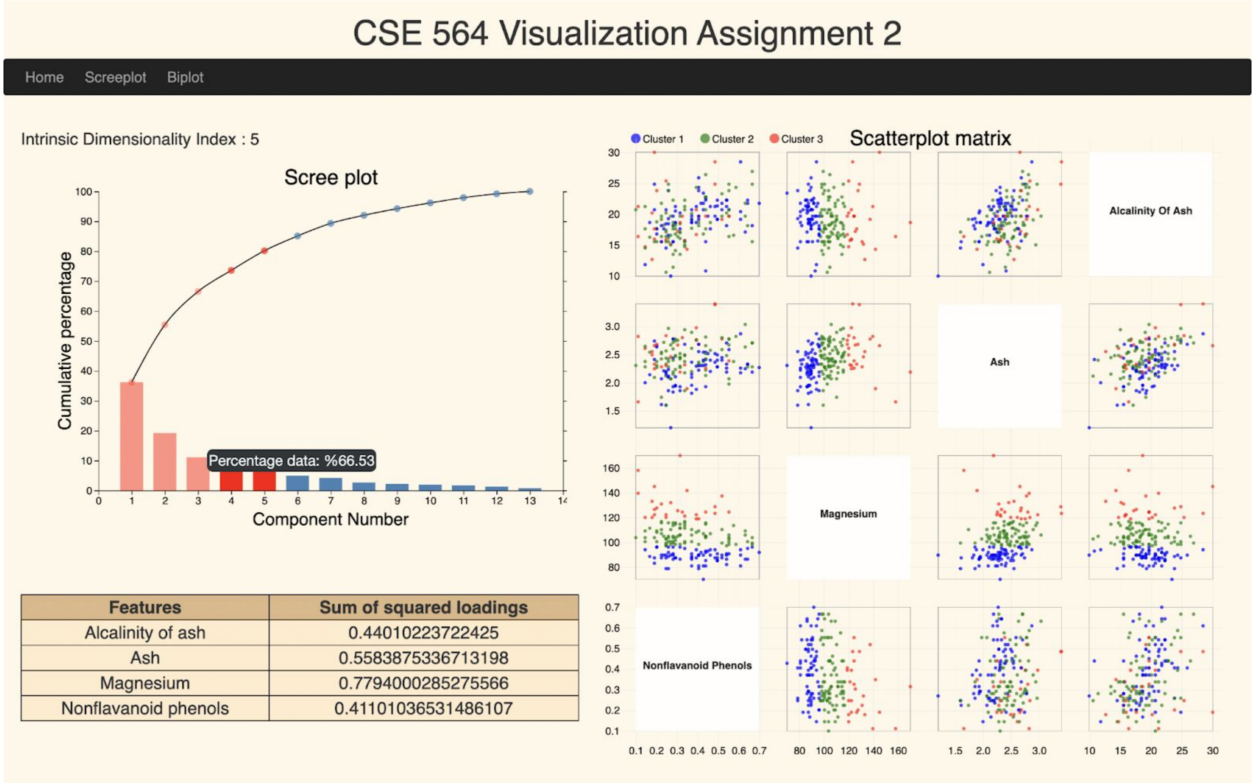
Intrinsic Dimensionality Index : 5

Features	Sum of squared loadings
Alcalinity of ash	0.44010223722425
Ash	0.5583875336713198
Magnesium	0.7794000285275566
Nonflavanoid phenols	0.41101036531486107

Task 2 Additional features:

The table and scatterplot matrix are displayed immediately after selecting the ‘di’ without refreshing the page. If we again select a new ‘di’ value the table and scatterplot matrix will be updated. This can be seen from (Image 5) and (Image 6)

Image 5

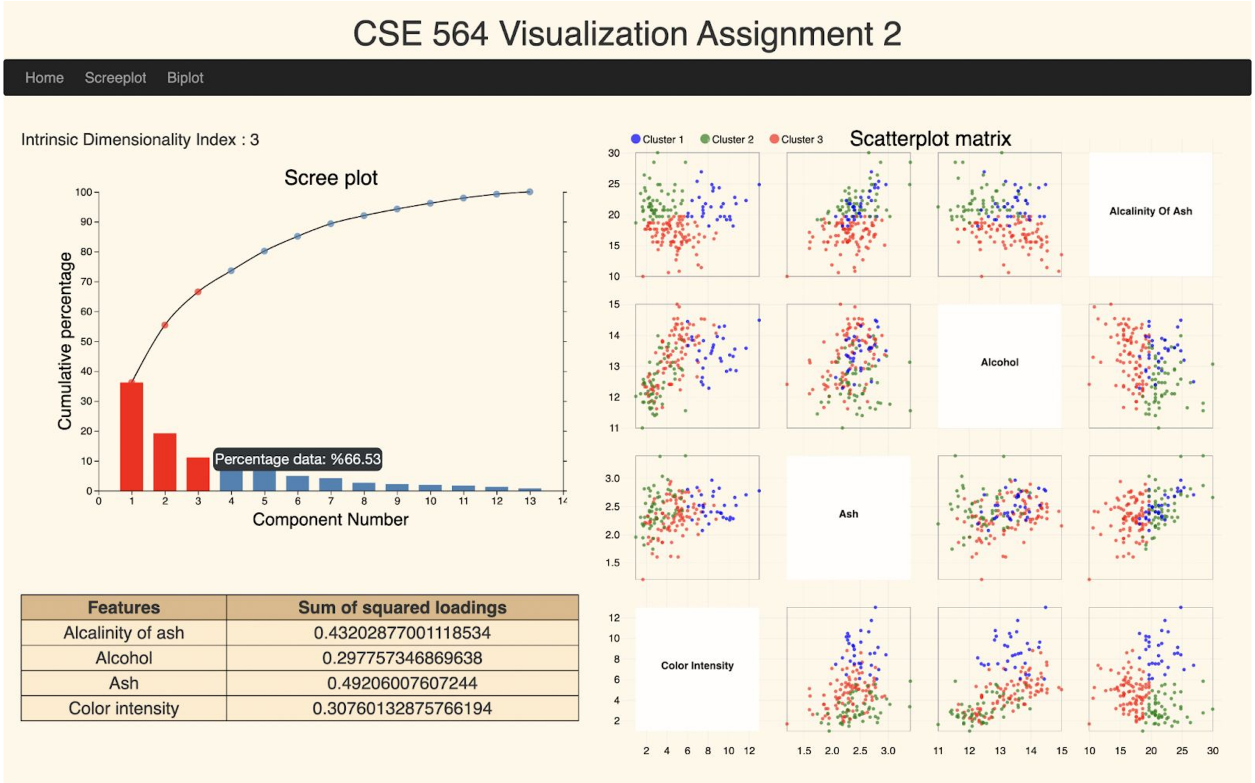


CSE 564 Visualization Assignment 2

Intrinsic Dimensionality Index : 5

Features	Sum of squared loadings
Alcalinity of ash	0.44010223722425
Ash	0.5583875336713198
Magnesium	0.7794000285275566
Nonflavanoid phenols	0.41101036531486107

Image 6



CSE 564 Visualization Assignment 2

Intrinsic Dimensionality Index : 3

Features	Sum of squared loadings
Alcalinity of ash	0.43202877001118534
Alcohol	0.297757346869638
Ash	0.49206007607244
Color intensity	0.30760132875766194

Task 1:

3. Biplot:

To see biplot, click on the biplot option present in the navbar (Image 7). Here X-axis coordinates are based on principle component 1 (PC1) and Y-axis coordinates are based on principle component 2 (PC2). The lines in the middle represent different features and their contribution to PC1 and PC2. If we hover on the tip of the lines represented by red circles, the name of the feature representing that line and its contribution in PC1 and PC2 will be displayed on the tip of the pointer (Image 8).

Image 7

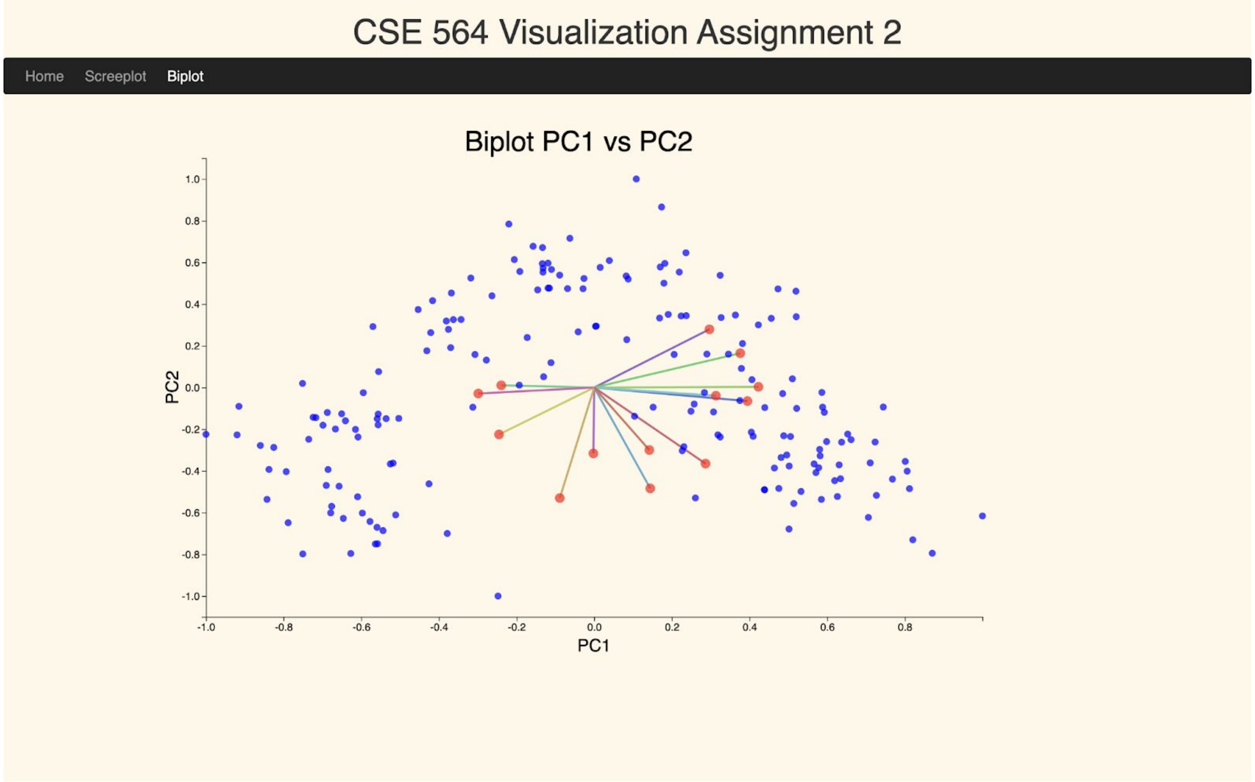
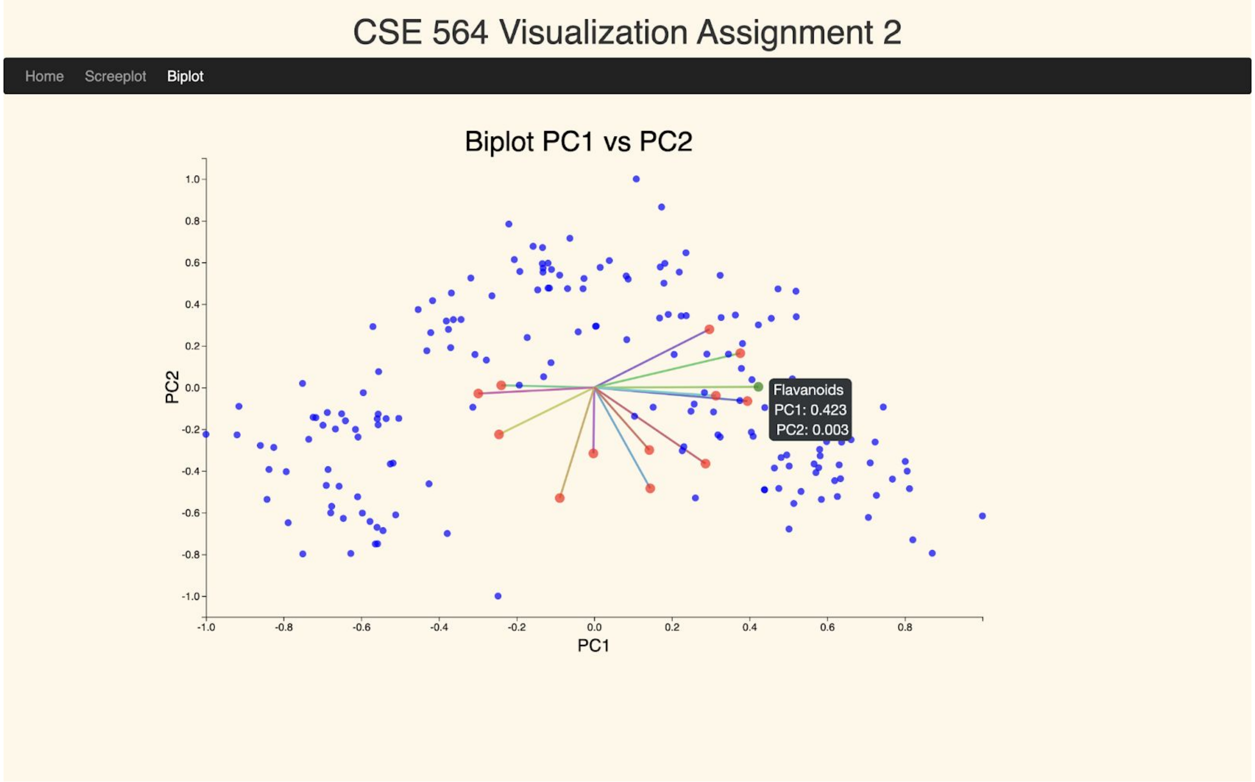

CSE 564 Visualization Assignment 2
Home
Screeplot
Biplot
Biplot PC1 vs PC2
PC2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
PC1

Image 8



CSE 564 Visualization Assignment 2

Part 2

Dataset:

https://www.kaggle.com/seymasa/rice-dataset-gonenjasmine/code

Reason:

The dataset selected for part 1 wasn't giving significant patterns in some plots like Parallel Coordinates Plot which is necessary to understand the basic working of any plot. I have displayed part 1 using this new dataset again. No changes from implementation point of view have been made for part 1 only the dataset is different.

About Data:

The data in the dataset was extracted from two kinds of rice (Gonen, Jasmine). The total number of entries is 18185 and the number of columns or attributes is 12. For the visualisation purpose we have taken only 1% of the original dataset by applying random sampling. The purpose of the original dataset is for checking classification models.

Attributes:

1. Area
2. MajorAxisLength
3. id
4. MinorAxisLength
5. Eccentricity
6. ConvexArea
7. EquivDiameter
8. Extent
9. Perimeter
10. Roundness
11. AspectRation
12. Class

Execution instructions:

Run app.py file to start the execution.