CSE-564 Visualization

Mini-Project#2(Task 1,2) Report

Piyush Soni *SBUID:113273375*

Dataset:

Data for this Lab was selected from Data Expo, which originally consists of flight arrival and departure details for all commercial flights within the USA, from 1987 to 2008. I have filtered the dataset for 1987. There are around 29 attributes out of which I have selected 10 numerical attributes which I thought would be good for this project. During pre-processing of data, I have removed rows with missing values and sample the data randomly for around 500 data points in order to avoid cluttering in plots.

Selected Attribute Information:

DepTime:	actual departure time (local, hhmm)
CRSDepTime:	scheduled departure time (local, hhmm)
ArrTime:	actual arrival time (local, hhmm)
CRSArrTime:	scheduled arrival time (local, hhmm)
FlightNum:	flight number
ActualElapsedTime:	in minutes
CRSElapsedTime:	in minutes
ArrDelay:	arrival delay, in minutes
DepDelay:	departure delay, in minutes
Distance:	in miles

Dataset before processing was pre-processed using Pandas and Numpy libraries of python and after storing in .csv visualized using ***d3.js, jQuery, JavaScript***, ***HTML***, ***CSS*** and for the backend I used ***Flask.*** All the source code, data are zipped in the same folder.

How to Run:

- Go to project directory, first install the requirements.txt using: ***pip install requirements.txt*** (at the start of the project I forgot to create new virtual_env so it might install all the libraries that were installed on my environment)
- open terminal and run: ***python main.py***
- Then open the browser and type ***localhost:<port_no.>*** which would be visible on the terminal when the app starts.

Preprocessing:

Following operations were done during pre-processing of the data:

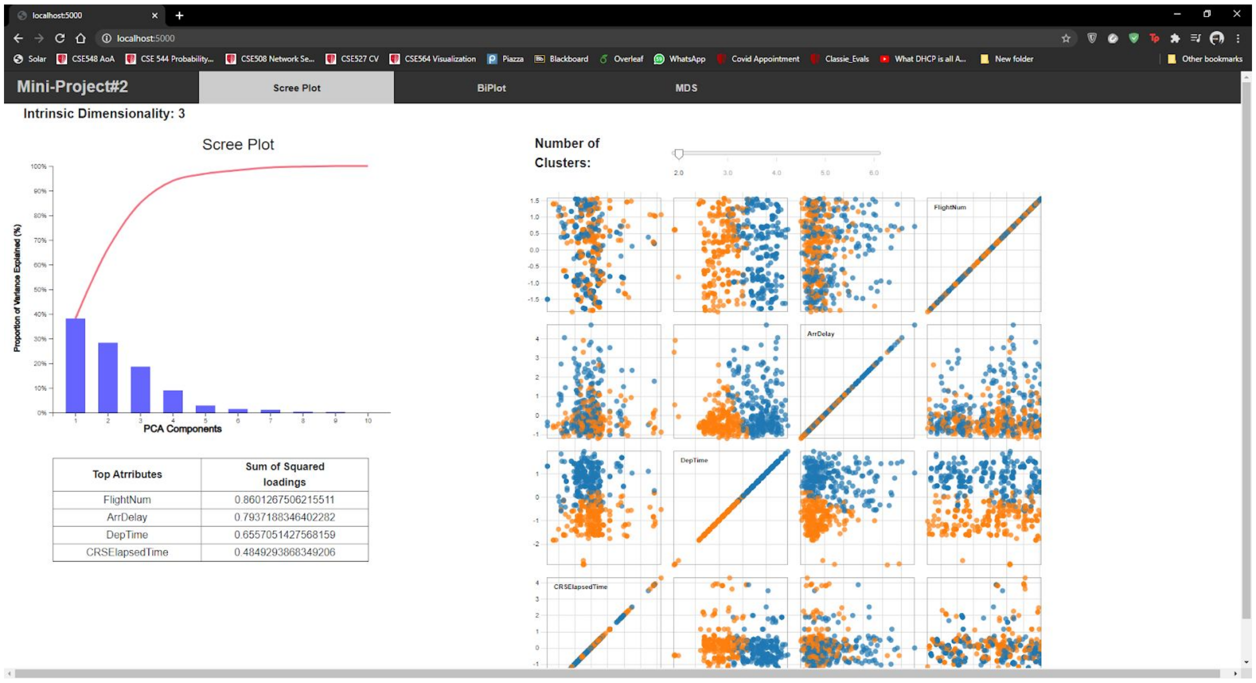
- Remove not so useful columns from the data. I removed around 19 attributes as they were not useful for the visualization task.
- Remove all the rows with missing values.
- Scale the dataset, so that each attribute would be in the same range which in turn helps in visualizing them properly.
- Randomly sample 500 data points from the data so it would help in avoiding cluttering in scatter plot matrix.

Tasks:

Tasks, that were expected from this project were:

Task 1: basic dimension reduction and data visualization with PCA

- This task is achieved with an interaction element. As in my implementation, there would be a default scree plot, Scatterplot Matrix and a table with top 4 attributes according to the highest Sum of Squared loadings.



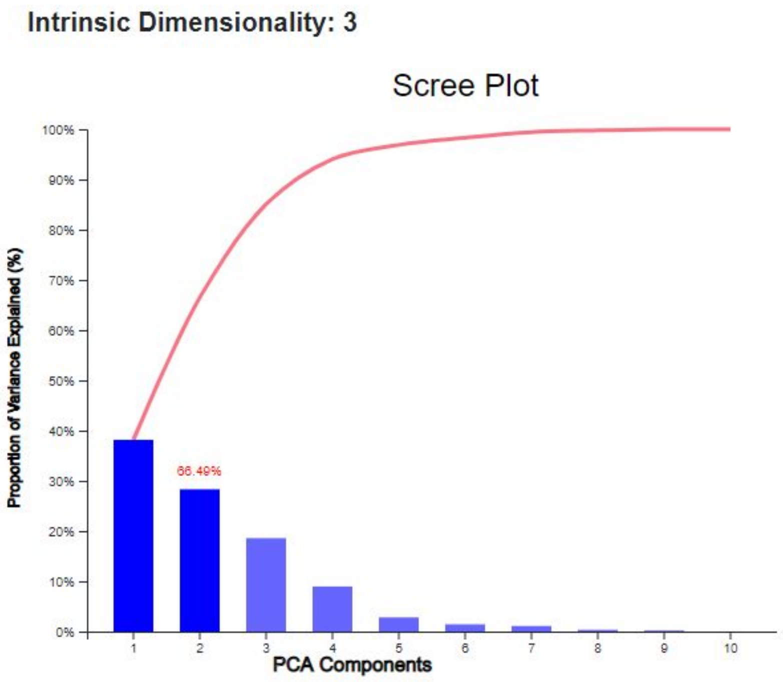
Intrinsic Dimensionality: 3

Top Atrributes	Sum of Squared loadings
FlightNum	0.8601267506215511
ArrDelay	0.7937188346402282
DepTime	0.6557051427568159
CRSElapsedTime	0.4849293868349206

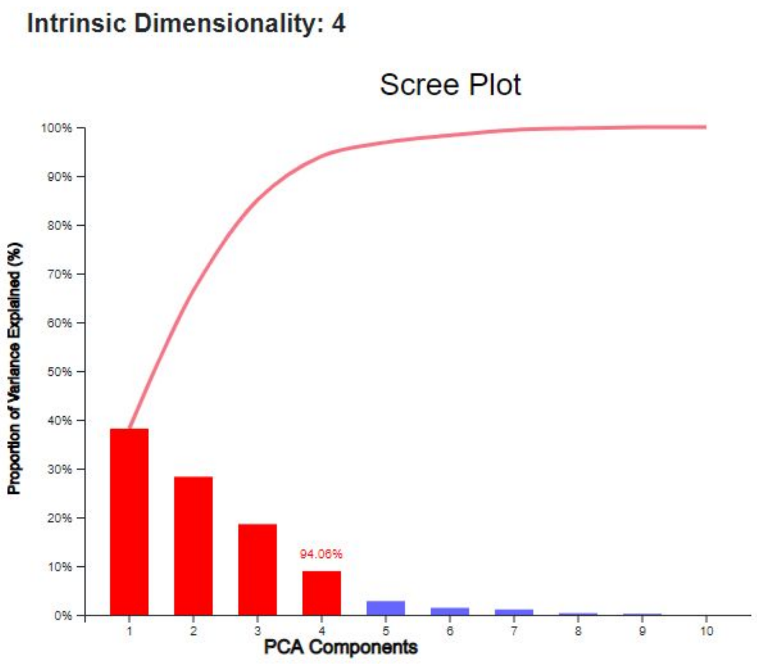
Landing page

- As seen in the above image, the scree plot is the one with a bar chart and a smooth curve, which denotes the cumulative sum of proportion of variance explained by PCA components from left to right. Each Bars in the scree plot represents % of variance explained by PCA components marked on X-axis.

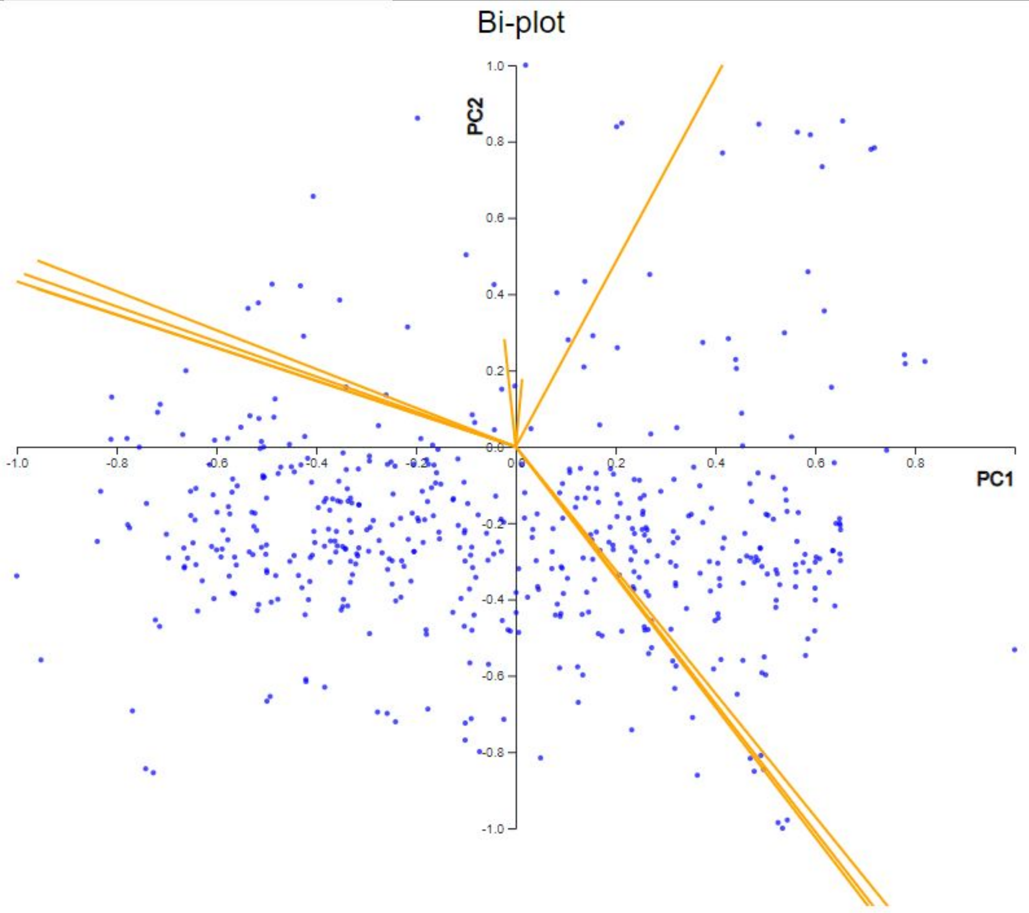
- If a user wants to change the intrinsic dimensionality, he/she can change it by clicking on the bars. If a user hover over the bar it darkens the bar and shows the cumulative % of variance explained till this PCA component on top of it. (default d_i : 3)



- If a user selected a bar it will highlight all the bars corresponding to the PCA components before that chosen in red color, update the the intrinsic_dimensionality on top of the plot, then update the top-4 attributes table according to that, also update the Scatter plot Matrix based on the updated Top - 4 attributes.

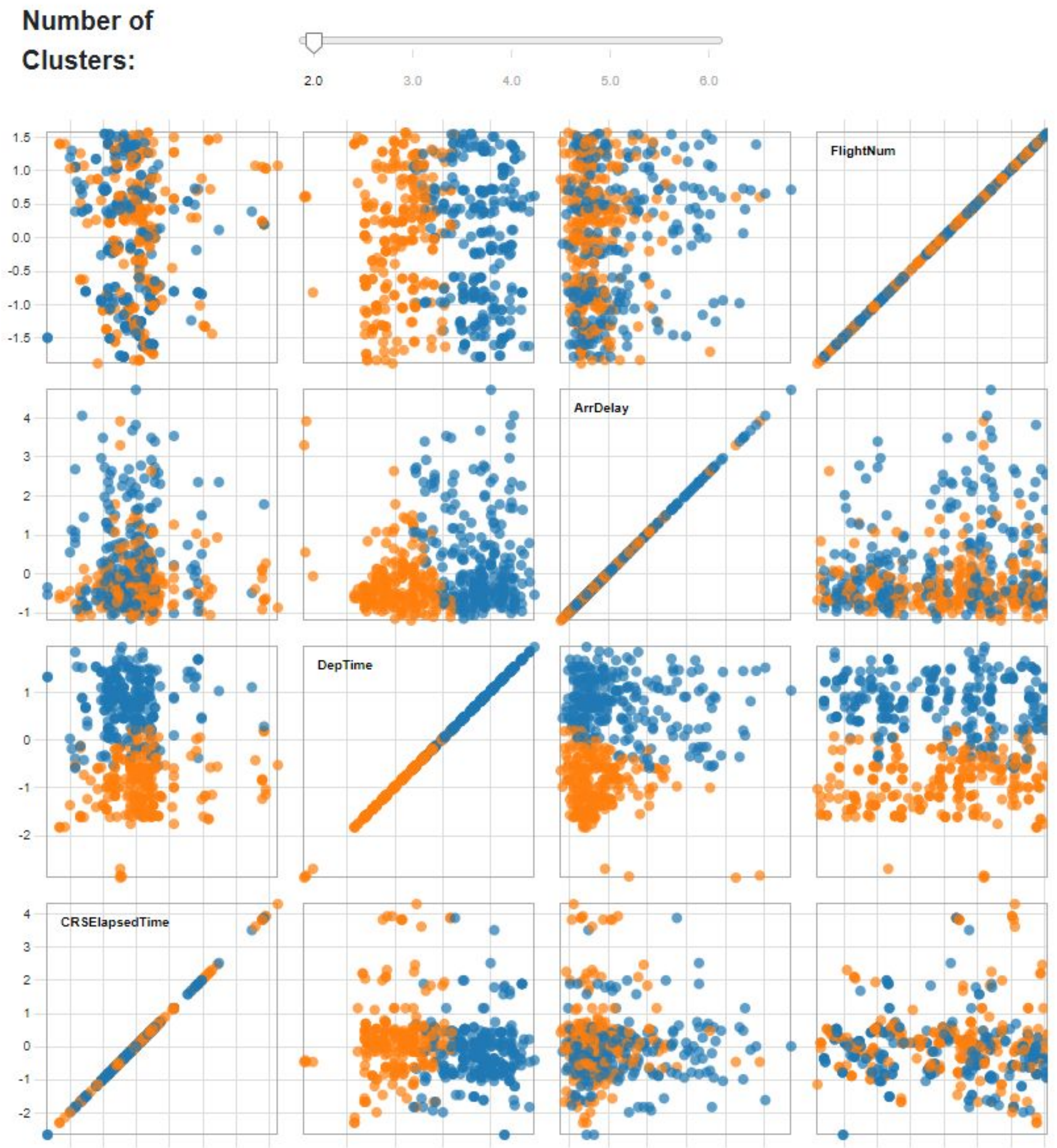


- I have also implemented a Bi-plot which basically plots all the data points in PC1 and PC2 space. It will also plot attributes axises, so that it would be easier for users to view which attributes are more affected by which PCA components. (orange lines shows the attributes in the PC1-PC2 coordinate space)



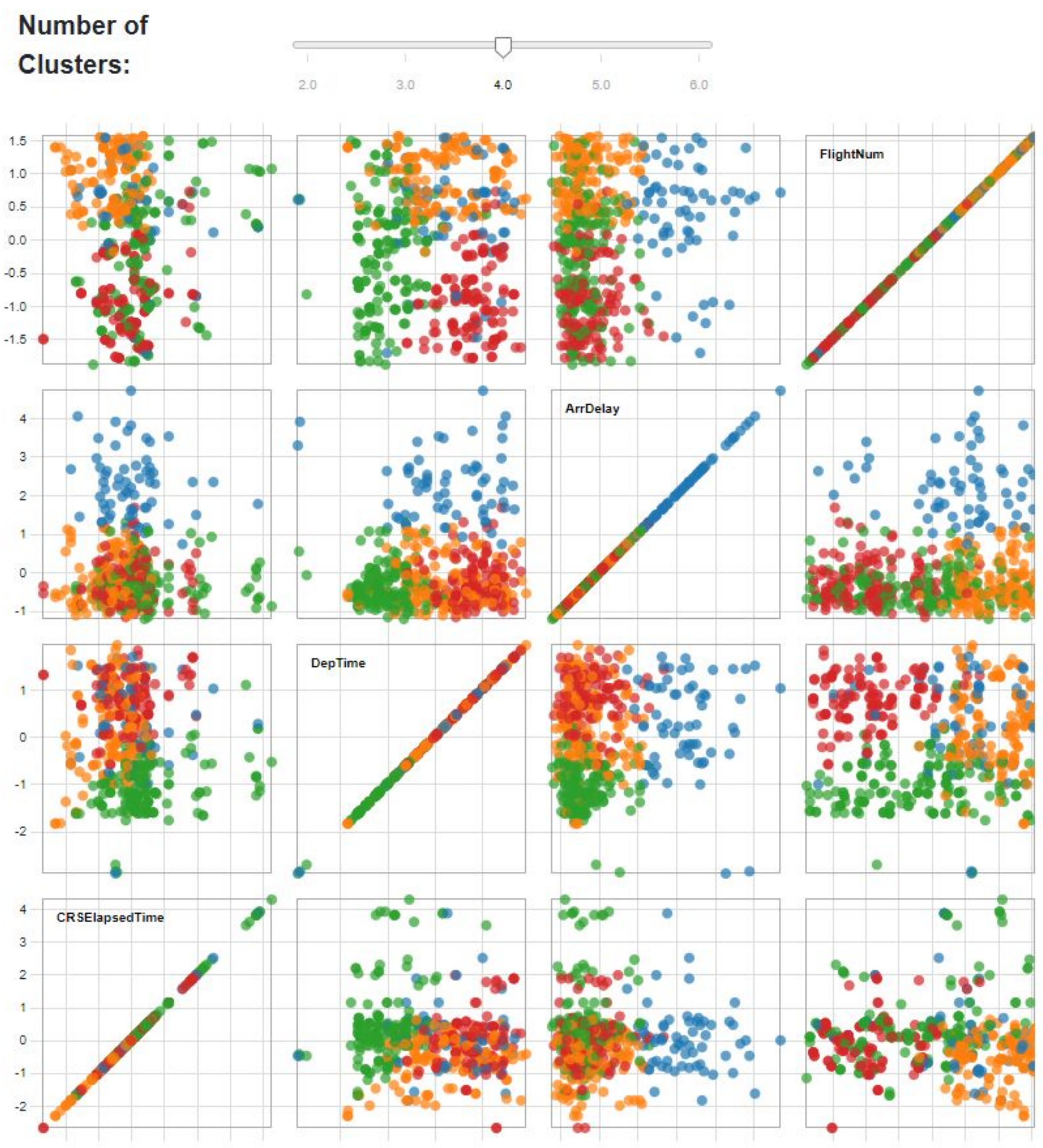
Task 2: Visualize the data with a scatterplot matrix

- In this task, we had to visualize the data points in scatterplot matrix and also show a table showing top 4 attributes according to the intrinsic dimensionality(d_i) that a user selected on Scree plot. Which is done as shown in the picture below. If a user selects a different attribute then this Scatter plot Matrix will also be updated. I chose the default number of clusters as 2. I also provide a slider to choose the number of clusters (calculated by k-means algorithm) as preferred by the user.



Scatterplot Matrix with default clusters, n=2

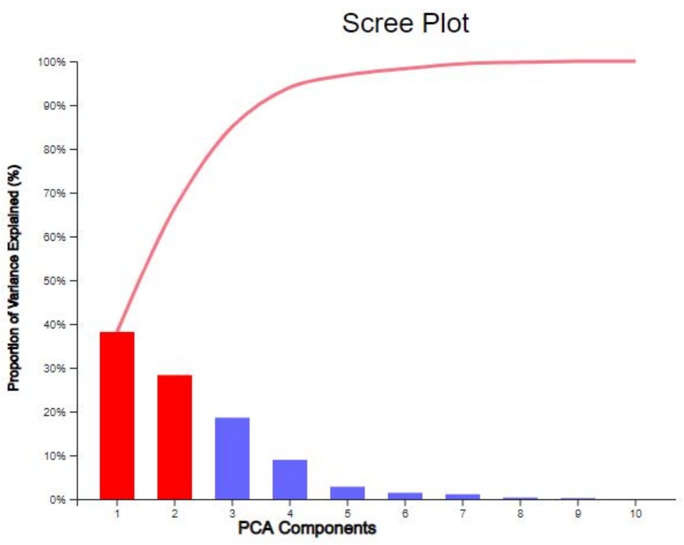
- If the user wants to see more clusters, it can be selected by the slider which is provided on top of the scatter plot matrix. As can be seen in the below screenshot, if a user drags the slider to 4th position, it will show 4-clusters in the Scatter plot matrix.



Scatterplot Matrix with 4 clusters (as chosen on slider on top of it)

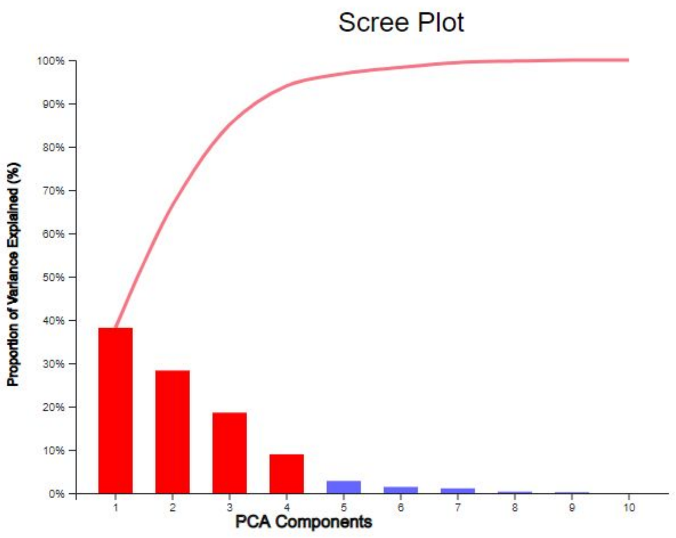
- As asked in task-2, if a user selected an intrinsic dimensionality from scree plot, it would also update a table which shows top-4 attributes based on their sum of squared loadings. It will also show, d_i value on top of the scree plot.

Intrinsic Dimensionality: 2



Top Atrributes	Sum of Squared loadings
ArrDelay	0.5909036287412404
FlightNum	0.48079550912516333
DepTime	0.4397092126603188
CRSElapsedTime	0.3319797382426811

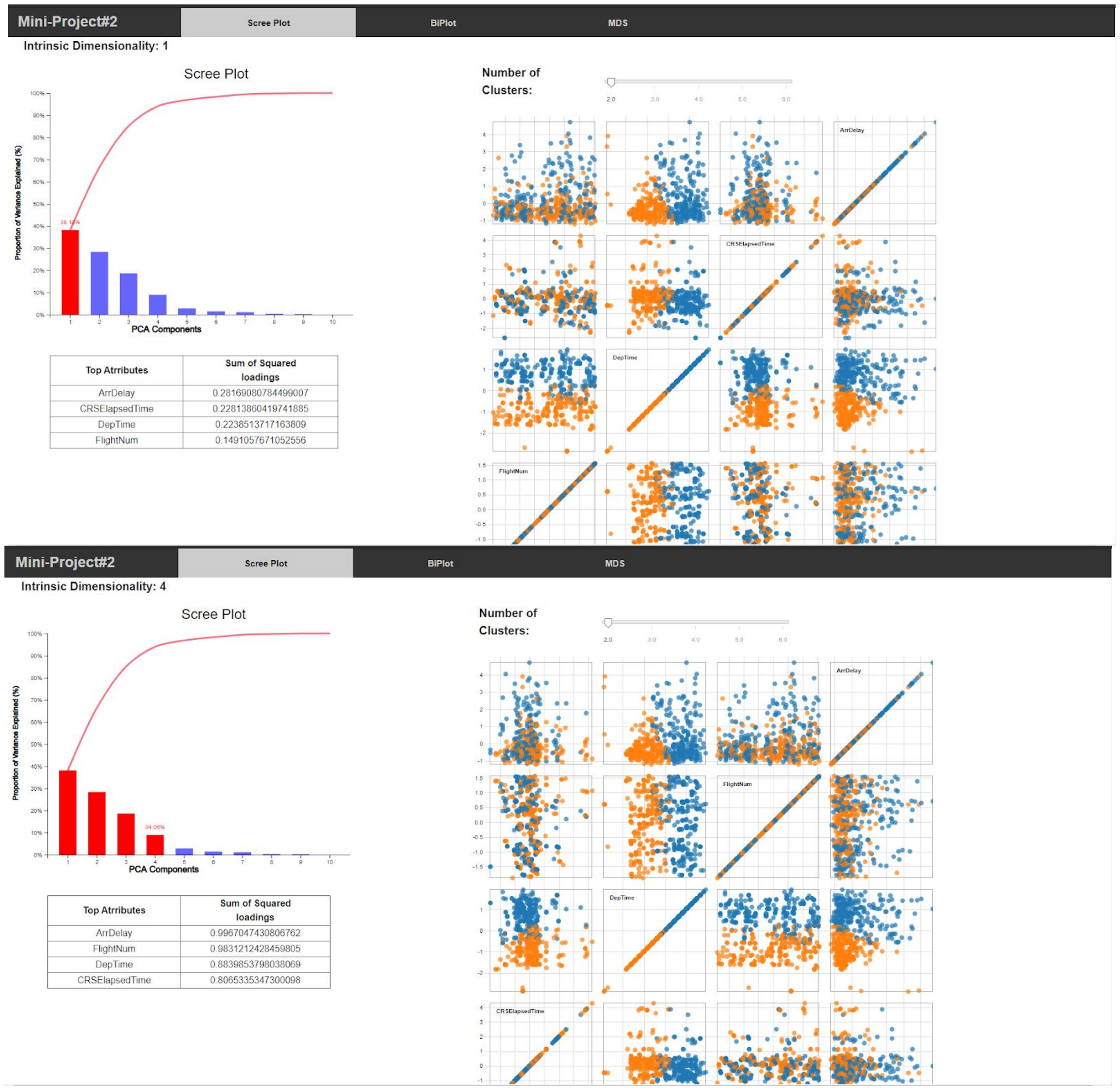
Intrinsic Dimensionality: 4



Top Atrributes	Sum of Squared loadings
ArrDelay	0.9967047430806762
FlightNum	0.9831212428459805
DepTime	0.8839853798038069
CRSElapsedTime	0.8065335347300098

Updated table on selecting new intrinsic dimensionality using Scree plot bars

- Whole update page looks like this when a user clicks on a bar:



Intrinsic Dimensionality: 1

Top Atrributes	Sum of Squared loadings
ArrDelay	0.28169080784499007
CRSElapsedTime	0.22813860419741885
DepTime	0.223851371716809
FlightNum	0.1491057671052556

Intrinsic Dimensionality: 4

Top Atrributes	Sum of Squared loadings
ArrDelay	0.9967047430806762
FlightNum	0.9831212428459805
DepTime	0.8839853798038069
CRSElapsedTime	0.8065335347300098

Before and after a user clicks on the bars to select new di value

- As can be seen in the above image, if a user clicks on the screeplot, intrinsic dimensionality is getting updated, table and scatter plot matrix are also being updated simultaneously based on the new sum of squared loadings values calculated using d_i.