CSE-564 Visualization

Mini-Project#2 Report

Piyush Soni *SBUID:113273375*

Video Link: Youtube Link, Drive Link

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:

- An active connection would be required in order to access d3.js and jQuery libraries.
- *[optional]* Go to project directory, install the requirements.txt using: ***pip install -r 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.

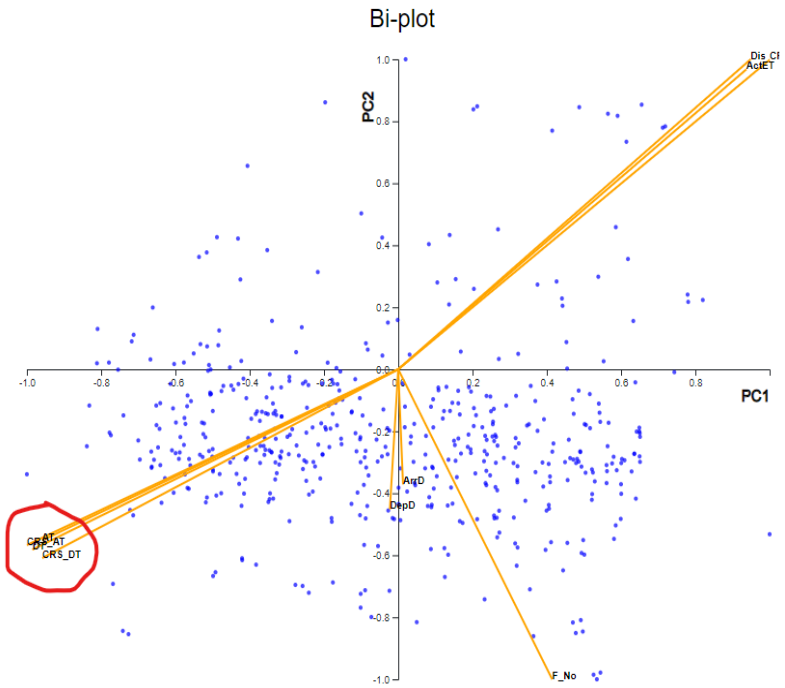
Preprocessing:

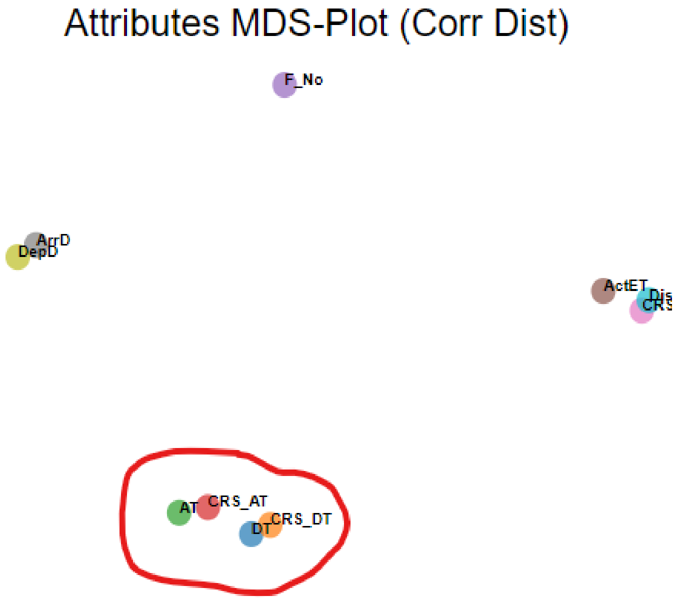
Following operations were done during pre-processing of the data:

- 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.

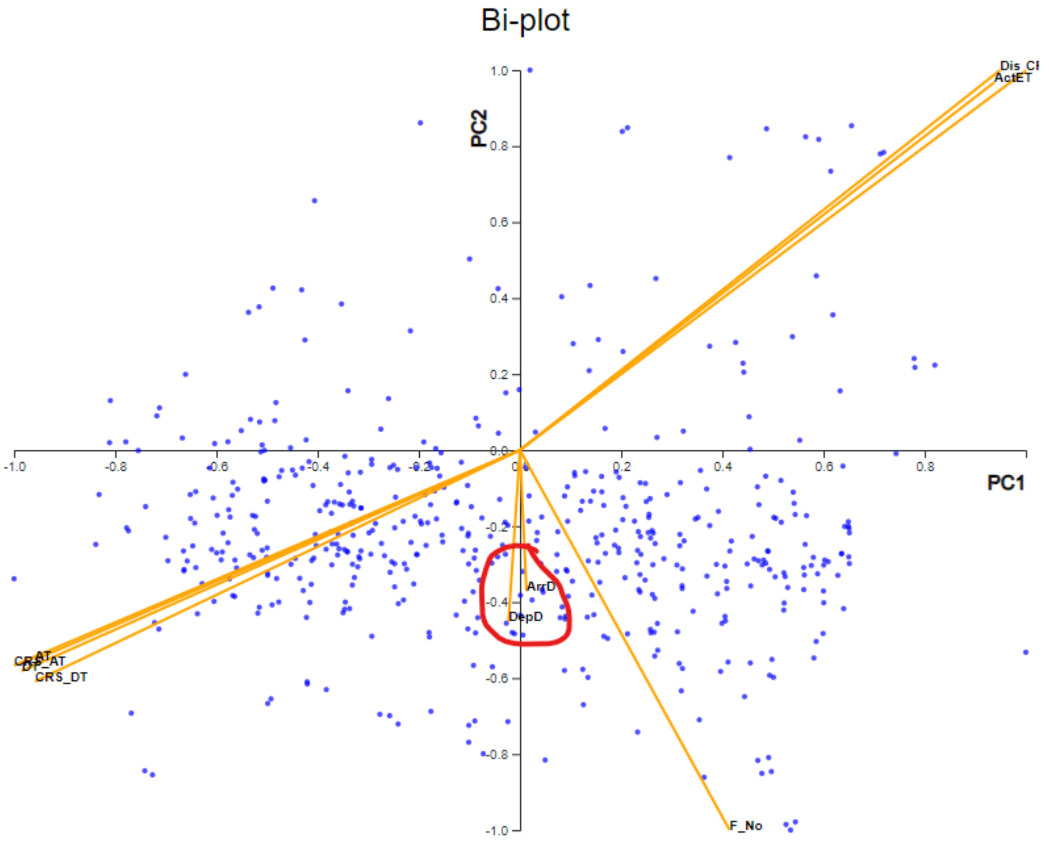
Observations:

- From Biplot we can see that CRS_DT, CRS_AT, AT, DT attributes are highly correlated to each other which can be cross checked from MDS-Correlation plot.

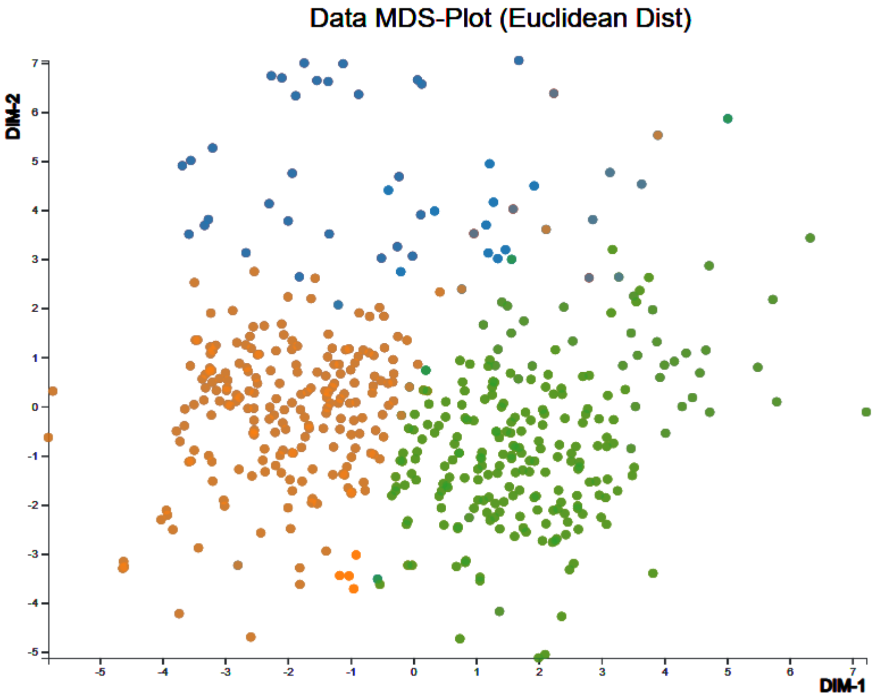




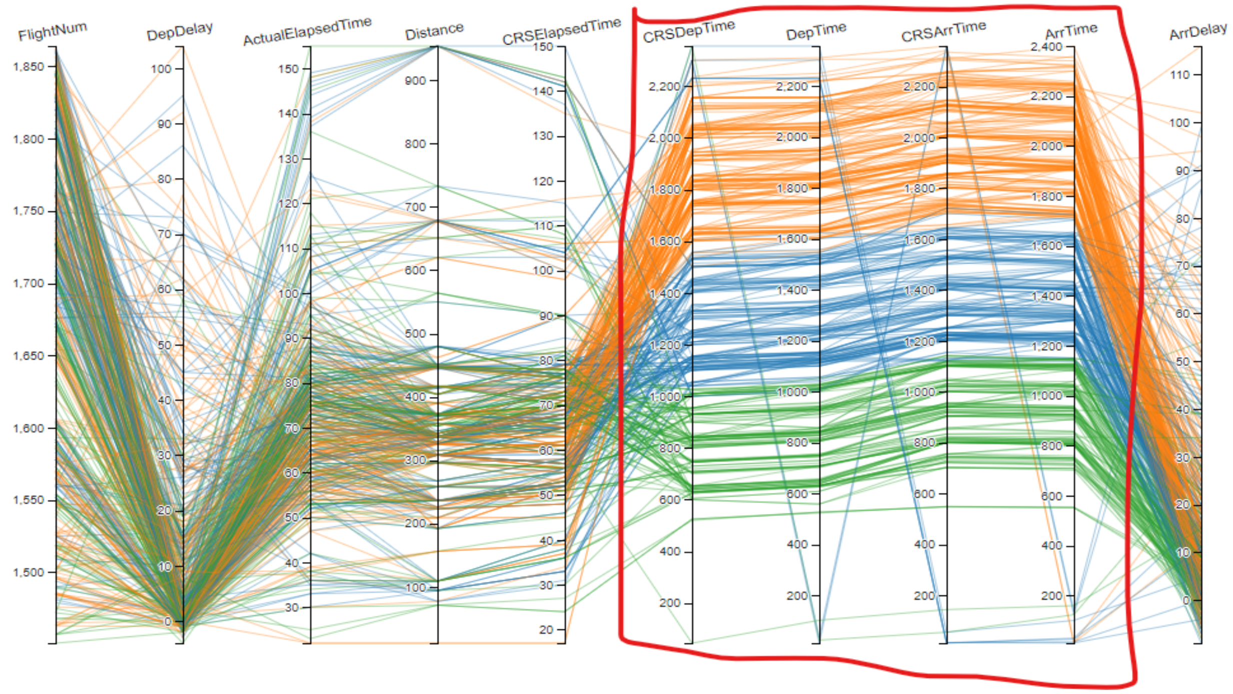
- ArrDelay and DepDelay are very influenced by PC2.



- Clusters forming in MDS with metric Euclidean Distance.

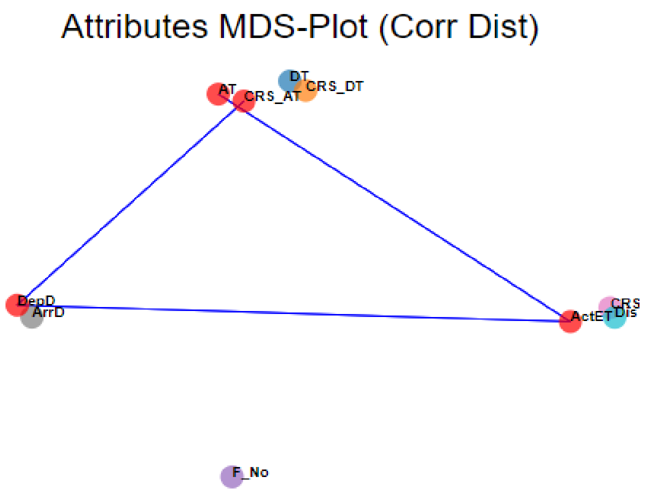


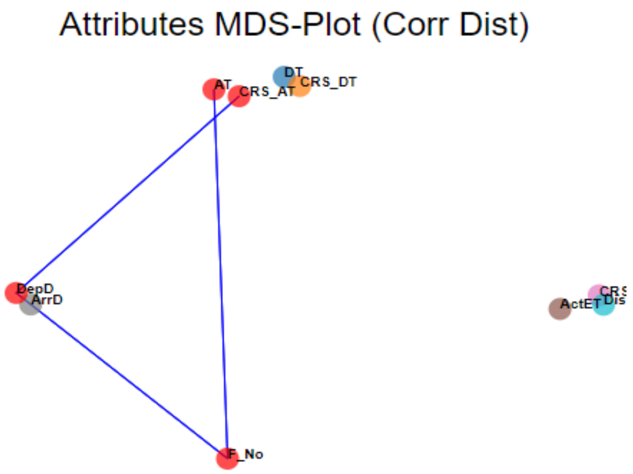
- CRS_ArrTime, ArrTime, DepTime, CRS_DepTime are seems to be correlated to each other, as the lines between these access in parallel coordinate plots are parallel to each other, which was also evident from Bi-PLot and MDS-Corr plot.



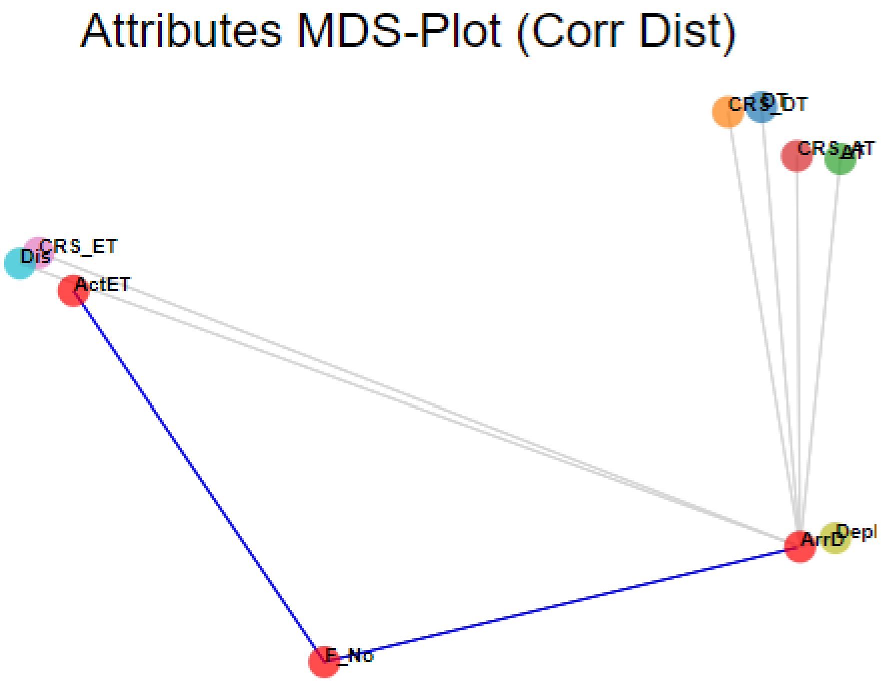
Implementation

- For the Bonus task, I implemented stack data structure, in order to make interaction more practical for the user. While selecting attributes on the MDS Corr plot, if user decided to change the sequence in middle of it, then the user can pop the element in the sequence by clicking back on it in the reverse order.





- During making a sequence on MDS-Corr plot for PCP ordering, if a User hover over a point it will automatically show all the possible attributes for the user to which user can move, in grey lines, in order to complete the sequence.


Attributes MDS-Plot (Corr Dist)
CRS_DT
DT
CRS_AT
AT
CRS_ET
Dis
ActET
ArrD
Depl
F_No