Lab Report - 2
CSE-564: Visualization

By: Naman Banati

Aim:

(i) To do basic dimension reduction and data visualization with PCA using Interactive Scree Plot and PCA- based Biplot

(ii) To use the PCA components < user selected dimensionality index and obtain 4 attributes with highest PCA loadings. Plot the attributes using a scatterplot matrix using k means clustering.

(iii) To construct MDS data plot using Euclidian distance using scatterplot (with points being clustered) and a MDS variable's plot using 1-|correlation| distance

(iv) To visualize the data in a Parallel Coordinates Plot (both numerical and categorical data) with user moveable axis ordering and polylines colored by cluster ID

Data Set: The same dataset used in Lab -1 is used for this lab also. The "Housing New York Units by Building" data set was taken from the NYC OpenData web portal and is provided by The Department of Housing Preservation and Development (HPD). HPD provides reports on building units and the features related to them. The Dataset was last updated on November 24th, 2020. The dataset has 4,732 records of building projects and 42 attributes describing them.

Link to Dataset: https://data.cityofnewyork.us/Housing-Development/Housing-New-York-Units-by-Building/hg8x-zxpr

How to run:

1. A python virtual environment was created in which the flask application was hosted. The flask application acts as the backend application. Navigate to the 'backend' directory and execute the following commands:
   i) source env/bin/activate
   ii) pip3 install -r requirements
   iii) export FLASK_APP=app3.py
   iv) flask run
   This will start the backend server

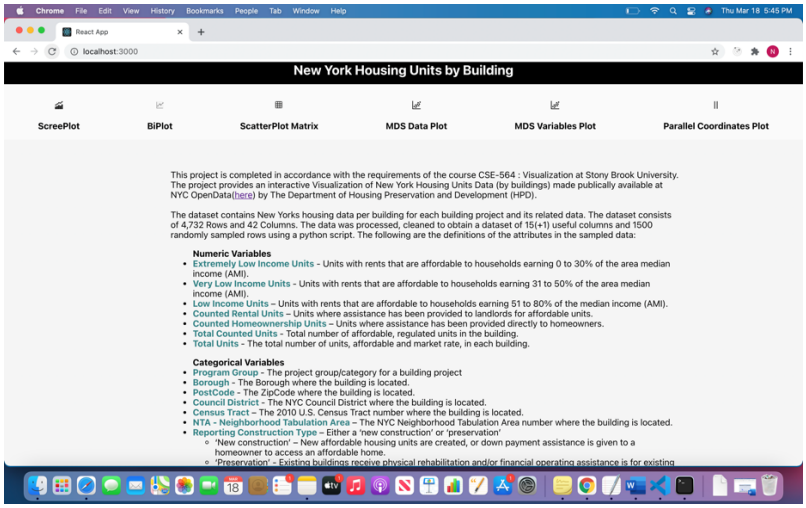
2. The front end of the application is made using React and d3. Each element on the application has been divided into component, molecules and atoms depending upon their usage to ensure modularity and code reusability. To run the application, navigate into the "frontend /visualization" folder.

   For the development run, type in "npm i" to install all react specific libraries. Then type in "npm start" to start the server.

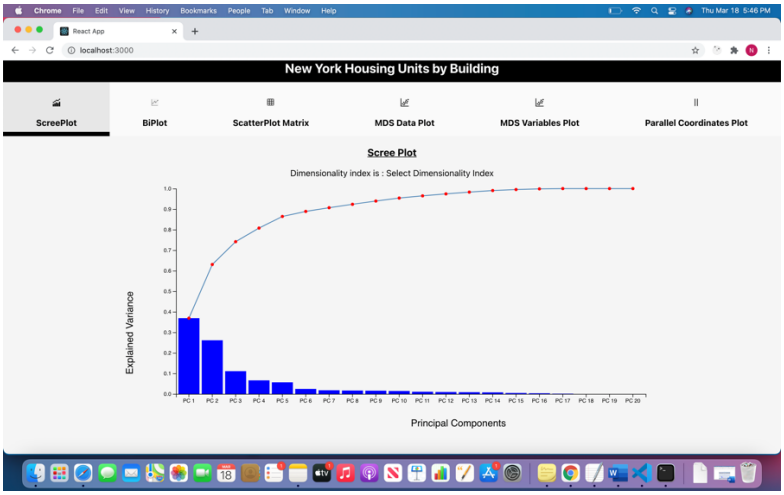
   Please have an active internet connection.

Implementation:

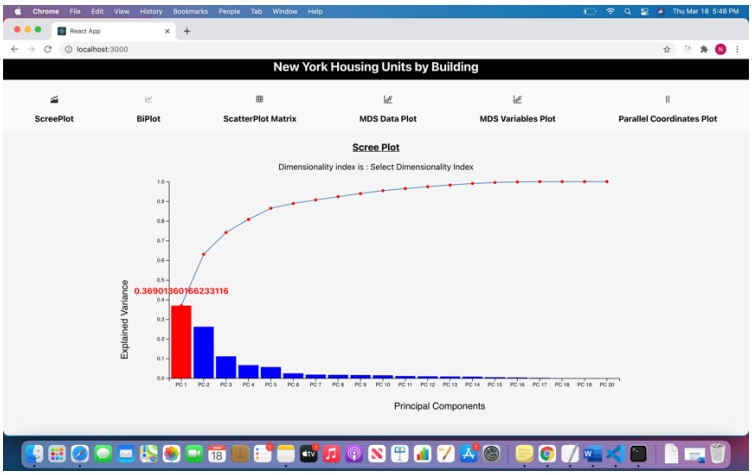
The home page of the application displays the details about the Data Set and the chart options present. A user has the option of plotting a ScreePlot, BiPlot, ScatterPlot Matrix, MDS Data Plot, MDS Variable's Plot or a Parallel Coordinates Plot. The user can click on the dashboard options to plot the charts.



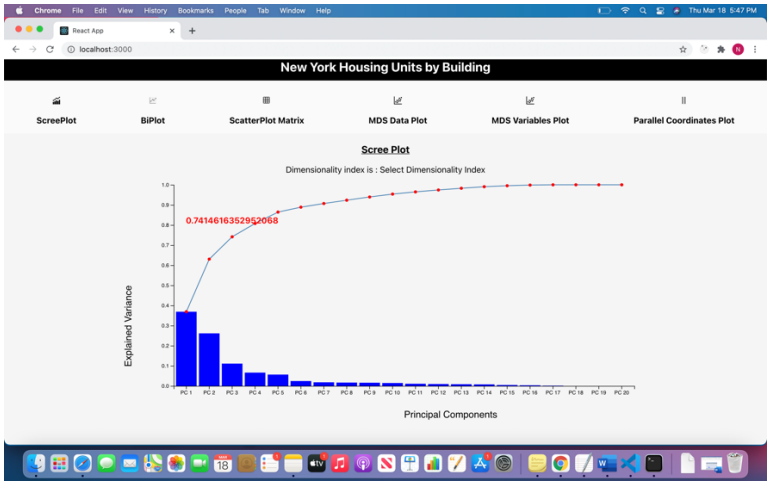
If the user selects a scree plot, a scree plot is plotted of the various Principal Components VS the Explainable Variance. A line chart (of screeplot) shows the cumulative variance. The cumulative explainable variance if all the principal components are selected is always equal to 1 or 100%.



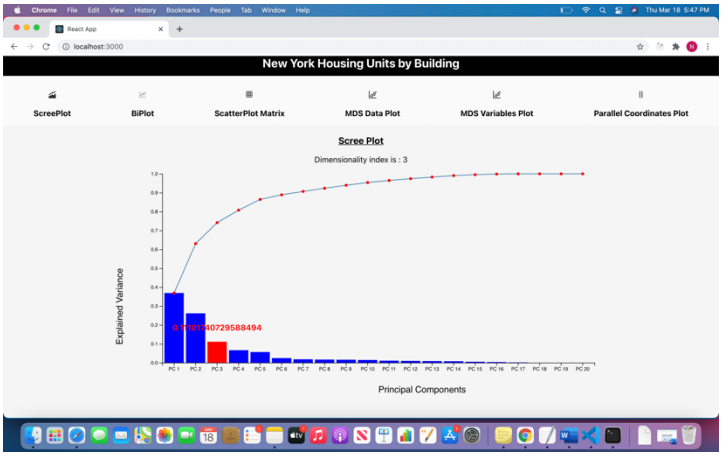
A user can hover over the bars of the scree plot to see the explainable variance as a pop-up text. Doing so also changes the color of the bar to red to give it a highlight effect.



A user can also hover on the dots on the line chart (in the scree plot) to see the cumulative explainable variance till that principal component.

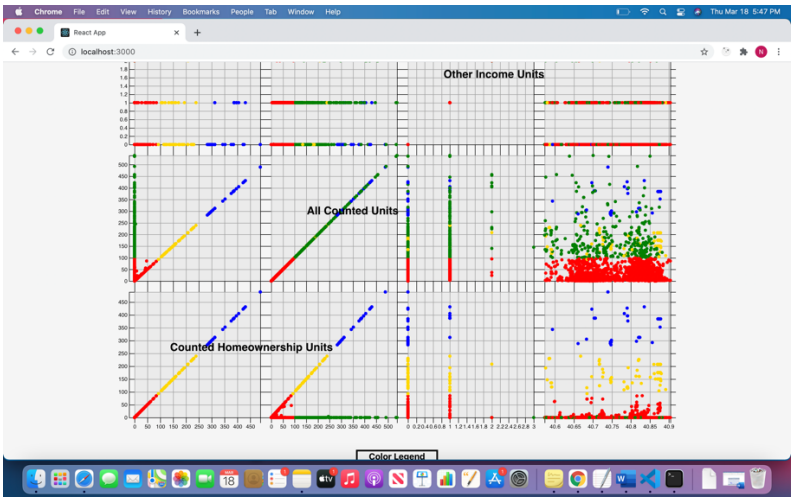


A user can click on any bar of the principal component to select the dimensionality index. The di is then show as a text on the screeplot.



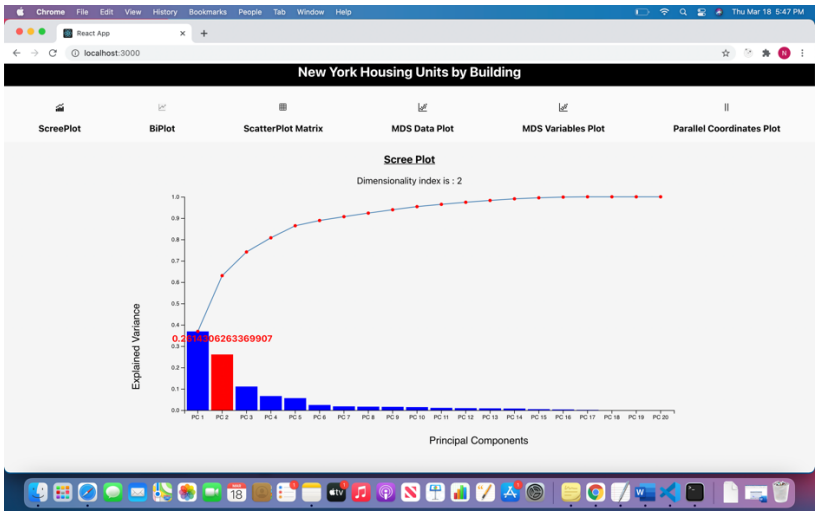
Now if the user selects to plot the scatterplot matrix, the dimensionality index is used to select the number of principal components used for calculating the sum of squares of loading. The 4 attributes/features with the maximum sum of square of their PCA components (loadings) are selected to plot the scatterplot matrix.

Since the user selected 3 as di, the first 3 PCA components are show with their loadings. The sum of square of loadings is used to determine the top 4 features.

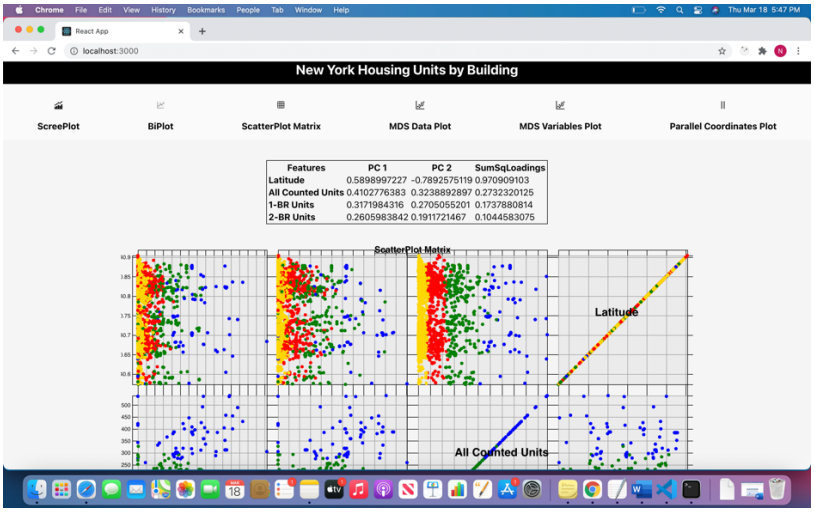


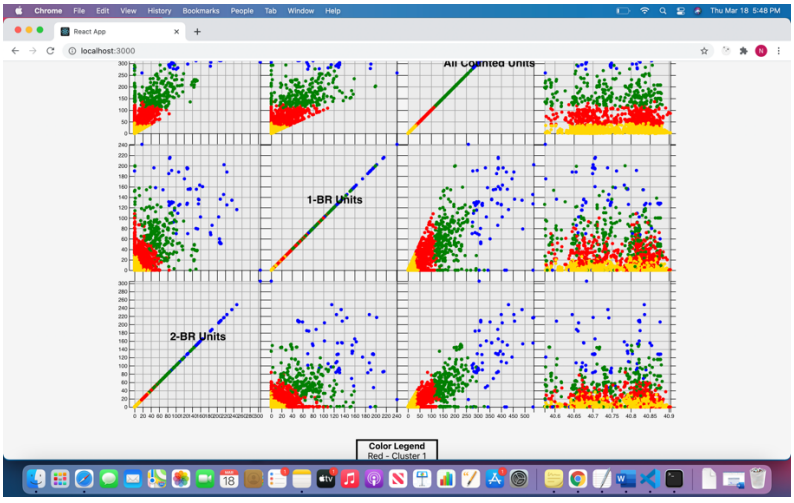
A 4x4 scatterplot matrix is plotted. K- means clustering is being used to cluster the points as well as for allocating a color. A color legend is provided for each cluster below the plot.

A user has the flexibility of selecting the di again any number of times from the scree plot.

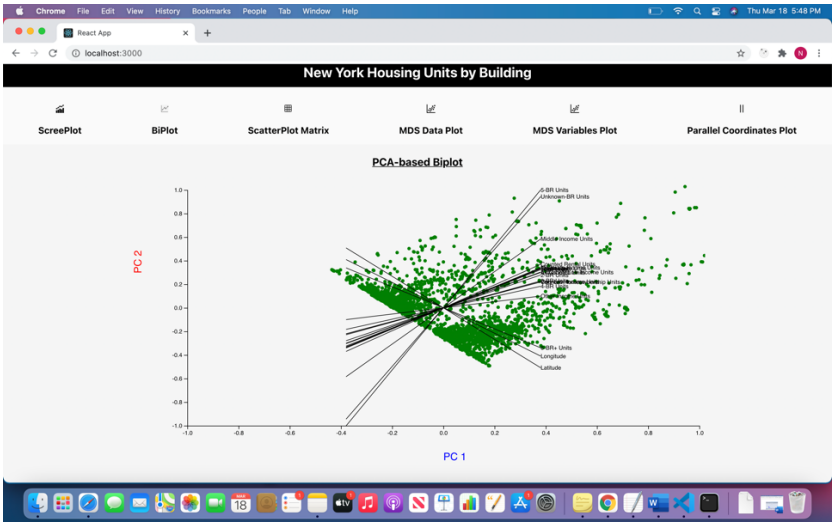


When a user does so and again plots the scatterplot matrix, we see that the number of principal components selected and the loadings change. This changes the features selected.





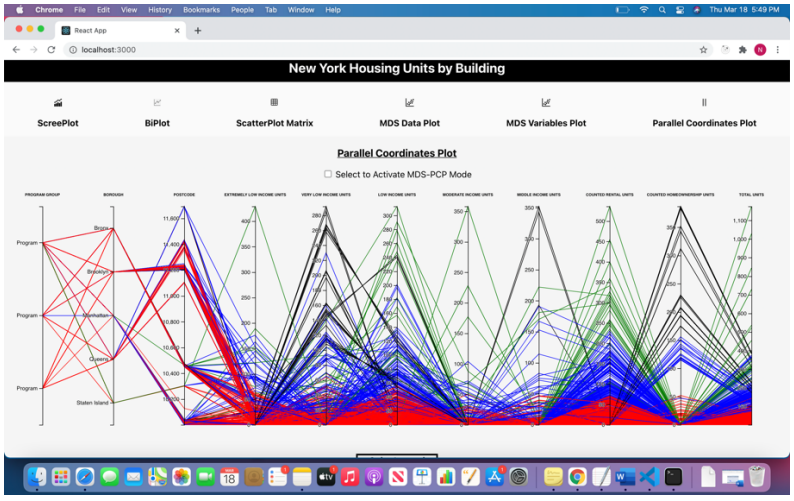
If the user clicks on the Biplot button of the dashboard, a Biplot is plotted. The plot is between principal component 1 vs principal component 2. The lines on the graph show the various attributes/ features.



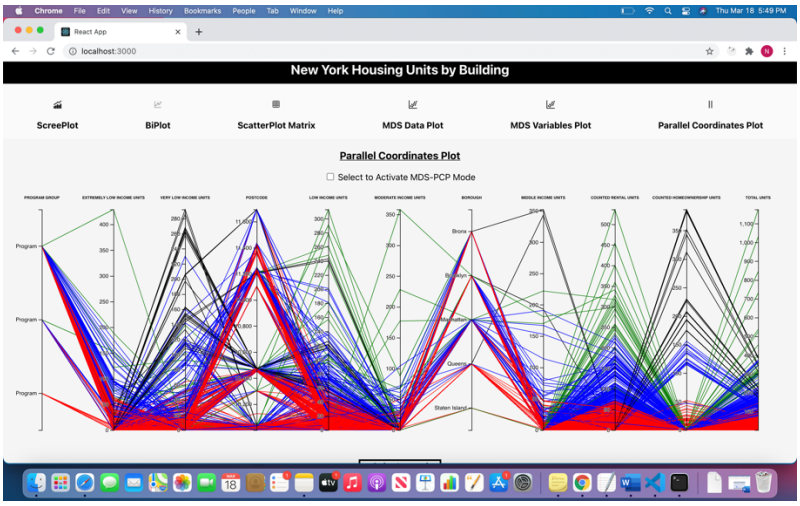
If a user clicks on MDS Data plot, a scatterplot is plotted which shows the sampled data points in a 2-D space. The same size for this plot was taken small as the computation time for MDS Data plot calculation is high. K-means clustering is used for clustering the points and coloring them.

When a user selects to plot a MDS Variable's plot, a scatterplot is plotted showing the exact same no of points as the features/attributes in the dataset. Each point has the feature name associated with it. This plot also has the feature that if the user selects points on the plot, they turn blue and the order of the selection is then used to plot the PCP for those plots selected (extra credit activity).

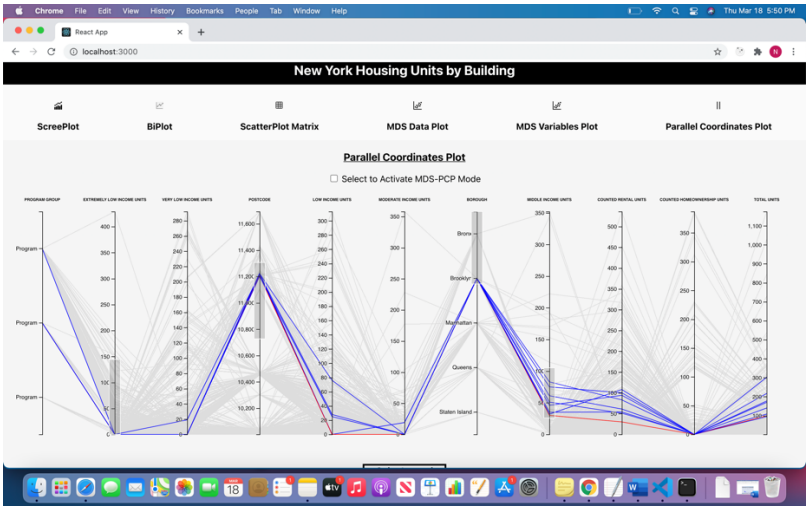
Next, if the user selects to plot a parallel coordinates plot, a PCP is plotted showing a mix of both numerical and categorical attributes as different movable axis's. There is also a checkbox button to activate/deactivate the MDS-PCP mode (extra credit activity functionality).



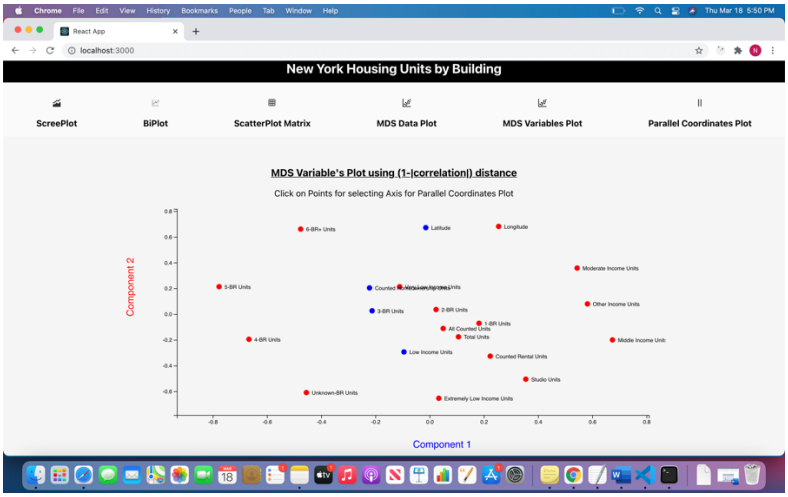
After reordering the axis's in the plot, the plot looks like this. Any order of the axis's (upto n!) can be made which provides a deep insight into the raw data and its relationships between features.



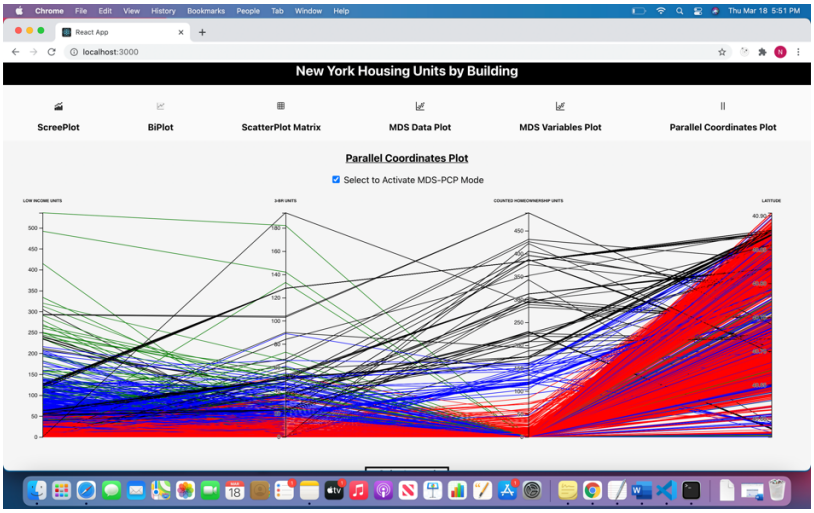
The plot also has the brushing/filtering functionality which can be used to select and view only the necessary data while filtering out other relations.



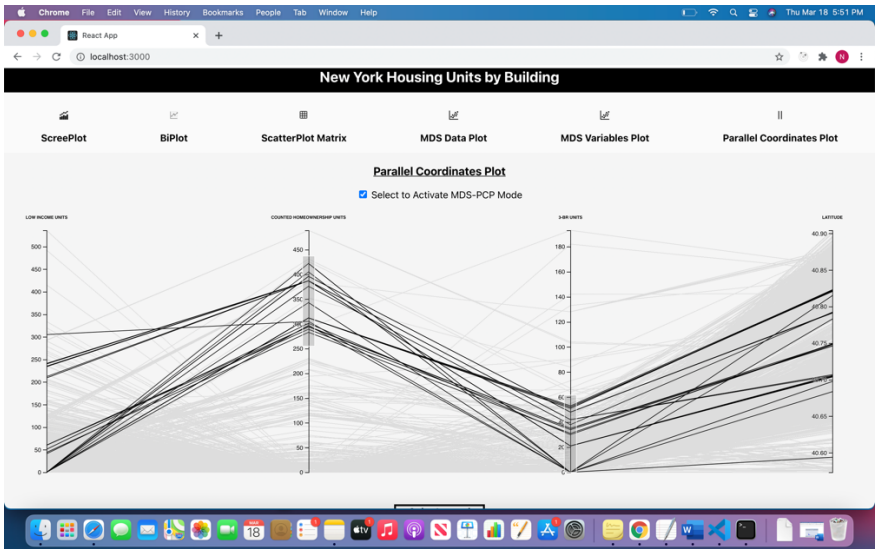
EXTRA CREDIT ACTIVITY:
If a user plots a MDS Variables plot and selects the points on the plots according to the attributes, the order of the user selection is maintained, and the points turn blue.



Now when the user plots a PCP and activates the MDS-PCP mode, a plot for the attributes selected in the MDS Variables plot are used for plotting a parallel coordinates plot.



K means clustering has been used to give color to the various polylines and a color legend is provided. Similar to the previous PCP's, we can move the axis's and apply filters for this plot also.



Upon deselecting the checkbox for the MDS-PCP mode, a normal PCP is plotted again.

Interesting Observations:

- From Plots:
  - The Biplot showed some interesting correlations between different features. 3-BR Units are closely related with Counted Homeownership Units.
  - ScatterPlot Matrix is symmetric across the diagonal as expected
- From Data via PCP:
  - Most of the buildings were made under Multi-Family Finance Program.
  - These buildings are very limited in Statin Island.
  - Moderate/Middle Income Units are very limited.
- During Implementation:
  - MDS Data Plot took a lot of computation time for 3500+ records, hence I used random sampling for reducing the number of data points
  - Sometimes the correlation function return NaN (or null) if the sample size for calculating the correlation between features is too small. Taking a relatively big sample size helped here (This is a limitation of the python corr() function)

Video Link:
https://youtu.be/eT5tuBylHgw

https://drive.google.com/file/d/1lF9dhU4G4SnanN2M1AQ68igjSuiUWTBf/view?usp=sharing