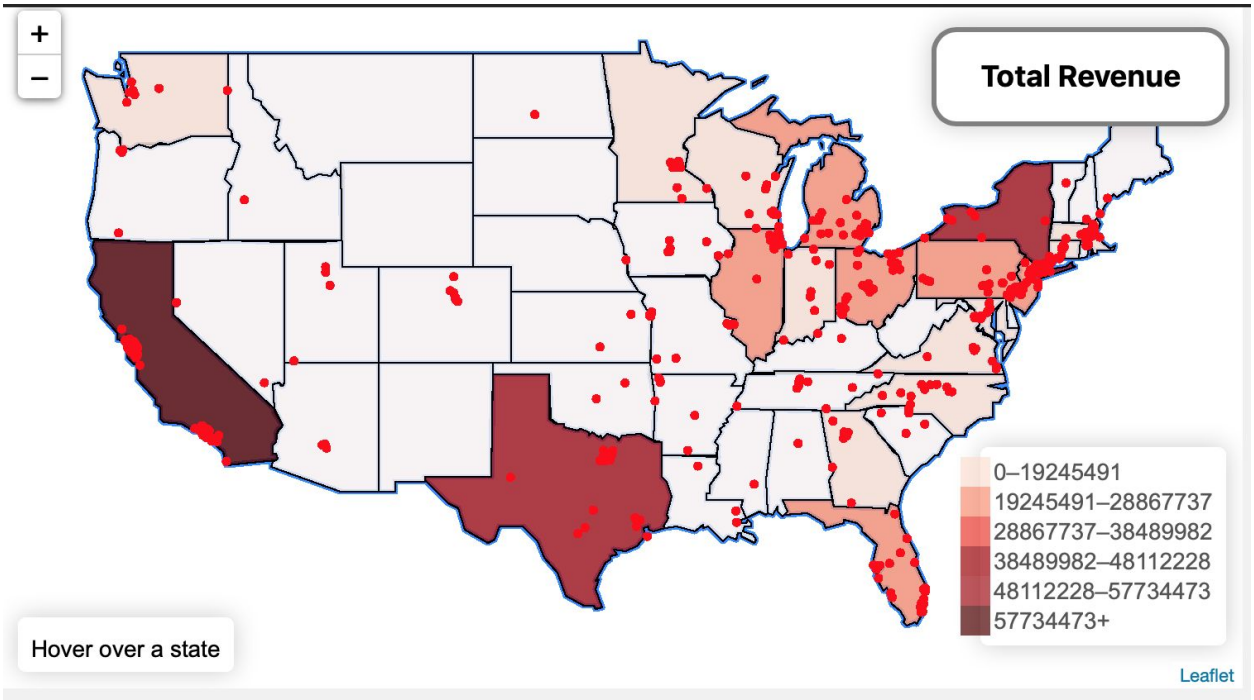
CSE 564 Final Project Report
By: Sabrina Margetic

For my final project, I focus on conveying a relationship between successful businesses and successful states. Previously, I had data on Fortune 1000 companies, this data included attributes of Change of Rank, Revenue, Revenue Change, Profit, Profit Change, Assets, Market Value, Employees, and Years on the Fortune 500 list. My additional data, conveying a successful state included, Total Revenue for the state, Federal Revenue provided to the state, State Revenue (revenue created by the state itself), Total Expenditure, Instructional Expenditure, GDP, Census, Median Household Income, Unemployment Rate, Average Iq, High School Graduate Percentage (percentage of population that graduated high school), Bachelors Graduate Percentage, and Masters Graduate Percentage. Since the original data was based on Fortune 1000 Companies in 2010, the attributes relating to the states were also found for 2010. The data was found on data.gov and kaggle.com.

To begin, the home page has a US map (located top left), a scatter plot (top right), parallel coordinates (bottom left), and another scatter plot (bottom right). The map created looks as such:

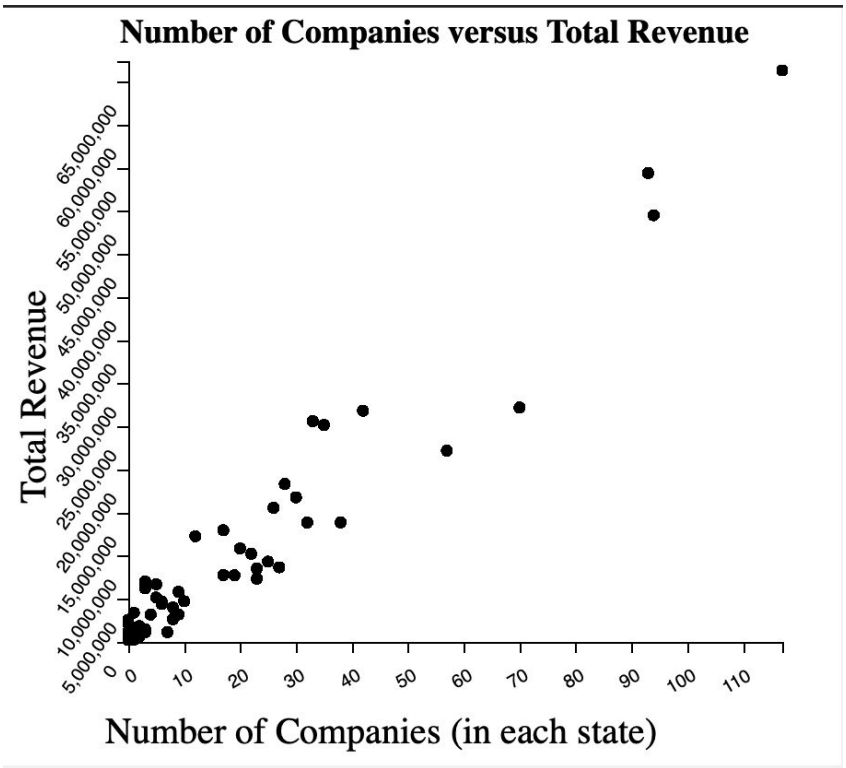


The map was created following a tutorial provided at: https://leafletjs.com/examples/choropleth/ from Leaflet. As you can see, the map has a water mark from Leaflet located at the bottom right.

Additionally, the top right of the map indicates which state variable we are looking at. The map is color coded according to the range located on the bottom right. Red dots are placed

in order to indicate the location of the fortune 1000 companies. When a state is hovered over, on the bottom left changes to indicate the number of companies and the value of that state attribute, for that state.

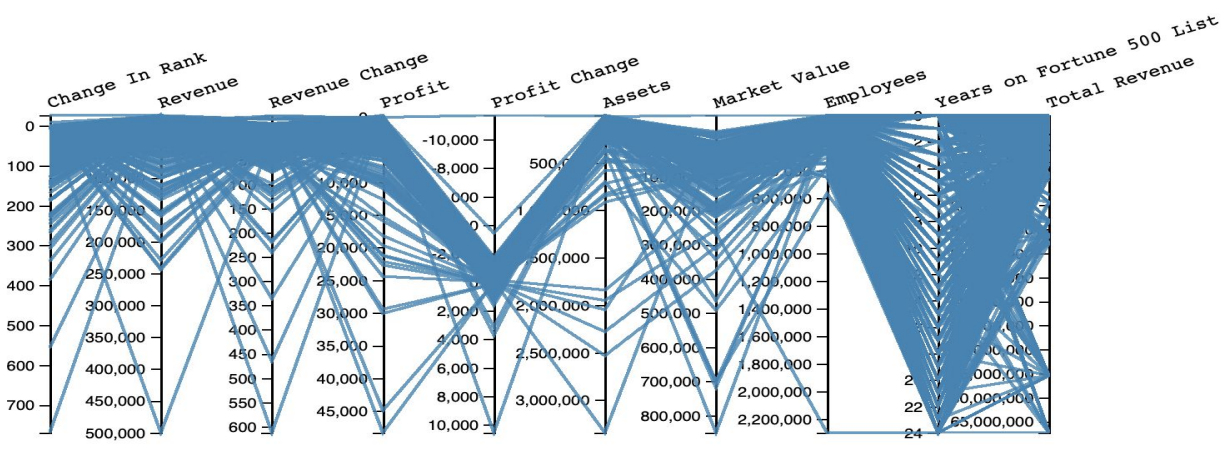
The second graph, the scatter plot, is used to depict the number of companies in a state versus values for a state attribute. An example of this is shown below.



This graph in particular shows a linear correlation between the total revenue of a state and the number of companies in a state (which are assumed to be considered successful because they are the top companies). This means that if you live in a city with more Fortune 1000 companies, the city is most likely to make more money.

Going back to the map, the upper right corner that says “Total Revenue”, from the previous example, is actually a drop down menu. When the menu is hovered over, it allows you to change the state variable you are looking at. This will consequently result in the scatter plots changing their y-axises as well.

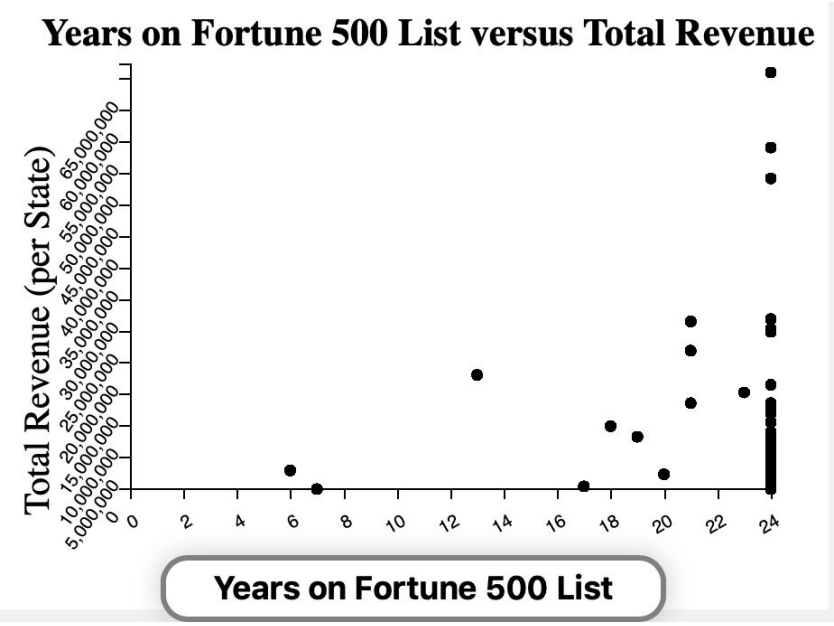
The parallel coordinate graph looks like such:



Code for this graph was based off of code found at: http://bl.ocks.org/syntagmatic/4020926 .

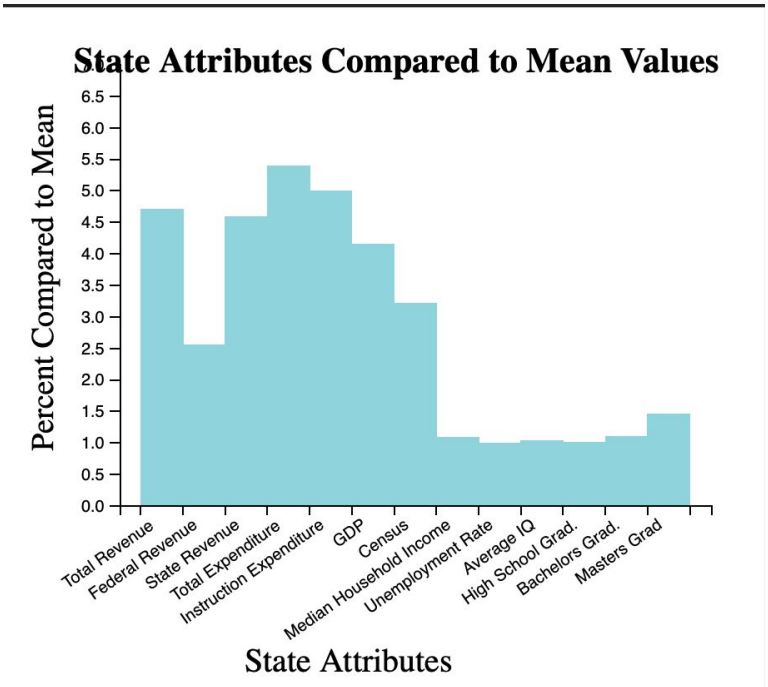
This coordinate graph plots the business attributes, and additionally the attribute specified by the map. When you hover over a segment, the rest turn gray, and you are able to clearly see the path it produces.

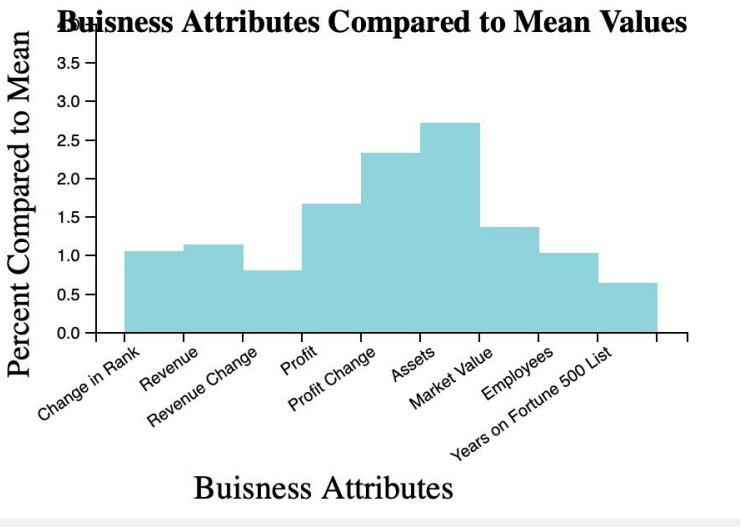
The bottom right scatter plot looks like such:



Similar to the previous map example, on the bottom is a drop down menu to which you can change the business attribute associated with it. The y axis is the state attribute from prior.

When a state is clicked on two bar graphs come to replace the scatter plots. They look as such:

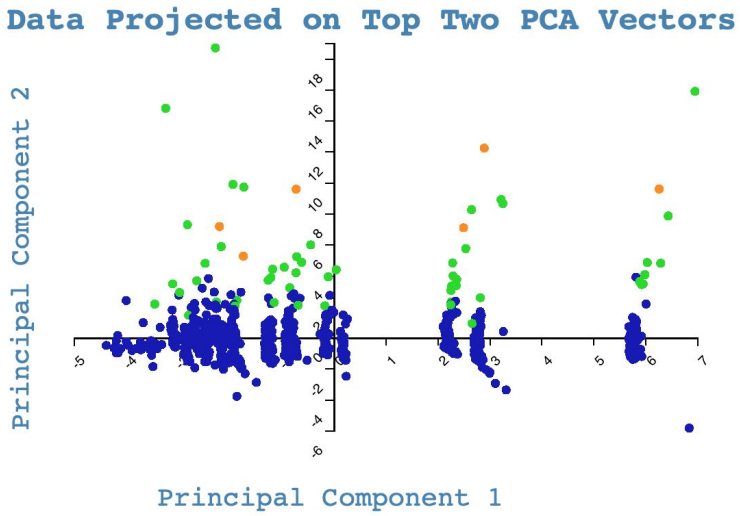




They are used to compare state and business attributes within the state to the mean values amongst all states.

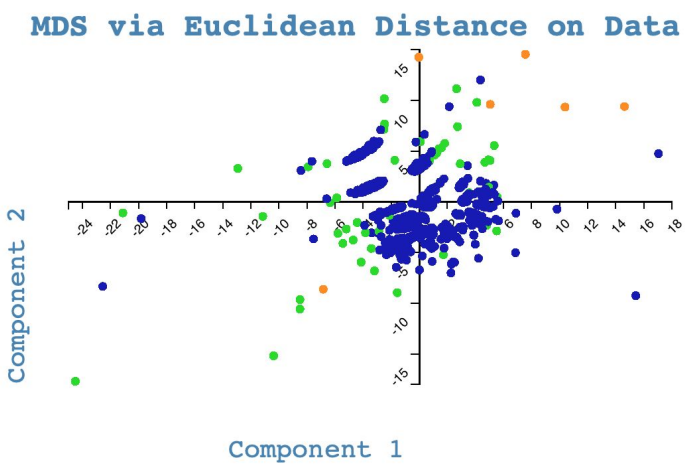
There are three additional tabs to the page, “Parallel Coordinates,” “PCA/MDS,” and “Scatter Plot Matrix.” The parallel coordinate graph looks similar to the one described above. However, now it shows all the state dimensions as well.

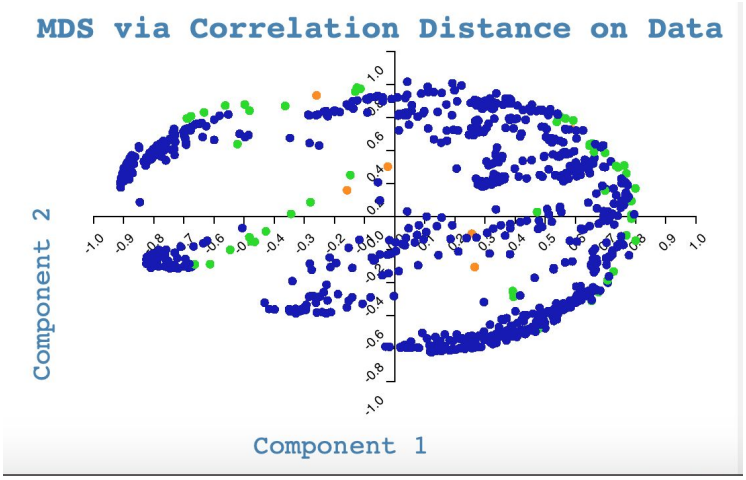
When the PCA/MDS tab is clicked, four graphs appear. The top left is a scree graph plotting the principal component numbers versus the eigenvalues. A red line is used to indicate when 75% of the variance has occurred. From the graph, you can tell that 75% of the variance would have occurred with 5 principal components. This comparably larger value (to the typical two or three value) is due to the dimension increase. The graph to the right is the projection of the data on to the top 2 principal component numbers. Given that the ideal amount of dimensions (after reduction) would be 5, it is understandable that this graph is so scattered (below).



The various colors on the graph are used to represent their cluster when k-means clustering is applied.

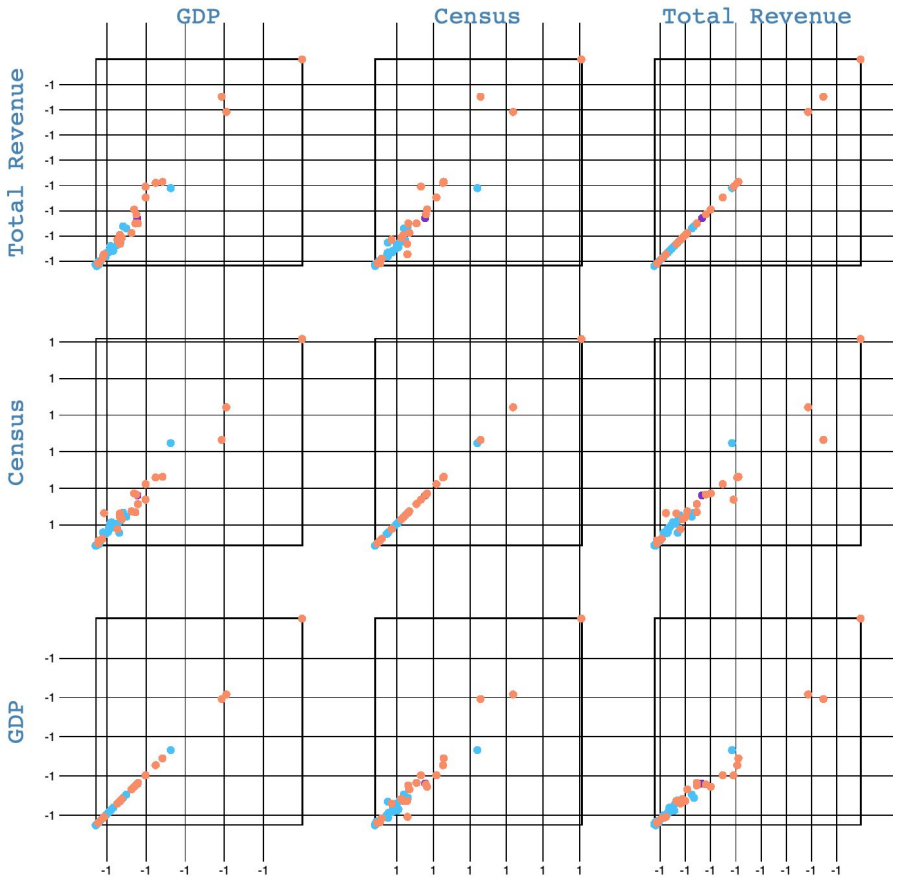
The bottom two graphs are MDS via Euclidean distance, and MDS via Correlation distance (shown below correspondingly).





From a previous lab, the MDS via Euclidean Distance data appeared more clustered around the center. Similarly, the MDS via Correlation Distance data appeared more precisely elliptical. These disparities are most likely caused by the increase in dimensions.

The last tab, the Scatter Plot Matrix, looks as such:



The scatter plot matrix is used to project the most highly correlated attributes. In this case, those attributes are GDP, Census, and Total Revenue. This allows us to get insight on state attributes, and understand that if we see a correlation between a business attribute and one of these state attributes, it also means a correlation to the other state attributes.