Final Project Report

By Kunal Kolhe

This is my report for the final project for my course: Visualization at Stony Brook University.

The final project was a culmination of all the assignments given to us during the semester.

Some of the key things we learned during the semester were:

D3.js

It is a library used for creating visualizations to be displayed in browsers. The language of D3.js as you might have guessed is JavaScript. It is extremely widely used and has a huge community who contributes to the development of examples and solving of errors on forums and stack overflow. It is a good language to have in your repertoire since it does not have a steep learning curve and is hugely powerful. If you want to show any interactive visualizations online, you definitely need to learn d3.js

New York Times is one of the most well-known users of d3.js. The creator of D3.js is a part of the NYT team.

Flask

Flask is a great micro web framework written in Python. It's language (Python) makes it easy to understand without much overhead of syntax.

It is also lightweight and again, does not have a great learning curve. It is easy to debug and since it runs on python, we can leverage the massive ease and power of python libraries for data processing.

Bootstrap

I learned Bootstrap during the completions for the assignments for this course. Again, it is very easy to understand and leverage. Simple bootstrap formatting gives us huge power to create beautiful and responsive websites. Bootstrap has many templates available to choose from and all of them very easy to modify to your requirements.

Dataset

I chose to work on a dataset from the UCI Machine Learning Repository.

It is the Statlog (Australian Credit Approval) Data Set.

This dataset has a good mix of Categorical, and numerical attributes. It has 690 data entries, 14 independent variables and 1 dependent variable.
I made a workable mapping based on this source here between the variables and their labels, since the data has been anonymized before posting it on the UCI Machine Learning Repository.

The Dataset contains missing values, so I cleaned them before proceeding to any visualizations.

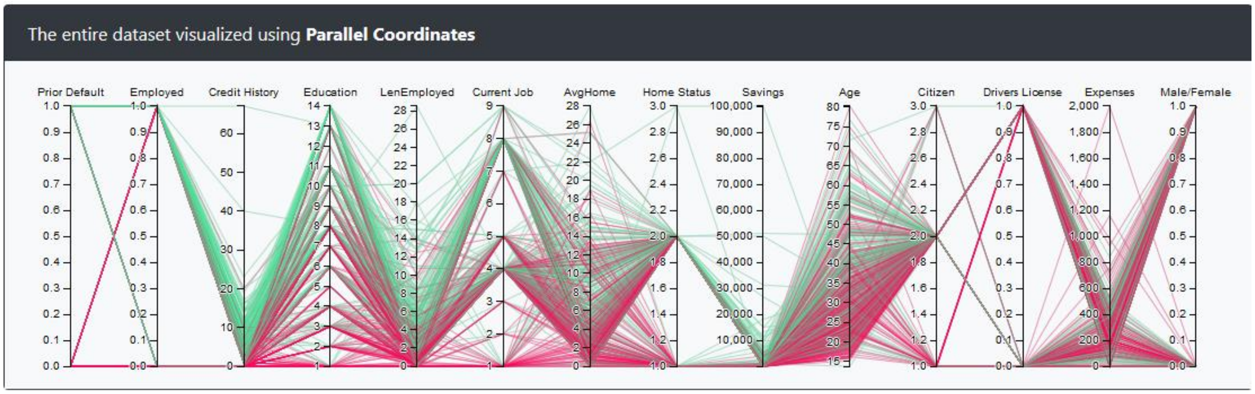
Visualizations

Before proceeding to build the visualizations in d3.js, I played around with the data. I made multiple plots using matplotlib.pyplot in python just to understand the data and how the final visualizations will look like and understanding the relationships between the data prior to starting the larger process of d3 visualizations.

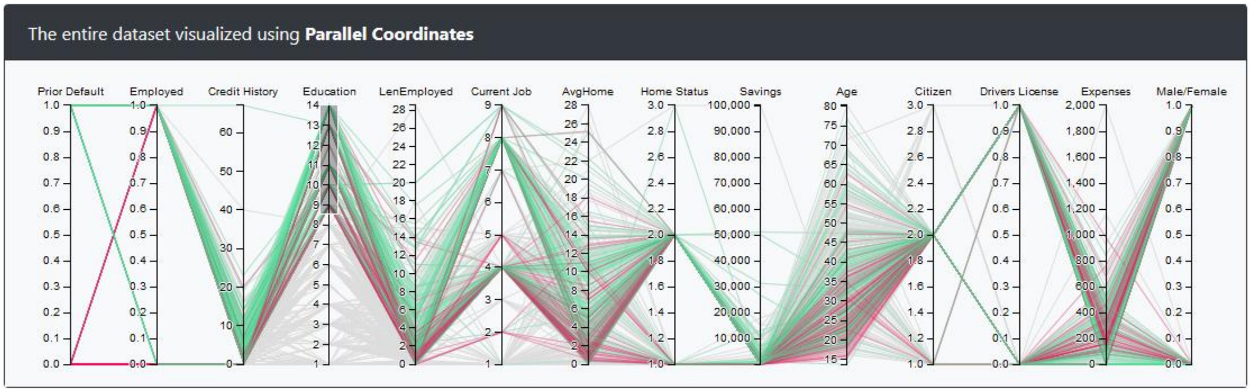
One of the important things I did was find the correlations of each of the independent variables against the target variable. This helped me understand which variables most contributed to the classification of the target variable.

One of the first plots I made in D3 was a parallel coordinate chart. This is extremely useful in terms of understanding how the relationships in the data work. All the datapoints are mapped out according to their values and general trends in the data begin to emerge.

The user can select a specific range on each axis and only the lines corresponding to that range will be emphasized.

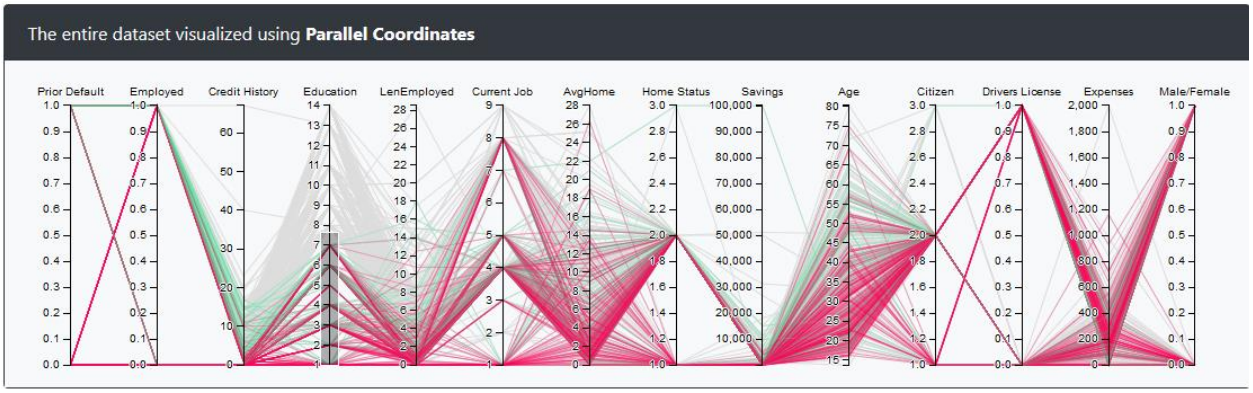

The entire dataset visualized using **Parallel Coordinates**

This image shows the entire dataset.


The entire dataset visualized using **Parallel Coordinates**

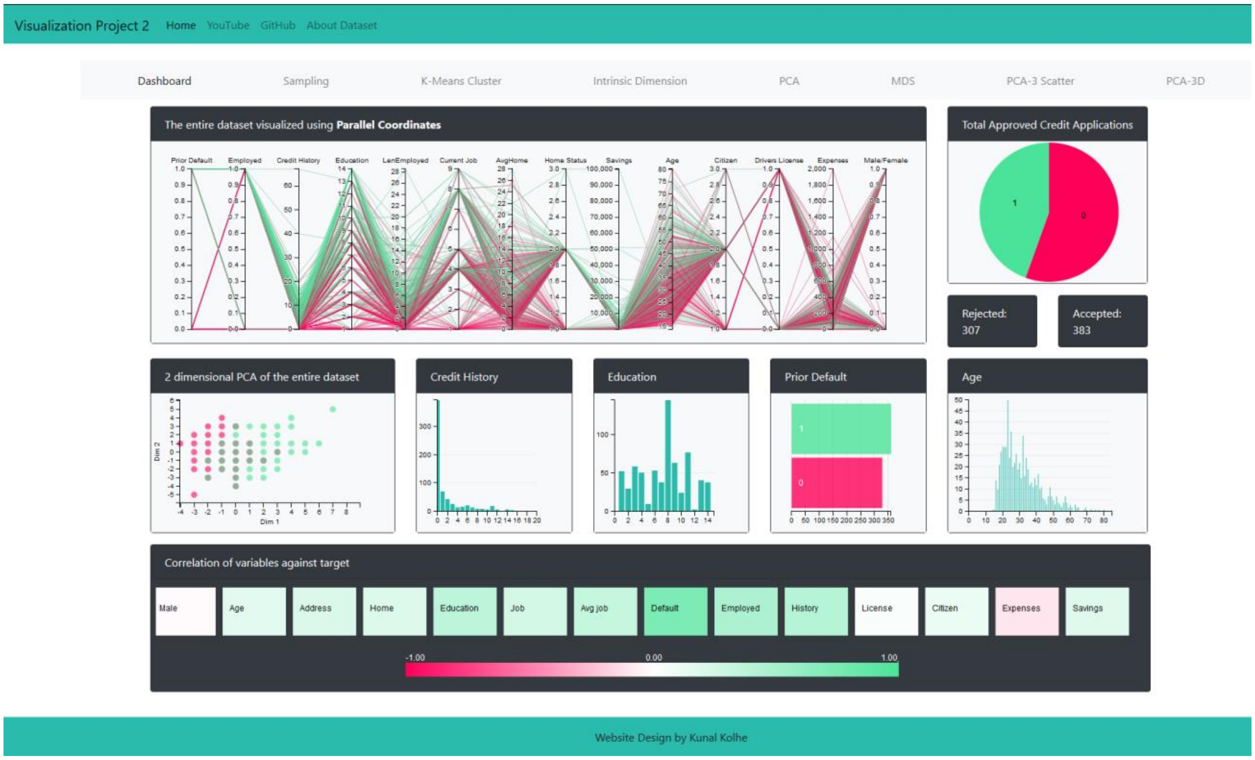
You can see that I have selected on the education axis and now you can see all the green lines which correspond to approved entries. You can see that most of the applications with higher education are approved. On the other hand, if the selection is dragged lower most of the applications are rejected.

This gives an important insight into the requirement for credit approval with respect to this field.



Again all the variables can be brushed together and individually.

Dashboard:



Rejected: 307

Accepted: 383

As you can see, the entire dashboard fits on one screen and gives a comprehensive view of the dataset.

All of the titles of the plots are explanatory and descriptive.

The color scheme is such that it will be still be differentiable to color-blind people.
I found a great tool to test this on this website: https://www.color-blindness.com/coblis-color-blindness-simulator/
I just uploaded this image and simulated how it will look for different disabilities. It is really interesting to know.

The entire Dashboard is responsive and intuitive.

The plots are linked together using crossfilter. It is a great library which allows us to easily filter data based on selections.

You can see further options on the menu bar on top in the page.

They are a feature of the previous assignment I had implemented and add value towards understanding of this dataset.

Here is a brief description of the other tabs:

Sampling: I just display how many rows remain after sampling 25% of the data using 2 techniques: Random Sampling and Stratified Sampling.

K-Means Cluster: The plots show the elbow point of K-Means clustering optimization.

Intrinsic Dimension: The plots show how many attributes capture 75% of the variance of the data.

PCA: The plots show the scatterplot matrix of the first 2 Principal Component Analysis vectors.

MDS: The plots are calculated using Multi-Dimensional Scaling of the entire dataset down to 2 dimensions.

PCA-3 Scatter: The plots are of the first 3 PCA dimensions in a scatter-plot matrix. The values can be brushed and the change is evident on the other plots.

PCA-3D: The plots are of a 3 Dimensional interactive plot of the first 3 PCA vectors.

The link to the code is: https://github.com/knl-kolhe/Visualization-Project

The link to the YouTube video of the project demo is: https://youtu.be/_A7MYfPbDHI