Documentation

Overview

The dataset chosen for this assignment is Electronic Health Record (EHR) dataset. The dataset consists of 5674 records, each defined by 34 attributes. The attributes can be classified into five broad categories, Demographics Information, Injury Information, Encounter Information, Other Flags, and Diagnosis Flags. Demographics attributes have basic attributes about the patients (or records) like gender, age etc, injury information attributes talk about things like date of injury, category of injury etc, encounter information has details about when the TBI occurred and diagnosis flags have information about the type of encounter (TBI or other).

The last class of attributes is highly sparse, most entries are 0. In this case visualizing these attributes will only add to the noise, so we decide to skip analysis for these attributes. Attributes like Patient ID and Date of Injury are skipped because they do not provide any domain related information.

This type of dataset is called a recorded dataset because it has records and a bunch of attributes about each record. These datasets are very common in the domain of data mining as some attributes (dependent) can be predicted using the remaining attributes (independent). In this assignment, we will see some interesting trends by visualizing up to four attribute information at a time. Given the special form of dataset, we will stick to the standard visualization techniques like bar graphs, histograms and scatter plots (2D and 3D).

Tools

As stated in the part one, the analysis in this assignment makes use of three visualization tools, D3, WebGL and Matlab (matplotlib library). Though Matlab is the most efficient tool for generating all the three kind of plots, the assignment also includes D3 and WebGL because it is good way to learn/explore theses tools.

D3 was used to generate a **couple-barplot**. We define a couple-barplot as two bar plots for an attribute (like Age) which differ from each other in the sense that the attribute data is distributed among both based on some other binary attribute (like Gender). We can extend this by taking the second attribute as a N-class attribute, but that would result in N different bar plots and would make the visual analysis difficult. Since we are using information from

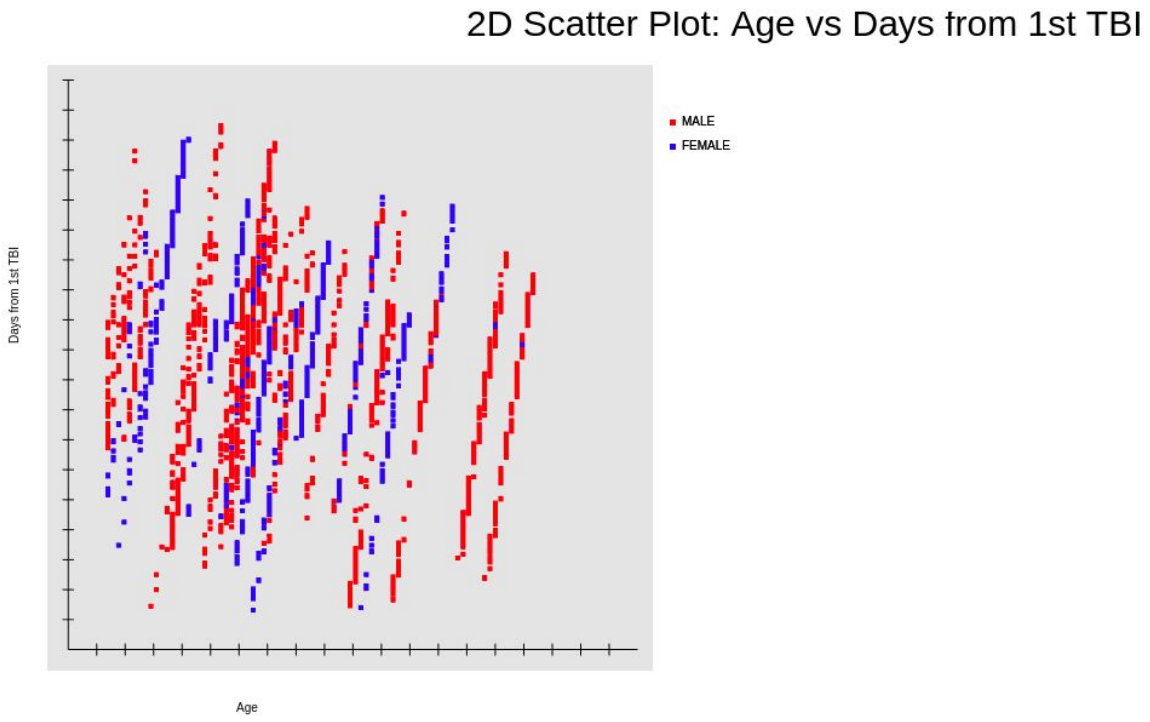
two attributes, it is a two-dimensional analysis. We choose Gender and Injury type as the second binary attribute.

We use WebGL to plot a 2D scatter plot for visualizing attributes "Age" and "Days from 1st TBI". WebGL was used just to explore the elementary graphics utility of this tool. We compare the figure generated using WebGL and the corresponding scatter plot given by python code.

Matplotlib is a python library for accessing Matlab visualization tools. Since python is faster and more scalable than Matlab, we use python for plotting the 2D scatter plots, 3D scatter plots and histograms.

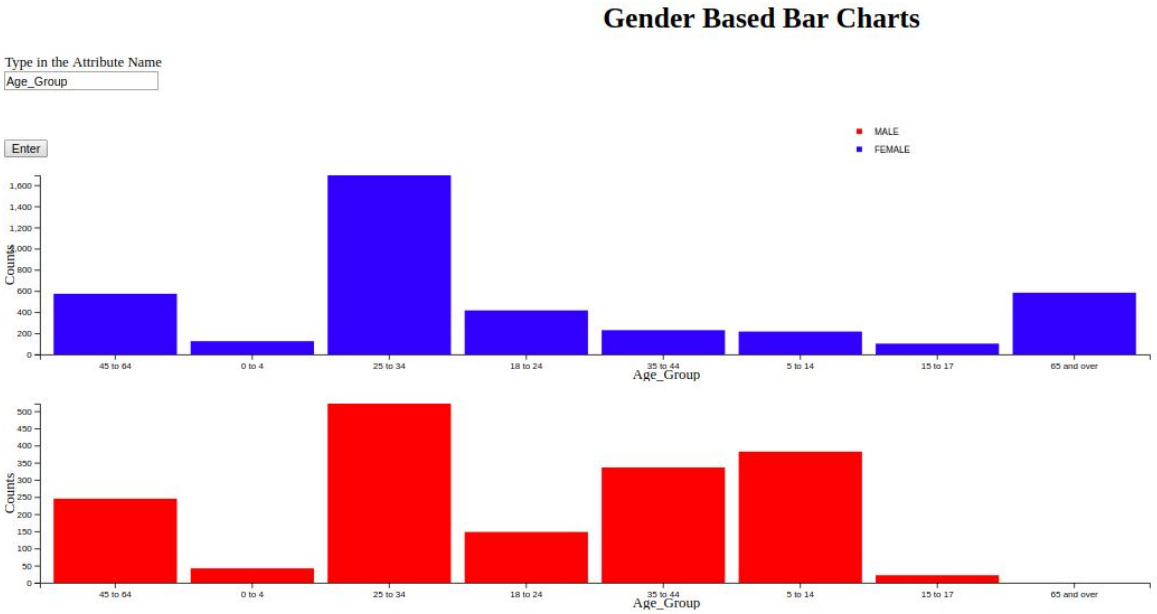
Figures

1. WebGL

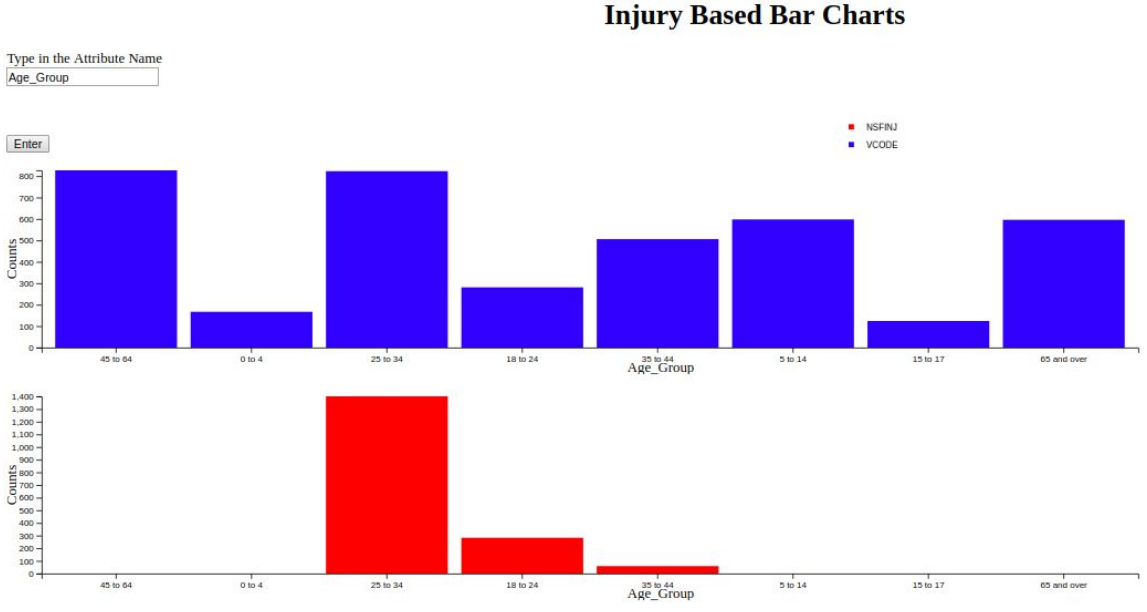


WebGL output for the 2D scatter plot between attributes **Age** and **Days from 1st TBI**

2. D3

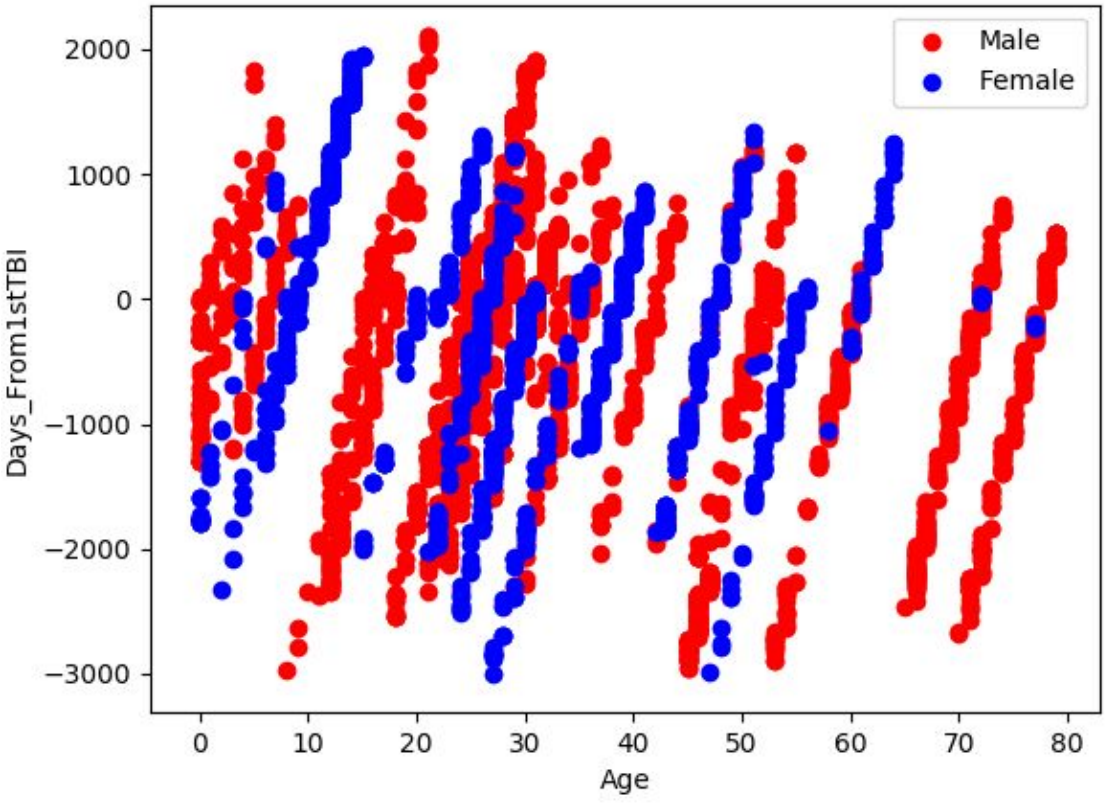


Couple-barplot for the **Age group** where the second attribute is **Gender.** Age group 24-35 has maximum patients in both the genders.

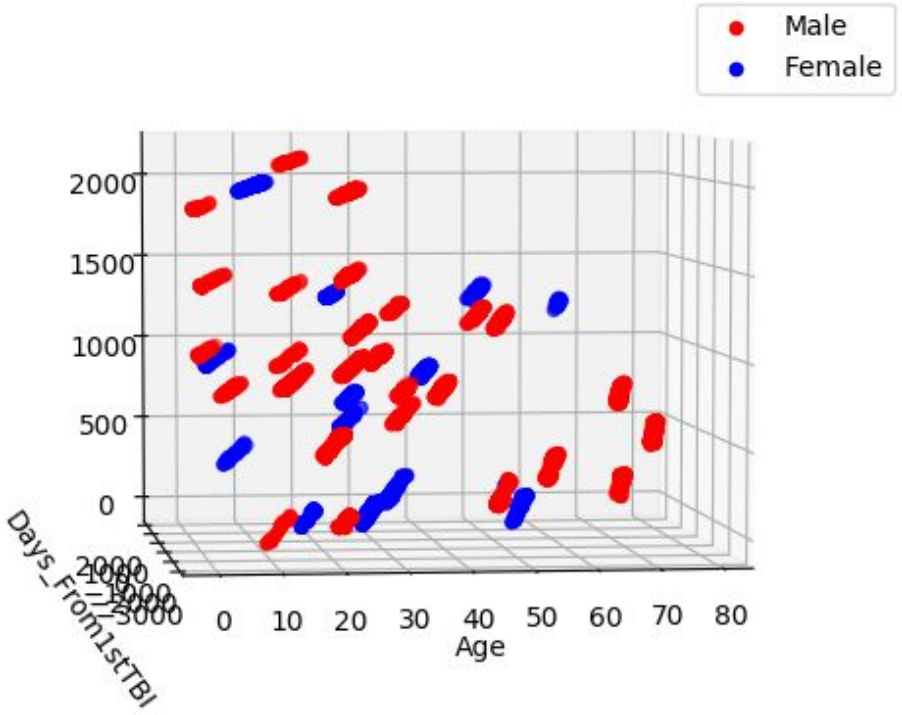


Couple-barplot for the **Age group** where the second attribute is **Injury type.** NSFNJ type patients are limited to 25-44 age group.

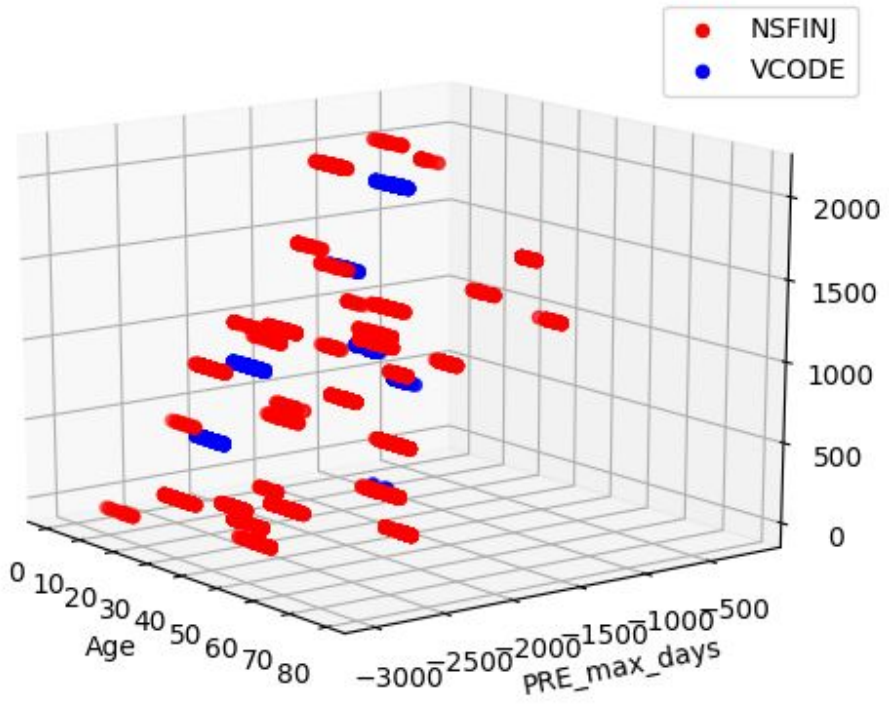
3. Matplotlib



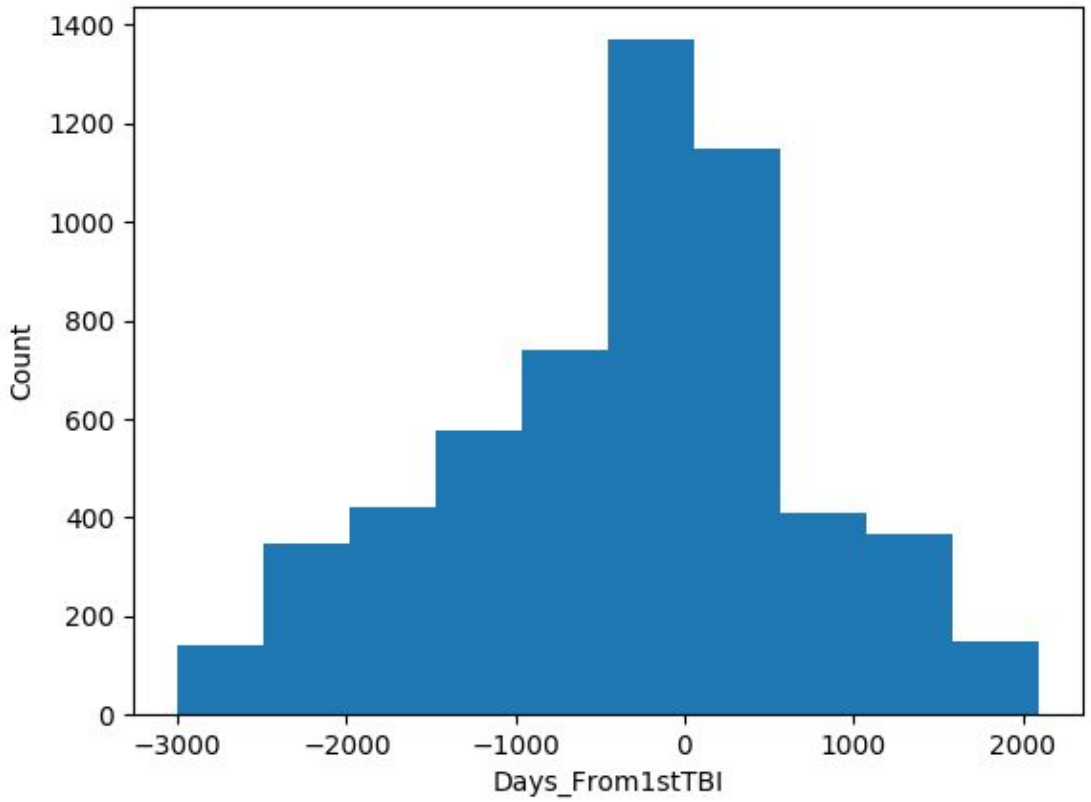
Age vs Days from 1st TBI scatter plot. This is simiar to the WebGL scatter plot.



3D scatter plot for attributes **Age**, **Days from 1st TBI** and **Post max days**. The data is further divided into two gender classes. It can be seen that these three attributes divides data into clear Male and Female clusters. Gender can be chosen as a dependent variable and can be predicted using these three attributes.



3D scatter plot with attributes **Age**, **Pre max days** and **Post max days.** Injuty label is the fourth attribute that further divides points into red and blue. It can be seen from the plot that the data forms red and blue clusters. This implies that the three attributes can be used to predict Injury labels for patients.



Histogram for Days from 1st TBI. It can be observed that most patients have value near 0 for this attribute.

Generating Figures

1. WebGl: Open the folder WebGL. Open proj1.html in Chrome browser. Click on ‘choose file’ and choose the ehr.csv file.

2. D3: Open D3 folder. Open a new server session on python. Append the GenderBar.html file path with server port. Enter the attribute name in textbox eg. ‘Age_Group’ and click on Enter.

3. Run the python scripts in Matplotlib folder.

Conclusion

We saw four different kinds of plots (2D scatter, couple-barplot, 3D scatter and histogram). The figures show some interesting statistics about the attributes. Though these plots were produced for many different attribute combinations, the final figures in this documentation are corresponding to the best attribute combinations.

As stated before since the dataset is record type data, matplotlib is the most efficient choice for the tasks performed in this assignment. D3 is a powerful visualization tool.
The D3 implementation has HTML elements like textbox and button. This makes D3 interactive. D3 is a better choice for the other two datasets.
WebGL was the most arduous out of all three. Since there are no primitives other than triangles. Drawing graph was a jarring task. Even a trivial task like drawing axes ticks required few lines of code.