Big Data Project Two

Matt Gilmore

Eric Wendt

Braden Meyerhoefer

Nabi Akram

Rahmana Muhammad

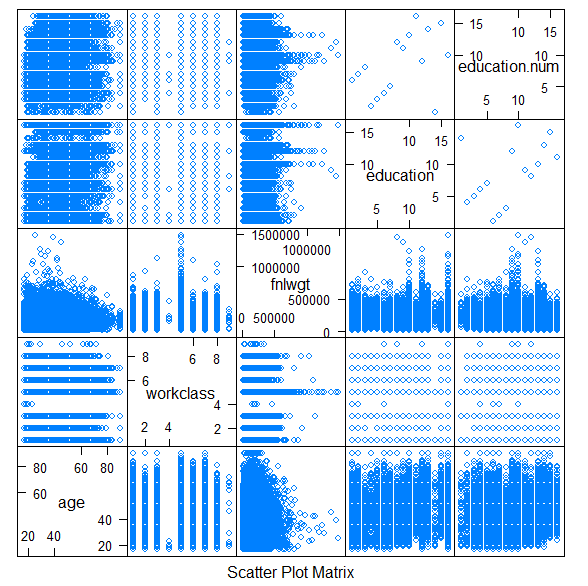
**Part 1**

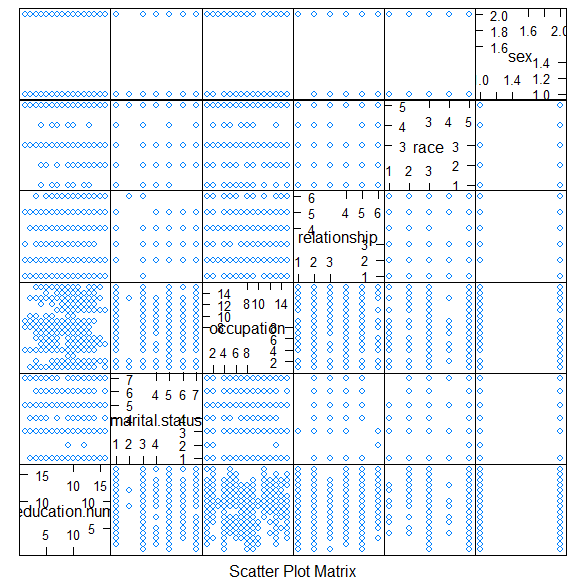
Below are the pairwise plots for all attributes of the dataset to view correlation patterns. The R functions used for this part are:

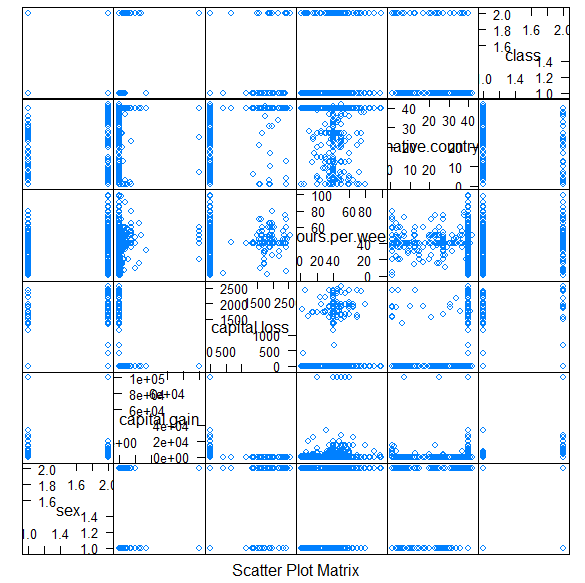
pairs(numeric\_data)

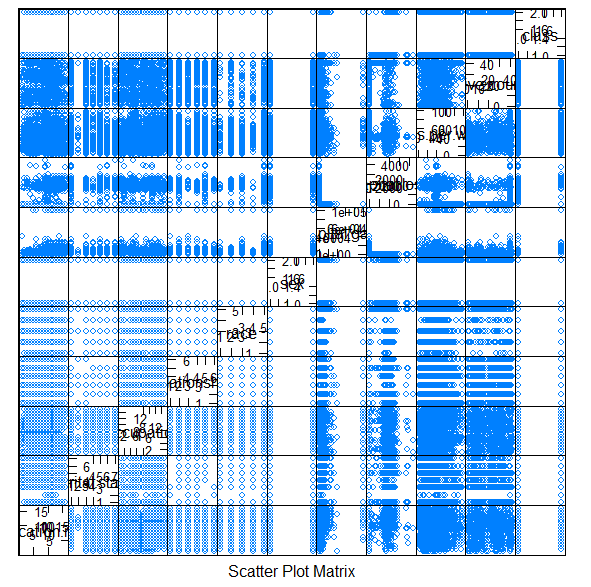
splom(~numeric\_data[1:4])

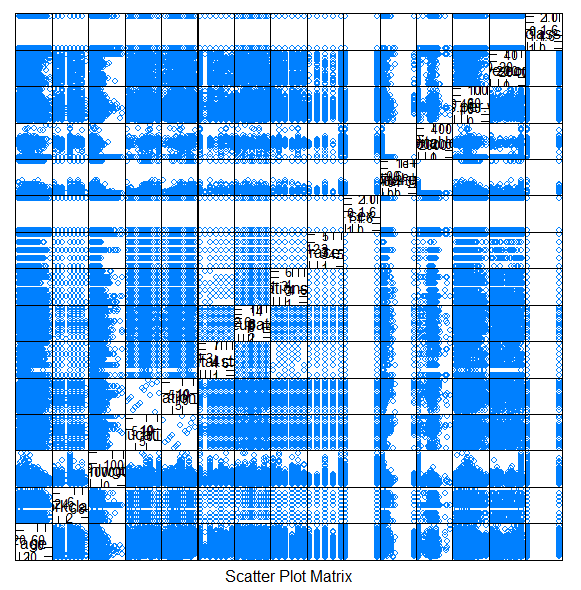
cor(numeric\_data[,i],numeric\_data[,j])











The visualization is difficult to discern information from, but we could tell that the lowest correlation variables were sex and class. We next had to use the cor() function to calculate exact correlation results among features for more precise measures. We found these three correlations as the most significant:

For column 4 and column 5 correlation is 0.35915294241097.

For column 5 and column 15 correlation is 0.335153952690942.

For column 8 and column 10 correlation is -0.582453690049836.

Thus the strongest correlation by far is between relationship and sex, where the correlation is negative.

**Part Two**

**Part Four**

For the prediction portion of this assignment, we used the following methods/objects from the sklearn library: LinearRegression, train\_test\_split, classification\_report, confusion\_matrix, and accuracy\_score.

Our exploration of the data using the various visualizations did not reveal any obvious correlations between the various fields, so we first tested training a linear model using all of the fields. Then, to see if we could make it better, we tried various combinations of fields that we intuitively thought would be good (like education and occupation for instance). Below are the results for each, we used used sklearn’s train\_test\_split to generate training and testing sets at the various breaks requested (50/50,60/40,70/30), but included in this report is just the 70/30 results as the larger training sets gave the best. Models were evaluated using sklearn’s various evaluation methods that we listed above. In the data below, first you will see the fields used to make the model, then the accuracy\_score, the confusion\_matrix, and the classification\_report. Not all tests are shown, but enough is included in the report to demonstrate our conclusion that the model performed best with all fields included, getting a weighted f1-score of 0.8. Our testing showed that the model’s performance depended not so much on which fields were used, but the number of fields used. This makes sense with our visualizations showing only a few clear correlations; the model does not have clear linear relations so it performs best when the most information is provided.

Everything

0.8267990582454704

[[7273 277]

[1415 804]]

precision recall f1-score support

0 0.84 0.96 0.90 7550

1 0.74 0.36 0.49 2219

accuracy 0.83 9769

macro avg 0.79 0.66 0.69 9769

weighted avg 0.82 0.83 0.80 9769

[‘age’,’education’]

0.7658921076875832

[[7461 89]

[2198 21]]

precision recall f1-score support

0 0.77 0.99 0.87 7550

1 0.19 0.01 0.02 2219

accuracy 0.77 9769

macro avg 0.48 0.50 0.44 9769

weighted avg 0.64 0.77 0.67 9769

['age','race','sex']

0.762206981267274

[[7405 145]

[2178 41]]

precision recall f1-score support

0 0.77 0.98 0.86 7550

1 0.22 0.02 0.03 2219

accuracy 0.76 9769

macro avg 0.50 0.50 0.45 9769

weighted avg 0.65 0.76 0.68 9769

['age', 'education','race','sex','hours-per-week']

0.7708056095813287

[[7345 205]

[2034 185]]

precision recall f1-score support

0 0.78 0.97 0.87 7550

1 0.47 0.08 0.14 2219

accuracy 0.77 9769

macro avg 0.63 0.53 0.50 9769

weighted avg 0.71 0.77 0.70 9769