Big Data Project Two

Matt Gilmore

Eric Wendt

Braden Meyerhoefer

Nabi Akram

Rahmana Muhammad

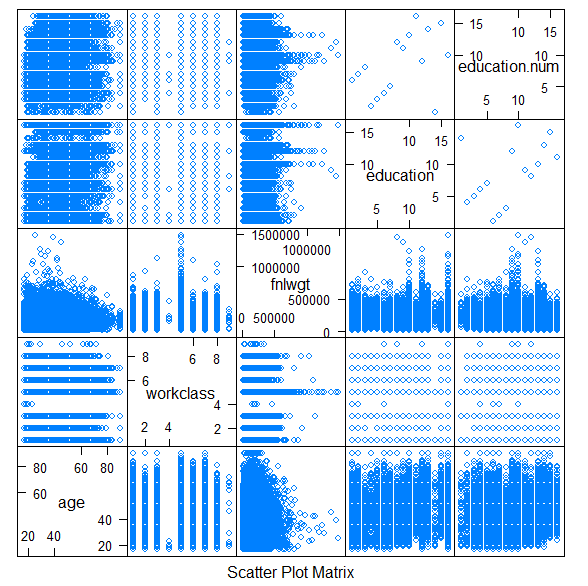
**Part 1**

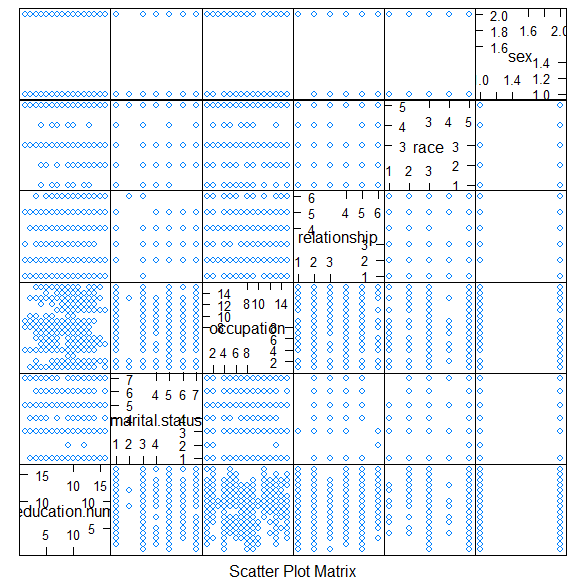
Below are the pairwise plots for all attributes of the dataset to view correlation patterns. The R functions used for this part are:

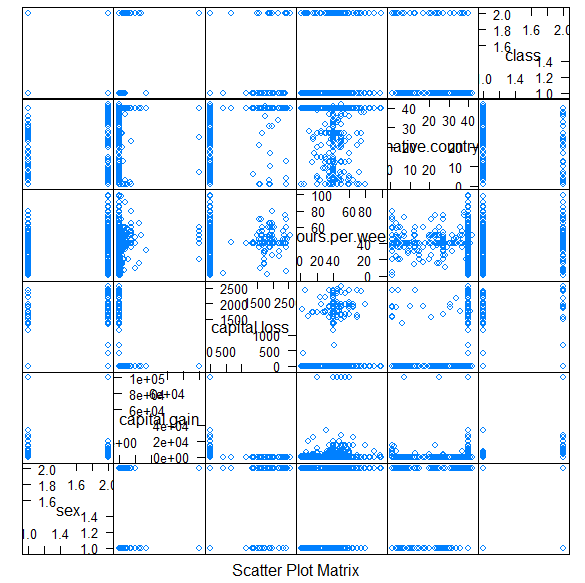
pairs(numeric\_data)

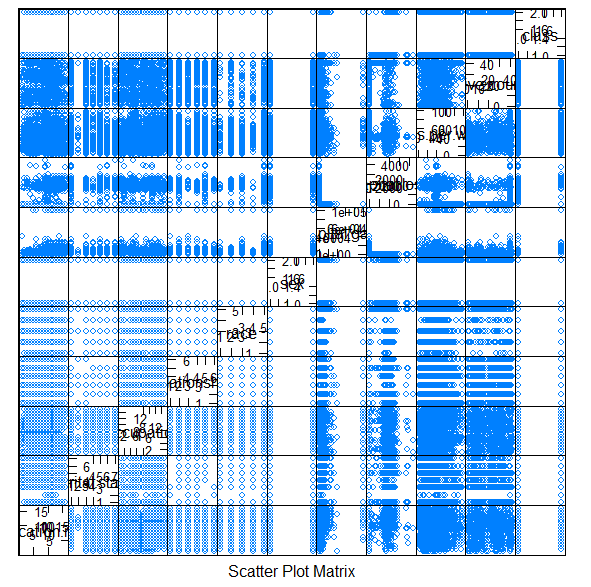
splom(~numeric\_data[1:4])

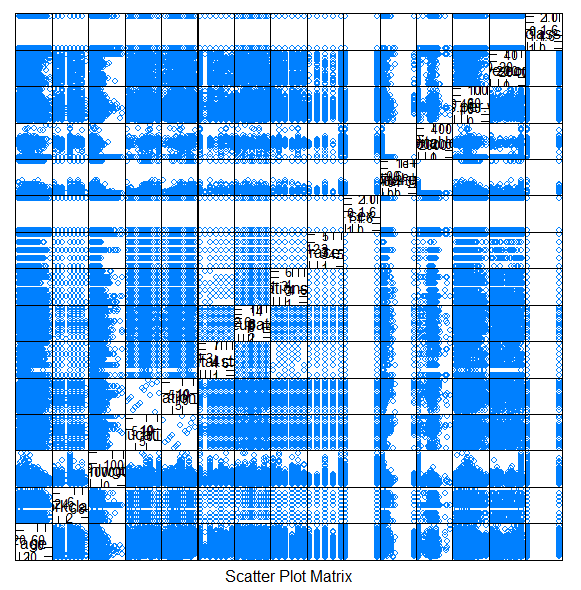
cor(numeric\_data[,i],numeric\_data[,j])











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**

In order to understand the data, we ran several statistic summarizations of the data.

First, summary(data):

age workclass

Min. :17.00 Private :22696

1st Qu.:28.00 Self-emp-not-inc: 2541

Median :37.00 Local-gov : 2093

Mean :38.58 ? : 1836

3rd Qu.:48.00 State-gov : 1298

Max. :90.00 Self-emp-inc : 1116

(Other) : 981

fnlwgt education

Min. : 12285 HS-grad :10501

1st Qu.: 117827 Some-college: 7291

Median : 178356 Bachelors : 5355

Mean : 189778 Masters : 1723

3rd Qu.: 237051 Assoc-voc : 1382

Max. :1484705 11th : 1175

(Other) : 5134

education.num marital.status

Min. : 1.00 Divorced : 4443

1st Qu.: 9.00 Married-AF-spouse : 23

Median :10.00 Married-civ-spouse :14976

Mean :10.08 Married-spouse-absent: 418

3rd Qu.:12.00 Never-married :10683

Max. :16.00 Separated : 1025

Widowed : 993

occupation relationship

Prof-specialty :4140 Husband :13193

Craft-repair :4099 Not-in-family : 8305

Exec-managerial:4066 Other-relative: 981

Adm-clerical :3770 Own-child : 5068

Sales :3650 Unmarried : 3446

Other-service :3295 Wife : 1568

(Other) :9541

race sex

Amer-Indian-Eskimo: 311 Female:10771

Asian-Pac-Islander: 1039 Male :21790

Black : 3124

Other : 271

White :27816

capital.gain capital.loss

Min. : 0 Min. : 0.0

1st Qu.: 0 1st Qu.: 0.0

Median : 0 Median : 0.0

Mean : 1078 Mean : 87.3

3rd Qu.: 0 3rd Qu.: 0.0

Max. :99999 Max. :4356.0

hours.per.week native.country

Min. : 1.00 United-States:29170

1st Qu.:40.00 Mexico : 643

Median :40.00 ? : 583

Mean :40.44 Philippines : 198

3rd Qu.:45.00 Germany : 137

Max. :99.00 Canada : 121

(Other) : 1709

class

<=50K:24720

>50K : 7841

Second, describe(data):

vars n mean sd

age 1 32561 38.58 13.64

workclass\* 2 32561 4.87 1.46

fnlwgt 3 32561 189778.37 105549.98

education\* 4 32561 11.30 3.87

education.num 5 32561 10.08 2.57

marital.status\* 6 32561 3.61 1.51

occupation\* 7 32561 7.57 4.23

relationship\* 8 32561 2.45 1.61

race\* 9 32561 4.67 0.85

sex\* 10 32561 1.67 0.47

capital.gain 11 32561 1077.65 7385.29

capital.loss 12 32561 87.30 402.96

hours.per.week 13 32561 40.44 12.35

native.country\* 14 32561 37.72 7.82

class\* 15 32561 1.24 0.43

median trimmed mad min

age 37 37.69 14.83 17

workclass\* 5 4.96 0.00 1

fnlwgt 178356 180802.36 88798.84 12285

education\* 12 11.81 2.97 1

education.num 10 10.19 1.48 1

marital.status\* 3 3.65 2.97 1

occupation\* 8 7.50 5.93 1

relationship\* 2 2.25 1.48 1

race\* 5 4.90 0.00 1

sex\* 2 1.71 0.00 1

capital.gain 0 0.00 0.00 0

capital.loss 0 0.00 0.00 0

hours.per.week 40 40.55 4.45 1

native.country\* 40 40.00 0.00 1

class\* 1 1.18 0.00 1

max range skew kurtosis

age 90 73 0.56 -0.17

workclass\* 9 8 -0.75 1.68

fnlwgt 1484705 1472420 1.45 6.22

education\* 16 15 -0.93 0.68

education.num 16 15 -0.31 0.62

marital.status\* 7 6 -0.01 -0.54

occupation\* 15 14 0.11 -1.23

relationship\* 6 5 0.79 -0.77

race\* 5 4 -2.44 4.87

sex\* 2 1 -0.72 -1.48

capital.gain 99999 99999 11.95 154.77

capital.loss 4356 4356 4.59 20.37

hours.per.week 99 98 0.23 2.92

native.country\* 42 41 -3.66 12.53

class\* 2 1 1.21 -0.53

se

age 0.08

workclass\* 0.01

fnlwgt 584.94

education\* 0.02

education.num 0.01

marital.status\* 0.01

occupation\* 0.02

relationship\* 0.01

race\* 0.00

sex\* 0.00

capital.gain 40.93

capital.loss 2.23

hours.per.week 0.07

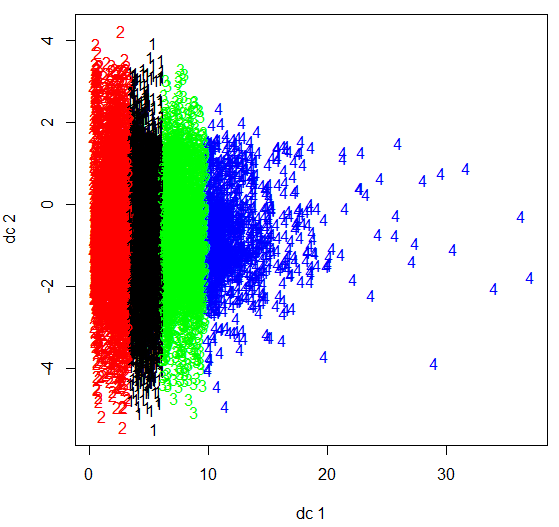
native.country\* 0.04

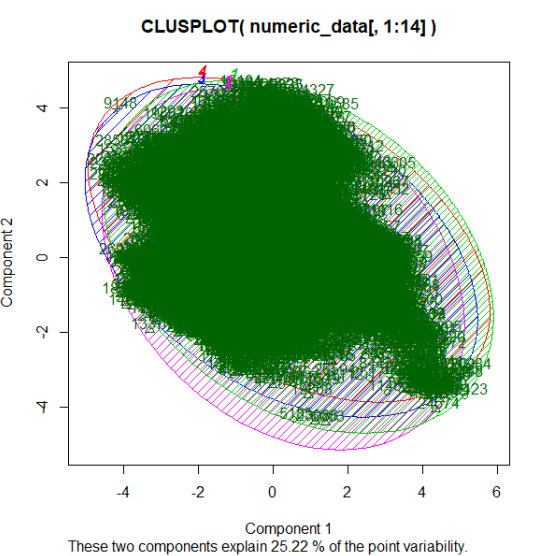
class\* 0.00

These statistics used the original data where the categorical variables were converted to numeric variables where applicable. This was done with the function data.matrix(data).

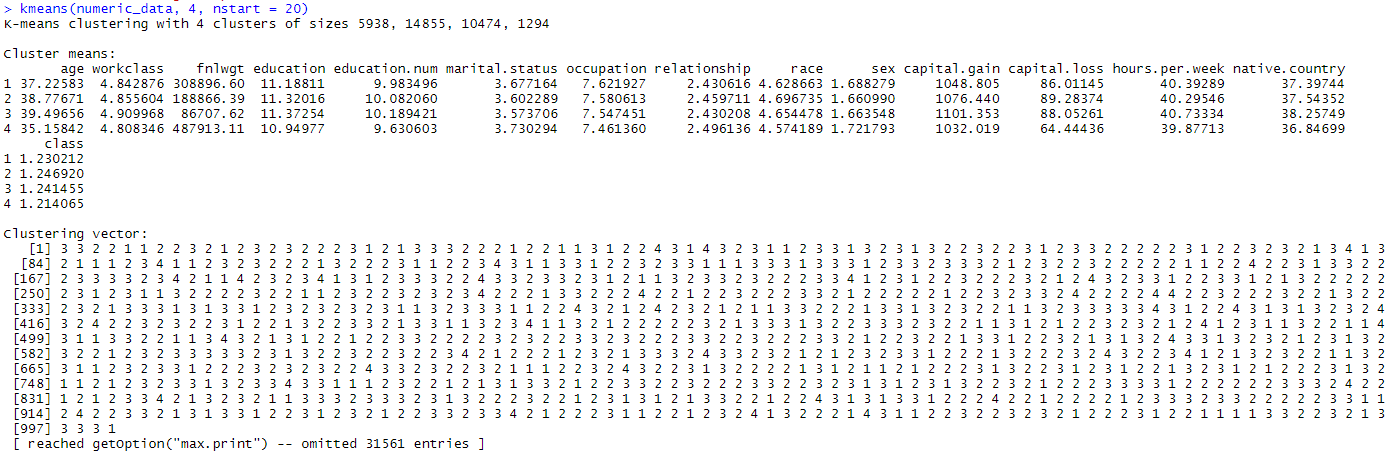
**Part Two**

**Plots**





**Tables**



**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

**What We Learned**

This project helped us learn that before complex functions can be applied to data sets, such as clustering or linear regression, you need to investigate and manipulate the data to prepare it. For instance, several of our variables were initially categorical and had to be converted to numeric. However, these mappings were not all within the same range and therefore not standardized. This would have thrown off our regression as some variables would have been weighted differently in this case. Thus, the data had to be scaled to the same range. Also, there were several missing or N/A values that had to be omitted or else it would throw off statistics for correlation. Furthermore, we needed to investigate correlation between variables to see how we could subset our data for regression. By viewing correlation between variables, we found that our data was not very correlated, and therefore we needed to use more variables to inform the model. Processes such as this helped improve our understanding of the data and data science.