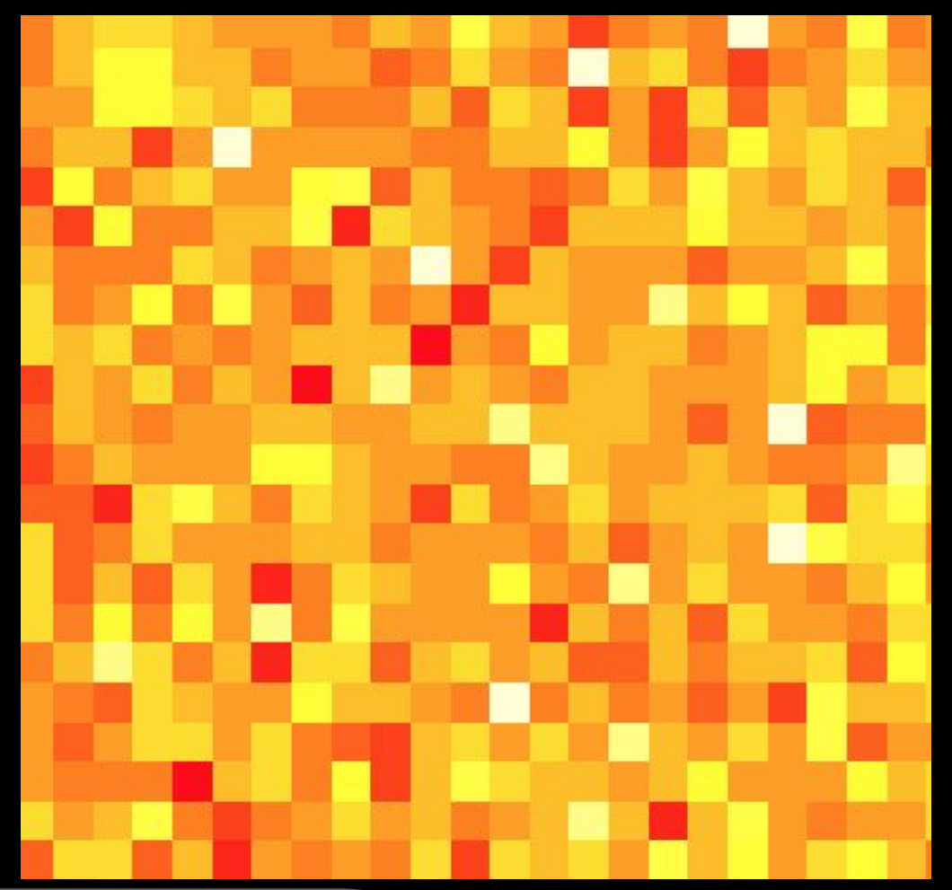
WEEESNAW

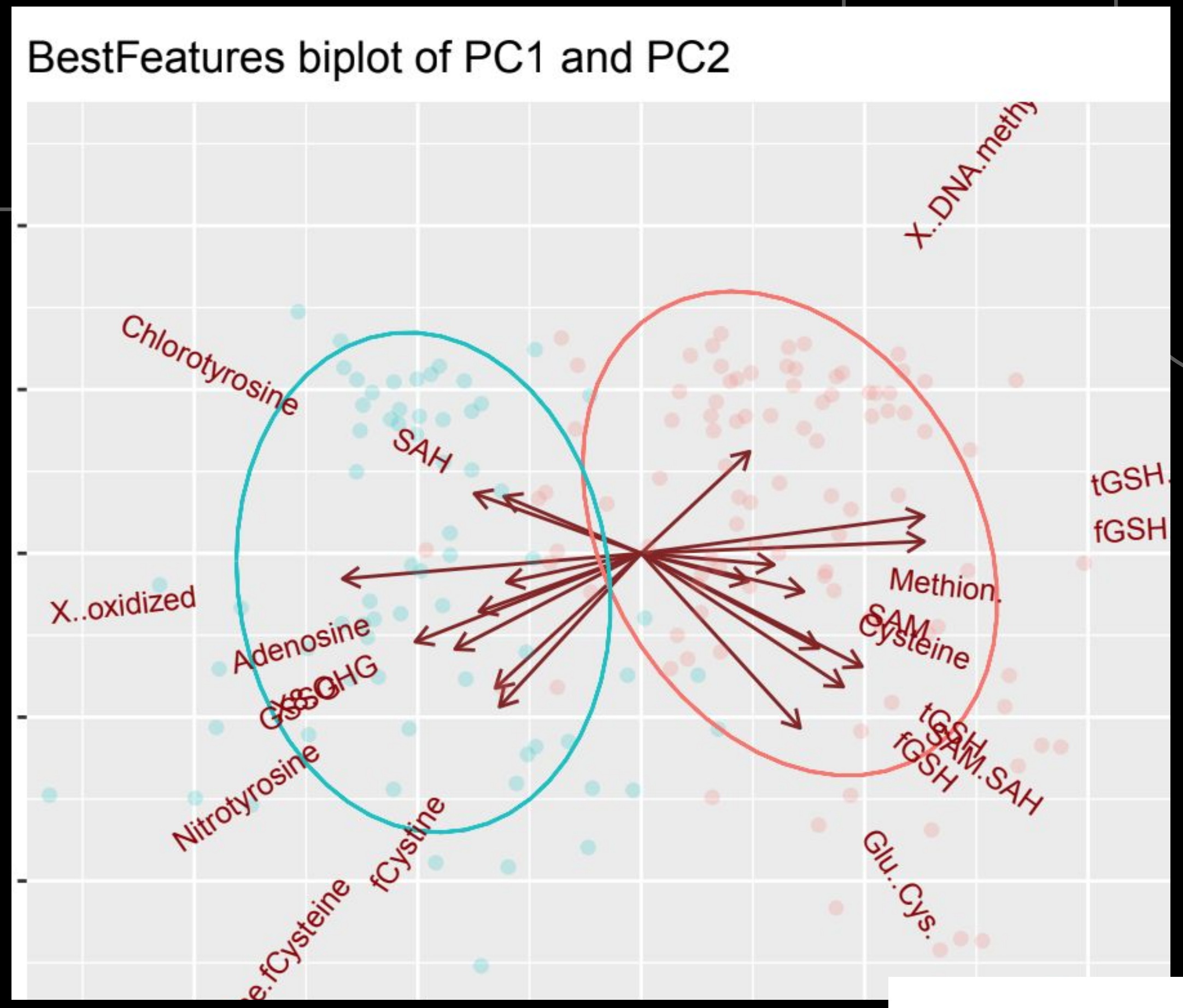
DATA ANALYSIS



Introduction

Our firm “Weesnaw” was hired as consultants to analyze the results of Dr. Hanh, and develop a computational model that analyzes and generates the prediction of Autism Spectrum Disorder (ASD) with biomarkers. We were given a data set with biomarker data and 67 samples that have ASD with the last column abelling it as ASD or not (has NEU - are neurotypical). Our task is to create an LDA model and see if we can improve t or find a better model.

```
                      Pr(>|z|)
ion.               0.000000375
                   0.000054982
                   0.001300615
SAH                0.000166221
NA.methylation     0.000000479
HG                 0.000000003
osine              0.000343522
eine               0.000006382
.Cys.              0.000224000
                   0.000000574
                   0.000003259
                   0.000000009
.GSSG              0.000000001
.GSSG              0.000000001
rotyrosine         0.000000006
otyrosine          0.000000017
tine               0.000002637
tine.fCysteine     0.000000693
xidized                4.0e-11
```



Data Description

Our training data consisted of 165 groups (rows) of patients who were diagnosed as being on the scale (ASD) vs. people who were not (NEU). There are 24 factors that represent different organic compounds found in blood samples.

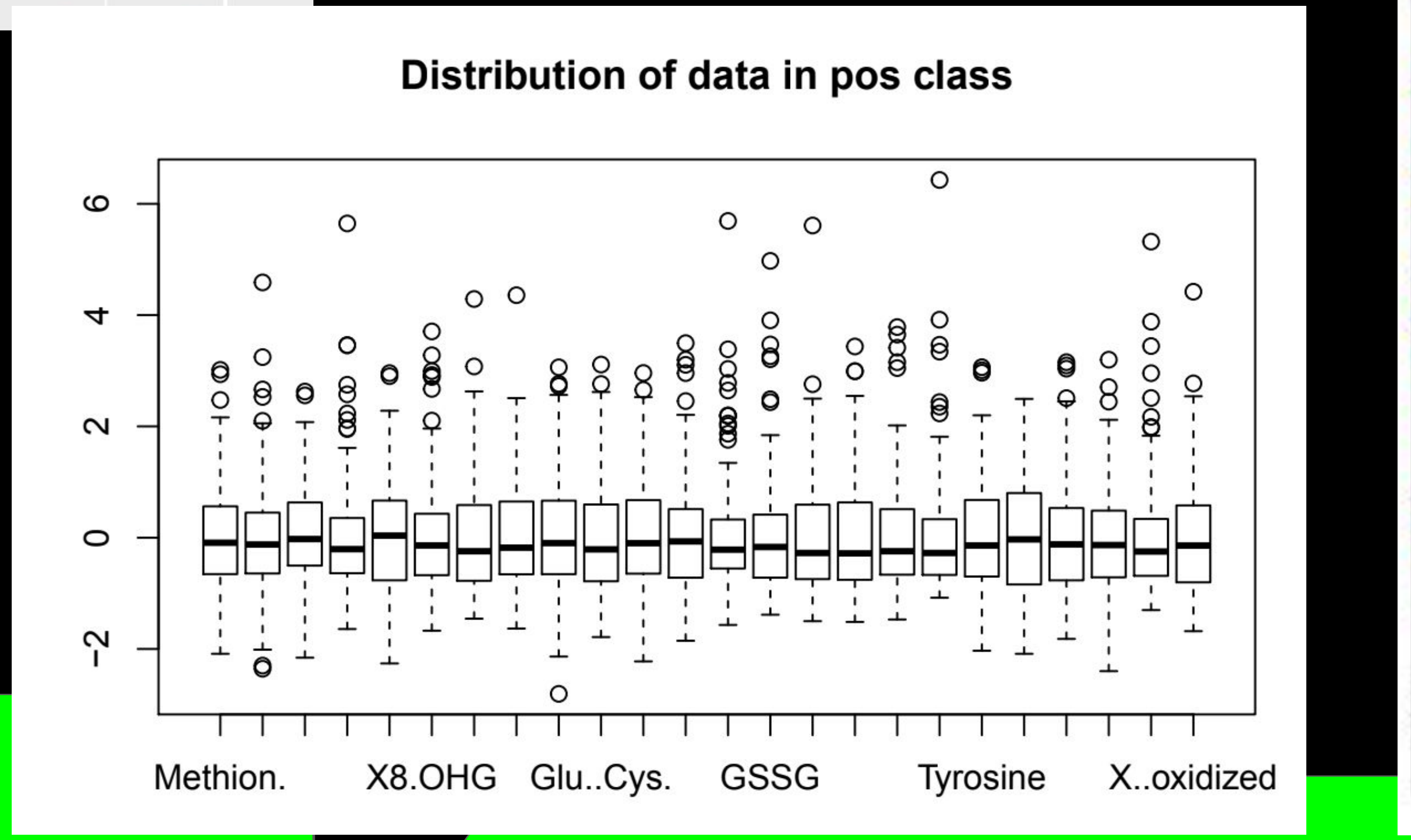
Our testing data consists of 41 groups (rows) of patients, and classifies them the same way our training data does (ASD vs. NEU based on 24 factors)

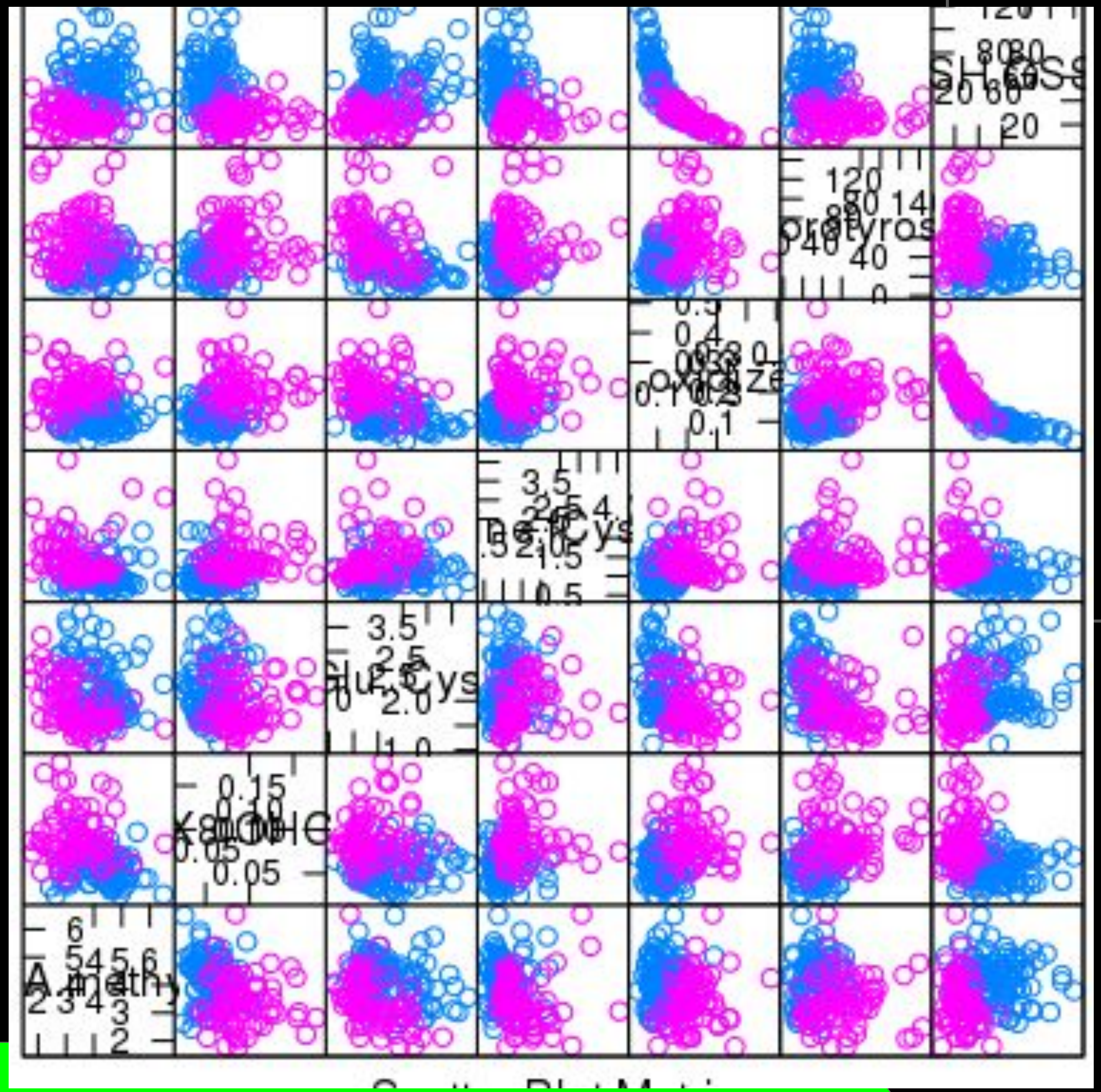
In our code we used ‘fulldat’ to store the 165 observations of the seven features chosen by Dr. Hahn. The seven features are specifically stored in ‘papervar’

Univariate Logistic Regression & Feature Challenge

- For the feature challenge we used the best features from the regression that were not the variables Dr. Hahn identified.
  - We used regression to calculate p-values which we used to determine feature importance

```
SVM(feat) Reference
87.8%         NEU ASD
Pred     NEU   24   4
         ASD    1  12
```





LDA VS SVM COMPARISON

	TRAIN SET			TEST SET		
	ASD	NEU	ALL	ASD	NEU	ALL
SVM	.984	.960	.970	.923	.857	.878
LDA	.940	.959	.949	.810	.885	.960
LOO	.938	.940	.939			

Alternative Models

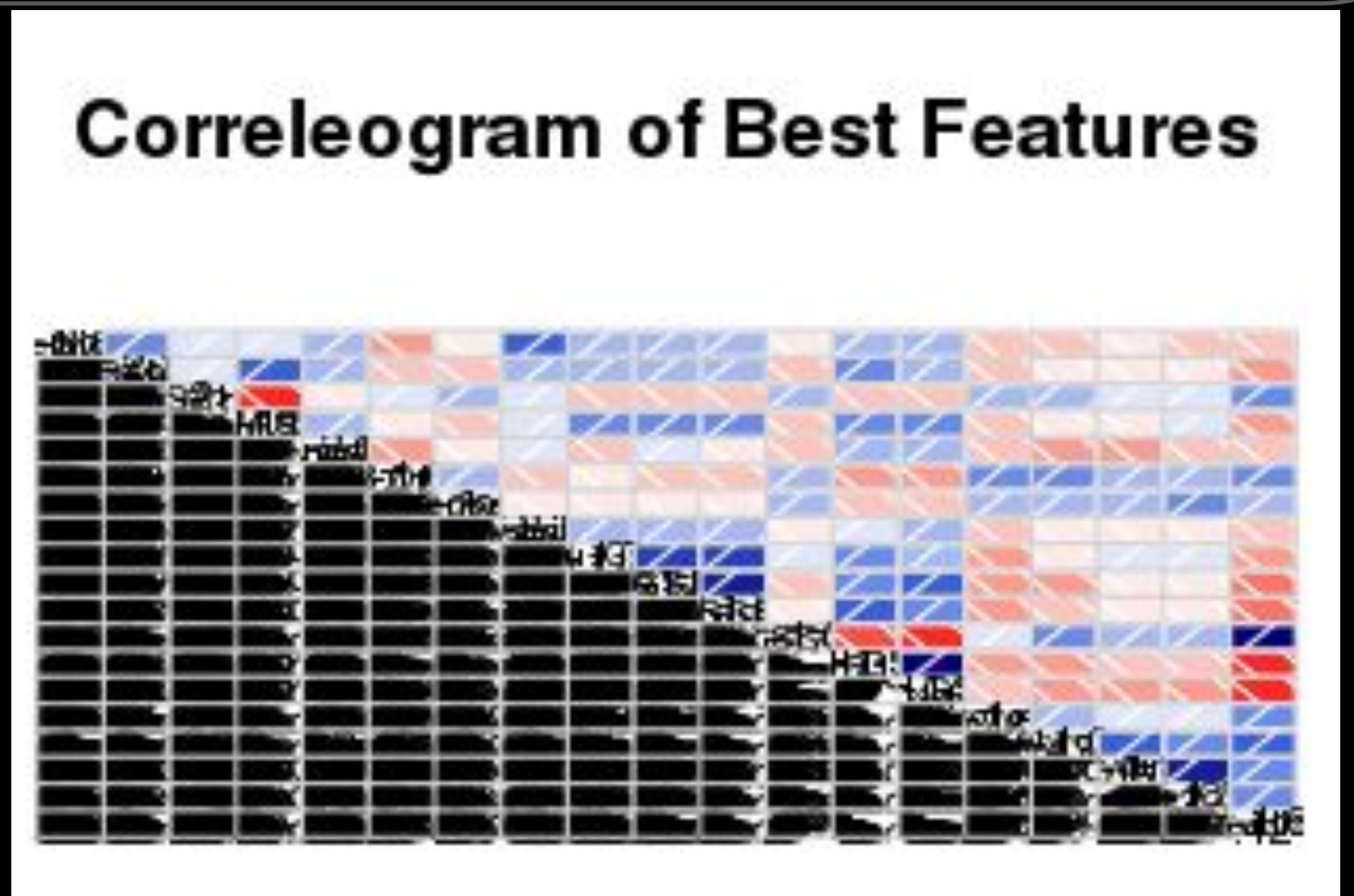
Ran multiple logistic regression on data in dataframe Train.df. Also ran SVM (87.8%) and PLSDA (88.5%.

```
SVM          Reference
              NEU  ASD
Pred    NEU    93   14
        ASD     5   53
```

```
PLSDA        Referenc
              NEU  AS
 Pred   NEU    22
        ASD     3   1
```

Additional Analysis

- Used ggplot function to examine different components determine the best way to separate data in seeing whic features are important for classification
- Used corrgram function to create a correlogram of the Best Features which compares through shading and presents points.
- Plotted the 7 paper variables against each other to see how separable they are
- plotted boxplots of both sets of features, scaled and unscaled to specifically see the variation in the range o values and to identify outliers.



Conclusion

In the end, Weesnaw has come to the conclusion that in this first round of analysis, that Dr. Hahn has picked out features as relevant as the ones Weesnaw has.

Our recommended model is LDA since it performed better on the testing data we chose, though SVM is still a good choice since it fit the training data better.

ENTATION BY: RAMIN CHOWDHURY, SEBASTIAN CASTILLO-SANCHEZ, TENZIN TASHI, MADISON CHAMBERLAIN