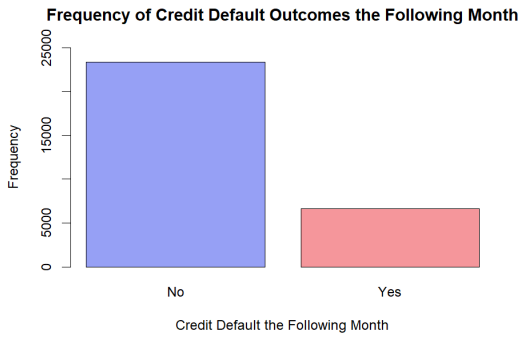
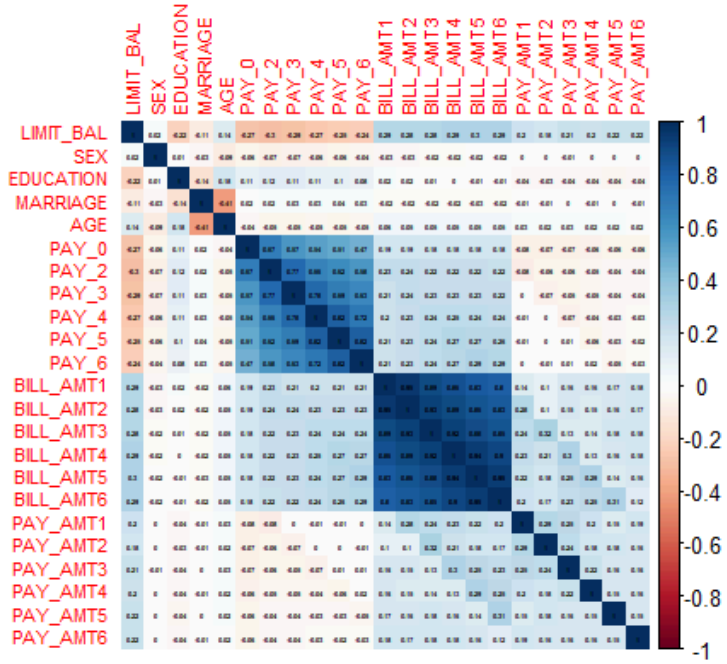
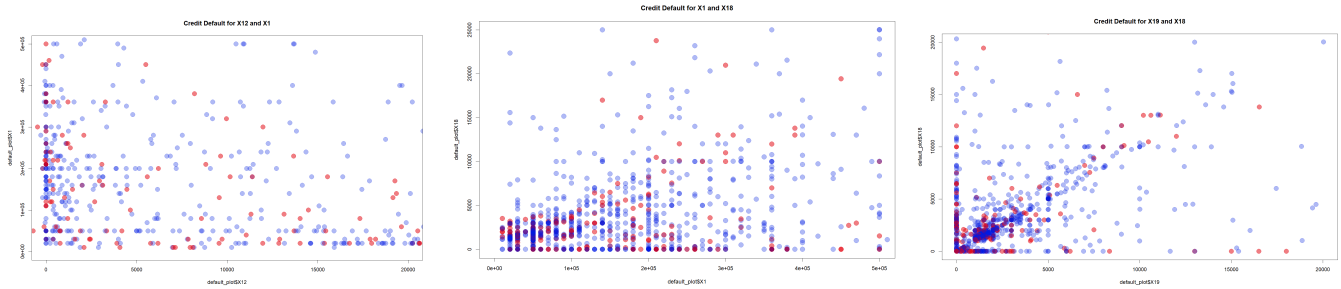
DS 5220 Milestone 2
EJ Wong and Ryan Houseman

EDA



We first want to consider the target variable for the problem we want to solve, credit default. Based on the population sample we obtained it seems that we have around 23,300 cases where the person does not default and about 6,600 cases where the person does default on their credit statement. In this case, we see that there is a large class imbalance and that we must be cautious considering generalized metrics like accuracy. We should consider additional metrics to understand and better score and predict the minority class (credit default).





Next we looked into the predictors to understand what they generally look like. When looking at the correlation between each of the predictors, we see that there is some positive correlation between payment amounts themselves and bill amounts themselves. It seems that there is some structure between these values which affect our modeling choices. In future steps we should look to test for significance of multicollinearity between these predictors before moving deeper beyond what is implemented below. Additional context into credit modeling may assist us to also qualitatively understand if these values are highly correlated with each other.

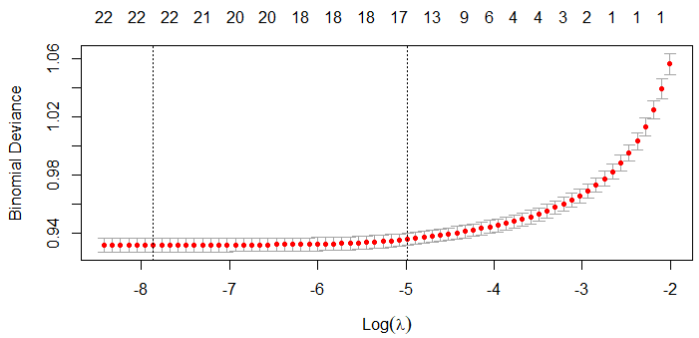
Logistic Regression Model

Initially we considered a basic logistic regression model consisting of all the predictors in the dataset. Overall we set this up as a baseline for which we believe we can increase the statistical power of our model by tuning with feature selection. In the initial logistic regression model, we can begin to get a sense of which variables might be useful predictors of credit default. In the summary below, we saw that a mix of variables with and without statistical significance in the logistic regression model. By implementing LASSO, many of these features will end up being zeroed out. We also performed 10-fold cross validation using the logistic regression model, and got a prediction error rate around 20%. We can use that metric as a baseline for model performance as we test different approaches.

```
Call:
glm(formula = Y ~ ., family = binomial, data = default)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.675e-01  1.212e-01  -5.510 3.59e-08 ***
ID          -1.338e-06  1.750e-06  -0.765 0.444396
X1          -7.615e-07  1.569e-07  -4.853 1.21e-06 ***
X2          -1.083e-01  3.069e-02  -3.530 0.000415 ***
X3          -1.010e-01  2.098e-02  -4.815 1.47e-06 ***
X4          -1.548e-01  3.171e-02  -4.883 1.05e-06 ***
X5           7.419e-03  1.779e-03   4.170 3.05e-05 ***
X6           5.771e-01  1.770e-02  32.611  < 2e-16 ***
X7           8.316e-02  2.019e-02   4.119 3.81e-05 ***
X8           7.173e-02  2.261e-02   3.172 0.001512 **
X9           2.478e-02  2.503e-02   0.990 0.322168
X10          3.336e-02  2.689e-02   1.240 0.214797
X11          7.990e-03  2.213e-02   0.361 0.718036
X12         -5.494e-06  1.136e-06  -4.836 1.33e-06 ***
X13          2.337e-06  1.505e-06   1.552 0.120572
X14          1.365e-06  1.323e-06   1.032 0.302295
X15         -8.861e-08  1.353e-06  -0.066 0.947770
X16          5.382e-07  1.522e-06   0.354 0.723630
X17          4.010e-07  1.196e-06   0.335 0.737316
X18         -1.363e-05  2.306e-06  -5.912 3.37e-09 ***
X19         -9.633e-06  2.095e-06  -4.599 4.24e-06 ***
X20         -2.723e-06  1.721e-06  -1.582 0.113614
X21         -3.967e-06  1.785e-06  -2.222 0.026286 *
X22         -3.333e-06  1.778e-06  -1.874 0.060864 .
X23         -2.065e-06  1.296e-06  -1.593 0.111076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

LASSO Tuned Logistic Model



We decided to implement LASSO to tune our model and complete feature selection. In this case we measured a loss function for logistic models, binomial deviance, to optimize the lambda penalty term which determines how many features LASSO inherently selects by zeroing them out. We found that there are two optimization points where this occurs where log lambda value is at -8 and -5 with 22 and 17 corresponding non-zero predictor values. We ended up selecting the model with 17 predictors as there is minimal loss added onto the model but there is a significant increase (approximately 0.02 binomial deviance) in statistical power as we have essentially dropped 5 predictors off of our optimized model.