Credit EDA

Matt Oehler

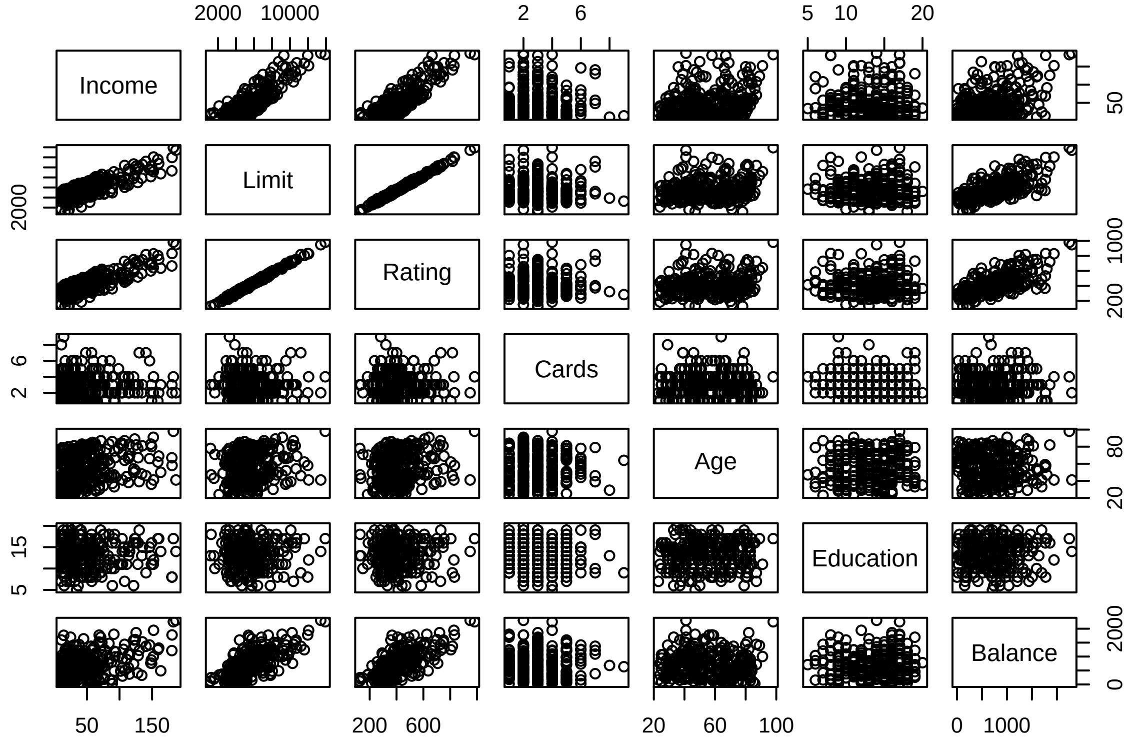
January 8, 2018

1. Goals of the Analysis

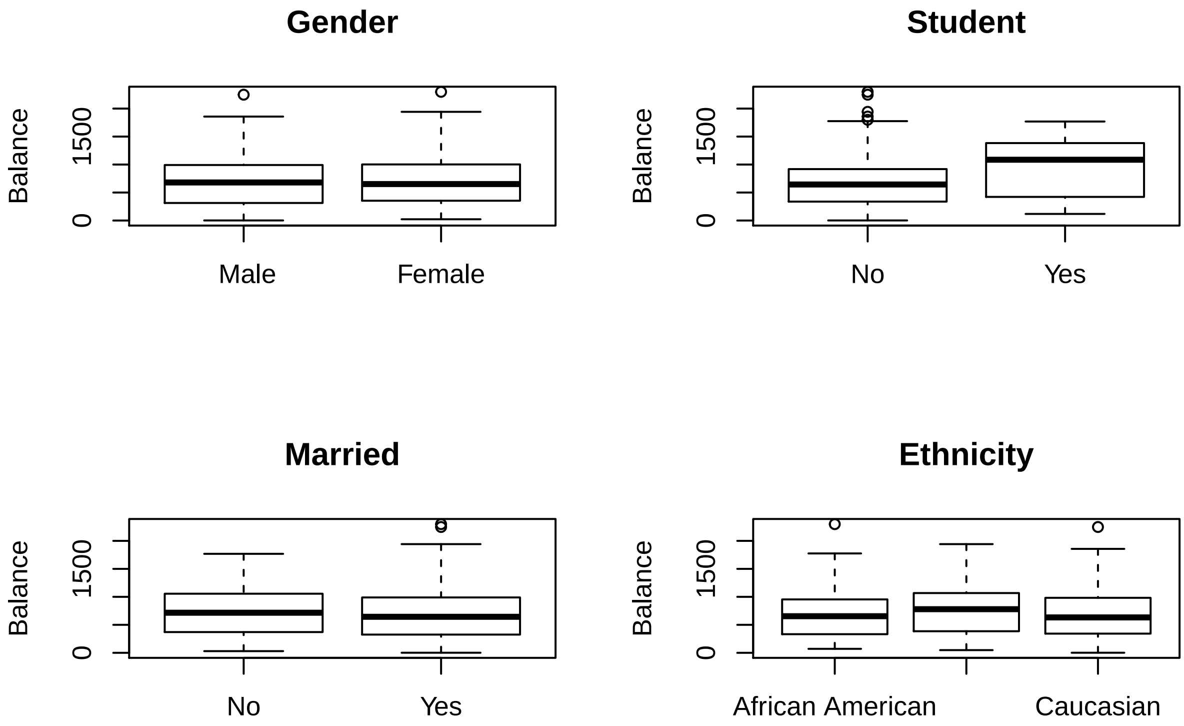
Credit card companies, like any other company, want to be profitable. In order to maximize their profits, a company wants to maximize the amount of card holders that have a moderate monthly balance, and minimize the amount of customers that have a large monthly balance. By analyzing this dataset, which contains the demographics and current balance of card holders, we hope to a) determine if there is a relationship between certain demographics of card holders and their current balance and b) create a model to predict a potential cardholder's balance so that the company can make a better decision about issuing that person a credit card.

2. Features of the Data

To start off the data exploration I created a pairs plot that shows the scatter plots for all quantitative variables in the data set. These variables include: income, credit limit, credit rating, number of credit cards, age, years of education and current balance. We can see right off the bat that credit limit and credit rating are very highly correlated. It also appears that the quantitative variables are all linearly related to balance. (Note: not all relationships are very strong).



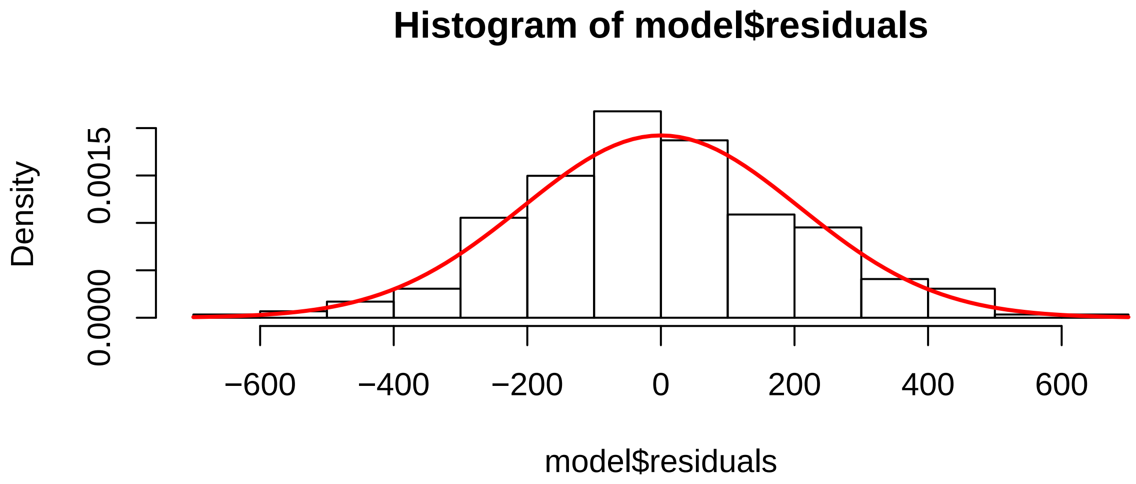
Next I made boxplots for all of the categorical variables and how they relate to current balance. Categorical variables include: gender, student status, marriage status, and ethnicity. Just by looking at the plots it seems that gender and marital status don't appear to have a large effect on balance. It does, however, appear that student and ethnicity might have an effect on balance.

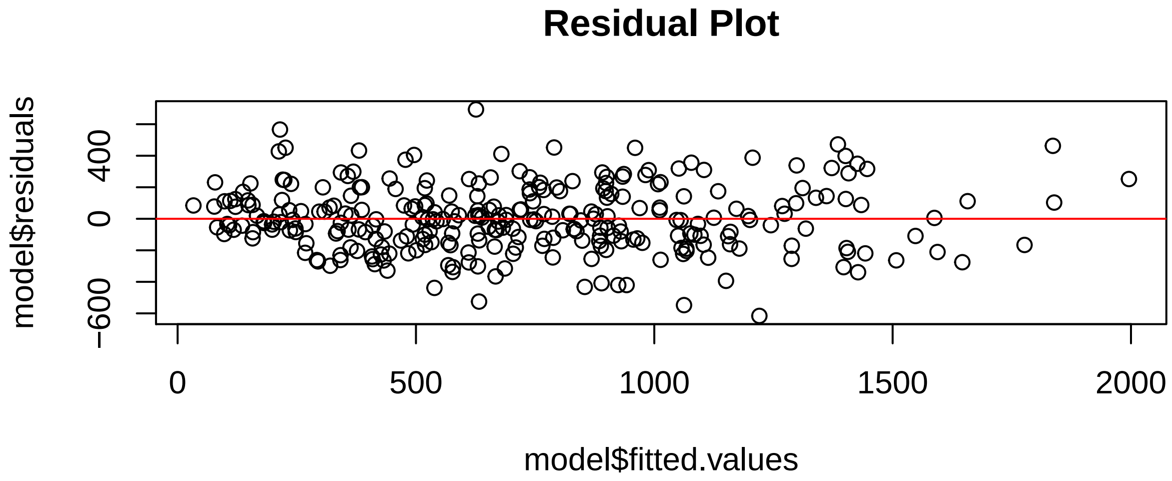


3. Statistical Method

I believe that multiple linear regression is an appropriate approach that can be used to answer our questions of interest because this data set has a single continuous response variable. We could use the model that we fit to calculate estimates of the relationship between certain variables and credit balance, and we could also use that model to predict the credit balance of a potential card holder. The assumptions of multiple linear regression model (normality, equal variance, and independance) mostly seem to hold as well, as shown in the following plots.

The historgram of residuals shows us that our residuals are approximately normally distributed. The plot of residuals vs fitted values shows that the variance for the residuals is approximately constant throughout the data set. The independence assumptions is validated by critical thinking, rather than through graphical representation. It makes sense that some of these variables are independent of one another (e.g. gender and ethnicity), but some variables such as credit limit and credit rating are very highly correlated and are therefore not likely to be independent. These issues can be dealt with when refining the statistical model by removing or transforming the variables that might cause issues.





4. Things I don't know

I'm not sure exactly how to best deal with issues that arise from the violation of the independence assumption. I know that the variance inflation factor can be used to decide which variables are worth removing from the model, but I'm not sure how to deal with cases where throwing out variables isn't a good solution.