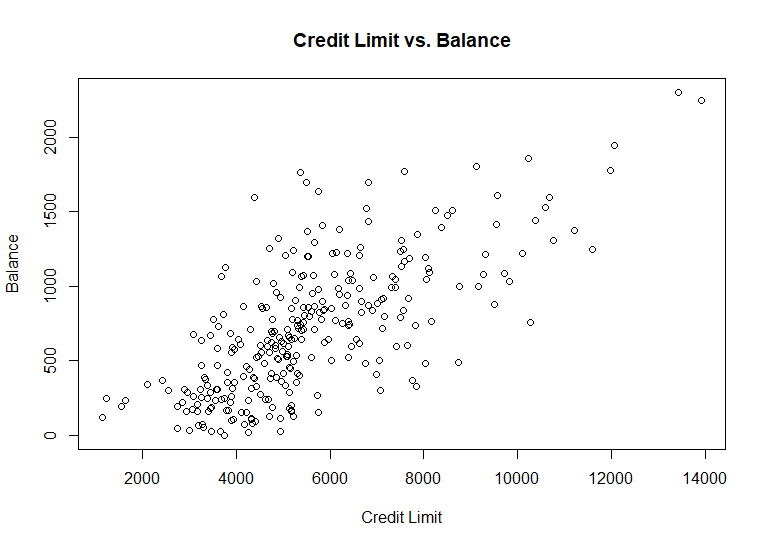
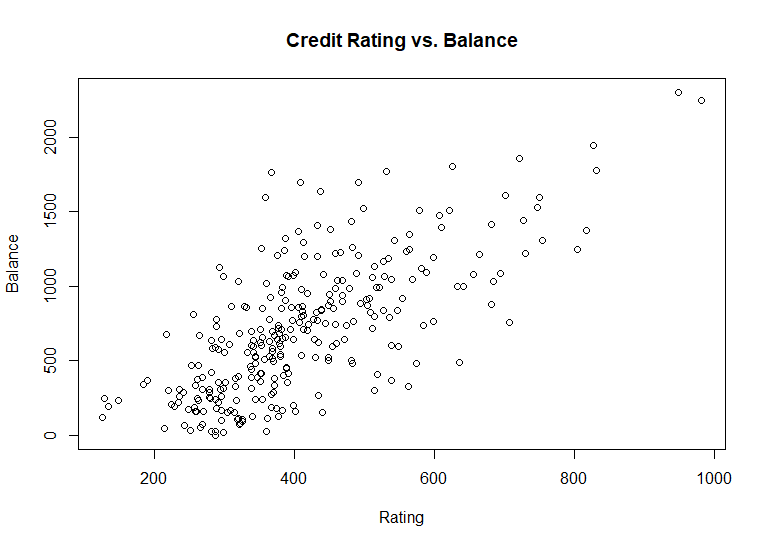
Jackson Curtis

Credit Risk EDA

Companies that issue credit cards profit by credit card users carrying moderate balances on their account. Large balances run the risk of bankrupting the customer and never getting paid off, while small balances don’t accrue interest for the issuing company. Therefore, we are interested in learning about what kind of balance a customer will carry, and how it relates to the other characteristics of the customer. By studying how things like income, age, and credit rating affect account balance, we can make predictions and quantify how confident we are that customers will behave in a specific way. When credit companies understand their customers, they will be able to avoid potential losses (due to lack of repayment) and focus advertising on their most profitable customers.

We have four categorical variables on our customers. They are ethnicity, gender, in school, and marital status. Each has good sample size for each possible category (male/female, Caucasian/Asian/etc). Of all possible combinations of categorical variables, only one combination has zero individuals (Asian, married, male, student).

 Analyzing the quantitative variables showed that many of the variables (income, limit, rating, and cards) were right skewed, but there were few points that would clearly be labeled as outliers or very different from the others. Analyzing the relationships between credit card balance and the explanatory variables turned up a few surprises. The graph on the left shows that carrying a large credit card balance is positively correlated with a good rating. While this surprised me, one explanation that I found was that credit limit is also positively correlated with balance, but the scales are much different. Those carrying large balances often have limits four times as big as their balance.



This problem seems well suited for a regression analysis. The relationships between the variables seem to be linear or that they could be made linear with transformations. The categorical variables have few levels and are pretty even. There are very few extreme observations in any of the variables, so it seems like many of the normal distribution assumptions will hold. Regression is extremely flexible for inference. By using regression we will be able to obtain confidence intervals on the effect of each explanatory variable as well as prediction intervals on individual observations. This will aid researchers in identifying candidates, and providing how estimates of how they will behave when selected.

One aspect of this analysis that concerns me is how to handle the extreme colinearity between credit rating and credit limit. The only solution that comes to mind is to throw out one of the variables, but that may be a bad option if it throws out useful information. In general, dealing with variables that aren’t necessarily observed, but instead assigned is something I’m cautious about.