Stat536 - HW1 Exploratory Data Analysis

Jacob Mortensen

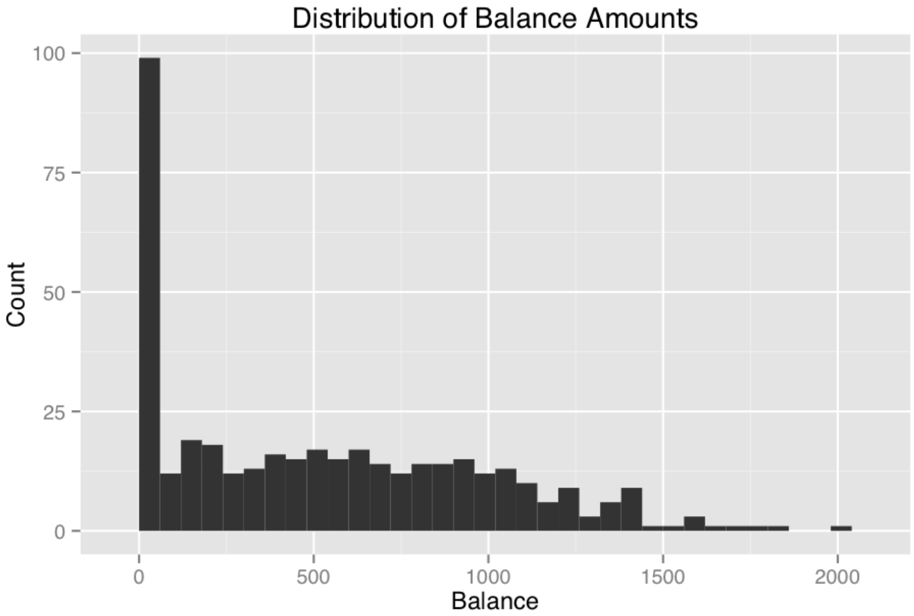
January 9, 2015

1. Goals of Analysis

Like any business, a credit card companies primary concern is their bottom line: they need to make enough money to stay afloat. The easiest way for them to make money is on interest charged on outstanding balances. This would make it seem that all they need to do to increase profits is to enroll those who will maintain the highest balances on their credit cards. However, those with the highest balances are also those who tend to default on their debt most often, and unlike other forms of monetary lending, credit card debt is unsecured, leaving credit card companies with no way to recoup their investment if borrowers default. Therefore, credit card companies have a vested interest in identifying which borrowers will maintain a low or moderate balance on their cards without creeping into the high balance terrain where they are more likely to default. This is a difficult process because credit card companies have access to only a limited amount of information about their applicants before issuing them a card. As a result our goal is to use this data set to create a model that predicts which borrowers are more likely to maintain a high balance in order to avoid lending to those who will lose money for the company so that we only issue cards to those applicants who will maintain a low to moderate balance on their cards.

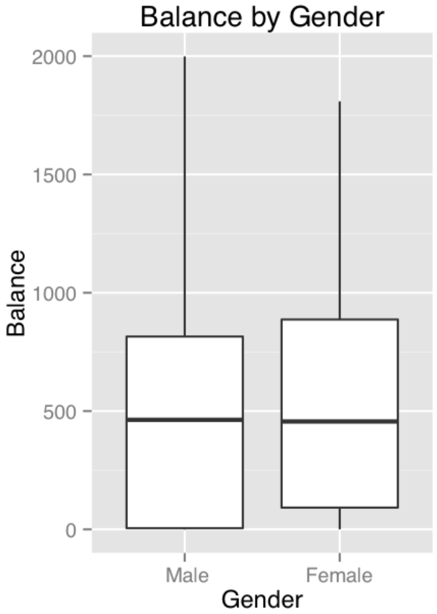
2. Summary of the Data

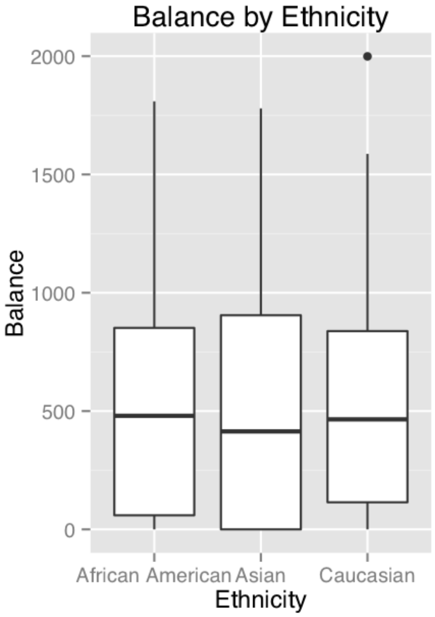
Our data set contains 11 variables with 400 observations. There are no missing values in the dataset. The data set contains variables that are both continuous and categorical, and so that will need to be taken into account when analyzing the data.



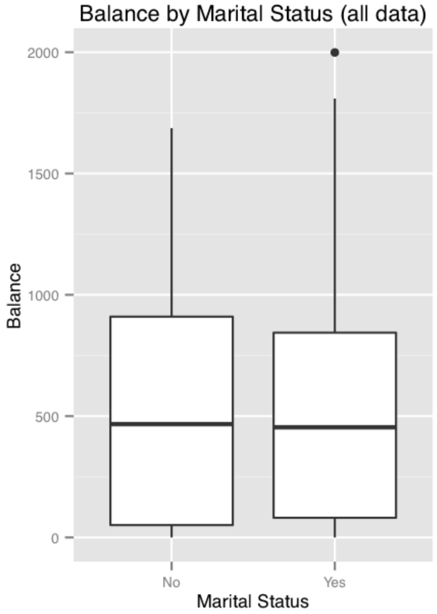
In this plot we get an idea for how the response variable (credit card balance) is distributed. We see that none of the values can be less than zero and as a result the data is highly skewed to the right, suggesting that we need to be careful in using any statistical method that assumes normality. This can also help us make decisions about what we should consider a high credit card balance.

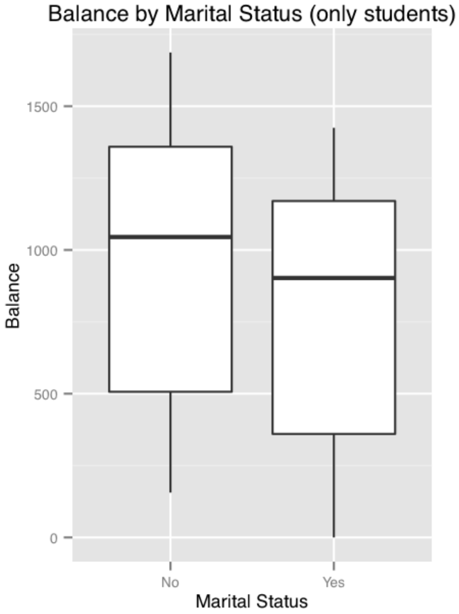
In order to examine potential relationships between credit card balance and the other variables I created several box plots and scatter plots. For sake of brevity, I have omitted most of those and chosen to include just a few that seemed most likely to yield interesting results.


Balance by Gender

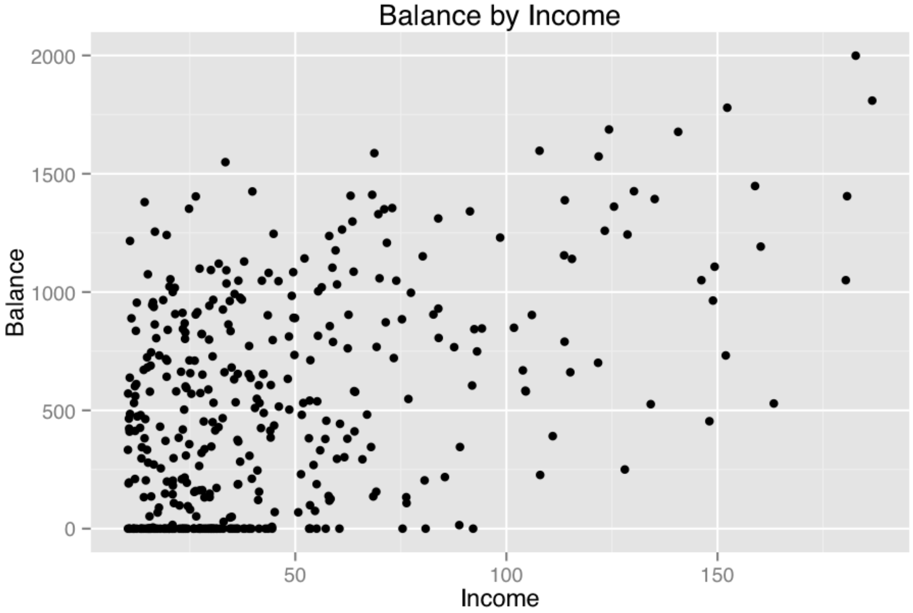

Balance by Ethnicity

Here we can see that neither ethnicity nor gender appears to have a significant relationship with balance, suggesting that these variables may not be significant contributors in our model. We will still test them to be certain, but my first impression is that these will not end up being included.

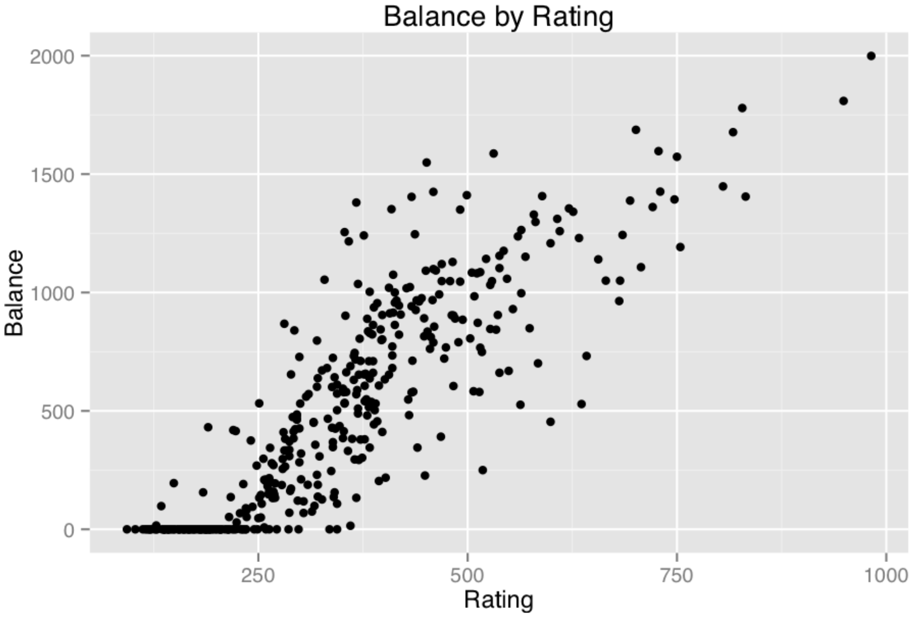

Balance by Marital Status (all data)


Balance by Marital Status (only students)

This plot shows that at first glance, marital status may not appear to have any relationship with balance, but if we condition the data on whether or not the holder is a student and then compare marital status, we observe a difference, indicating that an interaction may exist between the two variables.



This scatter plot shows us a couple of features of interest. The first is that there appears to be a positive correlation between credit card balance and income, so those with higher incomes tend to have higher credit card balances. The second is that income appears to be highly skewed, with a vast majority of the values falling below $75,000. Again, we will need to account for this when conducting our analysis.



This is the final plot that I will include. It shows that rating and balance have a strong positive correlation, which is unsurprising. If this rating is determined before the holder begins using their credit card, then this provides validation for the rating scheme the company is using, making it a very useful in predicting future balance.

In examining the rest of the plots, I did see that student and balance appeared to have a relationship, indicating that student will most likely be included in the model. Of course, we knew that anyway, since we showed that there appears to be an interaction between marital status and student, thus requiring student to be included in the model anyway. Most of the other variables did not appear to have a strong relationship with balance, suggesting that they may not be included in the final model, although, of course, we will test them to be certain.

3. Proposed Method

Based on the above information, it seems like an appropriate method of analysis for this dataset would be linear regression, although we may have to perform some transformations on the data to account for skewness. This would help us to accomplish our goal of determining high risk borrowers because once we have created a model, we can collect the appropriate information and plug it into our model, and if anything extends past the cut off we determine, then we will know which applications we should reject and to whom we should issue a credit card.

4. Things to Figure Out

One thing that I don't know how to do is how to determine the cutoff for high-risk borrowers. However, I feel like that might be more subject matter related than methodology related. In regards to methodology, something that I am not sure how to do is deal with the highly skewed and bounded data we see with the balance and income variables. Since these variables have such a high frequency of values on the lower end it violates some of our assumptions and could potentially make linear regression much more difficult. In previous classes I know we have used a log transform to account for skewness but I don't know if that is appropriate in this case, or if it is appropriate, if there is a better way to deal with that problem.