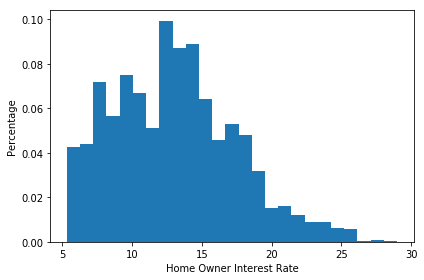
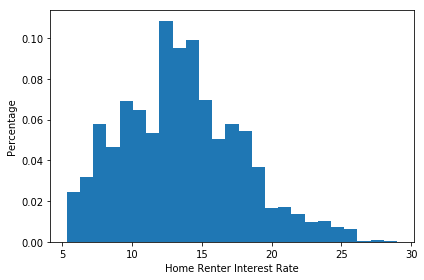
Inferential Statistical Analysis On Lending Club Data

The goal of my project is to predict if a borrower will default or pay back his/her loan on time as well as gather as many useful insights as possible. To predict loan default, I fit a Random Forest Model to the data gathered from the Lending Club data set. In this section, I want to cover my analysis that went into discovering other useful insights that might aid in assessing the best risks for lending money.

After fitting a Random Forest Model, interest Rate showed to be a significant predictor in determining default rate. I want to further investigate, using inferential statistics, any relationships or other findings between borrower groups. First, I looked at comparing homeowners and home renters to see if there were any differences in the interest rates charged. Interest rates for home owners average 13.026% and 13.576% for renters. Let’s look at some graphs to get a good look at the distributions.





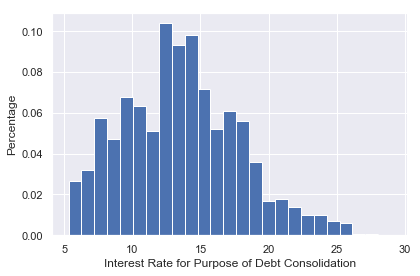
Just looking at the graphs and data it is difficult to tell if these groups are really different. To determine if these are significantly different I will use the following hypotheses using alpha=0.01:

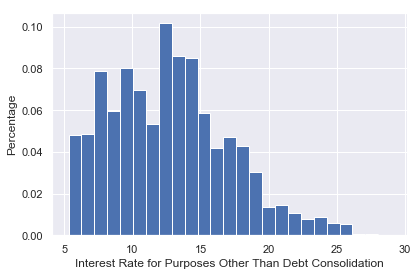
H\_o: The mean difference between the interest rate for home owners and home renters is 0

H\_a: The mean difference between the interest rate for home owners and home renters is not 0

First it’s important to note that the distributions, based on the graphs above, they don’t appear to fit a normal distribution so a t-test may not be appropriate. Instead I will use bootstrapping to test the null hypothesis. To do a bootstrap hypothesis test, I took the interest rates for both groups and subtracted out the mean for each respective group. Then I added back in the mean of interest rates for both groups together. This shifts the mean for each group to be equal to the overall mean. Now I can bootstrap sample from each of these groups (with shifted means) and compare the difference in means to the observed difference. Assuming the null hypothesis is true, I wouldn’t expect to find a significant amount of samples as or more extreme than the observed difference. From this test, I found that none of the 10,000 samples were more extreme than what was observed, so I conclude that there is a statistically significant difference between the means of each group. On average, home owners will pay 0.55% lower interest rate than home renters.

Next, I looked at interest rates between individuals requesting a loan for purposes of debt consolidation vs. any other reason. Below are histograms of the two groups.





As with the previous analysis, I used bootstrapping to compare the means of each group with the hypothesis:

H\_o: The mean difference between the interest rate for purposes of debt consolidation and anything else is 0

H\_a: The mean difference between the interest rate for purposes of debt consolidation and anything else is not 0

The p-value came in well below our stated alpha of .01, indicating there is a significant difference between mean interest rates for people who request a loan for debt consolidation and other purposes. On average, people who request a loan for debt consolidation pay a 0.87% higher interest rate.