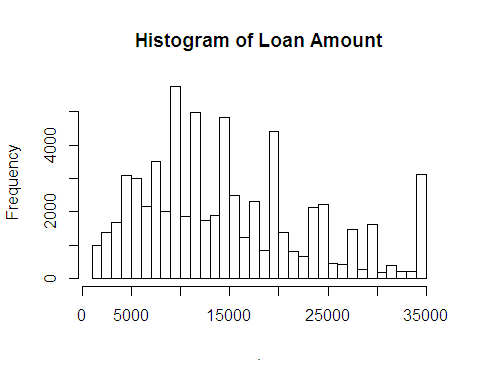
Lending Club Statistical Analysis

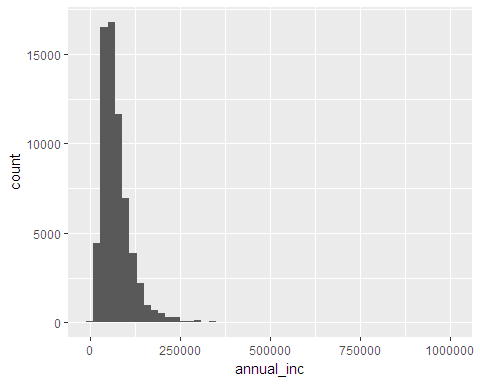
### The Data: What Questions Can Be Answered?

The dataset used for this analysis contains roughly 650,000 personal loans issued through Lending Club between 2014 and 2015. In order to operate within the memory capacity of my PC, I downsampled the data to 65,000 rows after cleaning.

Now that the data is clean and scaled down to a manageable size, the question we would like to find an answer to is this: are there certain aspects of a loan that would suggest a higher risk of defaulting? If so, how do these riskier loans impact the overall profitability of the lending service?

### Analysis

First, let's look at the distributions of loan amount and annual income as they are inextricably linked to a borrower's ability to pay the installments.

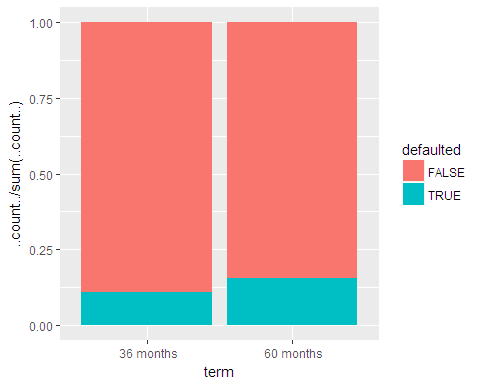


Loan amount is a slightly right tailed distribution with a large number of borrowers taking out the max loan amount of $35,000. We can see that borrowers typically apply for loans in multiples of $5,000. Annual income is also a right-tailed distribution with a very small number of borrowers earning between $250,000 and $1,000,000 annually and 75% of borrowers earning less than $90,000/year.

Next, calculating the default rate will allow us to compare the relative risk levels of different groups of loans.

## [1] 0.124109

Looking at the defaults, our dataset has a sample proportion of 12.4% of loans in default. We can use this as a benchmark when breaking down the default rate among subsets ot the total sample. Diving into the possible explanatory variables, we will first look at the loan term. Generally, a longer term on a loan reflects a desire to spread out repayment over a longer period which may be indicative of certain constraints in the borrower's monthly budget.



##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(sum(loans60), sum(loans36)) out of c(length(loans60), length(loans36))  
## X-squared = 301.28, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.04216650 0.05363397  
## sample estimates:  
## prop 1 prop 2   
## 0.1566660 0.1087658

Borrowers on a 60 month loan term defaulted at a rate of 15.9% while borrowers on a 36 month loan defaulted at 10.8%. Using a 2-sample test of equal proportions, we find that the null hypothesis of equal proportions should be rejected at a 95% confidence interval and that 60-month loans indeed have a higher default rate than 36 month loans. In searching for an explanation for the higher default rate among 60 month loans, we return to the topic of annual income and loan amount.

income\_60 <- loans[loans$term == " 60 months", "annual\_inc"]  
income\_36 <- loans[loans$term == " 36 months", "annual\_inc"]  
summary(income\_60)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9000 52000 70000 81146 96000 1000000

summary(income\_36)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6610 43500 61000 72896 88139 1000000

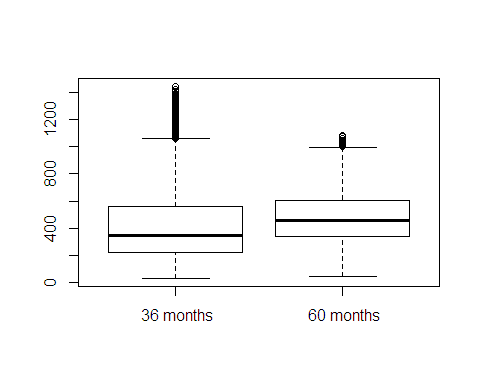
loanamt\_60 <- loans[loans$term == " 60 months", "loan\_amnt"]  
loanamt\_36 <- loans[loans$term == " 36 months", "loan\_amnt"]  
summary(loanamt\_60)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1500 14169 19200 20164 25000 35000

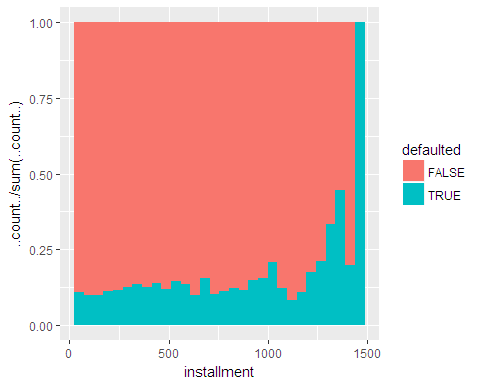
summary(loanamt\_36)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1000 6800 10000 12709 17000 35000

It turns out that borrowers on 60 month loans have a 14.8% higher median income than borrowers on 36 month loans. However, the median loan amount is 92% higher for 60 month loans, which is way out of proportion with the difference in income. This may suggest that the borrowers on 60 month loans are struggling to pay their monthly installments due to inadequate cash flow.

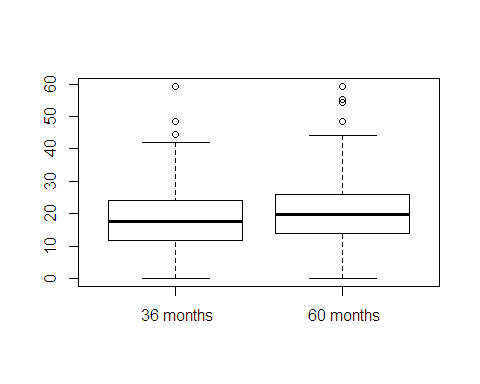


The above boxplot confirms that monthly installments are also higher for 60 month loans. Since we have a connection between 60 month loans and higher monthly installments, it will also help to show a density plot filled with default rates across different installment bins.

As expected, higher monthly installments on personal loans correlate with higher default rates. Another variable that may negatively impact a borrower's ability to pay off their loan is their total debt as a percentage of their annual income. Other things equal, a greater amount of total debt translates to a greater amount of interest paid on that debt, in addition to having to pay off the larger principal amount.

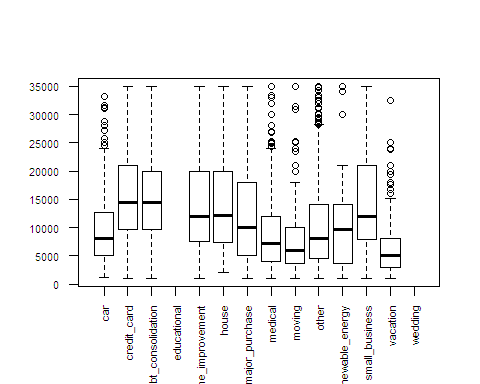
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 13.74 19.61 20.00 25.91 59.26

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 11.83 17.60 18.13 24.00 104.00



The median debt to income ratio for borrowers on 60-month loans is 19.61% compared to 17.6% for 36-month loans. This isn't a particularly compelling statistic, however debt to income shouldn't be ruled out as an explanatory variable on its own.

The next variable selected was loan purpose, which initially consisted of 14 factors but was reduced to 12 due to small sample sizes for certain factors. A series of boxplots below displays the distribution of loan amounts by loan purpose.



Small business loans tend to be the largest amount, followed by credit card loans and debt consolidation loans. Let's have a look at which loan purposes have the highest default rate using a table of default proportions:

##   
## FALSE TRUE  
## car 0.89674952 0.10325048  
## credit\_card 0.90506329 0.09493671  
## debt\_consolidation 0.86595761 0.13404239  
## educational   
## home\_improvement 0.89261917 0.10738083  
## house 0.83018868 0.16981132  
## major\_purchase 0.86933798 0.13066202  
## medical 0.85873016 0.14126984  
## moving 0.79575597 0.20424403  
## other 0.86185777 0.13814223  
## renewable\_energy 0.83783784 0.16216216  
## small\_business 0.81709402 0.18290598  
## vacation 0.85674931 0.14325069  
## wedding

Moving loans have the highest chance of defaulting at 20.4%, followed by small business loans at 18.3% and house loans at 17%. This is an interesting finding as moving loans are among the lowest average loan amounts. Somewhat unexpectedly, loans for credit card debt have the lowest default rate at 9.5% despite having one of the highest average loan amounts.

##   
## MORTGAGE OWN RENT  
## car 0.36137667 0.13766730 0.50095602  
## credit\_card 0.48388953 0.10964071 0.40646976  
## debt\_consolidation 0.49500013 0.09969212 0.40530775  
## educational   
## home\_improvement 0.75756023 0.15838032 0.08405946  
## house 0.32075472 0.12264151 0.55660377  
## major\_purchase 0.39634146 0.12891986 0.47473868  
## medical 0.45396825 0.10952381 0.43650794  
## moving 0.18302387 0.02652520 0.79045093  
## other 0.38495575 0.11980939 0.49523485  
## renewable\_energy 0.35135135 0.21621622 0.43243243  
## small\_business 0.42222222 0.12991453 0.44786325  
## vacation 0.36088154 0.12121212 0.51790634  
## wedding

Comparing homeownership types by loan purpose, we find that 79% of moving loans were taken out by renters while only 40.6% of credit card loans went to renters. The question that arises is whether or not renters default at a higher rate than the overall population.

##   
## FALSE TRUE  
## MORTGAGE 0.8940371 0.1059629  
## OWN 0.8724490 0.1275510  
## RENT 0.8541139 0.1458861

##   
## One Sample t-test  
##   
## data: def\_renters  
## t = 9.954, df = 26033, p-value < 2.2e-16  
## alternative hypothesis: true mean is greater than 0.124109  
## 95 percent confidence interval:  
## 0.1422874 Inf  
## sample estimates:  
## mean of x   
## 0.1458861

## # A tibble: 4 x 2  
## home\_ownership income  
## <fctr> <dbl>  
## 1 MORTGAGE 85193.74  
## 2 OWN 70149.53  
## 3 RENT 64920.84  
## 4 <NA> 78000.00

The proportion table above shows that renters have a default rate of 14.6% which is higher than our total sample proportion of 12.4%. Additionally, the t-test confirms that our null hypothesis of equal proportions should be rejected. As an explanation for what would cause renters to default more often, the above table shows that renters earn an average annual income of $64,920 compared to the "Own" and "Mortgage" groups at $70,150 and $85,194, respectively. Similarly, let's look at the annual incomes of different loan purposes:

## # A tibble: 12 x 2  
## purpose income  
## <fctr> <dbl>  
## 1 car 65319.65  
## 2 credit\_card 76164.04  
## 3 debt\_consolidation 74444.34  
## 4 home\_improvement 88311.22  
## 5 house 80387.31  
## 6 major\_purchase 78010.40  
## 7 medical 71158.25  
## 8 moving 65842.29  
## 9 other 69269.72  
## 10 renewable\_energy 65740.70  
## 11 small\_business 97547.28  
## 12 vacation 67803.73

It turns out that borrowers of moving loans had the second lowest mean annual income, which makes sense as so many of them are renters. It's likely that someone who takes out a loan in order to pay moving expenses has very little savings in addition to zero home equity (as they are usually renters), meaning the additional monthly expense of a personal loan is more likely to cause difficulty for them.

From a lender's standpoint, it is common to view a borrower's employment length as a measure of financial stability. A borrower who frequently leaves jobs may not have the consistent income required to make monthly payments on a loan, while a borrower who has been employed longer is not only likely to have higher earnings, but may have an easier time budgeting as they have been living on a consistent income for longer. Let's see if this holds true:

## # A tibble: 12 x 4  
## emp\_length default income med\_income  
## <fctr> <dbl> <dbl> <dbl>  
## 1 < 1 year 0.1395571 70934.52 60000  
## 2 1 year 0.1334600 71496.46 60000  
## 3 10+ years 0.1088417 82327.80 72000  
## 4 2 years 0.1271868 73167.03 60450  
## 5 3 years 0.1283994 73955.46 61250  
## 6 4 years 0.1165837 74721.25 62400  
## 7 5 years 0.1158650 75600.91 64500  
## 8 6 years 0.1328622 75068.99 63100  
## 9 7 years 0.1255479 75719.93 65000  
## 10 8 years 0.1342105 77116.76 65000  
## 11 9 years 0.1354246 76504.77 65000  
## 12 <NA> 0.1652026 49463.53 45000

As expected, income gradually increases as employment length increases, from a mean of $70,935 for borrowers with less than 1 year employed to $82,328 for those with 10 or more years of employment. Additionally, we can see the default rate gradually decrease in the same direction, with less than 1 year of employment defaulting 14% of the time and 10 or more years defaulting 10.9% of the time. What is even more convincing is the borrower's who listed "N/A" for their employment length: a mean income of just $49,464 and a default rate of 16.5%.

### Conclusion

Analyzing the Lending Club data set has revealed the key elements of good and bad loans. The dataset consisted of roughly 12.4% bad loans, but we were able to dive down and find certain characteristics that are associated with riskier loans. Some were pretty intuitive, such as annual income or debt-to-income ratio, while others may come as a surprise to someone with less domain knowledge, such as more defaults among longer loan terms or less defaults for loans that are used to pay off credit card debt. These variables will be used to build models that can predict the chances of a loan being successfully paid off as well as the potential payout, i.e. interest collected.