Predicting financial behavior, pipe dream or possible?

What is Lending Club?

Lending Club is a third party that facilitates direct peer-to-peer (P2P) lending. P2P lending can be both beneficial to the borrower and the lender by eliminating the need for banks or other big money institutions.

Everyday people can improve their financial portfolio from P2P lending as there are very few restrictions for investors (the creditor) to lend money and borrowers save money by using a digital platform. Going completely digital means fewer costs to pay for (lower interest rates), ease of setting up automatic payments, and quick approval times. The creditor/investor benefits from receiving interest payments and borrowers benefit from lower interest rates.

As with all lending practices, there is a risk on both the creditor and the borrower. Naturally, if the borrower defaults it lessens their ability to borrow in the future (mortgages, car financing, perhaps even passing credit check for employment). However, the upfront risk of actually losing cash belongs to the creditor. With this analysis, we are looking to minimize that risk by only investing in loans that we feel have the most likely outcome of being fully paid.

Anyone can be a passive investor in Lending Club and with this report we will discuss what can be understood from the public data that can be found on Lending Club's website as well as making a prediction on whether or not the current investments (loans) on Lending Club are likely to be paid off or not.

Lending Club's Data Exploration

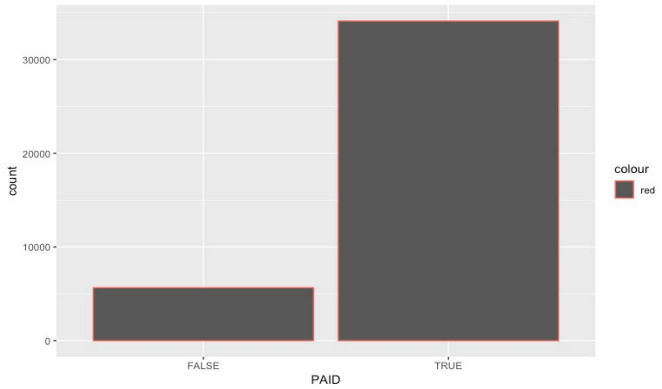
Lending Club has a downloadable dataset of loan observations from 2007-2011. From this data set, observations are categorized by many helpful variables. A sample of these variables includes: Home Ownership Status, Debt to Income Ratio, Annual Income, Interest Rates, and State of Residence (https://www.lendingclub.com/info/download-data.action).

A sample of a loan observation is below:

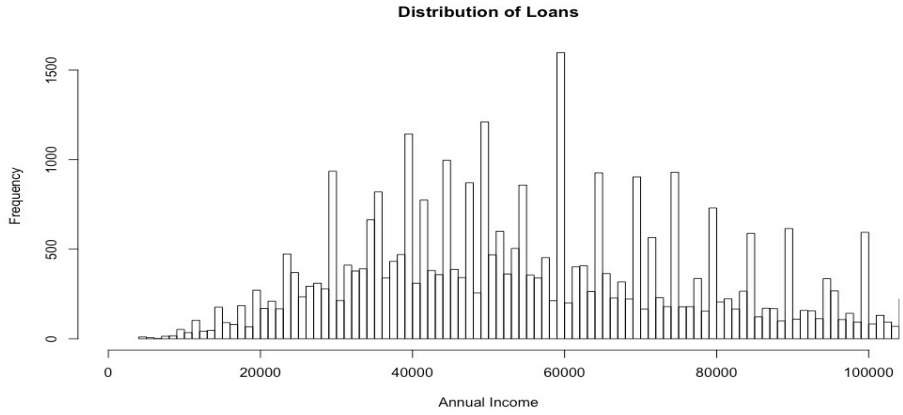
loan_amnt	term	int_rate	installment	installment_	grade	grade_nume	sub_grade	emp_length	home_owne	home_owne	annual_inc	loan_status	purpose
5000	36 months	10.65%	162.87	160	B	2	B2	10+ years	RENT	3	24000	Fully Paid	credit_card
2500	60 months	15.27%	59.83	60	C	3	C4	< 1 year	RENT	3	30000	Charged Off	car
2400	36 months	15.96%	84.33	80	C	3	C5	10+ years	RENT	3	12252	Fully Paid	small_busine
10000	36 months	13.49%	339.31	340	C	3	C1	10+ years	RENT	3	49200	Fully Paid	other
3000	60 months	12.69%	67.79	70	B	2	B5	1 year	RENT	3	80000	Fully Paid	other
5000	36 months	7.90%	156.46	160	A	1	A4	3 years	RENT	3	36000	Fully Paid	wedding

Data Visualizations

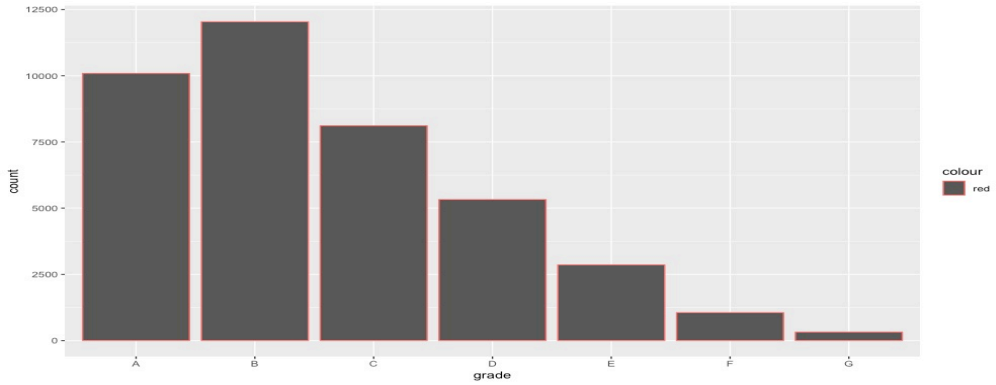
Out of the nearly 40,000 loan observations, it's safe to say that majority of these loans are paid off:



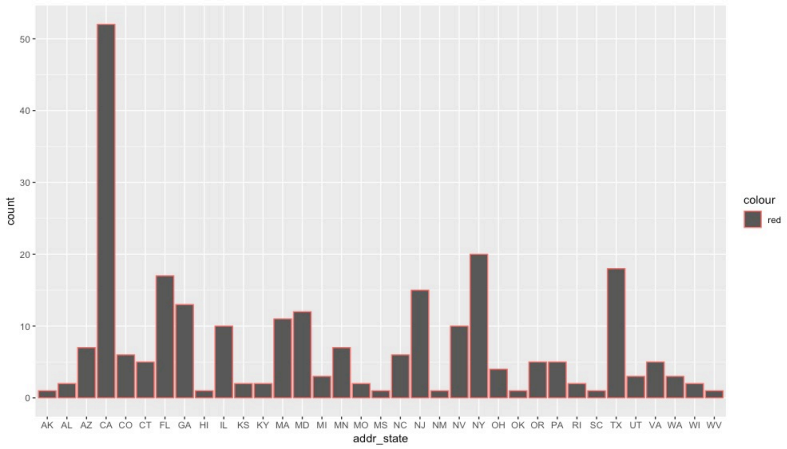
The distribution of annual income somewhat resembles a normal distribution:



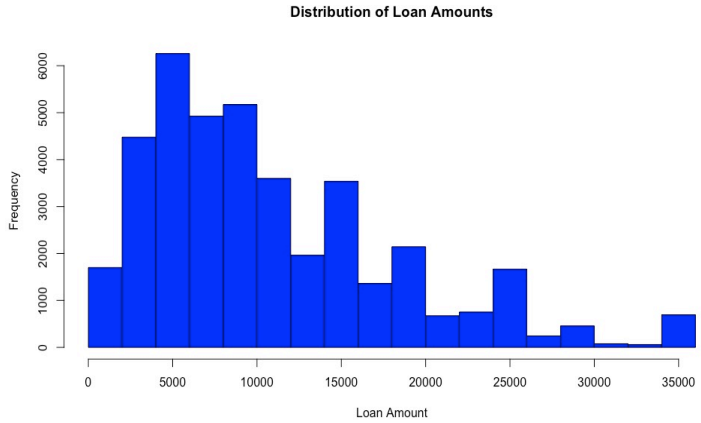
Grade Distribution:



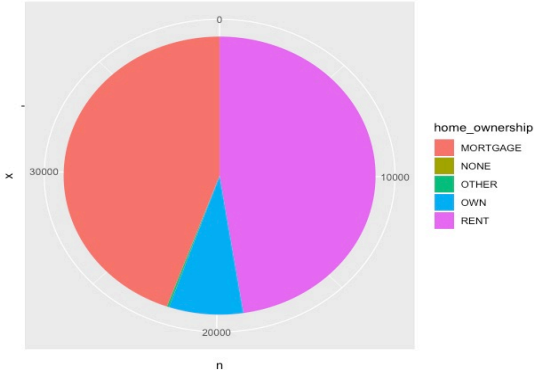
With the majority of borrowers living in California:



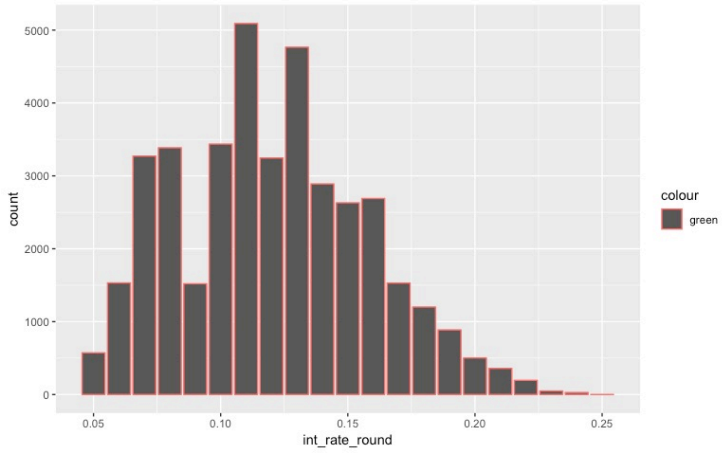
With the majority of Loans being between $5,000 and $10,000



The homeownership status is almost a 50/50 split between those that own/mortgage and rent



Distribution of interest rate is somewhat normal:



Summary of key takeaways:

Borrower's Information	Majority of Borrowers
Annual Income	$60,000
Credit Rating	B
Loan Amount	$5,000
State of Residence	CA
Credit Rating on Defaulted Loans	C

Creditworthiness Factors

Now that we have done data exploration on Lending Club's historical data and have a better understanding of borrowers financial makeup, we can now look to the future. The predictors that I rely on and will use as are:

- installment (payments made per month)
- annual income
- dti (debt to income ratio)
- interest rate
- grade (Credit worthiness on a grading scale A-G)

Using Linear Regression to Evaluate Predictors

With these variables, we can do simple linear regression to see if these variables are related to a loan being paid (vs being charged off/defaulted).

Installment amount:

```
> lm_installment<- lm(loandata$PAID ~ loandata$installment_round)
> summary(lm_installment)

Call:
lm(formula = loandata$PAID ~ loandata$installment_round)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8695  0.1341  0.1392  0.1451  0.1809

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.703e-01  3.238e-03 268.774  < 2e-16 ***
loandata$installment_round -3.942e-05  8.385e-06  -4.701  2.6e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3495 on 39784 degrees of freedom
Multiple R-squared:  0.0005552, Adjusted R-squared:  0.00053
F-statistic:  22.1 on 1 and 39784 DF,  p-value: 2.598e-06
```

Annual Income:

```
> lm_annualincome<- lm(loandata$PAID ~ loandata$annual_inc)
> summary(lm_annualincome)

Call:
lm(formula = loandata$PAID ~ loandata$annual_inc)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1961  0.1335  0.1434  0.1479  0.1572

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.419e-01  2.580e-03 326.353   <2e-16 ***
loandata$annual_inc 2.257e-07  2.746e-08   8.218   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3493 on 39784 degrees of freedom
Multiple R-squared:  0.001695,  Adjusted R-squared:  0.00167
F-statistic: 67.54 on 1 and 39784 DF,  p-value: < 2.2e-16
```

DTI:

```
> lm_dti<- lm(loandata$PAID ~ loandata$dti)
> summary(lm_dti)

Call:
lm(formula = loandata$PAID ~ loandata$dti)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8868  0.1238  0.1386  0.1517  0.1792

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8867555  0.0039064 226.998   <2e-16 ***
loandata$dti -0.0021977  0.0002622  -8.381   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3493 on 39784 degrees of freedom
Multiple R-squared:  0.001763,  Adjusted R-squared:  0.001738
F-statistic: 70.25 on 1 and 39784 DF,  p-value: < 2.2e-16
```

Interest Rate:

```
> lm_interest<- lm(loandata$PAID ~ loandata$int_rate_round)
> summary(lm_interest)

Call:
lm(formula = loandata$PAID ~ loandata$int_rate_round)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98728  0.04971  0.12368  0.17917  0.38260

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.079751   0.005811  185.80   <2e-16 *
loandata$int_rate_round -1.849417   0.046193  -40.04   <2e-16 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3427 on 39784 degrees of freedom
Multiple R-squared:  0.03873,   Adjusted R-squared:  0.03871
F-statistic:  1603 on 1 and 39784 DF,  p-value: < 2.2e-16
```

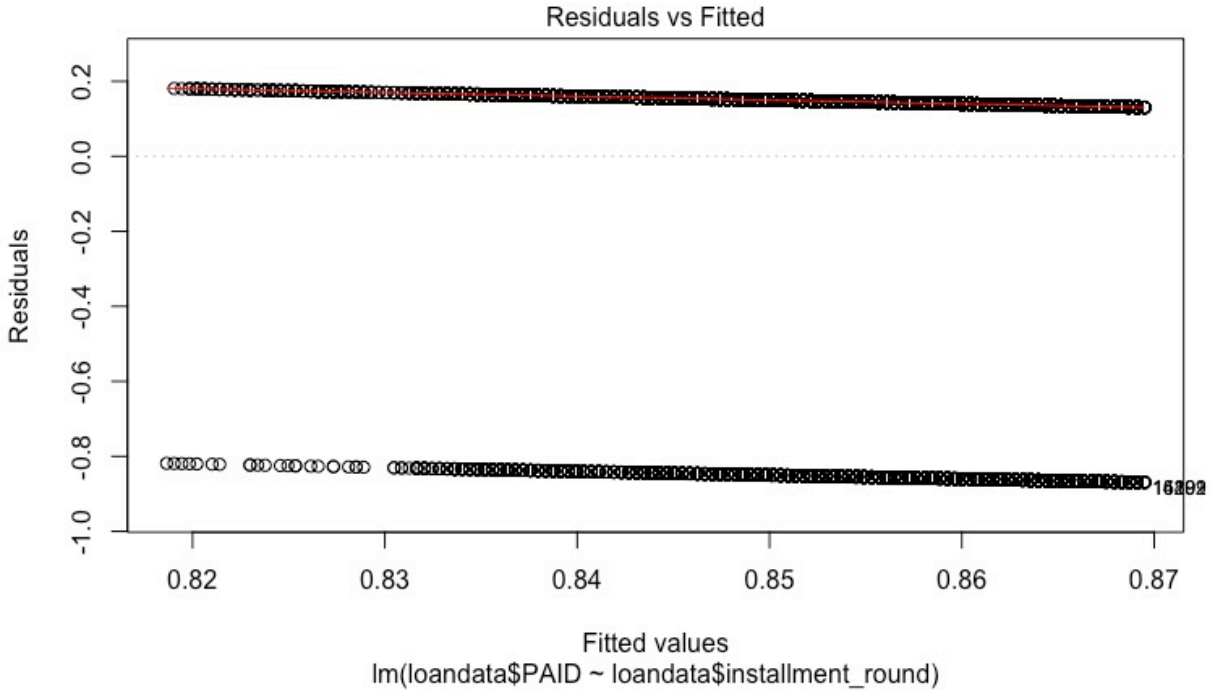
Grade:

```
> lm_grade<- lm(loandata$PAID ~ loandata$grade_numeric)
> summary(lm_grade)

Call:
lm(formula = loandata$PAID ~ loandata$grade_numeric)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93372  0.06628  0.11455  0.16281  0.35585

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.981977   0.003630  270.54   <2e-16 ***
loandata$grade_numeric -0.048261   0.001239  -38.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3431 on 39784 degrees of freedom
Multiple R-squared:  0.03673,   Adjusted R-squared:  0.03671
F-statistic:  1517 on 1 and 39784 DF,  p-value: < 2.2e-16
```

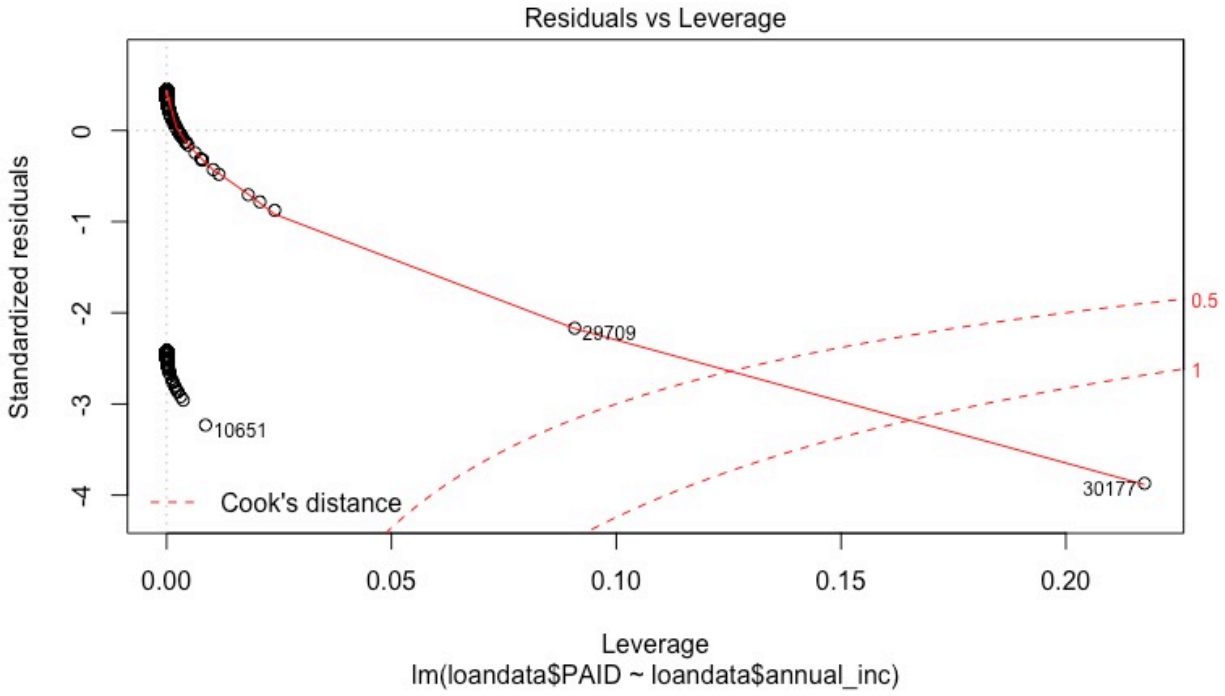
Fitting the Linear Regression Models:

After running the linear model on each of the predictors we're discovering that these variables indeed are relevant judging by the variable p-values however let's look at the fit of each model by looking at various plots:

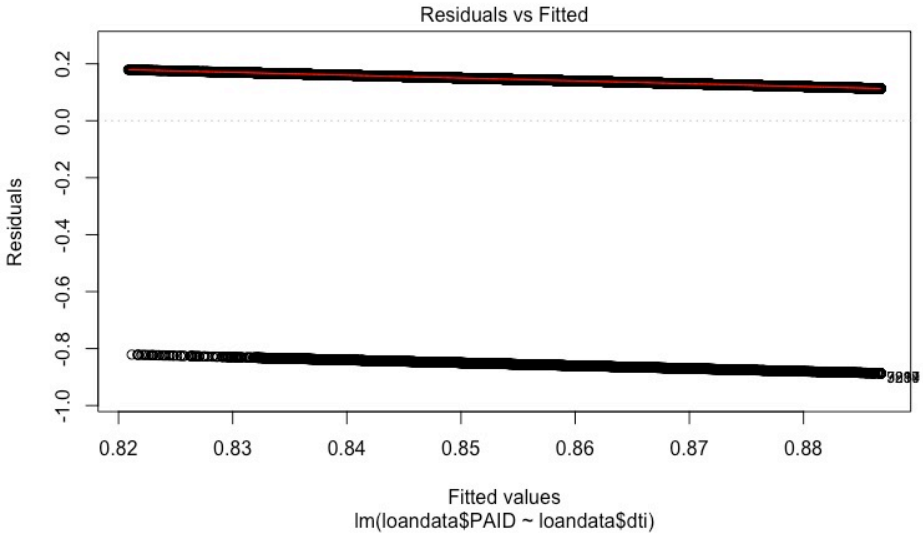
Installment:



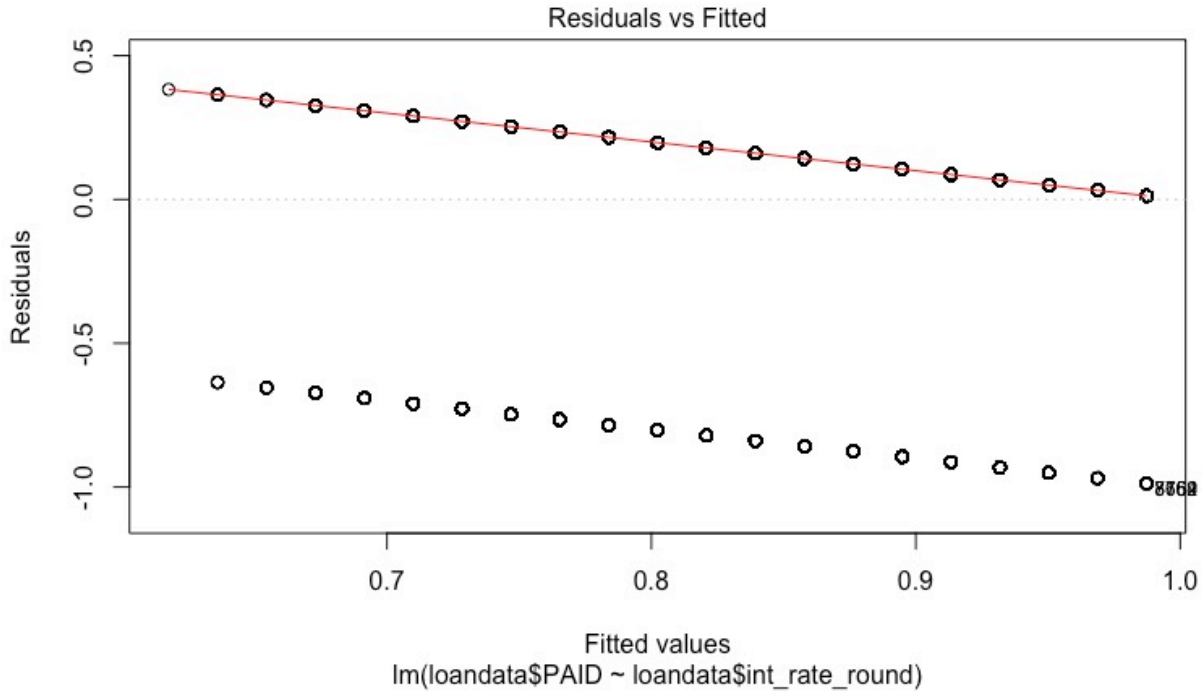
Annual Income:



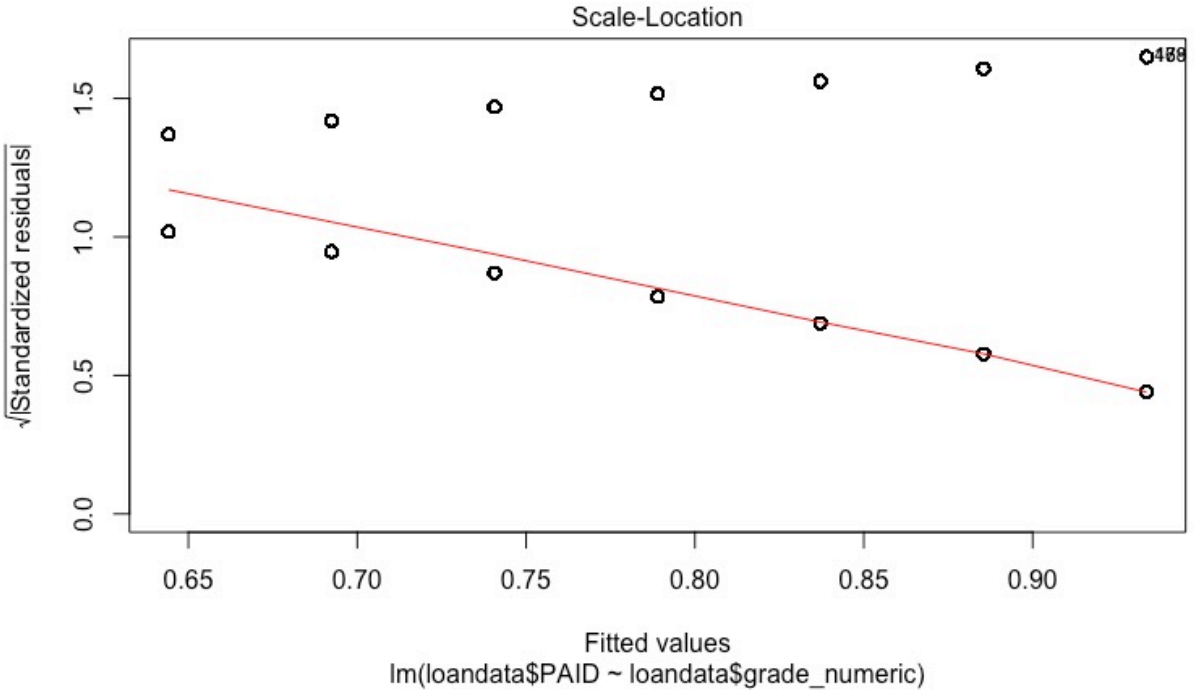
DTI:



Interest Rate:



Grade:



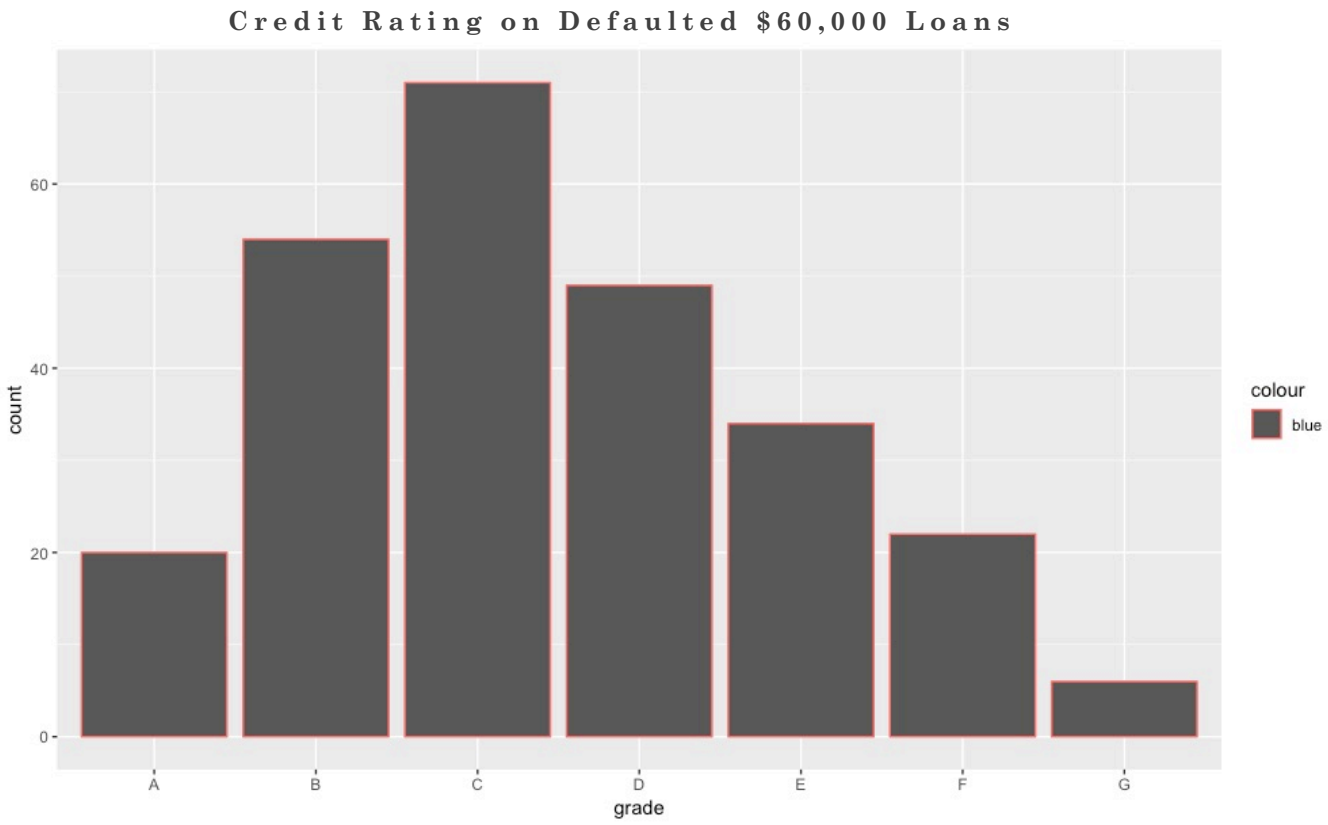
Conclusion and Reflection

It looks like, as I was afraid of, that the predictors **are** related (by looking at variable p-values) to whether or not the loan will be paid off. However after reviewing the Residual Standard Error, Multiple-R Squared and the F-Statistic, not too mention the poorly fitted diagnostic plots, I can confidently say that perhaps we can predict a borrowers financial behavior by casually evaluating their stats, however we definitely can not build a properly fitted model. Looks like I'll be going back to work tomorrow…

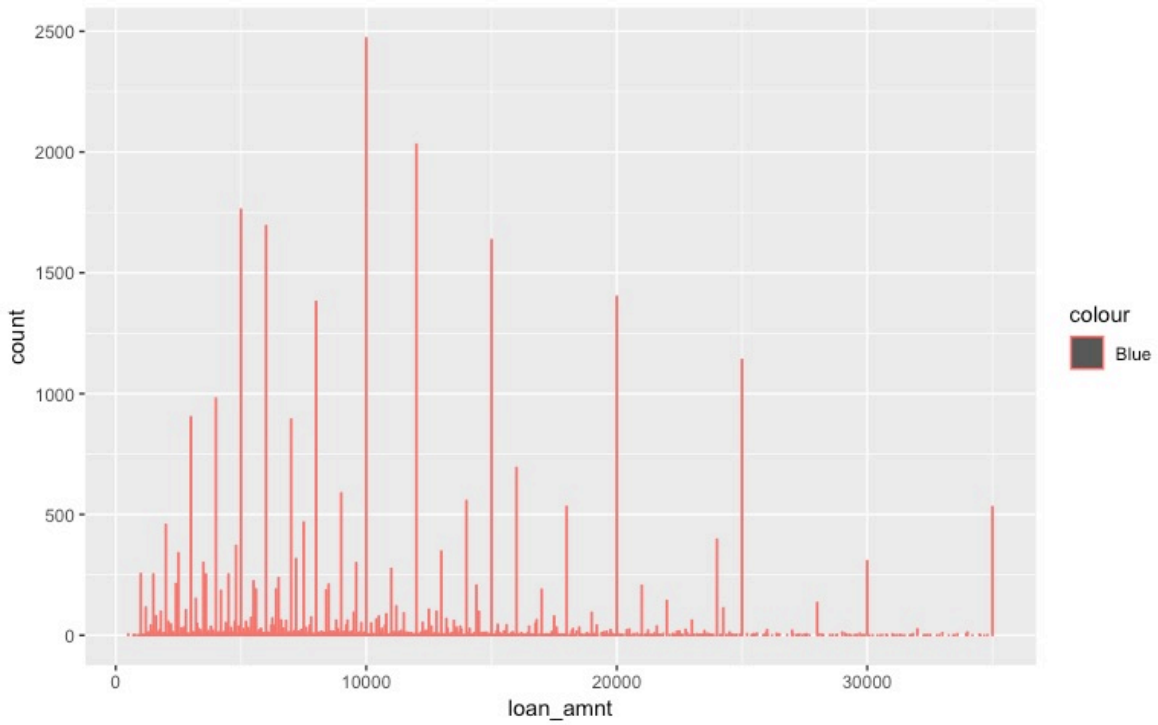
My reflections:

When it comes to predicting outcomes, predicting human behavior is probably slightly easier than predicting whether or not you're going to get into a car accident…too many outside factors. Outside factors beyond our control can negatively impact your financial status that we cannot read or predict by just looking at nominal or categorical factors (A medical emergency? An unexpected factory shut down? War?). With this in mind, I first approached this as my ticket to retirement. Enthusiastic and hopeful during the data exploration phase, I was floored to see how poorly fitted the models were. Which really made me wonder, why aren't the results what I expected? If this model was flawless, I'd quit my job in a heartbeat (along with others that are much more intelligent than I am). However, it could be that trying to predict human behavior using a machine is simply wishful thinking. Perhaps I should have stuck to modeling something with less of a 'human' factor… like how fast a different species of trees grow! But that wouldn't be nearly as fun or interesting.

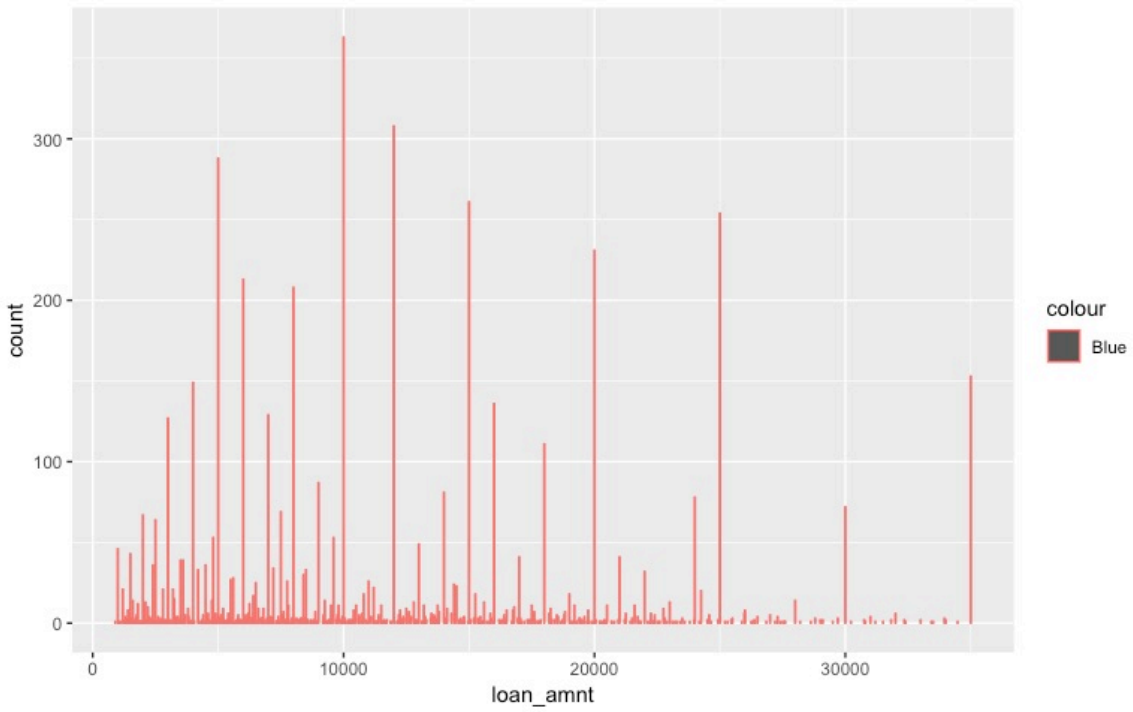
ADDITIONAL VISUALIZATIONS AND R SCRIPT



Distribution of Non Defaulted Loans



Distribution of Defaulted Loans


300
200
100
0
count
0
10000
20000
30000
loan_amnt
colour
Blue

R-Scripts I've written to help with my analysis----------

```
#purpose of this analysis is to understand what factors contribute to
defaults on loan payments and
#use those factors to predict whether or not a borrower will default

loandata<-LoanStats3a

#create binary category variable represented whether a loan has
defaulted (TRUE) or not (FALSE)
paid_column<- data.frame(PAID=(loandata$loan_status=="Fully
Paid"))
loandata<-cbind(loandata, paid_column)

#use logistic regression to predict whether a loan defaults or not
glm_loan<- glm(PAID ~ installment, family= binomial, data=loandata)
summary(glm_loan)


###########run model for multiple factors
formula<-PAID ~ installment  +   annual_inc  + dti + revol_util +

glm_test<-glm(formula, data=loandata, family="binomial")
summary(glm_test)
coefficients(glm_test)



#bring in market loans
market_loans<- primaryMarketNotes_browseNotes_1_RETAIL


loans_predict<-predict(glm_test, newdata=market_loans,
type="response")
loans_predict<-round(loans_predict,3)*100
print(loans_predict)

plot(loandata$installment, loandata$PAID)
curve(predict(glm_loan, data.frame(installment=x), type="response"),
add=TRUE)

#historical data all loan categories
hist(loandata, main="Distribution of Defaulted Loans",
xlab="grade_numeric")  #does all the data?
#loan amounts all
```

```
hist(loandata$loan_amnt, col="blue", xlab="Loan Amount",
main="Distribution of Loan Amounts")
hist(Loan_default$loan_amnt, col="yellow", main="Distribution of
Defaulted Loans by Loan Amount", xlab="Loan Amount")
options(scipen=5)
#annual income all##
hist(loandata$annual_inc, main="Distribution of Loans", xlab="Annual
Income",
    breaks=5000,xlim=c(0,100000))
#annual income by defaulted loans##
#create subset of charged off loans only
Loan_default<- loandata[which(loandata$PAID=='FALSE'),]

Loan_nondefault<- loandata[which(loandata$PAID=='TRUE'),]


hist(Loan_default$annual_inc, main="Distribution of Defaulted
Loans", xlab="Annual Income",
    breaks=5000, col= c("blue"), xlim=c(0,100000))
hist(Loan_nondefault$annual_inc, main="Distribution of Non
Defaulted Loans", xlab="Annual Income",
    breaks=5000, col= c("green"), xlim=c(0,100000))
#now that we see that $60000 annual income and lets subset even more
Loan_default_60000_income<-
Loan_default[which(Loan_default$annual_inc==60000),]
#let's see 60000 by grade distribution
#Load plyr Library
library(ggplot2)
grades_plot<-ggplot(data.frame(Loan_default_60000_income),
aes(x=grade, col="blue")) + geom_bar()
#lets see by state
state_plot<-ggplot(data.frame(Loan_default_60000_income),
aes(x=addr_state, col="red")) + geom_bar()

#grades all
grades_plot_all<-ggplot(data.frame(loandata), aes(x=PAID, col="red"))
+ geom_bar()

interest_rate_all<-ggplot(data.frame(loandata), aes(x=int_rate_round,
col="green")) + geom_bar()


#income_all
hist(loandata$annual_inc, main="Distribution of Annual Income",
xlab="Annual Income",
```

```
    breaks=5000, col= c("blue"), xlim=c(0,100000))
#now we see that Grade C defaults the most, create Grade C subset
Loan_default_C<- Loan_default[which(Loan_default$grade=="C"),]
Loan_nondefault_C<-
Loan_nondefault[which(Loan_nondefault$grade=="C"),]
#now graph loan_amt by grade C
grades_C_default<-ggplot(data.frame(Loan_default_C),
aes(x=loan_amnt, col="Blue")) + geom_bar()
grades_C_nondefault<-ggplot(data.frame(Loan_nondefault_C),
aes(x=loan_amnt, col="Blue")) + geom_bar()

ggplot(data.frame(loandata), aes(x=PAID, col="Blue"))+ geom_bar()
ggplot(data.frame(Loan_nondefault), aes(x=loan_amnt, col="Blue"))+
geom_bar()

#states all##
states_plot<-ggplot(data.frame(loandata_sample), aes(x=addr_state,
col="blue", main= "states")) + geom_bar()
#grades by defaulted loans only#
states_plot_default<-ggplot(data.frame(Loan_default),
aes(x=addr_state, col="blue", main= "states")) + geom_bar()

#piegraph for home ownership

#annual income lm test
lm_dti<- lm(loandata$PAID ~ loandata$dti)
summary(lm_dti)
predict_paid<-lm_dti$coeff[1]+ 17.1 * lm_dti$coeff[2]
#plot residuals to test fit
ggplot(data=loandata, aes(lm_dti$residuals))+
  geom_histogram(binwidth = 1, color="black", fill="purple4") +
  theme(panel.background = element_rect(fill="white"), axis.line.x =
element_line(),axis.line.y = element_line())
+ggtitle("Histogram for model residuals")


lm_interest<- lm(loandata$PAID ~ loandata$int_rate_round)
summary(lm_interest)
predict_paid<-lm_dti$coeff[1]+ 17.1 * lm_dti$coeff[2]
#plot residuals to test fit
ggplot(data=loandata, aes(lm_interest$residuals))+
  geom_histogram(binwidth = 0.05, color="black", fill="purple4") +
  theme(panel.background = element_rect(fill="white"), axis.line.x =
element_line(),axis.line.y = element_line())
+ggtitle("Histogram for model residuals")
```

```
library(dplyr)
df<-count(loandata,home_ownership)
library(ggplot2)

#pie chart

pie = ggplot(df, aes(x="", y=n, fill=home_ownership)) +
geom_bar(stat="identity", width=1)

pie = pie + coord_polar("y", start=0)


lm_installment<- lm(loandata$PAID ~ loandata$installment_round)
summary(lm_installment)
ggplot(data=loandata, aes(lm_installment$residuals))+
  geom_histogram(binwidth = .1, color="black", fill="purple4") +
  theme(panel.background = element_rect(fill="white"), axis.line.x =
element_line(),axis.line.y = element_line())
+ggtitle("Histogram for model residuals")
par(mfrow=c(2,2))
installment_plot<-plot(lm_installment)
annual_plot<-plot(lm_annualincome)

lm_annualincome<- lm(loandata$PAID ~ loandata$annual_inc)
summary(lm_annualincome)
ggplot(data=loandata, aes(lm_annualincome$residuals))+
  geom_histogram(binwidth = 1, color="black", fill="purple4") +
  theme(panel.background = element_rect(fill="white"), axis.line.x =
element_line(),axis.line.y = element_line())
+ggtitle("Histogram for model residuals")
lm_grade<- lm(loandata$PAID ~ loandata$grade_numeric)
summary(lm_grade)
plot(lm_dti)
```