Analyzing lendingclub.com Loan Data

What Kind of Customers Take Loans? For What Purpose?

John Karuitha

Friday, September 03, 2021

Contents

Background

Objectives

This data analysis project has the following objectives;

- Extract useful insights and visualize them in the most interesting way possible.
- Assess how long it takes for users to pay back their loan.
- Figure out what kind of people take a loan for what purposes.
- Discover the types of loans that have the highest or lowest risk of default.
- Build a model that can predict the probability a user will be able to pay back their loan within a certain period.

The Data

This dataset (source) consists of data from almost 10,000 borrowers that took loans - with some paid back and others still in progress. It was extracted from lendingclub.com which is an organization that connects borrowers with investors. We've included a few suggested questions at the end of this template to help you get started.

Understanding the Data

The dataset has 9578 rows and 14 variables. Table () below describes the variables.

Exploratory Data Analysis

In this section, I load the data and do exploratory data analysis. I start with data visualisation followed by summary statistics aimed at uncovering the insights stated in the objectives section above.

```
str(loans)
```

```
## spec_tbl_df [9,578 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ credit_policy    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ purpose          : Factor w/ 7 levels "all_other","credit_card",..: 3 2 3 3 2 2 3 1 5 3 ...
##  $ int_rate         : num [1:9578] 0.119 0.107 0.136 0.101 0.143 ...
##  $ installment      : num [1:9578] 829 228 367 162 103 ...
##  $ log_annual_inc   : num [1:9578] 11.4 11.1 10.4 11.4 11.3 ...
##  $ dti              : num [1:9578] 19.5 14.3 11.6 8.1 15 ...
##  $ fico             : num [1:9578] 737 707 682 712 667 727 667 722 682 707 ...
##  $ days_with_cr_line: num [1:9578] 5640 2760 4710 2700 4066 ...
##  $ revol_bal        : num [1:9578] 28854 33623 3511 33667 4740 ...
##  $ revol_util       : num [1:9578] 52.1 76.7 25.6 73.2 39.5 51 76.8 68.6 51.1 23 ...
##  $ inq_last_6mths   : num [1:9578] 0 0 1 1 0 0 0 0 1 1 ...
##  $ delinq_2yrs      : num [1:9578] 0 0 0 0 1 0 0 0 0 0 ...
##  $ pub_rec          : num [1:9578] 0 0 0 0 0 0 1 0 0 0 ...
##  $ not_fully_paid   : Factor w/ 2 levels "Fully Paid","Not Fully paid": 1 1 1 1 1 1 2 2 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   credit_policy = col_double(),
##   ..   purpose = col_character(),
##   ..   int_rate = col_double(),
##   ..   installment = col_double(),
##   ..   log_annual_inc = col_double(),
##   ..   dti = col_double(),
##   ..   fico = col_double(),
##   ..   days_with_cr_line = col_double(),
##   ..   revol_bal = col_double(),
##   ..   revol_util = col_double(),
##   ..   inq_last_6mths = col_double(),
##   ..   delinq_2yrs = col_double(),
##   ..   pub_rec = col_double(),
##   ..   not_fully_paid = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data Visualization

The Correlation Matrix

Figure () below shows the correlation matrix of the numeric explanatory variables and the response variable. I will refer to this correlation matrix in the analysis that follows.

What type of customers does the lender perceived to be risky? Are they as risky?

In this section, I use data vislualisation techniques assess the types of customers that the lender percieves to be riskiest and whether this perception is supported by the data. The interest rate is proxies the risk

Table 1: Variables Description

Variable	Class	Description
Credit_policy	Numeric	1 if the customer meets the credit underwriting criteria, and 0 otherwise.
Purpose	Character	The purpose of the loan. One of: small business, major purchase, home improvement, educational, debt_consolidation, credit card, and all other types lumped together
Int_rate (Interest Rate)	Numeric	The interest rate of the loan (more risky borrowers are assigned higher interest rates
Installment	Numeric	The monthly installments owed by the borrower if the loan is funded.
log_annual_inc	Numeric	The natural log of the self-reported annual income of the borrower.
dti	Numeric	The debt-to-income ratio of the borrower (amount of debt divided by annual income).
fico	Numeric	The FICO credit score of the borrower.
days_with_cr_line	Numeric	The number of days the borrower has had a credit line.
revol_bal	Numeric	The borrower's revolving balance; the amount unpaid at the end of the credit card billing cycle.
revol_util	Numeric	The borrower's revolving line utilization rate; the amount of the credit line used relative to total credit available.
inq_last_6mths	Numeric	The borrower's number of inquiries by creditors in the last 6 months.
delinq_2yrs	Numeric	The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
pub_rec	Numeric	The borrower's number of derogatory public records.
not_fully_paid	Numeric	1 if the loan is not fully paid; 0 otherwise.

assessment of the customer by the lender.

The correlation matrix in table () shows a high correlation between fico score and interest rates charged on loans (-0.71). Given that the interest rate has a very high correlation with fico score, it appears like the lender, to a large extent relies on fico scores to assess credit risk. Also, fico score has a high negative correspondence with revol_util(the amount of credit utilised relative to the total credit available) at -0.541. This observation means that individuals with high fico scores tend to utilize their lines of credit less than individuals with a low fico score. However, there are a few outliers as figure () shows. The outliers are individuals with high fico scores who also a high loan utilisation rate. These customers could be targeted with relationship banking to offer personalised services.

However, there is a high positive correlation between revol_util and interest rates. This observation implies that the lender views people with high loan utilisation as presenting a greater risk of default. However, as we have seen, people with high fico scores present low default risk and yet have high utilisation rate.

> Take away 1: Most people with a high fico score present have a low rate of loan utilisation. However, there are outliers representing customers with high fico scores and high loan utilisation. This segment of customers has a high return to risk ratio and could be targeted with personalised services.

Figure () supports this observation, showing that high fico scores correspond to lower loan delinquency.

> Take away 2: Segment customers with high loan utilisation rate using the fico scores. Charge a lower rate of interest to those customers with high fico scores and retain higher interest rates to customers with lower scores. In my assessment, the lender should use a fico cut off rate of 750 and a loan utilisation cap of 30%. The institution could then focus on customers who exceeed these two thresholds for targeted marketing.

```
loans %>% GGally::ggpairs(columns = c("int_rate", "installment",

                          "log_annual_inc",

                          "dti", "fico", "days_with_cr_line",

                          "revol_bal", "revol_util",

                          "inq_last_6mths", "delinq_2yrs",

                          "pub_rec", "not_fully_paid"))
```

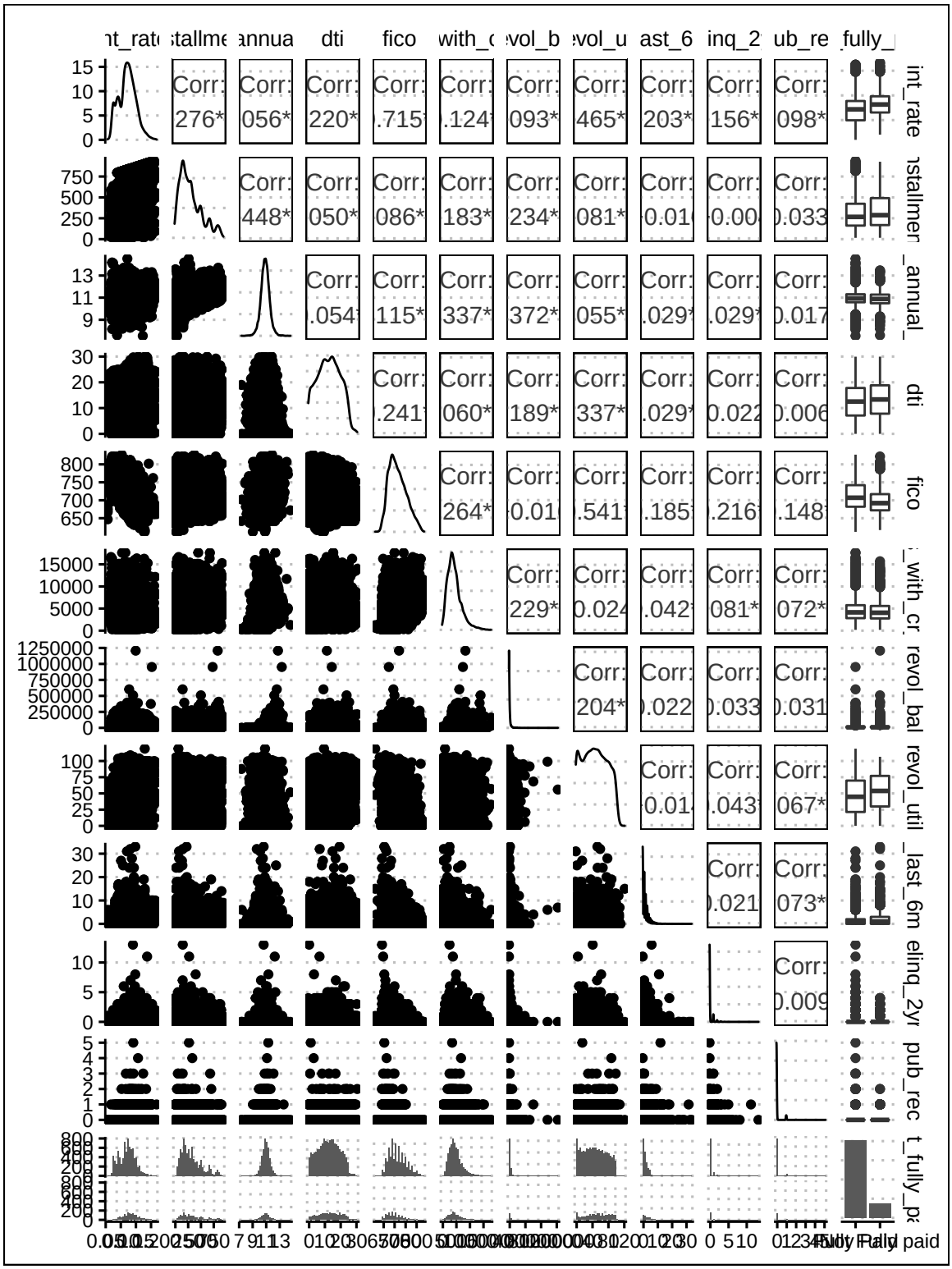
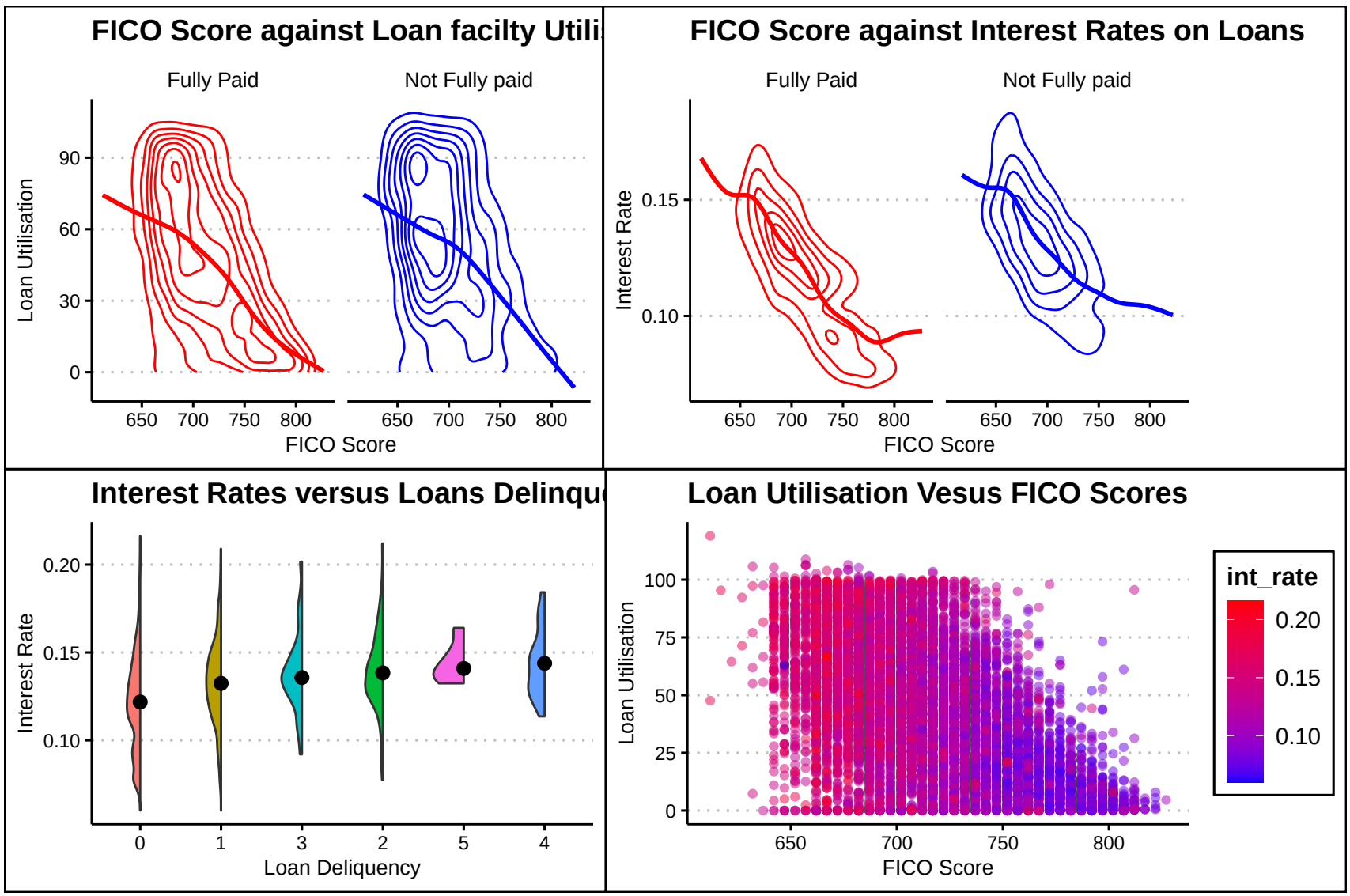


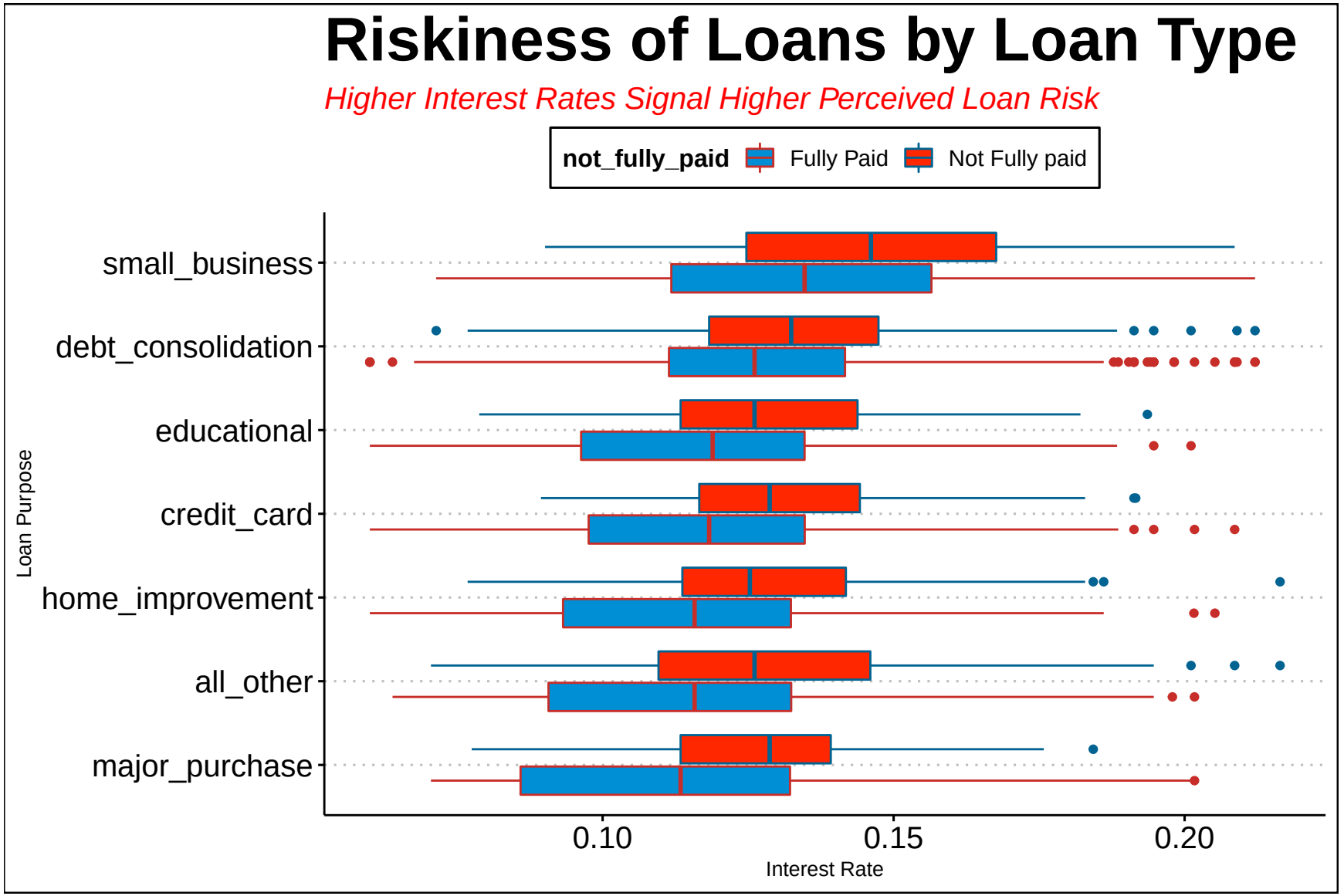
Figure 1: Correlation Matrix for Numeric Predictors

```
(loans %>%
  ggplot(mapping = aes(x = fico, y = revol_util, color = not_fully_paid)) +
  geom_density2d() +
  geom_smooth(se = FALSE) +
  scale_color_manual(values = c("red", "blue")) +
  labs(x = "FICO Score", y = "Loan Utilisation",
       title = "FICO Score against Loan facilty Utilisation") +
    theme(legend.position = "none") +
    facet_wrap(~ not_fully_paid) +
loans %>%
  ggplot(mapping = aes(x = fico, y = int_rate, color = not_fully_paid)) +
  geom_density2d() +
  geom_smooth(se = FALSE) +
  scale_color_manual(values = c("red", "blue")) +
  labs(x = "FICO Score", y = "Interest Rate",
       title = "FICO Score against Interest Rates on Loans") +
  facet_wrap(~ not_fully_paid) +
    theme(legend.position = "none")) /

(loans %>%
```

```
  group_by(factor(delinq_2yrs)) %>%
  filter(n() > 2) %>%
  ungroup() %>%
  ggplot(mapping = aes(x = fct_reorder(factor(delinq_2yrs), int_rate, median),
                       y = int_rate, fill = factor(delinq_2yrs))) +
  geom_half_violin(varwidth = TRUE, show.legend = FALSE) +
  stat_summary(fun = median, show.legend = FALSE) +
  labs(x = "Loan Deliquency", y = "Interest Rate",
       title = "Interest Rates versus Loans Delinquency") +


loans %>%
  ggplot(mapping = aes(x = fico, y = revol_util, color = int_rate)) +
  geom_point(alpha = 0.5) +
  labs(x = "FICO Score", y = "Loan Utilisation",
       title = "Loan Utilisation Vesus FICO Scores") +
  scale_color_gradient(low = "blue", high = "red"))
```


FICO Score against Loan facilty Utili
Fully Paid
Not Fully paid
Loan Utilisation
90
60
30
0
650
700
750
800
FICO Score
FICO Score against Interest Rates on Loans
Fully Paid
Not Fully paid
Interest Rate
0.15
0.10
650
700
750
800
FICO Score
Interest Rates versus Loans Delinqu
Interest Rate
0.20
0.15
0.10
0
1
3
2
5
4
Loan Deliquency
Loan Utilisation Vesus FICO Scores
Loan Utilisation
100
75
50
25
0
650
700
750
800
FICO Score
int_rate
0.20
0.15
0.10

```
loans %>% ggplot(mapping = aes(x = fct_reorder(purpose, int_rate, median),
                               y = int_rate, fill = not_fully_paid,
                               color = not_fully_paid)) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_fivethirtyeight() +
  scale_color_wsj() +
  labs(x = "Loan Purpose", y = "Interest Rate",
       title = "Riskiness of Loans by Loan Type",
       subtitle = "Higher Interest Rates Signal Higher Perceived Loan Risk") +
  theme(legend.position = 'top',
        plot.title = element_text(size = 30, face = "bold", colour = "Black"),
        plot.subtitle = element_text(size = 15, face = "italic", colour = "red"),
        axis.text = element_text(size = 15))
```


Riskiness of Loans by Loan Type
Higher Interest Rates Signal Higher Perceived Loan Risk
not_fully_paid
Fully Paid
Not Fully paid
small_business
debt_consolidation
educational
credit_card
home_improvement
all_other
major_purchase
Loan Purpose
0.10
0.15
0.20
Interest Rate

Summarising the Data

Basic Modelling

Conclusion

My analysis will consider these customers that have a high fico credit score but a high loan utilization rate. As noted above fico scores have a negative correlation with loan utilisation at -0.5413. Specifically, I examine the following issues.

- Do high fico-high loan utilisation clients pose substantial credit default risk?
- What types of loans do high fico-high loan utilisation clients demand?

To get going, I define define a high fico customer as having at least a score of 700. Customers with a loan utlisation rate over 30% are in the high loan utilisation category.

```
loans %>%

  filter(fico > 700 & revol_util > 30) %>%

  summarise(default = mean(delinq_2yrs))
```

```
## # A tibble: 1 x 1
##   default
##     <dbl>
## 1  0.0552
```

```
loans %>%

  filter(fico > 700) %>%

  summarise(default = mean(delinq_2yrs))
```

```
## # A tibble: 1 x 1
##   default
##     <dbl>
## 1  0.0674
```