Lending Club Loan Data by Malik Malik

Data Investigation and Wrangling

Let's take a look at the Lending Club dataset from Kaggle, which includes all Lending Club loans from 2007 to 2015.

The Lending Club loan data can be found here.

The question we want to answer is: Are there certain factors that affect chances of defaulting?

First, let's take a look at the dimensions of the dataset.

```
## [1] 887379     74
```

This is clearly a very large dataset, with 887,379 rows and 74 variables. I wonder if we could trim this dataset down?

Let's look at the 74 variables.

```
##  [1] "id"                          "member_id"
##  [3] "loan_amnt"                   "funded_amnt"
##  [5] "funded_amnt_inv"             "term"
##  [7] "int_rate"                    "installment"
##  [9] "grade"                       "sub_grade"
## [11] "emp_title"                   "emp_length"
## [13] "home_ownership"              "annual_inc"
## [15] "verification_status"         "issue_d"
## [17] "loan_status"                 "pymnt_plan"
## [19] "url"                         "desc"
## [21] "purpose"                     "title"
## [23] "zip_code"                    "addr_state"
## [25] "dti"                         "delinq_2yrs"
## [27] "earliest_cr_line"            "inq_last_6mths"
## [29] "mths_since_last_delinq"      "mths_since_last_record"
## [31] "open_acc"                    "pub_rec"
## [33] "revol_bal"                   "revol_util"
## [35] "total_acc"                   "initial_list_status"
## [37] "out_prncp"                   "out_prncp_inv"
## [39] "total_pymnt"                 "total_pymnt_inv"
## [41] "total_rec_prncp"             "total_rec_int"
## [43] "total_rec_late_fee"          "recoveries"
## [45] "collection_recovery_fee"     "last_pymnt_d"
## [47] "last_pymnt_amnt"             "next_pymnt_d"
## [49] "last_credit_pull_d"          "collections_12_mths_ex_med"
## [51] "mths_since_last_major_derog" "policy_code"
## [53] "application_type"            "annual_inc_joint"
## [55] "dti_joint"                   "verification_status_joint"
## [57] "acc_now_delinq"              "tot_coll_amt"
## [59] "tot_cur_bal"                 "open_acc_6m"
## [61] "open_il_6m"                  "open_il_12m"
## [63] "open_il_24m"                 "mths_since_rcnt_il"
## [65] "total_bal_il"                "il_util"
## [67] "open_rv_12m"                 "open_rv_24m"
## [69] "max_bal_bc"                  "all_util"
## [71] "total_rev_hi_lim"            "inq_fi"
```

```
## [73] "total_cu_tl"                  "inq_last_12m"
```

Alright, so we have a variable named called loan_status. This should tell us the status of each loan. Let's see how many of each loan status type there are, and if we can possibly cut down this large dataset based on what we see here.

```
##
##                                          Charged Off
##                                                45248
##                                              Current
##                                               601779
##                                              Default
##                                                 1219
## Does not meet the credit policy. Status:Charged Off
##                                                  761
##  Does not meet the credit policy. Status:Fully Paid
##                                                 1988
##                                           Fully Paid
##                                               207723
##                                      In Grace Period
##                                                 6253
##                                               Issued
##                                                 8460
##                                    Late (16-30 days)
##                                                 2357
##                                   Late (31-120 days)
##                                                11591
```

After taking a look at Lending Club's definitions for these loan statuses, we can break down the loans in this dataset to three groups:

- Loans that are no longer being paid for:
- Fully Paid: A loan has been fully repaid
- Default: A loan that is 121+ days past due
- Charged Off: A loan that is 150+ days past due and there is no reasonable expectation of further payments
- Loans that are still expected to be paid for:
- Current: A loan is up to date on all outstanding payments
- In Grace Period: Loan is past due but within the 15-day grace period
- Late (16-30 days): Loan is 16-30 days past due
- Late (31-120 days): Loan is 31-120 days past due
- Loans that did not meet the credit policy:
- Does not meet the credit policy. Status: Fully Paid
- Does not meet the credit policy. Status:Charged Off

For my analysis, I'm hoping to get a better understanding of what factors indicate a loan is more likely to default than others. As a result, I will only look at past loans.

Present loans are still in the process of payment, and I feel will not help me reach the most accurate conclusions. I also feel loans that don't meet credit policy won't benefit my analysis.

Let's subset the data so it only includes past loans, and then look at the dimensions.

```
##
##                                          Charged Off
##                                                45248
##                                              Current
##                                                    0
##                                              Default
##                                                 1219
## Does not meet the credit policy. Status:Charged Off
##                                                    0
##  Does not meet the credit policy. Status:Fully Paid
##                                                    0
##                                           Fully Paid
##                                               207723
##                                      In Grace Period
##                                                    0
##                                               Issued
##                                                    0
##                                    Late (16-30 days)
##                                                    0
##                                   Late (31-120 days)
##                                                    0
```

```
## [1] 254190     74
```

We can see here that we've successfully filtered the dataset to only include Fully Paid, Default, and Charged Off loans. The dataset has dropped from 887,379 rows to 254,190. We still have too many variables, though.

After looking at the variable dictionary that came alongside this dataset from Kaggle, I've decided to ignore all but 13 variables of the dataset. These 11 variables are the only ones I feel are required if I wish to engage in insightful analysis of this dataset.

The following are those 13 variables and their definitions:

- addr_state: The state provided by the borrower in the loan application
- annual_inc: Borrower's annual income.
- dti (Debt to Income Ratio): A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- emp_length: Employment length in years. 0 = less than a year.
- grade: LC assigned loan grade
- home_ownership: The home ownership status.
- installment: The monthly payment owed by the borrower if the loan originates.
- int_rate: Interest Rate on the loan.
- issue_d (Issue Date): The month which the loan was funded
- loan_amnt: The listed amount of the loan applied for by the borrower.
- loan_status: Current status of the loan
- purpose: A category provided by the borrower for the loan request.
- term: The number of payments on the loan. Values are in months and can be either 36 or 60.

Let's subset the dataset to only have those variables, and then look at the dimensions again.

```
## [1] 254190     13
```

We've now trimmed down the dataset to 13 variables.

One more thing. Before starting our analysis via plot generation, let's create a column, titled default_status, that will break down the past loans into two categories depending on their loan status:

- Default: Loans labeled as "default" or "charged off".

- Paid: Loans that were labeled as "Fully Paid".

By separating our loans this way, we'll be able to use plotting and other tools to see what variables are correlated with loans that go unpaid. Let's do that, and then look at the new dimensions.

```
# Group of paid loan statuses
default_group <- c("Default", "Charged Off")

# Creating the new variable
loans.2$default_status <- ifelse(loans.2$loan_status %in% default_group, "Default", "Paid")
```

```
## [1] 254190     14
```

We now have 14 variables, but take a quick look at our data to make sure this worked correctly.

```
##   addr_state annual_inc   dti emp_length grade home_ownership installment
## 1         AZ      24000 27.65  10+ years     B           RENT      162.87
## 2         GA      30000  1.00   < 1 year     C           RENT       59.83
## 3         IL      12252  8.72  10+ years     C           RENT       84.33
## 4         CA      49200 20.00  10+ years     C           RENT      339.31
## 5         AZ      36000 11.20    3 years     A           RENT      156.46
## 6         CA      48000  5.35    9 years     E           RENT      109.43
##   int_rate  issue_d loan_amnt loan_status        purpose       term
## 1    10.65 Dec-2011      5000  Fully Paid    credit_card  36 months
## 2    15.27 Dec-2011      2500 Charged Off            car  60 months
## 3    15.96 Dec-2011      2400  Fully Paid small_business  36 months
## 4    13.49 Dec-2011     10000  Fully Paid          other  36 months
## 5     7.90 Dec-2011      5000  Fully Paid        wedding  36 months
## 6    18.64 Dec-2011      3000  Fully Paid            car  36 months
##   default_status
## 1           Paid
## 2        Default
## 3           Paid
## 4           Paid
## 5           Paid
## 6           Paid
```

Okay, and let's see how many "Paid" and "Default" loans there are:

```
##
## Default    Paid
##   46467  207723
```

Let's take a look at our structure and summary before starting some exploratory univariate analysis

```
## 'data.frame':    254190 obs. of  14 variables:
##  $ addr_state    : Factor w/ 51 levels "AK","AL","AR",..: 4 11 15 5 4 5 5 44 4 5 ...
##  $ annual_inc    : num  24000 30000 12252 49200 36000 ...
##  $ dti           : num  27.65 1 8.72 20 11.2 ...
##  $ emp_length    : Factor w/ 12 levels "< 1 year","1 year",..: 3 1 3 3 5 11 6 1 7 3 ...
##  $ grade         : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 1 5 6 2 3 2 ...
##  $ home_ownership: Factor w/ 6 levels "ANY","MORTGAGE",..: 6 6 6 6 6 6 5 6 5 5 ...
##  $ installment   : num  162.9 59.8 84.3 339.3 156.5 ...
##  $ int_rate      : num  10.6 15.3 16 13.5 7.9 ...
##  $ issue_d       : Factor w/ 103 levels "Apr-2008","Apr-2009",..: 22 22 22 22 22 22 22 22 22 22 ...
##  $ loan_amnt     : num  5000 2500 2400 10000 5000 ...
##  $ loan_status   : Factor w/ 3 levels "Charged Off",..: 3 1 3 3 3 3 1 1 3 3 ...
##  $ purpose       : Factor w/ 14 levels "car","credit_card",..: 2 1 12 10 14 1 12 10 3 3 ...
```

```
##  $ term          : Factor w/ 2 levels " 36 months"," 60 months": 1 2 1 1 1 1 2 2 2 1 ...
##  $ default_status: chr  "Paid" "Default" "Paid" "Paid" ...
```

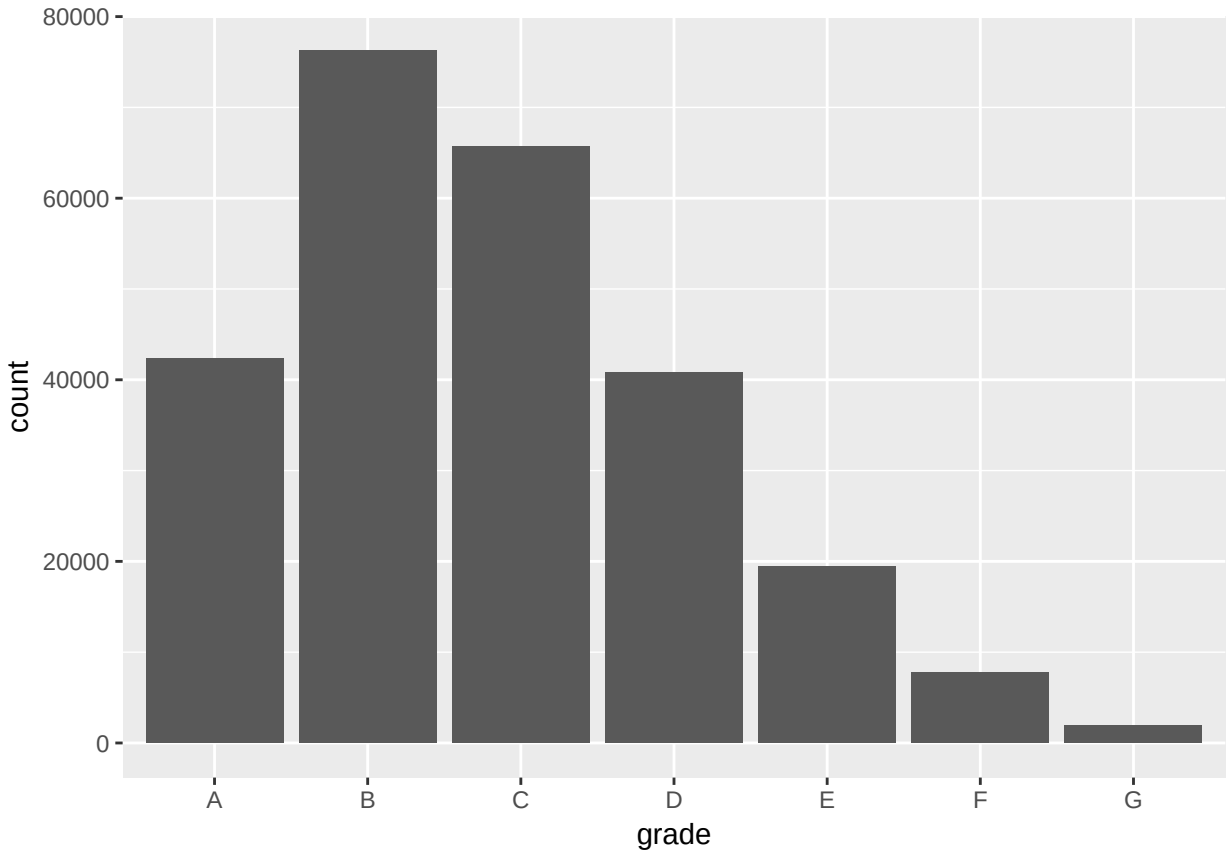
```
##    addr_state       annual_inc           dti            emp_length
##  CA     : 43321   Min.   :   3000   Min.   : 0.00   10+ years:77256
##  NY     : 21444   1st Qu.:  45000   1st Qu.:10.77   2 years  :23647
##  TX     : 19454   Median :  62000   Median :16.22   < 1 year :20975
##  FL     : 17640   Mean   :  72511   Mean   :16.56   3 years  :20484
##  NJ     :  9650   3rd Qu.:  87000   3rd Qu.:22.01   5 years  :18136
##  IL     :  9281   Max.   :8706582   Max.   :57.14   1 year   :16951
##  (Other):133400                                     (Other)  :76741
##  grade      home_ownership    installment         int_rate
##  A:42343   ANY     :     1   Min.   :  15.69   Min.   : 5.32
##  B:76263   MORTGAGE:125342   1st Qu.: 239.56   1st Qu.:10.74
##  C:65680   NONE    :    43   Median : 365.23   Median :13.53
##  D:40818   OTHER   :   141   Mean   : 418.27   Mean   :13.78
##  E:19387   OWN     : 22095   3rd Qu.: 547.55   3rd Qu.:16.55
##  F: 7739   RENT    :106568   Max.   :1424.57   Max.   :28.99
##  G: 1960
##      issue_d         loan_amnt          loan_status
##  Oct-2014:  8808   Min.   :  500   Charged Off: 45248
##  Jul-2014:  8614   1st Qu.: 7250   Default    :  1219
##  Apr-2014:  6744   Median :12000   Fully Paid :207723
##  Nov-2013:  6665   Mean   :13571
##  Oct-2013:  6653   3rd Qu.:18250
##  Dec-2013:  6644   Max.   :35000
##  (Other) :210062
##                purpose               term        default_status
##  debt_consolidation:149153    36 months:197373   Length:254190
##  credit_card       : 50309    60 months: 56817   Class :character
##  home_improvement  : 14976                       Mode  :character
##  other             : 14342
##  major_purchase    :  6279
##  small_business    :  4765
##  (Other)           : 14366
```

Univariate Plots

Let's begin our analysis by generating bar plots and histograms for our various variables.

Bar plots of categorical values

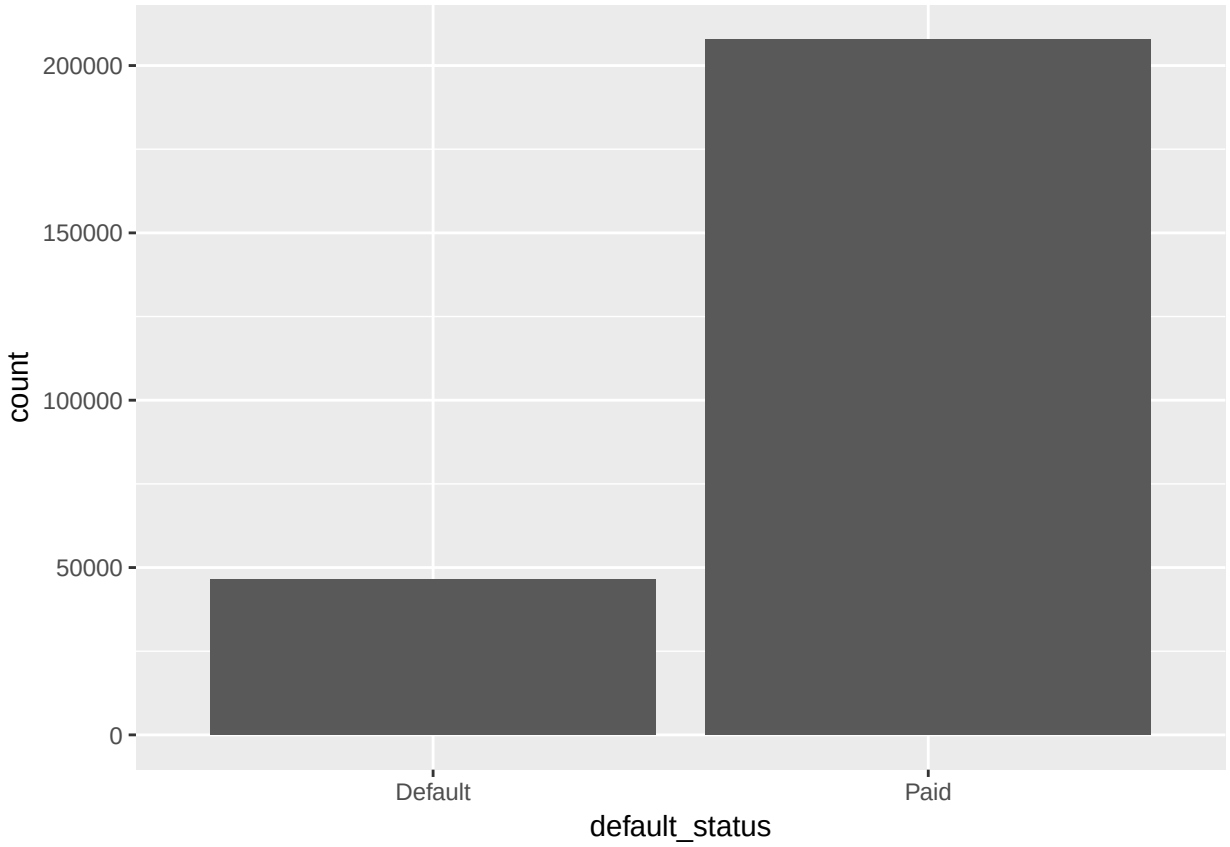
Grades




```
##     A     B     C     D     E     F     G
## 42343 76263 65680 40818 19387  7739  1960
```

This positively skewed distribution has most of its data from grades A to D.

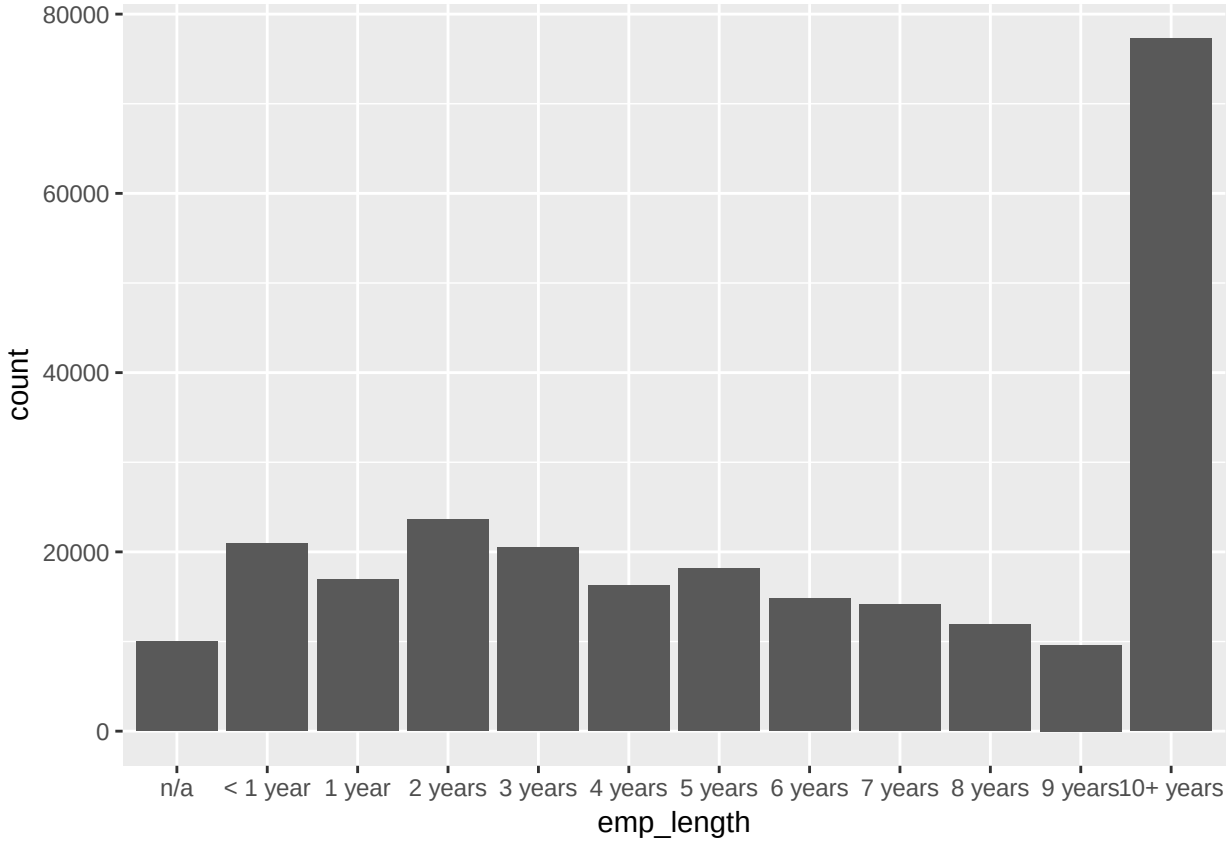
Paid Status




```
##
## Default    Paid
##   46467  207723
```

There are 207,723 loans that have been paid off, while 46,467 have either defaulted or been charged off.

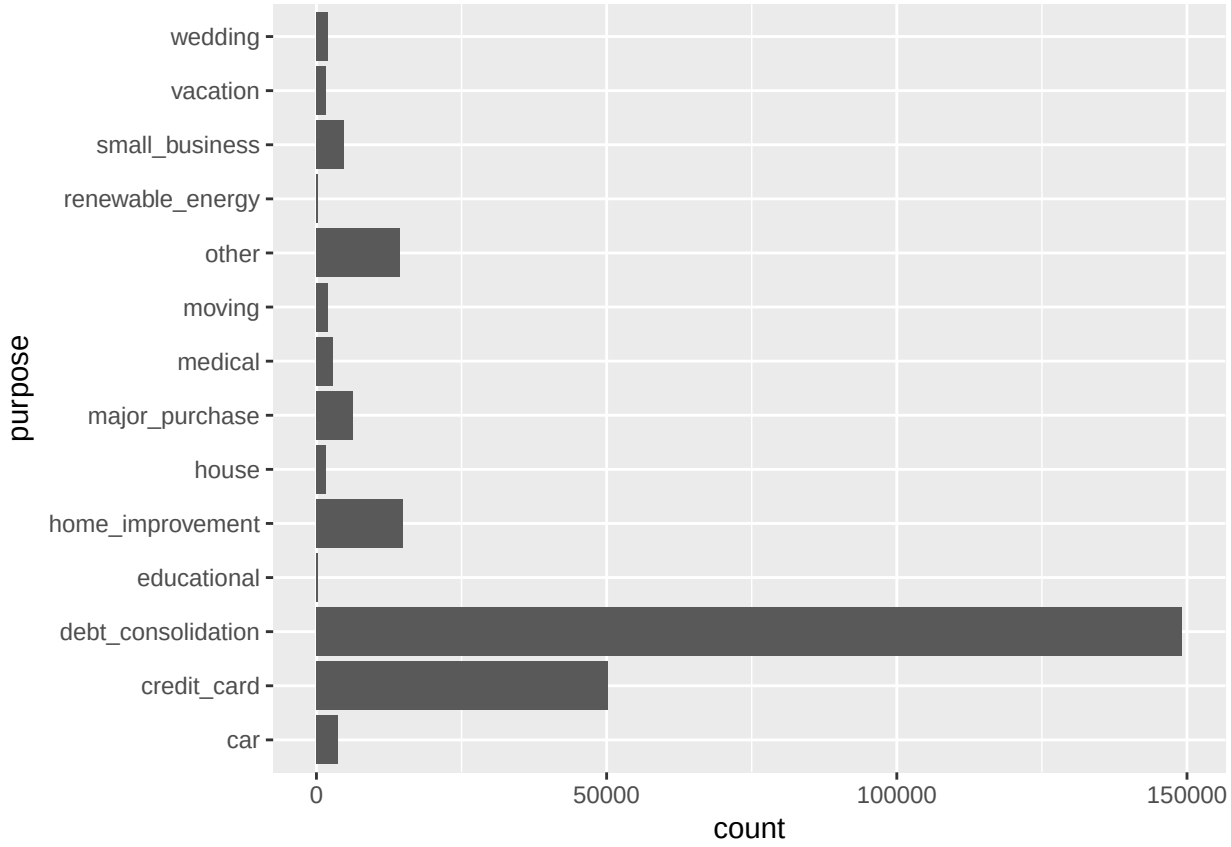
Employee Length




```
##       n/a  < 1 year    1 year   2 years   3 years   4 years   5 years
##      9968     20975     16951     23647     20484     16263     18136
##   6 years   7 years   8 years   9 years 10+ years
##     14816     14156     11922      9616     77256
```

Borrowers who have been employed for 10+ years are by far the most common. The rest are similar in occurrence. Also, it should be noted that "n/a" in this case likely means the borrower had no employment history at the time of applying for the loan.

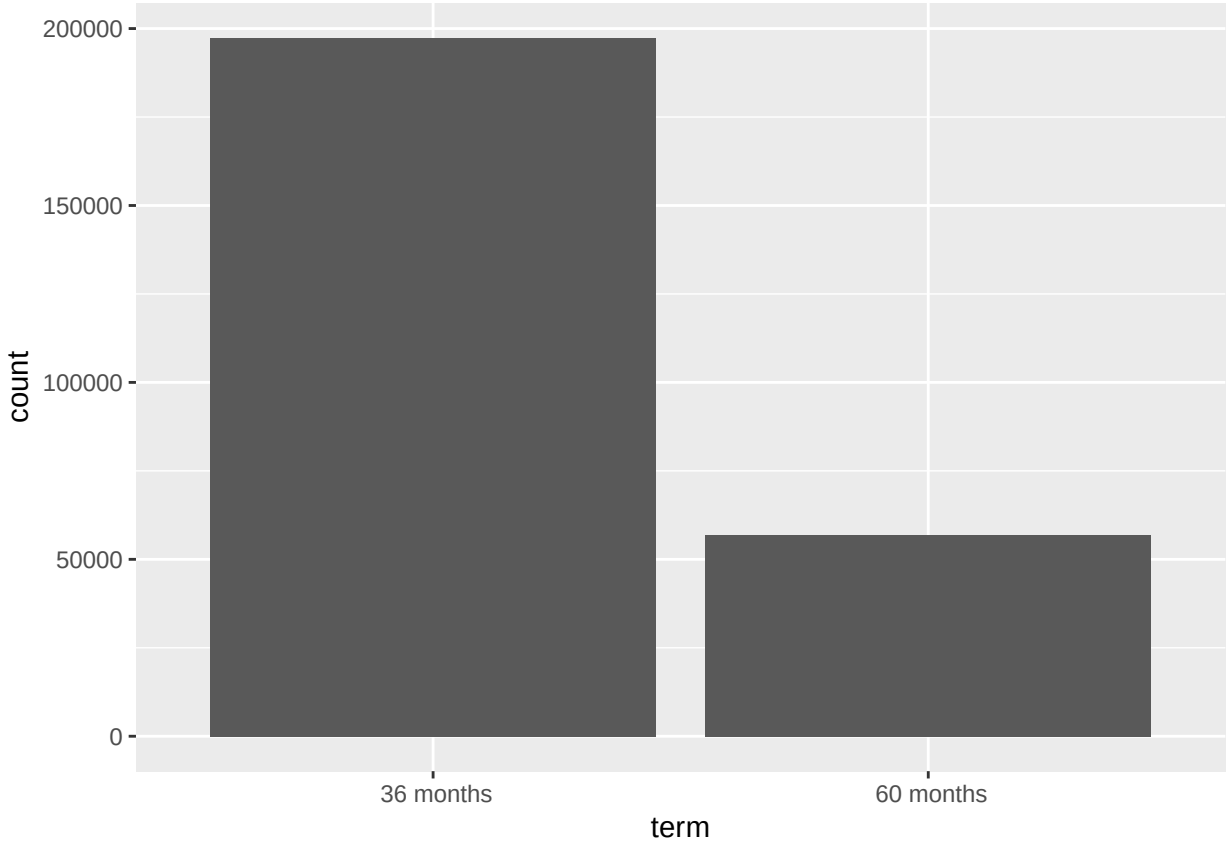
Purpose for loan




```
##
##                car        credit_card debt_consolidation
##               3656              50309             149153
##        educational   home_improvement              house
##                325              14976               1659
##     major_purchase            medical              moving
##               6279               2869               2039
##              other   renewable_energy     small_business
##              14342                267               4765
##           vacation            wedding
##               1596               1955
```

Debt Consolidation and Credit Card are the two biggest reasons for applying for Lending Club loans.

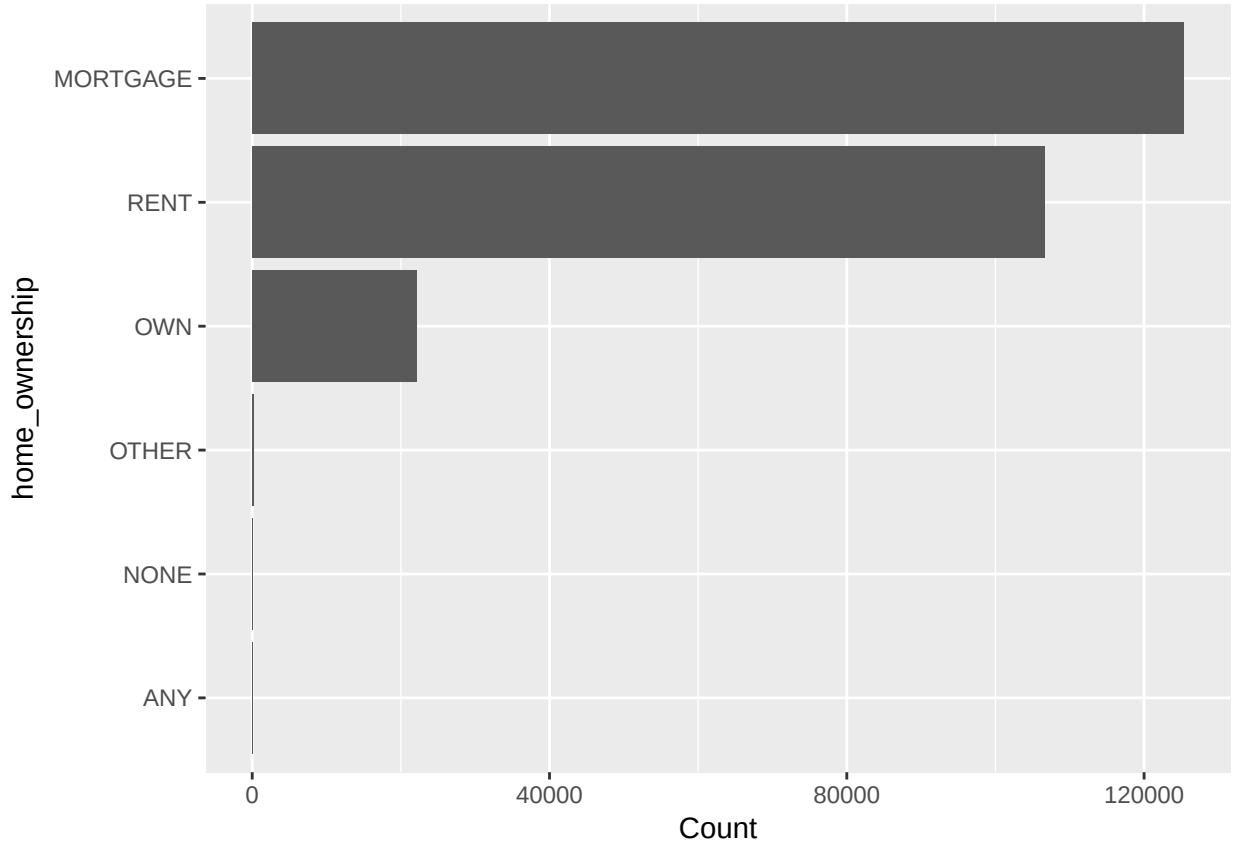
Terms




```
##  36 months  60 months
##     197373      56817
```

There are a lot more 36 month term loans than 60 months.

Home Ownership

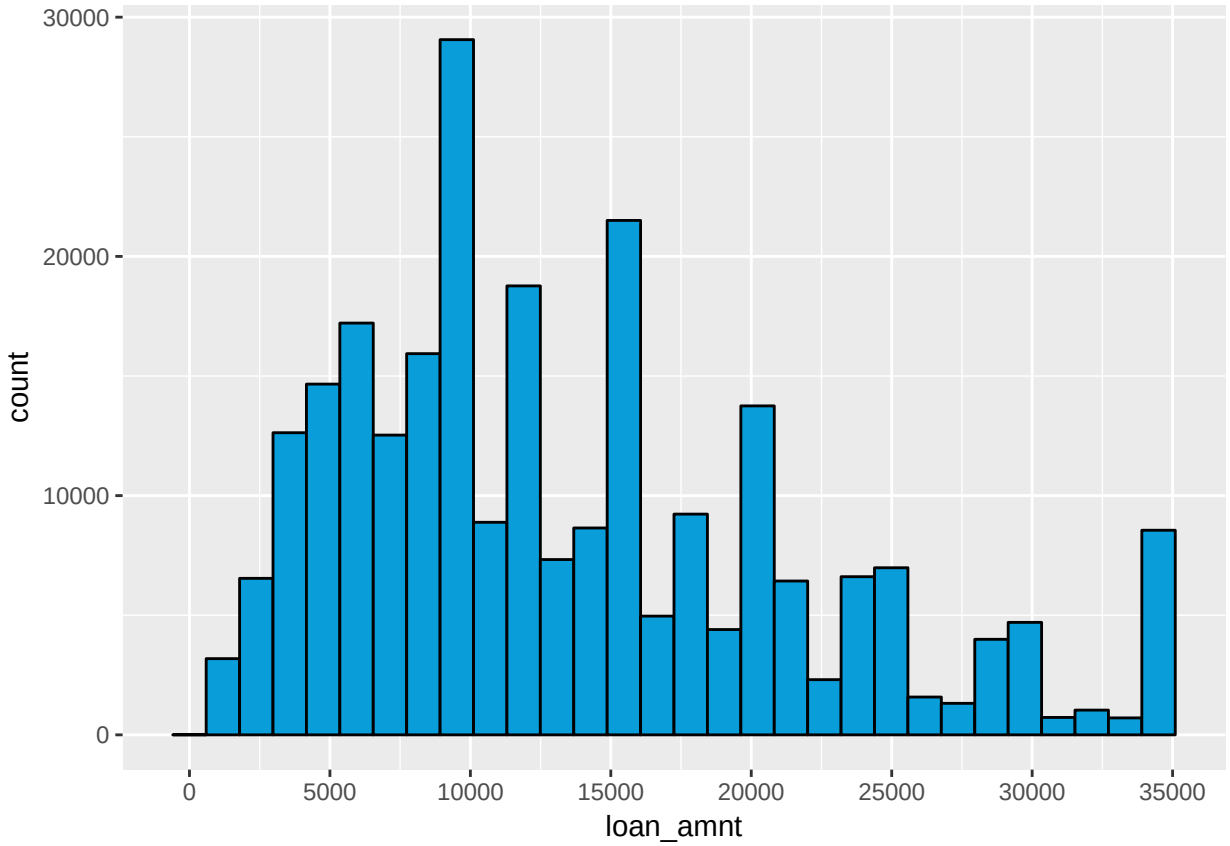



```
##       ANY MORTGAGE      NONE     OTHER       OWN      RENT
##         1   125342        43       141     22095    106568
```

Most borrowers either have a mortgage or rent, with a smaller amount of borrowers being home owners. "Other", "None", and "Any" are very few in comparison to the other three home ownership types.

Histograms of Continuous Variables

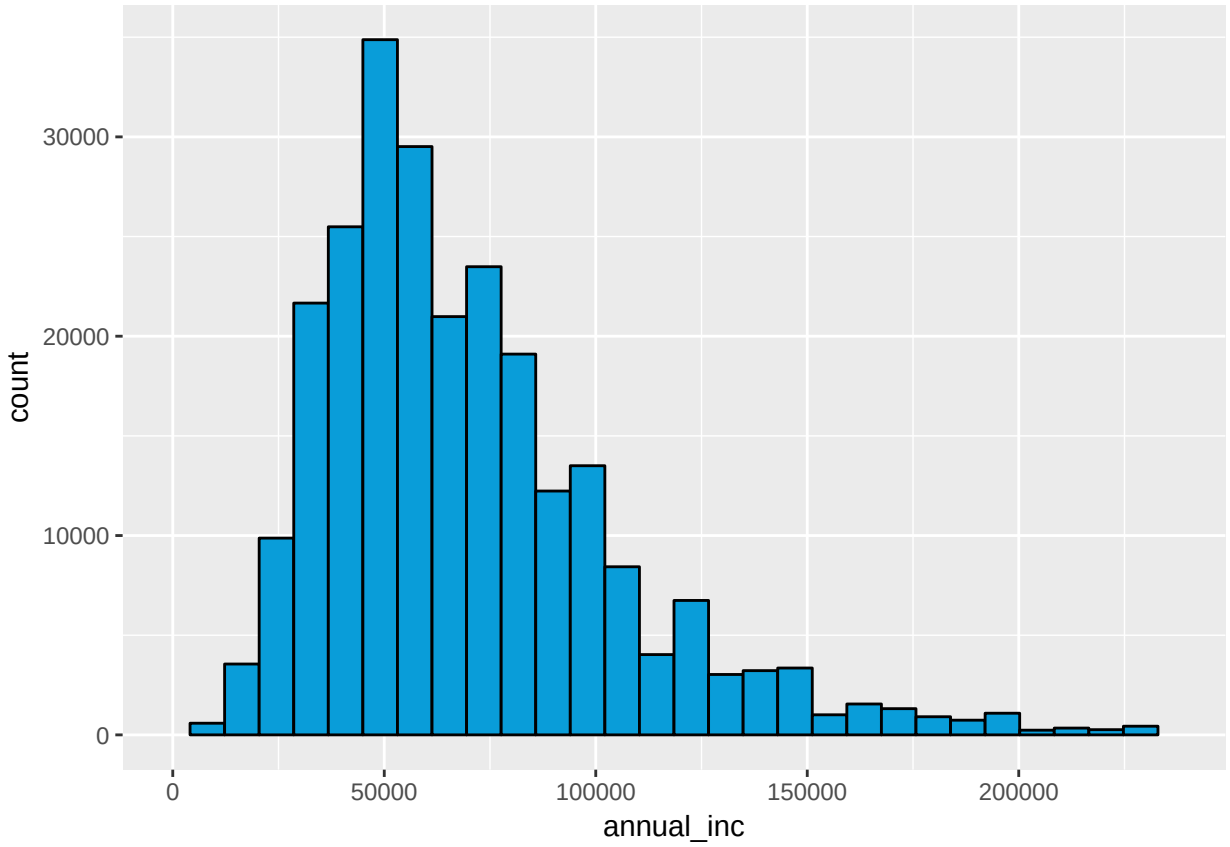
Loan Amount




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     500    7250   12000   13570   18250   35000
```

This right-tailed distribution shows that 75% of loan amounts are below 18,250, despite the max loan amount permitted by Lending Club being 35,000.

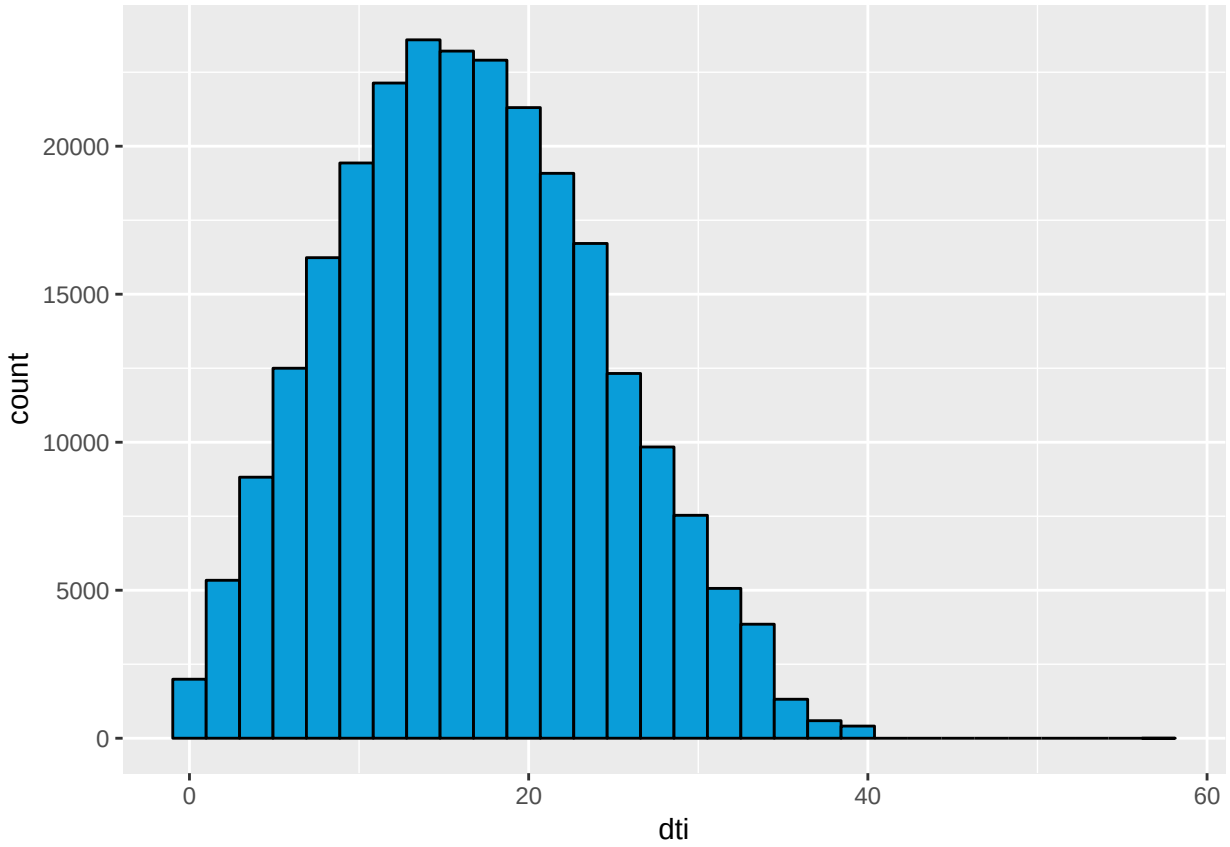
Annual Income




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3000   45000   62000   72510   87000 8707000
```

Considering that distribution of income in the US is positively skewed, it's not a surprise to see the same applies to a dataset of Lending Club borrowers. The median is 62,000, and 75% makes 87,000 or less.

Debt-to-Income Ratio

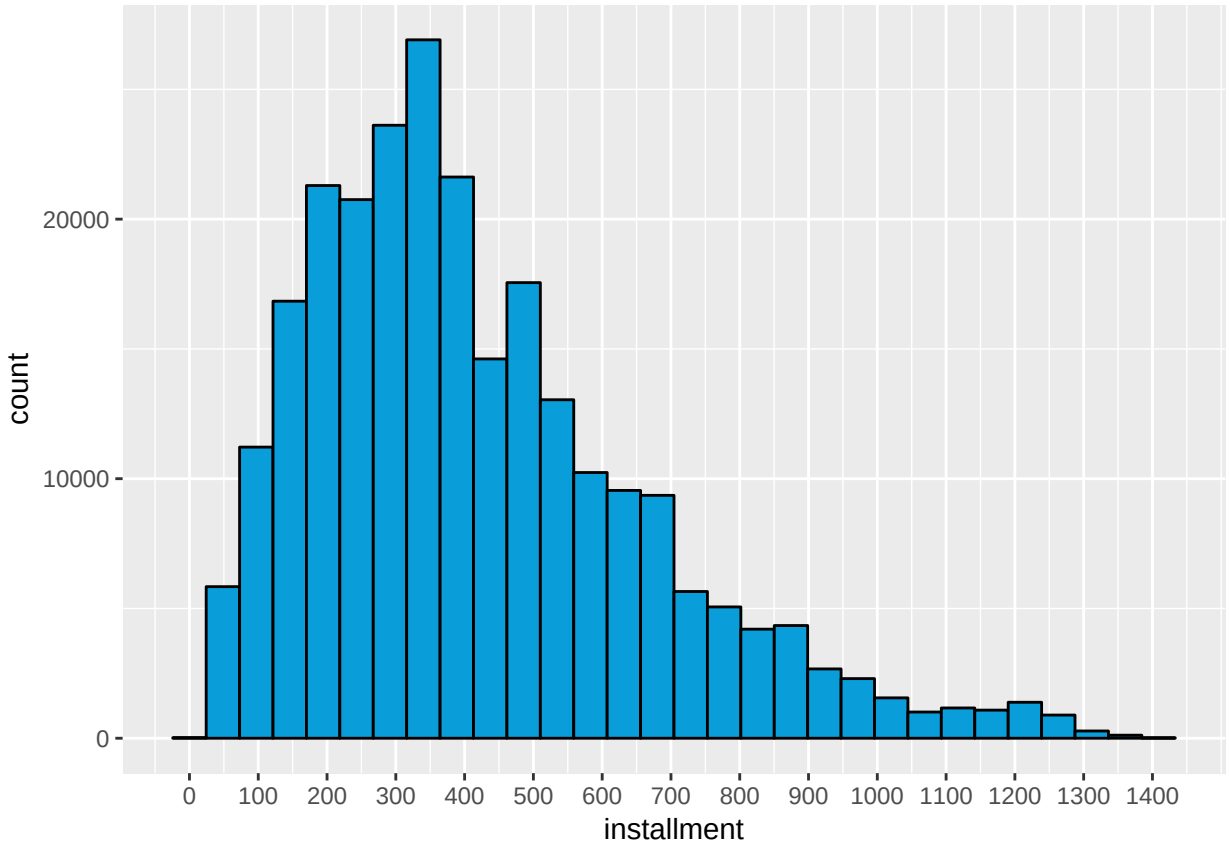



```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 -- 34.4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.77   16.22   16.56   22.01   57.14
```

Removing outliers beyond the 99th percentile, we have a normal distribution, with a median debt-to-income ratio of 16.22.

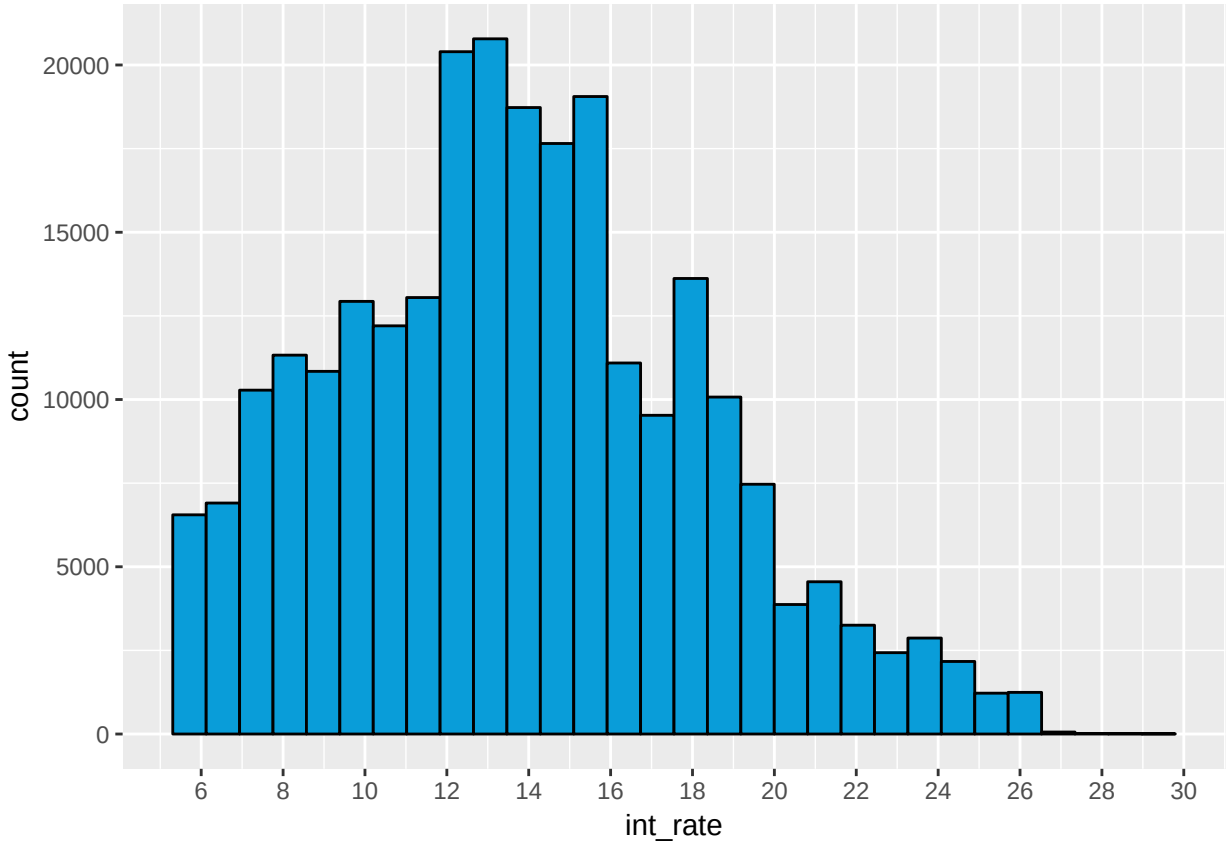
Loan Installment




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.69  239.60  365.20  418.30  547.60 1425.00
```

Loan installments are positively skewed, with a median of $365.20

Interest rates




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.32   10.74   13.53   13.78   16.55   28.99
```

The interest rate has a median value of 13.53, and a 16.55% is the 75th percentile. This is evident in the histogram, as there more values on the left/left-center of the distribution.

Univariate Analysis

What is the structure of your dataset?

The initial dataset contains 887,379 observations of 74 variables. There is a mix of continuous and categorical variables in the dataset. Since the dataset is quite large, I trimmed it. I cut it down to only have past loans (loans with loan statuses of "Default", "Paid Off", and "Charged Off") and 13 variables. The variables are: addr_state, annual_inc, dti, emp_length, grade, home_ownership, installment, int_rate, issue_d, loan_amnt, loan_status, purpose, term.

What is/are the main feature(s) of interest in your dataset?

The main feature of interest is the loan_status column, as that serves as the basis for the variable I create, default_status. I'm hoping to see what factors appear related to the default rate.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

I believe the grades variable will be of benefit, as such a variable essentially says what Lending Club thinks the loan's chances of defaulting are e.g. Is it a safe bet or risky?

Did you create any new variables from existing variables in the dataset?

Yes, I made a 14th variable called default_status. After trimming the dataset down to only include past loans, I made a variable that indicated whether or not a loan was paid off or were defaulted. If the loan's status was "Charged Off" or "Default", it would be given the label "Default". If it had the status "Paid Off", it would be placed in the "Paid" category.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Many graphs were positively skewed, like annual income, installment, and loan amount. The distributions for dti and annual_inc were initially problematic as they were heavily affected by a few outliers, so I set an x limit at the 99th percentile for both of them. Dti was shown to have a normal distribution after the outliers were removed.

I also factored the employee length variable in the process of trying to create it's bar chart, as the initial order of years shown in the chart wasn't in appropriate numerical order ('1o+ years' came immediately after '1 year')

Bivariate Plots

For the sake of a correlation plot, I'm going to make a variable called "default_status_int" that will have two values:

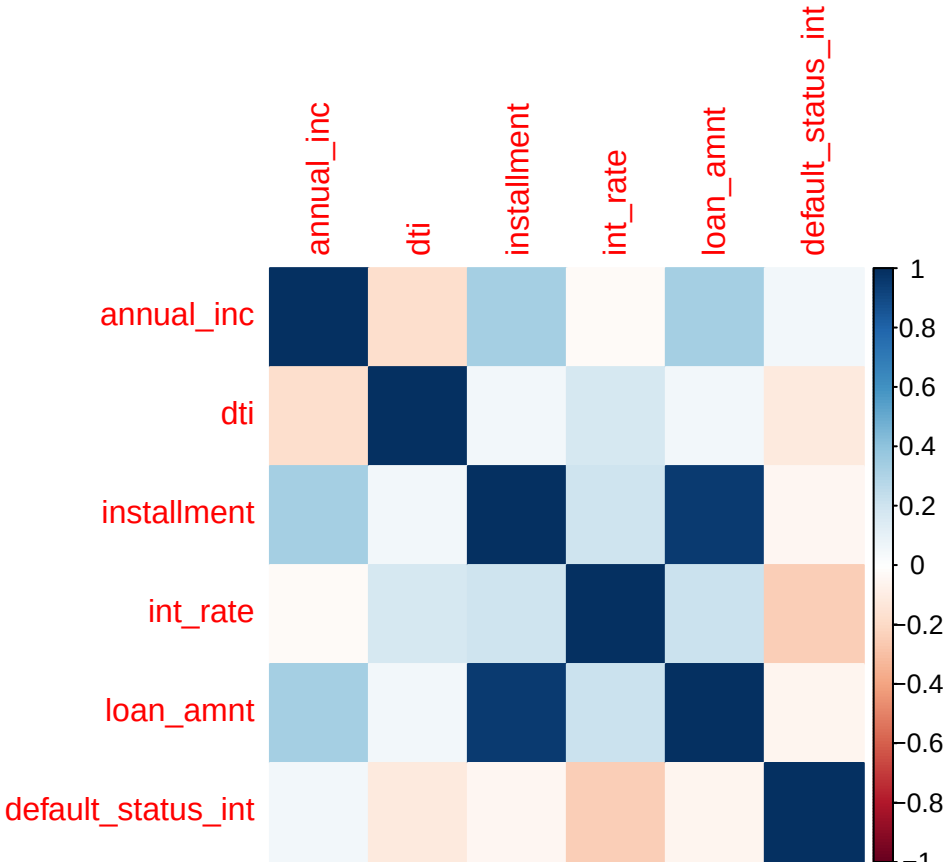
0 - If the loan has "Default" for default_status

1 - If the loan has "Paid" for default_status

Once I do this, I'll make a correlation plot. I'll be able to get a numerical value for correlations with default_status this way.

Correlation Plot

```
# Making the variable

loans.2$default_status_int <- ifelse(loans.2$default_status == "Paid", 1, 0)

# Making the corr plot

M <- cor(loans.2[,c("annual_inc","dti","installment", "int_rate", "loan_amnt",
        "default_status_int")]) # get correlations

# Plot matrix

corrplot(M, method = c('color'))
```



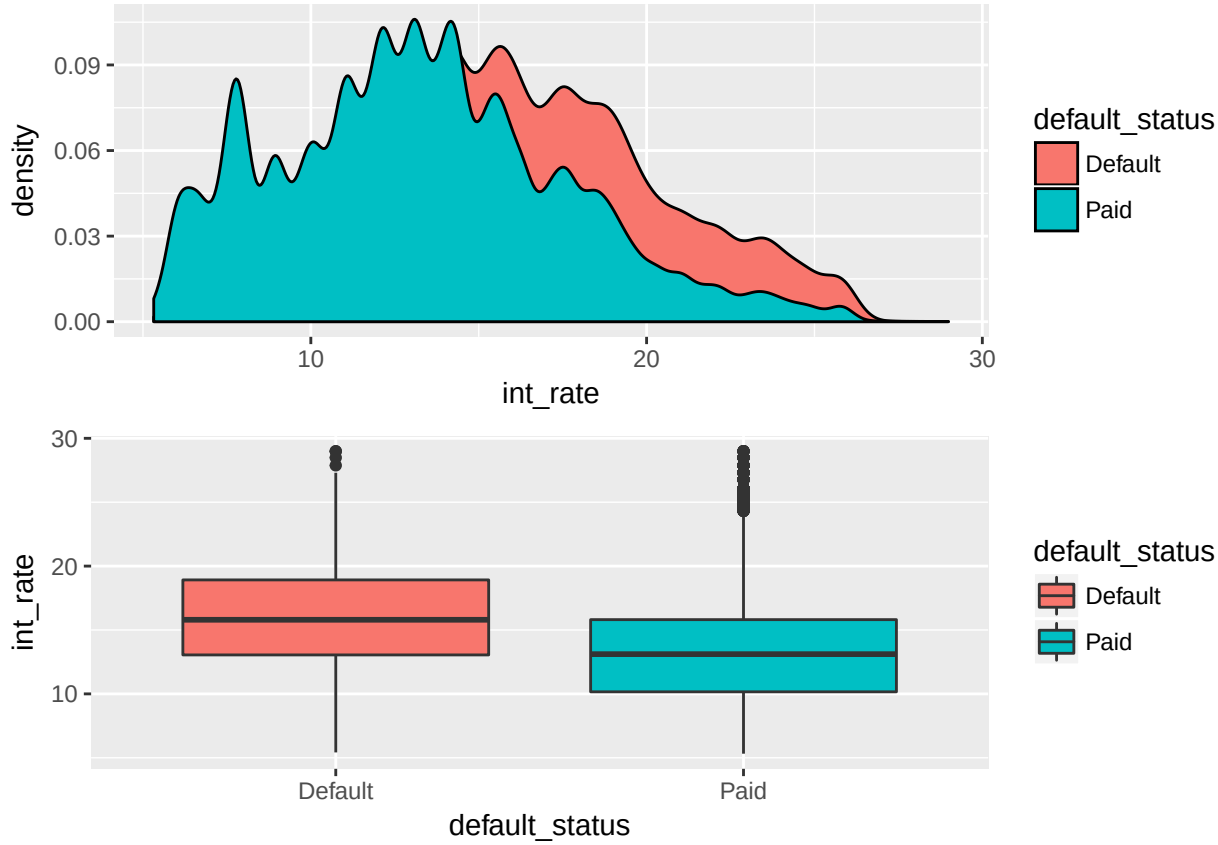

```
# Output correlation figures

as.data.frame(M)
```

```
##                     annual_inc         dti installment   int_rate
## annual_inc          1.00000000 -0.17176704  0.33003196 -0.0262819
## dti                -0.17176704  1.00000000  0.05363979  0.1728748
## installment         0.33003196  0.05363979  1.00000000  0.2026930
## int_rate           -0.02628190  0.17287484  0.20269302  1.0000000
## loan_amnt           0.33461379  0.05264769  0.95510093  0.2113808
## default_status_int  0.05938973 -0.11430378 -0.04403064 -0.2408903
##                      loan_amnt default_status_int
## annual_inc          0.33461379         0.05938973
## dti                 0.05264769        -0.11430378
## installment         0.95510093        -0.04403064
## int_rate            0.21138076        -0.24089033
## loan_amnt           1.00000000        -0.05829247
## default_status_int -0.05829247         1.00000000
```

Observations:

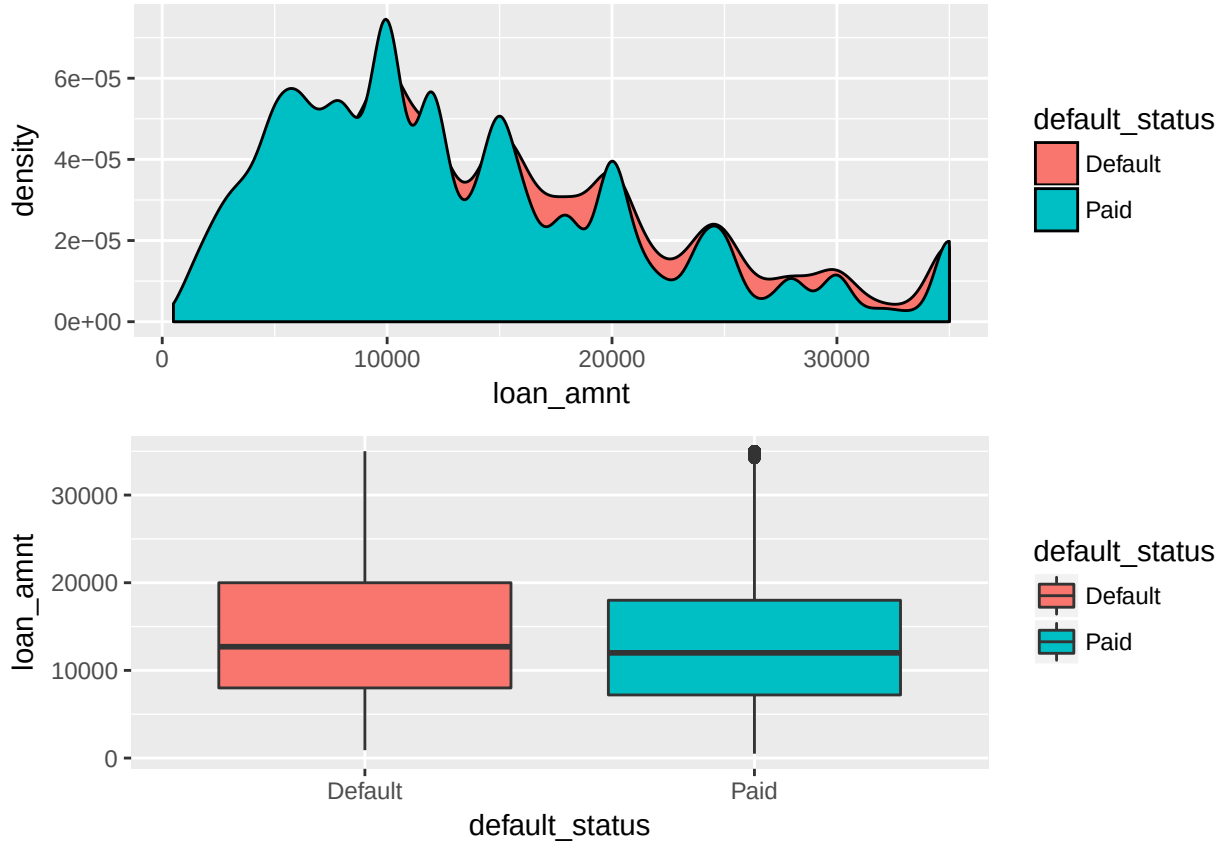
- There are no strong correlations with default status, but interest rate (-0.241) and dti (-0.114) have the highest correlation relative to the rest of the variables.
- Annual income appears to have minor correlations with installment (0.330) and loan_amount (0.334)
- Loan amount, unsurprisingly, has a near perfect linear correlation with installment (0.955). The bigger your loan amount, the more you're paying on a monthly basis.

Interest rate broken down by paid_status



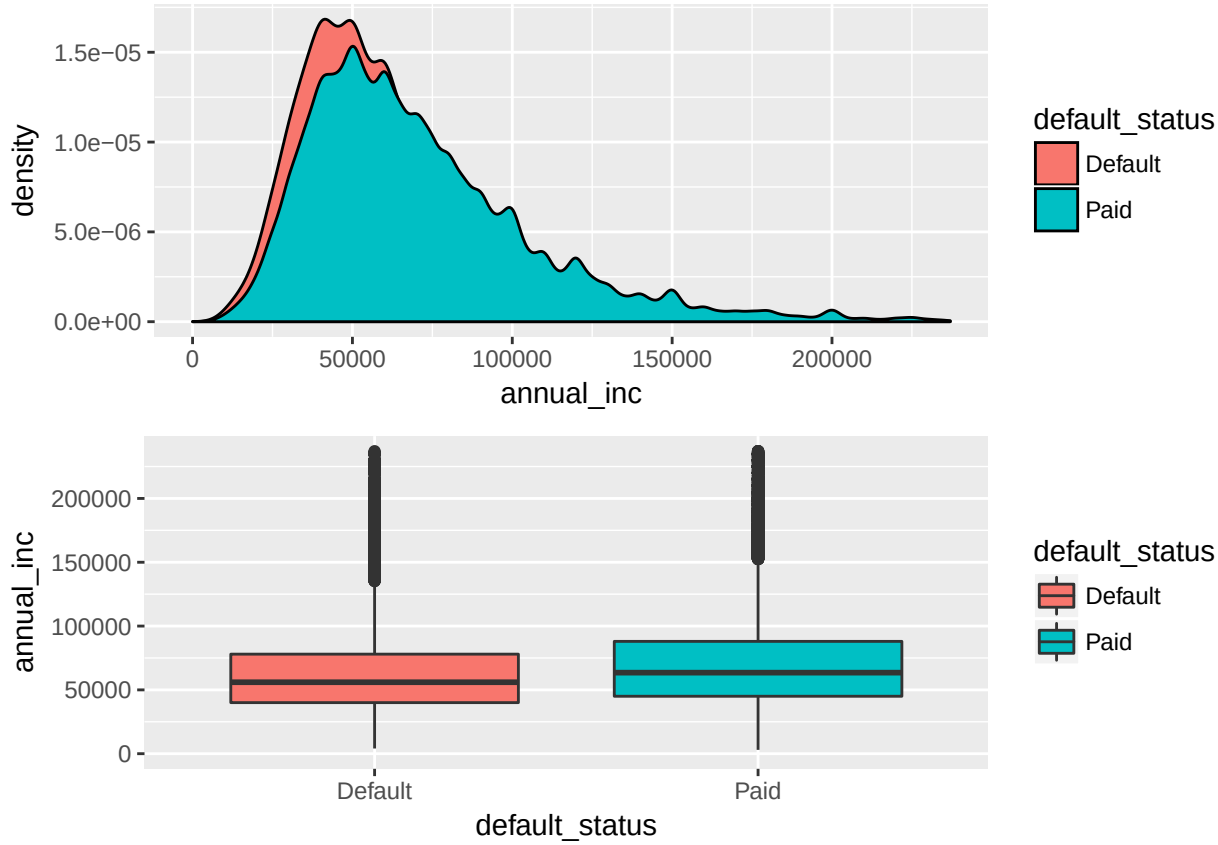
Interest rates are higher for borrowers who have defaulted on their loans.

Loan amount broken down by default_status



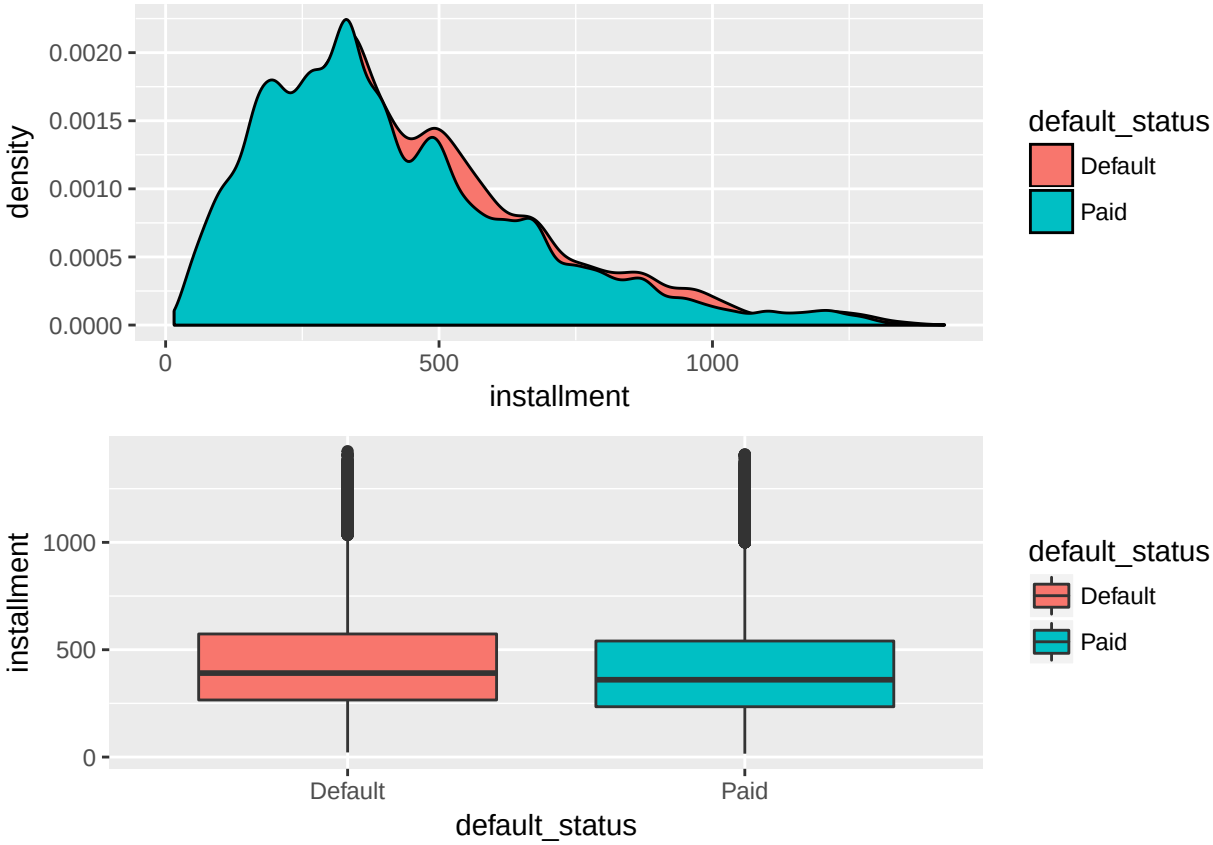
Loan amounts for borrowers who default on their loans tend to be slightly higher than borrowers who fully pay back their loans.

Annual income broken down by default_status



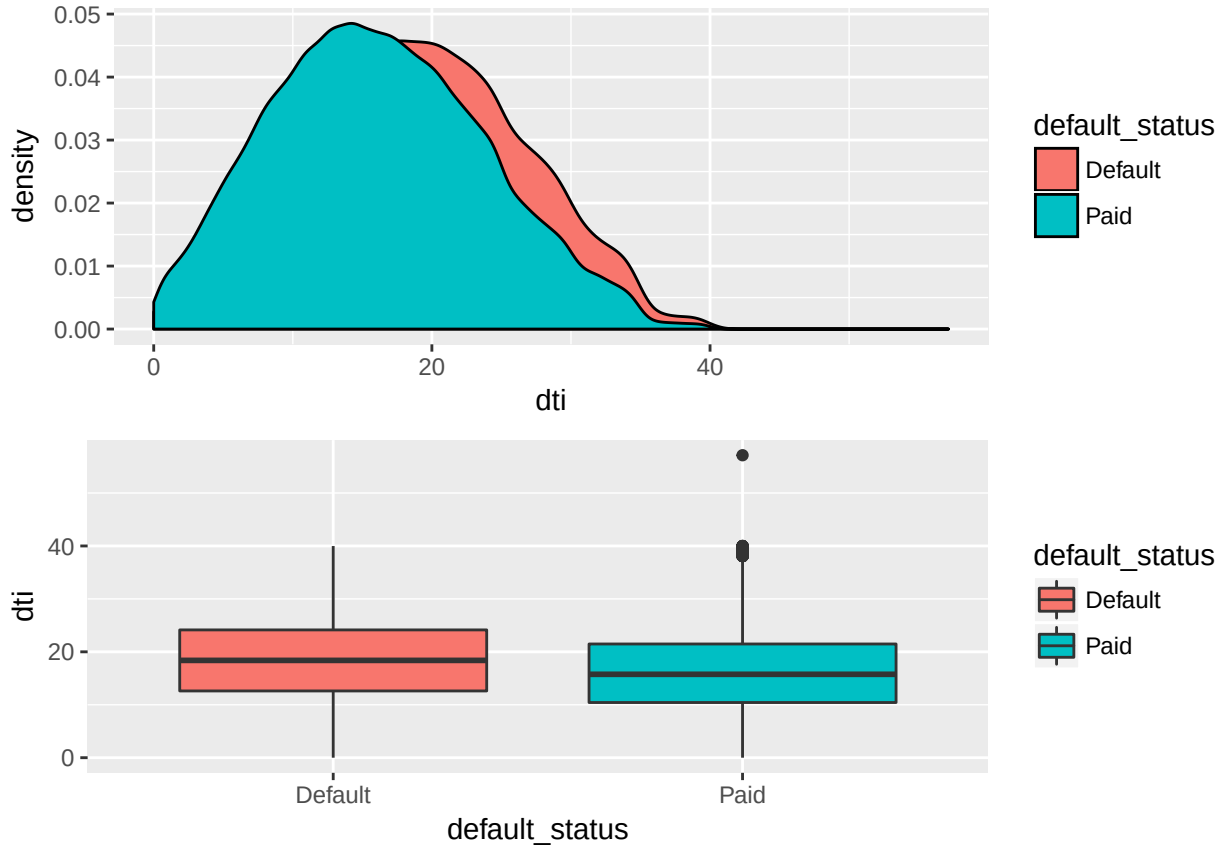
The mean annual income of borrowers who default is slightly lower than borrowers who pay back their loans.

Installment broken down by default_status



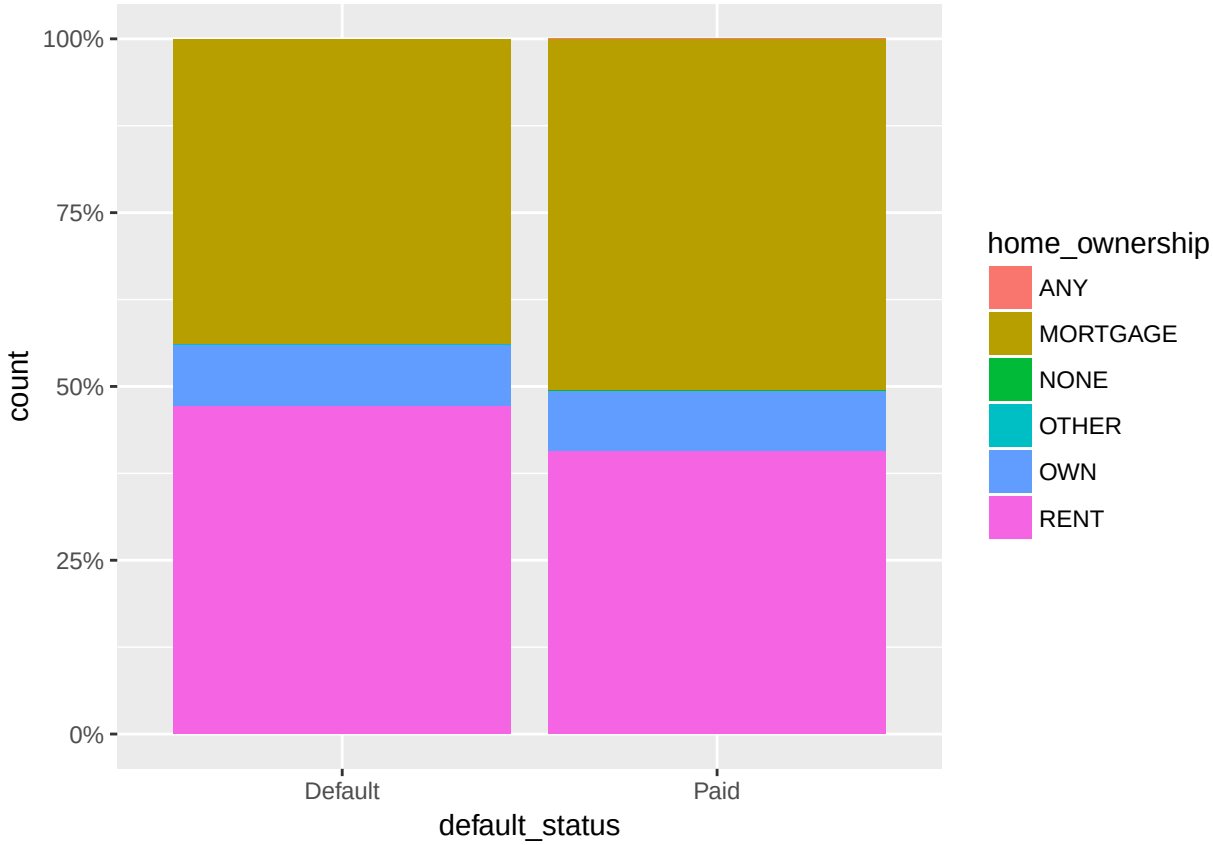
Consistent with what we saw with loan amounts, borrowers who default have slightly higher installments than borrowers who don't default.

Debt-to-income ratio broken down by default_status



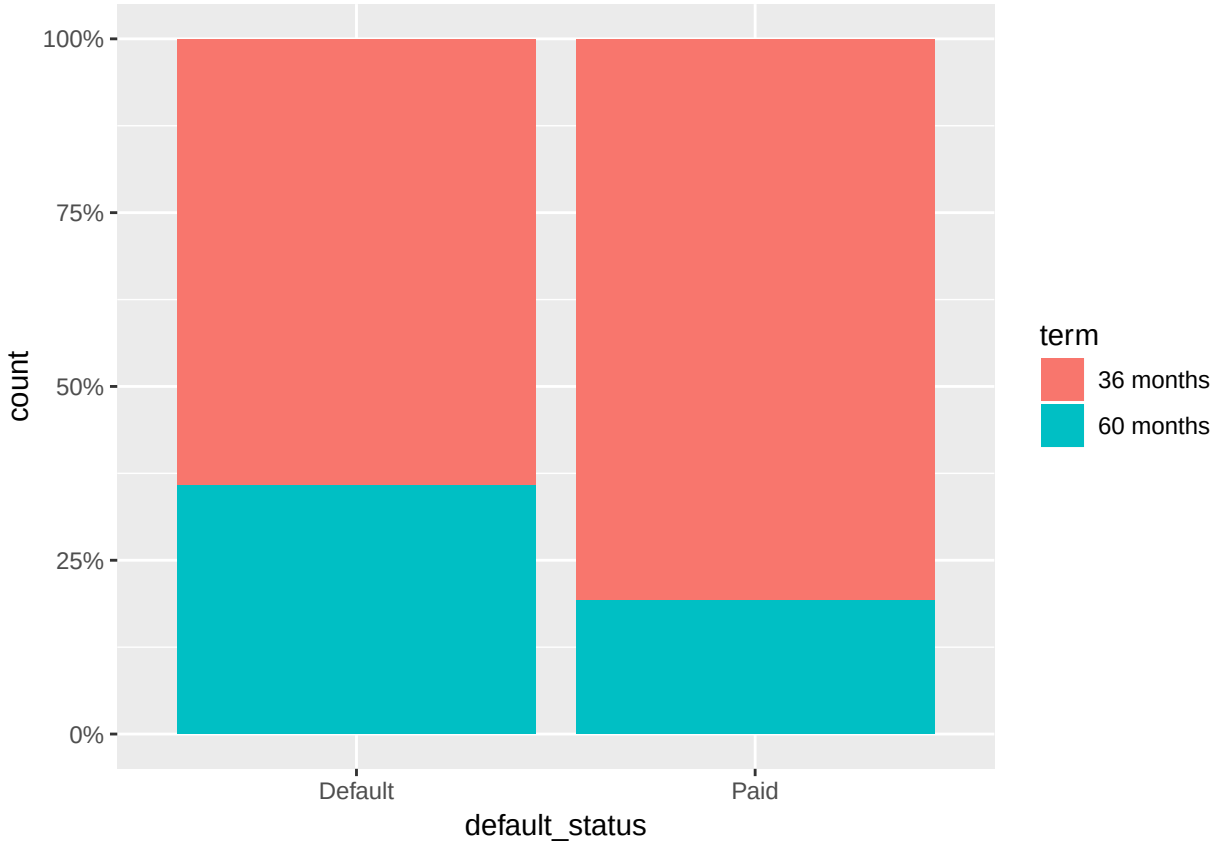
Debt-to-income ratio is higher for borrowers who default on their loans.

Default Status broken down by home_ownership



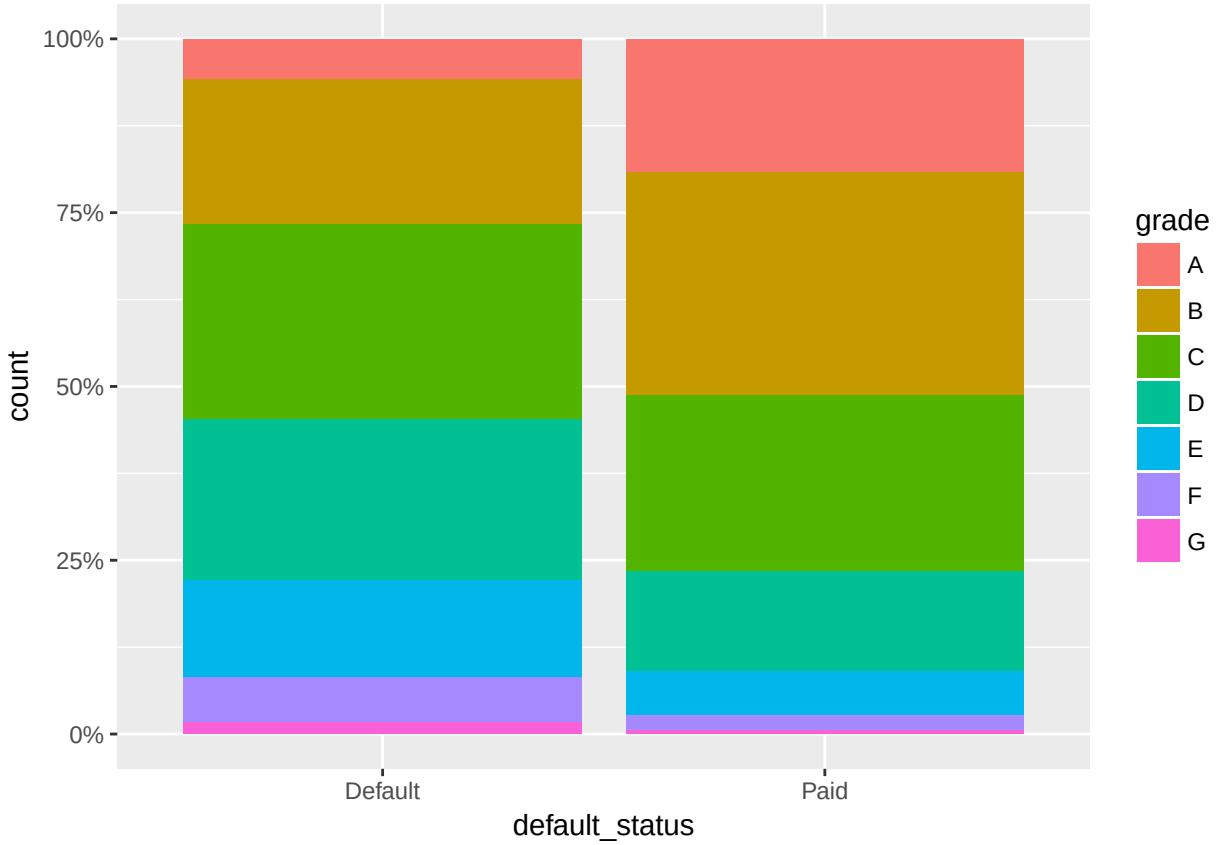
Home ownership split seems to be rather similar, but there are slightly more renters in the Default set of borrowers.

Default status broken down by loan terms



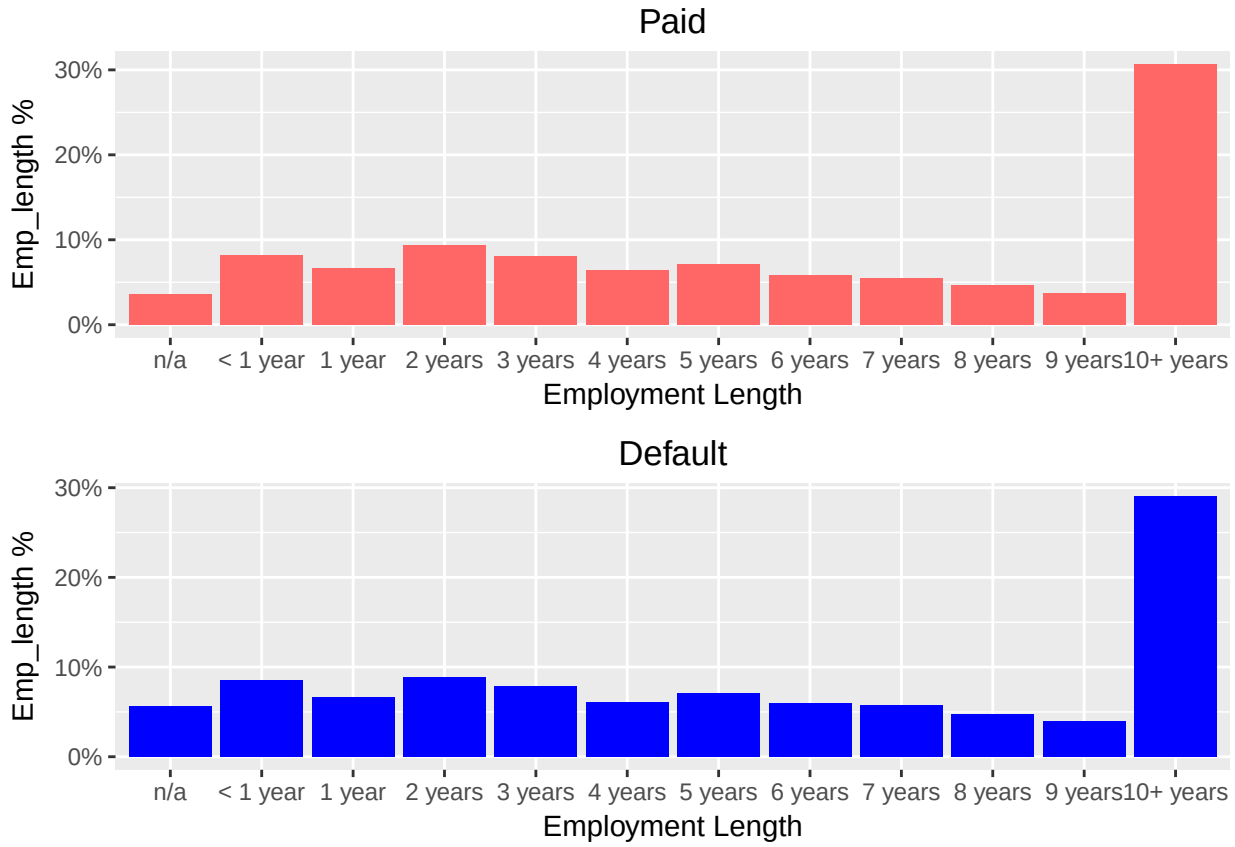
There is an increase in 60 month terms in the Default category of loans.

Default Status broken down by grade



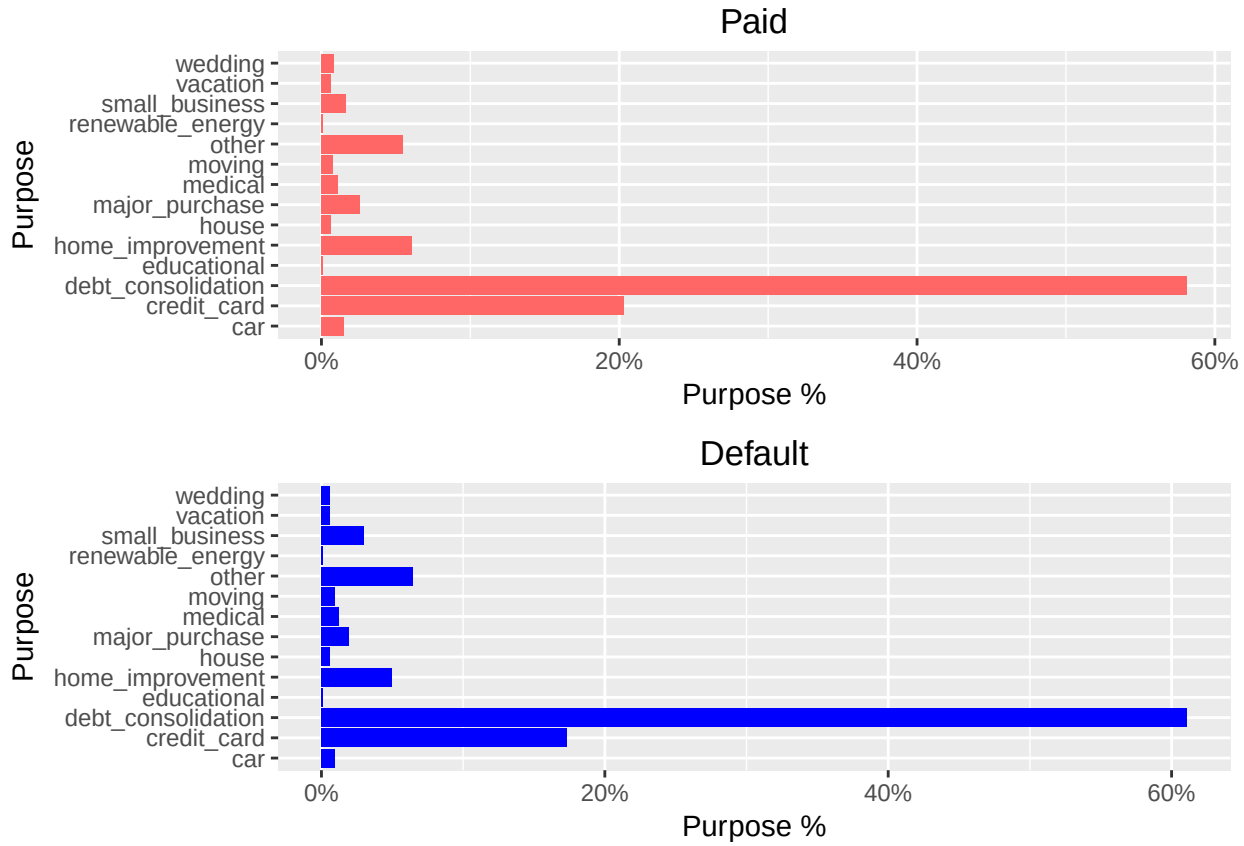
Here we can see that C, D, E, and F grades make up a higher proportion of the defaulted loans than they do for loans that were paid off.

Default Status broken down by proportion of employment length



Employment length split seems to be the same for both default and paid loans.

Default Status broken down by proportion of purpose



The distribution of the purpose variable is pretty similar, but we can see that loans that defaulted have slightly more borrowers looking to consolidate their debt.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Interest rate saw the most significant rise when comparing default loans vs paid loans. Debt-to-income ratio experiences a jump as well, but perhaps not to the extent of interest rate. Annual income appears to have a minor drop for default loans, while loan installment and loan amount experience a minor increase for default loans as well.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

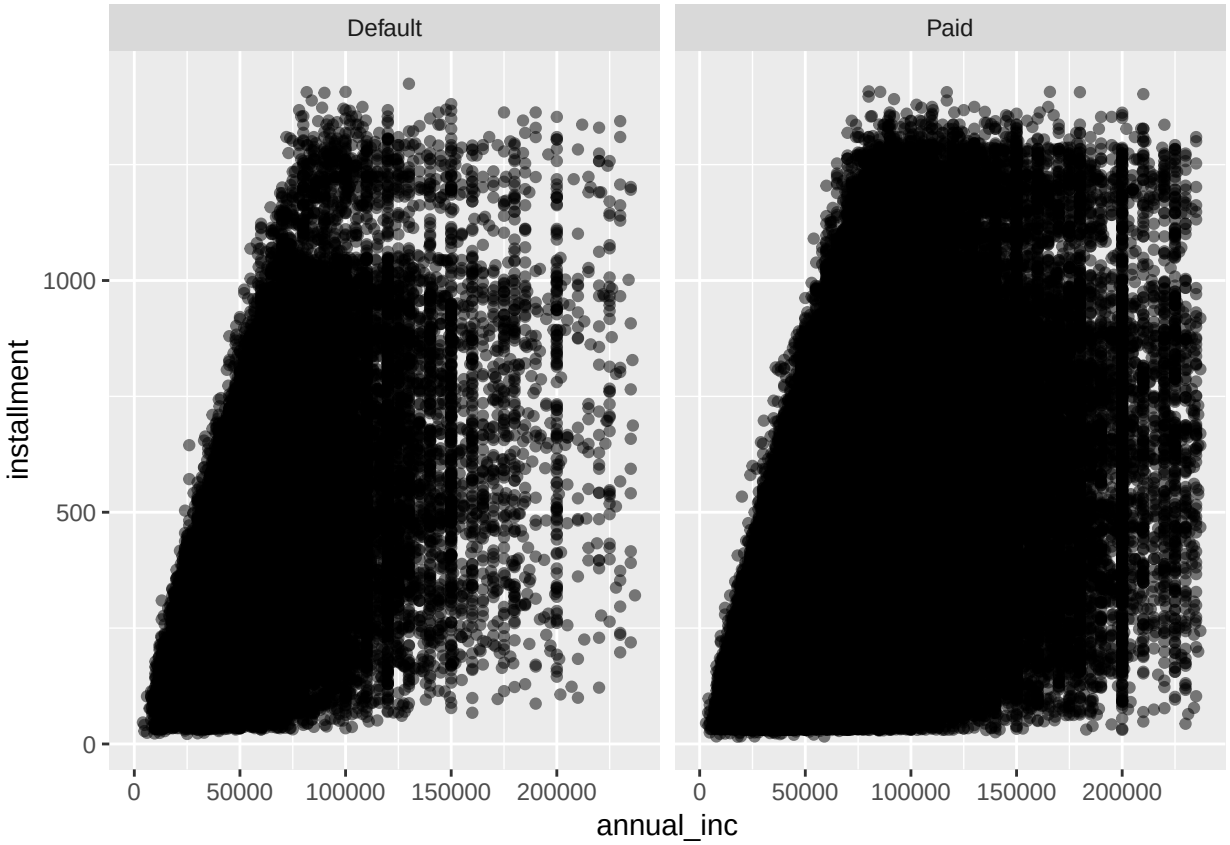
The corr plot indicated that there is a minor positive correlation between installment and annual income, at 0.22 Pearson's R. Installment and interest rate also have a positive correlation of 0.22.

What was the strongest relationship you found?

Between two continuous variables, the strongest relationship I found was between installment and annual income. Between categorical and continuous variables, interest rate and default_status was the strongest. Lastly, between two categorical variables, it appears that grades shift heavily from the top grades to the worst grades between paid loans and default loans.

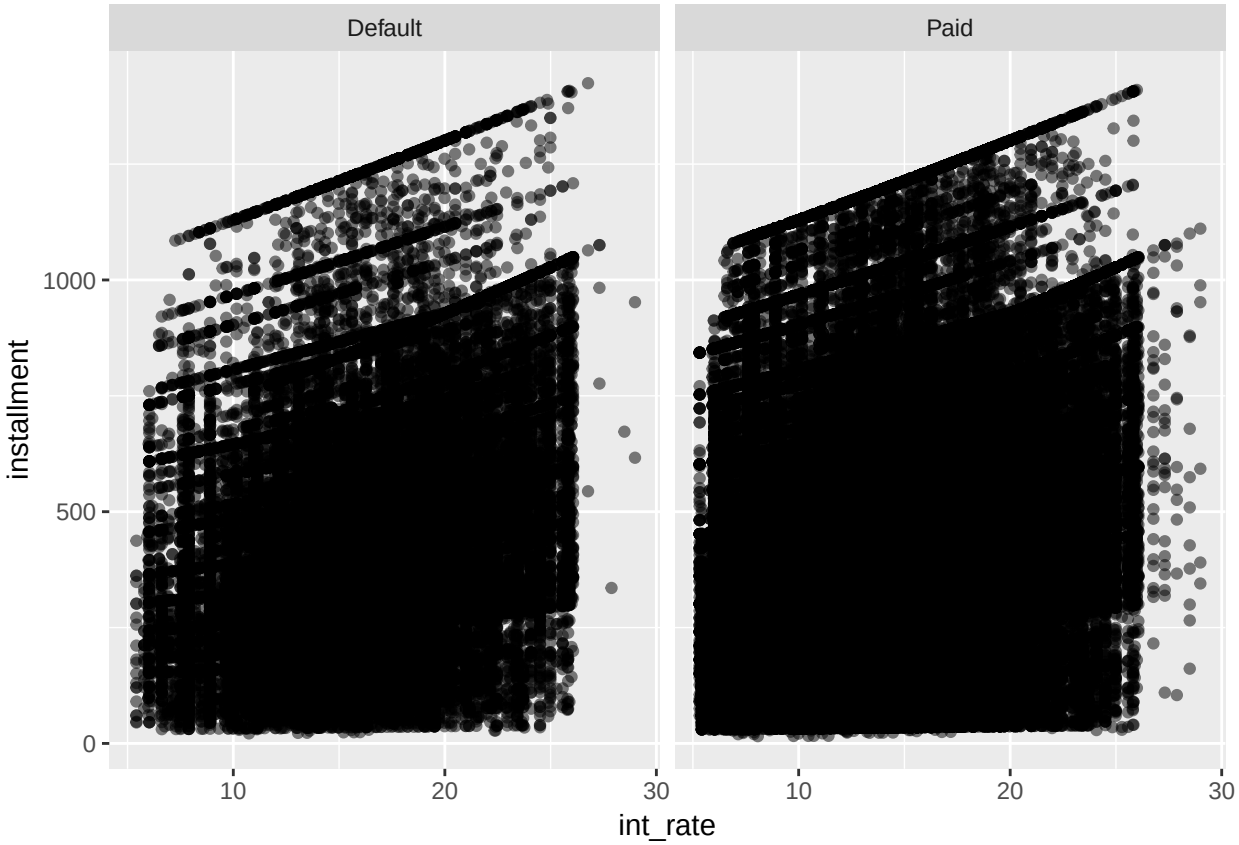
Multivariate Plots

Annual income and installment relationship broken down by default status



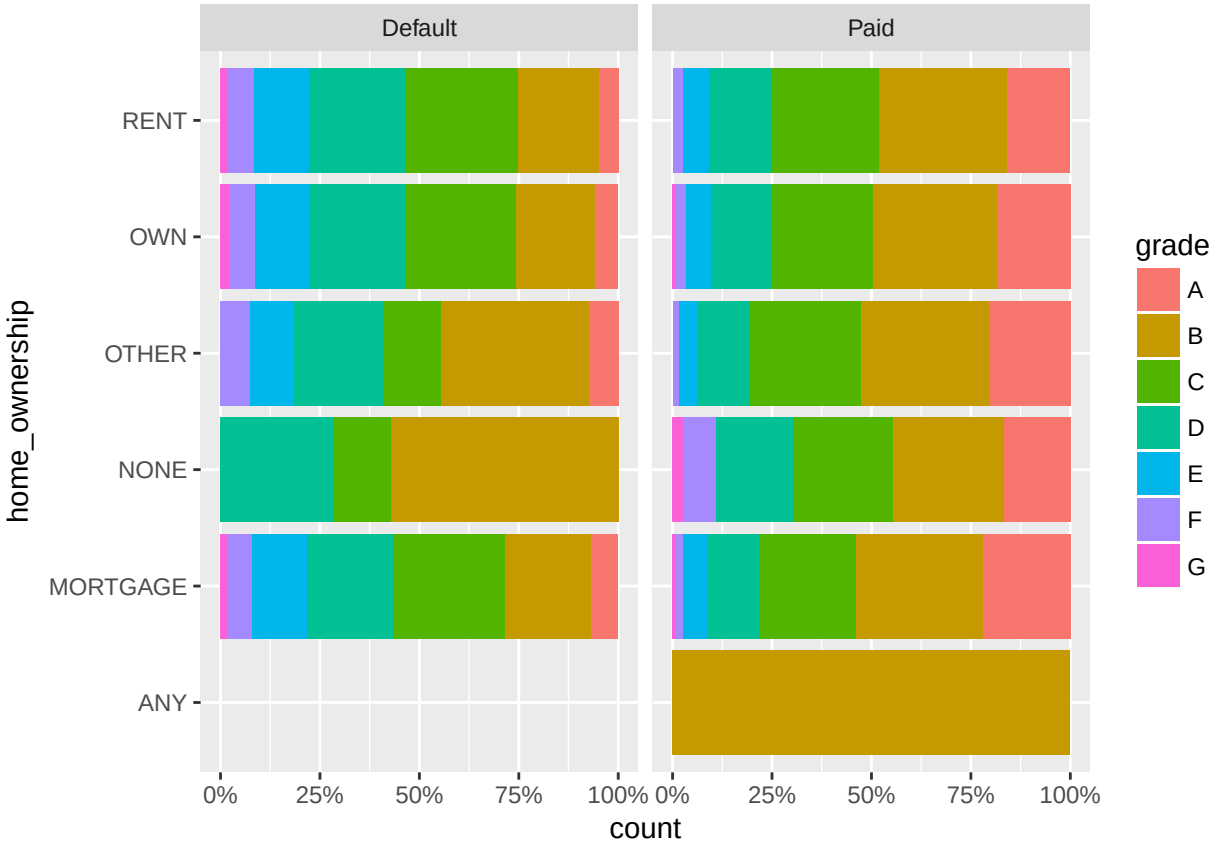
There doesn't appear to be any meaningful differences. There's a slight positive relationship between annual income and installment, but no new significant information appears when we compare loans that defaulted and loans that were paid for.

Installment and interest rate relationship, broken down by default status



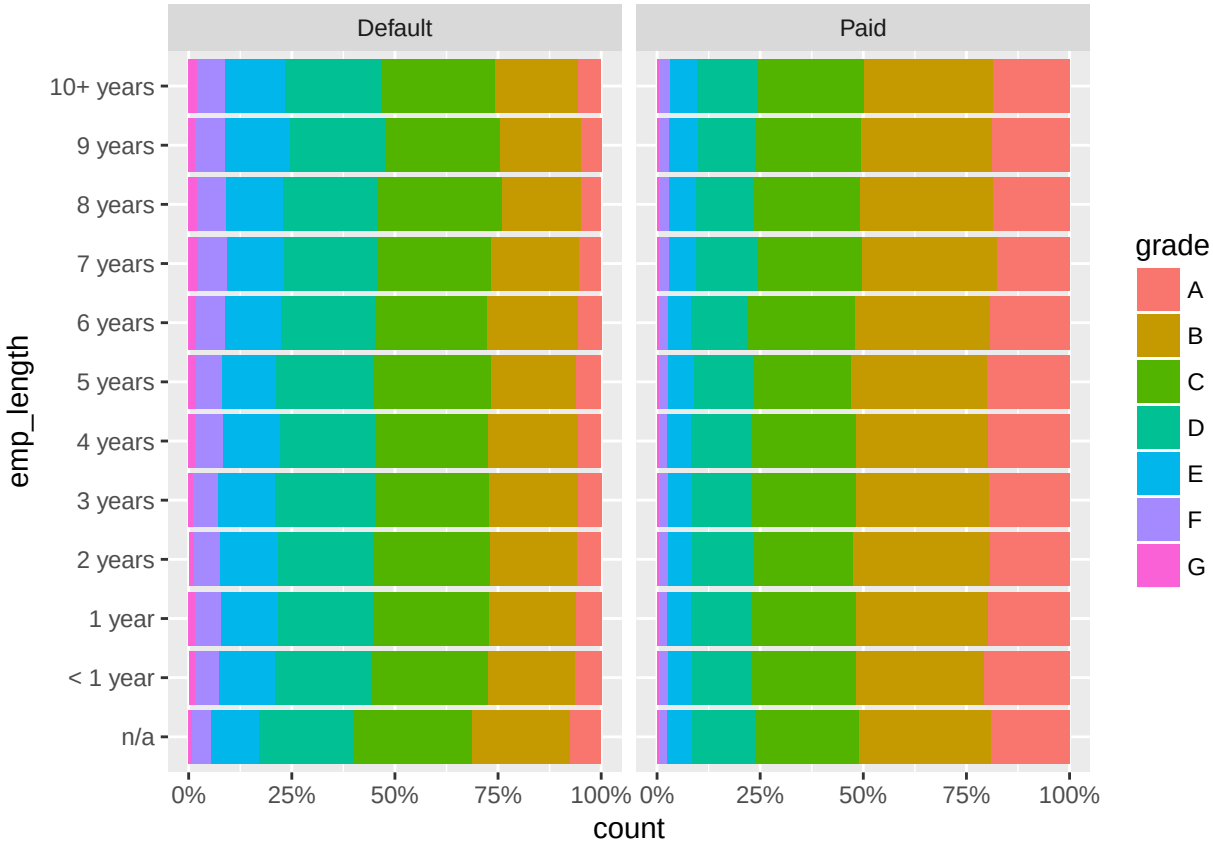
There doesn't appear to be anything meaningful here when we break down the relationship between int_rate and installment (which already doesn't have any serious correlation) broken down by default status.

Proportion of grades for each home ownership type, broken down by default status



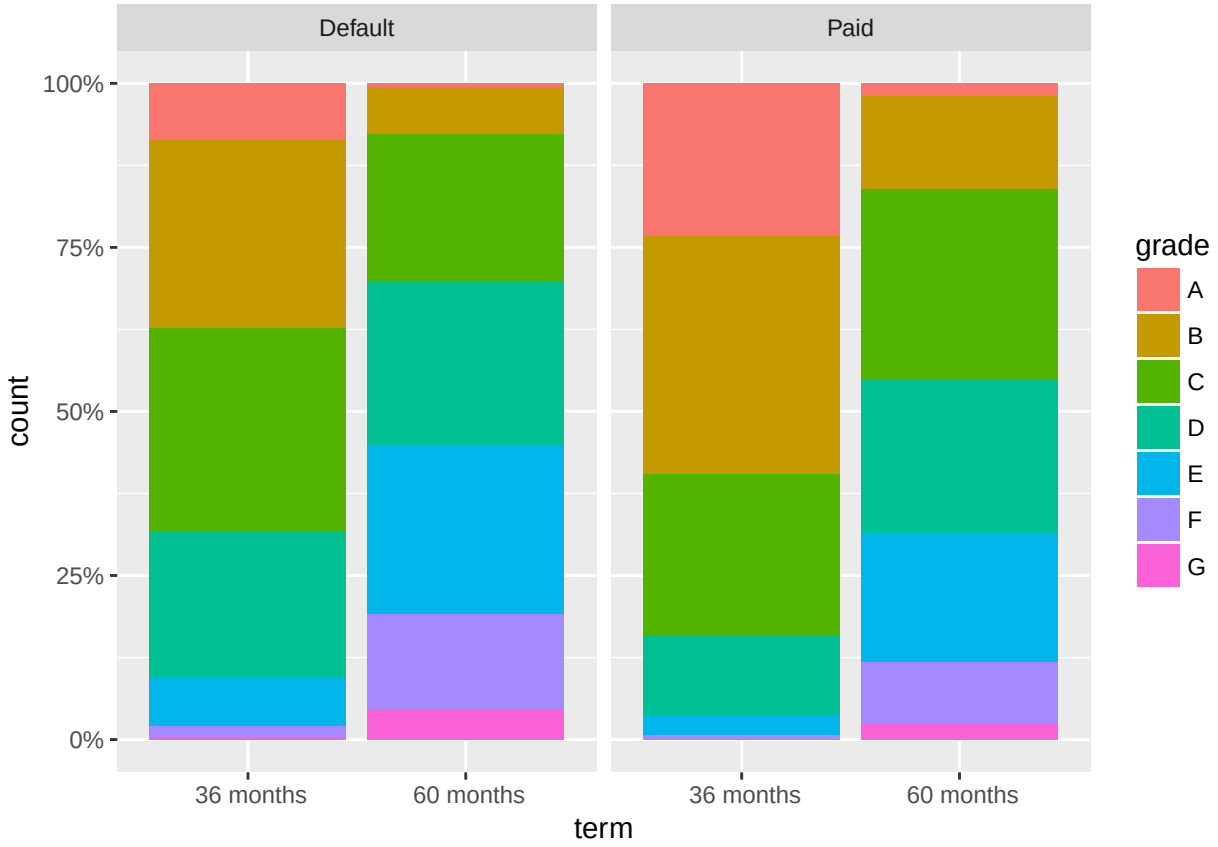
The relationship between grades and home ownership type is consistent among default status.

Proportion of grades for each length of employment, broken down by default status



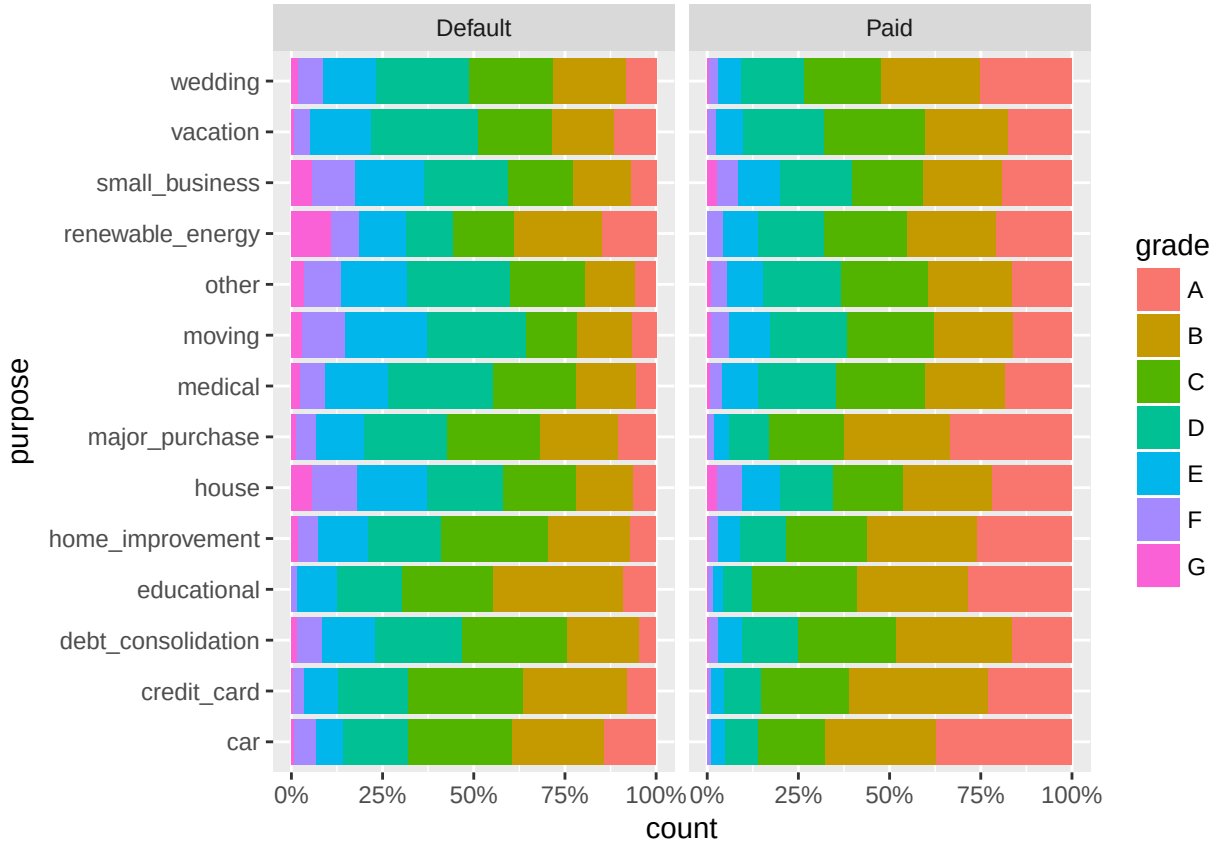
There isn't any new significant information here. It looks like the proportion of grades by year is pretty consistent throughout the years, and breaking them down to default and paid loans doesn't change much.

Proportion of grades for each term, broken down by default status



For both default and paid loans, 60 month loans see less higher grades like A and B, and increase in E/F/G loans, when compared to 36 month loans. We can see that the smallest chunk of A grades appears in the 60 month loans column for the Default loans. The biggest chunk of G grades appear in the same column.

Proportion of grades for each purpose, broken down by default status



It should be prefaced that it is difficult to confidently make claims here as there are far more data points for borrowers hoping to consolidate loans or pay off credit card debt than any other category. The "renewable_energy" purpose in the Default column has the biggest proportion of loans with G grades. Loans taken out for car purchases in the "Paid" column has the biggest proportion of A grades.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

When looking at the relationship between interest rate & installment, and installment and annual income, there didn't appear to be anything significant to take note of when using default status as a filter.

For the other two graphs, when looking at the breakdowns of grades for each default status, across all home ownership types or length of employment, there too appeared to be very little of note.

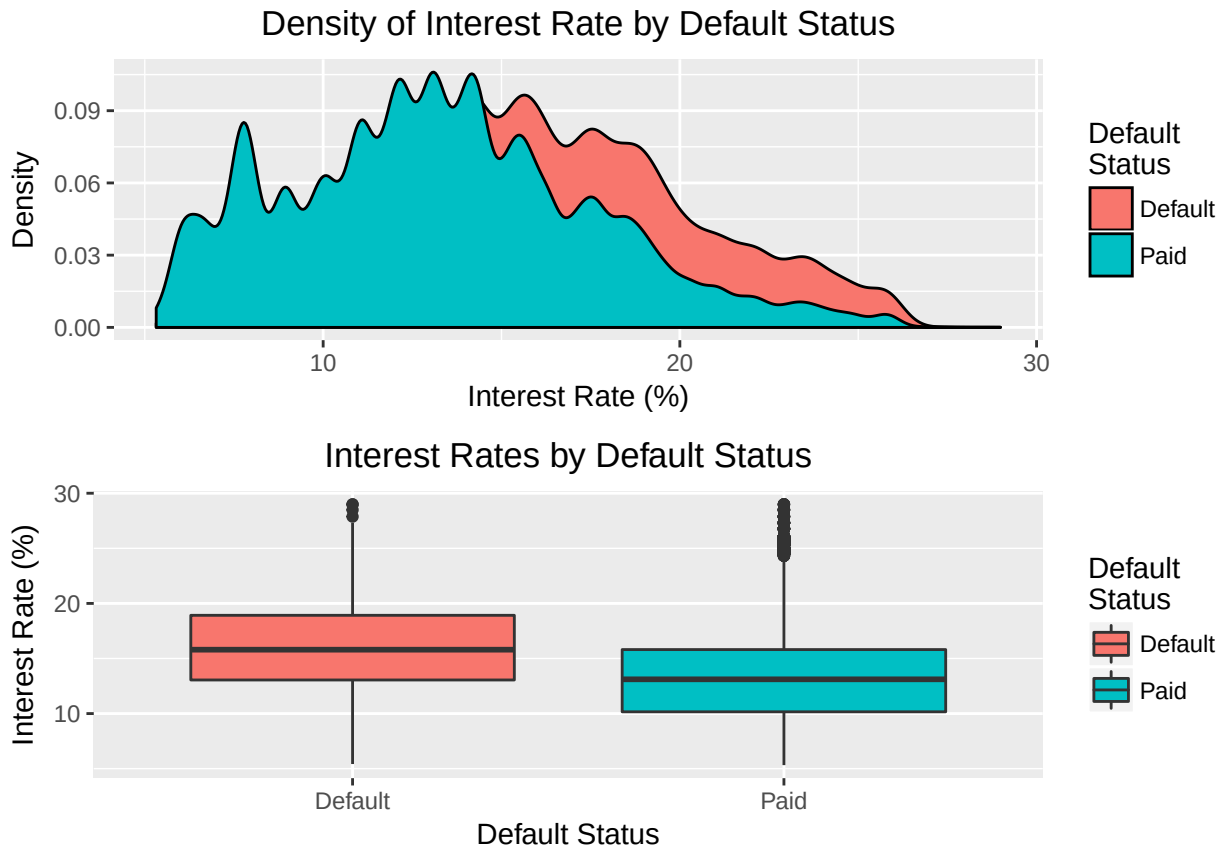
Were there any interesting or surprising interactions between features?

It's not surprising, but it's worth nothing that the 60 month terms among defaulted loans had a considerable amount of grade G loans.

The only surprising observation was that defaulted loans under the renewable energy purpose saw a significant increase in grade G loans, compared to other purposes.

Final Plots and Summary

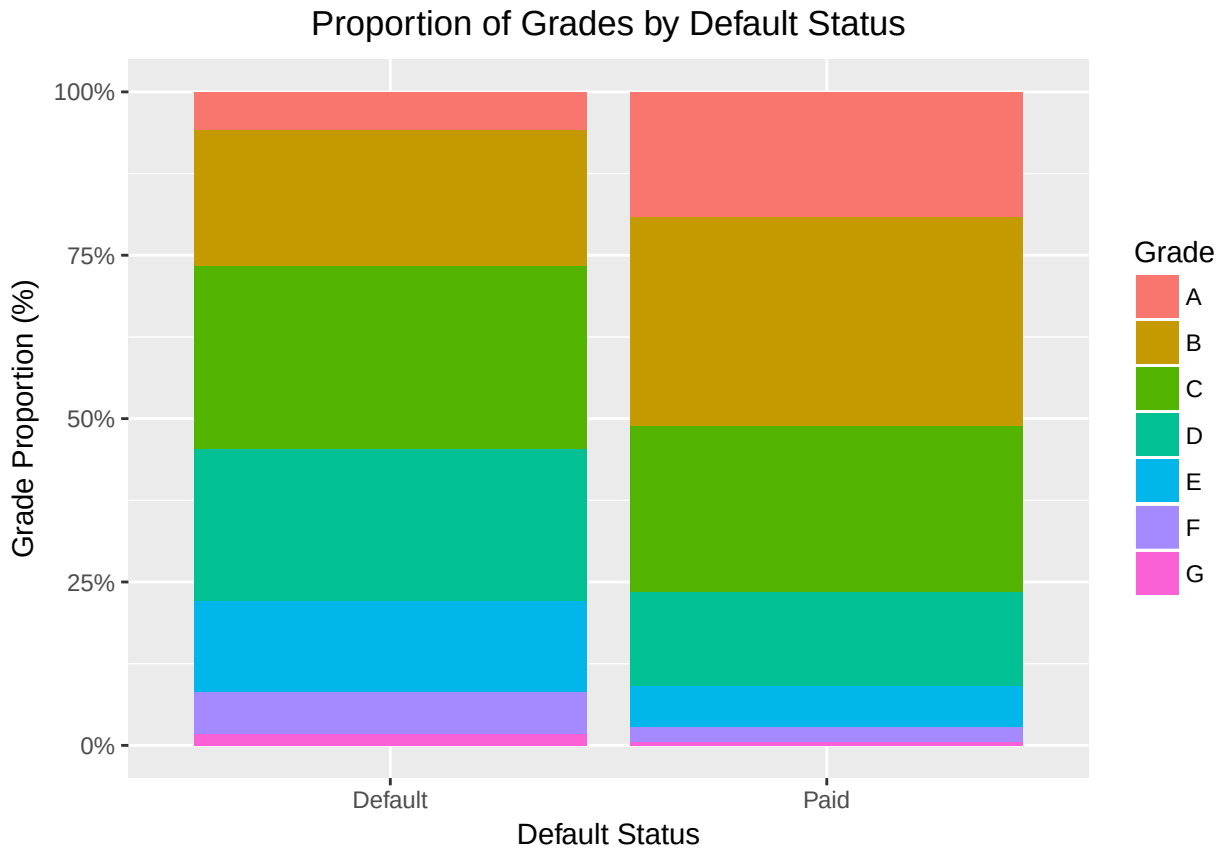
Plot One



Description One

This bivariate plot shows two plots - a KDE plot and a box plot. While interest rate's correlation with default status isn't very strong, it was the biggest among all of the continuous variables. These two plots clearly show that loans that default have higher interest rates than their paid counter parts. The KDE plot shows the default status further on the right, while the box plot shows that the median interest rate for defaulted loans is $> 20\%$, while $< 20\%$ for paid loans.

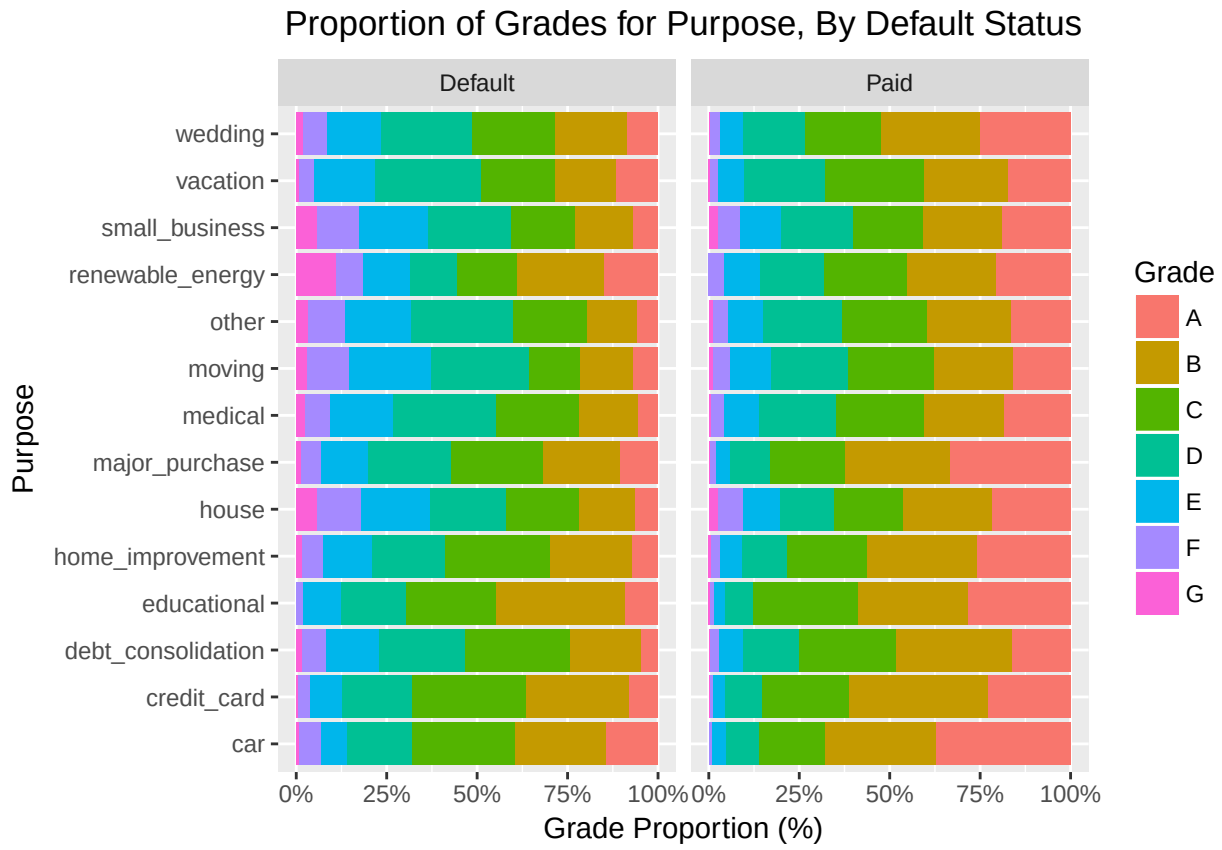
Plot Two



Description Two

This bivariate plot shows us the distribution of grades for each default status. Unsurprisingly, we can see that there are significantly less A grades for defaulted loans, alongside a decrease in B grades, and increase for C grade loans and lower.

Plot Three



Description Three

In this multivariate plot, we can see how the proportion of grades for each purpose changes between default and paid loans. Similar to the previous plot, A & B grades decrease in proportion for default loans but C grades and below increase. What is worth nothing here is where the proportion of A grades and G grades is the highest. Car loans that have been paid off have the highest proportion of A grade loans. Renewable energy loans that defaulted have the highest proportion of G grade loans.

Reflection

This dataset of Lending Club loan data initially had 887,379 variables and 74 variables. The data was cut down to only have 13 variables, and 254,190 observations (only past loans). 9 were continuous, 5 were categorical. I aimed to find out what variables were correlated with default rate, so I made a 14th variable called "Default Status" that indicated whether the past loan defaulted/charged off or was paid off.

For my univariate analysis, I made bar plots to look at the counts of categorical values, and histograms for continuous variables. Four of the five histograms were positively skewed - only debt-to-income ratio had a normal distribution. I had to set an x-limit at the 99th percentile for dti and annual income as their histograms were being affected by their extreme outliers. Also, employee length was factored, in order to get a bar chart that was in the correct numerical order, starting from "n/a" and ending at "10+ years".

When it came to bivariate analysis, I initially made a variable simply titled "default_status_int" that set an integer value for each default status. By doing so, I was able to get a numerical value for the correlation

between default status and the other 5 continuous variables. While no variable showed significant correlation with default status, the corr plot showed that interest rate and dti correlated the strongest with default status. KDE plots and box plots that were generated showed that those two variables were indeed the most correlated with default status. For categorical variables, grades significantly dropped in quality between default and paid loans, and the proportion of 60 month term loans experienced a significant jump for default loans.

For my multivariate analysis, I was interested to see how relationships between some variables changed across default and paid loans. Ultimately, I did not gain much insight here - relationships between variables like interest rate and installment were pretty consistent across default status. I also looked at how grade distributions across default status changed if you broke that down further by bringing in a third variable like employment length, term, home ownership and purpose. The only interesting observation I was made that renewable energy related loans that defaulted experience a relatively sharp increase in G rated loans, compared to other default loans.

I do not feel that many of these variables are strongly correlated with default rate. Interest rate and DTI are promising and worth investigating further. 60 month loans appears to have a notable increase among loans that defaulted. Even though we were able to see some slight differences in the default/paid loan means for annual income and loan amounts, I'm not convinced much weight needs to be put on those variables.

For the future, I would do four things:

1) I would try to incorporate loan issue date (issue_d) into my analysis, by transform the variable into an integer. By doing so, I could look at the distribution of default/paid off loans over time, and perhaps I could gain some interesting insights.

2) Utilize state data (addr_state) and see if there any states in particular that have a disproportionate amount of defaulted loans.

3) Incorporate the 600,000 present-day loans I removed into my analysis, as they might be able to find more variables strongly correlated with default rates.

4) Once I feel I have found enough variables that strongly correlate with default status, develop a linear model to predict default rates.