# Prosper Loan Data Analysis

*Mayukh Sarkar*

## Loading the packages

```
suppressMessages(library(devtools))
suppressMessages(library(ggplot2))
suppressMessages(library(ggthemes))
suppressMessages(library(dplyr))
suppressMessages(library(reshape2))
suppressMessages(library(memisc))
suppressMessages(library(gridExtra))
suppressMessages(library(RColorBrewer))
suppressMessages(library(magrittr))
suppressMessages(library(Kmisc))
suppressMessages(library(xtable))
suppressMessages(library(knitr))
suppressMessages(library(DT))
suppressMessages(library(scales))
suppressMessages(library(plotrix))
```

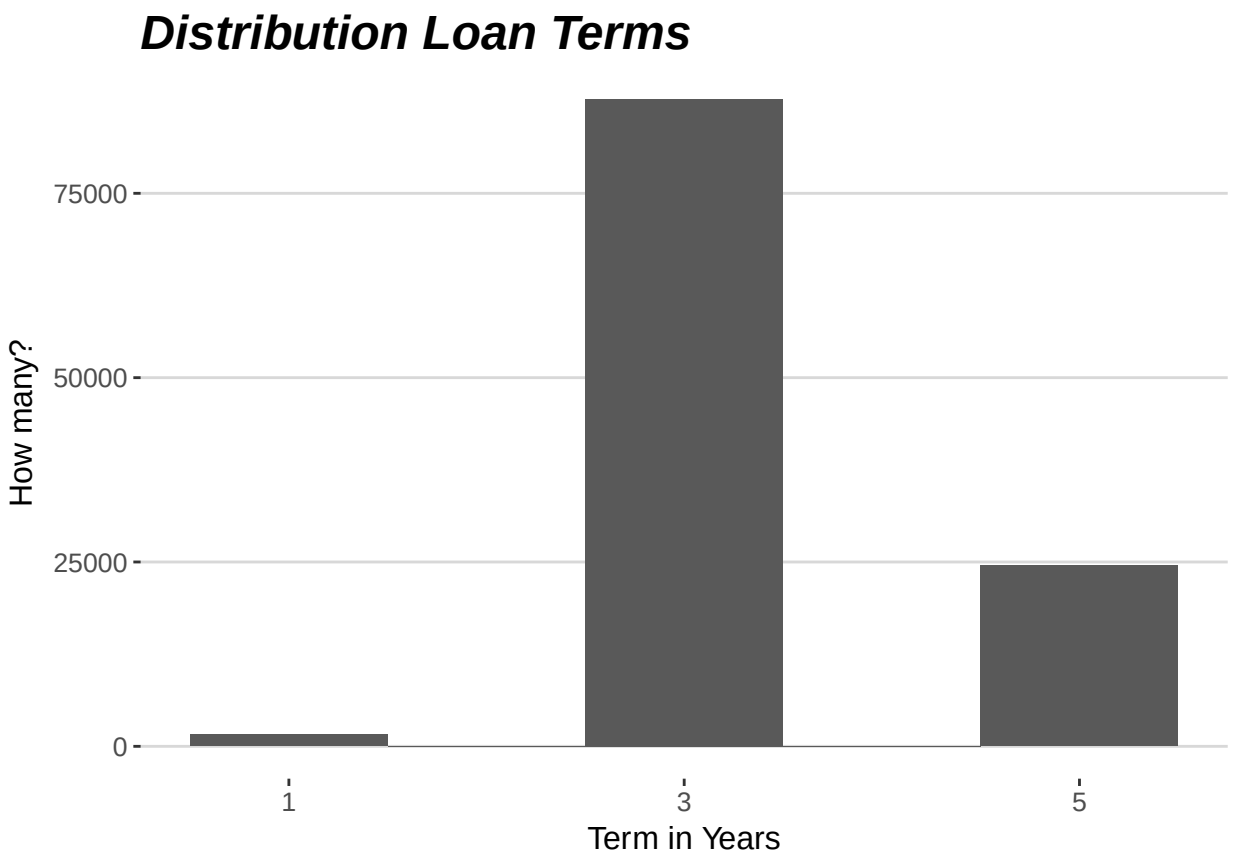
## Loading the dataset

```
loanData <- read.csv('prosperLoanData.csv')
```

## The first Qustion

The first question that needs to be asked is HOW LONG PEOPLE USUALLY OPT FOR LOAN? Let's answer this question with a histogram

```
loanData %>%
    ggplot(aes(x = Term / 12)) +
        geom_histogram(binwidth = 1) +
        theme_hc() +
        xlab('Term in Years') +
        ylab('How many?') +
        scale_x_continuous(breaks = seq(1, 5, 2)) +
        ggtitle("Distribution Loan Terms") +
        theme(plot.title = element_text(face = 'bold.italic',
                                        colour = "black", size=18))
```



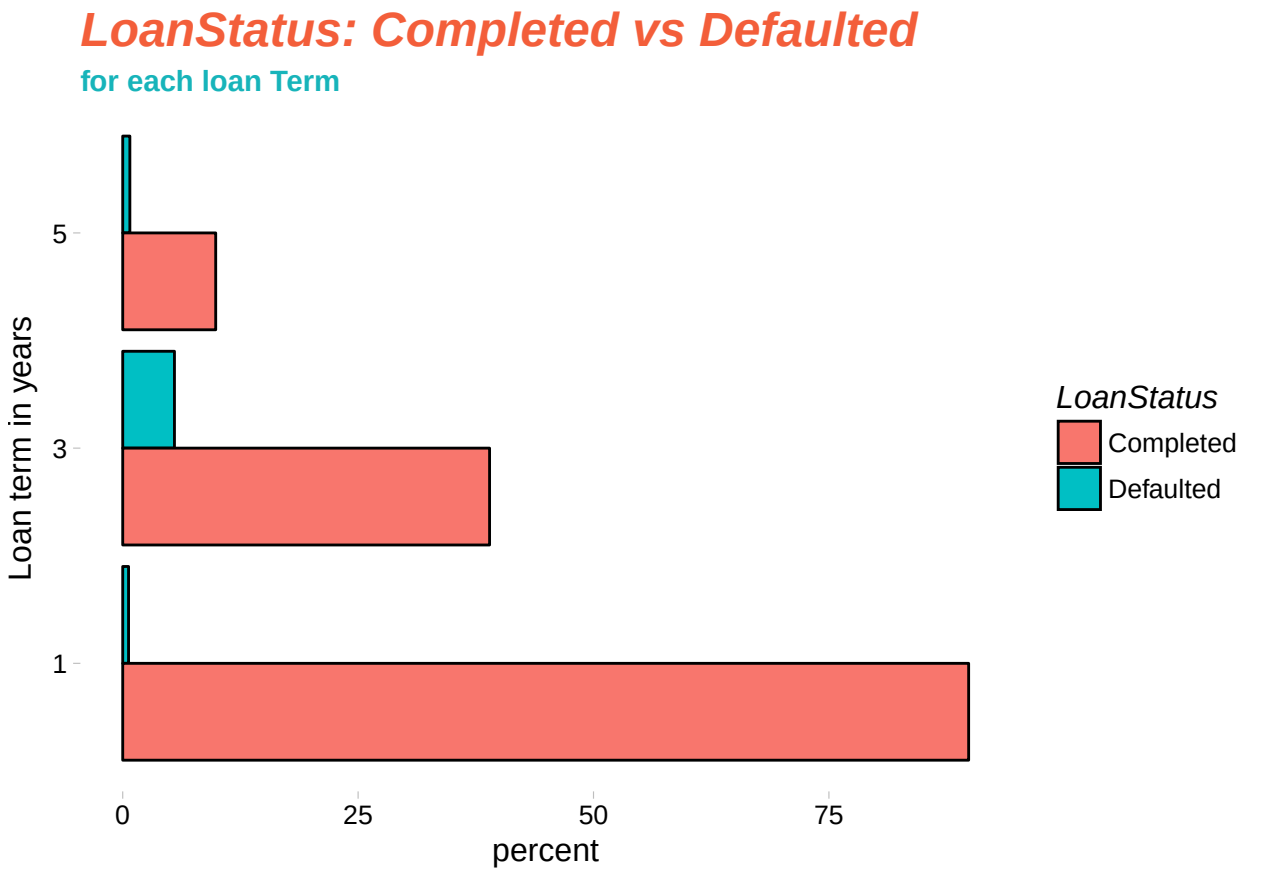

We can see that people don't really loan any amount for less than one year and the most popular loan amount is of 3 years although some people do choose for 5 years. Now lets assume something and check if it is correct or not. I assume that people who opted for 1 year would fail to repay their loan more as compared to the people who opted for 3 or 5 years. But is it true? So the next question is DOES PEOPLE REPAY THE LOANS BETTER WHEN THEY ARE GIVEN MORE TIME?

```
totalLoanTermes <- loanData %>%
    group_by(Term) %>%
    summarise(n = n()) %>%
    use_series(n)
totalLoanTermes <- rep(totalLoanTermes, each = 2)
loanData.two_status <- loanData %>%
    group_by(Term, LoanStatus) %>%
    summarise(n = n()) %>%
    filter(LoanStatus == 'Completed' | LoanStatus == 'Defaulted')
loanData.two_status$percent = (loanData.two_status$n / totalLoanTermes) * 100
```

Now let's plot the data

```
ggplot(aes(x = Term / 12, y = percent, fill = LoanStatus),
       data = loanData.two_status) +
    geom_bar(stat = 'identity', position="dodge", color = 'black') +
    scale_x_continuous(breaks = c(1, 3, 5)) +
    xlab('Loan term in years') +
    theme_pander() +
    coord_flip() +
    ggtitle("LoanStatus: Completed vs Defaulted",
            subtitle = "for each loan Term") +
```

```
    theme(plot.title = element_text(face = 'bold.italic', colour = '#F35E3A', size=18),
          plot.subtitle = element_text(face = 'bold', colour = '#17b4ba', size=11))
```




That's unusual because for **LoanStatus** of *Completed*, we see a trend but not in **LoanStatus** of *Defaulted* though. Now just because some customer's loan status is not Completed, doesn't mean his/her loan status is Defaulted. The reason we are exploring this is because we want to find if BANKS SHOULD FOCUS ON CUSTOMERS OPTING LOANS FOR SMALLER TERMS OR NOT? For this getting data of only two loan status won't be enough. So let's divide the customers into two groups

1. Good Customer
2. Bad Customer

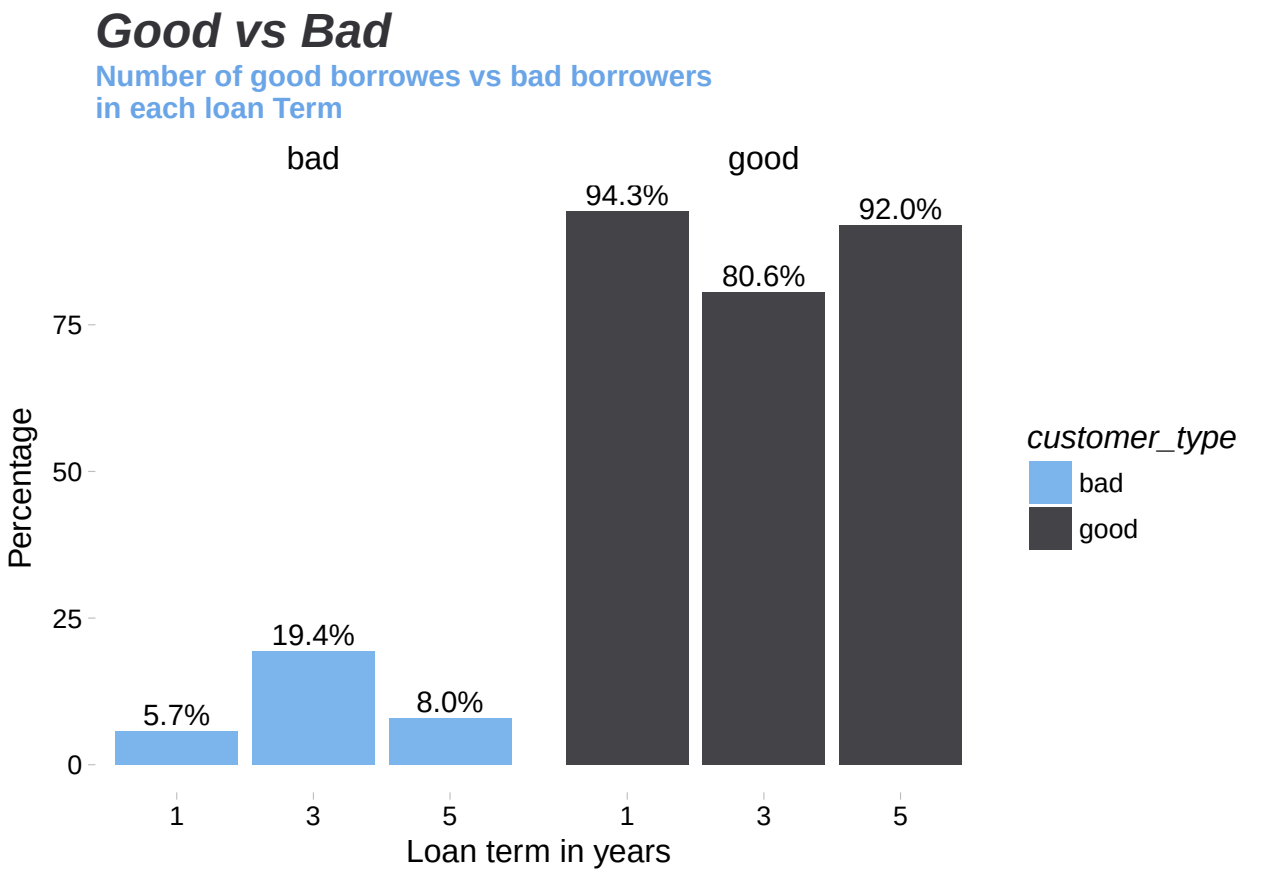
### The bigger picture

```
loanData.gb <- loanData %>%
    group_by(Term, LoanStatus) %>%
    summarise(n = n()) %>%
    mutate(customer_type = ifelse((LoanStatus == 'Current' |
                                   LoanStatus == 'Completed'|
                                   LoanStatus == 'FinalPaymentInProgress'),
                                  'good', 'bad')) %>%
    filter(LoanStatus != 'Cancelled') %>%
    mutate(freq = n / sum(n) * 100) %>%
    ungroup() %>%
    group_by(Term, customer_type) %>%
```

```
    summarise(n = sum(freq))
```

### Plotting the trend of different customer types

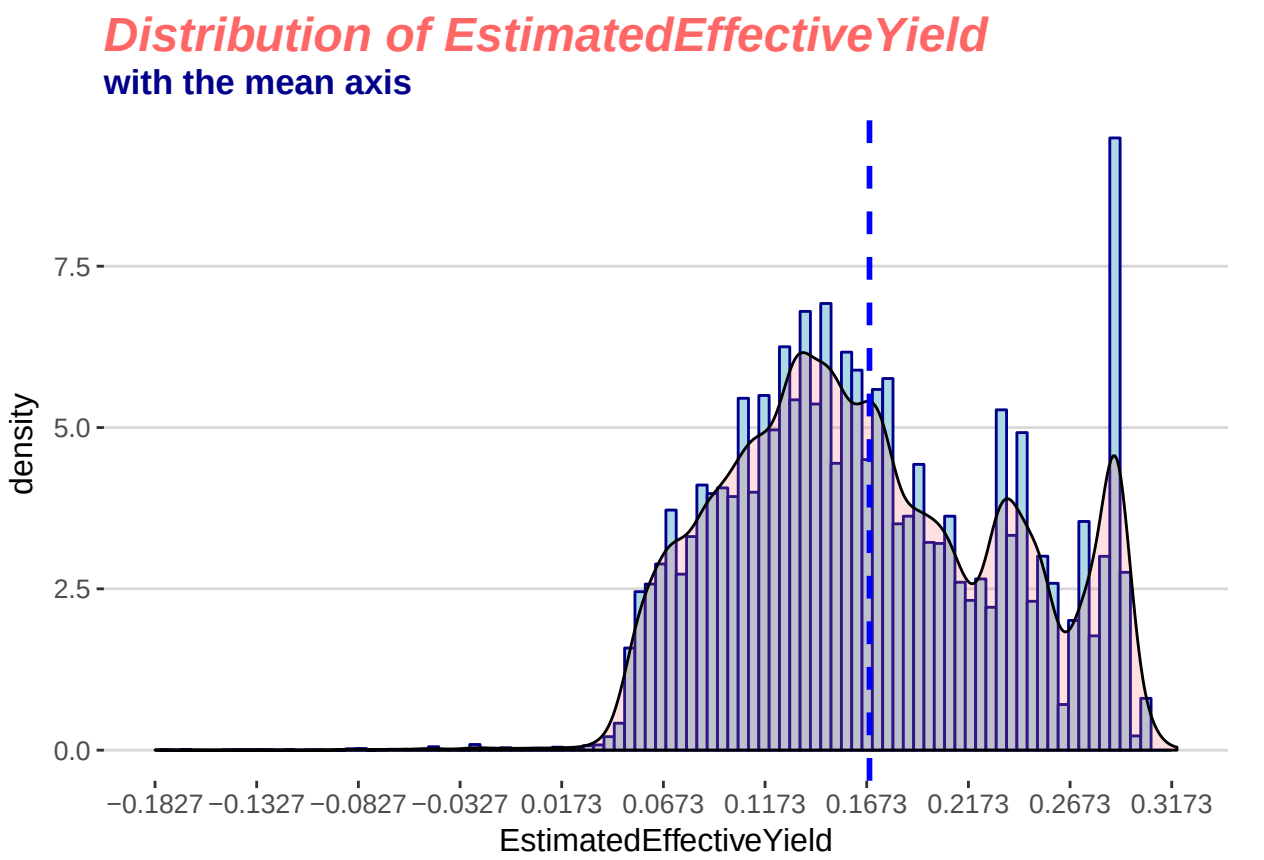
```
ggplot(aes(x = Term / 12, y = n, fill = customer_type), data = loanData.gb) +
    geom_bar(stat = 'identity', position="dodge") +
    theme_pander() +
    xlab('Loan term in years') +
    ylab('Percentage') +
    scale_x_continuous(breaks = c(1, 3, 5)) +
    geom_text(aes(label = sprintf("%2.1f%%", round(n, 2)), vjust = -.3), color="black") +
    facet_wrap(~customer_type) +
    scale_fill_hc() +
    ggtitle("Good vs Bad",
            subtitle = "Number of good borrowes vs bad borrowers
in each loan Term") +
    theme(plot.title = element_text(face = 'bold.italic', colour = '#333338', size=18),
          plot.subtitle = element_text(face = 'bold', colour = '#6aa5e7', size=11))
```




As we can that the short and long term prospects of the banks who issues the loan seems good because both in 1 and 5 year category, we see more number of good customers as compared to the 3 years. The number of good customers is however decreased a little from 1 year to 5 years but as compared to 3 years it is nothing. This may mean that banks should focus in long and short term customers as compared to medium term customers. But we should take this fact with a grain of slat because there might be other factors that may affect this trend. Only further exploration of the data would reveal that.

### EstimatedEffectiveYield - A better measure for a successful Lender

```
ggplot(aes(x = EstimatedEffectiveYield), data = loanData) +
    geom_histogram(aes(y = ..density..), bins = 100, na.rm = T,
                   color = 'darkblue', fill = 'lightblue') +
    theme_hc() +
    scale_x_continuous(limits = c(-0.1827, 0.3199),
                       breaks = seq(-0.1827, 0.3199, 0.05)) +
    geom_density(alpha=.2, fill="#FF6666", na.rm = T) +
    geom_vline(aes(xintercept=mean(EstimatedEffectiveYield, na.rm=T)),
               color="blue", linetype="dashed", size=1) +
    ggtitle("Distribution of EstimatedEffectiveYield",
            subtitle = "with the mean axis") +
    theme(plot.title = element_text(face = 'bold.italic', colour = '#FF6666', size=18),
          plot.subtitle = element_text(face = 'bold', colour = 'darkblue', size=13))
```



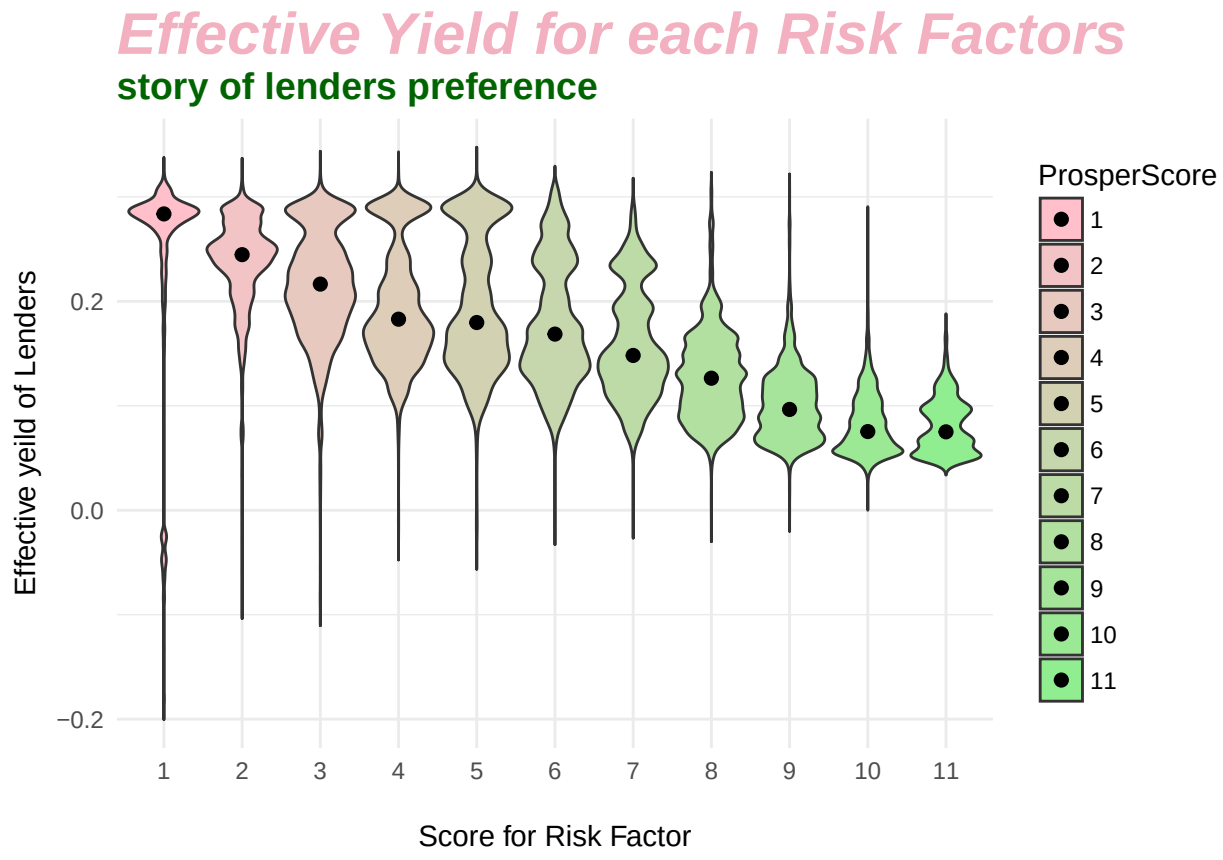

**EstimatedEffectiveYield** is said to be better estimate for the lenders than the interest rate because the interest includes *processing fees*, *uncollected interest due to borrower being chargedoff*. Plus it also doesn't include *late fines*. Hence EstimatedEffectiveYield takes account for all these things and it is thus a better measure. Above we are trying to see the distribution of the EstimatedEffectiveYield and we can see that it is multimodal. We see the most popular EstimatedEffectiveYield is around 0.3 while the mean is around 0.17 represented by the blue dotted line. The multimodal pattern shows that there are multiple EstimatedEffectiveYield that is popular. Strangely we can also see that the some customers have negative EstimatedEffectiveYield. This may mean a lot of things. This may mean that their BorrowerRate is a lot lower than their *service fee rate* or these customer's *uncollected interest on chargeoff* is lot more or they just never payed the late fee and payed back the loans along with the interest always on time.

### Does Lenders prefer borrowers with better Prosper Score ?

Now lets see what is the distribution of EstimatedEffectiveYield depending on the different **ProsperScore**. This is important because we want to answer a question, i.e., IF LENDERS GET MORE EstimatedEffectiveYield IF THEY HAVE BETTER ProsperScore ?

We are using violin plot instead of box plot for this.

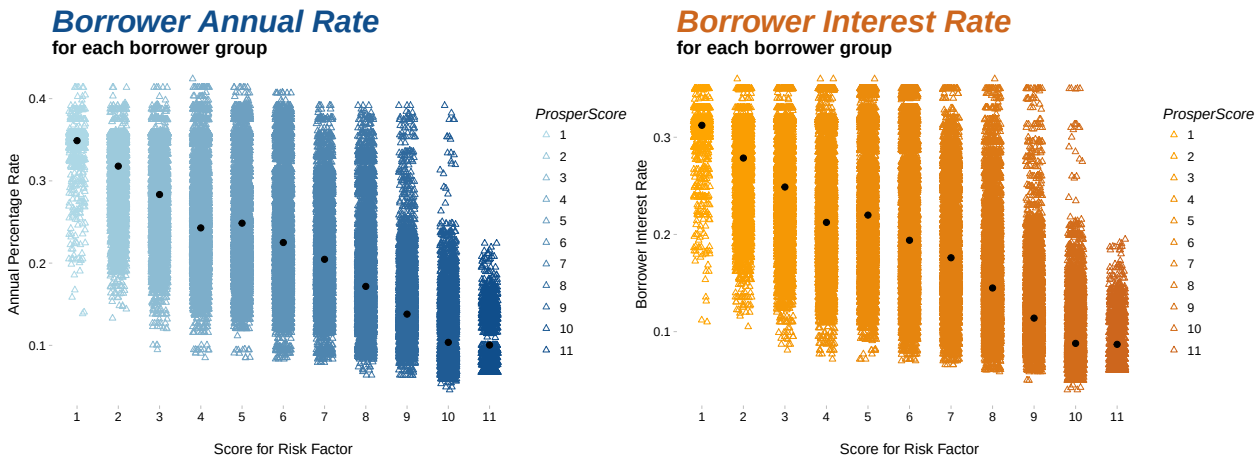
```
loanData$ProsperScore <- factor(loanData$ProsperScore)
ggplot(aes(x = ProsperScore, y = EstimatedEffectiveYield, fill=ProsperScore),
       data = subset(loanData, !is.na(loanData$ProsperScore) &
                             !is.na(loanData$EstimatedEffectiveYield))) +
    geom_violin(trim = F, scale = "width") +
    stat_summary(fun.y=median, geom="point", size=2, color="black") +
    scale_fill_manual(values=colorRampPalette(c("pink", "lightgreen"))(11)) +
    theme_minimal() +
    xlab('\nScore for Risk Factor') +
    ylab('Effective yeild of Lenders') +
    ggtitle("Effective Yield for each Risk Factors",
            subtitle = "story of lenders preference") +
    theme(plot.title = element_text(face = 'bold.italic', colour = '#f3b0c0', size=22),
          plot.subtitle = element_text(face = 'bold', colour = 'darkgreen', size=14))
```




Well this is interesting. We see an wonderful trend here. Here more score for the risk factor means better the borrower and lesser score for risk factor means poor prospects from the borrowers. We can see that for lower ProsperScore distribution of effective yield in a lot more than the higher ProsperScore. This may mean that lenders charges a variety of interest rate from the borrower with poor prospects as compared to borrowers with better prospect. We can also notice how median (represented by the black dot) is decreasing

as ProsperScore is increasing. This may mean that lenders give more relaxations to borrowers with better ratings as compared to borrowers with poor rating. Does that mean lenders trust and like borrowers with better ProsperScore! Let's do a little more analysis to reveal it more. The reason we need more exploration on this is because EstimatedEffectiveYield includes more things such as late fine and doesn't include processing fee and others. So more EstimatedEffectiveYield for lesser ProsperScore borrowers may be due to high late fines because lesser ProsperScore borrowers are more prone to fail to repay their loan on time each month. So, Let's see if borrower's interest rate shows the same trend for each ProsperScore categories or not because interest rates doesn't include late fines.

```
borrower_apr <- ggplot(aes(x = ProsperScore, y = BorrowerAPR, color=ProsperScore),
       data = subset(loanData, !is.na(loanData$ProsperScore) &
                             !is.na(loanData$BorrowerAPR))) +
    geom_jitter(shape=2, position=position_jitter(0.2)) +
    stat_summary(fun.y=median, geom="point", size=2, color="black") +
    scale_color_manual(values=colorRampPalette(c("lightblue", "dodgerblue4"))(11)) +
    theme_pander() +
    xlab('\nScore for Risk Factor') +
    ylab('Annual Percentage Rate') +
    ggtitle("Borrower Annual Rate",
            subtitle = "for each borrower group") +
    theme(plot.title = element_text(face = 'bold.italic', colour = 'dodgerblue4', size=25),
          plot.subtitle = element_text(face = 'bold', colour = 'black', size=15))

borrower_rate <- ggplot(aes(x = ProsperScore, y = BorrowerRate, color=ProsperScore),
       data = subset(loanData, !is.na(loanData$ProsperScore) &
                             !is.na(loanData$BorrowerRate))) +
    geom_jitter(shape=2, position=position_jitter(0.2)) +
    stat_summary(fun.y=median, geom="point", size=2, color="black") +
    scale_color_manual(values=colorRampPalette(c("orange", "chocolate3"))(11)) +
    theme_pander() +
    xlab('\nScore for Risk Factor') +
    ylab('Borrower Interest Rate') +
    ggtitle("Borrower Interest Rate",
            subtitle = "for each borrower group") +
    theme(plot.title = element_text(face = 'bold.italic', colour = 'chocolate3', size=25),
          plot.subtitle = element_text(face = 'bold', colour = 'black', size=15))

grid.arrange(borrower_apr, borrower_rate, nrow = 1, ncol = 2)
```



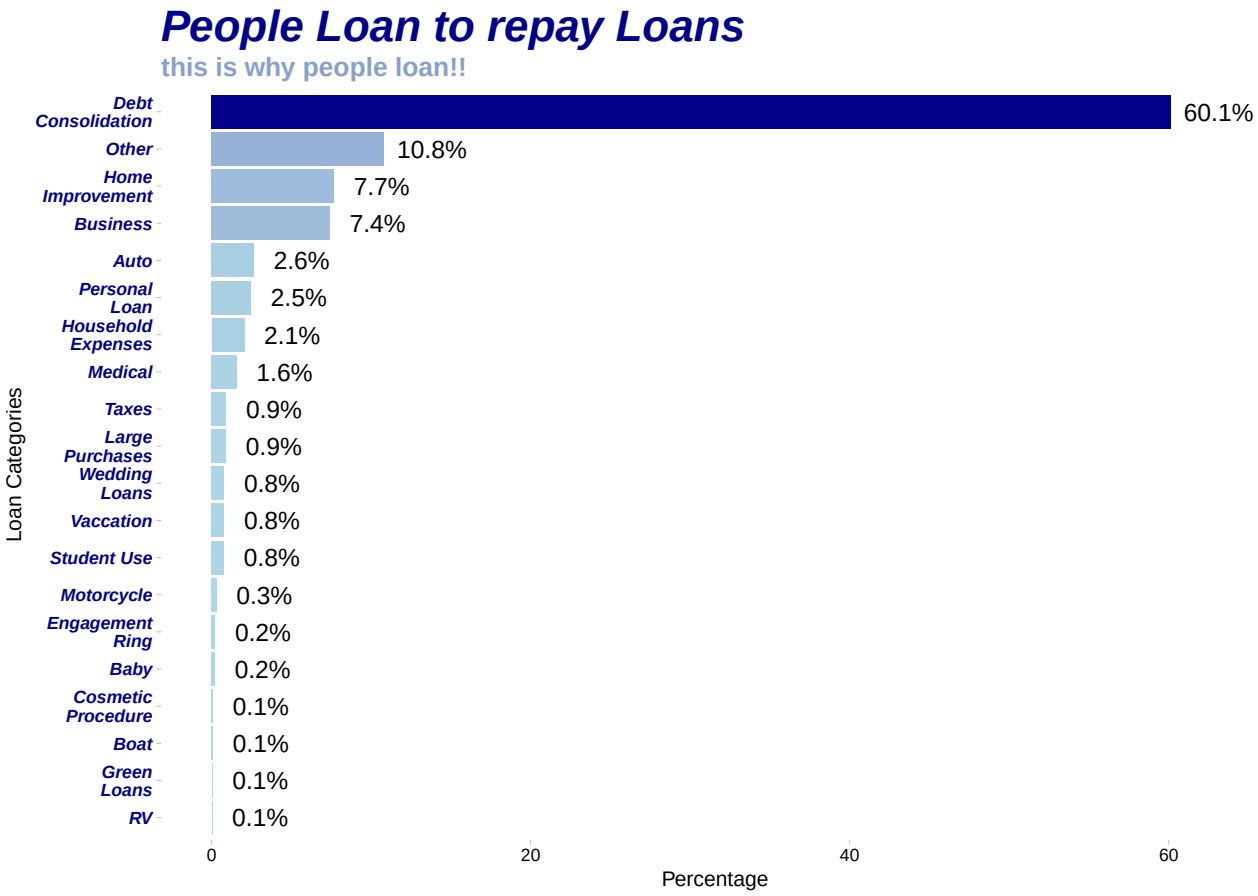

Finally things are revealed much better now! We can clearly observe that for both *BorrowerAPR* and

*BorrowerRate* which are metric for interest rates, we see a declining trend as the *ProsperScore* is increasing. This justifies the fact even more that lenders somehow prefers to charge less for all the borrowers with better ProsperScore as compared to borrowers with inferior ProsperScore. Here food for thought is that is it moral too? Well I guess no amount of EDA can reveal it.

But are we missing the real question here? We are performing all these analysis on the Prosper loan data to answer several questions but what is the most important thing for a loan ? It's **BORROWERS** and what is the most important question related to borrowers?

## What people want loans for?

```
categories <- c("Debt\nConsolidation", "Home\nImprovement", "Business",
                "Personal\nLoan", "Student Use", "Auto",
                "Other", "Baby", "Boat",
                "Cosmetic\nProcedure", "Engagement\nRing", "Green\nLoans",
                "Household\nExpenses", "Large\nPurchases", "Medical",
                "Motorcycle", "RV", "Taxes", "Vaccation", "Wedding\nLoans")
mapBorrowerCategory <- function(categoryNumber) {
    ifelse(categoryNumber == 0, 'na', categories[categoryNumber])
}

loanData.reasons <- loanData %>%
    group_by(ListingCategory..numeric.) %>%
    summarise(n = n()) %>%
    filter(ListingCategory..numeric. != 0) %>%
    mutate(category = mapBorrowerCategory(ListingCategory..numeric.),
           freq = n / sum(n) * 100) %>%
    arrange(n) %>%
    dplyr::select(-ListingCategory..numeric.)

loanData.reasons$category <- factor(loanData.reasons$category)
ggplot(aes(x = reorder(category, freq), y = freq, fill = freq), data = loanData.reasons) +
    geom_bar(stat = 'identity', position="dodge") +
    theme_pander() +
    scale_fill_gradient(low = 'lightblue', high = 'darkblue') +
    coord_flip() +
    guides(fill=FALSE) +
    geom_text(aes(label = sprintf("%2.1f%%", round(freq, 2))),
              color="black", size = 5, nudge_y = 3) +
    ylab('Percentage') +
    xlab('Loan Categories') +
    theme(axis.text.y = element_text(face = 'bold.italic', colour = 'darkblue')) +
    ggtitle("People Loan to repay Loans",
            subtitle = "this is why people loan!!") +
    theme(plot.title = element_text(face = 'bold.italic', colour = 'darkblue', size=25),
          plot.subtitle = element_text(face = 'bold', colour = '#87a0cc', size=15))
```




## Loan for Loans !!

It is really strange that people took loans to reimburse loans/debts and lenders also issued loans to these people. Isn't is a bad idea to take laons to pay for loans? Aren't we falling into a vicious cycle? We can also see that Home Inprovement is above Medical which may mean well insured population or negligence toward health. The other section with more than 10% of the loans would remain unknown. Why the loan reason would be unknown? Does it indicate any illegal reason to opt for loan! Let's explore this more. Let's see the Occupation and Employment status of Borrowers for all the loans falling into the *others* section.

## Who are others??

```
other.borrowers <- loanData %>%
    filter(ListingCategory..numeric. == 7) %>%
    group_by(EmploymentStatus, Occupation) %>%
    summarise(n = n()) %>%
    filter(!is.na(Occupation)) %>%
    ungroup() %>%
    arrange(EmploymentStatus, desc(n)) %>%
    ungroup() %>%
    group_by(EmploymentStatus) %>%
    top_n(n = 5, wt = n) %>%
    ungroup()
```

Table 1: Employed

| Occupation | n |
|---|---|
| Other | 1579 |
| Professional | 789 |
| Administrative Assistant | 280 |
| Teacher | 256 |
| Computer Programmer | 247 |

Table 2: Full-time

| Occupation | n |
|---|---|
| Other | 622 |
| Professional | 326 |
| Computer Programmer | 139 |
| Teacher | 125 |
| Administrative Assistant | 115 |

Table 3: Part-time

| Occupation | n |
|---|---|
| Other | 23 |
| Sales - Retail | 13 |
| Administrative Assistant | 8 |
| Food Service | 7 |
| Nurse (RN) | 4 |
| Student - College Graduate Student | 4 |
| Student - College Senior | 4 |

Table 4: Self-employed

| Occupation | n |
|---|---|
| Other | 152 |
| Professional | 32 |
| Investor | 19 |
| Construction | 16 |
| Sales - Commission | 16 |

Table 5: Others

| EmploymentStatus | Occupation | n |
|---|---|---|
| Other | Other | 352 |
| Other | | 95 |
| Retired | Other | 127 |
| Retired | Engineer - Electrical | 2 |

We can see some really strange things happening here. In fact there are so many strange things that I have to list them down

**Firstly** what is the difference between *Employed* and *Full-time*?

**Secondly** there are too many others. The complete sample of the data is taken for those borrowers whose purpose for borrowing is *Others*. But there are *Others* for *Occupation* and *EmploymentStatus* too.

**Thirdly** how can someone be *Retired* but still has *Occupation Others* ?

**Lastly** there are some dubious borrowes in the *Others* table. We can see that there are 352 cases where the borrowing reason, EmploymentStatus, Occupation all are *Others* and 95 cases where *Occupation* is left blank.

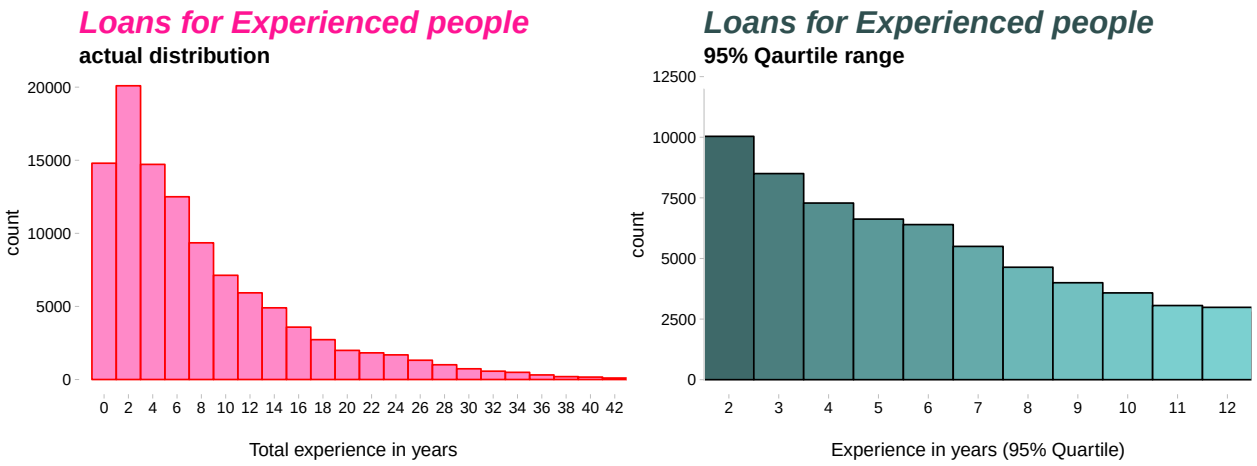
These above points, specially the last one says not all the bowrrower's information is revealed or in some cases they seemed to be fake too. I guess it is a partial dead end for us to see why people mention *Other* as a reason while opting for loan.

Let's now move into something more computational. Let's first see after working for how many yaers, people like to opt for a loan? We can then move into some deeper level of analysis.

## Does experienced people opt for loan lesser ?

```
borrower.experience <- loanData %>%
    filter(!is.na(EmploymentStatusDuration))

hist1 <- ggplot(aes(x = EmploymentStatusDuration / 12), data = borrower.experience) +
         geom_histogram(binwidth = 2, color = 'red', fill = 'deeppink', alpha = 1/2) +
```

```
         theme_pander() +
         scale_x_continuous(breaks =
                                seq(min(borrower.experience$EmploymentStatusDuration),
                                    max(borrower.experience$EmploymentStatusDuration),
                                        2)) +
         coord_cartesian(xlim = c(0, 41)) +
         xlab('\nTotal experience in years') +
         ggtitle("Loans for Experienced people",
                 subtitle = "actual distribution") +
        theme(plot.title = element_text(face = 'bold.italic', colour = 'deeppink', size=20),
          plot.subtitle = element_text(face = 'bold', colour = 'black', size=14))

hist2 <- ggplot(aes(x = EmploymentStatusDuration / 12), data = borrower.experience) +
         geom_histogram(binwidth = 1, color = 'black', aes(fill = ..count..)) +
         theme_pander() +
         coord_cartesian(xlim = c(2, 12)) +
         scale_x_continuous(breaks = seq(2, 12, 1)) +
         xlab('\nExperience in years (95% Quartile)') +
         scale_fill_gradient("Count", low = "darkslategray1", high = "darkslategrey") +
         guides(fill=FALSE) +
         ggtitle("Loans for Experienced people",
                 subtitle = "95% Qaurtile range") +
        theme(plot.title = element_text(face = 'bold.italic', colour = 'darkslategrey', size=20),
          plot.subtitle = element_text(face = 'bold', colour = 'black', size=14))
grid.arrange(hist1, hist2, ncol = 2, nrow = 1)
```



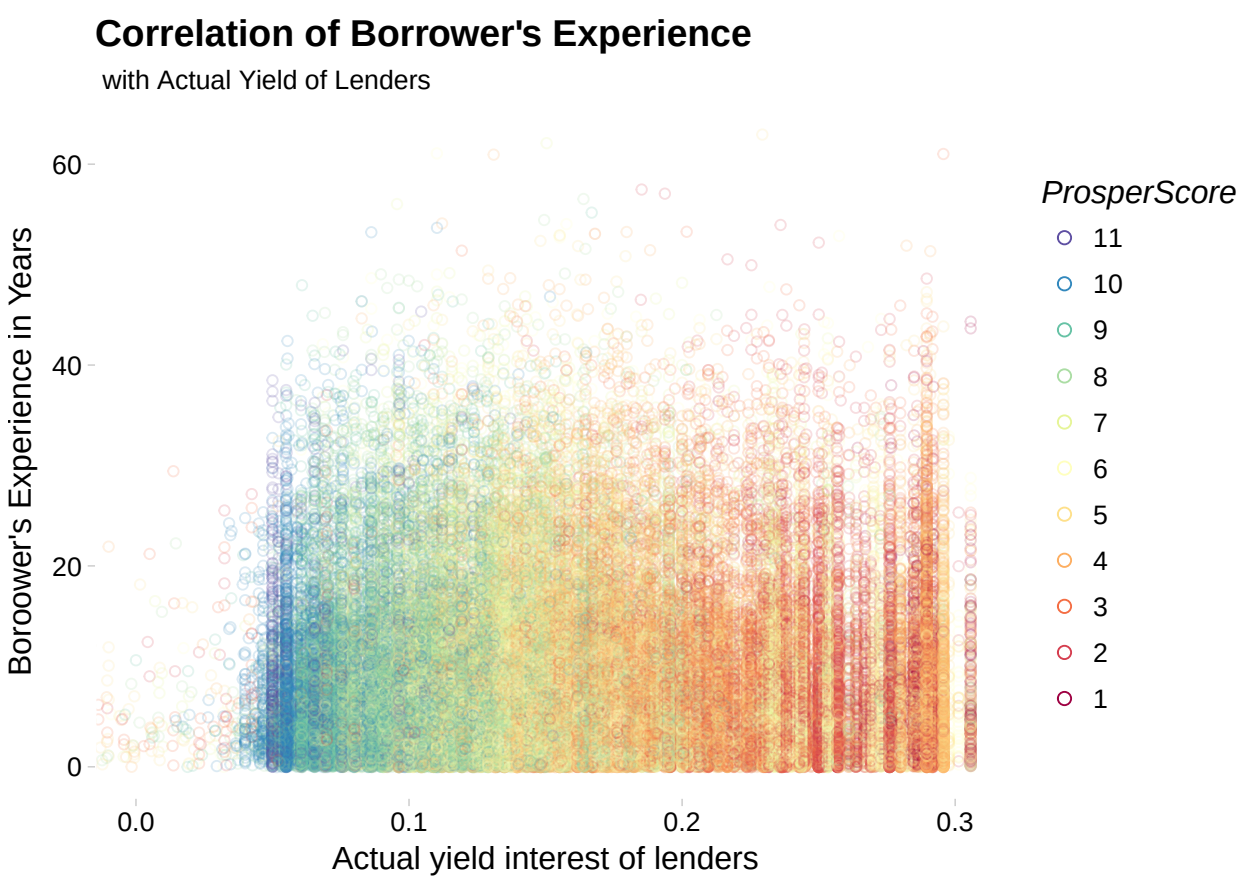

## I was right ..:)

Well this was really intersting. We can see that our assumption was overall correct because the as people gain experience in their jobs, lesser they opt for the loans. This may be due the fact that as people gain experinece their salary also increases and hence the lesser reason they find to opt for loans or the reason can be something different. We can also see in the right histogram of 95% quartile, most people opt for loans when they have almost 2 years of experience.

Lets explore it even more let's see the correlation between number of month a person has expericence and the EstimatedEffectiveYield variable that we have explored earlier. The question that we want to explore - DOES LENDERS ASK FOR LESS INTEREST FORM THE BORROWERS WHO ARE MORE EXPERIENCED? This can be true because people with more job experience should have more potential to replay their loan

better because they have higher paying jobs and hence their ProsperScore would be higher. And as we have seen that borrowers with better prosper score pay lesser to the lenders and lenders somehow prefer them.

```
ggplot(aes(y = EmploymentStatusDuration / 12,
           x = EstimatedEffectiveYield,
           color = ProsperScore),
       data = subset(loanData, !is.na(loanData$ProsperScore))) +
    geom_point(na.rm = T, position = 'jitter', alpha = 1 / 6, shape = 1) +
    scale_color_brewer(type = 'div', palette = 'Spectral', direction = 1,
                       guide = guide_legend(title = 'ProsperScore', reverse = T,
                                            override.aes = list(alpha = 1, size = 2))) +
    coord_cartesian(xlim = c(0.0, 0.3)) +
    xlab('Actual yield interest of lenders') +
    ylab('Boroower\'s Experience in Years') +
    theme_pander() +
    ggtitle("Correlation of Borrower's Experience",
            subtitle = " with Actual Yield of Lenders")
```




As it seems from the scatter plot that the pattern seems to have no correlation. It means our assumption was not correct. Borrowers with better **EmploymentStatusDuration** don't seem to get any special relaxation from lenders in terms of interest each month. This can be further confirmed by checking the *Perarson's correlation Coefficient*. We see that the correlation coefficient is almost 0 (-0.0233).

```
with(subset(loanData, !is.null(EmploymentStatusDuration) &
                !is.null(EstimatedEffectiveYield)),
     cor.test(EmploymentStatusDuration / 12, EstimatedEffectiveYield))
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  EmploymentStatusDuration/12 and EstimatedEffectiveYield
## t = -6.7924, df = 84832, p-value = 1.11e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03003902 -0.01658791
## sample estimates:
##         cor
## -0.02331452
```

This also says that even though the true correlation is not ture and alternative hypothesis is accepted, their is some **serious statistical evidence of significance**. But if we look into the CI, it is within the range of -0.03 to -0.016 which is very small. Good R value is said a value $< -0.3$ or value $> 0.3$. This value is definitely not that large. Judging from the context atlest it is not. So we can say that there is **no practical significance.** Hence we can not tell with any confirmation that More Exprerienced Lenders end up paying Less/More interest to the Lenders.

## How people loan for their Homes ?

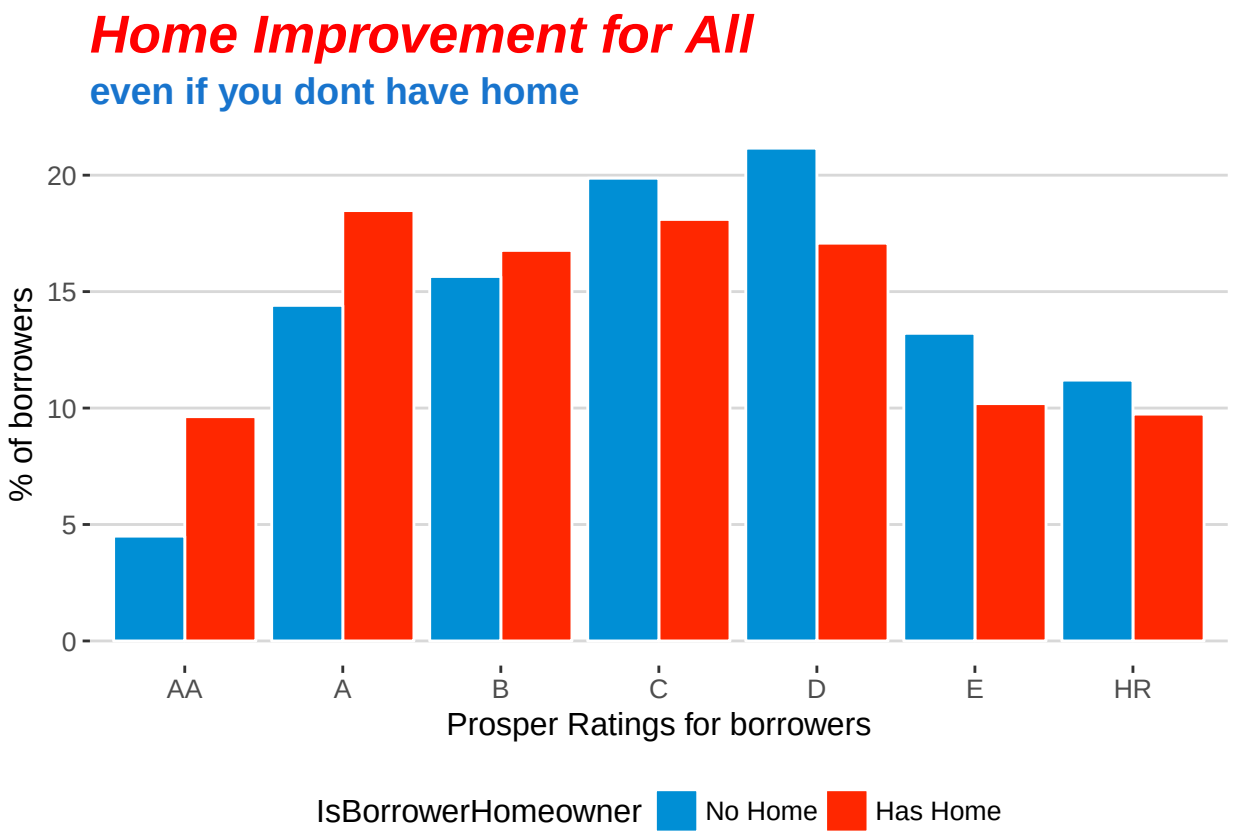
Here we are going to explore the people for two category.

1. First those who are opting for loan to rennovation of home when they have a house.
2. Second those who opt for home loans even though they don't have house.

```
home.loans <- loanData %>%
    filter(ListingCategory..numeric. == 2,
           ProsperRating..Alpha. != "",
           IsBorrowerHomeowner != "") %>%
    group_by(IsBorrowerHomeowner, ProsperRating..Alpha.) %>%
    summarise(n = n()) %>%
    mutate(freq = round(n / sum(n) * 100, 2))
levels(home.loans$IsBorrowerHomeowner) <- c('No Home', 'Has Home')
home.loans$ProsperRating..Alpha. <-
    factor(home.loans$ProsperRating..Alpha., ordered = T,
           levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR'))
home.loans <- arrange(home.loans, ProsperRating..Alpha.)
```

### Now let's plot the data

```
ggplot(aes(x = ProsperRating..Alpha., y = freq,
           fill = IsBorrowerHomeowner), data = home.loans) +
    geom_bar(stat = 'identity', position="dodge",
             color = 'white') +
    theme_hc() +
    scale_fill_fivethirtyeight() +
    xlab('Prosper Ratings for borrowers') +
    ylab('% of borrowers') +
    ggtitle("Home Improvement for All", subtitle = "even if you dont have home") +
    theme(plot.title = element_text(face = 'bold.italic', colour = 'red', size=20),
          plot.subtitle = element_text(face = 'bold', colour = 'dodgerblue3', size=14))
```




### Are Investors partial to Borrowers with better rating ??

Here we visualizing an interesting trend. We are seeing for people with ProsperRating of *AA, A & B*, number of people opted for loan mentioning the reason for laon as **Home Improvement** but still has no home is more than the people who has home and opted for loan for Home Improvement and truly a home owner. But for people with poor rating, the trend in opposite and expected also. People who don't have any house should not get any loan mentioning the loan reason of House Improvement. These whole thing shows not only lenders give much preference to Rating over Verification and KYC (Know Your Customer) but this also shows irrespective of rating of borrowers, lenders dont care much about laon reason as a whole. The following code proves it even more. We can see that there are not a single borrowers who mentioned their loan purpose to be Home Improvement when they didn't have their own house and their loan was not approved. That's a strong evidence to show that investors don't care much about how much borrowers fudge or fake their loan purpose.

```
a <- loanData %>%
    filter(ListingCategory..numeric. == 2,
           LoanStatus == 'Cancelled')
print.numeric_version(dim(a)[0])
```

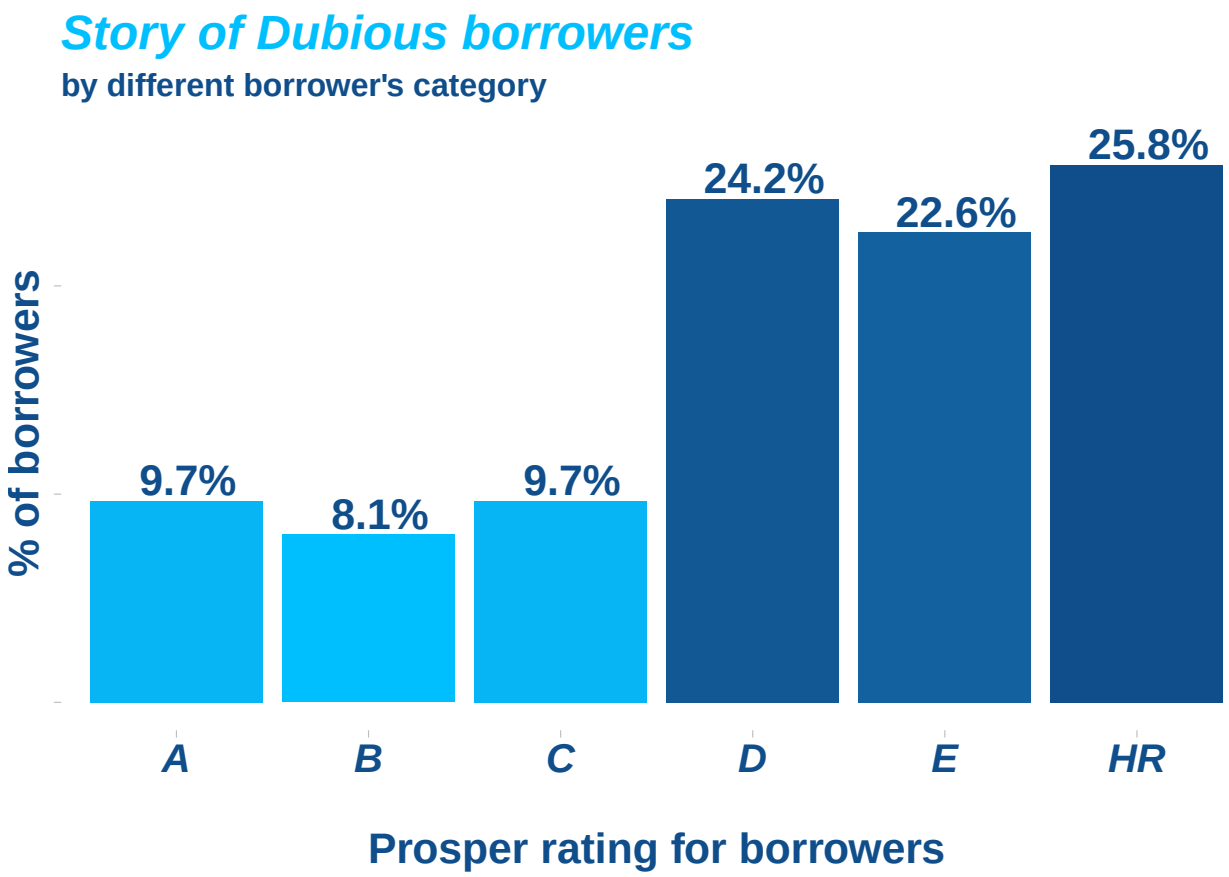
```
## <0 elements>
```

### What's up with the people having dubious income sources ?

```
dubious.borrowers <- loanData %>%
    filter(ListingCategory..numeric. == 7,
```

```
           Occupation == 'Other',
           EmploymentStatus == 'Other',
           IncomeVerifiable == 'False') %>%
    group_by(ProsperRating..Alpha.) %>%
    summarise(n = n()) %>%
    mutate(freq = round(n / sum(n) * 100, 2))
dubious.borrowers$ProsperRating..Alpha. <-
    ordered(dubious.borrowers$ProsperRating..Alpha.,
            levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR'))
dubious.borrowers <- arrange(dubious.borrowers, ProsperRating..Alpha.)
```

```
ggplot(aes(x = ProsperRating..Alpha., y = freq, fill = freq),
       data = dubious.borrowers) +
    geom_bar(stat = 'identity', position="dodge") +
    theme_pander() +
    scale_fill_continuous(low = 'deepskyblue1',
                          high = 'dodgerblue4') +
    guides(fill=FALSE) +
    xlab('\nProsper rating for borrowers') +
    ylab('% of borrowers') +
    geom_text(aes(label = sprintf("%2.1f%%", round(freq, 2))),
              color="dodgerblue4", size = 5.5, nudge_y = 1,
              fontface = 'bold', nudge_x = 0.06) +
    theme(axis.text.x = element_text(face = 'bold.italic',
                                     colour = 'dodgerblue4',
                                     size = 15),
          axis.text.y=element_blank(),
          axis.title.x = element_text(face = 'bold',
                                      colour = 'dodgerblue4',
                                      size = 16),
          axis.title.y = element_text(face = 'bold',
                                      colour = 'dodgerblue4',
                                      size = 16),
          plot.title = element_text(face = 'bold.italic', colour = 'deepskyblue1', size=18),
          plot.subtitle = element_text(face = 'bold', colour = 'dodgerblue4', size=12)) +
    ggtitle("Story of Dubious borrowers", subtitle = "by different borrower's category")
```




### Prosper rating system seems legit !!

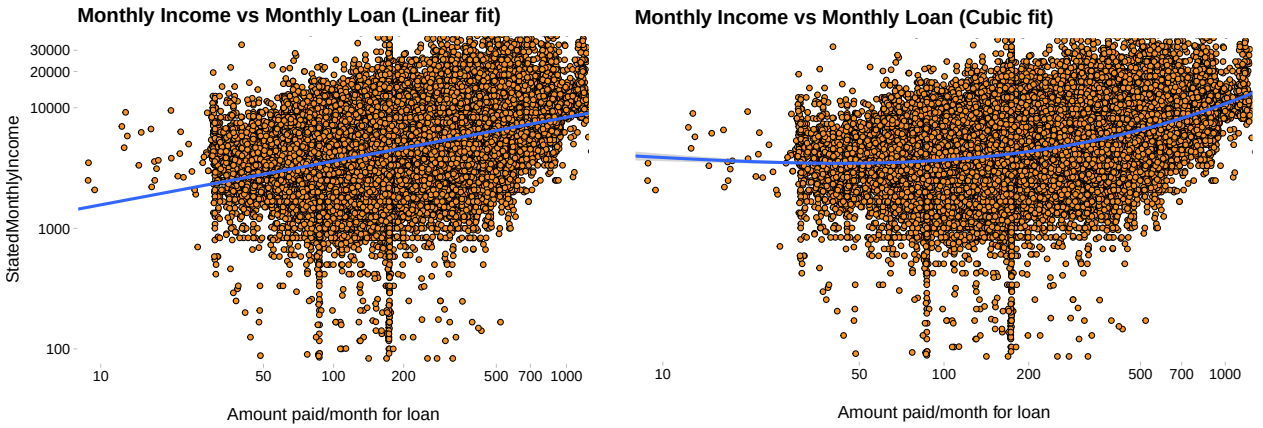
Although we can't see any specific pattern among the dubious borrowers but we can see one thing very well. Overall borrowers with better ratings tends to be less dubious than borrowers with poor ratings. One more interesting thing is that the above barchart does not inlude borrowers with $AA$ rating. It means that the best borrowers are really best and the prosper rating really works.

### Are Lenders greedy ?

I always wanted to do this. But from this dataset we can answer this question a lot of way. One of the way is to check if the lenders asked for money if the borrowers income was high. Let's see if the correlation is substantial.

```
base <- loanData %>%
    filter(StatedMonthlyIncome != 0.0,
           MonthlyLoanPayment != 0.0,
           ProsperRating..Alpha. != "") %>%
    ggplot(aes(y = StatedMonthlyIncome,
               x = MonthlyLoanPayment)) +
    geom_jitter(color = 'black',
                fill = '#F79420', shape = 21) +
    coord_cartesian(ylim = c(100, 30000), xlim = c(10, 1000)) +
    scale_y_continuous(trans = log10_trans(),
                       breaks = c(100, 1000, 10000, 20000, 30000)) +
```

```
    scale_x_continuous(trans = log10_trans(),
                       breaks = c(10, 50, 100, 200, 500, 700, 1000)) +
    theme_pander()
scatter1 <- base + geom_smooth(span = 0.3, method = 'lm') +
    theme(plot.margin=unit(c(0, 1, 0, 0), "cm")) +
    xlab('\nAmount paid/month for loan') +
    ggtitle("Monthly Income vs Monthly Loan (Linear fit)")
scatter2 <- base + geom_smooth(span = 0.3,
                               method = 'lm',
                               formula = y ~ splines::bs(x, 3)) +
    theme(axis.title.y=element_blank(),
          axis.text.y=element_blank(),
          axis.ticks.y=element_blank()) +
    xlab('\nAmount paid/month for loan') +
    ggtitle("Monthly Income vs Monthly Loan (Cubic fit)")
grid.arrange(scatter1, scatter2, nrow = 1, ncol = 2)
```




Here I am trying to see the relationship of *MonthlyLoanPayment* with *StatedMonthlyIncome.* We can see from the above statterplot that most of the *MonthlyLoanPayment* is distribued from 10 to 1000 and the *StatedMonthlyIncome* is distributed from 100 to 30000. Both the scales are transformed in log scale and we can clearly observe two things, namely,

```
1. There is defineite a strong positive correlation
between monthly income and monthly loan amount.

2. Cubic relationship seems a better fit that simple linear model.
```

## Are we sure that they are greedy ?

Now we can see that there was definitely a strong correlation between the two variables but are we sure? Let's find the **Correlation Coefficient** to analyse it more.

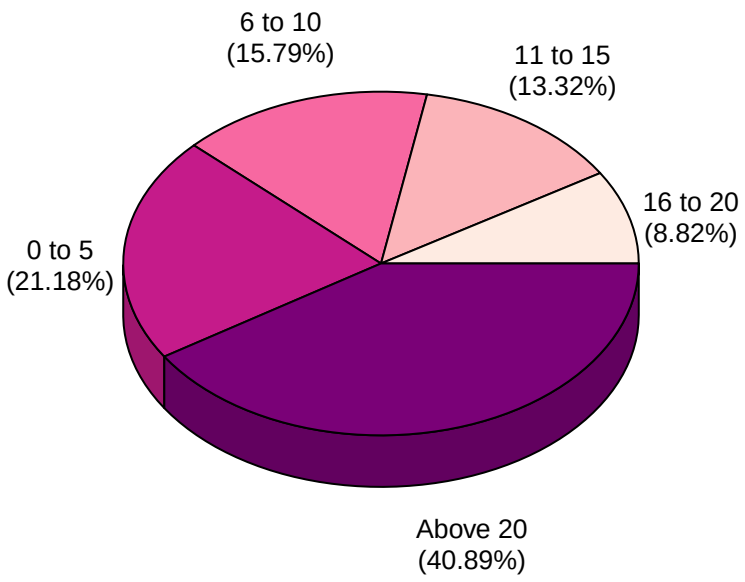
```
with(loanData, cor.test(MonthlyLoanPayment, StatedMonthlyIncome, method = "pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  MonthlyLoanPayment and StatedMonthlyIncome
## t = 67.764, df = 113940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.1912423 0.2024055
## sample estimates:
##       cor
## 0.1968303
```

Well we can't really say that there is a strong correlation looking at the value of R which is almost 0.2. Usually it is said to be of high statistical importance if it is more than 0.3 or less than -0.3. But we can see that the value is still accpeptable with somewhat positive correlation with the population Conficence Interval being more than 0. The strong t-statistics of 67.76 and small p-value shows that the statistical significance of alternative hypothesis is very strong. But this can also be somewhat practical significance too, atleast to be included in a linear model.

```
rec <- loanData %>%
    group_by(Recommendations) %>%
    summarise(n = n(),
              AvgLoan = median(MonthlyLoanPayment)) %>%
    mutate(freq = AvgLoan / sum(AvgLoan) * 100)


A <- cut(rec$Recommendations, breaks = c(-1, 5, 10, 15, 20, 40),
    labels= c("0 to 5", "6 to 10", "11 to 15", "16 to 20", "Above 20"))
rec$recrange <- A

recommend <- rec %>%
    group_by(recrange) %>%
    summarise(total = round(sum(freq), 2))
k <- recommend[order(recommend$total), ]
plotrix::pie3D(sort(recommend$total),
               labels = sprintf("%s\n(%s%%)", k$recrange, sort(recommend$total)),
      col=brewer.pal(5, "RdPu"),
      main = "Monthly Loan for differnt recommendations",
      theta = pi / 3, labelcex = 0.8)
```

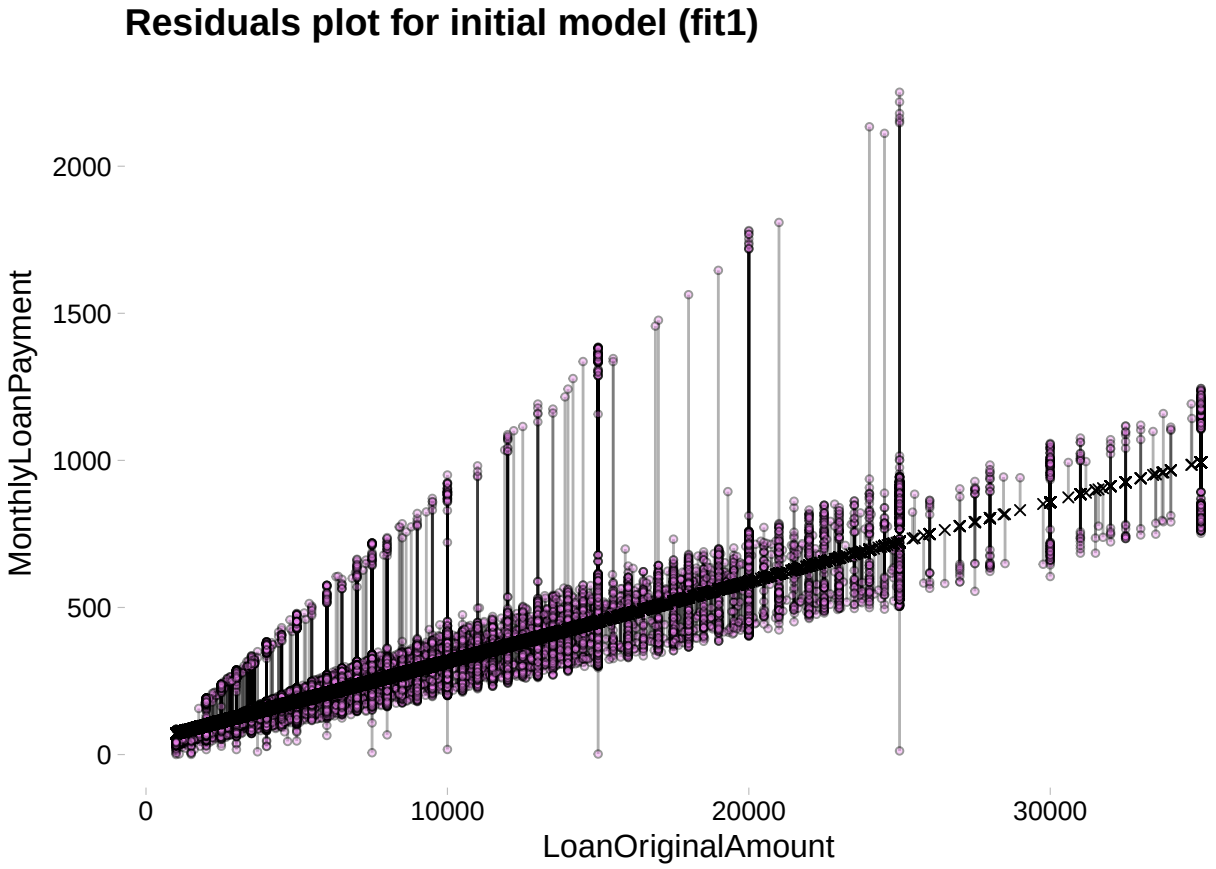


As we can see that people with *Above 20* recommendations has the highest number of *MonthlyLoanPayment* percentage as compared to any other group. But this still can't assure us that people with higher recommendations pay more loan each month. This question was relevant because we want to find out IF PEOPLE WITH MORE RECOMMENDATION ARE POWERFUL AND RICH ENOUGH TO PAY MORE LOAN MONTHLY? The answer is we can't be really sure.

## Let's predict the Monthly Loan Payment

There are multiple number of models that we can create with different number of input or independent variables. Let's try to create a linear model and see how we can improve it.

```
# Selecting columns that we require
model <- loanData %>%
    dplyr::select(MonthlyLoanPayment,
                  LoanOriginalAmount,
                  DebtToIncomeRatio,
                  ProsperRating..numeric.,
                  Term,
                  LoanOriginationQuarter) %>%
    filter(!is.na(MonthlyLoanPayment), !is.na(LoanOriginalAmount),
           MonthlyLoanPayment != 0.0, LoanOriginalAmount != 0.0,
           MonthlyLoanPayment != "", LoanOriginalAmount != "",
           !is.na(DebtToIncomeRatio), DebtToIncomeRatio != "",
           !is.na(ProsperRating..numeric.))

# Creating the first Linear Model

fit1 <- lm(MonthlyLoanPayment ~ LoanOriginalAmount, data = model)

# Predict from the model
model$predicted <- predict(fit1)

# Calculate the residuals for accuracy
model$residuals <- residuals(fit1)

# Visualize how accurate the model (fit1)

ggplot(data = model, aes(x = LoanOriginalAmount, y = MonthlyLoanPayment)) +
    geom_segment(aes(xend = LoanOriginalAmount , yend = predicted), alpha = 0.3) +
    theme_pander() +
    geom_point(shape = 21, fill = 'violet', size = 1, alpha = 0.4) +
    geom_point(aes(y = predicted), shape = 4) +
    ggtitle("Residuals plot for initial model (fit1)")
```




```
summary(fit1)
```

```
##
## Call:
## lm(formula = MonthlyLoanPayment ~ LoanOriginalAmount, data = model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -710.32  -33.57   -3.00   27.64 1528.86
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.556e+01  4.716e-01   96.61   <2e-16 ***
## LoanOriginalAmount 2.708e-02  4.177e-05  648.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.23 on 77112 degrees of freedom
## Multiple R-squared:  0.845,  Adjusted R-squared:  0.845
## F-statistic: 4.203e+05 on 1 and 77112 DF,  p-value: < 2.2e-16
```

*LoanOriginalAmount* should be the most important factor to decide what would be the *MonthlyLoanPayment*. As we can see that the correlation has very good *Adjusted R-squared* value of 0.845 or almost 85%. The *Residual Standard Error* is however quite high of almost75 which indicates the average performace of the linear model. We can visualise how the residuals are spread from the above plot. Their distribution is quite high. So the bigger question is that can we improve this model. Improving the model would ensure the following things

1. Higher Adjusted R-squared value
2. Lower Residual Standard Error
3. Less variation/ Less dispersed residual graph

The key to improve any model liner/non-linear is to increase the number of independent variables. The choice of independent variable should be very logical. We should not include the variable which are dependent on the predicting variable. The independent variables should be truly independent. Here are the following variables that are choosen to improve the model.

1. LoanOriginalAmount: This should be the most important one
because how much the borrower has to pay each month
should be dependent on the total amount that is loaned

2. DebtToIncomeRatio: This is important because lenders may
ask more money each month who has a record of high DebtToIncomeRatio
because they are considered borrowers with poor prospect and
lenders might charge them more if giving lonas

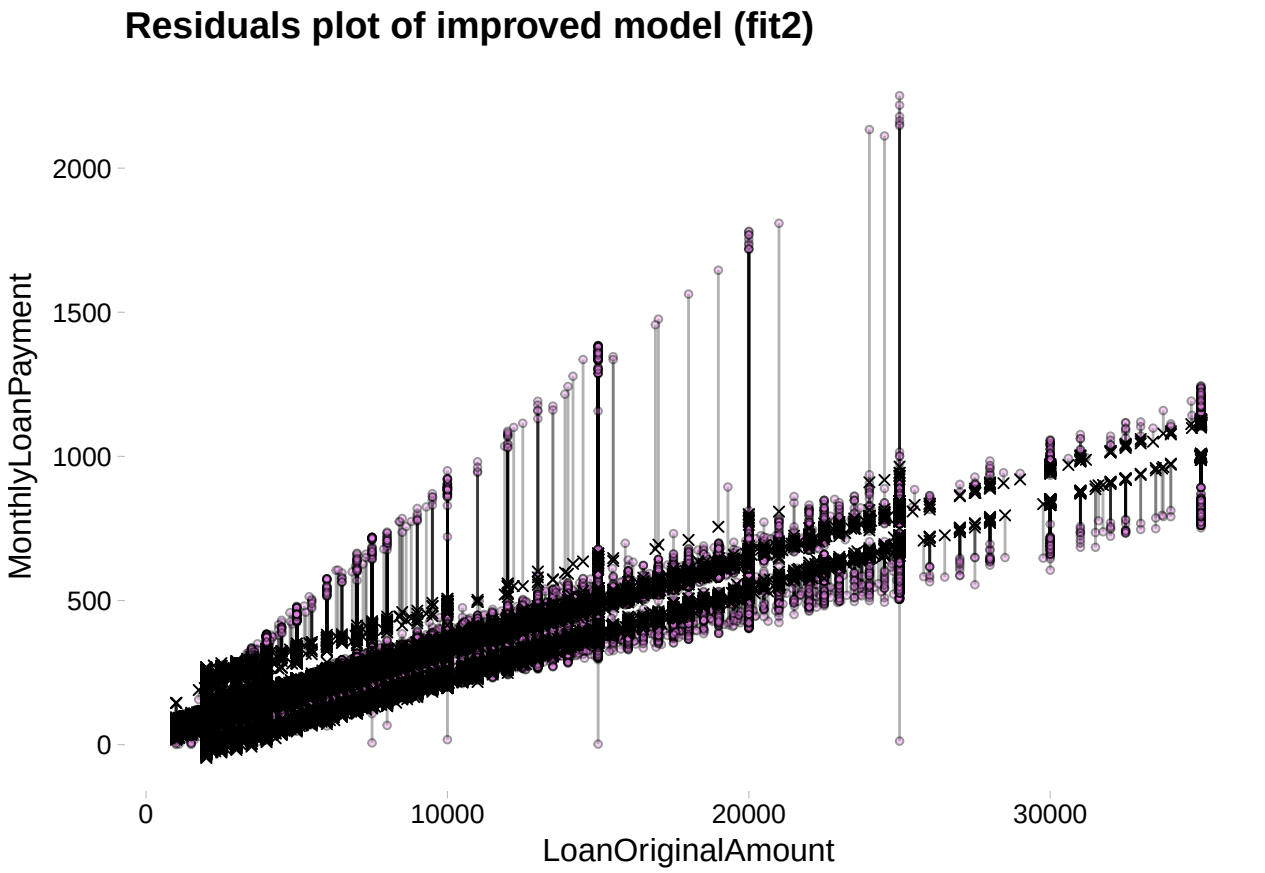
3. ProsperRating..numeric.: This is obvious because the we have shown
already that lenders trust the prosper rating a lot over anything else
and hence borrowers with poor ratings should pay more each
month as compared to better ratings.

4. Term: Term decides how long the borrower is opting for the loan
and hence longer the Term shorter should be monthly payment.

5. LoanOriginationQuarter: Lenders may ask for different monthly loan payment
on different time of the year because of dynamic interest rate
and changing macro economic factors on different time of the year.

```
# Selecting columns that we require
model <- loanData %>%
    dplyr::select(MonthlyLoanPayment,
                  LoanOriginalAmount,
                  DebtToIncomeRatio,
                  ProsperRating..numeric.,
                  Term,
                  LoanOriginationQuarter) %>%
    filter(!is.na(MonthlyLoanPayment), !is.na(LoanOriginalAmount),
           MonthlyLoanPayment != 0.0, LoanOriginalAmount != 0.0,
           MonthlyLoanPayment != "", LoanOriginalAmount != "",
           !is.na(DebtToIncomeRatio), DebtToIncomeRatio != "",
           !is.na(ProsperRating..numeric.))

# Creating the first Linear Model

fit2 <- lm(MonthlyLoanPayment ~ LoanOriginalAmount +
               DebtToIncomeRatio +
               ProsperRating..numeric. +
               Term + LoanOriginationQuarter,data = model)

# Predict from the model
model$predicted <- predict(fit2)

# Calculate the residuals for accuracy
```

```r
model$residuals <- residuals(fit2)

# Visualize how accurate the model (fit1)
# summary(fit2)
ggplot(data = model, aes(x = LoanOriginalAmount, y = MonthlyLoanPayment)) +
    geom_segment(aes(xend = LoanOriginalAmount , yend = predicted), alpha = 0.3) +
    theme_pander() +
    geom_point(shape = 21, fill = 'violet', size = 1, alpha = 0.4) +
    geom_point(aes(y = predicted), shape = 4) +
    ggtitle("Residuals plot of improved model (fit2)")
```




We can see a lot less variation in the residual. Moreover when we summarize the fit2 for our improved model, we see that the value of *Residual standard error* decreased a lot to almost 49 and *Adjusted R-squared* incresed to almost 94%. This was a huge jump from our previous model. Hence we can say that fit2 predicts the monthly loan amount better than the fit1 model. This also proves that *LoanOriginalAmount* is not only the deciding factor of *MonthlyLoanPayment*. There are other factors too that contribute to this. From the `summary(fit2)`, we observe that the slope values like below

1. `LoanOriginalAmount: 3.124e-02`
2. `DebtToIncomeRatio:  3.211e+00`
3. `ProsperRating..numeric.: -1.050e+01`
4. `Term: -4.993e+00`

We can see that as we might have assumed, *LoanOriginalAmount* & *DebtToIncomeRatio* has positive solpe which is obvious because if these two increse, lenders ask for more money from the borrowers. On the other hand, *Term* has a negative slope which is logical because if term increses, monthly loan payment should decrease.

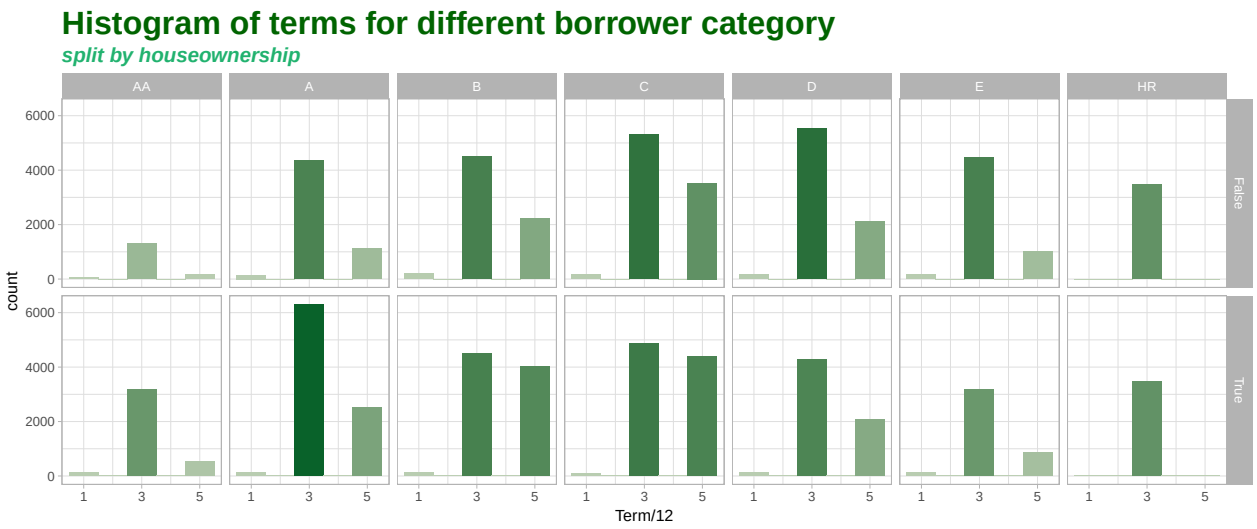
We can also see that the both maximum and minimum residual for our improved model is lesser than our previous model.

## Some Final Thoughts

Let's select 3 plots from what we have discussed and elaborate them bit further.

### Term's distribution for each Prosper Category

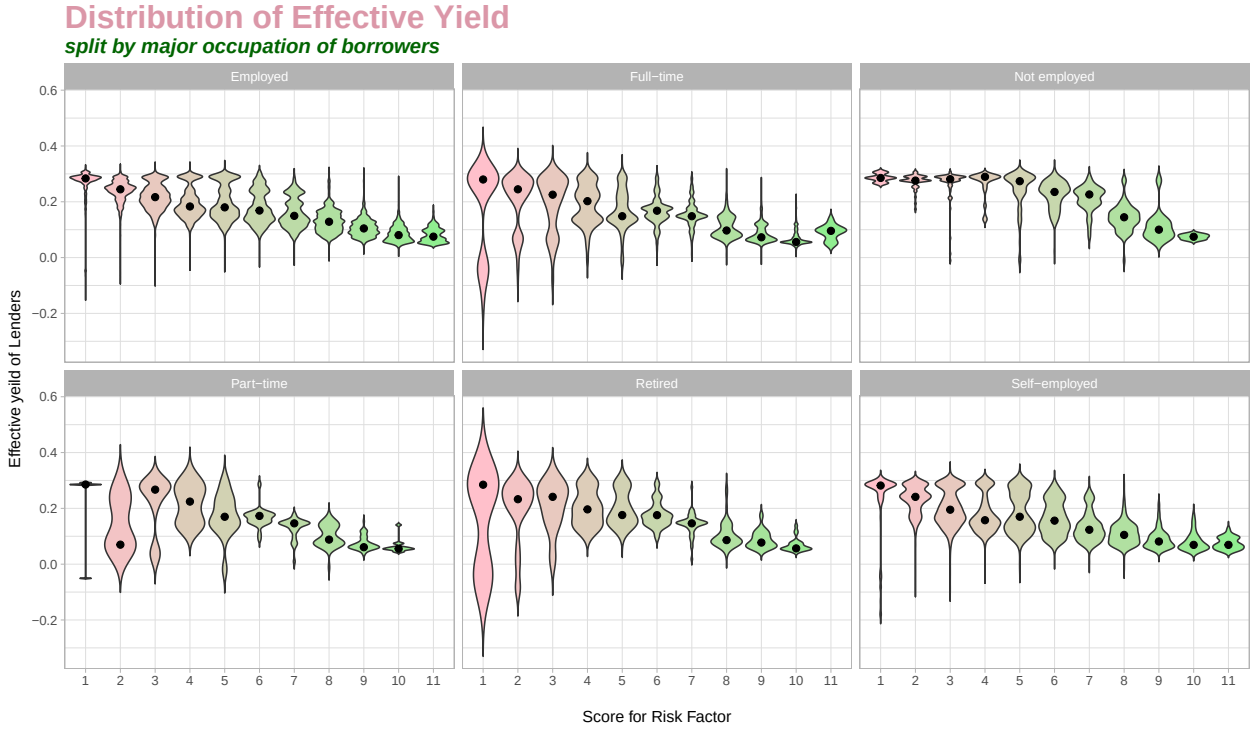
```
ggplot(subset(loanData, loanData$ProsperRating..Alpha. != ""), aes(x = Term / 12)) +
    geom_histogram(binwidth = 1, aes(fill = ..count..)) +
    scale_x_continuous(breaks = seq(1, 5, 2)) +
    facet_grid(IsBorrowerHomeowner ~ ordered(ProsperRating..Alpha.,
            levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR'))) +
    guides(fill=FALSE) +
    theme_light() +
    scale_fill_continuous_tableau(palette = "Green", space = "Lab") +
    ggtitle("Histogram of terms for different borrower category",
            subtitle = "split by houseownership") +
    theme(plot.title = element_text(face = 'bold', colour = 'darkgreen', size=22),
          plot.subtitle = element_text(face = 'bold.italic', size = 14, colour = "#21b26f"))
```




As we can see the above mentioned plot is extension of the first plot of the histogram of different terms. Now it is split over serveral important variables. Firstly we are distributing it by each borrower category to see if there is any change in the pattern of loan terms for different type of borrowers. We can see along all the borrower types, 3 years is the most common borrowing year as we have already seen. We can see that people with $AA$ rank loan very less, while people of $C$ and $D$ category seems to loan most and all of their most poplular loan amount is 3 years. Moreover row-wise we split the plot in 2 sections namely, *True and False* indcating Houseowner or not. We can see a lot more loan for houseowners with $AA$ and $A$ category while for others no houseowners seems to loan same or even more for low rating borrowers than houseowners. This may indicate that wheather a borrower would get more loan or not is dependent on the fact that they own their own house or not but only for high ranking borrowers. For low ranking borrowers, owning their own house doesn't affect much.

**EstimatedEffectiveYield variation on Employment type**

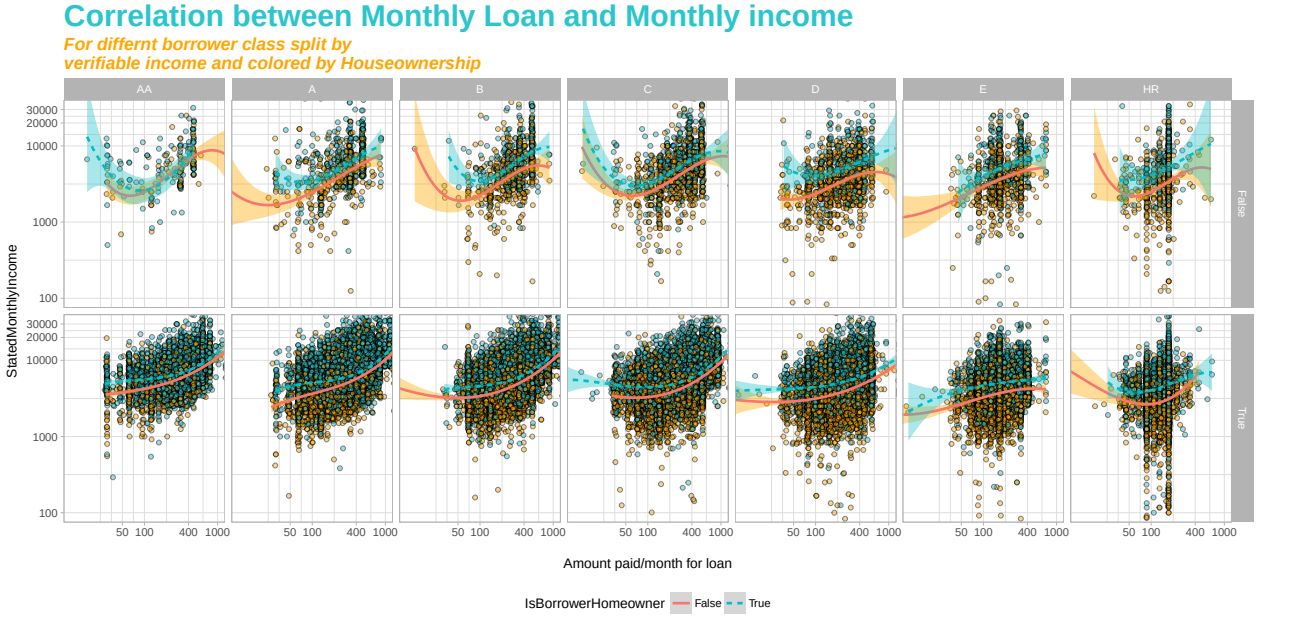
```
loanData$ProsperScore <- factor(loanData$ProsperScore)
ggplot(aes(x = ProsperScore, y = EstimatedEffectiveYield, fill=ProsperScore),
       data = subset(loanData, !is.na(loanData$ProsperScore) &
                             !is.na(loanData$EstimatedEffectiveYield) &
                             loanData$EmploymentStatus != "" & loanData$EmploymentStatus != "Other")) +
    geom_violin(trim = F, scale = "width") +
    stat_summary(fun.y=median, geom="point", size=2, color="black") +
    scale_fill_manual(values=colorRampPalette(c("pink", "lightgreen"))(11)) +
    theme_light() +
    xlab('\nScore for Risk Factor') +
    ylab('Effective yeild of Lenders') +
    guides(fill = F) +
    facet_wrap(~EmploymentStatus) +
    ggtitle("Distribution of Effective Yield",
            subtitle = "split by major occupation of borrowers") +
    theme(plot.title = element_text(face = 'bold', colour = '#db97a8', size=22),
          plot.subtitle = element_text(face = 'bold.italic', size = 14, colour = "darkgreen"))
```




EstimatedEffectiveYield is measusred here which is the extention of the 5th plot of distribution EstimatedEffectiveYield for all ProsperScores. We found that as prosper socre increses, lenders seemed to demand comparatively lesser than the borrowers with poor ProsperScore. Now we would see how this varies with the profession of the borrowers. We see *Not employed* people tends to have very less variation along all the prosperscore ranges. This may be due to the fact that they are not employed and hence return form them is kept fixed. We see almost the same pattern in the *Self-employed* people. We see on the other hand a lot of variation in *Full-time* and *Retired* category. Probably because there are scopes of different level of returns from them. *Self-employed* is somehow moderate as it should be because of its limited return values from them while *Employed* remained to be strange because if it's amibuity with *Full-time* as discussed. Overall we see a lot of ups and downs for the lenders' return if the borrowers are retired or full time as compared to others.

### Correlation of MonthlyIncome vs MonthlyLoanPayment

```
base <- loanData %>%
    filter(StatedMonthlyIncome != 0.0,
           MonthlyLoanPayment != 0.0,
           ProsperRating..Alpha. != "") %>%
    ggplot(aes(y = StatedMonthlyIncome,
               x = MonthlyLoanPayment)) +
    geom_jitter(color = 'black', shape = 21, alpha = 1/2) +
    coord_cartesian(ylim = c(100, 30000), xlim = c(10, 1000)) +
    scale_y_continuous(trans = log10_trans(),
                       breaks = c(100, 1000, 10000, 20000, 30000)) +
    scale_x_continuous(trans = log10_trans(),
                       breaks = c(50, 100, 400, 1000))

base + geom_smooth(span = 0.3, method = 'lm', formula = y ~ splines::bs(x, 3),
                   aes(linetype = IsBorrowerHomeowner, colour = IsBorrowerHomeowner)) +
                                    theme(axis.title.y=element_blank(),
                                          axis.text.y=element_blank(),
                                          axis.ticks.y=element_blank()) +
    xlab('\nAmount paid/month for loan') +
    aes(fill = factor(IsBorrowerHomeowner)) +
    facet_grid(IncomeVerifiable ~ ordered(ProsperRating..Alpha.,
            levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR'))) +
    theme_light() +
    scale_fill_manual(values=c("#FFAA00", "#2BC5CC")) +
    guides(fill = F) +
    theme(legend.position="bottom",
          plot.title = element_text(face = 'bold', colour = '#2BC5CC', size=24),
          plot.subtitle = element_text(face = 'bold.italic', size = 14, color = "orange")) +
    ggtitle("Correlation between Monthly Loan and Monthly income",
            subtitle = "For differnt borrower class split by
verifiable income and colored by Houseownership")
```



This plot talks about so much and it is hard to talk about all of these. The color is by houseownership while the rows are split by verifiable income or not and columns are as usual borrowers' rating. We are trying to find the correlation of monthly income vs monthly loan payment. The cubic line fits in almost all of them and we see both the geom smoothing lines are following each other nicely. This means weather a borrower owns a house or not is not much of importance for the correlation. But we definitely see a more linear relationship for all the categories for the borrowers having verifiable income as compared to non verifiable income. This means that if income is verified, then their is a linear change in monthly loan payment and monthly income as compared non verifiable income borrowers. We also see that for borrowers with verifiable income the relation goes from linear to cubic as the rating of the borrowers goes down.

## Reflection

**Struggle** with this dataset were many. I was at first intimidated by it seer size of number of rows. I honestly have never seen any CSV file which is this big and contains this many variables. However the real relief was the CSV file that contained the meaing of all the variables in the dataset. But the biggest struggke that I faced in this dataset was actually common in most of the datasets that contains this many variables. Some of the variables seems to ahve the same meaning for example *BorrowerAPR*, *BorrowerRate*, *LenderYield*, *EstimatedEffectiveYield*. These terms seemed to have similar meaning but not exactly the same. Now the was which one to include in our analysis and which is more inportant than other? Although with little bit more understanding I solved it but it took some time. Another struggle was to establish right combination of variables to create accurate linear regression model for minimum residual. This problem could have been solved using *Gradient Decend* but beacuse it being out of scope I had to try different combination manually. This also took some time.

**Successes** were achieved even after the above mentioned struggle because of some detailed analysis of the dataset and the CSV file supplied that expalins the variables in the dataset. Linear Regression model's accuracy was checked by carefully taken the variables and then checking the *Adjusted R-squared* values and the *Residual standard error* and plotting the Residual standard error. By doing this I could actually improve the regression model substantially already discussed earlier.

Although some **ideas for future exploration** can improve this analysis substantially. I really think regrssion model can be improved further using *Gradient Decend* to better approximate the solope and the intercept of the line. I also believe that is some information hidden inside the delinquency variables and late payment variables in *CurrentDelinquencies, AmountDelinquent, DelinquenciesLast7Years* and *OnTimeProsperPayments, ProsperPaymentsLessThanOneMonthLate, ProsperPaymentsOneMonthPlusLate* with respect the Monthly loan amount which can be explored further.

However one thing is confirmed, Financial Dataset is really intensive stuff.