Prosper Loan Data Analysis

By: Yougender Chauhan

Lekha Ajit Kumar

Cal State LAlogo_shield

California State University, Los Angeles

Submitted to Dr. Shilpa Balan

INDEX

Part A: Introduction

The purpose of this project is to comprehensively investigate a dataset comprising loan information sourced from Prosper.com, covering the period from 2005 to 2014. The data originates from an undisclosed bank and is centered around various aspects related to the bank's customers. Although the specific identity and details of the bank remain unknown, the dataset is extensive, encompassing more than 110,000 individual loan entries, each characterized by 81 distinct variables.

In this study, we focus our analysis on 17 specific variables within the dataset. The primary objective is to uncover valuable insights into four key aspects:

Individuals Making Loan Requests ("**who**"):

By exploring certain variables, we seek to understand the demographic and personal characteristics of the individuals applying for loans. This includes aspects such as age, employment status, and other relevant factors that contribute to the "who" behind the loan requests.

Amounts Requested ("**what**"):

Our investigation will delve into variables related to the amounts requested by loan applicants. This involves understanding the range, distribution, and trends in the requested loan amounts, shedding light on the "what" aspect of the loan transactions.

Types of Loans Applied For ("**why**"):

We aim to analyze specific variables that provide insights into the types of loans applied for by Prosper clients. Understanding the motivations and purposes behind loan applications contributes to unraveling the "why" behind the borrowing behavior.

Criteria Influencing Loan Approvals ("**how**"):

The study will focus on variables that illuminate the criteria influencing approvals for Prosper clients. This encompasses factors such as credit scores, employment history, and other relevant metrics, aiming to reveal the "how" behind the decision-making process for approving loans.

In summary, our exploration aims to extract meaningful information from the Prosper.com loan dataset, shedding light on the characteristics of loan applicants, the dynamics of requested loan amounts, the motivations driving loan applications, and the factors influencing the approval process. By concentrating on these key variables, we seek to contribute valuable insights to understanding lending practices and customer behavior within the specified timeframe and financial context.

REFERENCES:

1. Article title: Personal loans through Prosper, trusted since 2005

URL: https://www.prosper.com/personal-loans

Date published: 2005

2. Article title: An Analysis of Bank Financial Strength Ratings and Credit Rating Data.

URL: https://www.mdpi.com/2227-9091/9/9/155

Date published: August 26, 2021

3. Article title: How can banking data analysis mitigate financial risks?

URL: https://www.linkedin.com/advice/0/how-can-banking-data-analysis

Website title: How Data Analysis Can Mitigate Banking Risks.

Date published: September 18, 2023

Part B: DATASET(S)/ DATA SOURCE URL:

https://www.kaggle.com/datasets/henryokam/prosper-loan-data/data

Part C: Data Description

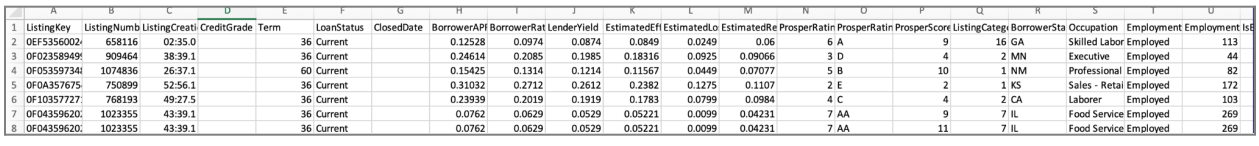
Data Description of Used Columns

NAME	DESCRIPTION	Example value
CreditScoreRangeLower	Lowest credit score number	640
CreditScoreRangeUpper	Highest credit score number	659
CurrentCreditLines	No of credit lines	5
CurrentDelinquencies	Number of accounts delinquent at the time the credit profile was pulled.	2
DebtToIncomeRatio	The debt-to-income ratio of the borrower at the time the credit profile was pulled.	0.17

EmploymentStatus	The employment status of the borrower at the time they posted the listing.	Self-employed
EstimatedReturn	The estimated return assigned to the listing at the time it was created.	0.06
IncomeRange	The income range of the borrower at the time the listing was created.	$25,000-49,999
ListingCategory	The category of the listing that the borrower selected when posting their listing	0
LoanOriginalAmount	The origination amount of the loan.	9425
Occupation	The Occupation selected by the Borrower at the time they created the listing.	Professional
OpenCreditLines	Number of open credit lines at the time the credit profile was pulled.	2

IsBorrowerHomeOwner	A Borrower will be classified as a homeowner if they have a mortgage on their credit profile or provide documentation confirming they are a homeowner.	True
EstimatedLoss	Estimated loss is the estimated principal loss on charge-offs.	0.0249
PublicRecordsLast12Months	Number of public records in the past 12 months when the credit profile was pulled.	0
EffectiveYield	Effective yield is equal to the borrower's interest rate	0.138
RevolvingCreditBalance	Dollars of revolving credit at the time the credit profile was pulled.	1444.000000

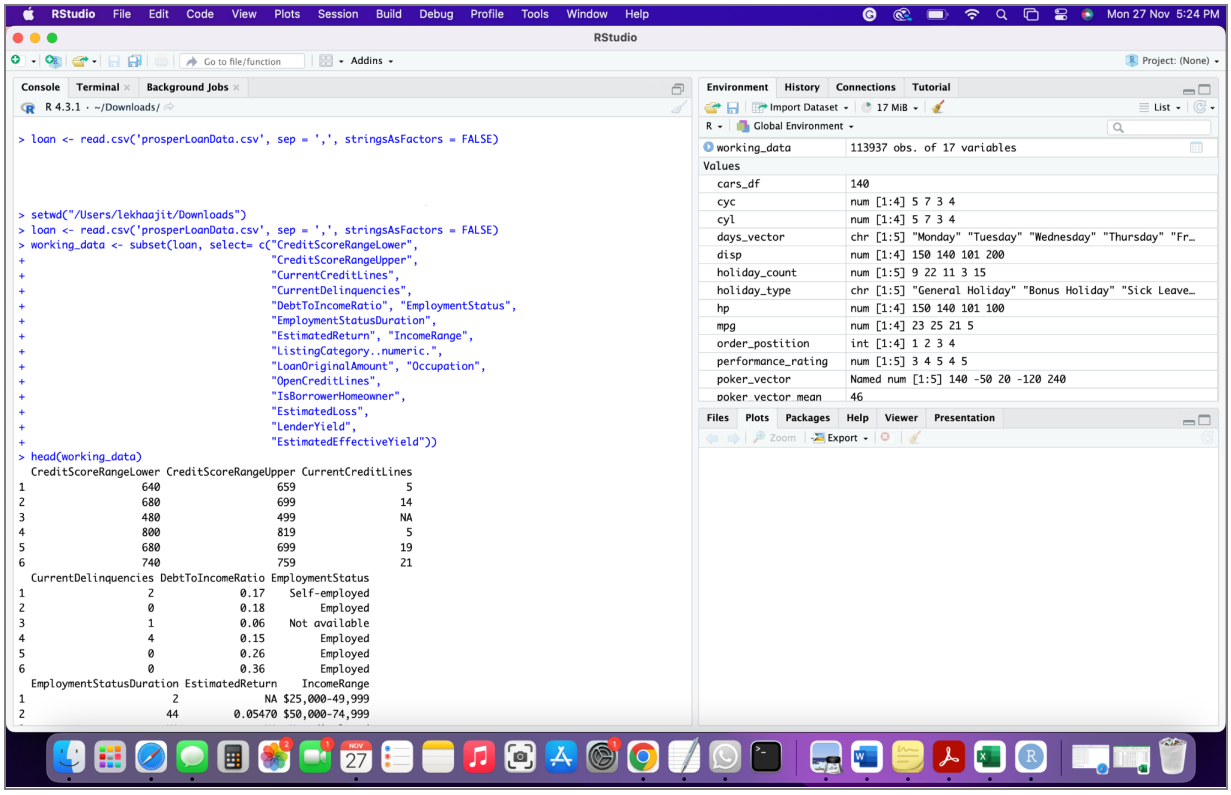
Screenshot of the dataset:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	ListingKey	ListingNumb	ListingCreati	CreditGrade	Term	LoanStatus	ClosedDate	BorrowerAPI	BorrowerRat	LenderYield	EstimatedEf	EstimatedLo	EstimatedRe	ProsperRatin	ProsperRatin	ProsperScore	ListingCateg	BorrowerSta	Occupation	Employment	Employment IsE
2	0EF5356002	658116	02:35.0		36	Current		0.12528	0.0974	0.0874	0.0849	0.0249	0.06	6	A	9	16	GA	Skilled Labor	Employed	113
3	0F023589495	909464	38:39.1		36	Current		0.24614	0.2085	0.1985	0.18316	0.0925	0.09066	3	D	4	2	MN	Executive	Employed	44
4	0F053597348	1074836	26:37.1		60	Current		0.15425	0.1314	0.1214	0.11567	0.0449	0.07077	5	B	10	1	NM	Professional	Employed	82
5	0F0A357675	750899	52:56.1		36	Current		0.31032	0.2712	0.2612	0.2382	0.1275	0.1107	2	E	2	1	KS	Sales - Retai	Employed	172
6	0F1035772	768193	49:27.5		36	Current		0.23939	0.2019	0.1919	0.1783	0.0799	0.0984	4	C	4	2	CA	Laborer	Employed	103
7	0F043596203	1023355	43:39.1		36	Current		0.0762	0.0629	0.0529	0.05221	0.0099	0.04231	7	AA	9	7	IL	Food Service	Employed	269
8	0F043596203	1023355	43:39.1		36	Current		0.0762	0.0629	0.0529	0.05221	0.0099	0.04231	7	AA	11	7	IL	Food Service	Employed	269

Part D: Data cleaning

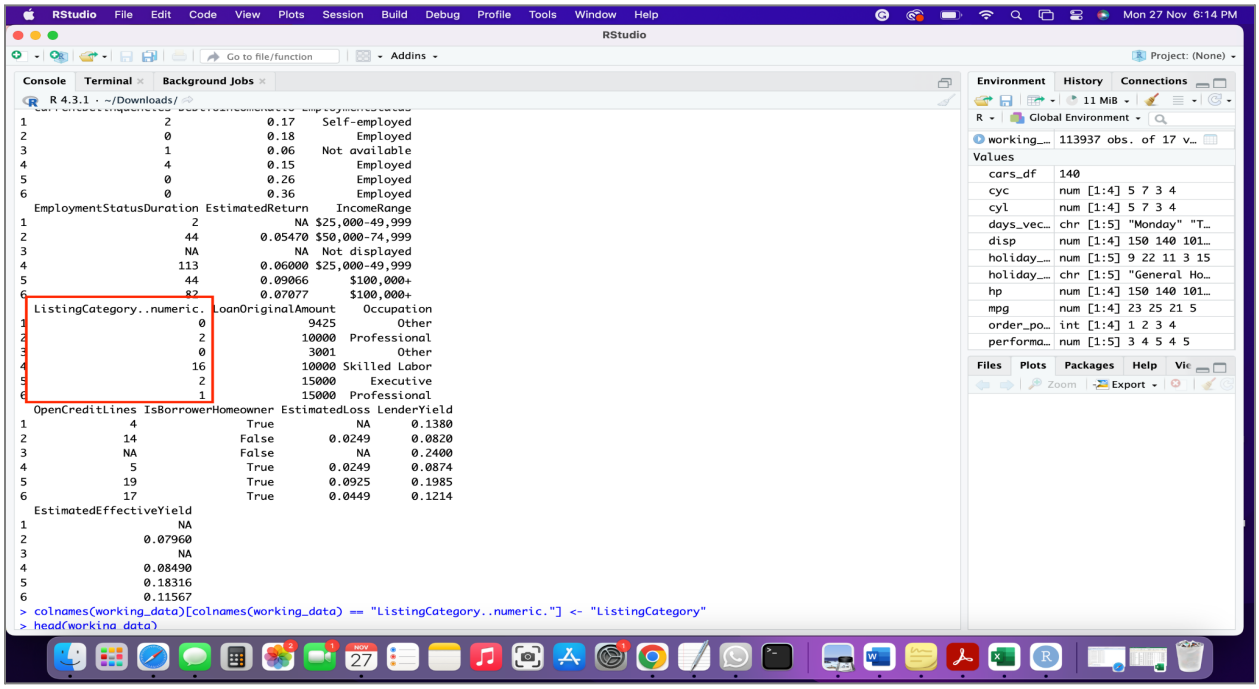
1. Selecting 17 out of 81 rows. Our dataset has 81 columns from which we are using only 17 columns. In this R code, it reads a CSV file into a data frame (Loan) and then creates a new data frame (working_data) containing a subset of columns from the original data frame.



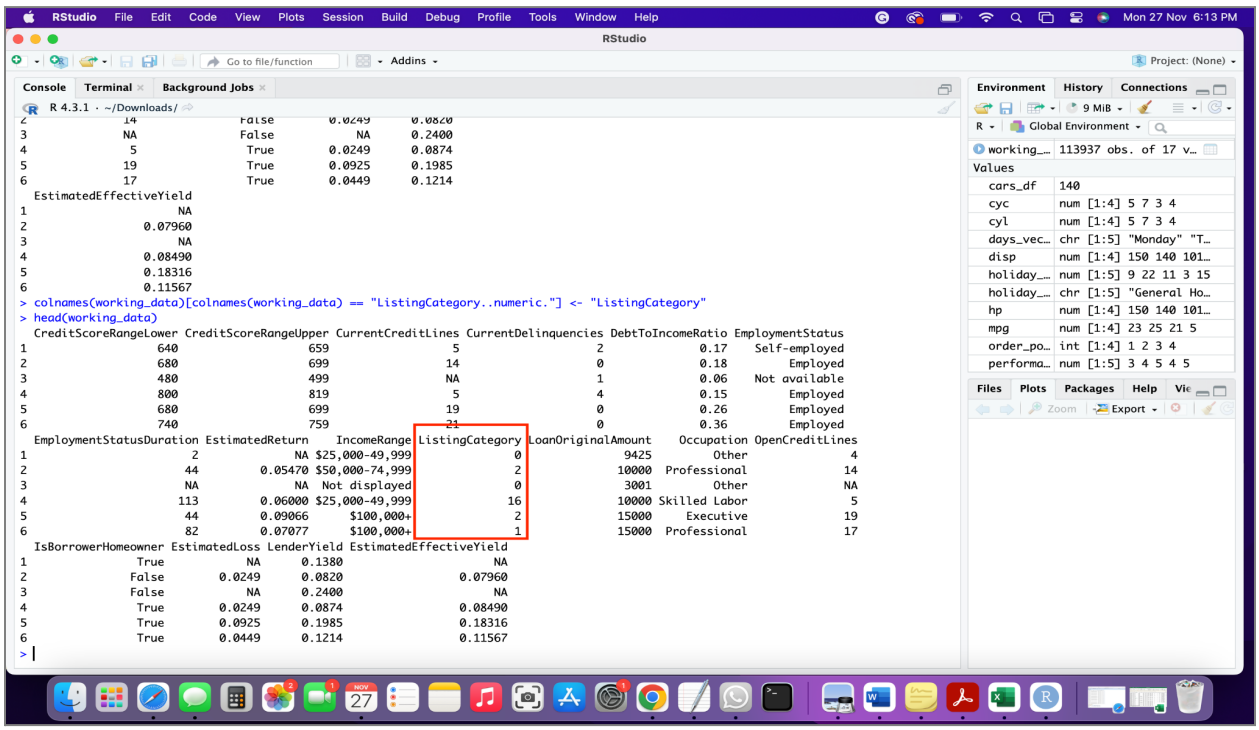
```
> loan <- read.csv('prosperLoanData.csv', sep = ',', stringsAsFactors = FALSE)



> setwd("/Users/lekhaajit/Downloads")
> loan <- read.csv('prosperLoanData.csv', sep = ',', stringsAsFactors = FALSE)
> working_data <- subset(loan, select= c("CreditScoreRangeLower",
+                                        "CreditScoreRangeUpper",
+                                        "CurrentCreditLines",
+                                        "CurrentDelinquencies",
+                                        "DebtToIncomeRatio", "EmploymentStatus",
+                                        "EmploymentStatusDuration",
+                                        "EstimatedReturn", "IncomeRange",
+                                        "ListingCategory..numeric.",
+                                        "LoanOriginalAmount", "Occupation",
+                                        "OpenCreditLines",
+                                        "IsBorrowerHomeowner",
+                                        "EstimatedLoss",
+                                        "LenderYield",
+                                        "EstimatedEffectiveYield"))
```

2. Replace column name. Here we are renaming the column in working_data from "ListingCategory..numeric." to "ListingCategory." It's a way of cleaning up or making the column name more concise and user-friendly.

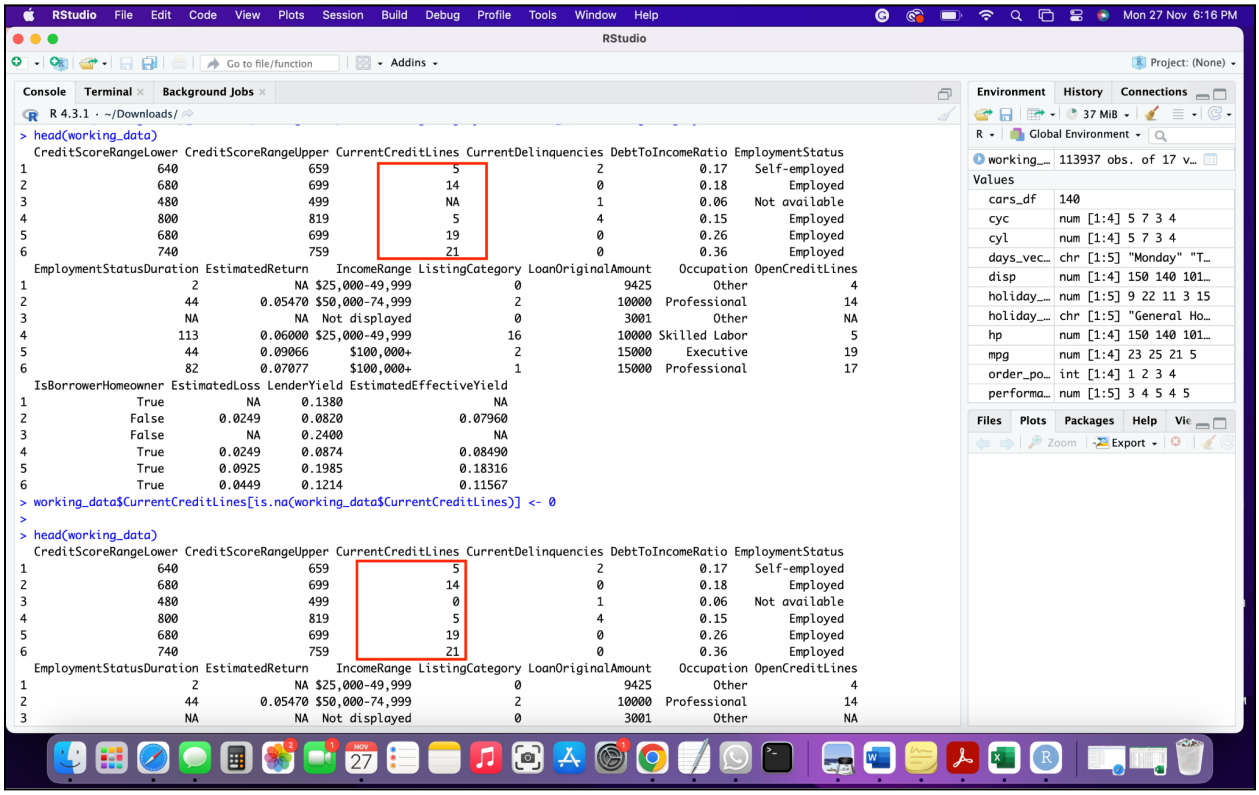
BEFORE



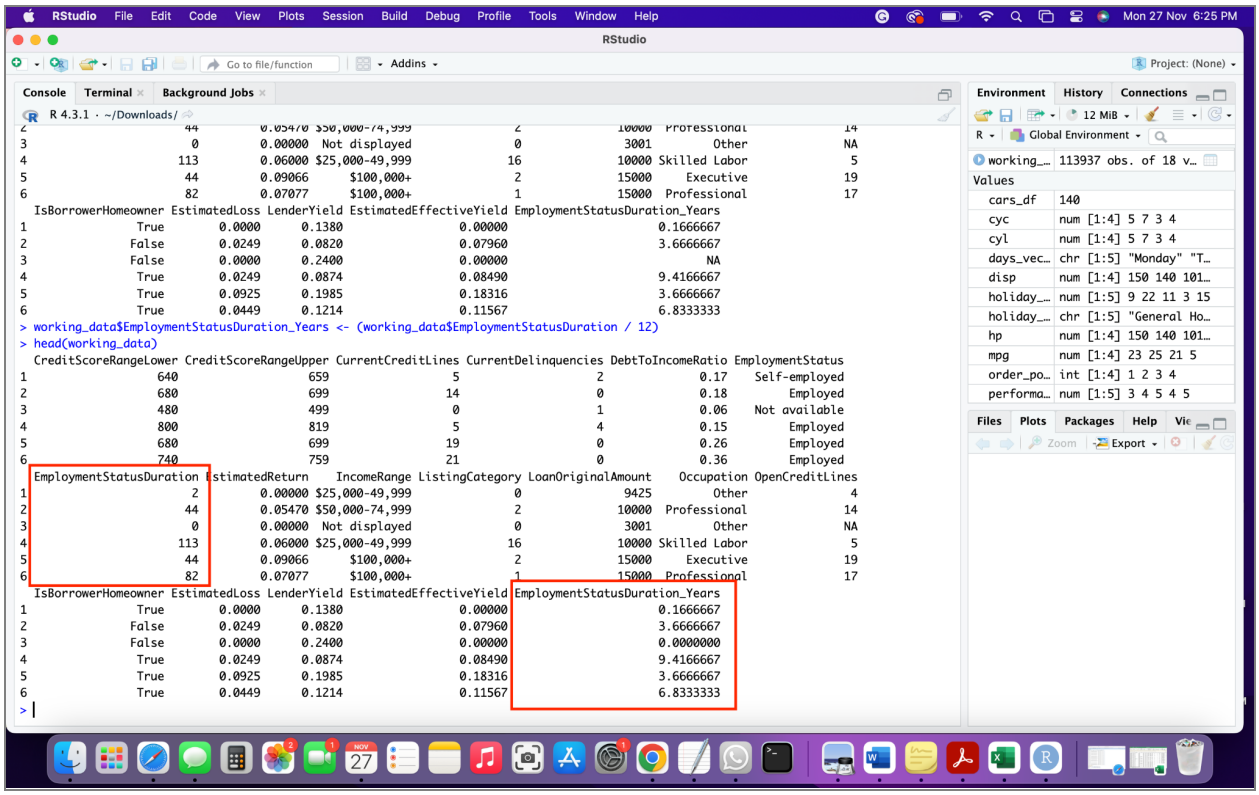
AFTER



3. Replacing NA with 0. This line of code replaces any missing values with 0 in the "CurrentCreditLines" column of the working_data data frame. It's a common approach to handling missing data, providing a default or placeholder value in place of the missing values for analysis.



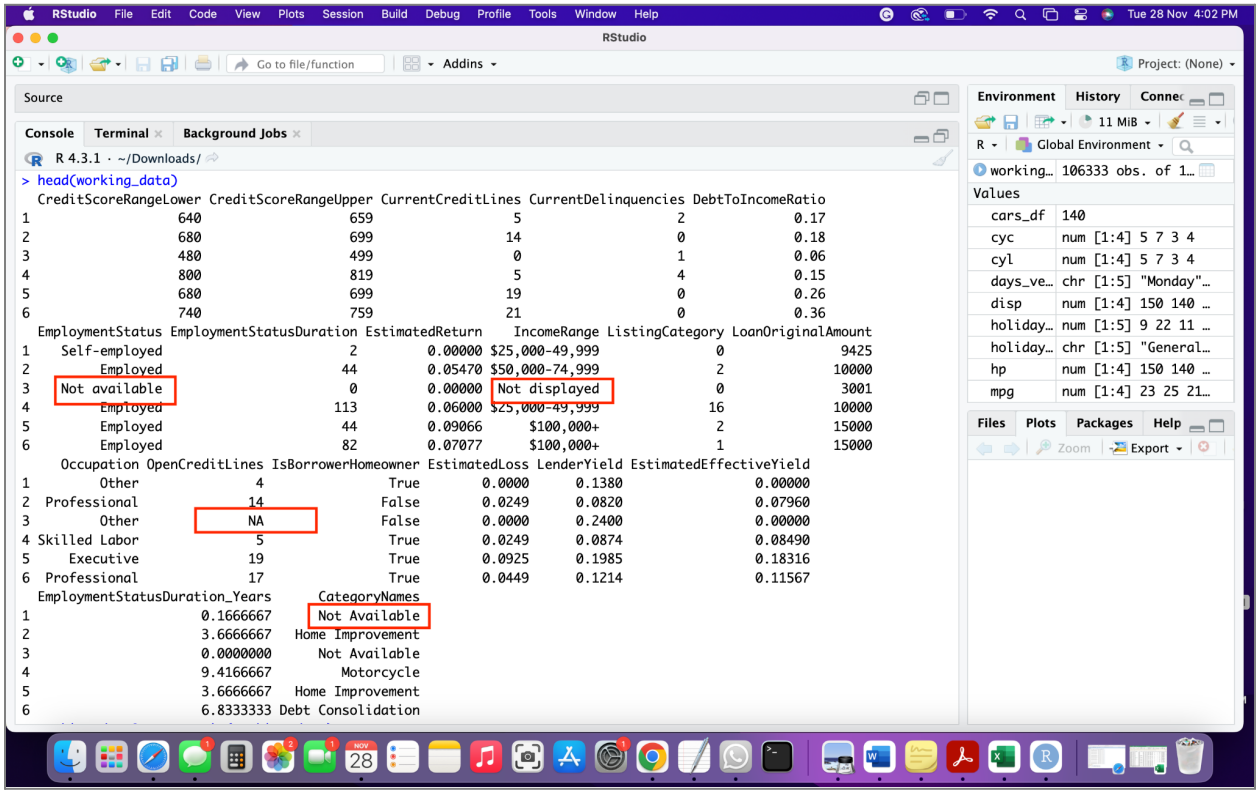
4. Employment duration column added and converted months to years. So, the purpose of the following code was to convert the values in the EmploymentStatusDuration column from months to years and store the result in a new column called EmploymentStatusDuration_Years. This kind of transformation was done because we were dealing with time durations in different units. To make the data more interpretable or consistent, we did this conversion.



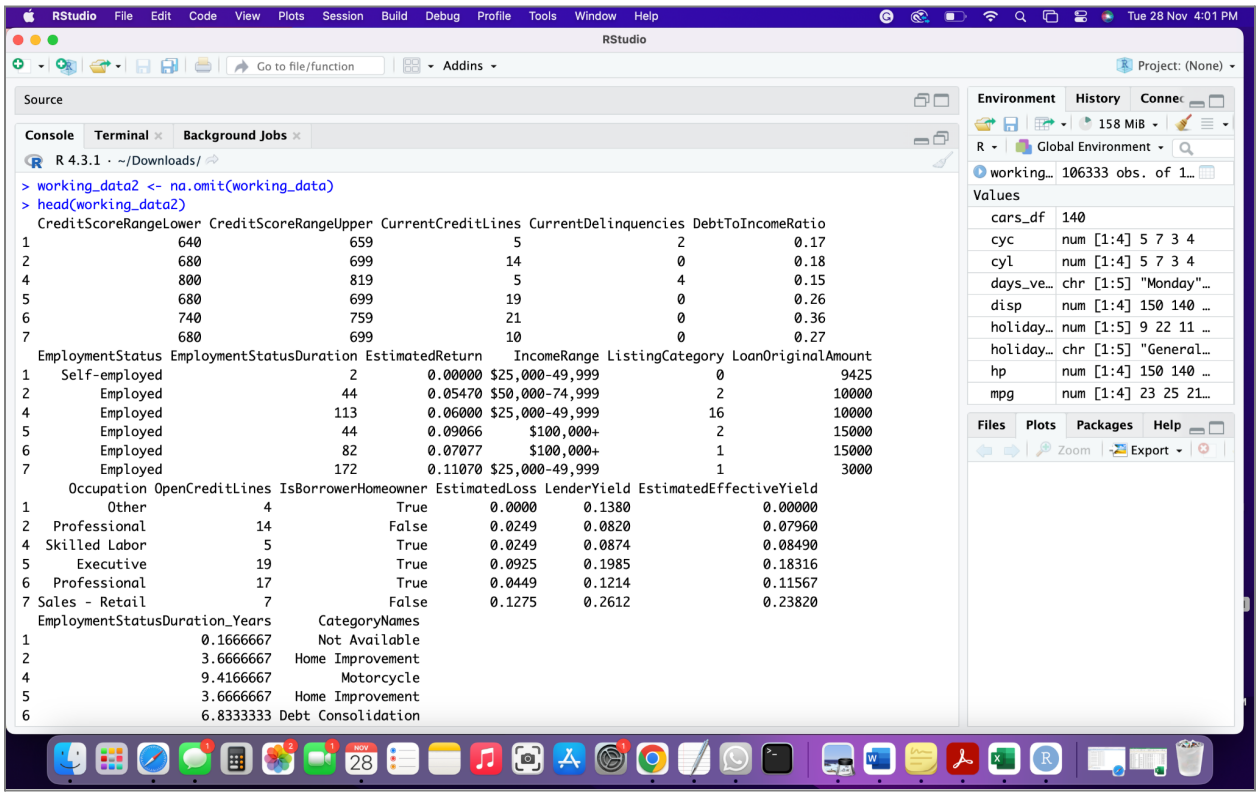

```
> working_data$EmploymentStatusDuration_Years <- (working_data$EmploymentStatusDuration / 12)
> head(working_data)
  CreditScoreRangeLower CreditScoreRangeUpper CurrentCreditLines CurrentDelinquencies DebtToIncomeRatio EmploymentStatus
1                   640                   659                 5                   2              0.17    Self-employed
2                   680                   699                14                   0              0.18         Employed
3                   480                   499                 0                   1              0.06    Not available
4                   800                   819                 5                   4              0.15         Employed
5                   680                   699                19                   0              0.26         Employed
6                   740                   759                21                   0              0.36         Employed
  EmploymentStatusDuration EstimatedReturn    IncomeRange ListingCategory LoanOriginalAmount    Occupation OpenCreditLines
1                        2         0.00000 $25,000-49,999               0              9425         Other              4
2                       44         0.05470 $50,000-74,999               2             10000  Professional             14
3                        0         0.00000  Not displayed               0              3001         Other             NA
4                      113         0.06000 $25,000-49,999              16             10000 Skilled Labor              5
5                       44         0.09066      $100,000+               2             15000     Executive             19
6                       82         0.07077      $100,000+               1             15000  Professional             17
  IsBorrowerHomeowner EstimatedLoss LenderYield EstimatedEffectiveYield EmploymentStatusDuration_Years
1                True        0.0000      0.1380                 0.00000                      0.1666667
2               False        0.0249      0.0820                 0.07960                      3.6666667
3               False        0.0000      0.2400                 0.00000                             NA
4                True        0.0249      0.0874                 0.08490                      9.4166667
5                True        0.0925      0.1985                 0.18316                      3.6666667
6                True        0.0449      0.1214                 0.11567                      6.8333333
>
```

5. Drop Nulls: We had missing values in our records. Dealing with missing data can complicate our data analysis. Removing rows with missing values simplifies the analysis process hence, we dropped the rows with missing values.

BEFORE



AFTER



Part E: Analysis & Visualizations

1. **Question of Analysis -** How does the identified correlation between occupations and loan amounts, along with the rapid density decrease as loan amounts rise, support the claim that occupation is not a determinant for Prosper loans?

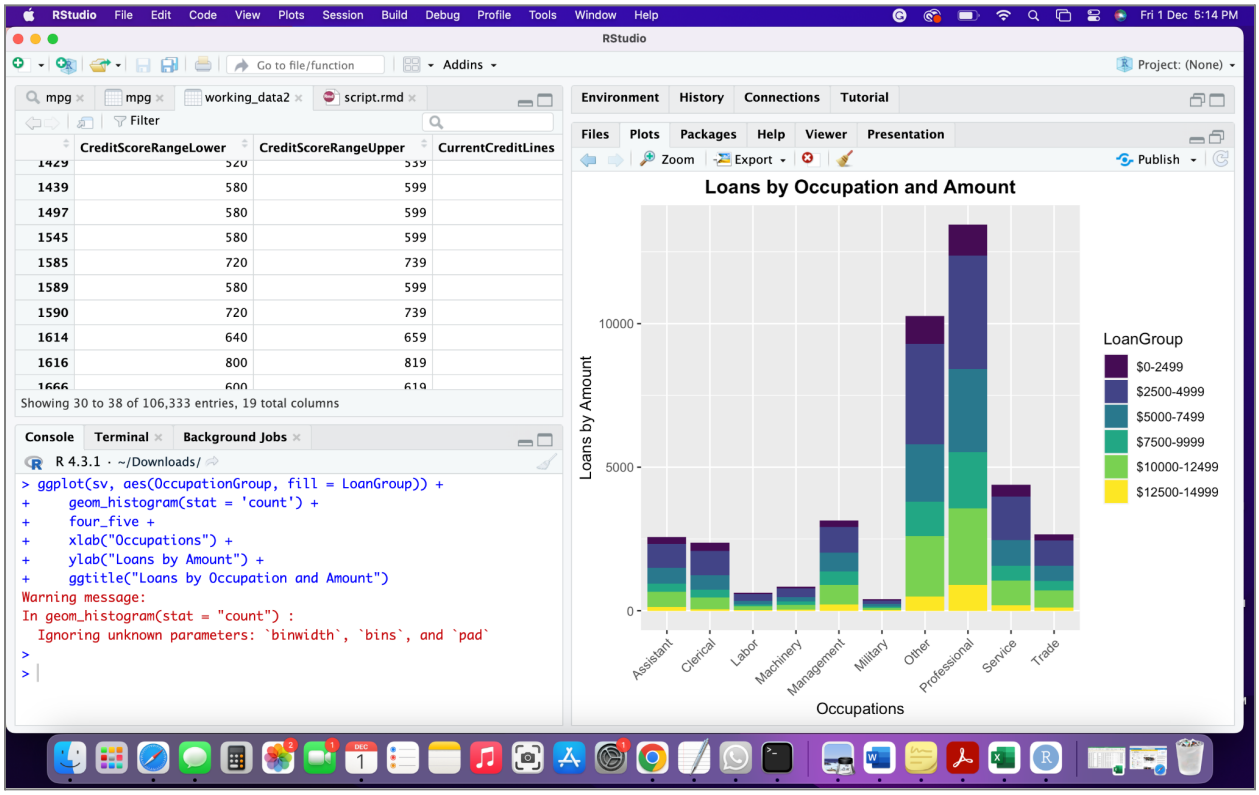
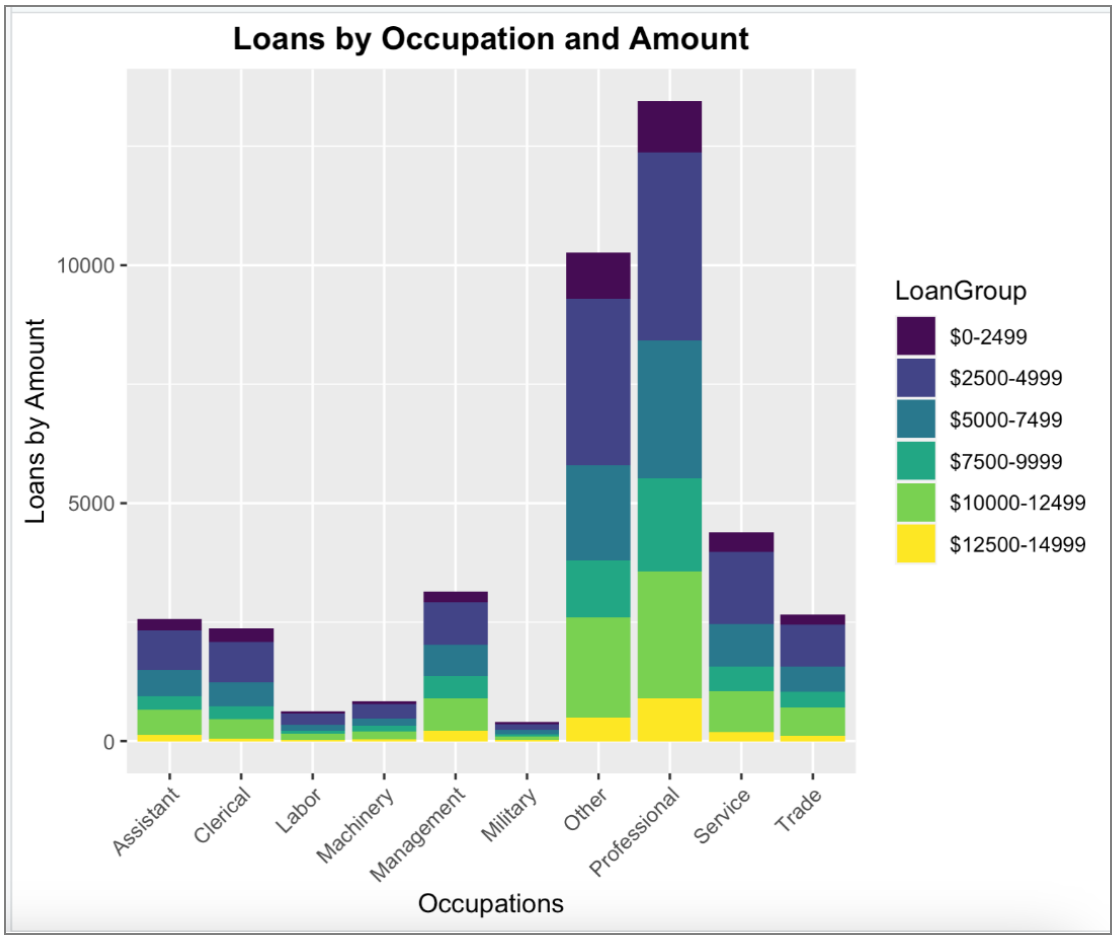


Figure 1. Loans by Occupation and Amount.



Code

```
ggplot(sv, aes(OccupationGroup, fill = LoanGroup)) +
+geom_histogram(stat = 'count") +
+four_five +
+xlab("Occupations") +
+ylab ("Loans by Amount") +
+ggtitle("Loans by Occupation and Amount")
```

We chose this plot because it answers the questions of who(loan requestor) and what(loan amount). It is very easy to see the relationship between occupations and loan amounts. All occupations are clearly represented which agrees with the assertion that occupation is not a factor in qualifying for a loan with Prosper. The division of volume by loan amount is clear as well. We

can see as the loan amount increases the density drops off rapidly which agrees with the assertion that to be successful as a loan requestor the loan should be less than $15,000.

2. **Question of Analysis -** Is there a correlation with respect to specific credit score, loan amount and, DTI?

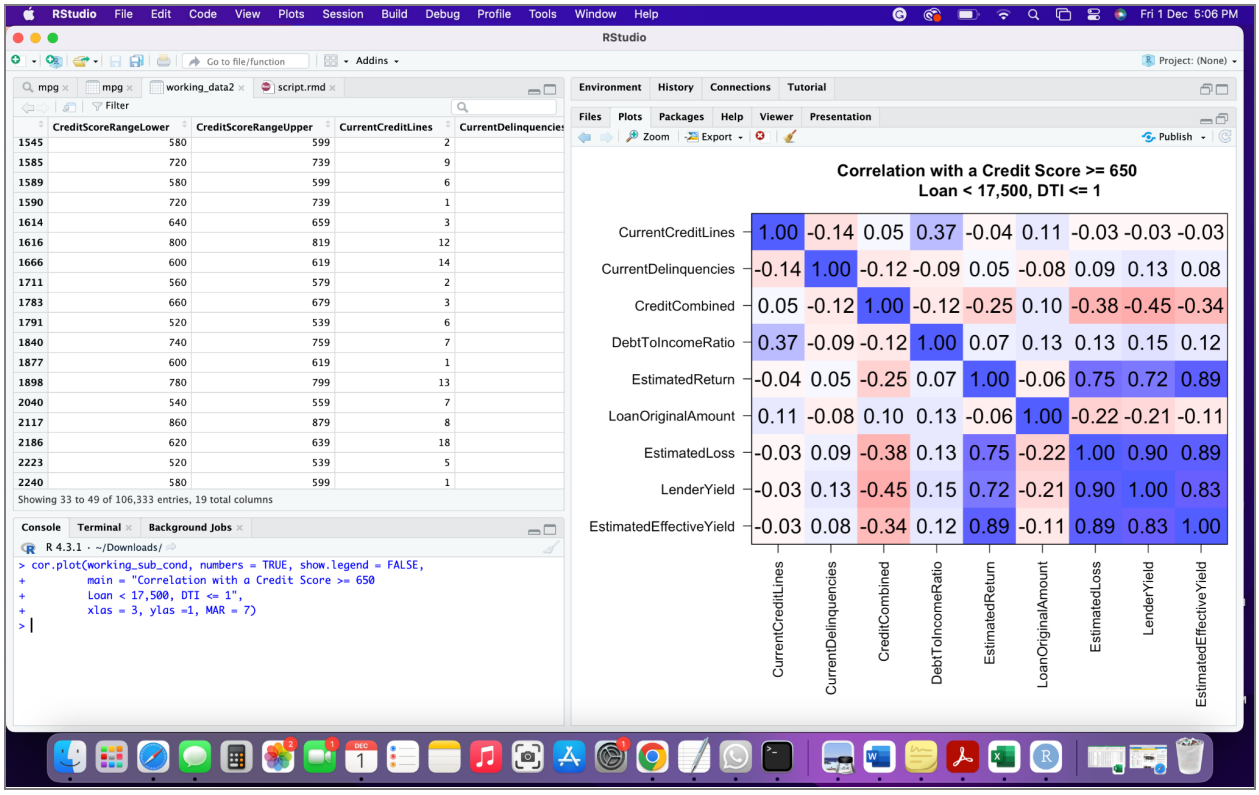
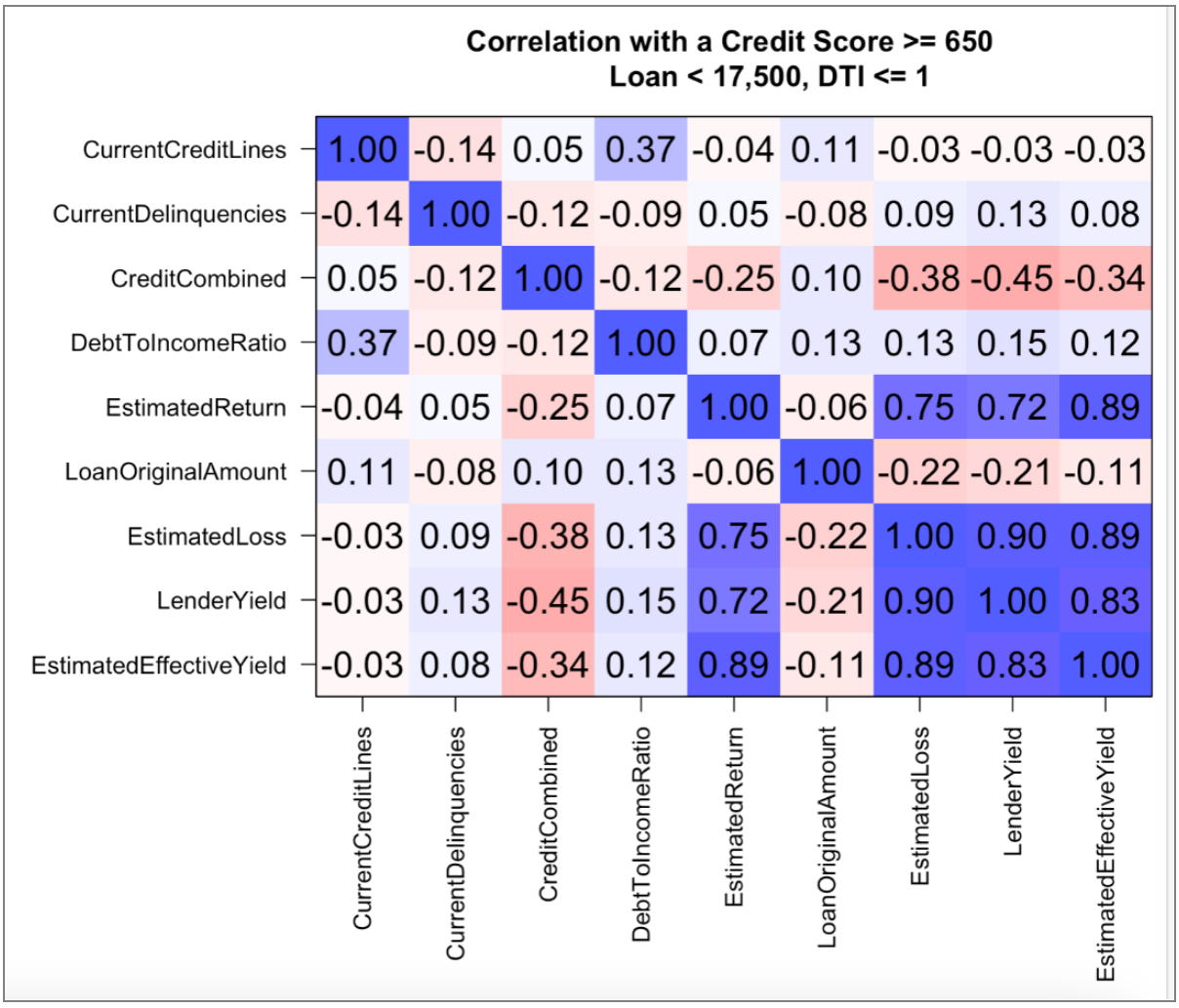


Figure 2. Correlation with Credit Score >= 650, Loan < 17,500, DTI <=1.



Code

```
cor.plot(working_sub_cond, numbers = TRUE, show. legend = FALSE,
+main = "Correlation with a Credit Score >= 650
+Loan < 17,500, DTI <= 1".
+xlas = 3, ylas =1, MAR = 7)
```

There is a difference of 0.01 - 0.27 between the plot without conditions and the plot that has conditions set at a credit score of at least 650, a loan amount less than $17,500 and where the DTI is less than or equal to 1. That change in significance is very predictable because the largest variance in significance is between CurrentCreditLines and DebtToIncomeRatio. These two

variables have a very limited scope of significance in that as CurrentCreditLines rises the significance of DTI falls. DTI shows the most significance when it is 1 or lower. It will show less significance as the value falls and equally so at a the value greater than 1. Putting it simply, the more Lines of credit you have the more debt you obtain. When you reach a debt of greater than your income the number no longer has significant value in regard to the number of credit lines.

3. **Question of Analysis -** How does the density chart analysis reveal the distribution of loans across various loan amount and what insights does the observation offer?

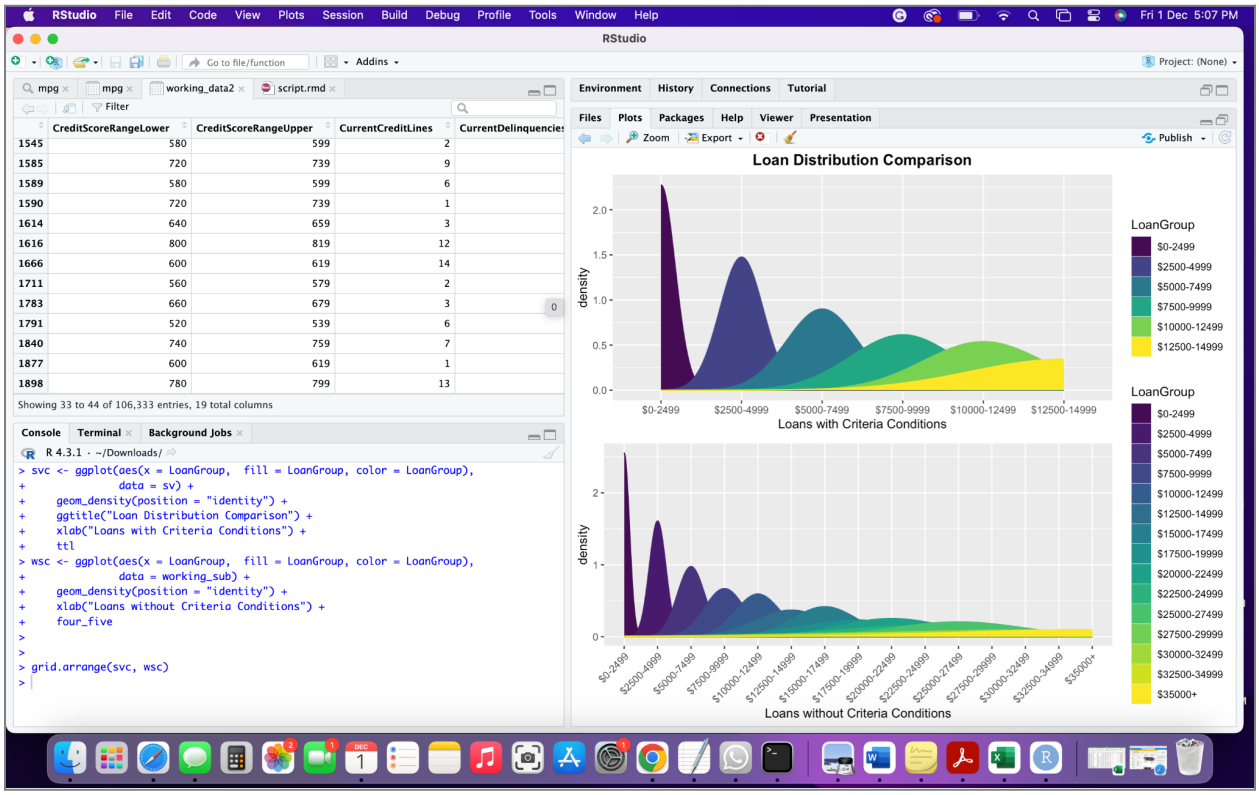
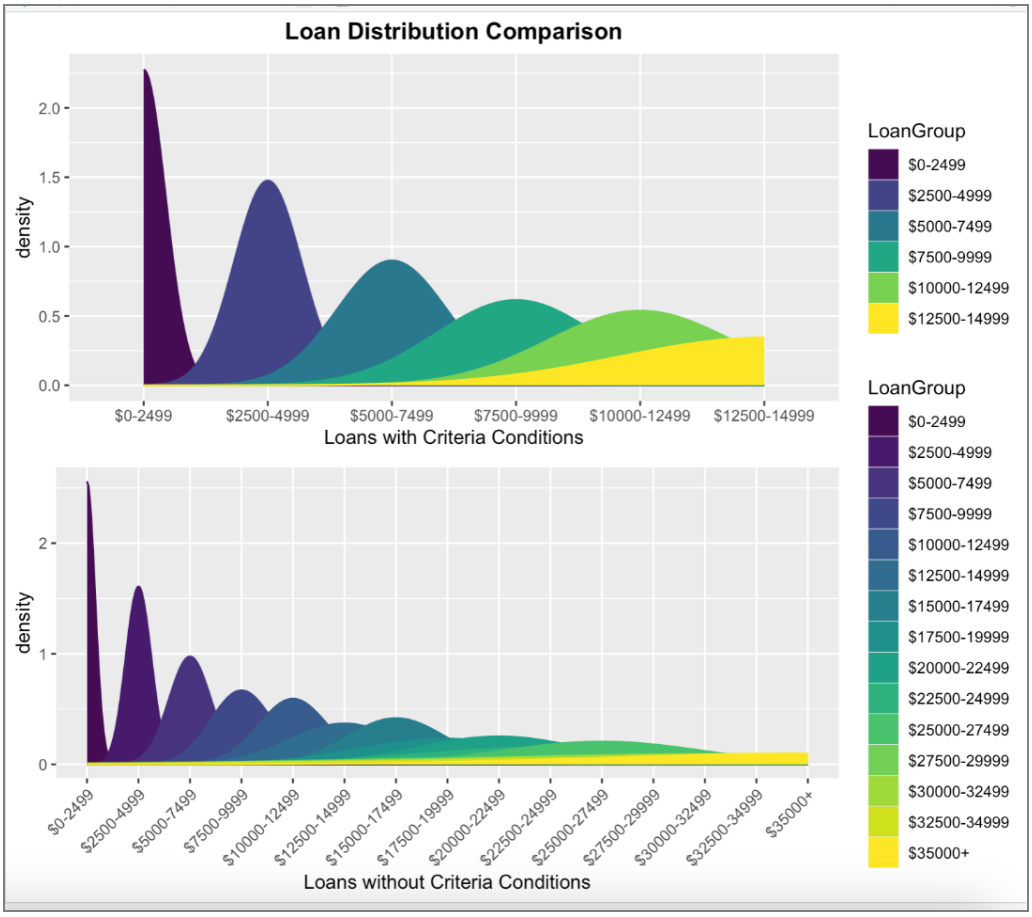


Figure 3. Loan Distribution Comparison.



Code

```
svc <- ggplot(aes(x = LoanGroup, fill = LoanGroup, color = LoanGroup),data = SV) +
geom_density(position ="identity") +
ggtitle("Loan Distribution Comparison") +
 xlab("Loans with Criteria Conditions") +
 ttl
> wsc <- ggplot(aes(x = LoanGroup, fill = LoanGroup, color = LoanGroup),
data = working_sub) +
geom_density(position = "identity") +
xlab("Loans without Criteria Conditions") + four_five
> grid.arrange(svC, Wsc)
```

We used a density chart here to see the number of loans for bin ranges of loan amount. The criteria set to restrict the loan amount to less than $15,000 is very solid here. We see that once the

loan amount reaches $17,500 the plot starts to flatten out rapidly. The criteria established to confine loans below $15,000 becomes evident. Notably, the data reveals a significant shift in loan dynamics as the amount surpasses $17,500, with the plot displaying a rapid flattening trend. This observation raises important questions about the impact of loan thresholds, prompting a deeper exploration into the factors influencing borrower behavior and the effectiveness of such constraints on loan amounts.

4. **Question of Analysis -** How does the scatter plot analysis reveal the distribution of loans across different catagory with respect to the loan amount?

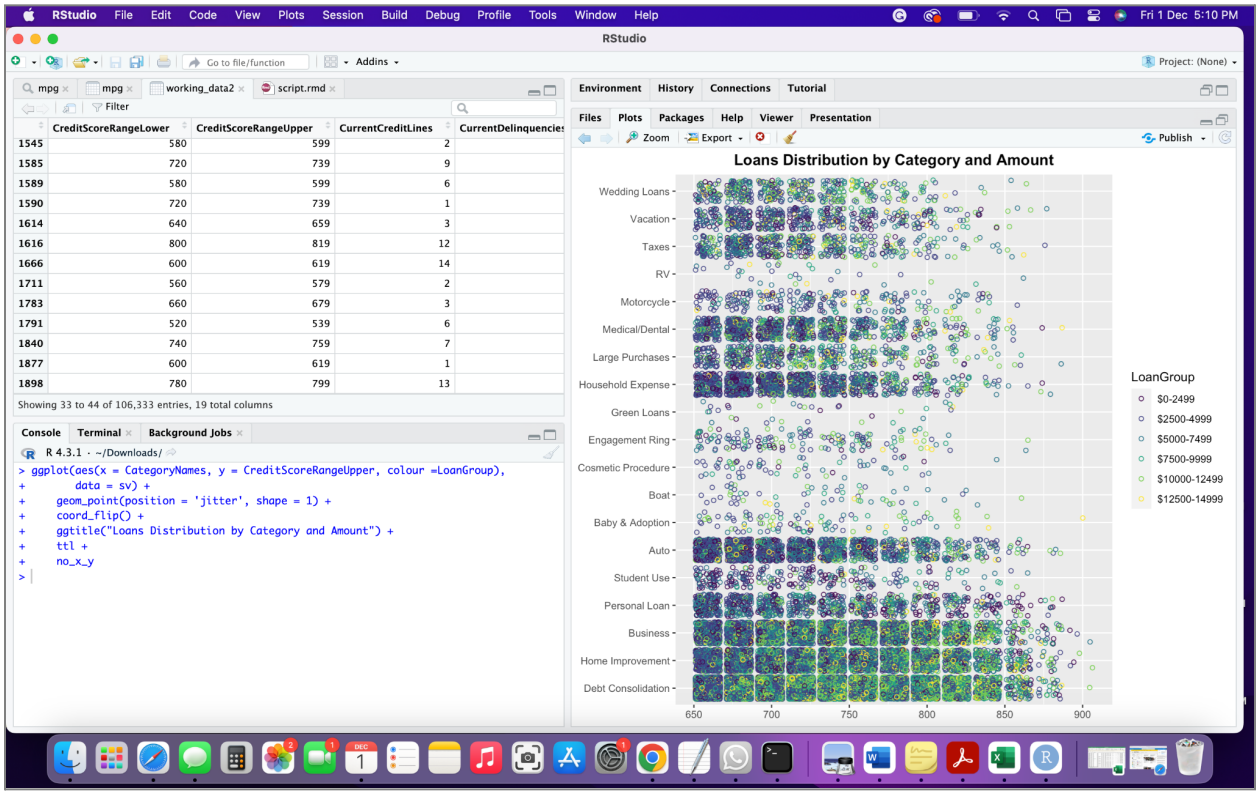


Figure 4. Loan Distribution by Category and Amount.

Code

```
ggplot(aes(x = CategoryNames, y = CreditScoreRangeUpper, colour =LoanGroup),
data = SV) +
geom-point (position = 'jitter', shape = 1) +
coord_flip() +
ggtitle("Loans Distribution by Category and Amount") +
ttl +
no_x_y
```

Because of the non-linear relationship between our loan requestor and Prosper loan approval criteria we take an approach of narrowing the scope according to what we have learned about the requestor. Here we see the dispersion of loan for those whose credit score is greater than 650,

debt to income ratio is less than .30, income is greater than $25,000 and loan request is less than $15,000. Across distinctlyvdefined loan categories we see that the primary loan requests are for Business, Debt Consolidation, or Home Improvement.

5. **Question of Analysis -** What are the Loan distribution characteristics and outlier patterns among different Category types for diffrent customers?

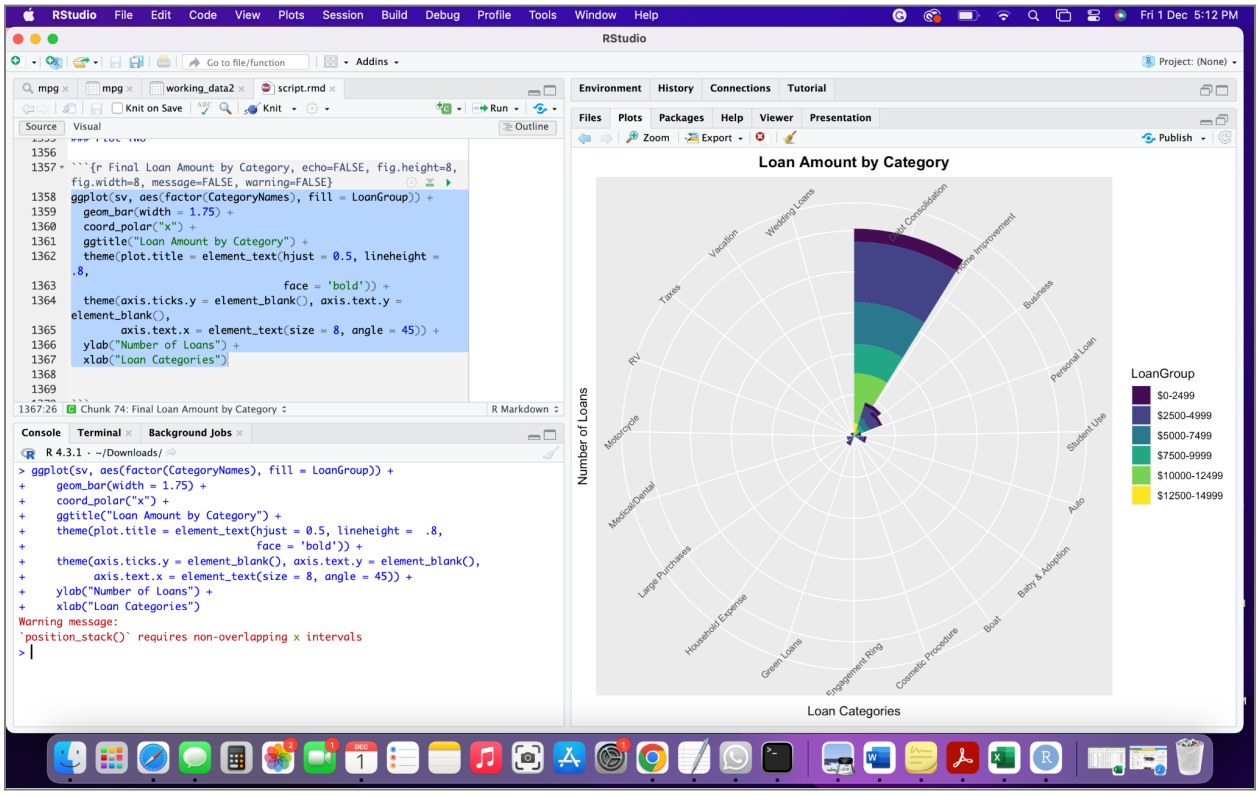


Figure 5. Loan Amount by Category.

This plot was chosen to answer the questions of what(loan amount), and why(loan category). We know the answer to the who of Prosper loans this plot shows that we will be more successful in the loan application process if the loan is for debt consolidation. However, if the loan is for home improvement or business and the amount requested is less than $7,500 we will be much more likely to be approved than for other loan categories

Part F: Statistical summary

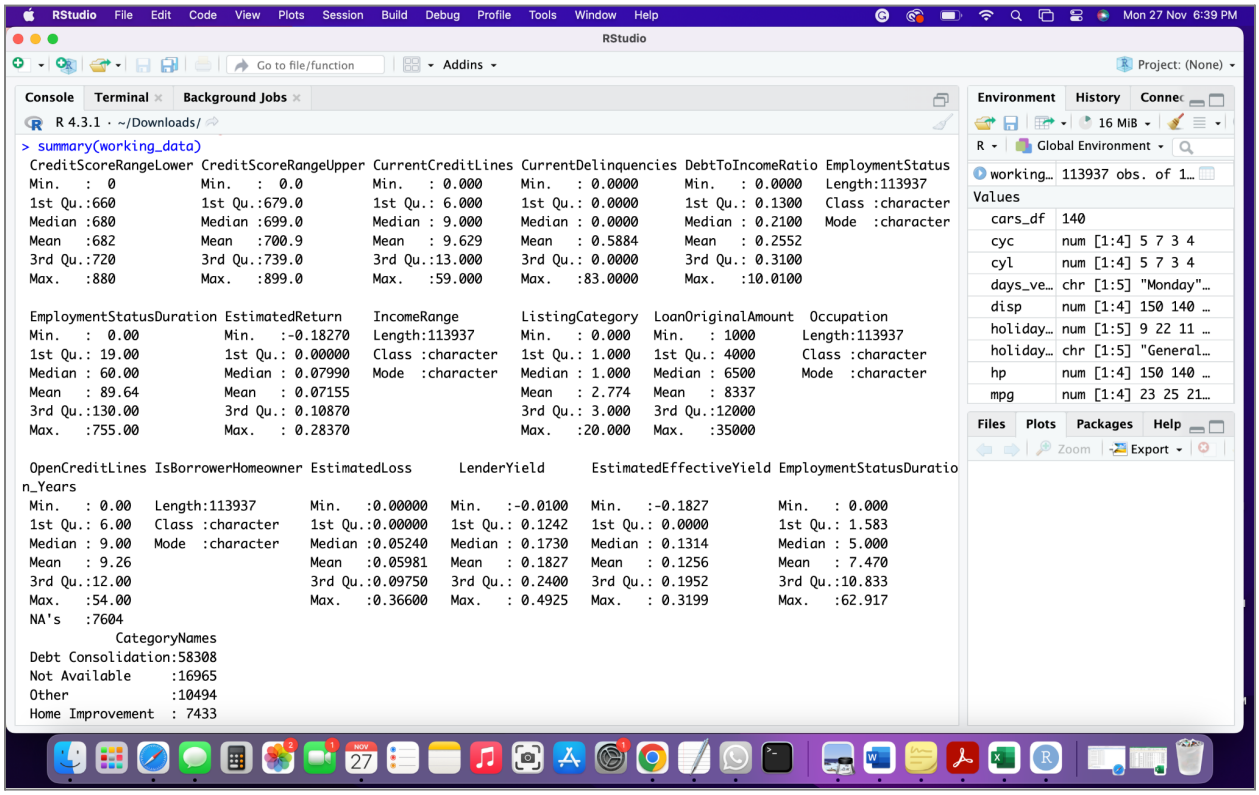


Figure - Descriptive Statistics for assignment data.

The dataset encompasses a comprehensive set of information, including credit scores, delinquencies, loan amounts, current credit lines, and various other fields. This richness in data allows for a holistic understanding of the borrowers' financial profiles.

By exploring various fields in the dataset, lenders and analysts can derive insights into patterns, trends, and potential risk factors that may impact the creditworthiness of individuals.

Credit Delinquencies:

The fact that the majority of the data shows zero current delinquencies is a positive indicator of

financial health among the individuals in the dataset. This suggests that a significant portion of the borrowers is managing their credit responsibilities well.

While the mean value is low (0.5884), it's essential to note that the presence of a maximum value of 83 indicates a tail in the distribution, suggesting that there are outliers or extreme cases where borrowers are facing a higher number of current delinquencies. Investigating these outliers could provide valuable insights into potential risk factors.

Loan Amounts:

The average loan amount of around 8337 provides a central tendency for the dataset. However, the fact that the median is lower than the mean suggests a positively skewed distribution. This implies that there are a few loans with significantly higher amounts, pulling the mean upward. Understanding the distribution of loan amounts is crucial for lenders to assess the range and diversity of loans. It's also essential for risk management, as larger loans may pose different risk profiles.

Credit Scores:

The wide range of credit scores, with the majority falling within the mid to high range, indicates a diverse creditworthiness among borrowers. This diversity is essential for lenders to tailor their risk assessment strategies and offerings.

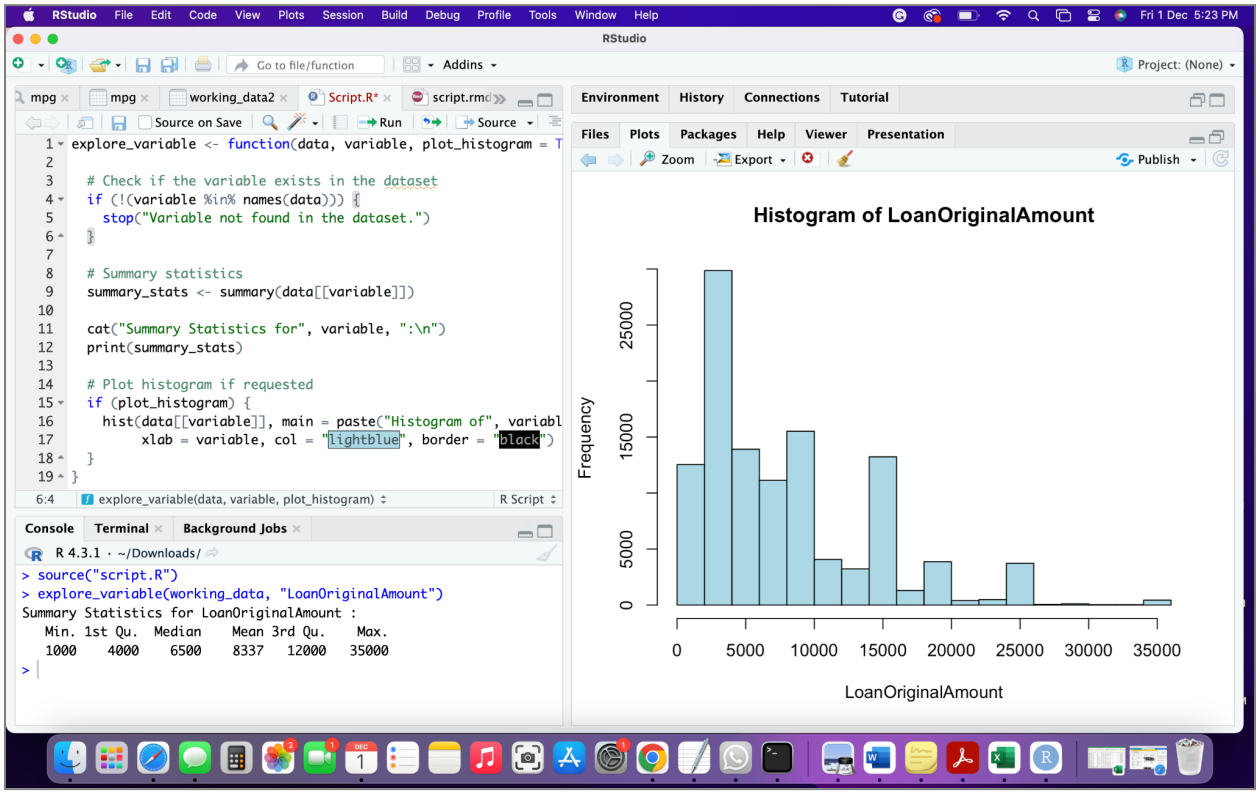
The variation in credit scores also emphasizes the importance of considering other factors, such as income, employment history, and debt-to-income ratio, for a more nuanced evaluation of an individual's financial health.

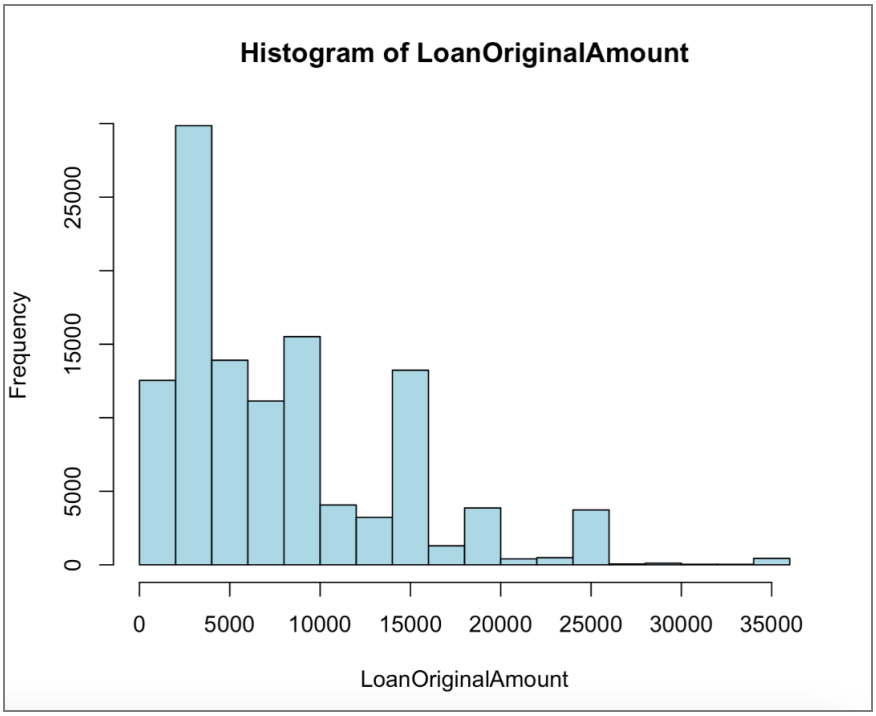
Current Credit Lines:

The variation in current credit lines, with an average around 9.629, highlights the diversity in the available credit among the individuals in the dataset. This information is valuable for understanding how borrowers utilize and manage their existing credit.

In conclusion, this dataset provides a wealth of information that can be leveraged to make informed decisions in the lending and financial domain. Exploring individual fields and their relationships can lead to deeper insights, helping stakeholders better understand the factors influencing creditworthiness and financial stability.

Script/Functions





Code:

```
explore_variable <- function(data, variable, plot_histogram = TRUE) {

  # Check if the variable exists in the dataset
  if (!(variable %in% names(data))) {
    stop("Variable not found in the dataset.")
  }

  # Summary statistics
  summary_stats <- summary(data[[variable]])

  cat("Summary Statistics for", variable, ":\n")
  print(summary_stats)

  # Plot histogram if requested
  if (plot_histogram) {
    hist(data[[variable]], main = paste("Histogram of", variable),
         xlab = variable, col = "lightblue", border = "black")
  }

}
```

let's create a user-defined function in R that could be useful for studying this dataset. This function will calculate some basic statistics and visualizations for the specified variable. We'll call it explore_variable. This function will take three parameters: the dataset, the variable to explore, and a boolean flag indicating whether to plot a histogram or not.

This function takes a dataset (data), the variable you want to explore (variable), and an optional argument plot_histogram to decide whether to plot a histogram or not. It then prints summary statistics for the variable and, if specified, plots a histogram.

You can call this function for any variable in your dataset, and it will give you a quick overview of the variable's distribution and summary statistics.

CONCLUSION:

The goal of this study was to answer the questions of who, what, why and how Prosper grants loans. The dataset used holds 113,937 entries across 81 variables. The total number of variables used across this study both original and created totaled 26.

The choice of those specific variables was made to give the most meaning to the data by using a more traditional approach to the criteria of the loan process.

From there each was evaluated parallel to the goals of the study. I was able to determine that while certain occupations have a higher volume of loans, all occupations were represented if they were within the parameters that were defined as preferred.

The final preferred criteria is:

- All Occupations
- Credit Score >= 650

- Current Delinquencies < 1
- Debt to Income Ratio < .30
- Income > $25,000
- Loan Amount < $15,000
- Loan Category of Debt Consolidation, Business or Home Improvement.

There are so many more insights to be gained from this data set and I have just scratched the surface. Other things to explore could be how Prosper defines a good risk or how many of Propser's customers are repeat customers. It would also be interesting to take this data and develop a model to help investors determine the best loans to back and what loans to avoid.