Communicate Data Findings
Loan Data from Prosper

A report with me **exploratory** data analysis

Advanced Data Analysis

egFWD - Udacity

BY:

Mohamed Najm

mohamednajm250@gmail.com

Introduction

This the third project for the advanced data analysis and it's about visualization and using Loan Data from Prosper dataset and doing an Investigation for it and gets some insights.

Dataset description

This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others.

This data dictionary explains the variables in the data set. we are not expected to explore all of the variables in the dataset! Focus your exploration on about 10-15 of them.

So, going through this notebook I tried to visual the data and get insights from it and want to use it to answer some questions like:

- What is the frequency of EmploymentStatus value exist?

- What is the frequency of IncomeRange values exist?

- What's the most status for a loan?

- What is the Distribution of BorrowerRate?

- What is the Distribution of Investors?

- What is the Distribution of LoanOriginationDate?

- What is the distribution of loan over days, months and years?

- What is the Distribution of LoanOriginalAmount?

- Is there any correlation between the features?

And more question.

What I did?

First, going through data wrangling process which As we known include three phases: gathering, Assessing and define and cleaning.

- Gathering: the data by downloading it and import it using pandas.

- Assessing and define: Assessing the dataset by two ways: Visually and programmatically to explore it and then define all what it need to cleaned which will be shown later.

- Data cleaning: in this phase clean all what mentioned above.

Here are some problems I found at the dataset:

- 871 duplicated rows at ListingNumber, ListingKey and LoanKey

- many columns with null values

- many columns we do not need so drop them

- some columns need to turn its type into date type and maybe split it into columns if needed like: ListingCreationDate, ClosedDate and DateCreditPulled

- some categries columns need to apply get dummies on it

- split some columns into two columns like: LoanOriginationQuarter

- remove the ($) from the columns IncomeRange

Second, extract the features of interest

Which are: Term, LoanStatus, BorrowerRate, EmploymentStatus, IncomeRange, Investors, Recommendations, LoanOriginationQuarter, LoanOriginationDate, LoanOriginalAmount, StatedMonthlyIncome, DelinquenciesLast7Years, LenderYield, ProsperRating (Alpha), IsBorrowerHomeowner, PercentFunded, ListingCategory(numeric)

Third, Descriptive Statistics

Let's starting with the summary statistics for the numerical and categorical features before explanatory analysis.

1- The Numerical varibles

2- The categorical variables

What is the structure of your dataset?

After cleaning the data, there are 110723 rows of data and after drop columns which not in are interest so, there are 17 columns which are: Term, LoanStatus, BorrowerRate, EmploymentStatus, IncomeRange, Investors, Recommendations, LoanOriginationQuarter, LoanOriginationDate, LoanOriginalAmount, StatedMonthlyIncome, DelinquenciesLast7Years, LenderYield, IsBorrowerHomeowner, ListingCategory (numeric).

12 numeical cloumns and 5 non numerical ones which are: LoanStatus ,EmploymentStatus ,IncomeRange, LoanOriginationQuarter and the last one is a datetime64 type

What features in the dataset do you think will help support your investigation?

- I think that the borrowers rating will have the highest impact on chances of default. Also I expect that LenderYield and loan amount will play a major role and maybe the category of credit. Prosper rating will depend on stated income and employment status.

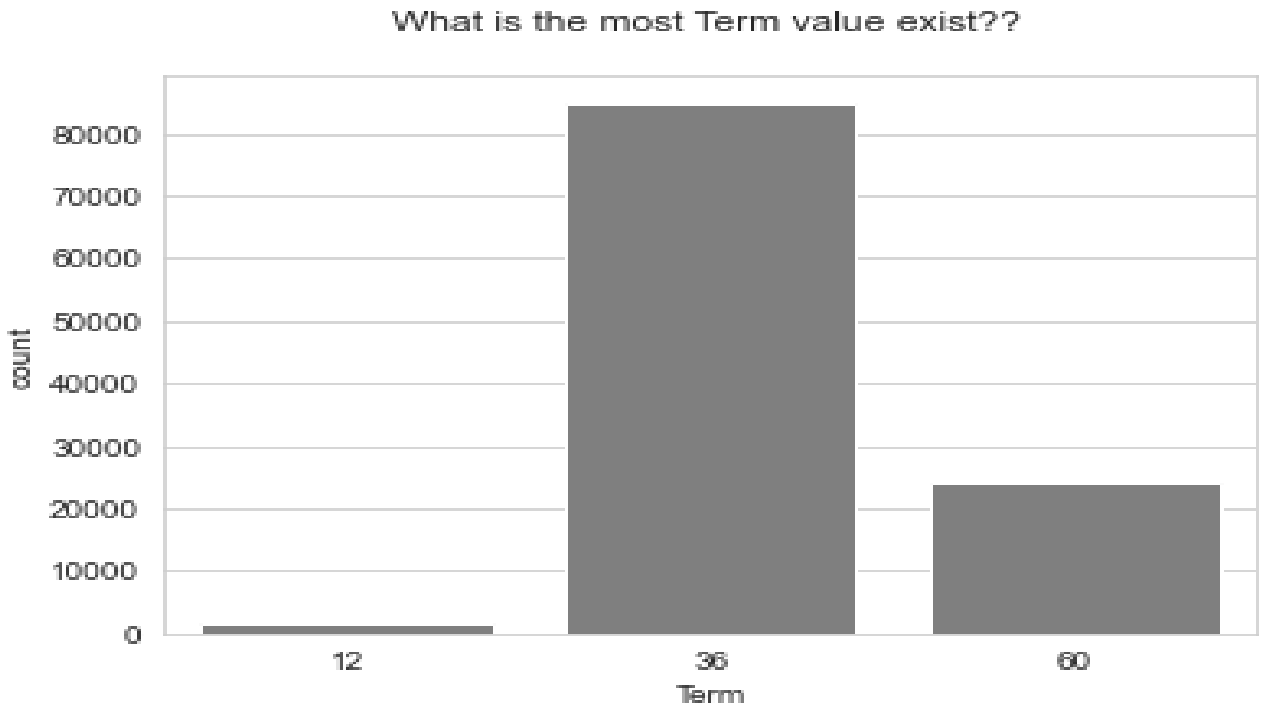
And finally Exploration [the main target]

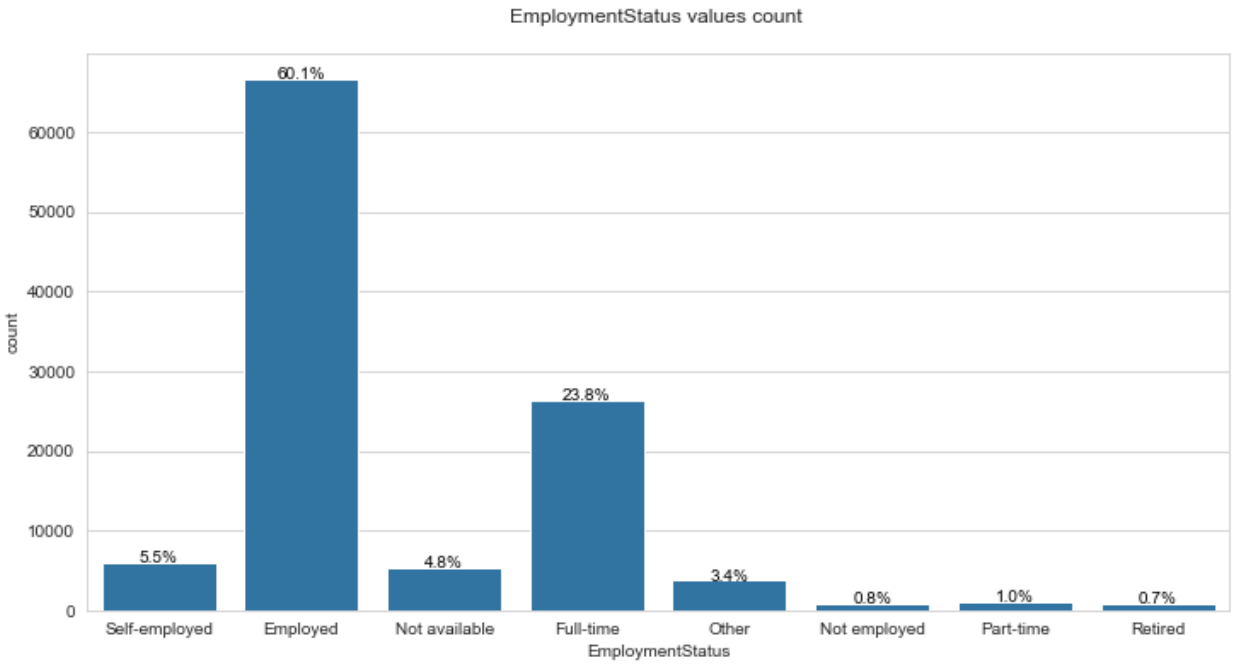
- Exploration the data to get some insights from graphs to support them, the exploration process include three phases:

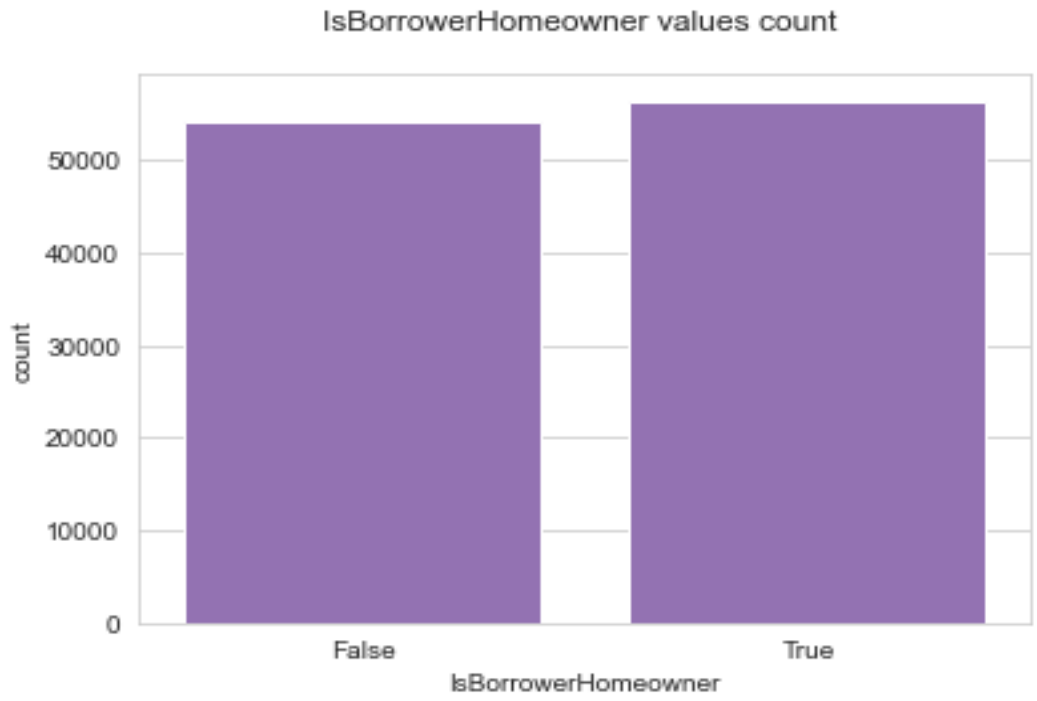
- Data Analysis: get insight and answer question using the dataset.

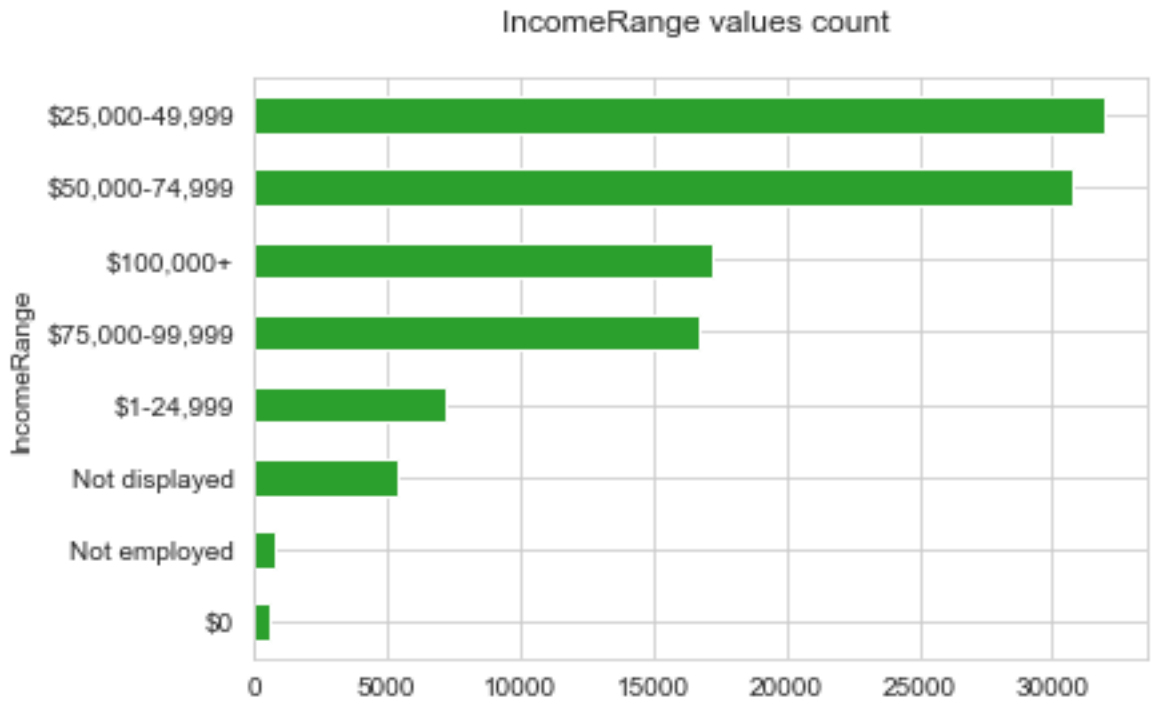
1- Univariate Exploration

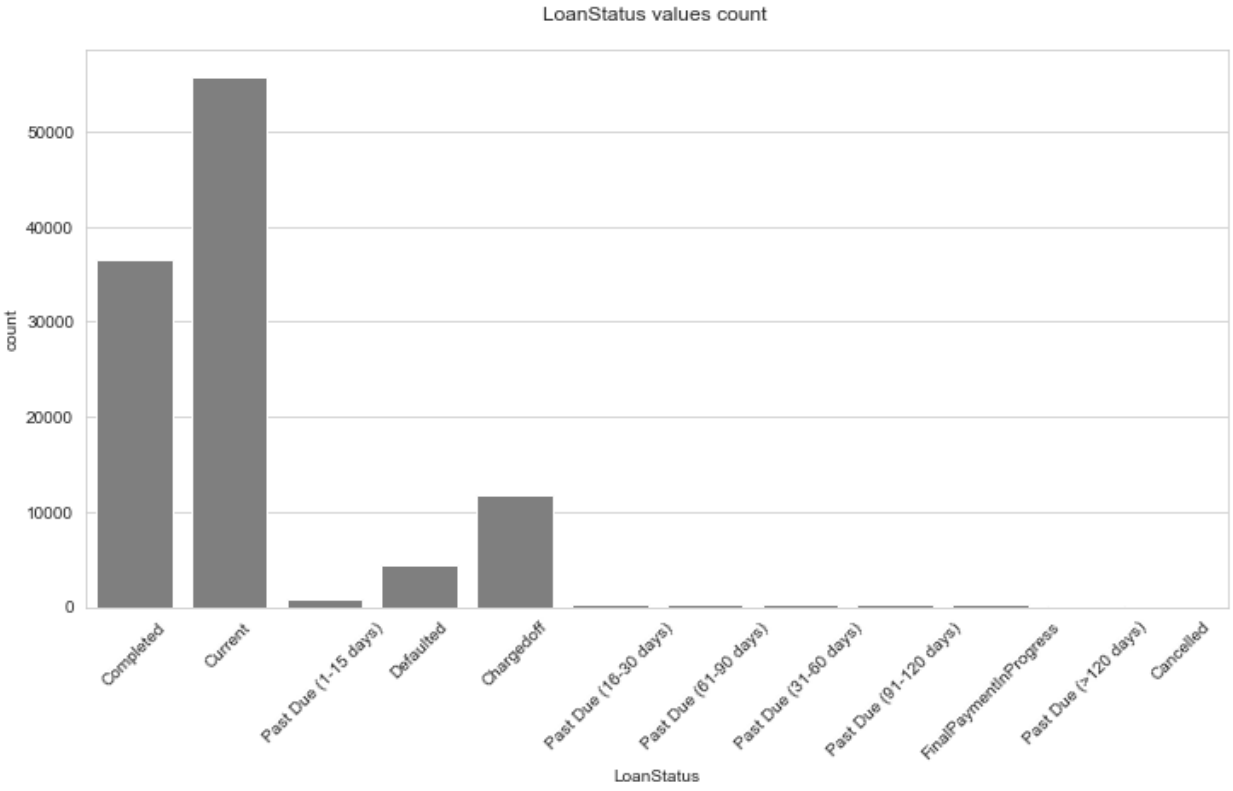
In this section, investigate distributions of individual variables. If you see unusual points or outliers, take a deeper look to clean things up and prepare yourself to look at relationships between variables.

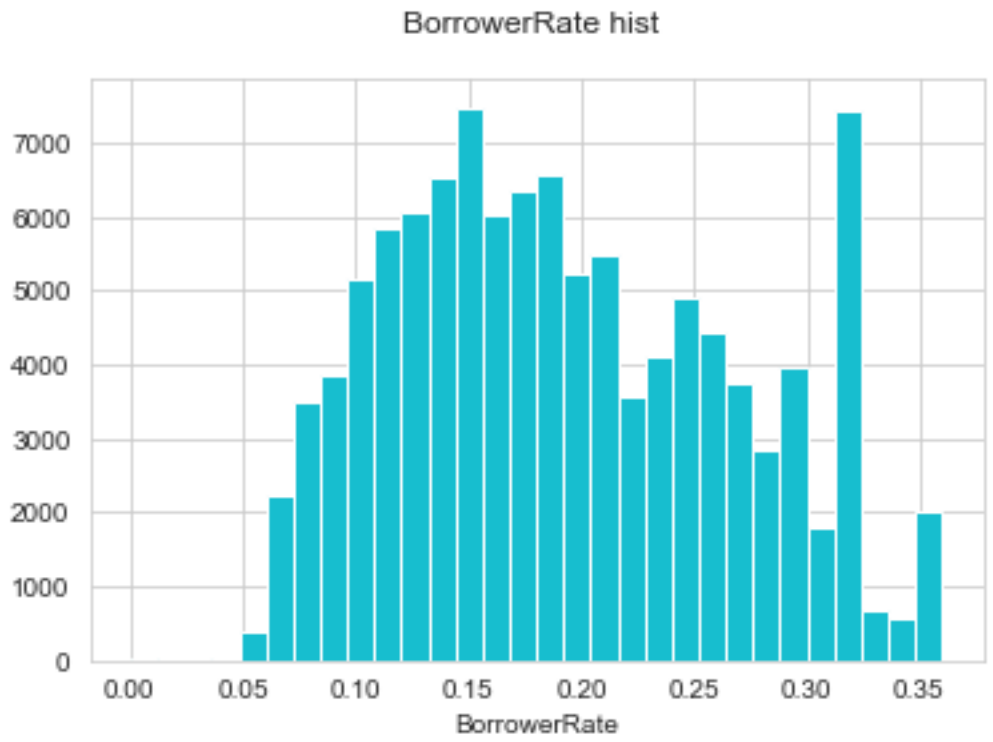


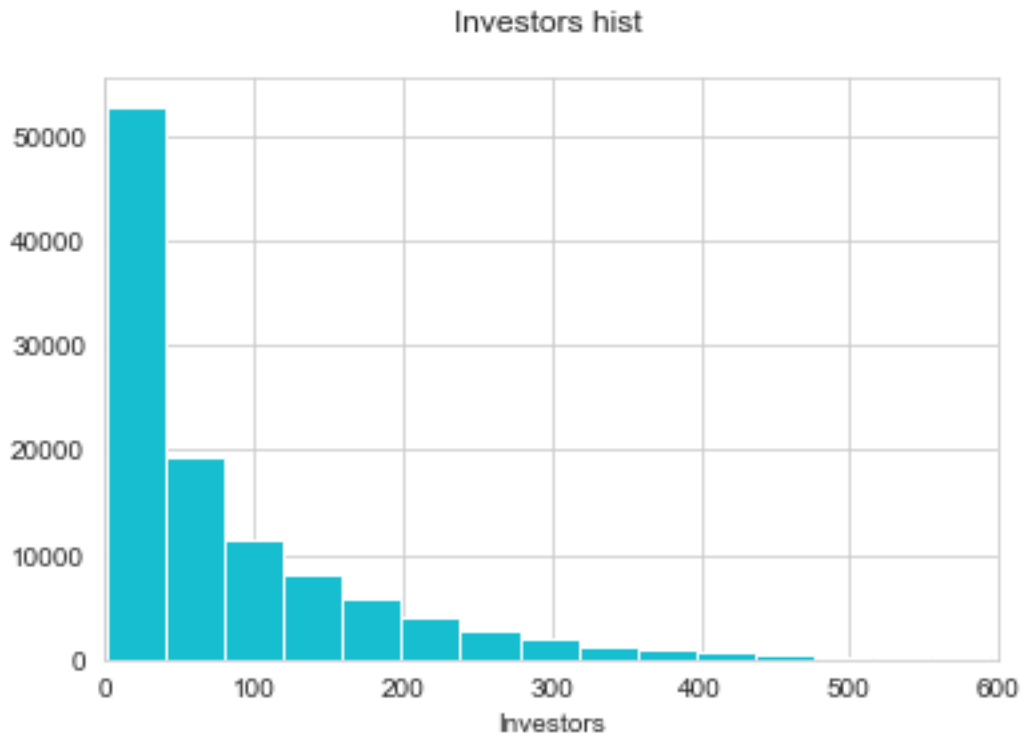

EmploymentStatus values count
60.1%
23.8%
5.5%
4.8%
3.4%
0.8%
1.0%
0.7%
60000
50000
40000
30000
20000
10000
0
count
Self-employed
Employed
Not available
Full-time
Other
Not employed
Part-time
Retired
EmploymentStatus

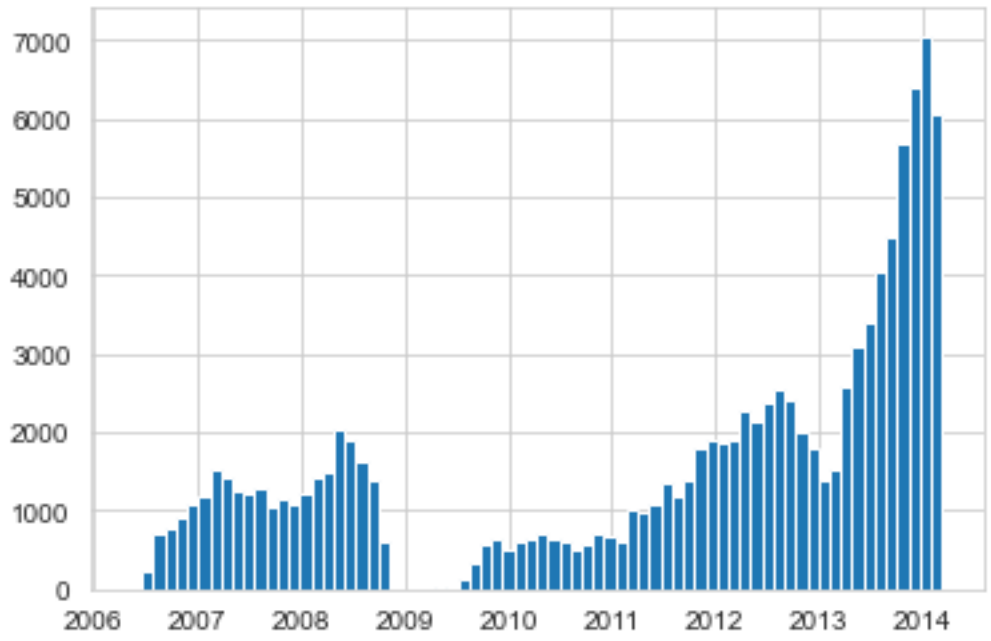

IsBorrowerHomeowner values count
50000
40000
30000
20000
10000
0
count
False
True
IsBorrowerHomeowner


IncomeRange values count
IncomeRange
$25,000-49,999
$50,000-74,999
$100,000+
$75,000-99,999
$1-24,999
Not displayed
Not employed
$0
0
5000
10000
15000
20000
25000
30000

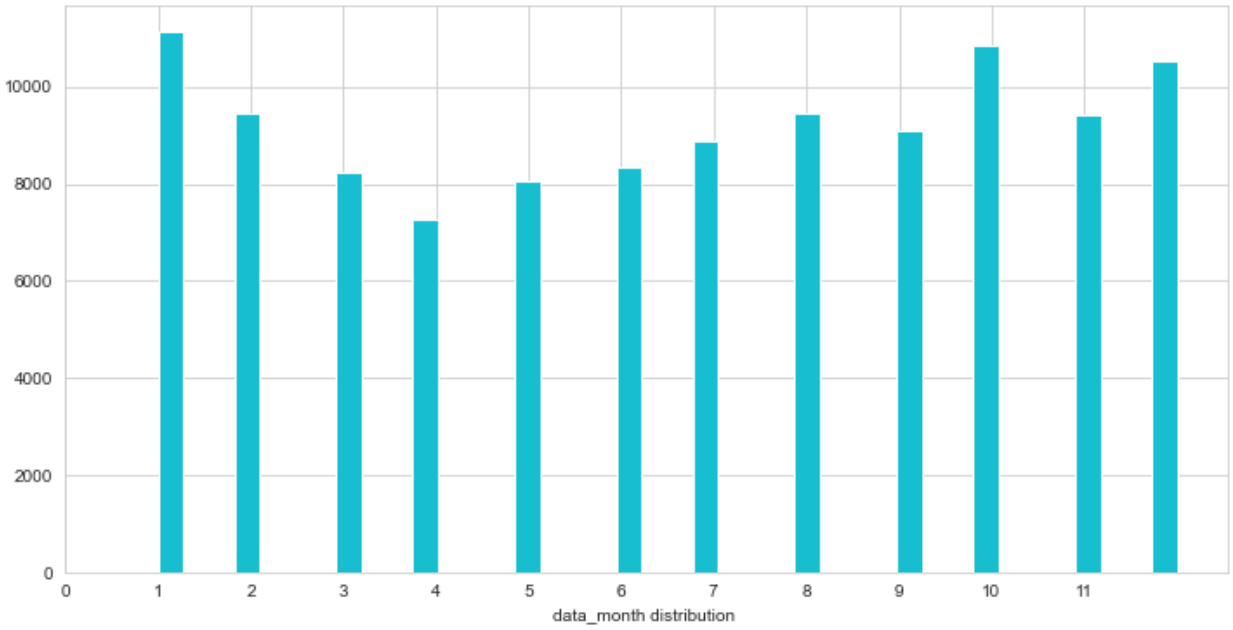

LoanStatus values count
count
50000
40000
30000
20000
10000
0
Completed
Current
Past Due (1-15 days)
Defaulted
Chargedoff
Past Due (16-30 days)
Past Due (61-90 days)
Past Due (31-60 days)
Past Due (91-120 days)
FinalPaymentinProgress
Past Due (>120 days)
Cancelled
LoanStatus

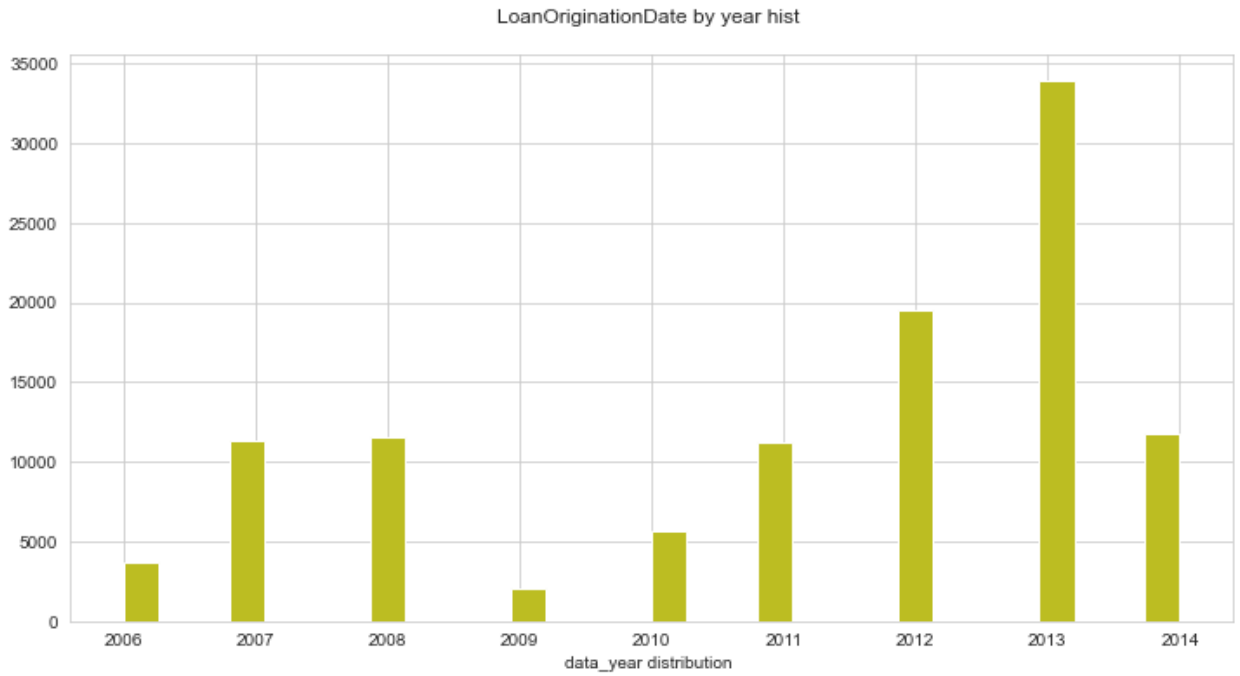

BorrowerRate hist
7000
6000
5000
4000
3000
2000
1000
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
BorrowerRate

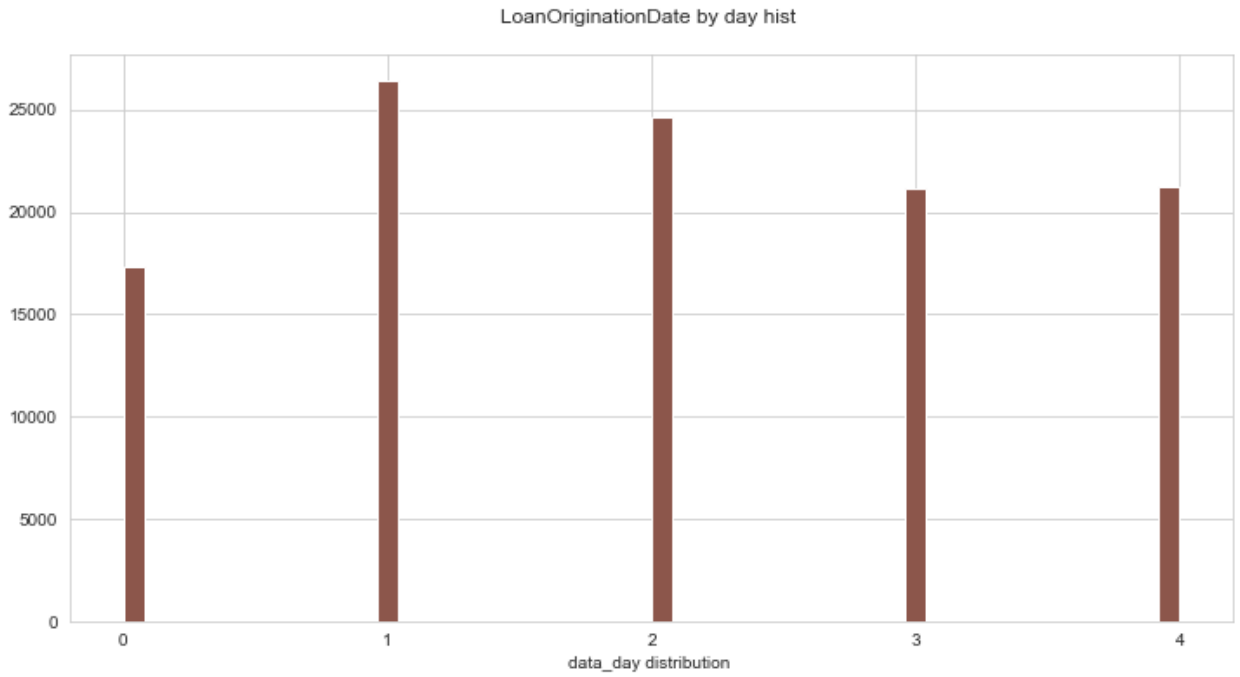

Investors hist
50000
40000
30000
20000
10000
0
0
100
200
300
400
500
600
Investors

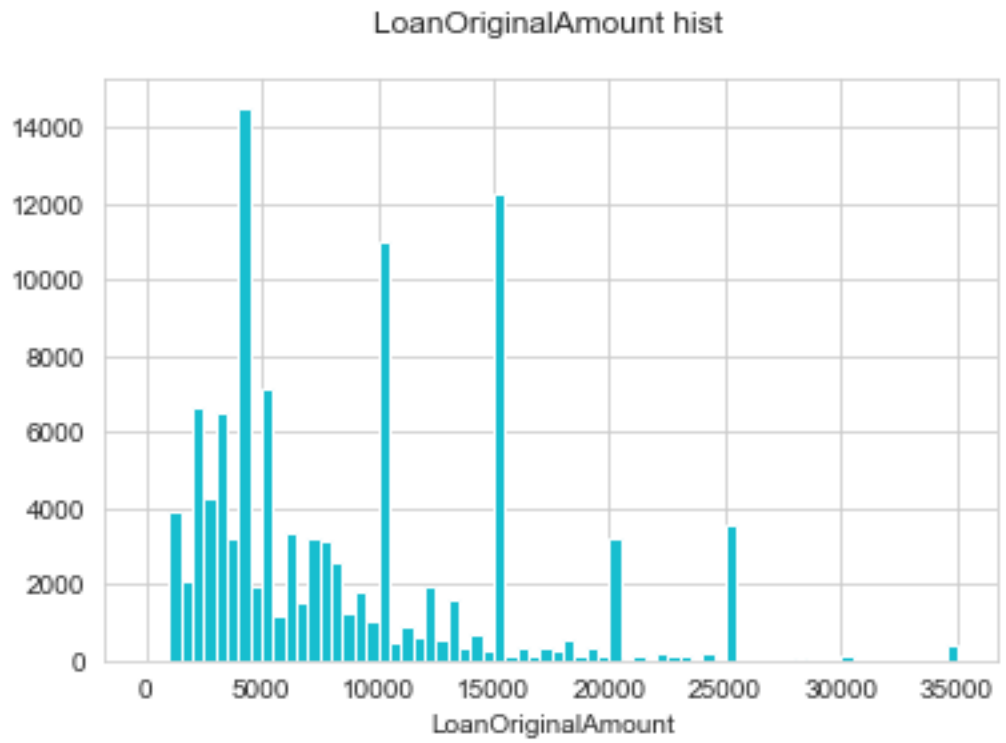


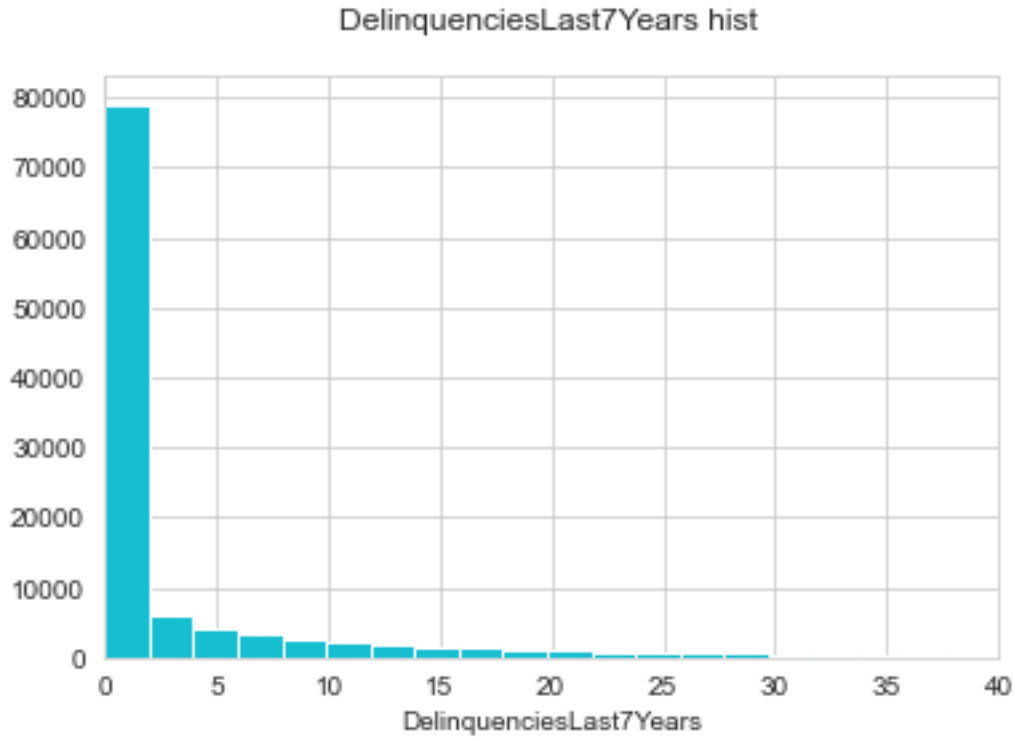
LoanOriginationDate by month hist




LoanOriginationDate by year hist
35000
30000
25000
20000
15000
10000
5000
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
data_year distribution


LoanOriginationDate by day hist
25000
20000
15000
10000
5000
0
0
1
2
3
4
data_day distribution


LoanOriginalAmount hist
14000
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000
35000
LoanOriginalAmount


DelinquenciesLast7Years hist
80000
70000
60000
50000
40000
30000
20000
10000
0
0
5
10
15
20
25
30
35
40
DelinquenciesLast7Years

- Univariate_Exploration_Conclusions

first part of the exploration is finished, some questions have been answeres above by graphs like:

What is the frequance of EmploymentStatus value exist ?

What is the frequance of IncomeRange values exist ?

What's the most status for a loan ?

what is the Distribution of BorrowerRate?

what is the Distribution of Investors ?

what is the Distribution of LoanOriginationDate ?

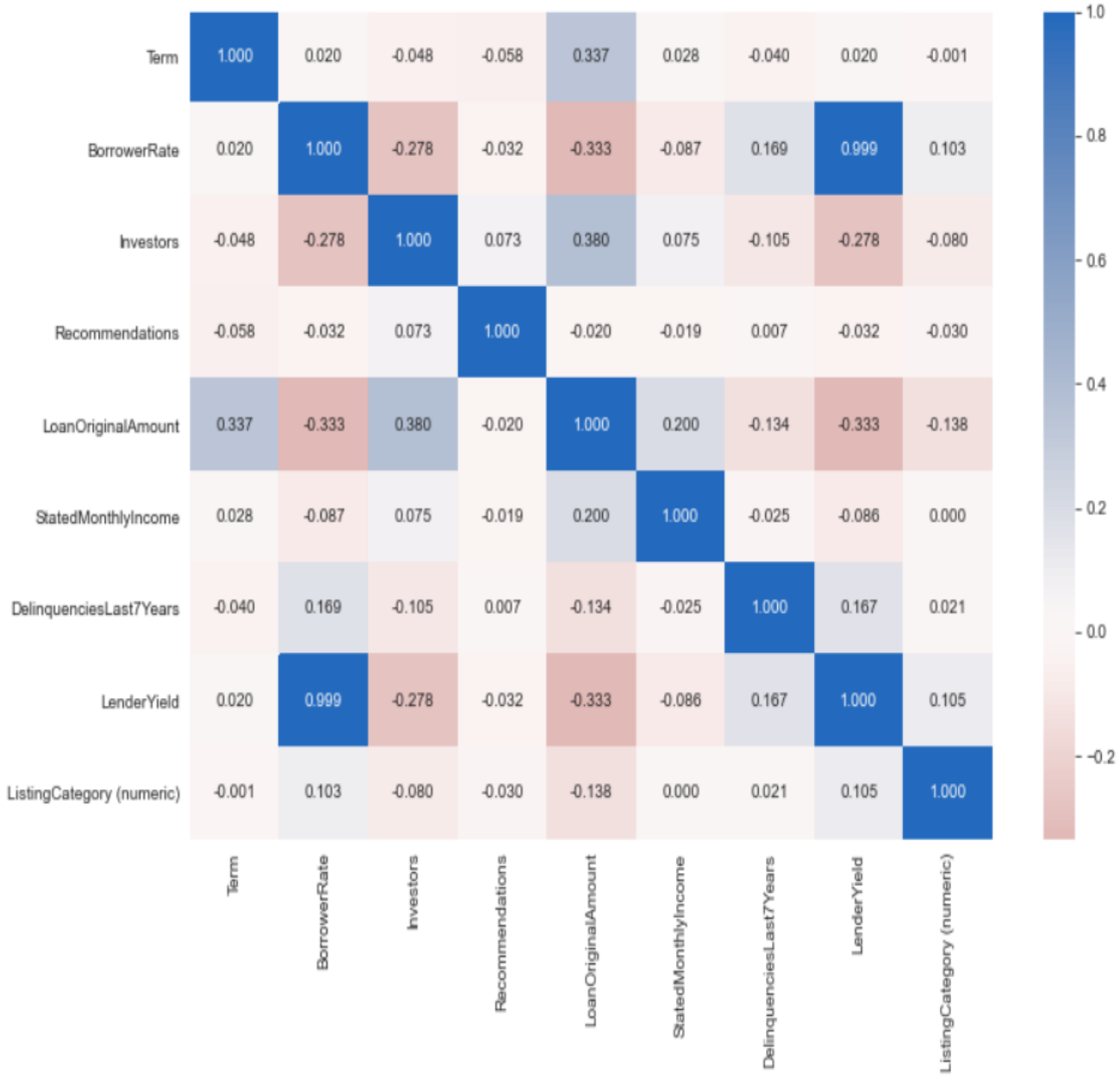
What is the distribution od loan over days, months and years?

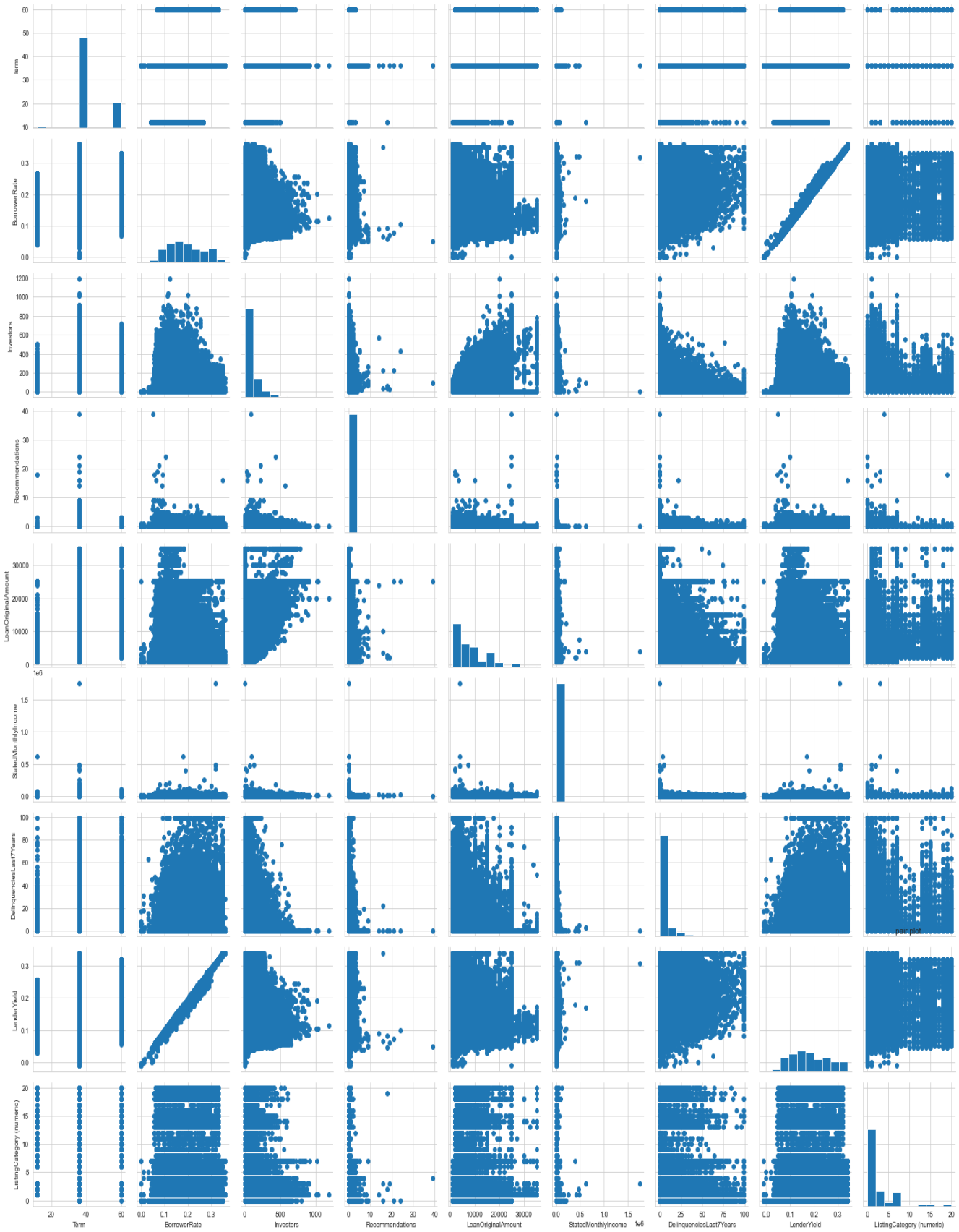
What is the Distribution of LoanOriginalAmount ?

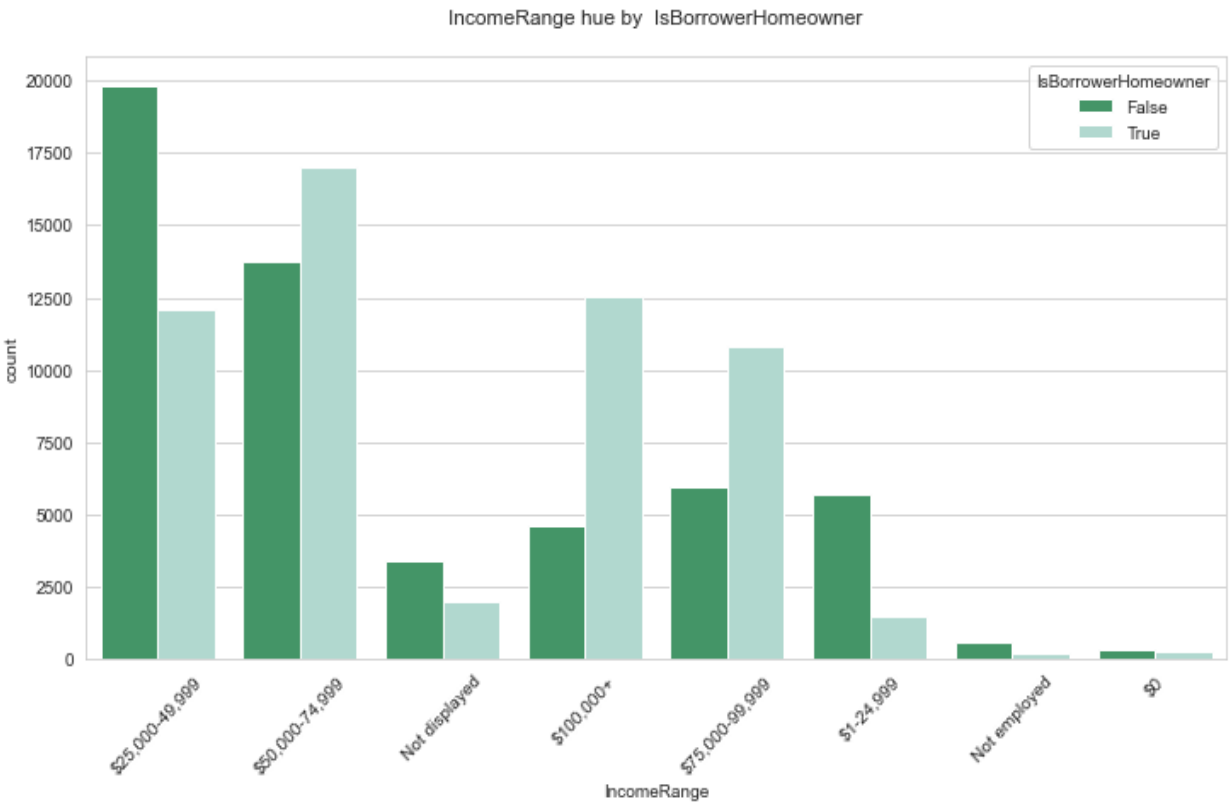
2- Bivariate Exploration

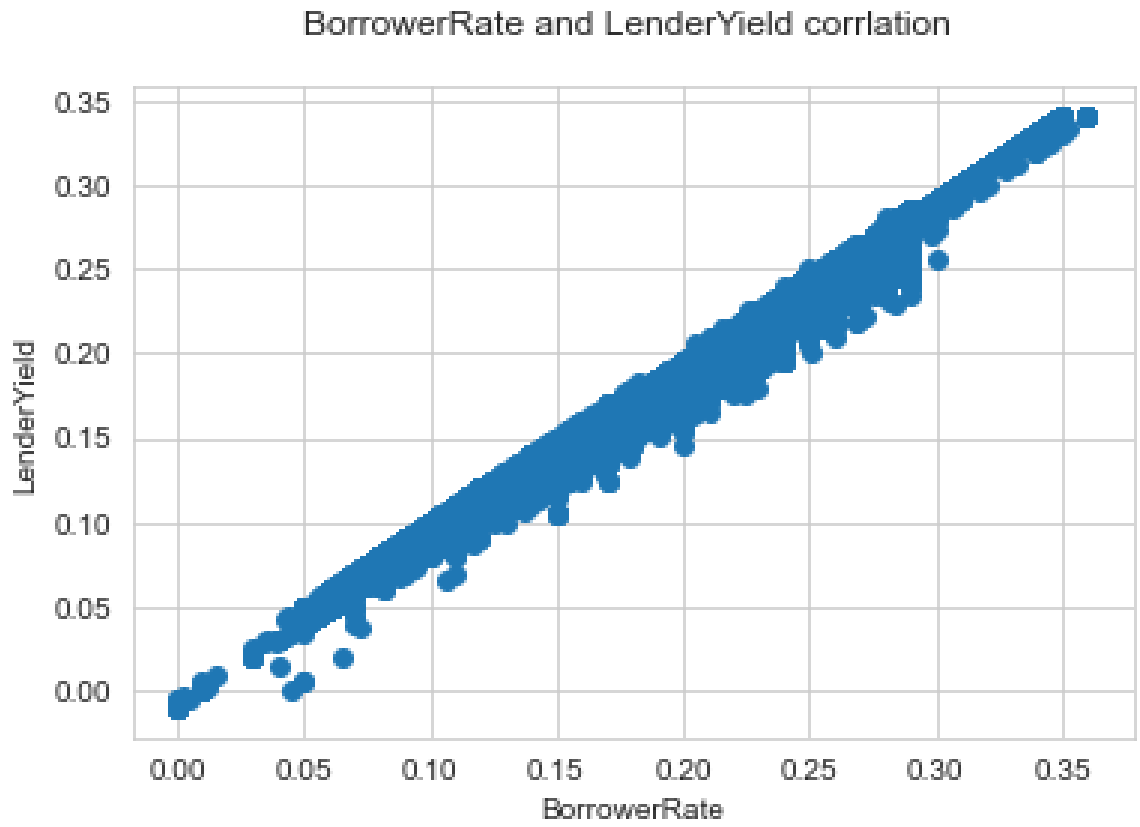
Now, To start off with bivarite exploration,

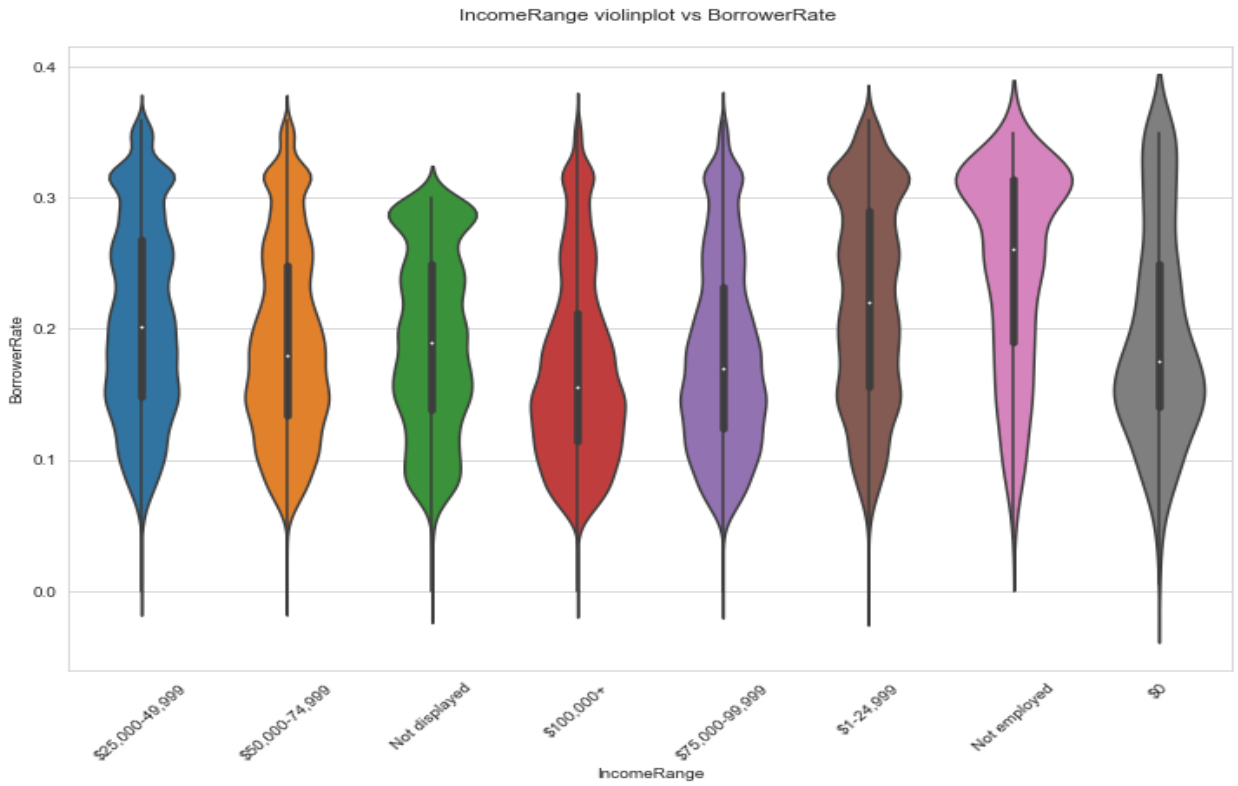
In this section, investigate relationships between pairs of variables in your data.

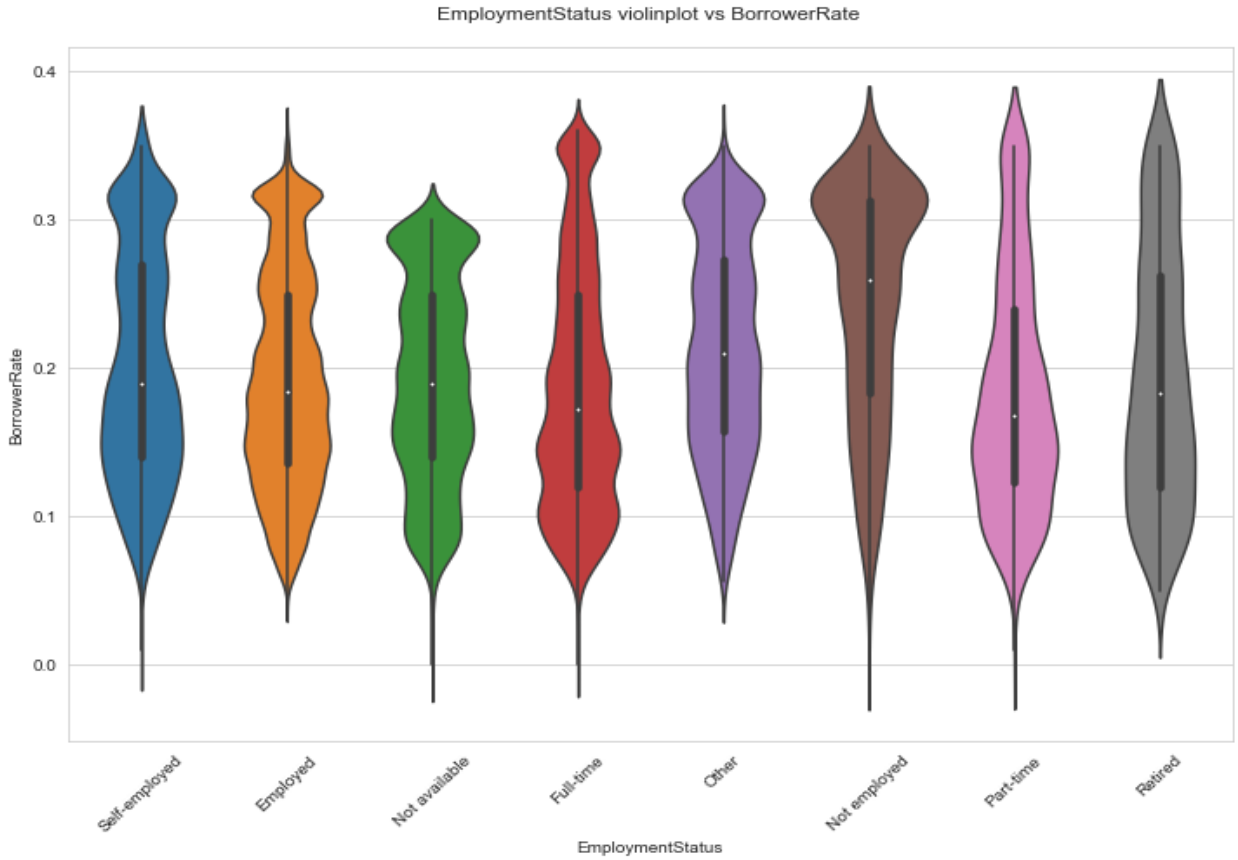


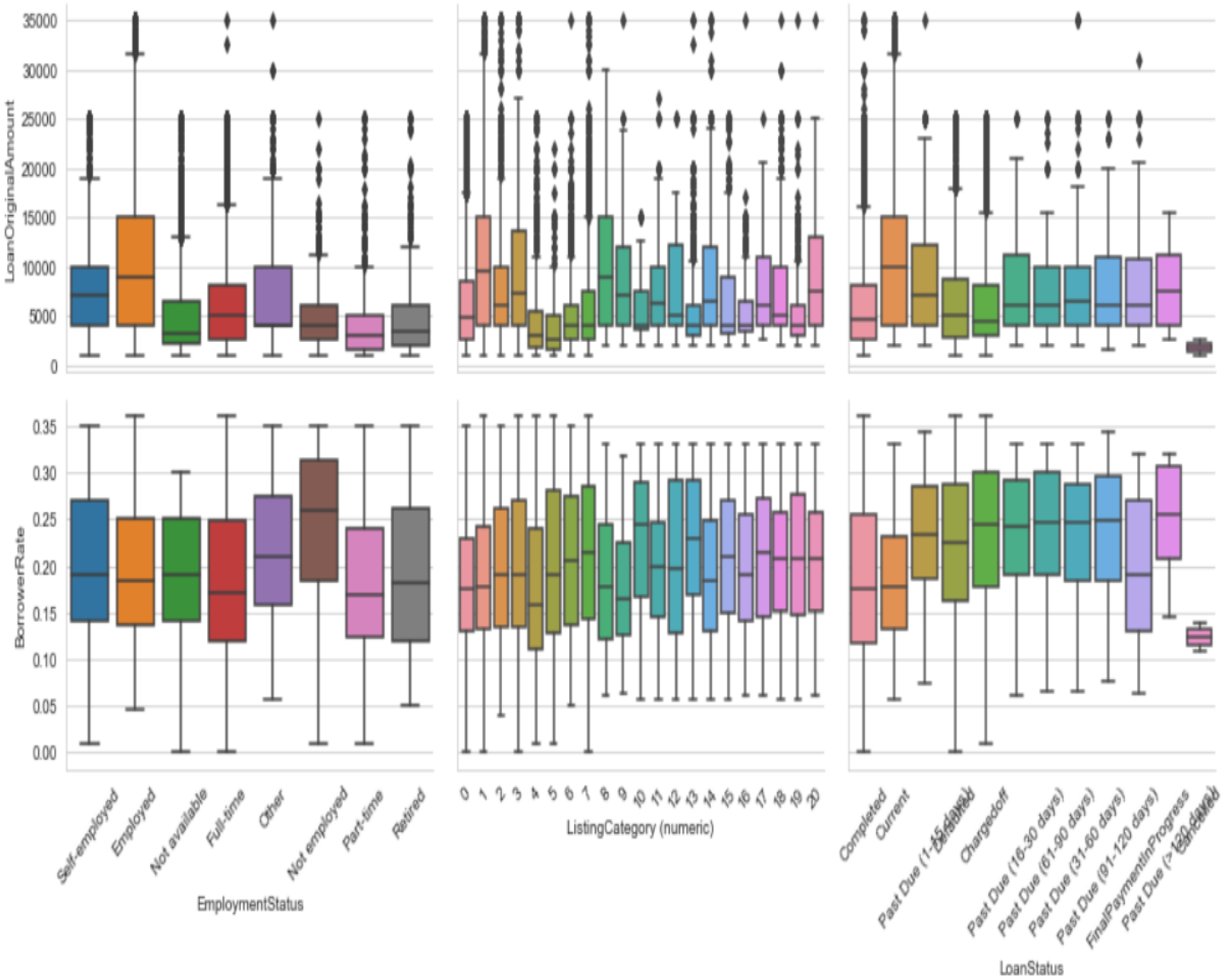

Term
BorrowerRate
Investors
Recommendations
LoanOriginalAmount
StatedMonthlyIncome
DelinquenciesLast7Years
LenderYield
ListingCategory (numeric)
1e6
pair plot

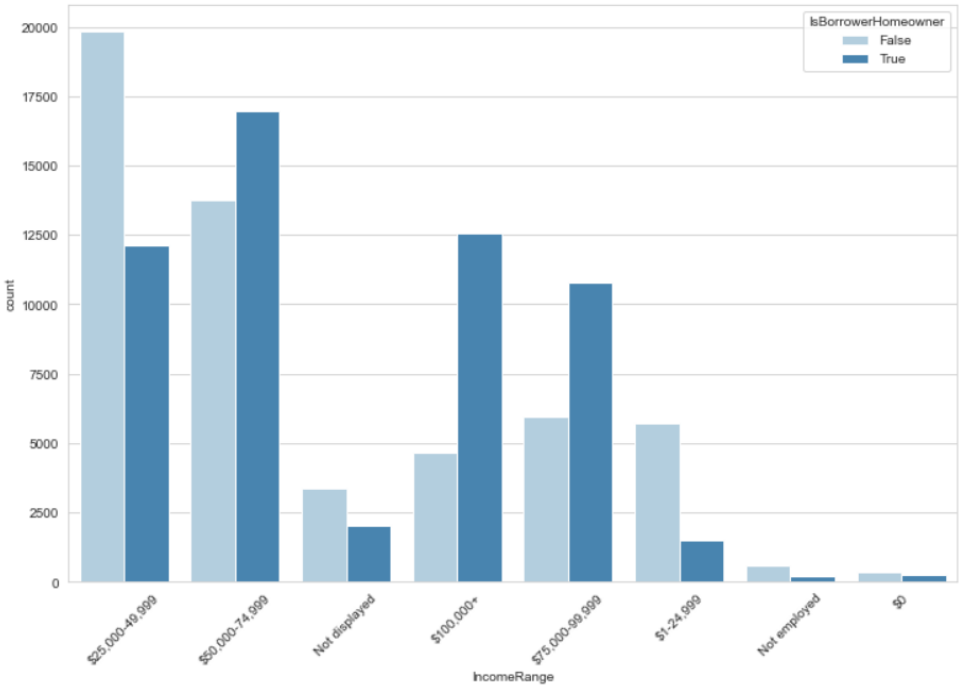

IncomeRange hue by IsBorrowerHomeowner
IsBorrowerHomeowner
False
True
count
20000
17500
15000
12500
10000
7500
5000
2500
0
$25,000-49,999
$50,000-74,999
Not displayed
$100,000+
$75,000-99,999
$1-24,999
Not employed
$0
IncomeRange


BorrowerRate and LenderYield corrlation
LenderYield
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
BorrowerRate

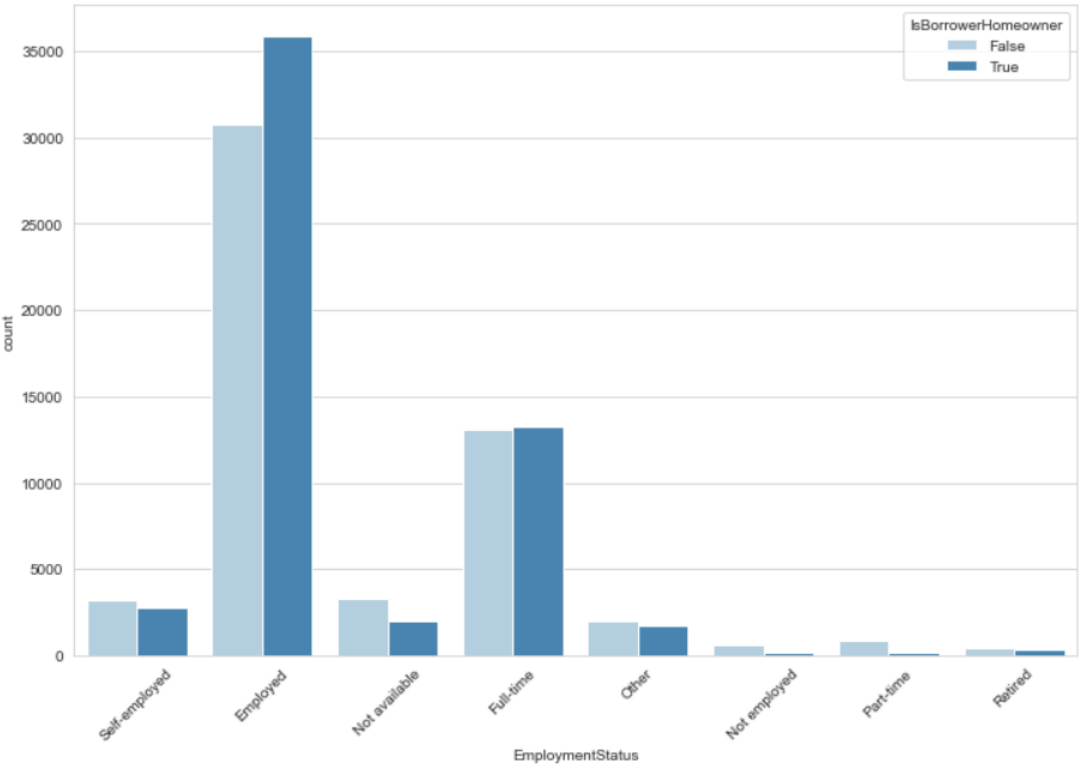

IncomeRange violinplot vs BorrowerRate
BorrowerRate
0.4
0.3
0.2
0.1
0.0
$25,000-49,999
$50,000-74,999
Not displayed
$100,000+
$75,000-99,999
$1-24,999
Not employed
$0
IncomeRange

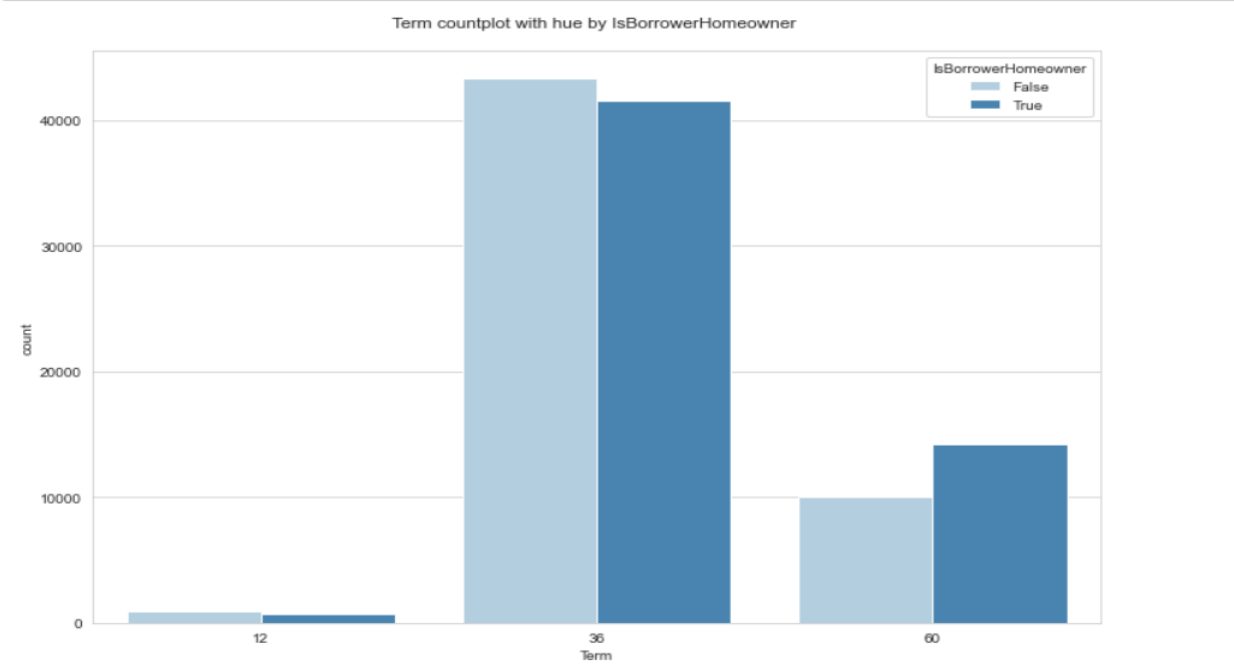

EmploymentStatus violinplot vs BorrowerRate
BorrowerRate
0.4
0.3
0.2
0.1
0.0
Self-employed
Employed
Not available
Full-time
Other
Not employed
Part-time
Retired
EmploymentStatus


LoanOriginalAmount
35000
30000
25000
20000
15000
10000
5000
0
BorrowerRate
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Self-employed
Employed
Not available
Full-time
Other
Not employed
Part-time
Retired
EmploymentStatus
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
ListingCategory (numeric)
Completed
Current
Chargedoff
Past Due (16-30 days)
Past Due (61-90 days)
Past Due (31-60 days)
Past Due (91-120 days)
FinalPaymentInProgress
LoanStatus



EmploymentStatus countplot with hue by IsBorrowerHomeowner





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?

- From the heatmap I found that:

    - Interestingly, there is no strong correlation between variables in this dataset except between BorrowerRate and lenderYield.

    - There is some positive correlation between LoanOriginalAmount and Term, LoanOriginalAmount and StatedMonthlyIncome, LoanOriginalAmount and number of Investors, number of delinquencies and borrowers rate, Recommendations and TotalProsperLoans.

    - There is also negative correlation between lenderYield and number of investors, LoanOriginalAmount and lenderYield,

LoanOriginalAmount and borrower rate and number of investors and borrower rate.

- IncomeRange with hue by IsBorrowerHomeowner, three section have the true hue over false one and income between 25,000 and 49,999 was the max including number of who has loans.

- Employment status of individuals with lower ratings tends to be 'Not employed', 'Self-employed', 'Retired' or 'Part-time'.

- Borrower rate for individuals with low rating is higher. High monthly income corresponds to higher rating.

- and answer quation like:

  - What is the EmploymentStatus for them?

  - Is hue by BorrowerHomeowner show diffrence in a spacific status?!

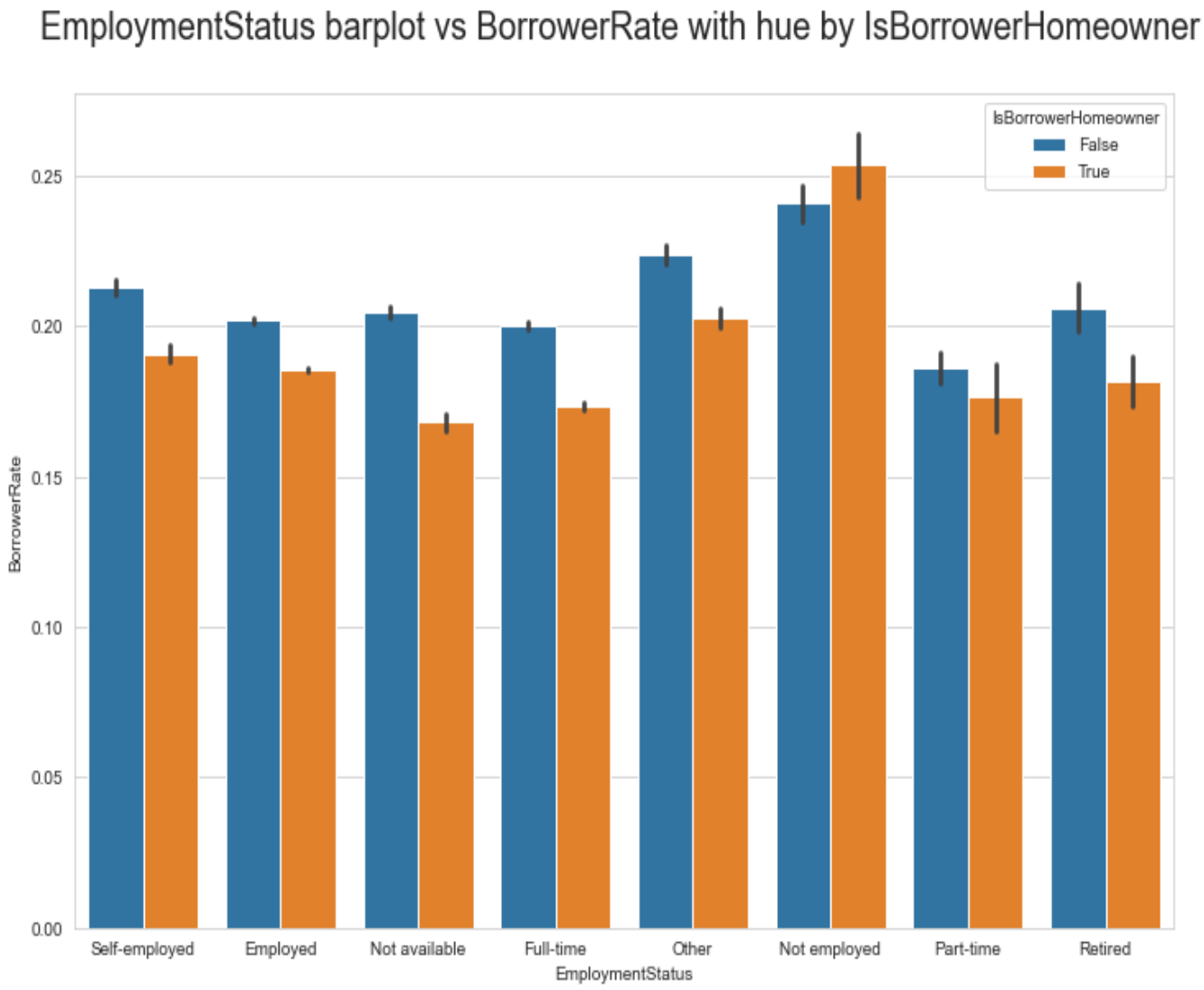
  - Is IsBorrowerHomeowner change over Term values?

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

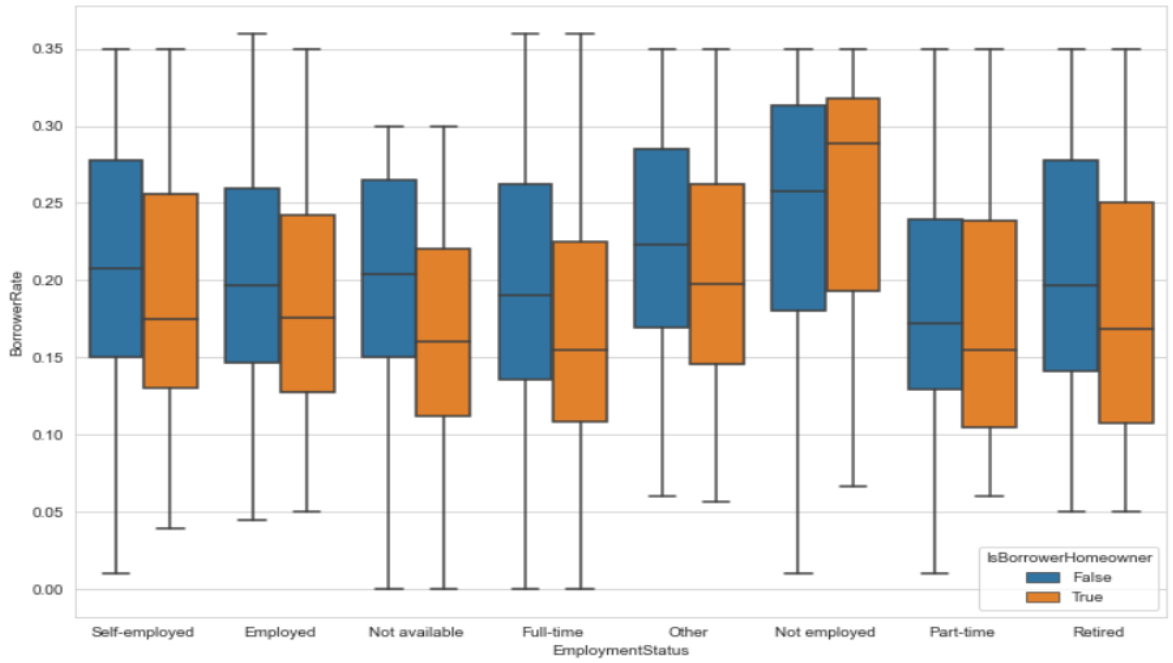
- Strong correlation between BorrowerRate and lenderYield, and interesting thing I've observed is that mid-Term (36) is the most frequent Term.

3- Multivariate Exploration

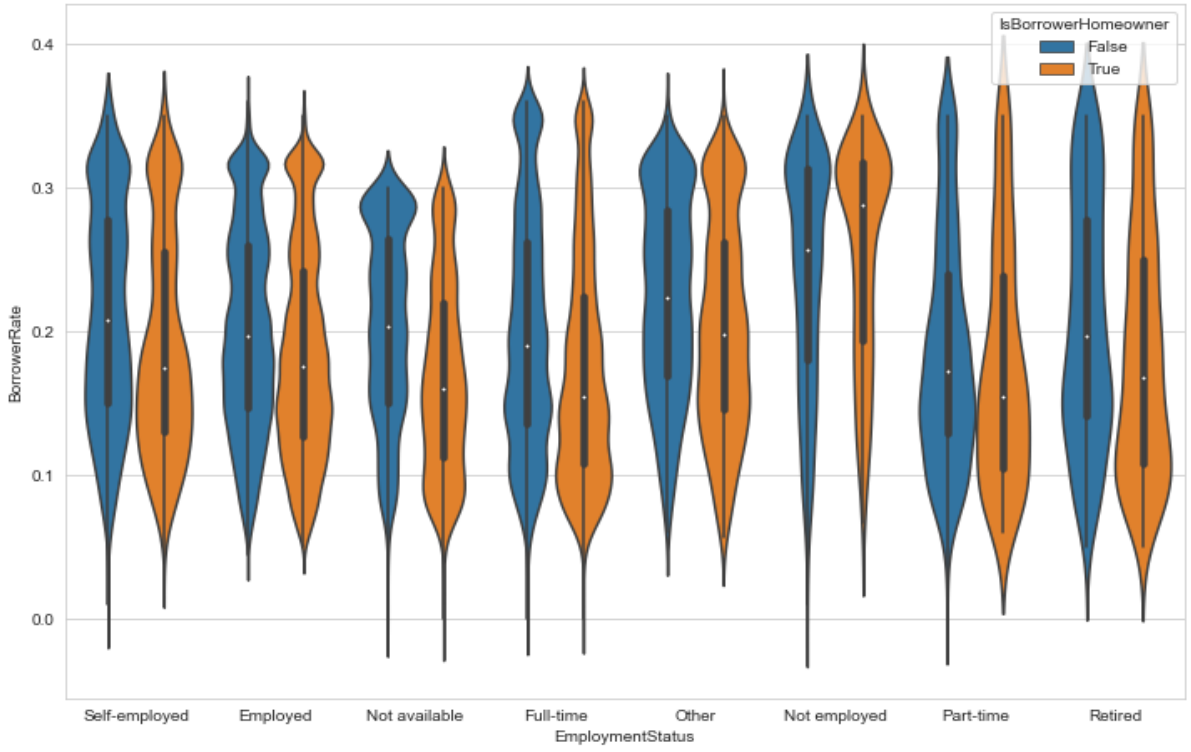
- Create plots of three or more variables to investigate your data even further. Make sure that your investigations are justified, and follow from your work in the previous sections.

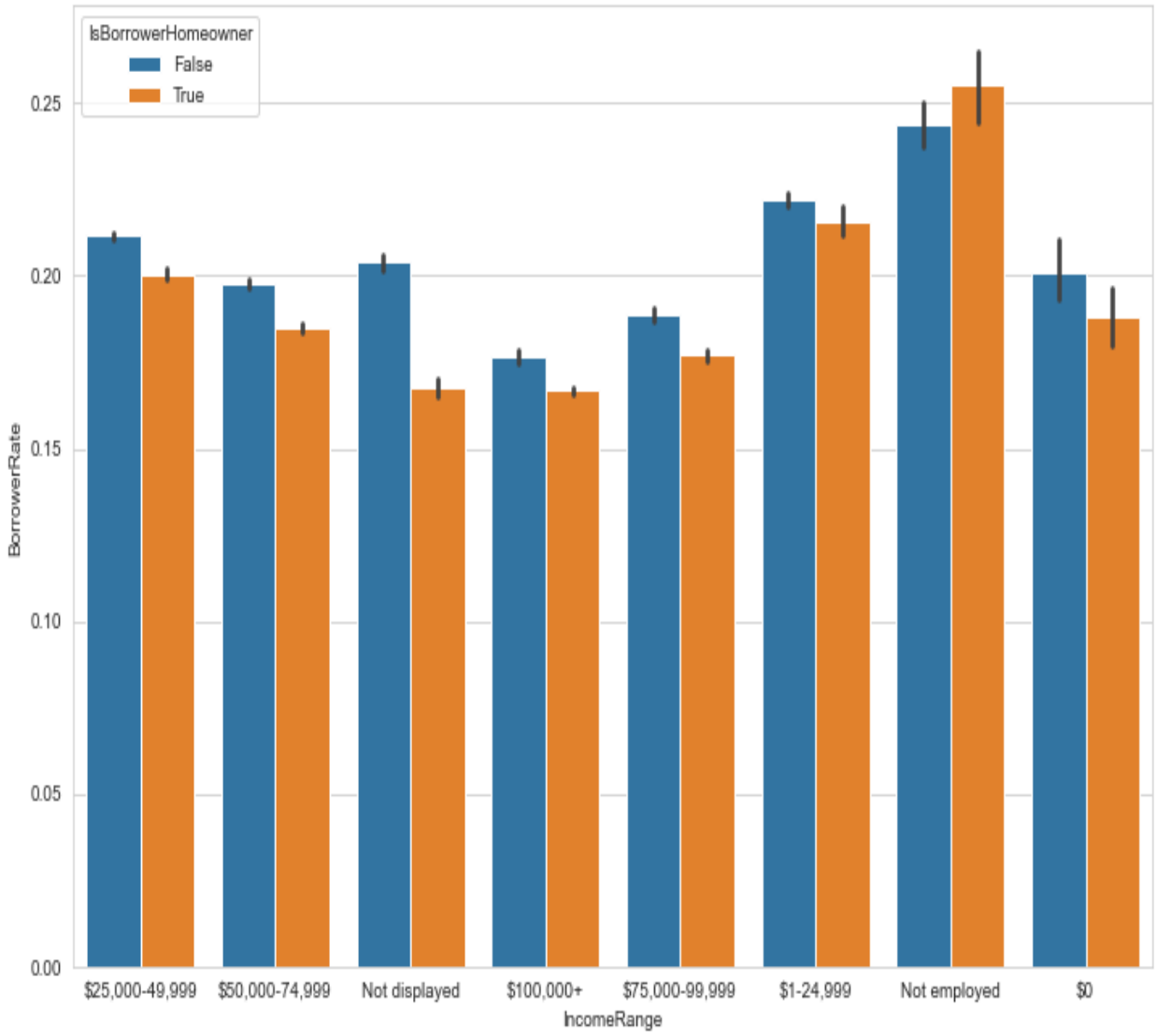


EmploymentStatus boxplot vs BorrowerRate with hue by IsBorrowerHomeowner

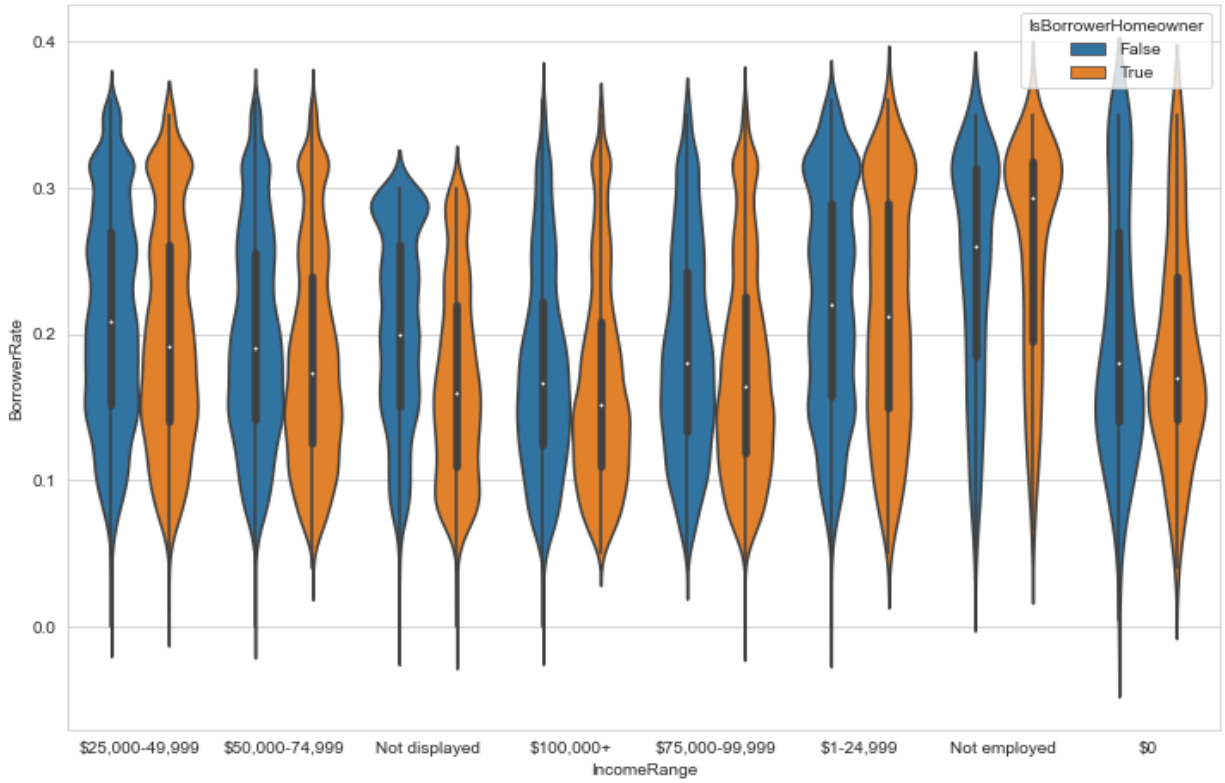


EmploymentStatus violinplot vs BorrowerRate with hue by IsBorrowerHomeowner

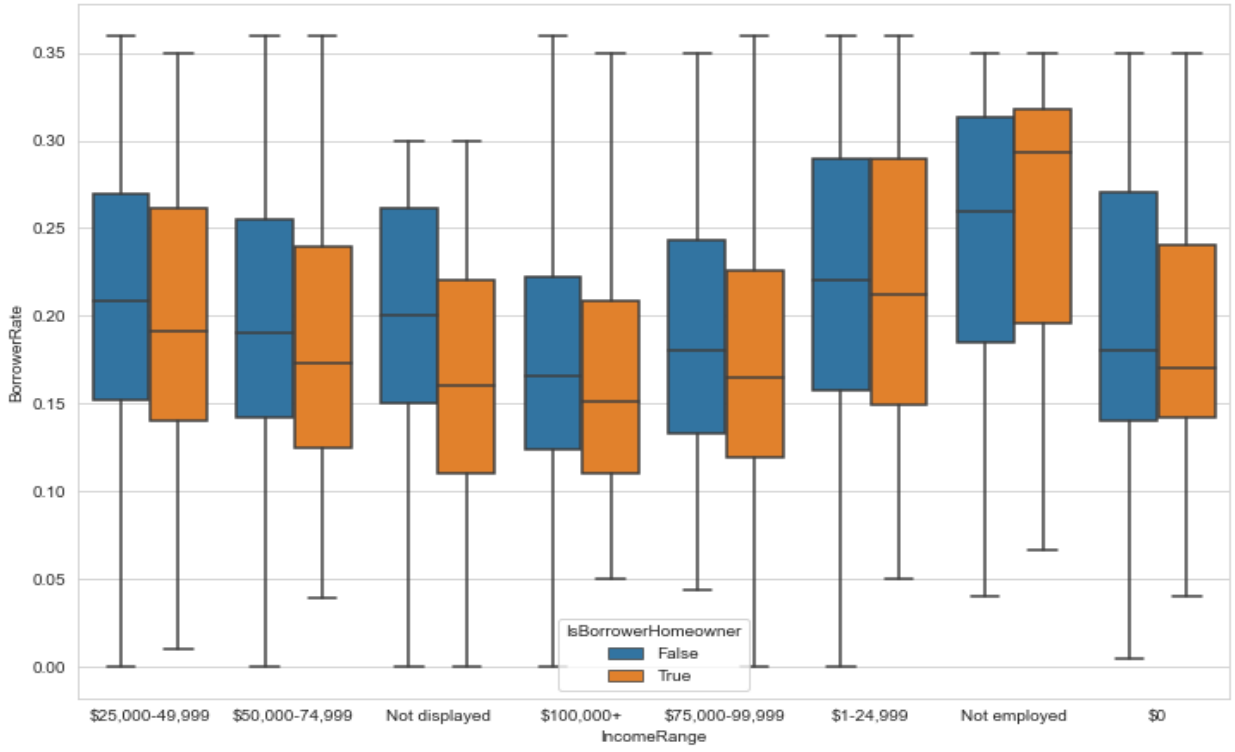


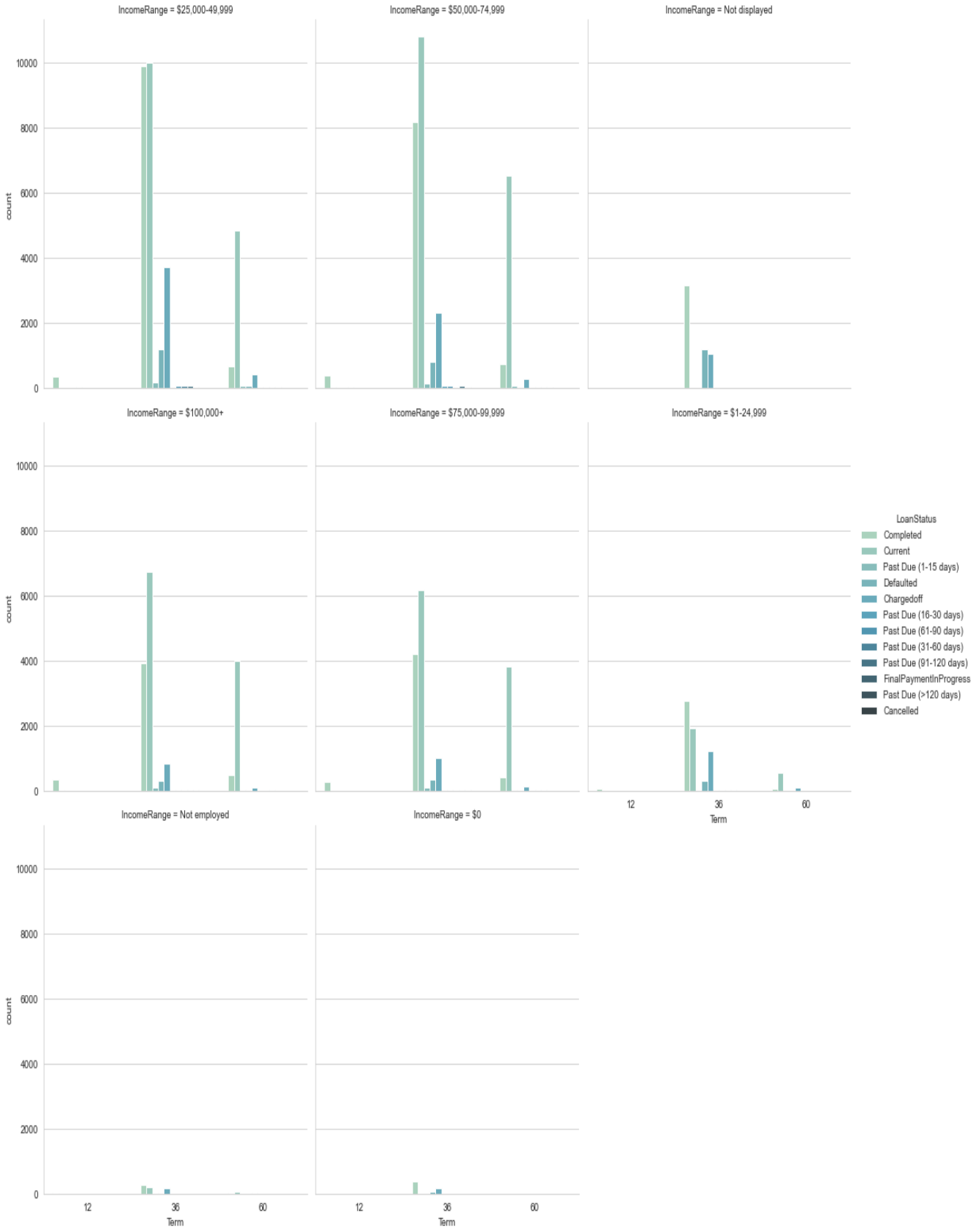

IncomeRange barplot vs BorrowerRate with hue by IsBorrowerHomeowner
IsBorrowerHomeowner
False
True
BorrowerRate
0.25
0.20
0.15
0.10
0.05
0.00
$25,000-49,999
$50,000-74,999
Not displayed
$100,000+
$75,000-99,999
$1-24,999
Not employed
$0
IncomeRange

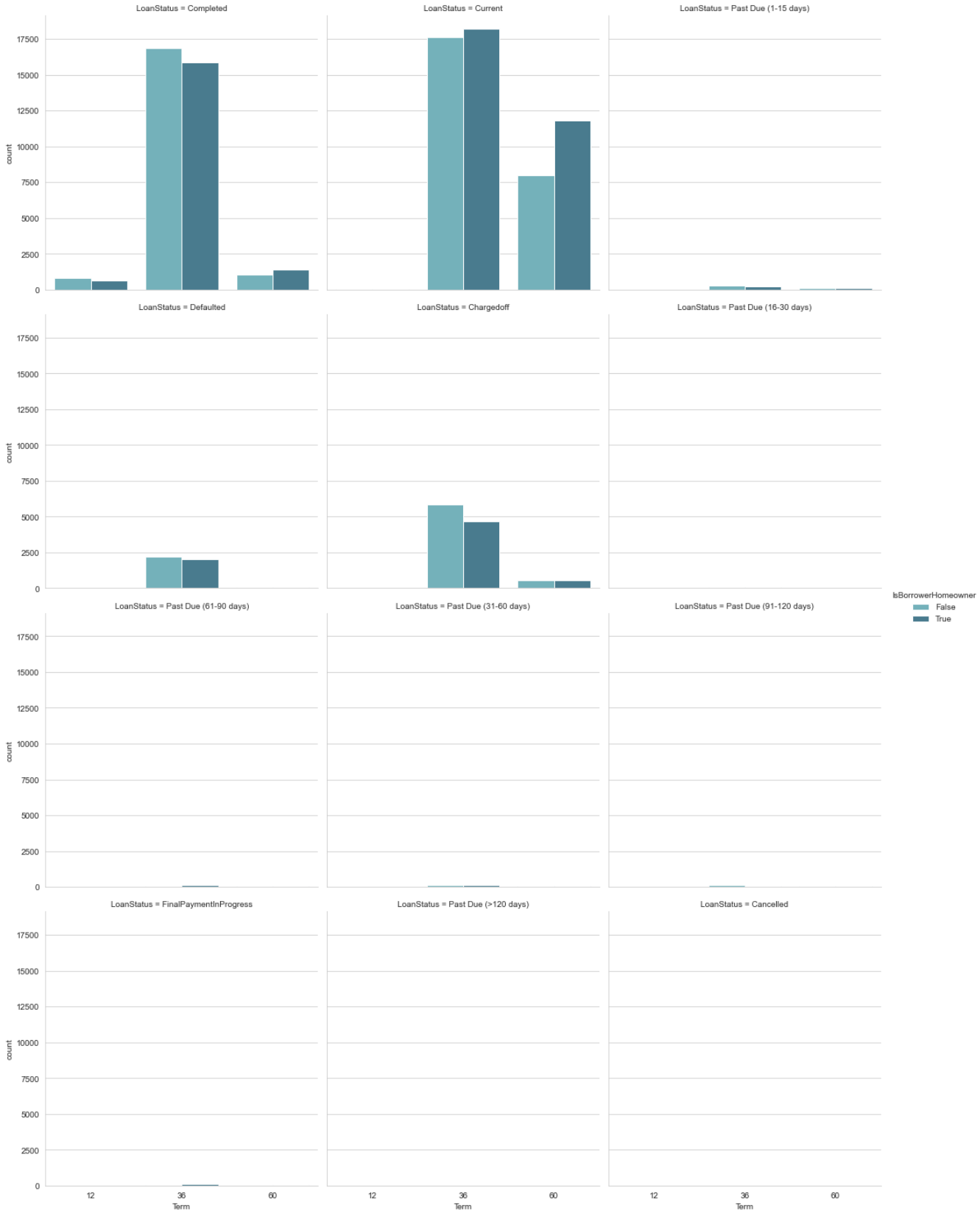
IncomeRange violinplot vs BorrowerRate with hue by IsBorrowerHomeowner

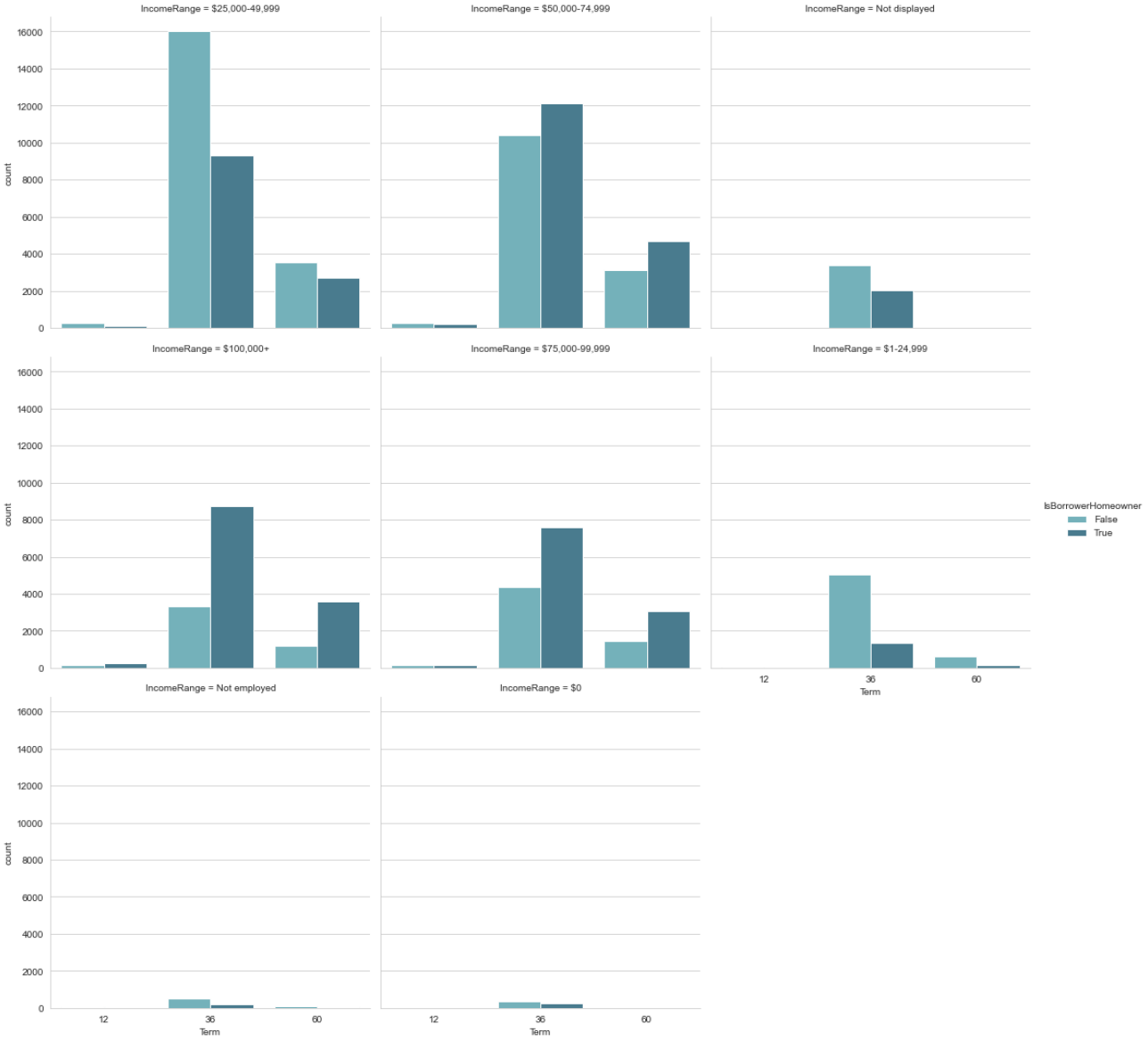


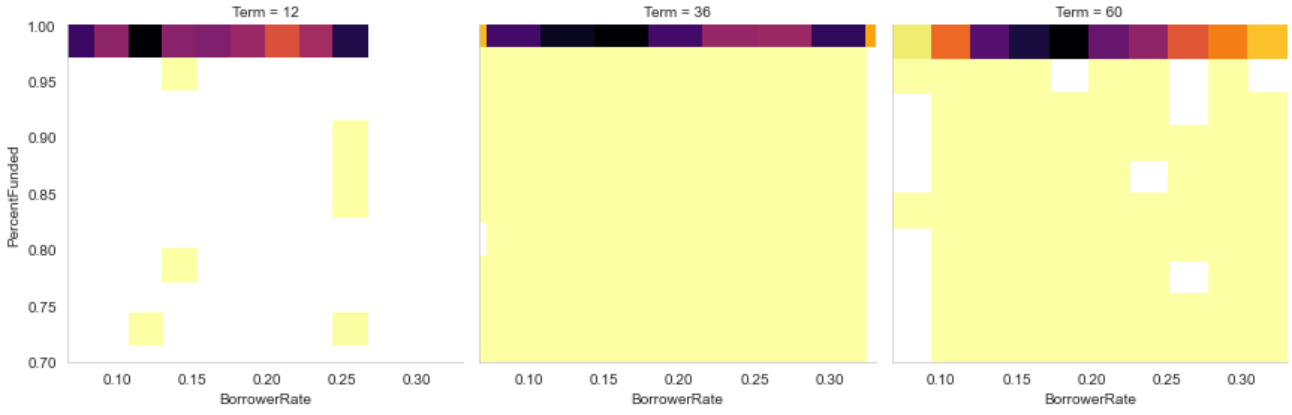
IncomeRange boxplot vs BorrowerRate with hue by IsBorrowerHomeowner




IncomeRange = $25,000-49,999
IncomeRange = $50,000-74,999
IncomeRange = Not displayed
IncomeRange = $100,000+
IncomeRange = $75,000-99,999
IncomeRange = $1-24,999
IncomeRange = Not employed
IncomeRange = $0
count
Term
12
36
60
0
2000
4000
6000
8000
10000
LoanStatus
Completed
Current
Past Due (1-15 days)
Defaulted
Chargedoff
Past Due (16-30 days)
Past Due (61-90 days)
Past Due (31-60 days)
Past Due (91-120 days)
FinalPaymentInProgress
Past Due (>120 days)
Cancelled


LoanStatus = Completed
LoanStatus = Current
LoanStatus = Past Due (1-15 days)
LoanStatus = Defaulted
LoanStatus = Chargedoff
LoanStatus = Past Due (16-30 days)
LoanStatus = Past Due (61-90 days)
LoanStatus = Past Due (31-60 days)
LoanStatus = Past Due (91-120 days)
LoanStatus = FinalPaymentInProgress
LoanStatus = Past Due (>120 days)
LoanStatus = Cancelled
count
0
2500
5000
7500
10000
12500
15000
17500
12
36
60
Term
IsBorrowerHomeowner
False
True


IncomeRange = $25,000-49,999
IncomeRange = $50,000-74,999
IncomeRange = Not displayed
IncomeRange = $100,000+
IncomeRange = $75,000-99,999
IncomeRange = $1-24,999
IncomeRange = Not employed
IncomeRange = $0
count
Term
IsBorrowerHomeowner
False
True


Term = 12
Term = 36
Term = 60
PercentFunded
BorrowerRate

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?

- Our initial assumptions were strengthened. The outcome of credit depends on IncomeRange, Term, Employment status. Defaulted credits tend to be larger than completed for all ratings except the lowest ones. Not Employee has highest BorrowerRate. midterm (36 month) is the highest BorrowerRate. Income range over than $75,000 tend to have mid_term(36).

Were there any interesting or surprising interactions between features?

- Looking back on the point plots, it doesn't seem like there's a systematic interaction effect between the features. However, the features also aren't fully independent and have weak corrloration. But it is interesting in something like the BorrowerRate plot for each one against IncomeRange, Term and Employmentstatus.

The End

That’s all you can see my work and enjoy it, and if there anything I can do I will be more than happy to connect.

Connect with me:

- Gmail: mohamednajm250@gmail.com

- Linked In: https://www.linkedin.com/in/mohamed-najm-aa6a00158/

- Github: https://github.com/Mohamwd-Najm