(Prosper Loan Data)

by (John Beniamin Kostandy Shenoda)

Preliminary Wrangling

> *A dataset that contains information about loans, its composed of information about the loan amoint, the borrower categeory, the loan period and so on.*

	LoanOriginalAmount	BorrowerAPR	StatedMonthlyIncome	Term	ProsperRating (Alpha)	EmploymentStatus	MonthlyLoanPayment	Investors	Occupation	CreditGrade	EmploymentStatusDuration
0	9425	0.16516	3083.333333	36	NaN	Self-employed	330.43	258	Other	C	2.0
1	10000	0.12016	6125.000000	36	A	Employed	318.93	1	Professional	NaN	44.0
2	3001	0.28269	2083.333333	36	NaN	Not available	123.32	41	Other	HR	NaN
3	10000	0.12528	2875.000000	36	A	Employed	321.45	158	Skilled Labor	NaN	113.0
4	15000	0.24614	9583.333333	36	D	Employed	563.97	20	Executive	NaN	44.0

What is the structure of your dataset?

> *dataset is composed of 81 columns with 11397 notices in the rows(not all columns with the same number of notices, but i ll investigate only 10 features (columns) witch are the most important*

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

> *the main features are the amount of money borrowed, the period and categeorising the borrower*

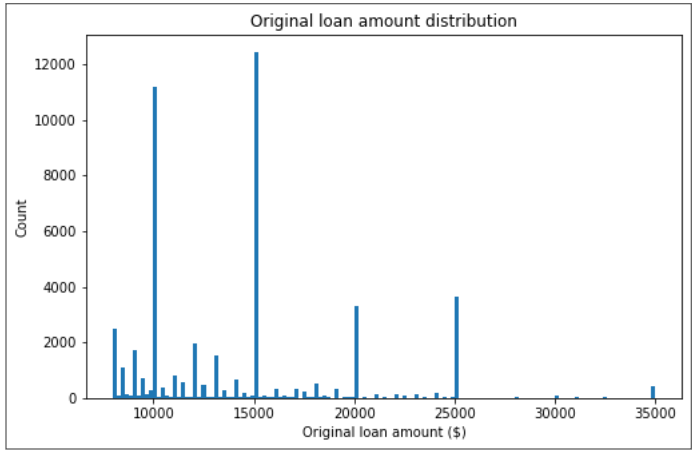
What features in the dataset do you think will help support your investigation into your feature(s) of interest?

> *stated monthly income, loan original amount,*

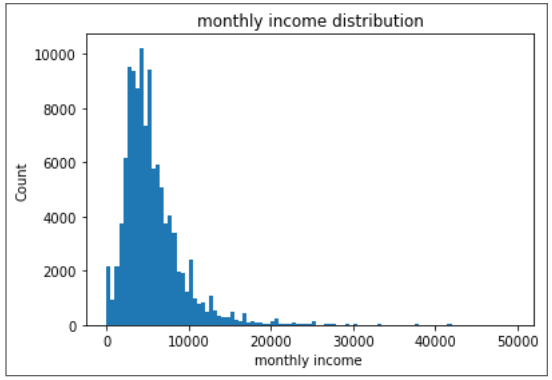
Univariate Exploration

> *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.*

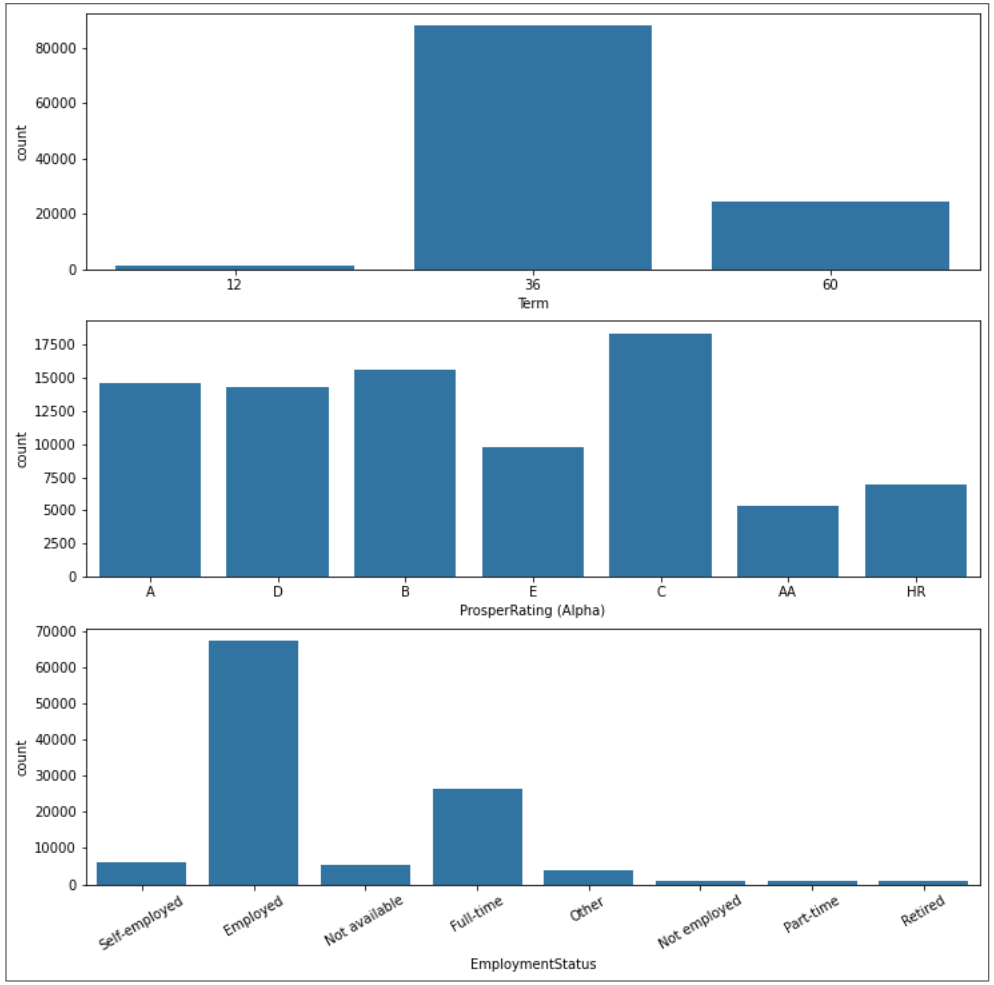
we will start with investigating some variables like loan original amount, monthly income, employment status


Original loan amount distribution
Count
Original loan amount ($)
0
2000
4000
6000
8000
10000
12000
10000
15000
20000
25000
30000
35000

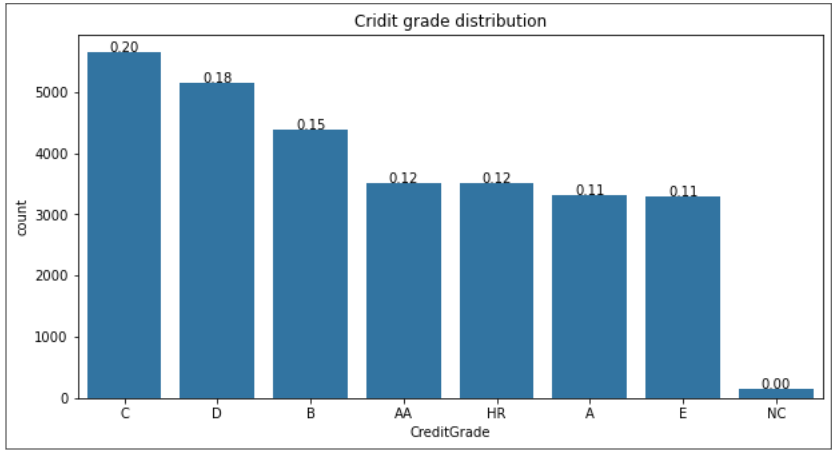
we can notice that the largest spikes are at 10, 15, 20, 25 ks, while the smaller spikes noticed at 7, 9, 13 ks

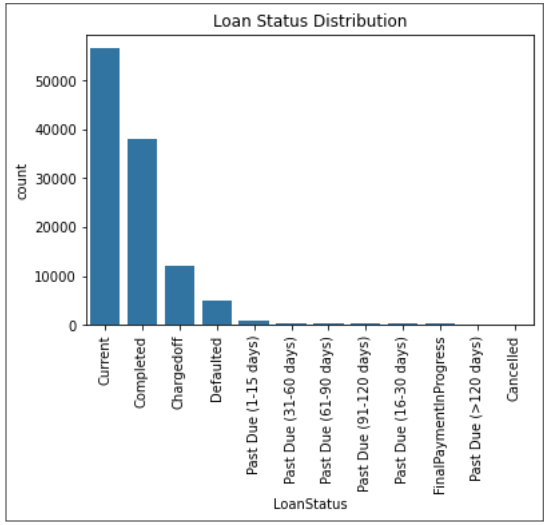

monthly income distribution
Count
monthly income
0
2000
4000
6000
8000
10000
0
10000
20000
30000
40000
50000

we can notice that monthly income distribution is right skewed

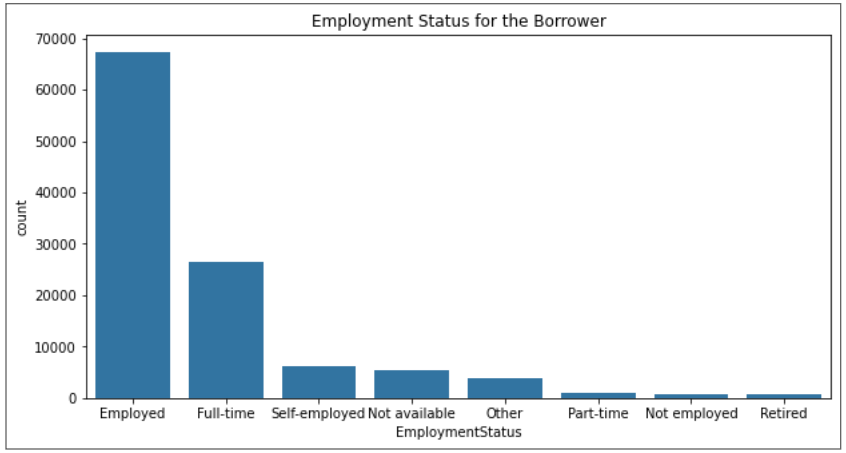

count
80000
60000
40000
20000
0
12
36
60
Term
count
17500
15000
12500
10000
7500
5000
2500
0
A
D
B
E
C
AA
HR
ProsperRating (Alpha)
count
70000
60000
50000
40000
30000
20000
10000
0
Self-employed
Employed
Not available
Full-time
Other
Not employed
Part-time
Retired
EmploymentStatus

we can notice that most of the loans are for a 36 months period, taken by fulltime jobs holder, with a ratind (a,b,c,d)

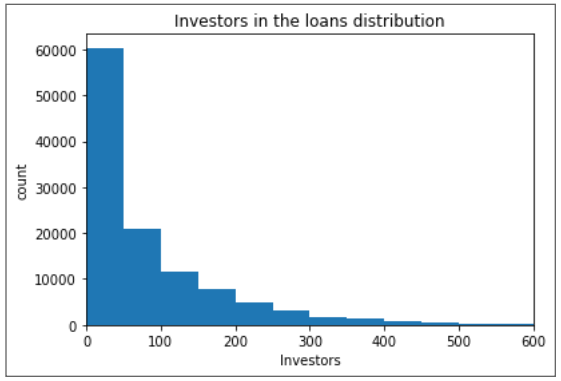

Cridit grade distribution
count
CreditGrade
0.20
0.18
0.15
0.12
0.12
0.11
0.11
0.00
C
D
B
AA
HR
A
E
NC
0
1000
2000
3000
4000
5000


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

we can notice that the most amount of lians are current, which cal=n lead us to further invedstigations


Employment Status for the Borrower
70000
60000
50000
40000
30000
20000
10000
0
count
Employed
Full-time
Self-employed
Not available
Other
Part-time
Not employed
Retired
EmploymentStatus

```
Text(0, 0.5, 'count')
```



Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

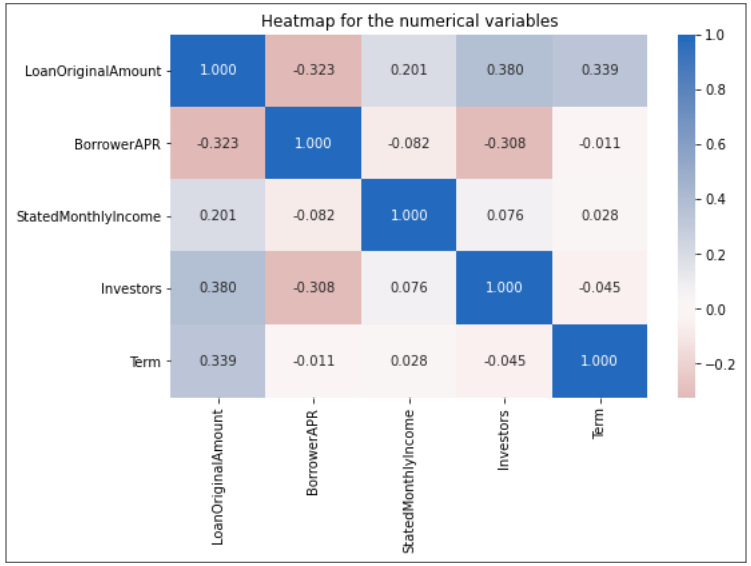
> *there is no unusual, its logic that loan increases as the investors increased and as the customer is employed!*

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

> *no there is no need for more operations !*

Bivariate Exploration

> *In this section, investigate relationships between pairs of variables in your data. Make sure the variables that you cover here have been introduced in some fashion in the previous section (univariate exploration).*


Heatmap for the numerical variables
LoanOriginalAmount
BorrowerAPR
StatedMonthlyIncome
Investors
Term
1.000
-0.323
0.201
0.380
0.339
-0.323
1.000
-0.082
-0.308
-0.011
0.201
-0.082
1.000
0.076
0.028
0.380
-0.308
0.076
1.000
-0.045
0.339
-0.011
0.028
-0.045
1.000
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

we can notice some strong corelations between investors & loan amount, term& loan amount. this may need further invistigation

```
Text(0.5, 1.0, 'matrix relations between numerical variables')
```

we can record some notices here, a -ve relationship between investors and borrower apr & monthly income

the peak of the long term loans is in the rating c in case of emplyed person

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?

> *the employed people as they have the most loans, the less term of loan they have!*

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

> *I am astonished about the result that self employed people have less loan as the employed!*

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.

we can notice more the -ve relation that we stated above

the -ve relationship is more clearly appearing

the loan goes in a logterm as the income aand loan amount increase

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?

> *we find that the relation ship between loan amount and oan term increases as the APR increases, which mean as the trust in the customer increase you can lend him more money for more time to pay!*

Were there any interesting or surprising interactions between features?

> *data analytics is always interesting as you can read facts through numbers, !*

> *At the end of your report, make sure that you export the notebook as an html file from the `File > Download as... > HTML` menu. Make sure you keep track of where the exported file goes, so you can put it in the same folder as this notebook for project submission. Also, make sure you remove all of the quote-formatted guide notes like this one before you finish your report!*