Part I - (Exploring the 'Loan Data from Prosper' dataset)

by (Muhammad Tareq Mahmud)

Introduction

> 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 dataset will be used to answer certain questions like:
>
> 1. What factors affect a loan's outcome status?
> 2. What affects the borrower's APR or interest rate?
> 3. Are there differences between loans depending on how large the original loan amount was?

Preliminary Wrangling

In [1]:

```
# import all packages and set plots to be embedded inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb

%matplotlib inline
```

> Load in your dataset and describe its properties through the questions below. Try and motivate your exploration goals through this section.

In [2]:

```
df= pd.read_csv('loan.csv')
print(df.shape)
print(df.isnull().sum().sum())
df.dropna()
```

```
(113937, 81)
1364086
```

Out[2]:

ListingKey	ListingNumber	ListingCreationDate	CreditGrade	Term	LoanStatus	ClosedDate	BorrowerAPR	BorrowerRate	LenderYield	...	LP_ServiceFees	LP_CollectionFees	LP_GrossPrincipalLoss	LP_NetPrincipalLoss	LP_NonPrincipalRecoverypayments	PercentFunded	Recommendations	InvestmentFromF

0 rows × 81 columns

In [3]:

```
df.head()
```

Out[3]:

	ListingKey	ListingNumber	ListingCreationDate	CreditGrade	Term	LoanStatus	ClosedDate	BorrowerAPR	BorrowerRate	LenderYield	...	LP_ServiceFees	LP_CollectionFees	LP_GrossPrincipalLoss	LP_NetPrincipalLoss	LP_NonPrincipalRecoverypayments	PercentFunded	Recommendations
0	1021339766868145413AB3B	193129	2007-08-26 19:09:29.263000000	C	36	Completed	2009-08-14 00:00:00	0.16516	0.1580	0.1380	...	-133.18	0.0	0.0	0.0	0.0	1.0	0
1	10273602499503308B223C1	1209647	2014-02-27 08:28:07.900000000	NaN	36	Current	NaN	0.12016	0.0920	0.0820	...	0.00	0.0	0.0	0.0	0.0	1.0	0
2	0EE9337825851032864889A	81716	2007-01-05 15:00:47.090000000	HR	36	Completed	2009-12-17 00:00:00	0.28269	0.2750	0.2400	...	-24.20	0.0	0.0	0.0	0.0	1.0	0
3	0EF5356002482715299901A	658116	2012-10-22 11:02:35.010000000	NaN	36	Current	NaN	0.12528	0.0974	0.0874	...	-108.01	0.0	0.0	0.0	0.0	1.0	0
4	0F023589499656230C5E3E2	909464	2013-09-14 18:38:39.097000000	NaN	36	Current	NaN	0.24614	0.2085	0.1985	...	-60.27	0.0	0.0	0.0	0.0	1.0	0

5 rows × 81 columns

What is the structure of your dataset?

> The dataset contains 113937 rows and 81 variables/columns in total. This dataset contains 1364086 null values in total.

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

> I have selected 17 columns out of the 81 from the original dataset which contain 113937 rows. This is a big dadaset and in certain columns there are some NaN values which were dropped. After dropping the Nan values from those columns, we had 83520 rows and 17 columns.

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

> I am going with with the full dataset but there are certain columns which will be used frequently: *'Term','LoanStatus','BorrowerAPR','BorrowerRate','LenderYield','ProsperScore','ListingCategory (numeric)','Occupation','EmploymentStatus','IsBorrowerHomeowner','CreditScoreRangeLower', 'CreditScoreRangeUpper','InquiriesLast6Months','TotalInquiries','IncomeRange','LoanOriginalAmount','MonthlyLoanPayment'*

Univariate Exploration

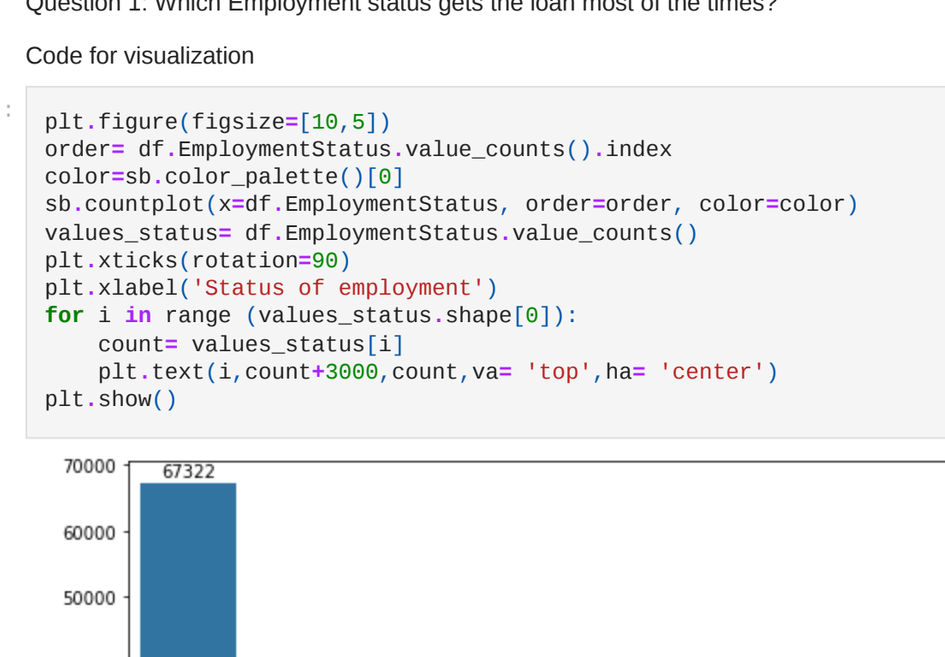
> In this section, investigate distributions of individual variables. If you see unusual points or outliers, take a deeper look to clean things up and prepare yourself to look at relationships between variables.

Question 1: Which Employment status gets the loan most of the times?

Code for visualization

In [4]:

```
plt.figure(figsize=[10,5])
order= df.EmploymentStatus.value_counts().index
color=sb.color_palette()[0]
sb.countplot(x=df.EmploymentStatus, order=order, color=color)
values_status= df.EmploymentStatus.value_counts()
plt.xticks(rotation=90)
plt.xlabel('Status of employment')
for i in range (values_status.shape[0]):
    count= values_status[i]
    plt.text(i,count+3000,count,va= 'top',ha= 'center')
plt.show()
```



Observation 1:

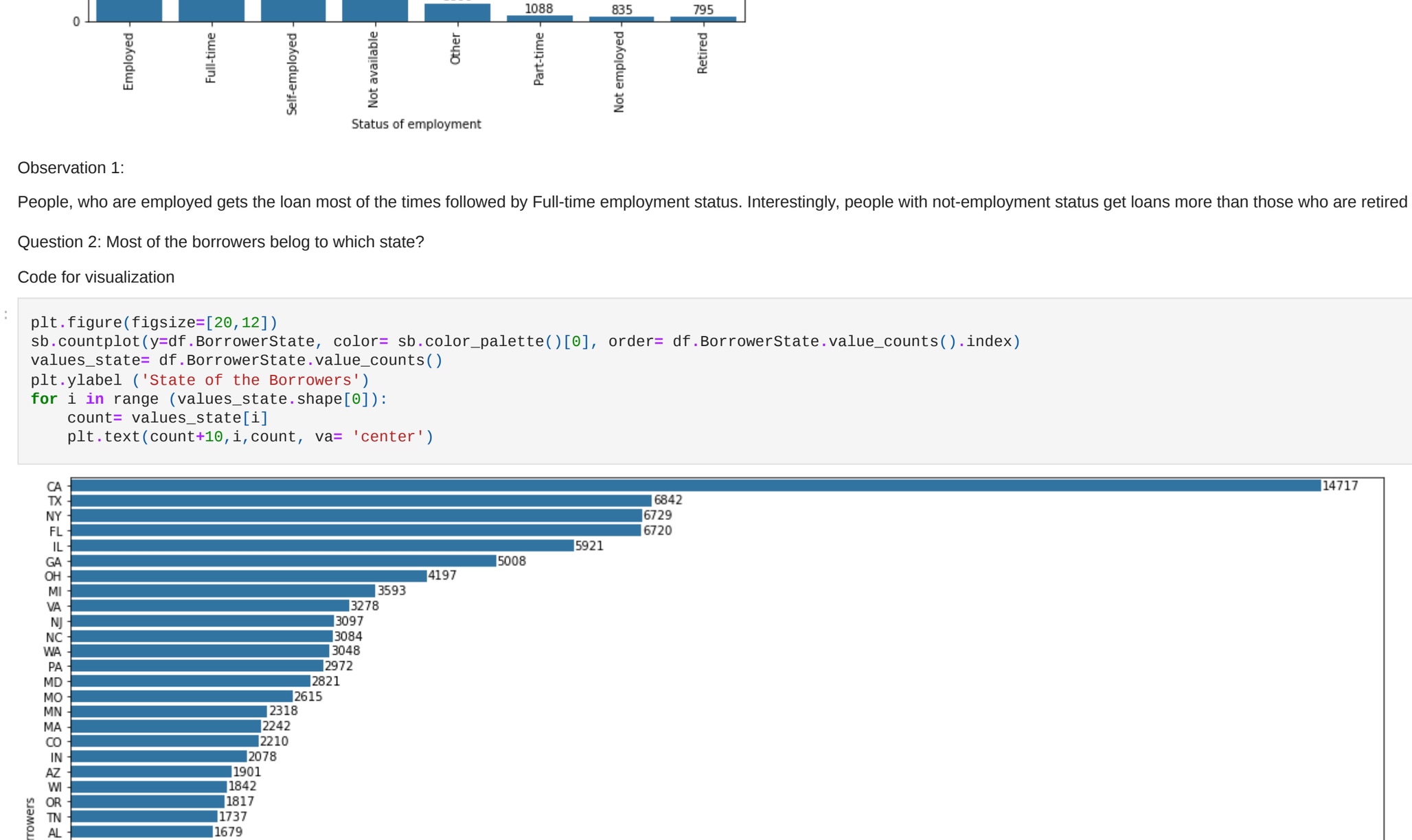
People, who are employed gets the loan most of the times followed by Full-time employment status. Interestingly, people with not-employment status get loans more than those who are retired or doing part-time jobs.

Question 2: Most of the borrowers belog to which state?

Code for visualization

In [5]:

```
plt.figure(figsize=[20,12])
sb.countplot(y=df.BorrowerState, color= sb.color_palette()[0], order= df.BorrowerState.value_counts().index)
values_state= df.BorrowerState.value_counts()
plt.ylabel ('State of the Borrowers')
for i in range (values_state.shape[0]):
    count= values_state[i]
    plt.text(count+10,i,count, va= 'center')
```



Observation 2:

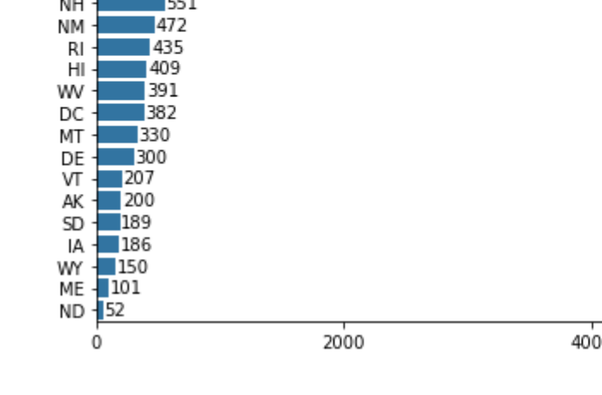
The above chart displays that most of the borrower who takes loan are from California, followed by New York and Texas

Question 3: If most of the borrowers are homeowners?

Code for visualization

In [6]:

```
home_owner_sorted= df.IsBorrowerHomeowner.value_counts()
home_owner_sorted
label= ['Own a home', 'Do not own a home']
plt.pie(home_owner_sorted, labels= label, startangle=90, counterclock=False, wedgeprops={'width':.4});
plt.axis('square');
```



Observation 3:

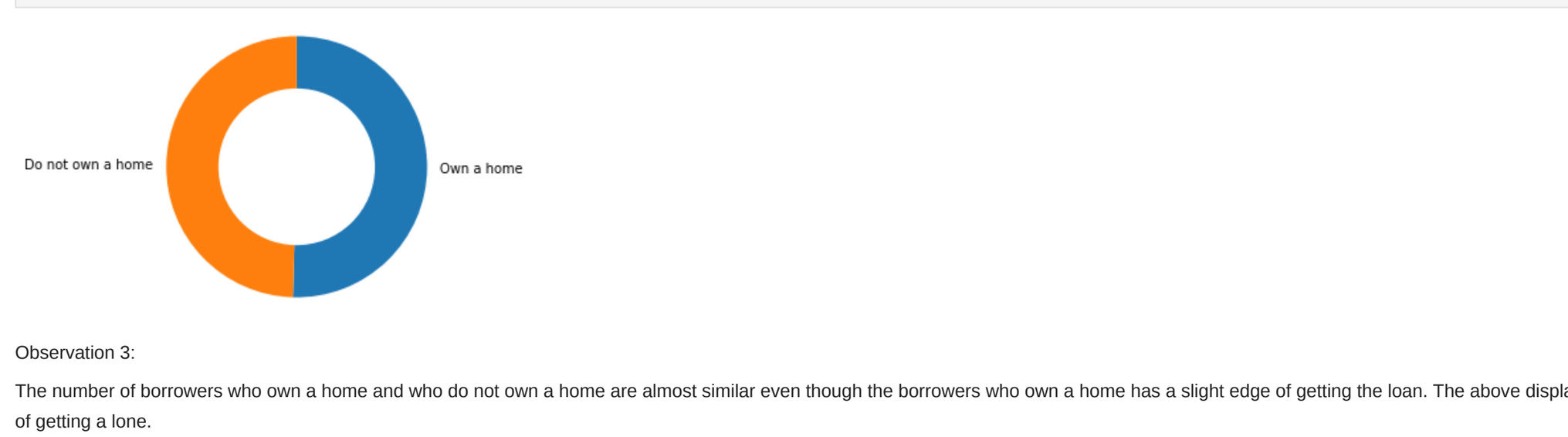
The number of borrowers who own a home and who do not own a home are almost similar even though the borrowers who own a home has a slight edge of getting the loan. The above displayed Donut chart shows that even though most of the borrower own a home but owning a home is not the only criteria of getting a lone.

Question 4: Interest rate of the loans?

Code for visualization

In [7]:

```
bins= np.arange(.04,.36+.4,.01)
plt.figure(figsize=[20,5])
plt.hist(data=df,x='BorrowerRate',bins=bins, alpha=.6);
plt.xlim(0.03,.36);
xticks=[.03,.05,.10,.15,.2,.25,.30,.31,.32,.33,.35]
plt.xticks(xticks,xticks);
plt.xlabel('The most common interest rate');
```



Observation 4:

The above Histogram displays that the most of the clients or borrows received the loan with an interest rate between .31 to .32.

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

> Used the value_counts() on the BorrowerState and EmploymentStatus to figure out most of the borrowers belong to which state and to which employment status. Later their values and index were used to produce barchart.

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.

Bivariate Exploration

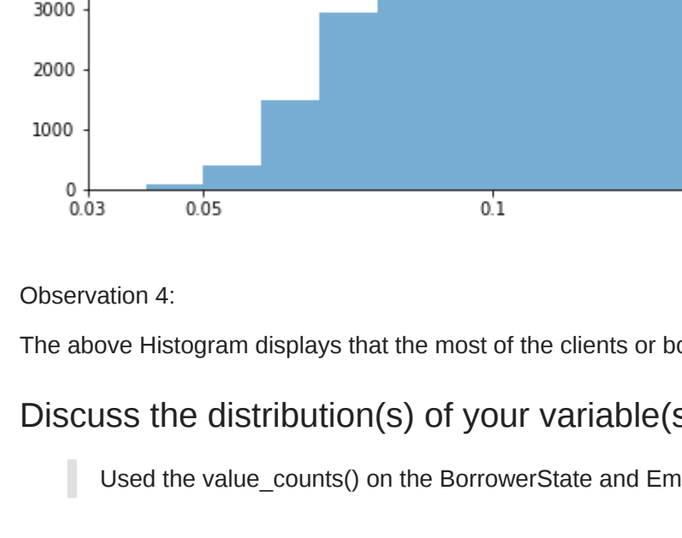
> In this section, investigate relationships between pairs of variables in your data. Make sure the variables that you cover here have been introduced in some fashion in the previous section (univariate exploration).

Question 5: Borrower with which Income Range get Highest amount of Loan (on an average/median)?

Code for visualization

In [8]:

```
sb.boxplot(data=df, y= 'IncomeRange',x= 'LoanOriginalAmount')
plt.xticks(rotation=85);
```



Observation 5:

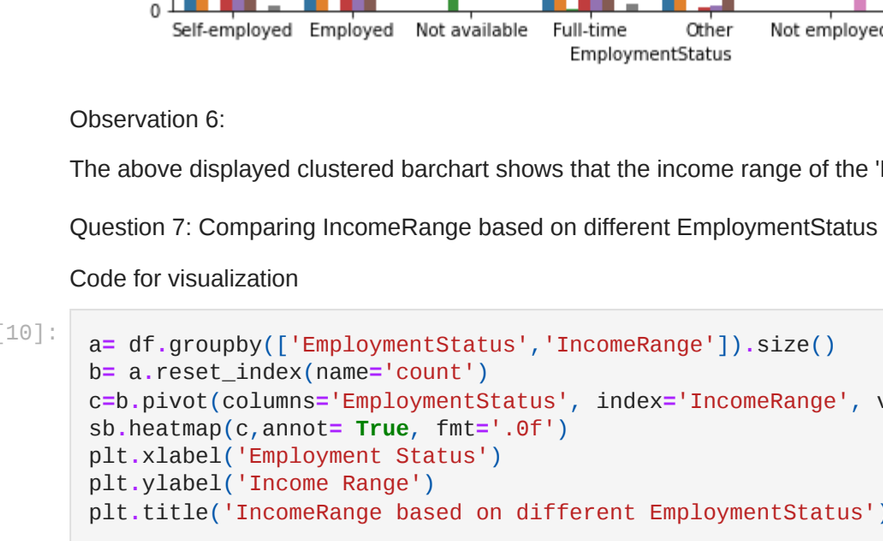
The above Boxplot is displaying the relation between the Income Range and Original Loan Amount. It projects that if the income range in more than or equal to $100,000 then the borrower gets the highest amount of loan (on an average). The overall boxplot is trying to depicts that how much a borrower will get as loan amount is dependent on his/her/their income range.

Question 6: What is the income range of different Employmnet Status?

Code for visualization

In [9]:

```
plt.figure(figsize=[10,5])
sb.countplot(data=df, x='EmploymentStatus', hue='IncomeRange');
plt.legend( title = "Income Range");
```



Observation 6:

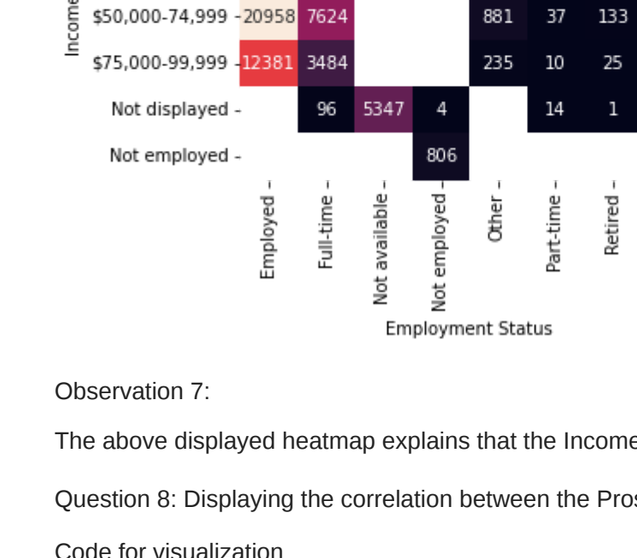
The above displayed clustered barchart shows that the income range of the 'Employed' Employmnet status is the hoghest among all the other employment status.

Question 7: Comparing IncomeRange based on different EmploymentStatus

Code for visualization

In [10]:

```
a= df.groupby(['EmploymentStatus','IncomeRange']).size()
b= a.reset_index(name='count')
c=b.pivot(columns='EmploymentStatus', index='IncomeRange', values='count')
sb.heatmap(c,annot= True, fmt='.0f')
plt.xlabel('Employment Status')
plt.ylabel('Income Range')
plt.title('IncomeRange based on different EmploymentStatus');
```



Observation 7:

The above displayed heatmap explains that the Income Range of those who are Employed is higher than any other Employment Status. It depicts that thise who are Employed earn more than any other group.

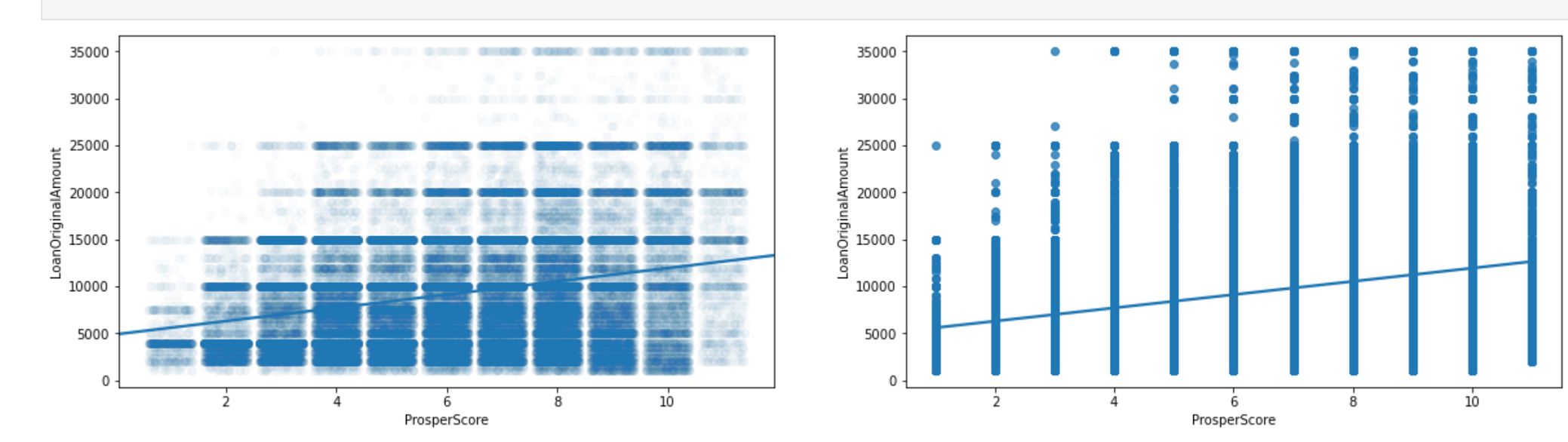
Question 8: Displaying the correlation between the ProsperScore and LoanOriginalAmount

Code for visualization

In [11]:

```
plt.figure(figsize=[20,5])
plt.subplot(1,2,1)
sb.regplot(x= df.ProsperScore, y=df.LoanOriginalAmount, truncate= False,x_jitter= .4, scatter_kws={'alpha':.02})

plt.subplot(1,2,2)
sb.regplot(x= df.ProsperScore, y=df.LoanOriginalAmount);
```



Observation 8:

The above displayed scatterplot shows that a positive correlation exists between the ProsperScore and LoanOriginalAmount. With the increase of ProsperScore (better prosper score), the chances of LoanOriginalAmount increases.

Question 9: Is there a correlation between LoanOriginalAmount and BorrowerRate?

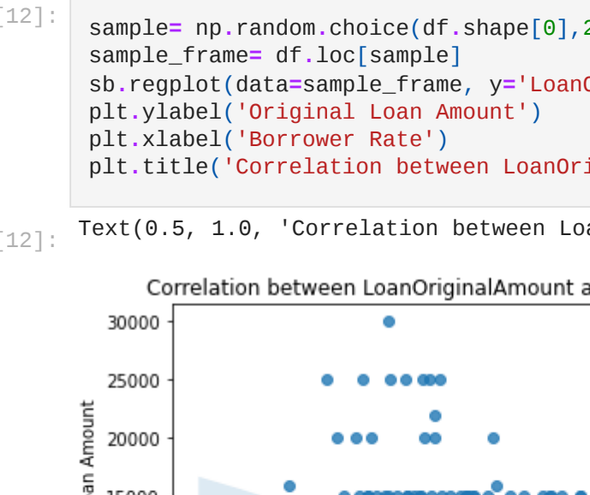
Code for visualization

In [12]:

```
sample= np.random.choice(df.shape[0],200,replace=False)
sample_frame= df.loc[sample]
sb.regplot(data=sample_frame, y='LoanOriginalAmount', x='BorrowerRate')
plt.ylabel('Original Loan Amount')
plt.xlabel('Borrower Rate')
plt.title('Correlation between LoanOriginalAmount and Borrower Rate')
```

Out[12]:

```
Text(0.5, 1.0, 'Correlation between LoanOriginalAmount and Borrower Rate')
```



Observation 9:

It seems that there is a negative correlation associated between BorrowerRate and Original Loan Amount. It depicts, when the Borrower Rate increases, the loan Original amount decreases because of the assumed financial risk.

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?

> Those who belong to the Employed employment status earns more than any other employment status. And those whose income range is more than 100,000, their chance of getting the loan is higher than any other income range.
>
> ### Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?
>
> One can assume the borrower rate based on the original loan amount. As the loan amount increases the borrower rate decreases. and at the same time, there is a positive correlation between Original loan amount and prosper score. mWith the increased prosper score the chance of getting a high first loan amount increases.

Multivariate Exploration

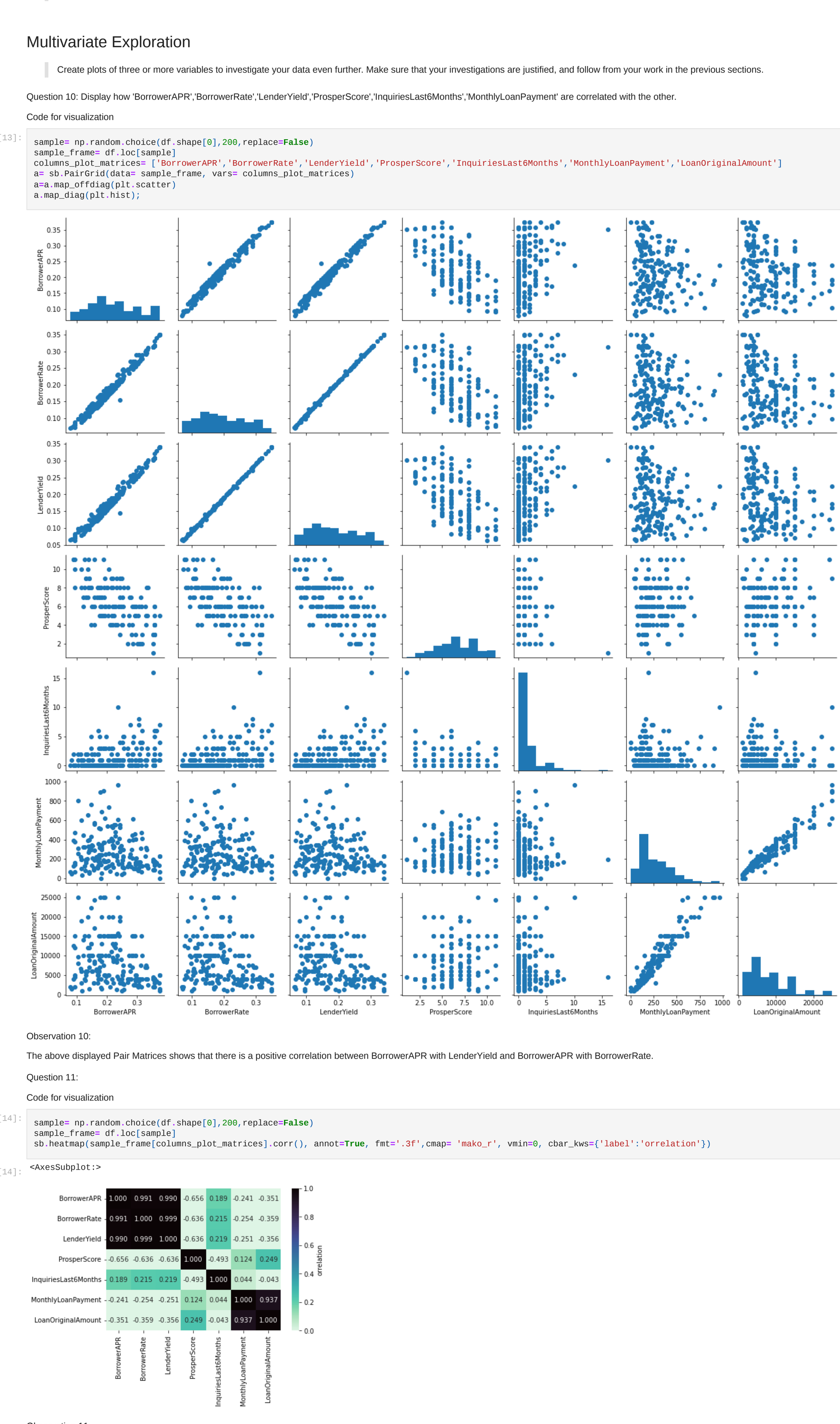
> Create plots of three or more variables to investigate your data even further. Make sure that your investigations are justified, and follow from your work in the previous sections.

Question 10: Display how 'BorrowerAPR','BorrowerRate','LenderYield','ProsperScore','InquiriesLast6Months','MonthlyLoanPayment' are correlated with the other.

Code for visualization

In [13]:

```
sample= np.random.choice(df.shape[0],200,replace=False)
sample_frame= df.loc[sample]
columns_plot_matrices= ['BorrowerAPR','BorrowerRate','LenderYield','ProsperScore','InquiriesLast6Months','MonthlyLoanPayment','LoanOriginalAmount']
a= sb.PairGrid(data= sample_frame, vars= columns_plot_matrices)
a=a.map_offdiag(plt.scatter)
a.map_diag(plt.hist);
```



Observation 10:

The above displayed Pair Matrices shows that there is a positive correlation between BorrowerAPR with LenderYield and BorrowerAPR with BorrowerRate.

Question 11:

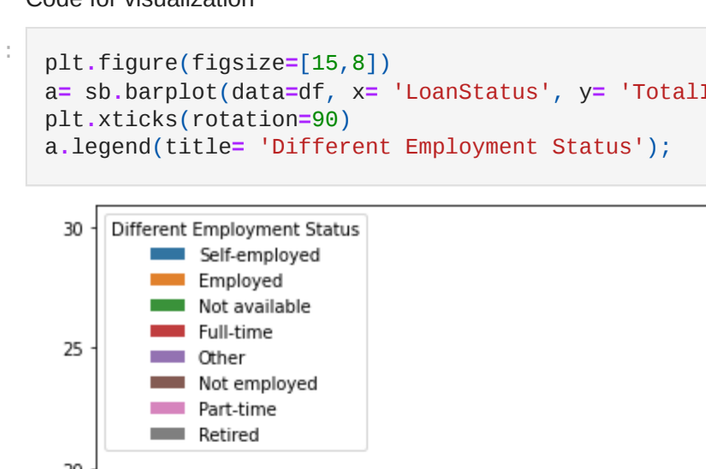
Code for visualization

In [14]:

```
sample= np.random.choice(df.shape[0],200,replace=False)
sample_frame= df.loc[sample]
sb.heatmap(sample_frame[columns_plot_matrices].corr(), annot=True, fmt='.3f',cmap= 'mako_r', vmin=0, cbar_kws={'label':'orrelation'})
```

Out[14]:

```
<AxesSubplot:>
```



Observation 11:

The above projected heatmap is trying to confirm the same output the Plot Matrices projected. It is depicting that there is a positive correlation between BorrowerAPR with LenderYield and BorrowerAPR with BorrowerRate.

Question 12: What is the mean 'total inqueries'of the different Loan Status based on different ' Employment Status'?

Code for visualization

In [15]:

```
plt.figure(figsize=[15,8])
a= sb.barplot(data=df, x= 'LoanStatus', y= 'TotalInquiries',hue= 'EmploymentStatus')
plt.xticks(rotation=90)
a.legend(title= 'Different Employment Status');
```

Observation 12:

The above displayed adoapted bar chart shows that the mean 'total inquiries'of the Loan Status 'Past Due (16-30 days)'is higher within the 'Full Time Employment Status'. The mean of the 'Final Payment in Progress' Loan Status is higher within the 'Other' employment Status.

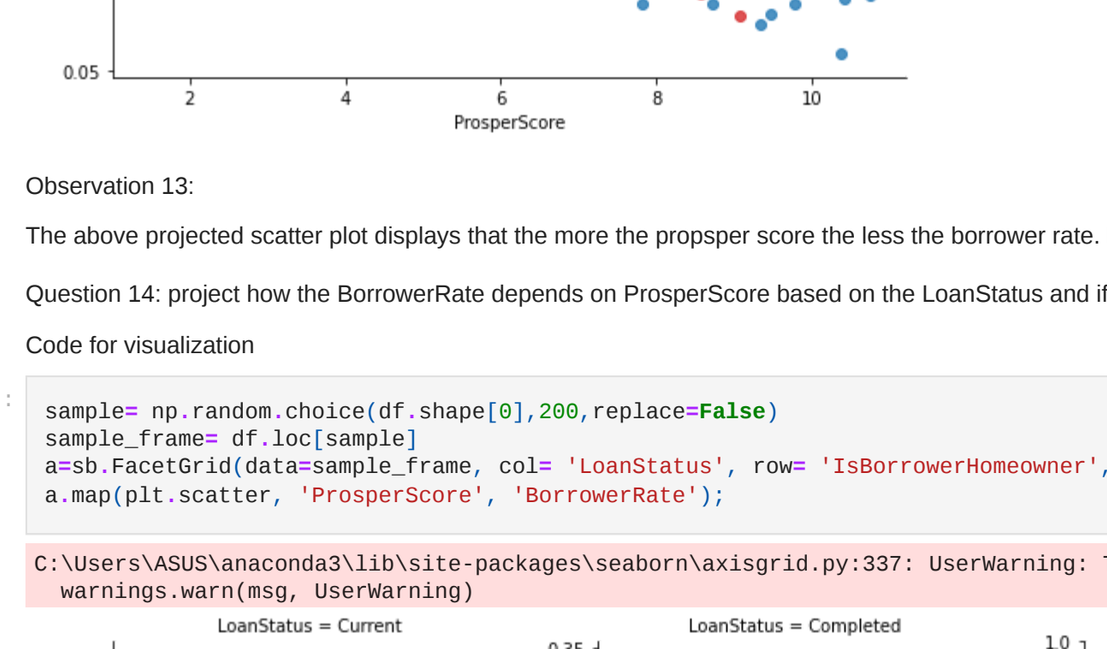
Question 13: How the 'Borrower Rate ' of different 'LoanStatus' are distributed based on 'ProsperScore'

Code for visualization

In [16]:

```
sample= np.random.choice(df.shape[0],100,replace=False)
sample_frame= df.loc[sample]
a= sb.FacetGrid(data=sample_frame, hue= 'LoanStatus',size=5, aspect=1.5)
a=a.map(sb.regplot,'ProsperScore','BorrowerRate', x_jitter=.0, fit_reg=False)
a.add_legend(title='Different Loan Status');
```

```
C:\Users\ASUS\anaconda3\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



Observation 13:

The above displayed scatter plot displays that the more the propsper score the less the borrower rate. It can be said that the BorrowerRate is higly dependent on the ProsperScore.

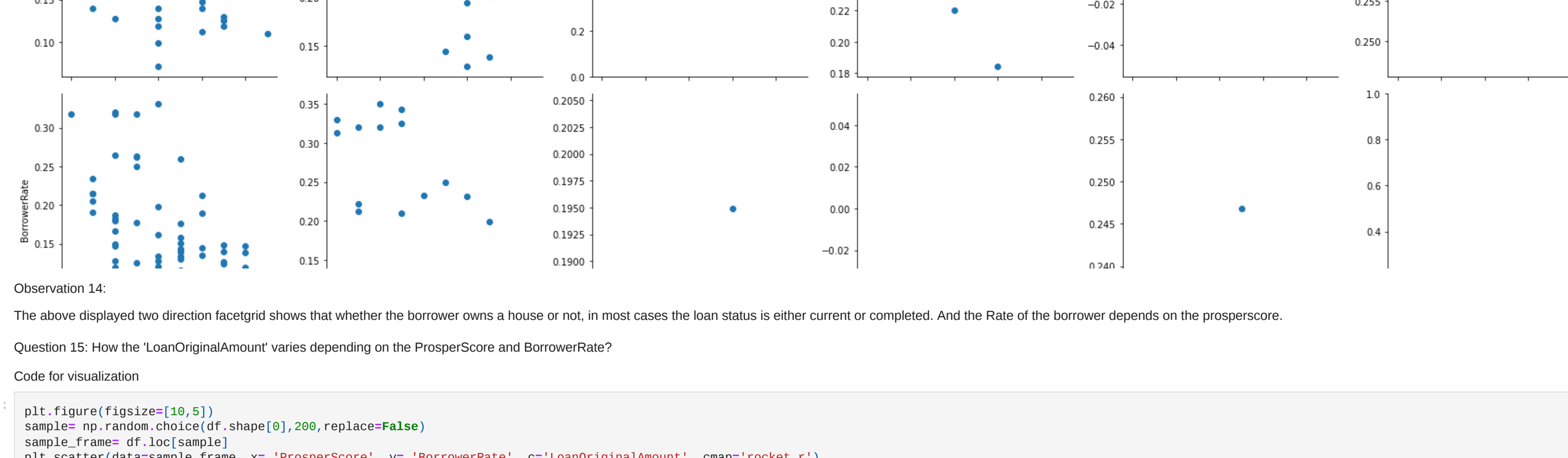
Question 14: project how the BorrowerRate depends on ProsperScore based on the LoanStatus and if the borrower has a house or not

Code for visualization

In [17]:

```
sample= np.random.choice(df.shape[0],200,replace=False)
sample_frame= df.loc[sample]
a=sb.FacetGrid(data=sample_frame, col= 'LoanStatus', row= 'IsBorrowerHomeowner', size=4, margin_titles=True, sharey=False)
a.map(plt.scatter, 'ProsperScore', 'BorrowerRate');
```

```
C:\Users\ASUS\anaconda3\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



Observation 14:

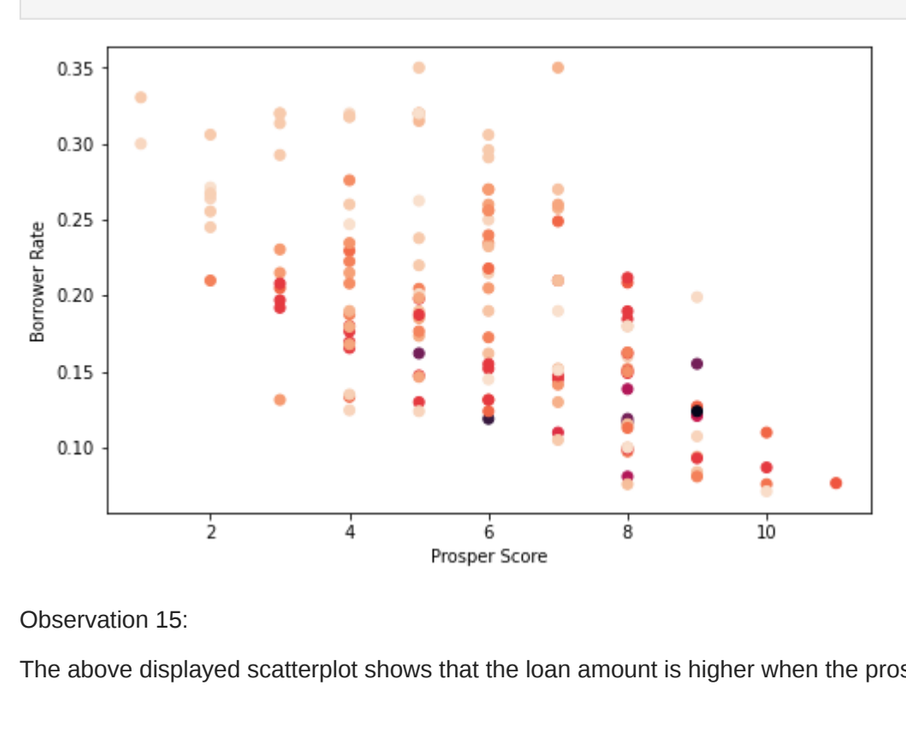
The above displayed two direction facetgrid shows that whether the borrower owns a house or not, in most cases the loan status is either current or completed. And the Rate of the borrower depends on the prosperscore.

Question 15: How the 'LoanOriginalAmount' varies depending on the ProsperScore and BorrowerRate?

Code for visualization

In [18]:

```
plt.figure(figsize=[10,5])
sample= np.random.choice(df.shape[0],200,replace=False)
sample_frame= df.loc[sample]
plt.scatter(data=sample_frame, x= 'ProsperScore', y= 'BorrowerRate', c='LoanOriginalAmount', cmap='rocket_r')
plt.colorbar(label='Loan Original Amount')
plt.xlabel('Prosper Score')
plt.ylabel('Borrower Rate');
```



Observation 15:

The above displayed scatterplot shows that the loan amount is higher when the prosper score is between 5 to 10 and at the same time the Borrowe Rtae is also lower. When the prosper score is low, the borrower rate is higher and at the same time the loan amount is lower.

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?

> Getting a loan is not depedent on whether a brrower owns a house or nat rather it depends on prosper score, employment status and income range.

Were there any interesting or surprising interactions between features?

> It is surprising to see that the borrower with full time employment status are more inquired than the other employment status. And, the Loan Original Amount and Borrower Rate depend on the Borrower's prosper score

Conclusions

> Those who belong to the 'Employed' employment status earns more than any other employment status. And those whose income range is more than 100,000, their chance of getting the loan is higher than any other income range. One can assume the borrower rate based on the original loan amount. As the loan amount increases the borrower rate decreases. And at the same time, there is a positive correlation between Original loan amount and prosper score. With the increased prosper score the chance of getting a high first loan amount increases. Getting a loan is not depedent on whether a brrower owns a house or nat rather it depends on prosper score, employment status and income range.It is surprising to see that the borrower with full time employment status are more inquired than the other employment status. And, the Loan Original Amount and Borrower Rate depend on the Borrower's prosper score

In [ ]:

In [ ]: