(Prosper Loan Dataset Visual Analysis)

by (Ujwala K)

Investigation Overview

Mainly my focus will be to answer below questions:-¶
1)Factors responsible for loan status outcome such as completed or cancelled or past-due etc
2)Factors affecting borrower's APR(which usually includes broker fees, closing costs, rebates, and discount points) and borrower's interest rate
3)Large loan amount differentiation factors(Are there differences between loans depending on how large the original loan amount was?)

Dataset Overview

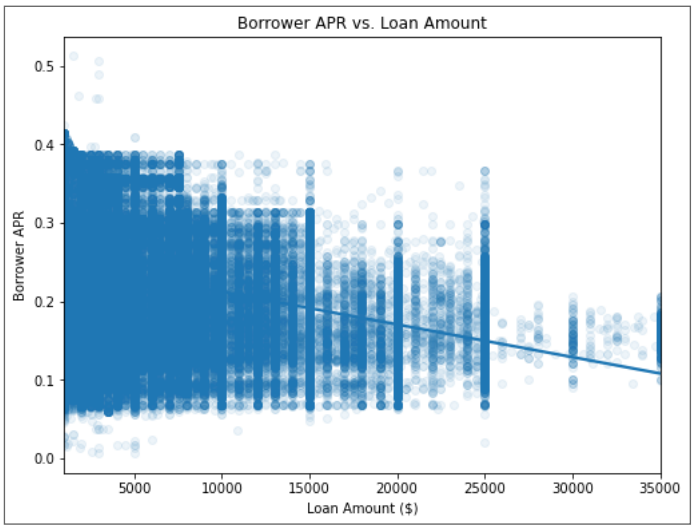
This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others.

Borrower APR vs. Loan Amount

> *This plot again shows the same observation as of above correlation map that range of APR decrease with the increase of loan amount. Overall, the borrower APR is negatively correlated with loan amount.*

In [6]:

```
plt.figure(figsize = [8, 6])
sb.regplot(data = df, x = 'LoanOriginalAmount', y = 'BorrowerAPR', scatter_kws={'alpha':0.08});
plt.xlabel('Loan Amount ($)')
plt.ylabel('Borrower APR')
plt.title('Borrower APR vs. Loan Amount');
```



BorrowerAPR and LoanOriginalAmount relation with categorical Term and Credit-ProsperScores

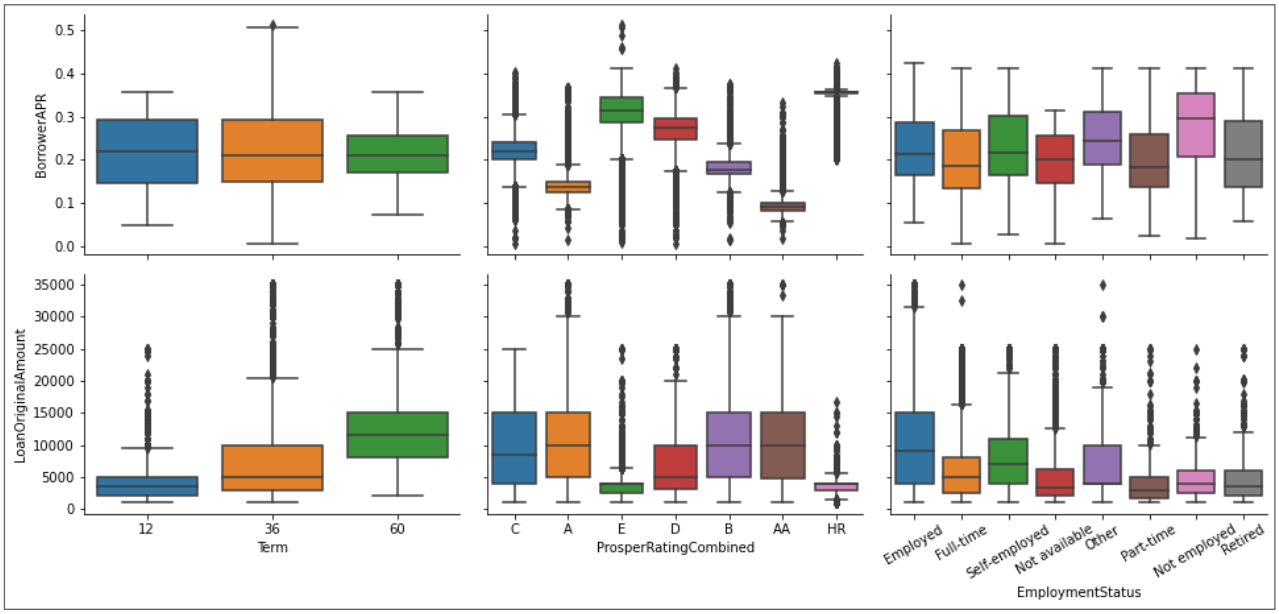
1)Employmentstatus with Employment definitely has LoanAmount sactioned more as per box plot quartiles , however BorrowerRate seems almost same for all employtment statuses.
2)Borrowers with better rating also have higher loan amount.
3)For higher Loan Amounts , term is also high.

In [7]:

```
# plot matrix of numeric features against categorical features.


plt.figure(figsize = [10, 10])
g = sb.PairGrid(data = df, y_vars = ['BorrowerAPR', 'LoanOriginalAmount'],
                x_vars = cat_vars, height = 3, aspect = 1.5)
g.map(sb.boxplot);
plt.xticks(rotation=30);
```

```
<Figure size 720x720 with 0 Axes>
```



Heatmap of Median Estimated Returns by Credit Rating and Income Range

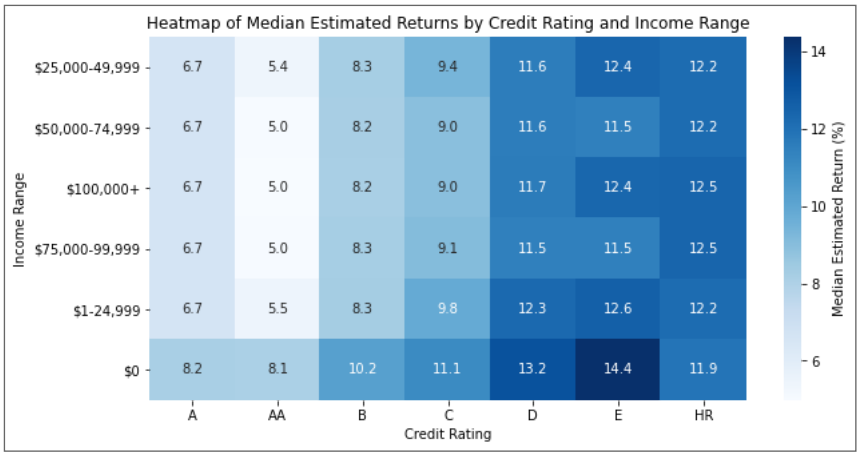
From the heat map we can see that for lower credit ratings, and lower incomes we have higher estimated returns. However income range is not influencing the estimated returns by credit rating category.

In [8]:

```
plt.figure(figsize = [10,5])

cat_med = df.groupby(['ProsperRatingCombined', 'IncomeRange']).median()['EstimatedReturn']*100
cat_med = cat_med.reset_index(name = 'EstimatedReturnMedian')
cat_med = cat_med.pivot(index = 'IncomeRange', columns = 'ProsperRatingCombined', values = 'EstimatedReturnMedia

sb.heatmap(cat_med, annot = True, fmt = '.1f', cmap = "Blues", cbar_kws = {'label' : 'Median Estimated Return (%
plt.xlabel('Credit Rating')
plt.ylabel('Income Range')
plt.title('Heatmap of Median Estimated Returns by Credit Rating and Income Range');
```

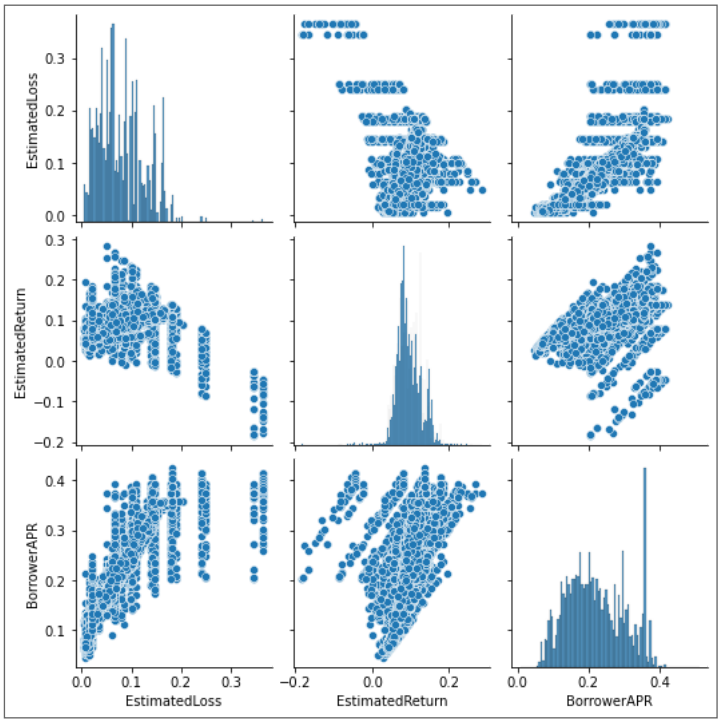


Pairplot for BorrowerAPR and Estimated loss , Return relationship

Its clear from the plots above that BorrowerAPR has positive relationship with EstimatedLoss and EstimatedReturn , BorrowerAPR increases with both of them.

In [9]:

```
sb.pairplot(df[['EstimatedLoss', 'EstimatedReturn', 'BorrowerAPR']]);
```



LoanStatus and Estimated loss , Return relationship

From above Box plots ,we can observe that ChargedOff and Defaulted Loan Statuses Borrowers have higher EstimatedLoss and EstimatedReturn percentage

In [10]:

```
g = sb.PairGrid(data = df, y_vars = ['EstimatedReturn', 'EstimatedLoss'],
                x_vars = ['LoanStatus'] , height=7)
g.map(sb.boxplot);
plt.xticks(rotation=90);
```