Part II - Prosper Loans Analysis

By Bessan Hussein

My Github: @Bessan Hussein

My LinkedIn: @Bessan Hussein

Investigation Overview

The primary objective of this investigation is to conduct a comprehensive exploratory data analysis (EDA) on a loan dataset to uncover key trends, patterns, and relationships among various loan attributes. By visualizing and analyzing the data, we aim to gain insights into loan performance, borrower characteristics, and market dynamics. These findings will inform strategic decision-making, risk assessment, and the development of targeted loan products.

Dataset Overview and Executive Summary

On November 24, 2008, the SEC found Prosper to be in violation of the Securities Act of 1933. As a result of these findings, the SEC imposed a cease and desist order on Prosper ... In July 2009, Prosper reopened their website for lending ("investing") and borrowing after having obtained SEC registration for its loans ("notes"). After the relaunch, bidding on loans was restricted to residents of 28 U.S. states and the District of Columbia. Borrowers may reside in any of 47 states, with residents of three states (Iowa, Maine, and North Dakota) not permitted to borrow through Prosper

The dataset comprises loan-level information including loan amounts, interest rates, borrower demographics, employment details, and loan performance metrics. This EDA focuses on understanding loan characteristics, borrower behavior, and market trends.

```
# 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 sns

# suppress warnings from final output
import warnings
warnings.simplefilter("ignore")
```

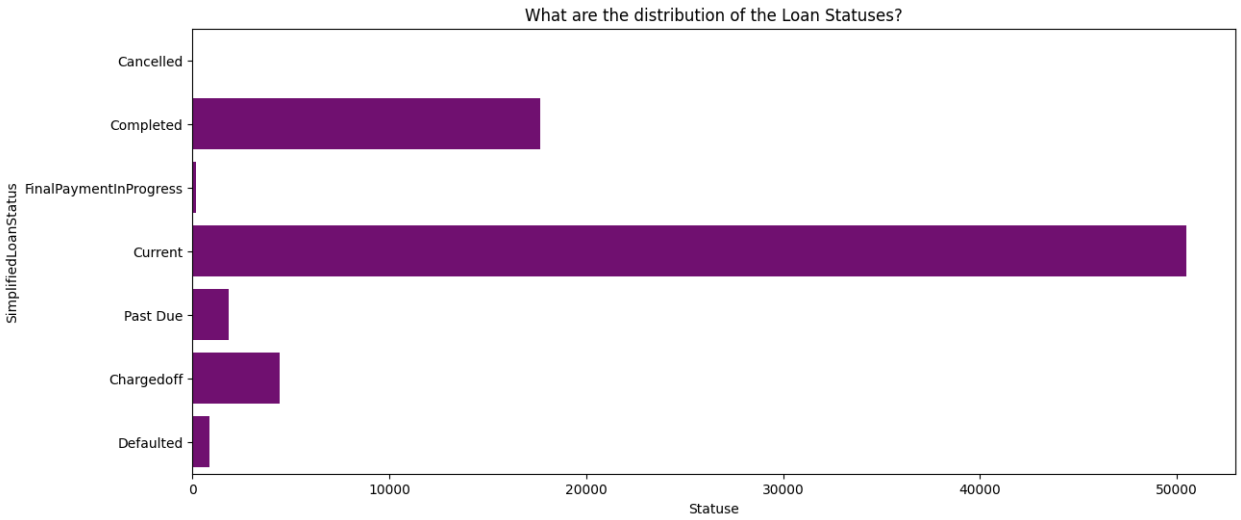
load dataset

```
df = pd.read_csv("./data/prosperLoanDataCleaned.csv") ## Load the csv
into pandas dataframe
```

What are the distribution of the Loan Statuses?

```
## add orderring to the loan status and include the cancelled ones
loan_status_order = ['Cancelled', 'Completed',
'FinalPaymentInProgress', 'Current',
                     'Past Due', 'Chargedoff', 'Defaulted']

plt.figure(figsize=(14, 6))
sns.countplot(data = df, y = 'SimplifiedLoanStatus', color = 'purple',
order=loan_status_order)
plt.xlabel('Statuse')
plt.title('What are the distribution of the Loan Statuses?');
```



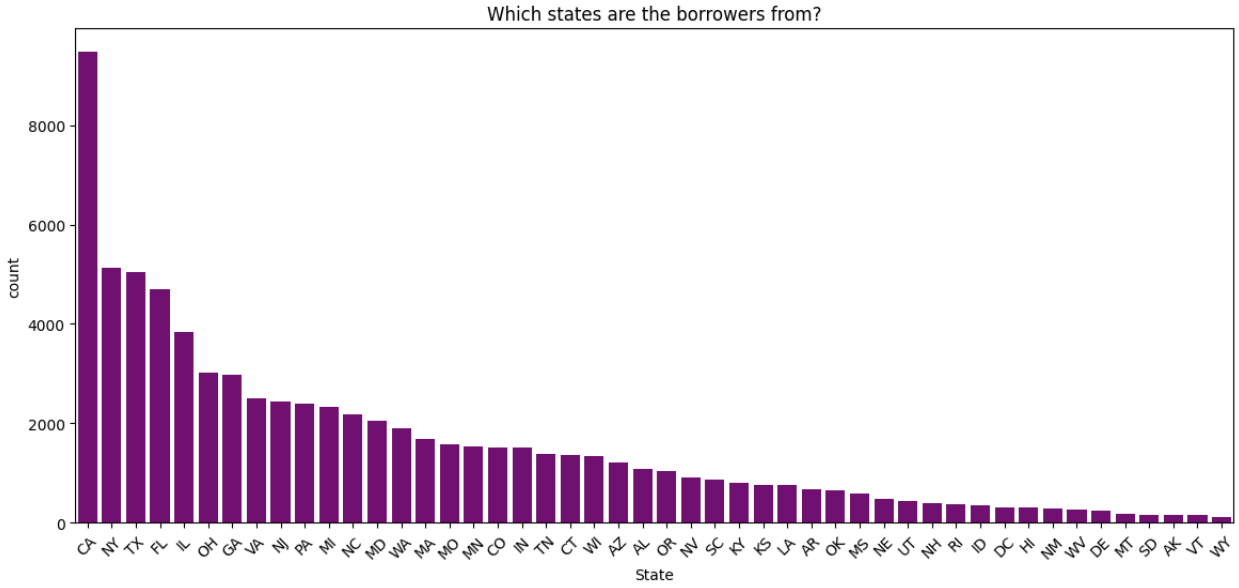
It seems that the majority status among the loan statuses is the `current` which is 50462, While the completed are success story they are `17675`. The chargedoff however are 4444 cases they are failed loans. While the `deafulted` are the cases in danger of `chargedoff` they reach 885. And the past due are slightly larger 1835 but still not in danger of charge off.

The chraged off and the deafulted represntts the actual risks for these loans.

Nothing unusual with this distribution.

Which states are the borrowers from?

```
counts_borrower = df['BorrowerState'].value_counts()
plt.figure(figsize=(14, 6))
ax = sns.barplot(counts_borrower, color = 'purple')
plt.title('Which states are the borrowers from?')
plt.xlabel('State')
plt.xticks(rotation = 45);
```



California boasts the highest number of borrowers among all US states, followed closely by Texas, New York, and Florida. On the other end of the spectrum, states like South Dakota, Alaska, Vermont, and Wyoming have significantly lower borrower counts.

California is an unpper bound outlier, however this is numerically and valid values that will not be handled.

What are the employment statuses of students who are taking loans?

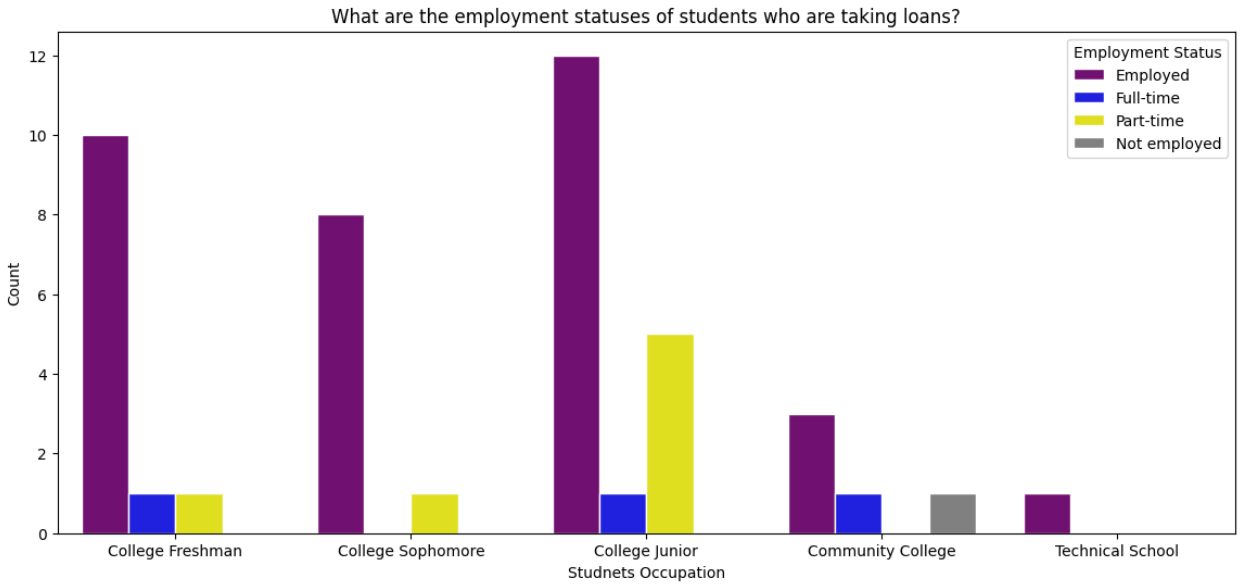
```
## List of students occupations ordered by their natural order
study_occupations = ['Student - College Freshman',  'Student - College
Sophomore',  'Student - College Junior',
 'Student - Community College', 'Student - Technical School']

## Labels for the visualization
labels = ['College Freshman', 'College Sophomore', 'College Junior',
'Community College', 'Technical School']

## the natural order of employment staus
employment_status_hue_oredr = ['Employed', 'Full-time', 'Part-time',
'Not employed']

plt.figure(figsize=(14, 6))
sns.countplot(data=df[df['Occupation'].isin(study_occupations)],
x='Occupation', palette=['purple','blue', 'yellow',  'grey'],
              hue='EmploymentStatus',
hue_order=employment_status_hue_oredr,  order = study_occupations,
edgecolor = 'white')
plt.title('What are the employment statuses of students who are taking
loans?')
```

```
plt.xlabel('Studnets Occupation')
plt.ylabel('Count')
plt.xticks(ticks= study_occupations, labels= labels)
plt.legend(title='Employment Status');
```



Employment is widespread among students taking loans, with the majority falling into the "Employed" category across all educational levels.

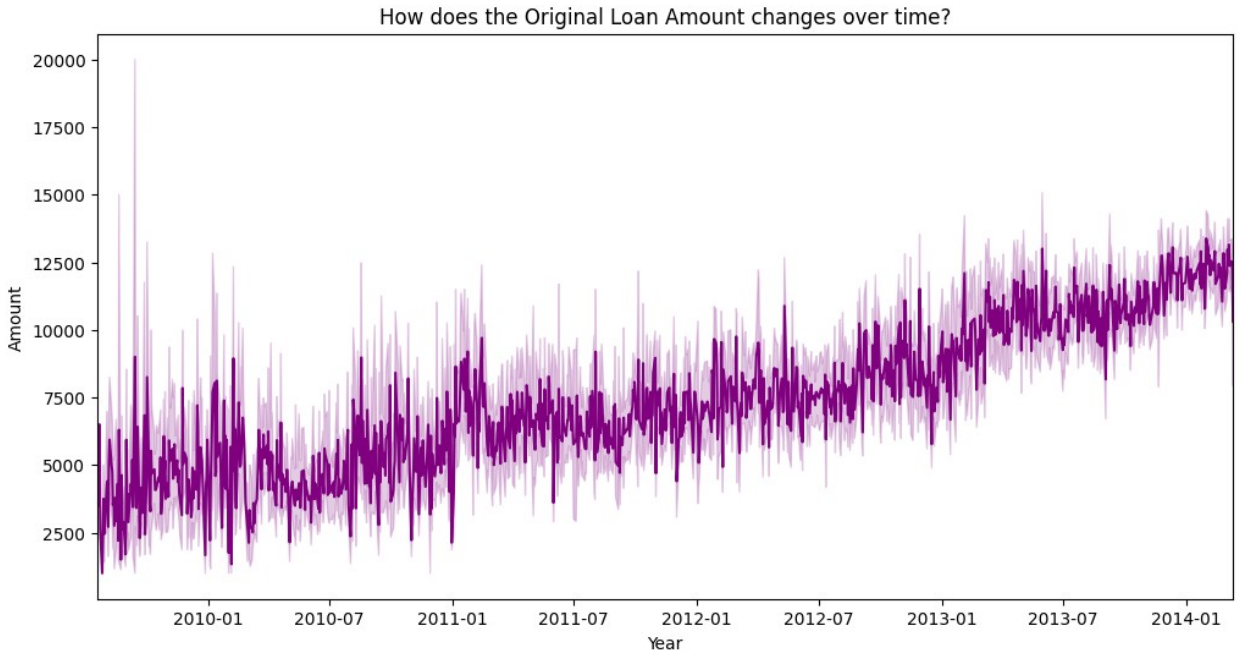
Part-time work appears to be the most common arrangement, indicated by the higher count of "Part-time" compared to "Full-time" in most categories. Expected since they are students.

Community college students exhibit a notably higher proportion of full-time employment compared to other groups.

The "Technical School" category shows a lower overall employment rate and a higher percentage of students who are not employed.

How does the Original Amount of Loan change over time?

```
df['LoanOriginationDate'] = pd.to_datetime(df['LoanOriginationDate'])
## convert to datetime datatype
plt.figure(figsize=(12, 6))
sns.lineplot(data= df, x = 'LoanOriginationDate', y
='LoanOriginalAmount', color = 'purple')
plt.ylabel('Amount')
plt.xlabel('Year')
plt.xlim(pd.to_datetime('2009-07-20'), pd.to_datetime('2014-03-12'))
## the max and min dates of the Loan Origination Date
plt.title('How does the Original Loan Amount changes over time?');
```



There is a noticeable increase in loan amounts over the years, suggesting that borrowers have been taking out larger loans as time progresses.

Interestingly, the year 2009 stands out with some of the highest loan amounts, indicating that during this period, there were notably larger loans compared to other years.

How does the original loan amount vary across different income ranges and loan statuses?

```
loan_status_order = ['Completed', 'FinalPaymentInProgress', 'Current',
'Past Due', 'Chargedoff', 'Defaulted']

## Facet Plotting three variables SimplifiedLoanStatus,
LoanOriginalAmount and IncomeRange
grid = sns.FacetGrid(data = df, col = 'SimplifiedLoanStatus', col_wrap
= 3, col_order= loan_status_order)
grid.map(sns.barplot, 'LoanOriginalAmount', 'IncomeRange', color =
'Purple')
grid.set_titles("{col_name}")
grid.set_axis_labels('Loan Original Amount', 'Income Range')
plt.suptitle('How does the original loan amount vary across different
income ranges and loan statuses?', y=1.02)
grid.figure.set_size_inches(14, 6);
```

As income range increases, the loan original amount also tends to increase. This is evident in the length of the bars across different income ranges.

The "Current" and "Completed" statuses, where the highest loan original amount is not always associated with the highest income range.

Higher income ranges might be associated with fewer defaults, which is worth investigating.

Across all loan statuses, the highest Loans amounts are associated with the 100,000+ income range.

How do different terms group with the selected variables?

```
selected_columns = ['LoanOriginalAmount',
'BorrowerAPR','DebtToIncomeRatio']

## Plot Matrix
grid = sns.pairplot(data=df, vars=selected_columns, hue='Term',
palette=['purple', 'yellow', 'b'], plot_kws = {'s':2})
plt.suptitle('How do different terms group with the selected
variables?', y=1.02)
grid.figure.set_size_inches(14, 6);
```

Longer terms (36 and 60 months) tend to be associated with higher loan original amounts.

There doesn't seem to be a strong relationship between term and borrower APR or debt-to-income ratio.

Loan Original Amount vs. Borrower APR have a weak positive correlation, suggesting that larger loans might have slightly higher APRs.