Data Visualization Process

I chose the Loan data from Prosper, and chose the Loan status as the dependent variable to explore the what factors in the dataset impact what a client's loan status will be.

In the exploratory data visulaization notebook, I began with looking into the properties of the dataset and what each column represents by using pandas functions such as .info, .describe and .unique to see what type of data was in the independent variables I was interested in investigating further.
I then converted certain columns into categories for further analysis.
I used 3 types of explorations in this notebook, namely univariate, bivariate and multivariate and summed up my findings for each explorations under each one.

Univariate Exploration

I began with univariate exploration in which I used histograms bar chart subplots and bar graphs. In this exploration I looked into Borrower Rates, used the subplots to look at various categorical variables such as the loan status, term and employment status and grouped the income range and plotted it on a bar graph. I then visualized all the null values and dropped all the columns I won't be using for analysis to get a better visual of which columns have null values.

Bivariate Exploration

In this exploration I began with exploring the numbers of homeowners by loan status and then moved on to creating a heatmap with all the numeric independent variables I was investigating to see what the correlation between the values was. Next, I plotted matrices of scatter plots and then pair grids of boxplots. Scatter plots on estimated returns and losses were plotted against rates but were not included in the final analysis as the data of the two columns had many missing values and wouldn't give an accurate representation. Finally I moved on to create boxplot and violin plots for clearer representation of relevant variables.

Multivariate Exploration

Here I plotted point plots to investigate the relationship between 3 variables with the loan status being the constant and represented in each graph against other independent variables.

Summary Findings

The findings that were chosen to be presented in the explanation were the histograms that gave an overview of its numbers for the borrower rates. The homeowner graph was included to give clear views of its impact by loan status. The boxplot and violin plots indicating credit score impact on loan status showed a visual summary statistics of its effect on loan status. The final three graphs indicated the effects income, employment status and length of the loan on the status and each graph provided information on how each factor had an impact on the loan status which was elaborated in the explanation visualization.

Resources Used

- https://stackoverflow.com/questions/52456874/drop-rows-on-multiple-conditions-in-pandas-dataframe
- https://thispointer.com/python-pandas-how-to-drop-rows-in-dataframe-by-conditions-on-column-values/
- https://nbconvert.readthedocs.io/en/latest/
- https://revealjs.com/
- https://github.com/hakimel/reveal.js#installation