‘*Understanding Student Debt Repayment*’ - Adrian Halarewicz

Problem Statement:

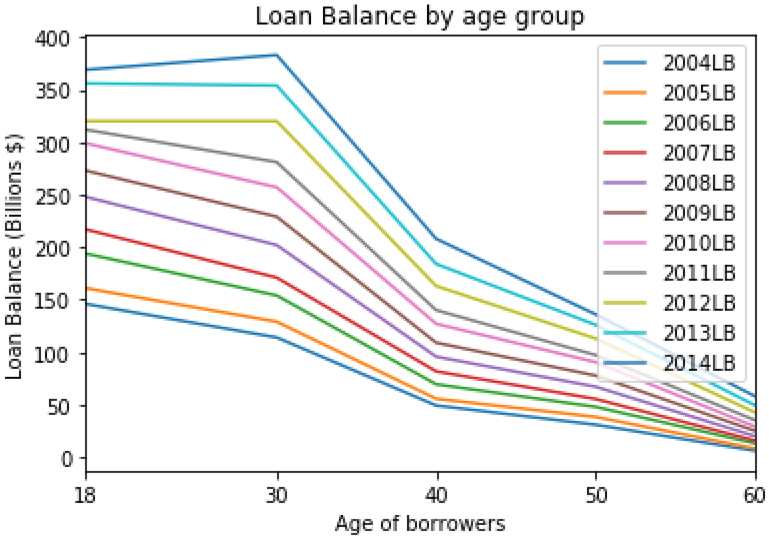
The purpose of this project is to help students obtain a better understanding of the burden student loans impose and predict a borrower’s ability to repay the debt. Many students blindly take on student loans without an accurate or realistic understanding of how the debt will be repaid. Many, if not all, students do not know their approximate after college income before starting their education. Therefore, it is difficult to predict the effect of student loans once a student enters re-payment.

The client is any person applying for a student loan or any person currently enrolled in a student loan program. The client expected to take on student loans will have a strong sense of what field of study and the career they will go into. They will know which city or region they want to live and work in after graduation. The client will know how much of the tuition costs they will be able to pay and how much they are expecting to borrow in the form of student loans. The borrower will also know basic metrics that lending institutions use to classify their borrower’s ability to repay debt.

This model will help the client by providing meaningful insight and guidance about choosing a field of study, how much debt they could potentially handle, and if the costs associated with their school of choice are a financially stable investment. Additionally, the model will predict if the borrower will or will not default on their loan. Understanding their future financial standing and knowing the likelihood of defaulting on loans could strongly influence a student’s decision to attend a particular school and take on student debt.

This project will help students better understand the risk and responsibilities associated with student loan debt. The problem will be solved by analyzing a few key factors that best define a borrower. The model will make predictions based on the borrower’s financial history and will consider metrics such as the amount borrowed, interest rates, and previous credit history. Additionally, the model will report to the borrow how they compare to the rest of the population in terms of occupation, expected salary within the field, and the cost of living for their city.

Meaningful visualizations have been created for the findings to effectively share information with the user. The visualizations are built from current and historical data. We can observe the patterns and trends, but these patterns and trends are not guaranteed to persist in future data. The following figure shows the continuous rise of outstanding student loan balances by age group for a 10 year period. This visualization suggests that outstanding loan balances will continue to rise as students take on more and more debt. As a disclaimer, the total loan balance does not account for the sample size of the borrowers for each observed year. This visualization could indicate two possible situations: Individual borrowers are taking on larger amounts of debt, or there is an increase in the number of borrowers taking on a consistent amount of debt.



Description of the Dataset:

The data for this project came from a variety of different sources pertinent to Student Loan data in the US. The raw CSV files include data for mean income by age, median income by age, number of loans distributed, loan balance by age, earnings by occupation, and cost of living information. The data regarding occupation income data by age was produced from US Census data, while loan specific data came from the federal student aid database and the data.world website. The cost of living index data was obtained from kaggle.

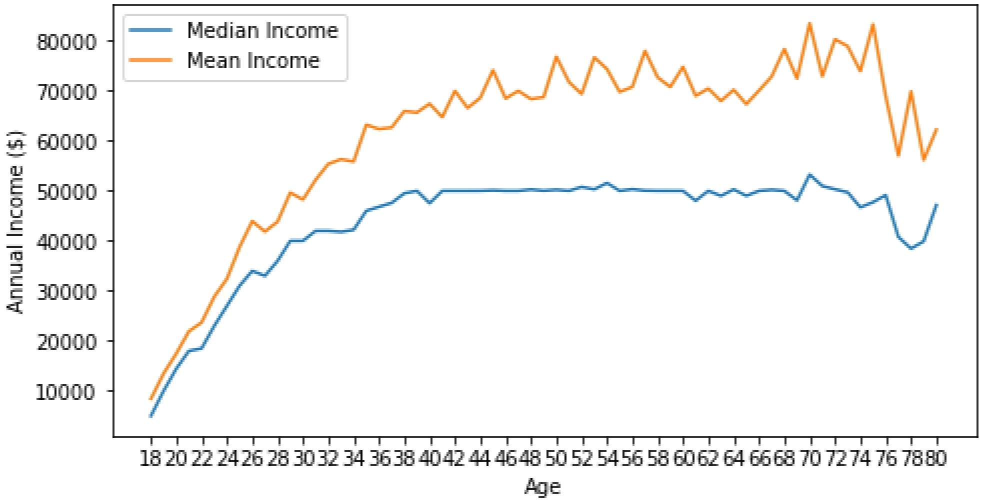
Many of the files were imported in an already clean format, but the census data required a significant amount of wrangling to construct a ‘*tidy*’ (Hadley Wickham, *Tidy Data*) Multi-Indexed DataFrame.

The data for the Machine Learning component of this project was obtained from Lending Club’s public loan data. The data was downloaded directly from the Lending Club website as zipped CSV files. The Lending Club dataset contains over 2 million rows of loan observations. Each row of the dataset corresponds to a single loan issued by lending club, and each column represents relevant features describing the history and status of the loan and its borrower. This data will be used to train a model that will predict if a borrower will or will not default on a particular loan.

EDA and Summary of Findings:

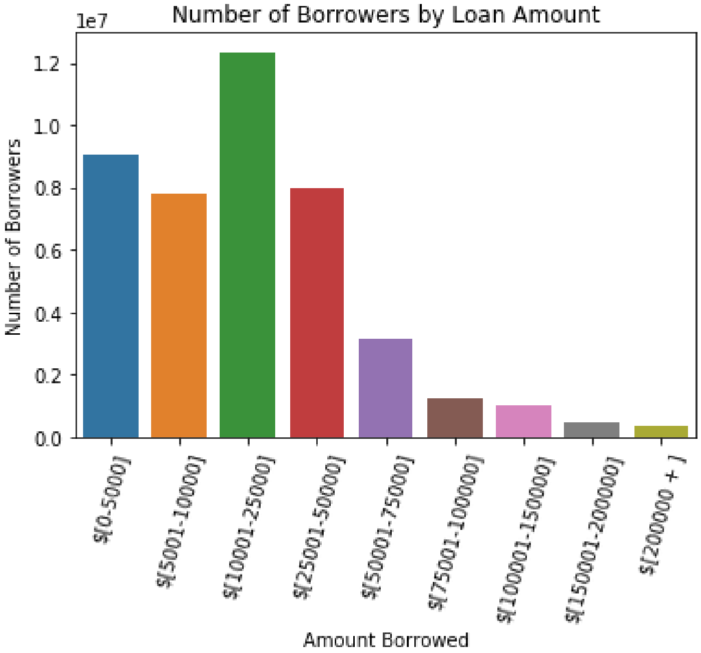
There are many factors to consider when deciding to take on student loan debt. Obtaining a better understanding of expected income can help borrowers better predict their debt to income ratio and their ability to pay back a loan. Asking the right questions can help a borrower explore data and obtain meaningful insights about their future financial standing.

The primary concern with debt is one’s ability to repay the debt. To help explore the debt to income ratio of borrowers, the data for mean and median income can provide meaningful insight. **How does income increase with age, and does having more years of experience lead to higher compensation?**



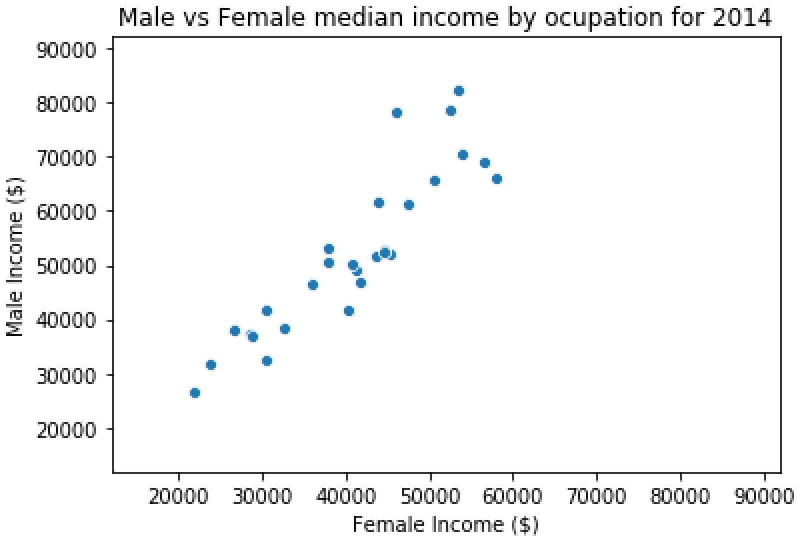
For employees between the age of 18 and 40, there is a direct correlation between compensation and age due to having more years of professional experience. For employees over the age of 40, there is a plateau in the median data that shows there is no significant difference between 20 and 50 years of experience. The mean trend line continues to increase as it accounts for employees in careers that allow for continued growth and increasing compensation. There is a plateau in median income due to much of the workforce being employed in low skill jobs with minimal room for growth.

In addition to better understanding a borrower’s expected income, it will benefit the borrower to better understand where they stand in the student loan debt landscape. **What does the distribution of debt look like, and how will a borrower’s amount of debt compare to other borrowers?**



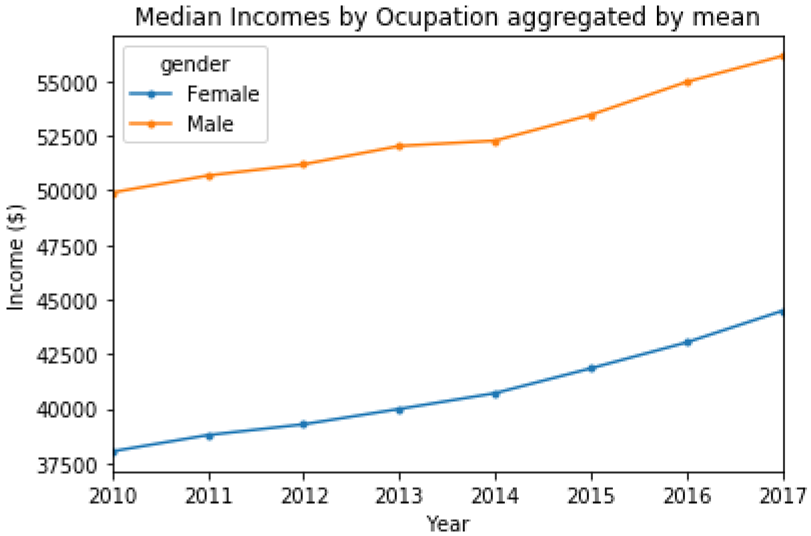
A majority of borrowers hold less than $50,000 of student loan debt. This amount of debt is tolerable with respect to the mean and median income data. For most cases, the amount of student loan debt does not exceed the median annual income. For the cases of borrowers with significantly more debt, those borrowers are typically students pursuing higher education for a specialized career. Those with more student loan debt are more likely to work in higher-paying careers with compensation and growth that justifies the additional investment in their education.

The mean and median data discussed above provides a very general approximation for the amount of income a borrower can expect. Student loan borrowers tend to already have a sense of direction within their education and their desired career path, so it would be more beneficial to look at the median income for specific occupations and how the data differs for males and females. **What is the relationship between median income for males and females within each occupation for a given year?**



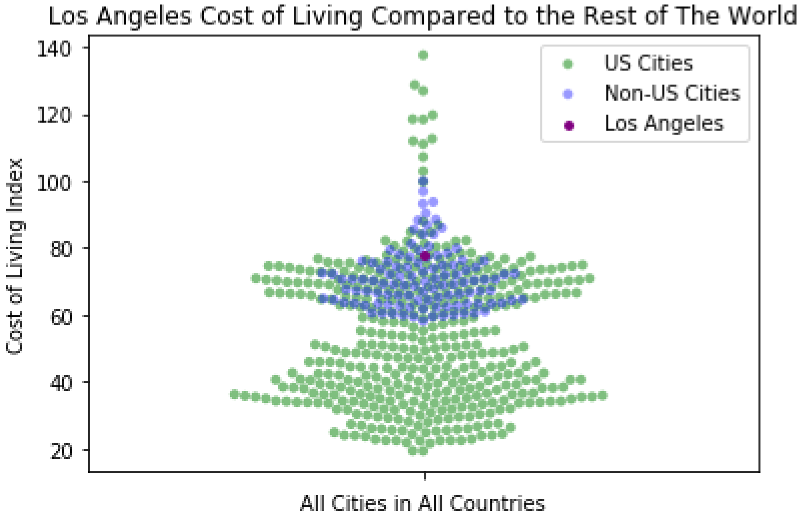
The scatter plot describes a wage gap between males and females. In the scatter plot, each point represents a single occupational category. The horizontal position of a point describes the median income for females in that single occupational category. The vertical position of the point corresponds to the median income of males in that same occupational category. The trend line for this scatter plot has a slope that is greater than 1. This implies that men are compensated more than their female counterparts in the same line of work.

Seeing the presence of a wage gap in the data, it would be beneficial to explore trends in the difference in wages. **What are the trends in the wage gap and how did the incomes for males and females compare over a recent 10-year span?**



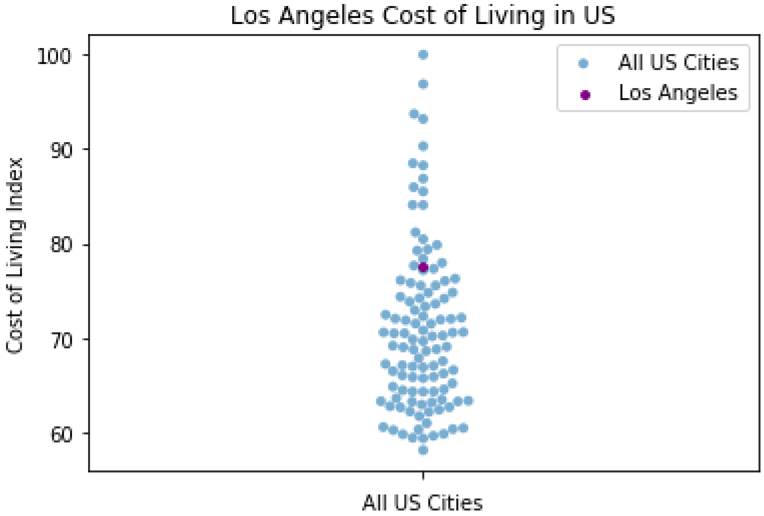
When an average of the median income by occupation is computed separately for males and females, there are two clear trends when consecutive years are plotted as a time-series. The first of the two trends show that the average compensation across the many different occupations is increasing over time. This is good because employees are getting paid more for doing the same types of work. The second trend is in the even amount of separation between the two almost parallel curves. Both the male and female curves increase with time, but the curve describing median income for males remains approximately $12,000 above the median income curve for females. The difference in the curves appears to remain constant without diverging.

Finally, we can examine the cost of living to better inform a borrower about future financial standing. Because this analysis is built on data collected for the US, the borrower should understand how the cost of living differs in the US from other countries. **How does the cost of living index in US cities compare to other cities throughout the world?**



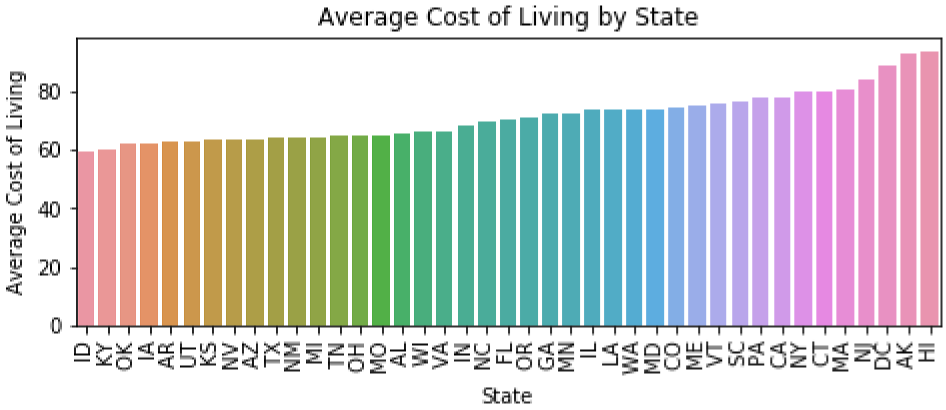
The cost of living throughout the world appears to be a bimodal distribution, which separates the world into two groups. The left clustered peak represents countries with a lower cost of living and the right clustered peak represents countries with a higher cost of living. The US cities appear to be entirely in the group with a higher cost of living. A bee-swarm plot provides a better visualization of the distribution. Again, the US is clearly in the swarm of cities with a higher cost of living with a similar distribution to its respective swarm.

Furthermore, A borrower will want to know how the cost of living in their particular city compares to other cities in the US. **How does the cost of living in a single city compare to other cities in the US?**



The answer will focus on Los Angeles, CA as a single example. This beeswarm plot shows the cost of living in Los Angeles as it compares to other cities in the US. It is clear that Los Angeles is among the US cities with a higher cost of living, but is surprisingly not the highest. This graph can be overlayed with the world data to understand how Los Angeles compares to other cities throughout the world.

A final thought for the borrower would be to compare the average cost of living in different states. **How does the average cost of living vary by state?**



This final graph shows the borrower how the cost of living differs from state to state. The bar plot was constructed by grouping the cities by state and aggregating them by their mean.

With a better sense for expected income by occupation and the associated cost of living for each city, borrowers will be more informed when taking on student loans. Though the return on investment is biased for men due to the wage gap, it is possible to repay debt and not be crippled by an extreme debt to income ratio. By structuring your education around a high skill occupation, a student loan borrower will be able to repay their debt as they progress through their professional career.

Statistical Analysis:

The DataFrame containing the cost of living index for cities throughout the world can be treated and processed as a sample for determining the true mean of the cost of living. The statistical analysis for this project will focus on the cost of living data and obtaining a sense of the true mean for the cost of living throughout the United States.

First, we can examine the distribution of the data points representing the cost of living index for individual cities. Because the scope of this project is built around student loan data in the United States, the data for the rest of the word will not play much of a significant role in this statistical analysis. It would be beneficial for the user to understand how the cost of living index of their particular city compares to the average cost of living index in the US.

Assuming there are cities in the US that are not accounted for in the cost of living index data, a challenge appears when trying to determine the average cost of living index throughout all cities in the US. To solve this problem, different methods of statistical inference can be applied to our sample. This project utilizes a Frequentist, Bootstrap, and Bayesian approach to statistical inference to determine the mean cost of living index in the US.

For the three individual methods of inference, the results are as follows:

Frequentist	[69.26 , 72.55]	95% Confidence Interval
Bootstrap	70.9	Most probable mean
Bayesian	68.86	Most probable mean

Machine Learning:

Once the Lending Club data for all loans from 2007 to Q3 of 2019 has been collected and cleaned, the data set is ready for the application of machine learning. Several different machine learning algorithms were utilized to explore and make predictions on the dataset. Supervised Learning algorithms such as Lasso Regression, Random Forest, and SVM were used to identify key features of the dataset and predict loan status, while unsupervised machine algorithms such as KMeans clustering and Primary Component Analysis were used to explore the data points.

Because the cleaned and wrangled dataset contains 51 features, Lasso Regression would play a vital role in determining which features have the most predictive power when considering loan status. The Lasso regression revealed that the amount of outstanding principal (out_prncp) and the last payment amount (last_pymnt_amnt) were the features with the largest direct correlation to a loan entering default. On the other hand, loan amount (loan_amnt), total payment (total_pymnt), and total interest received (total_rec_int) were the features with the most negative correlation to loans entering default. Swarm plots for these features show the distribution of the respective features for loans not in default overlayed with the respective features of loans in default.

To further explore the data set and look for trends in the individual data points, KMeans clustering was applied to the data set. In order to determine the best number of clusters for the dataset, the sum of squares was computed for different values of n_clusters. Using the knee/elbow method revealed that 3 or 4 clusters significantly reduced the squared sum of distances from each point to their respective cluster centers. Because the elbow in the graph was not sharp enough to exactly pinpoint the best number of clusters for KMeans, silhouette analysis was also applied to the dataset. In combination with the Primary Component Analysis that reduced the dimensionality of the dataset to two features, the scatter plot of the reduced dataset showed four sets of clustered data points.

To make initial predictions on the default status of loans, a Random Forest was trained with its default hyperparameters. When evaluating the model, the confusion matrix and specifically the number of True Positives revealed the model's lack of ability to correctly predict if a loan will default. More specifically, the number of True Positives and False Positives were both zero, indicating that the model predicted that all loans will not default. After further investigation, the cause of the problems is the presence of a severe class imbalance in the target column of the data set.

A sample of the dataset revealed that 619,995 loans did not default and only 853 loans did default. This is understandable as Lending Club only wants to issue loans that will be repaid in-full and avoid default, therefore minimizing the number of loans that enter default. To fix the issue of class imbalance, SMOTE from imbalanced-learn was applied to oversample the number of default loans. Doing so will decrease the overall accuracy of the model, but will increase the model’s ability to correctly predict if a loan will default. From a business perspective, it is not a major issue if loans that will not default are classified as a loan that will default, but it would be very detrimental to the lending institution if loans that will default go unrecognized and are classified as a loan that will be successfully repaid. Therefore, it is in the best interest of the lending institution to maximize the model’s ability to correctly identify the loans that will default.

After implementing the Synthetic Minority Oversampling Technique to the training data, the newly trained model performed with 71% accuracy. But more importantly, the number of True Positives increased and the number of False Positives decreased in the confusion matrix. The area under the ROC curve was calculated to be 0.7, which indicates the model has a significant level of predictive power.

To practice building a supervised classifier and tuning its hyperparameters, a Support Vector Machine was also implemented to predict if a loan will or will not default. To maximize the number of True Positives that the SVM is able to predict, several modifications had to be made to the data and the hyperparameters.

Unlike a random forest, a support vector machine requires the features to be scaled in order to avoid distance bias within the features. The features of the dataset were scaled with scikit-learn’s StandardScaler, and the initial model was trained on the data. Again, due to class imbalance of the target variable, SMOTE was used to oversample the data for loans that did default. This improved the model because the model no longer predicted that all loans will not default.

To further improve the model, Grid Search Cross Validation was implemented to identify the best hyperparameters for the SVM. After individually training models on the scaled data with all of the possible combinations of *C* and *gamma*, the GridSearcCV returned the best combination of hyperparameters.

Due to the size of the Lending club dataset, the SVM was trained on a sample of the original data set. A pipeline was created to scale the data, identify the best parameters with grid search and make predictions on the data with an SVM trained with the best combination of hyperparameters. With the pipeline running successfully, the complete data set can be fed in the data to improve the quality of the model and its predictive power.