**SBA Loans Analysis**

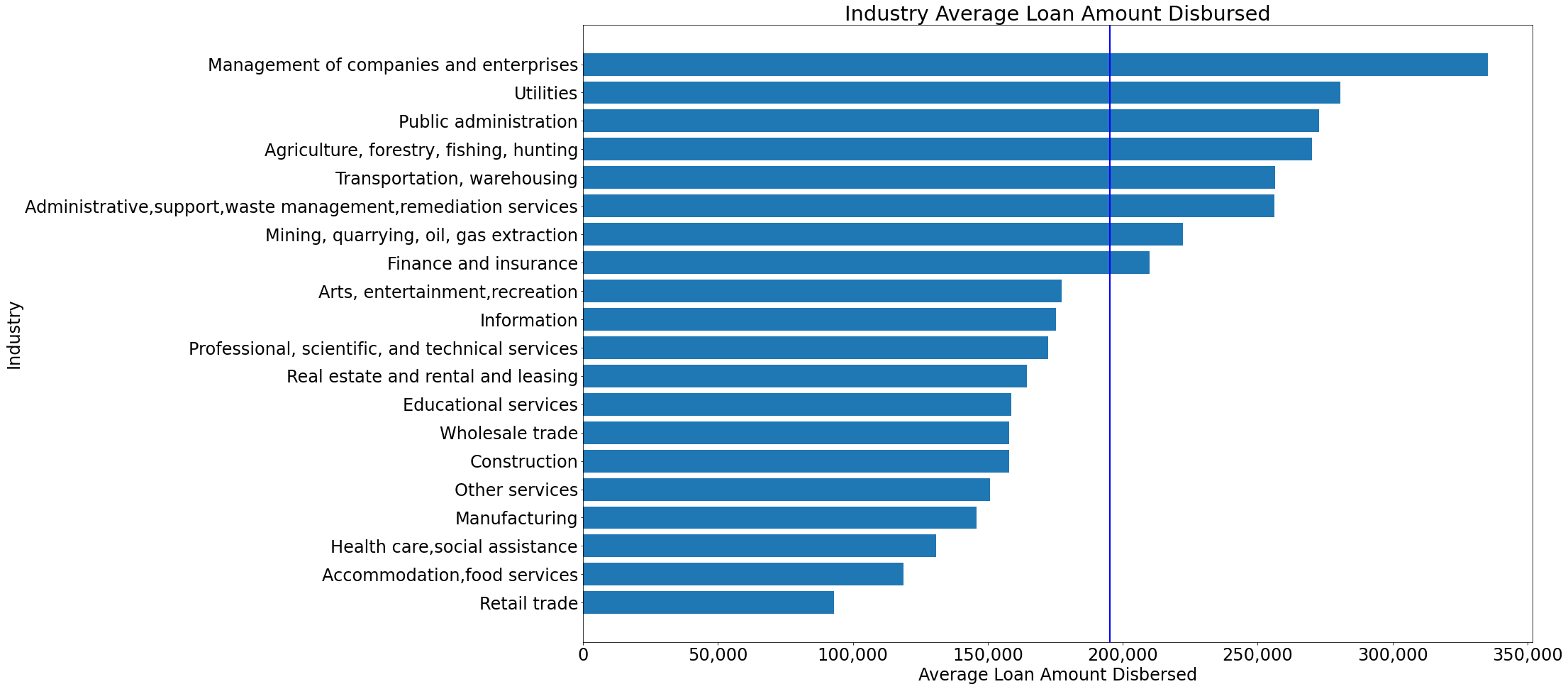
By: Eduardo Ortiz, Morgan Prukop, and Deautaun Ross

If you’re a small business owner in need of financial assistance, getting a loan from the Small Business Administration (SBA) is quite a viable option in your arsenal. In this article, we would like to share our findings on loan applications for the SBA to ensure that you are better prepared if a loan is something you are considering as a business decision.

We obtained the dataset for this analysis through Kaggle, which can be found [here](https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied). Briefly summarizing the dataset, it came in the form of tabular data, which consisted of 27 columns and a little over 800 thousand rows, each row representing a loan application, with each column describing the details of each application. Most applications ranged between the early 1980’s to the early 2010’s and the locations of all applicants are located within the United States.

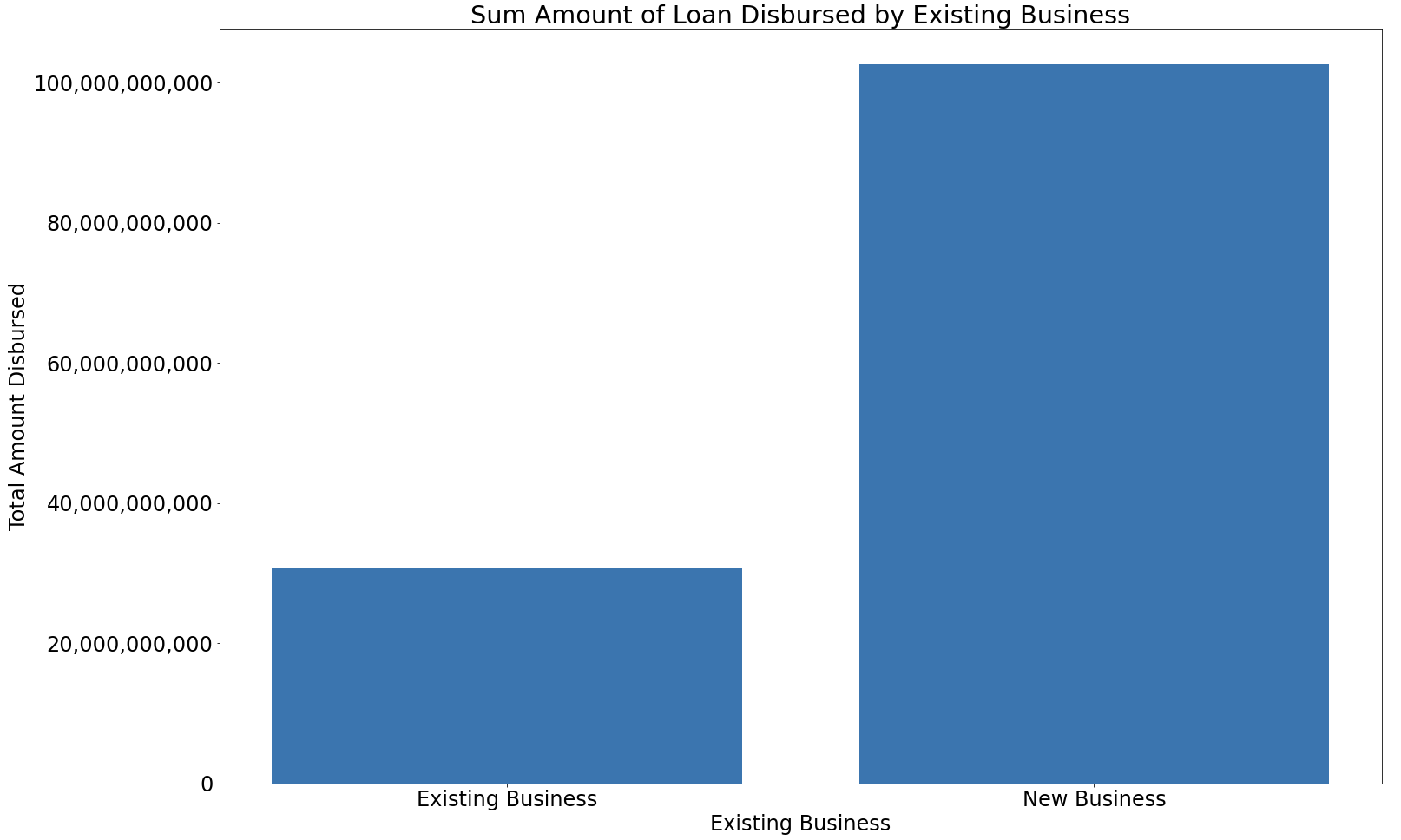
Before we begin the analysis, there are a few questions that need to be asked in order to guide us through the analysis journey. We wanted to know three things: “What factors affect how much the loan is for?”, “What factors tend to cause a loan to be defaulted on?”, and, “Can we build a model to predict whether a loan will be defaulted on or not?” With these three questions in mind, we can start taking a deeper look at this dataset.

Starting with the first question, “What factors affect how much the loan is for?”, we took a look at three different factors: industry, existing business, and franchise. For industry, we analyzed how much money a business would apply for a loan based on the industry they are in.



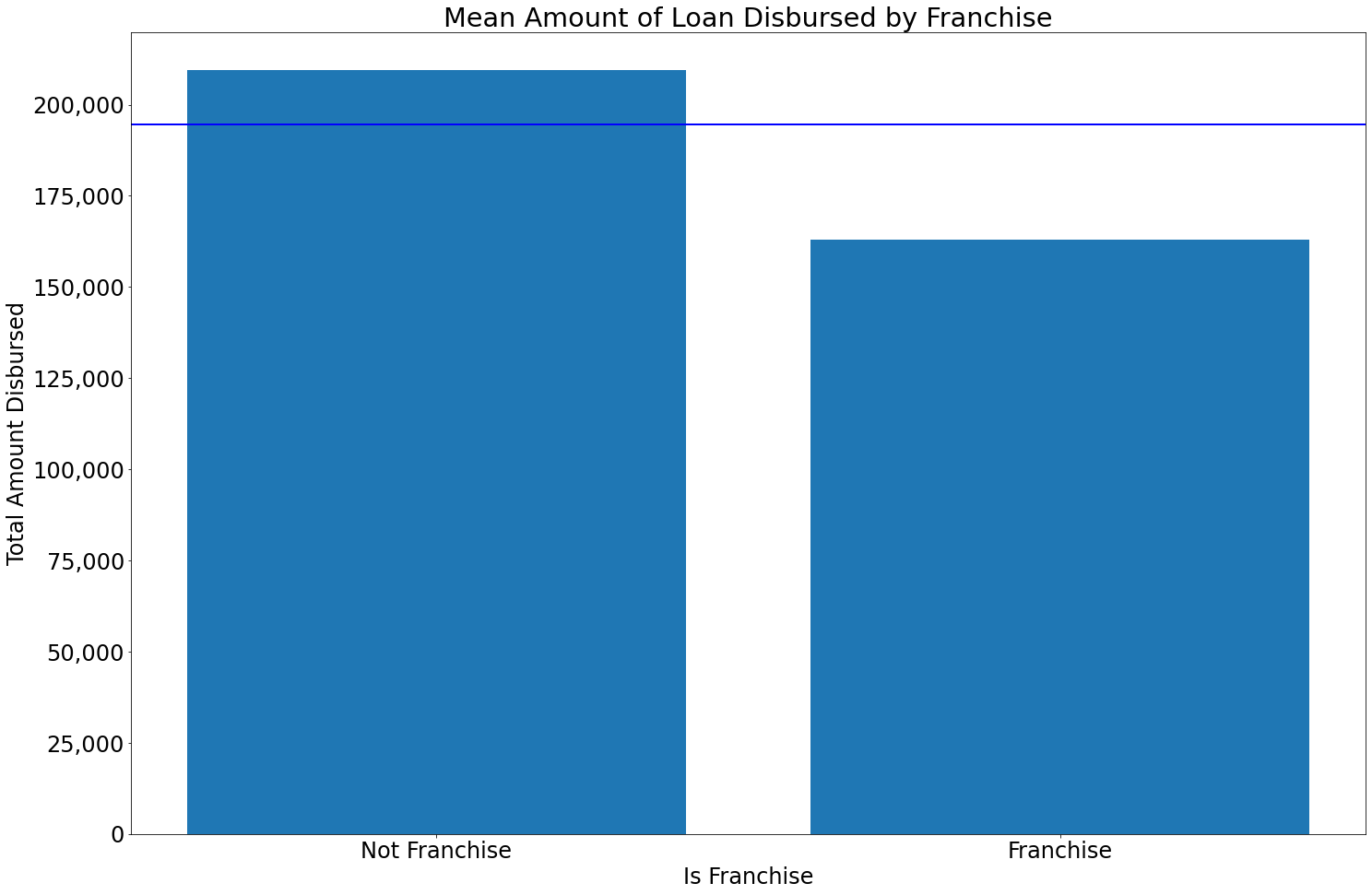
The figure above shows us that the businesses that take out the biggest loans are primarily in industries that handle natural resources, human resources, and other people’s money. The main piece of information that we get from this graph is that there is a significant difference between industries in terms of average loan amounts given to these businesses. This is just something to consider if you think that the industry you are in might have an impact on how much money other business owners in your industry are applying for.

Moving on to the next factor, for existing businesses, we looked at how much money businesses apply for on their loans based on if it was a new business, or an already existing business.



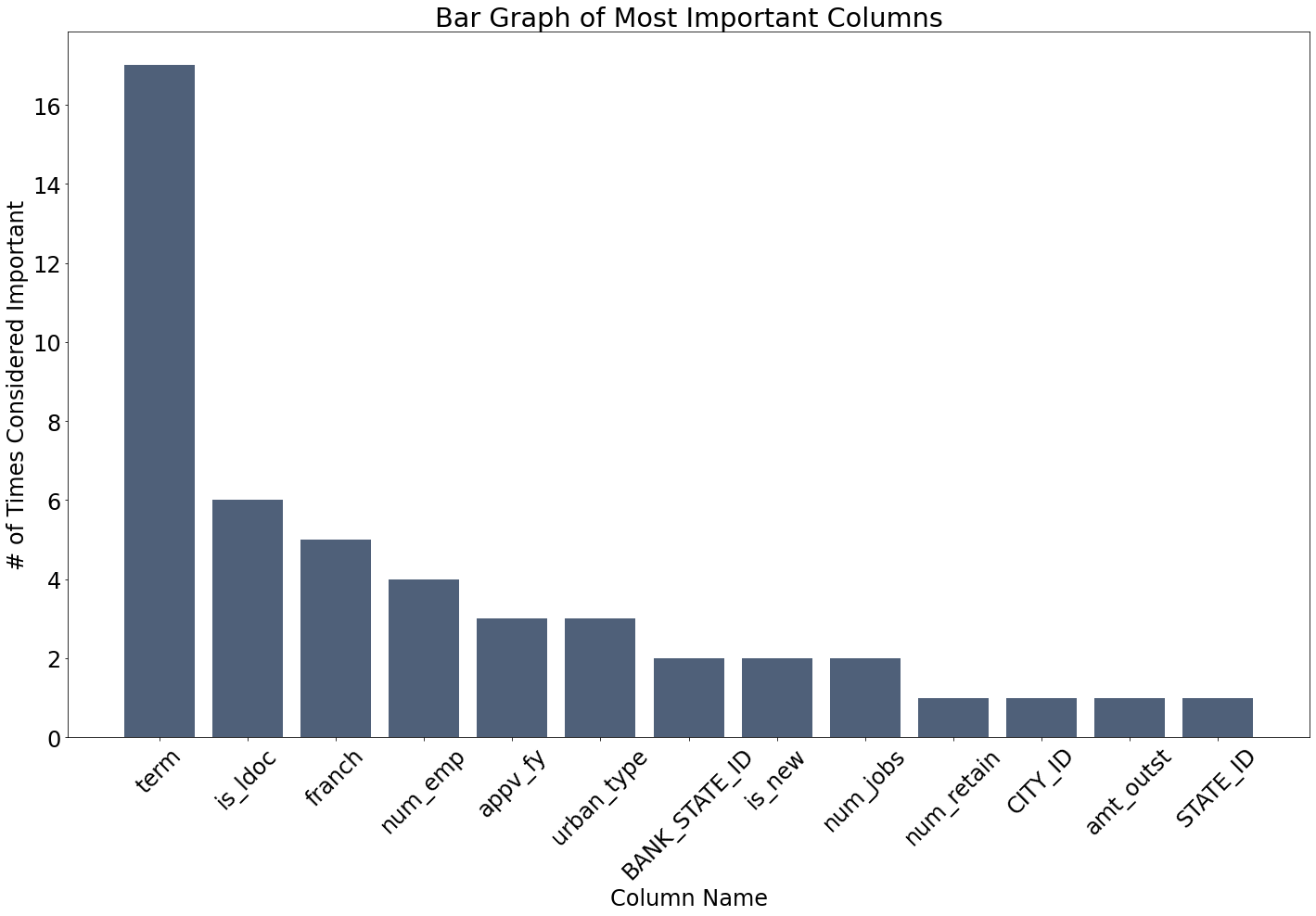
One important thing to mention about the figure just above is that, in our dataset, new businesses account for only 30% out of all loan applications, however, when their loan amounts are summed, new businesses account for more than triple the loan disbursement amount. The main takeaway for this graph is that, because of some external factors not included in the dataset, if you are applying for a loan as a new business owner, you may be looking at quite a hefty up-front investment in your startup.

Moving on to the final factor, we tried to determine if there is a trend on whether your business being a franchise or not affects the amount of money disbursed for a loan.



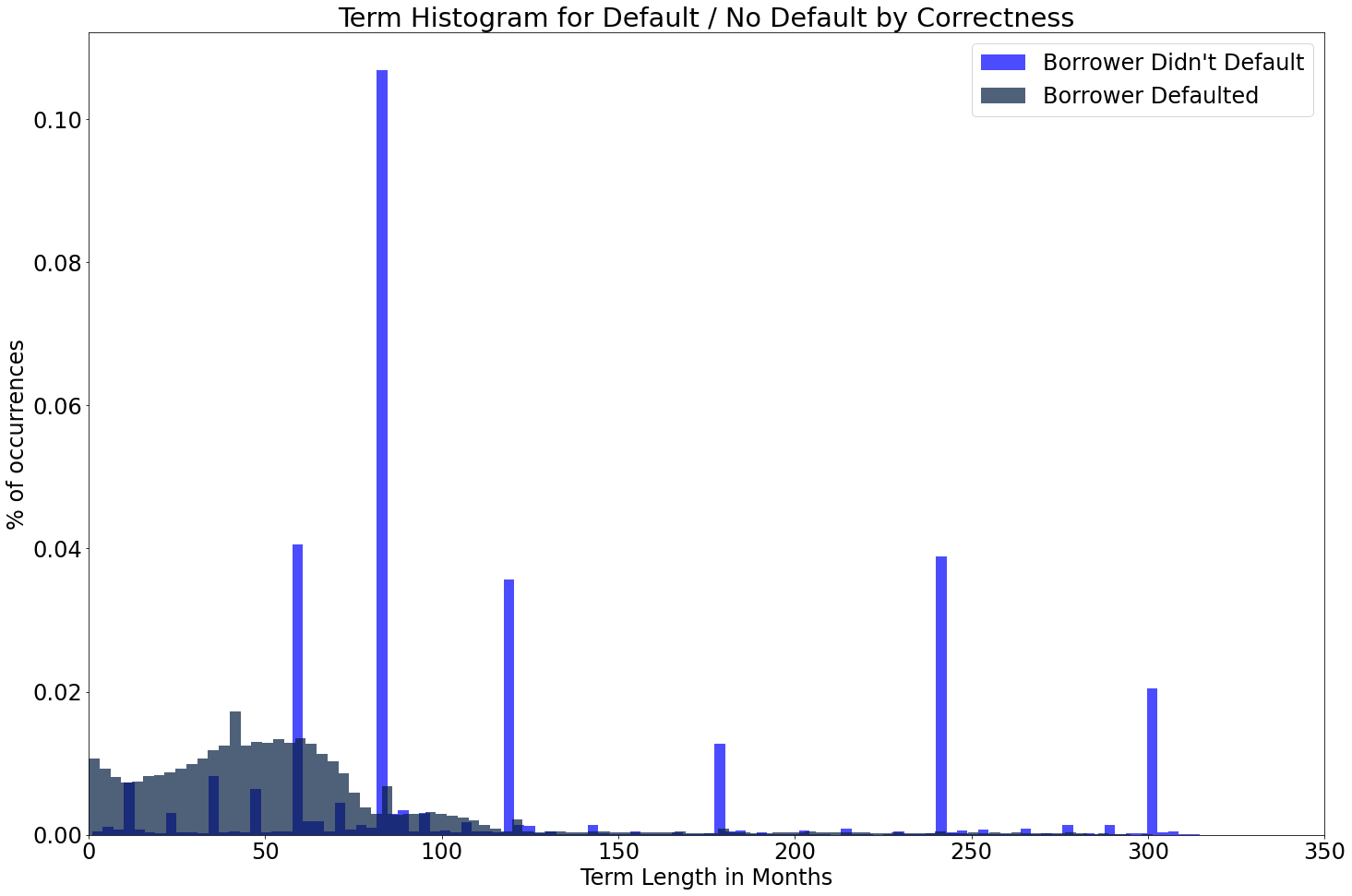
While the figure just above does not show as big of a difference compared to the previous two graphs, the difference in this figure is not trivial. Typically, franchises apply for smaller loans when compared to businesses that are not considered a franchise. What this graph tells us is that if you are starting your business as a franchise, you typically do not need as much money compared to non-franchises. Now that we have talked about factors that show how much money a loan should normally be for, let us move on to the next question.

For the second of our three questions, “What factors tend to cause a loan to be defaulted on?”, we came up with a method to determine which columns (factors) in our data set would be the most insightful in helping to determine if a loan would be defaulted on or not.



To explain how we got the figure directly above, we took all 27 columns in our data set and generated every single combination of those columns. We then fitted a logistic regression model to each combination of columns and recorded the score of the model. Score, in this context, means how well did that specific combination of columns do in predicting whether a borrower defaulted on their loan or not. For example, if a model gave us a score of 0.47, then that means that this specific combination can predict defaults with 47% accuracy. After going through every combination, we only looked at the combination which achieved the highest score.

We fitted every single column combination 17 times, and out of those 17 times, the column named, “term” appeared 17 times. This means that, using our method of determining important factors, we found that the loan term is the most crucial determinant of loan defaults. Since the loan term was interpreted as being such an important factor, we decided to take a closer look at that column.

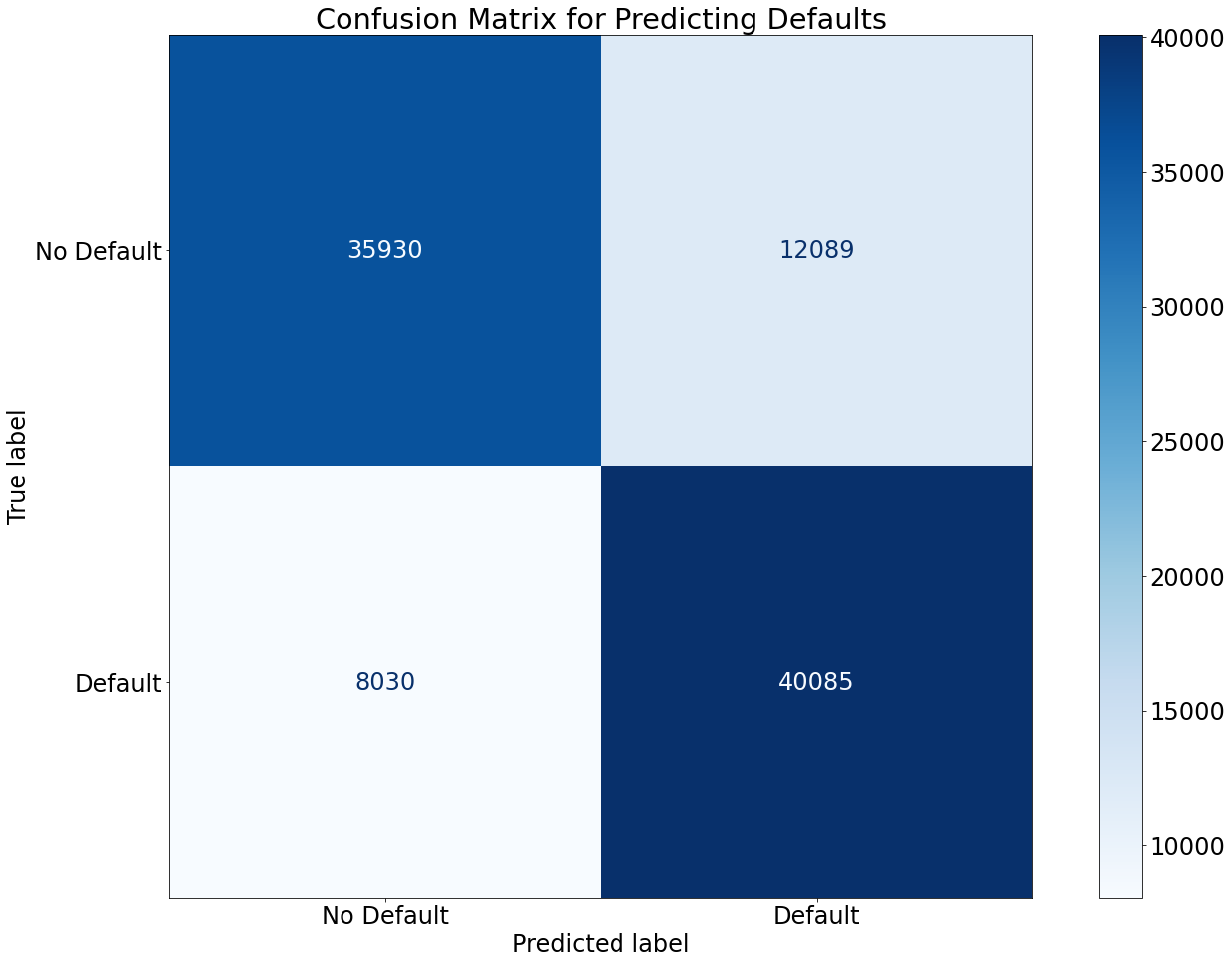


In the figure above, we plotted a histogram of the column, “term”. However, we also split the histogram up into two different sections: a histogram of the term lengths of borrowers who defaulted, and a histogram of the term lengths of borrowers who did not default. As shown in the above figure, there seems to be a very contrasting characteristic between the two categories. People who do not default almost always have term lengths of 60, 80, 120, 180, etc. months.

One very important thing to mention for this column, is that, when we looked at the documentation this dataset came with from Kaggle, it was very vague in describing the details of each column. What this means is that, for the column, “term”, it is quite intriguing that there are term lengths of 11 months or less, as it doesn’t make sense from a common sense standpoint that a bank would give out a loan and expect you to pay it back in under a year. This led us to believe that when borrowers defaulted on their loan, their actual term length was not entered into the dataset, but rather the month that they defaulted on. This means that, while term is a good column to try to determine defaults, it is virtually useless in real world scenarios, as we are not able to predict if loan terms affect defaults. But for the sake of the next question, we still used it to try to make a model, although its real world applications will be limited.

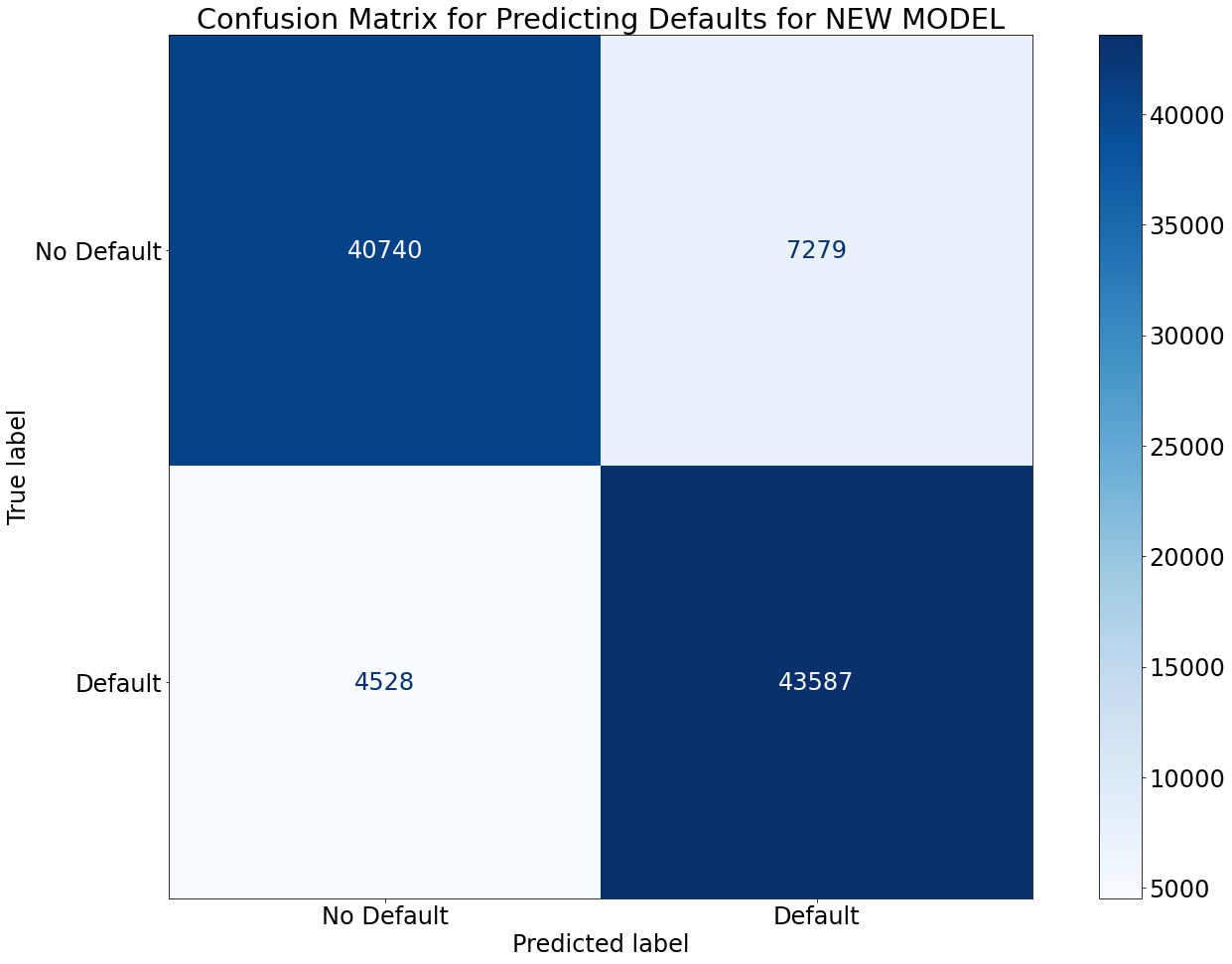
Building off the results from the previous question leads us to the final question, “Can we build a model to predict whether a loan will be defaulted on or not?” As stated in the findings of the previous question, the real world applications will be limited, as most of the columns in the data contain information that is only available after someone has defaulted.

Similar to what we did to determine the most important factors, we used a logistic regression model to fit the variable, “term”. The score for this model was, on average, 0.79, which means that it is 79% accurate.



Looking at the figure above, the logistic regression model does a decent job of “predicting” a default depending on loan term. This graph tells us that the model is better at predicting defaults rather than non-defaults. This makes sense, as back in the previous figure with the spikes, the regression model splits that histogram in half vertically and clumps one side of the graph as borrowers who defaulted and the other side as borrowers who did not default. In our testing, the model usually makes that split at a term length of 70 months.

Just for fun, we decided to squeeze as much accuracy as possible out of the term length variable, and made our own simple model that predicted if any borrower was in the spikes shown in the histogram above, then it would categorize them as being someone who didn’t default. When we did this, this is the confusion matrix it gave:



This simple model increased our average accuracy of “prediction” by around 9% on average from 79% with the logistic regression model to 88% with this new model. The main takeaway for this question is that there does not seem to be any information in the dataset that we can use to predict whether future borrowers will default on their loan or not. This model we made is purely to see how well we could get our score to be.

In conclusion, we found that, term, while flawed, gave the best model accuracy for predicting defaults. We also found that non-franchise businesses make higher than the average loan disbursement and new businesses usually apply for more hefty loans.