Capstone Milestone Report 1:

Problem statement: This report attempts to determine which characteristics of home loan applications are the driving factors into whether that application is accepted or not. This can be useful to organizations that take financial risk by giving out loans to prospective homeowners.

The dataset was publicly available and contained 500,000 records with training features and training labels(acceptance) as well as 500,000 records that had features but no labels, which was to serve as the test set. There were missing values for several features, which were dealt with in different ways. Depending on the importance of the feature that had missing values(as determined through visualization methods and feature analysis which will be discussed later), different strategies were used. For example, for strongly predictive features like applicant income, missing values were imputed using the median, while missing values for weakly predictive features like population were not dealt with at all because those features simply ended up being excluded from the model. Additional data wrangling was required in order to get the features in formats that lent themselves well for classification analysis. Log transformations were applied to the applicant income and loan amount features, which changed those features to having strongly skewed distributions to perfectly normal distributions. Also, as I got further into my analysis, I began to use feature engineering methods to further wrangle the dataset but that will be discussed in a later report.

The data exploration process consisted of several types of analysis and visualization. Correlation matrices between numerical features were created and analyzed to get an initial sense of which features were most strongly correlated both with each other and with the actual acceptance of the application. Then, these relationships were visualized using both python and PowerBI.

These visualizations can be found in the full version of my report that I am uploading to my GitHub account. I am also uploading powerpoint slides that highlight my findings.