Problem/Overview:

Financial institutions (e.g. banks) refuse a countless number of loan/credit applications to people in need each day, but they have a cause for doing so. These institutions put themselves at monetary risk by lending credit to others, especially with no guarantee of the money's return. As a result, financial institutions are meticulous and stringent on the qualifications required to request large loans. This means that they will only approve credit applications to clients that have a strong record for repayment, as well as proven financial means to repay the loans.

Although this strategy minimizes the risk of a loan default, it prevents the newer clients from receiving the loans they need. This is especially true since the only way to obtain good credit is to consistently spend smaller amounts of credit over a longer period. Those individuals uninformed about credit are unable to build their credit early, resulting in difficulty applying for larger loans at a later age. As a result, some of these individuals are forced to become patrons for unreliable lenders.

To address that, some financial institutions have taken the risk and strived to expand the financial inclusion to those with insufficient or non-existing credit histories. These institutions utilize a variety of alternate data to gauge and measure their clients' risk profile.

My objective for this project will be to experiment with a sample of credit data provided by Home Credit Group, an international financial institution focused on lending to the population with little to no credit history. I will analyze the data and generate a predictive model that can reliably classify each client as high-risk or low-risk for credit loans.

Potential Clients:

The potential clients for this project include two groups: financial institutions and loan applicants. All financial institutions can utilize this data to expand their loans applicant pool. If the model can reliably predict high-risk and low-risk loan applicants via alternate data, then the risk of defaults can be minimized. With a larger approved applicant pool, financial institutions will earn larger returns. Additionally, when a client has a positive loan experience, they are more likely to return as a reoccurring applicant. As a result, this would yield higher returns both short-term and long-term for the financial institutions.

Loan applicants, especially applicants with near non-existent credit history, can utilize this information to see and understand alternative attributes that can help them develop towards becoming a low-risk applicant despite short credit histories. However, this information is applicable across as larger spectrum, essentially to anyone who is applying for a loan they may not currently qualify for with a normal financial institution's requirements. This may help the applicant reliably acquire the money they need.

Data:

This dataset was provided by the Home Credit Group to a currently ongoing Kaggle Competition. The data consists of one training dataset supplemented with six other datasets of varying information. No exact time-frame is provided for, but there are records for up to 8 years of prior credit history relative to the time of application data. Overall, the model will be trained with over three-hundred thousand unique applicants, and any associate information. This data contains over a total of three hundred attributes, consisting of a mixture of both categorical and quantitative information.

Approaching the Problem:

The first focus will be to select the columns of interest within the pool of three attributes. This part of the analysis will be done inside an external excel file provided by the Kaggle Competition containing descriptions for each category in the data. To do this, I will first focus on removing columns that provide redundant information (e.g. if desired, I can prove them by checking for multicollinearity). Next, I will return to the updated categories and select specific categories that could make practical sense in relaying information (e.g. behavioral, geographical, etc.) of the clients' characters.

After I have narrowed down the categories to only ones of interest, I will begin feature engineering, to construct more informative columns and further reduce the number of columns. Once feature engineering is completed, I will move onto cleaning the data. This involves thoroughly checking the data for wrongly inputted data or any missing values. This pattern of feature engineering and data cleaning will be performed on one dataset at a time.

After all the datasets have been worked on, I will merge the datasets to the training data, making sure to address any null values that may appear at that time. The last part is to dive into the data and begin an in-depth visual and statistical analysis, along with addressing any new issues that may be discovered.

Deliverables:

With the project, I hope to deliver well-documented code with visuals and statistical analysis that allow me to construct a strong presentation for my results.