Capstone Project 1 Data Wrangling

GitHub project available at
https://github.com/mhardcastle0/Springboard/blob/master/lending_club_capstone/data_wrangling.ipynb

The goal of this project is to determine whether macroeconomic variables at the time that an unsecured personal loan is originated are useful predictors of loan outcomes. To determine this requires data on personal loan originations and default statuses, in addition to macroeconomic variables of interest.

The primary datasource is the Lending Club dataset, which is hosted as a flat file on Kaggle. Preparing the data source required little modification: each row represents a loan with an origination month, term length, delinquency status, and several other credit- and performance-related metrics. The data contains several fields that may be relevant to analysis, so all fields from the file were maintained. Loan amounts at origination, origination dates, and loan statuses were evaluated for unexpected values and were all found to be in-line with expectation. Additionally, no missing values that would impact analysis were observed.

For proxies of macroeconomic well-being, daily S&P 500 stock index values and monthly unemployment rates were sourced from the Federal Reserve Economic Data (FRED) API. As loan all other data sets are at the monthly level, the S&P 500 data was sampled to include only the prices at the beginning and end of each month, in addition to the average monthly value. This data contained several missing values, as stocks are not traded on some weekend and holiday days; for any day that was missing stock S&P 500 data, the data from the previous available day was used. The monthly unemployment rates sourced from FRED were usable without accommodation of outliers or missing values.

The bankruptcy data, sourced from an Excel file provided by Epiq Global, required the most modification. Several formatting issues were present, including merged cells containing the year portion of each period and summary and location-specific rows of data that needed to be dropped. The file required reading into a Pandas dataframe, transposing the data, removing superfluous columns, forward-filling the years to accommodate data missing due to merged cells, and merging the years and months into a single datetime column.

Finally, all of the data sources were merged together. All original Lending Club data was maintained, into which the unemployment rate, beginning- and end-of-month S&P 500 values, average monthly S&P 500 values, and monthly bankruptcy quantities were joined.