Create a short document (1-2 pages) in your github describing the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.

My dataset is a from Lending Club and it is data on loans that were serviced through the company. There are approximately 880000 rows and 75 columns. I chose this data because it was a greater than 100,000 rows and there were many features that could be studied. In order to prepare my dataset, I took the following steps to clean it and learn a little bit about it.

First, I looked at the list of the features and compared it to the data dictionary. I used “describe()” on the dataframe to determine how many values were available in each of the columns. There were 19 columns that had less 80% of values populated and this would have been very difficult to fill or interpolate or fill. I removed these columns from the clean data set. Looking at the values that were several columns that contained non-unique values. I removed all the columns that had less than 5 unique values.

I next browsed through the data dictionary and compared the column names for features that would indicate something about predicting a loan. Many features were redundant or containing information that did not add any value. I removed these columns. There were 18 remaining columns.

Several columns had text data that included slight variations in spelling or the categories could be simplified. I used “unique()” on each of these features, examined the number of distinct values, and then created a dictionary and then used “map()” to simplify the features. There was a feature named “emp_length and it contained objects with text for the length of a loan applicant’s employment. I removed the text so that all of these values became float values, and if the value did not exist a zero was inserted. Lending club groups all people who have been employed more than 10 years into one group.

Another feature that I simplified was the “home_ownership” feature. I simplified and mapped all categories. I grouped “mortgage” and “own” as one group. Rent as a distinct group. And all other categories into “other”. The “int_rate” feature was originally an object with a percent sign. I removed the percent sign and converted the value to a float.

The most important thing I did to clean the data was to examine the “loan_status” and determine which values would be considered a Good loan and which values were considered bad. This is important because it is what will determine the quality of the loan. The were 11 values, which I had to map to a “good” or bad status. Here is how they were mapped: *'Current': True, 'Fully Paid':True, 'Charged Off': False,*

'Default':False, 'Late (31-120 days)':False, 'In Grace Period':False, 'Late (16-30 days)':False, 'Does not meet the credit policy. Status:Fully Paid':True, 'Does not meet the credit policy. Status:Charged Off':False, None:False, 'Issued':True.

This data should be able to be analyzed and my goal is to find features that can accurately predict the quality of a loan.