Joseph Petersen - October 2018 Springboard Data Science Career Track Cohort

Capstone 1 Data Wrangling Report

In general, the dataset was very clean and little wrangling was warranted. The dataset consisted of blood donation data collected in Taiwan. Columns are “Recency (months)”, “Frequency (times)”, “Monetary (cc donated)”, “Time (months)”, and “whether he/she donated blood in March 2007”. The columns aren’t very descriptive, but an associated file had column explanations.

Questions:

1. The column names were renamed for clarity and ease of use (i.e. “Time (months)” to “months_since_first” and “Recency (months)” to “months_since_last”) and also reindexed into a more logical, readable structure. Initial exploratory analysis included examining summary statistics, the distribution of the response column, and both box plots and histograms of each column.
2. The dataset was clean with no null values.
3. Outliers were detected in the plots, but the determination was that they were actual data and could be useful.