Springboard Assignment: Create a Google Doc (1-2 pages) describing the data wrangling steps you took to clean the dataset. Include answers to these questions in your submission:

a. What kind of cleaning steps did you perform?
b. How did you deal with missing values, if any?
c. Were there outliers, and how did you handle them?

Prediction of hospital readmission rate for patients with an existing diagnosis based on factors measured at time of intial admission.¶

The dataset I will be using was originally constructed by researchers at Virginia Commonwealth University to see whether the decision to take a measurement of HbA1c (a test to measure glucose in diabetic patients) during hospitalization led to lower rate of hospital readmission. The HbA1c test is considered by the researchers to be a proxy for a more active management of the diabetes in the patient.

In this analysis, I will repurpose the data set to perform a correlational analysis/ predictive modeling to determine whether one or more variables are predictive of whether or not a patient will be readmitted to the hospital within 30 days of discharge. Time allowing, the analysis may be extended to see whether prediction of readmission after 30 days can also be predicted.

Please see the Jupyter notebook located at: https://github.com/janineb/hospital-readmission-prediction/blob/master/diabetes_dataset_cleaning.ipynb for all the details on the actual data cleaning steps used.

High level Overview of the Data Cleaning Process

The dataset, 'Diabetes 130-US hospitals for years 1999-2008 Data Set' is an open dataset was obtainable at:
:https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#
After initial overview of the data, missing values were found in several of the columns. A table was constructed showing, for each feature, the % missing values, the # of unique values for each , the first value, and a list of the unique values.

The 'weight' column was dropped because it was missing 97% of the values.

Because there is potential value to a predictive model of the category, ‘none’, or ‘unknown’ when a value is either not given or was not known by the researcher, the decision was made to retain these as a category.

Two columns were missing about half of their values: payer code and medical specialty. No obvious correlation between the columns was found (for example, no medical specialty stood out as overwhelmingly missing). While the most common value of each column could be used as a fill, it was decided to replace the missing value with ‘none’ and experiment with different fill methods at the time of predictive modeling to see whether this has a positive, negative, or no effect.

Target Column preparation

The target column is 'readmitted'. To answer the original question, I transformed the column to binary (0 for 'No' and '>30', 1 for '<30').

Less than 30 days ('<30') is the standard for assessing what constitutes ‘hospital readmission’: CMS defines a hospital readmission as "an admission to an acute care hospital within 30 days of discharge from the same or another acute care hospital."Readmissions-Reduction-Program". www.cms.gov. 2016-02-04. Retrieved 2016-03-01.Also: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/2015-ACR-MIF.pdf

This is the criteria that makes sense for the initial mock business case discussed in the proposal.

However, I might want to extend the analysis for practice. A Kaggle competition which used the same data set (a closed competition, so no kernels posted) suggested that the expected output should include '>30' as a separate category (see https://www.kaggle.com/c/diabetes-hospital-readmission#evaluation). I will create a second column to include this as a separate category then decide whether to do multiclass analysis later on if time allows. This will contain an additional designation of 2 for ‘>30’ which will no longer be grouped with 0.

Outliers

I did not identify any outliers in the data set.