Data Wrangling

The data for this project comes from the Heart Disease Data Set from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). The data is split into 4 parts, with each file corresponding to patient information from a specific locale (Cleveland, Hungary, Switzerland, and Long Beach). The dataset consists of 14 columns, described below:

- age
- sex (0 = female, 1 = male)
- cp (chest pain type; 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
- trestbps (resting blood pressure in mm Hg)
- chol (serum cholesterol in mg/dl)
- fbs (fasting blood sugar > 120 mg/dl; 0 = false, 1 = true)
- restecg (resting electrocardiographic results; 0 = normal, 1 = ST-T wave abnormality, 2 = probable/definite left ventricular hypertrophy)
- thalach (maximum heart rate)
- exang (exercise induced angina; 0 = false, 1 = true)
- oldpeak (ST depression induced by exercise relative to rest)
- slope (slope of peak exercise ST segment; 1 = upsloping, 2 = flat, 3 = downsloping)
- ca (number of major vessels colored by fluoroscopy; 0-3)
- thal (MPI defects; 3 = normal, 6 = fixed defect, 7 = reversible defect)
- num (heart disease diagnosis; 0 = absence, 1-3 = presence)

The data is in csv format with null values represented by '?' strings. I imported the data for each of the 4 locales directly from the repository using pandas, making sure to specify na_values='?'. I also specified the column names, as they were missing from the files in the repository.

Upon inspecting the data, I noticed that certain columns in some datasets contained a significant amount of missing values. After calculating the percentage of missing values in each column for each dataset, I chose to drop columns that consisted of more than 40% missing values. These columns corresponded to the slope, ca, and thal columns of the Hungary dataset, the fbs, ca, and thal columns of the Switzerland dataset, and the slope, ca, and thal columns of the Long Beach dataset. The Cleveland dataset was mostly complete, with no column containing more than 1.3% missing values. For the categorical columns (sex, cp, fbs, restecg, exang, slope, ca, thal), I assigned missing values their own category with a value of -1. For numeric columns (age, trestbps, chol, thalach, oldpeak), I filled missing values with the column means from their respective dataset. If the data proves skewed upon further analysis, I may choose to return and fill these missing values with column medians instead.

For this project, I am primarily interested in detecting the presence or absence of heart disease and not the severity, so I converted the num column to be values of 0 or 1 instead of 0 through 4. After applying this transformation, certain datasets revealed themselves to be very skewed, with the Switzerland dataset, for example, consisting of 93% patients with heart

disease. Building models for individual datasets with such skewed data would likely not work well, so I decided to focus on models for the Cleveland dataset and a combined analysis of all the locales. These datasets exhibit a 46/54 and 55/45 split of patients without heart disease and with heart disease, respectively. For the combined analysis, I performed an inner join on all the datasets, keeping only the columns I had not thrown away in any one of them. These columns were the age, sex, cp, trestbps, chol, restecg, thalach, exang, oldpeak, and num columns.

When inspecting the summary statistics of the combined dataset, I noticed something strange: The minimum values for the trestbps and chol columns were 0. This made no sense, since those columns represented patient blood pressure and cholesterol levels. Examining further revealed that the Switzerland dataset in particular consisted almost entirely of 0 entries for chol. Presumably, these entries were missing values that were filled in as 0 in the files. To correct this, I replaced the 0 values in those columns with the column mean without taking the 0 values into account.

Visualizing the data for the combined dataset revealed the presence of a considerable number of outliers, particularly for the trestbps and chol columns. Inspecting these values revealed them to be within reason, as I had found reports of more extreme values online. For now, I decided to keep the data as it was, but if the outliers prove to be a problem later on, I may choose to apply a log transformation to those columns. The notebook that I used to perform this data wrangling can be found as data_wrangling.ipynb, and the cleaned datasets were saved as csv files into this repository.