Data Wrangling

Loading and filtering data

The first step taken after loading the olympic data set was to merge it with the NOC data set, which identifies the country, region, and potentially exact location of the olympic athletes. The data set was then filtered to only include athletes from the summer olympics, as the data for the winter olympics was incomplete. This filtering left approximately 222,000 observations for olympic athletes.

Missing Values

However, many of the observations were incomplete for non-medaling athletes from the 1920 olympics and before. As a result, the incomplete observations were dropped from the data set, leaving approximately 167,000 observations.

Additionally, there are many missing values from the 'Medal' column as the majority of competing olympic athletes do not receive a medal for their event. This column was left alone as the data is needed to better predict body type for olympic athletes.

Outliers

Looking at the 167,000 observations in the data set, there are some obvious outliers. Within all competing olympic athletes, there are athletes that have competed at age 10 (in 1896, winning a bronze medal in parallel bars) and there are athletes that have competed in their 90s (in equestrian sports). There are also olympians that have medalled in their 50s and 60s. From weight and height, there are olympians who have been very short comparatively (measuring in at 127cm) and some who towered over their peers (at 226 cm), while the majority of athletes measured near the median height of 175 cm. Similarly, weight also had several outliers, from athletes who weighed in at 25 kg to athletes who weighed in at 214 kg (median is 70 kg). As each of these outliers represents a valid data point, all outliers were retained in the data set.

Additional data wrangling

One of the problems with the current data set is the inclusion of multiple categorical variables. To better analyze these categorical variables, the Sport column, Sex column (male/female), and Medal column were expanded into binary columns using one hot encoding. This final, expanded data set was used for all further analysis.