For my part of the project, I worked on cleaning up the dataset. The horse races have several important features that are characters, primarily the name of the horse, jockey, and trainer. Each is important because each person/animal will have an associated race history. For example, W M Lai, might be particularly good as a jockey and therefore have better overall outcomes with any given horse. Therefore, in order to make this qualitative data meaningful, we have feature engineering new corresponding columns to indicate the average percentile rank as performance indicators.

This was done so, by chopping up each piece of the data into individual races. Then the ranks were converted into percentiles by way of rank/total number of horses competing in the individual race. Note that horses that were unable to complete or disqualified for some reason were given a ranking equivalent to last place, meaning their percentile rank is 100. By looping through all of the data and checking for the corresponding names, we are able to assign a metric that shows us how well each jockey, trainer, or horse performs in general, throughout all races.