Springboard Capstone Project: Data Wrangling

Vikash Patel

My data wrangling experience with my specific project data sets certainly helped provide some practical experience dealing with missing data. The data sets were already properly structured and labeled, with the training sales data being very basic in its recording, however the weather data included large amounts of missing temperature and precipitation data that needed to be resolved.

The weather data that I am working with is critical for forecasting sales, so it was important that I dealt with the missing data in the most thorough manner. I initially extracted only relevant columns of interest from the weather data, like dates, station identifiers, average daily temperate, and daily precipitation total. I believe this to be enough information to incorporate into the model that will be forecasting sales. At first, I attempted to forward fill missing temperature values after sorting values by station and date, so that the previous day’s temperature and/or precipitation would carry over. However, this became an issue when some stations did not have any initial temperatures recorded for the beginning of the global date period (January 1st, 2012) for all stations. Upon exploring the data further, I found that there was one station that only had temperature and precipitation data recorded for the month of September 2013. I decided to drop any records that did not have actual temperature or precipitation data. If I were to have used a method to fill in missing values, being either forward filling, backward filling, or even taking average of surrounding values, my weather data would have been compromised by inaccurate data, especially in the case for the station mentioned earlier with only 1 month worth of recorded data. Dropping any records with missing temperatures and/or precipitation totals will also prevent anomalies from being falsely ignored.

Wrangling through and cleaning the weather data provided a practical lesson in dealing with missing values, and an understanding of how certain methods used to fill missing values can impact the resulting data set statistics. From this clean weather data, I plan on identifying the types of weather each station is accustom to, so that abnormal or severe weather is easily identifiable for each station.