Data Wrangling

1. The two main steps for data cleaning were:
    a. Handling missing values or NaN values
    b. Detecting and removing outliers in data

2. As always, there were missing values in the data. A list was created containing the columns with missing values to see which columns had missing values in them. Since only 3 columns had NaN values and none of them were important for prediction/analysis purposes, those 3 columns were dropped from the dataframe.

3. There were outliers in the data for some of the numeric columns. Outliers were detected using two methods – box plots and z-score:
    a. Box plots clearly show the data points and how many lie too far away from all the other points.
    b. Z-score from the “scipy” package and “stats” module is a convenient of detecting outliers in data. All observations having a z-score less than 3 were removed from the dataframe (3 is a convenient number)