Data wrangling for Capstone project 1

1. What kind of cleaning steps did you perform?

   Step1. Understand the data: column types, data frame shape, correlation between the variables.

   Step2. Remove outliers

   Step3. Imputing missing values

2. How did you deal with missing values, if any?

I have combined the train and test set for this step and I looked at categorical variables and numeric variables separately.

Methods of imputation for categorical variables:

1. If number of missing values is less than 5, use mode to replace missing values
2. If number of missing values is more than 5, use "None" to replace missing values

Methods of imputation for numeric variables:

1. If number of missing values is less than 25, use "mean" to replace missing values
2. If number of missing values is more than 5, use "median" to replace missing values

3. Were there outliers, and how did you handle them?

I dealt with the outliers by mostly reading the graphs, I deleted the obvious outliers from the data.