Solution: Clean the Data

This lesson gives a detailed review of the solution to the challenge from the previous lesson.

We'll cover the following

Solution

```
def clean_data(df):

    df = df.dropna() # dropping all rows with null values

    # A list of all columns on which outliers need to be removed
    out_list = ['median_house_value', 'median_income', 'housing_median_age']

    quantiles_df = (df.quantile([0.25,0.75])) # computing 1st & 3rd quartiles

    for out in out_list: # traversing through the list

        Q1 = quantiles_df[out][0.25] # Retrieving value of 1st quartile
        Q3 = quantiles_df[out][0.75] # Retrieving value of 3rd quartile

        iqr = Q3 - Q1 # computing the interquartile range

        lower_bound = (Q1 - (iqr * 1.5)) # computing lower bound
        upper_bound = (Q3 + (iqr * 1.5)) # computing upper bound

        col = df[out] # Storing reference of required column

        col[(col < lower_bound)] = lower_bound # Assign outliers to lower bound

        col[(col > upper_bound)] = upper_bound # Assign outliers to upper bound

    return df

# Test Code

df = pd.read_csv('housing.csv')

df_res = clean_data(df.copy())

print(df_res)
```

Explanation

A function `clean_data` is declared with `df` passed to it as a parameter.

On **line 3**, the `dropna()` function of the `DataFrame`, which automatically finds and removes all NaN containing rows, is used.

On **line 6**, a `list` that contains all the columns of the dataset from which outliers need to be removed is declared.

On **line 8**, the `quantile` function of the `DataFrame` is used to find the **first** and **third** quartile to help us compute the lower and upper bound for outliers.

On **line 10**, a `for` loop is used to traverse through the list. On each iteration, the columns in the list get processed for outliers.

On **lines 12 & 13**, the **1st** and **3rd** quartile values are retrieved for the required column.

On **line 15**, the *interquartile range* is calculated from the quartile values.

On **lines 17 & 18**, the lower and upper bound values are calculated using the ***IQR*** value calculated above.

On **line 20**, the reference for the required current column is stored in a variable for the removal of identified outliers.

On **line 22**, those values of the current column, which are below or less than the lower bound value, are assigned that same lower bound value to get them in the required range.

On **line 24**, those values of the current column, which are above or greater than the upper bound value, are assigned that same upper bound value to get them in the required range.

A quiz awaits you in the next lesson.