Submission of assignment for:-

<https://www.springboard.com/workshops/data-science-career-track/learn#/curriculum/451>

**Project: Capstone Project 1 - Data Wrangling**

Honestly, the data that I picked up was more or less clean and in-shape to perform further analysis.

I would like to share some useful methods that I would have used in case I needed to perform data wrangling on messy data.

1. DATA CLEANING

df.columns = ['a','b','c'] - Renames columns

pd.isnull() - Checks for null Values, Returns Boolean Array

pd.notnull() - Opposite of s.isnull()

df.dropna() - Drops all rows that contain null values

df.dropna(axis=1) - Drops all columns that contain null values

df.fillna(x) - Replaces all null values with x

df.fillna(s.mean()) - Replaces all null values with the mean (mean can be replaced with almost any function from the statistics section)

df.replace(1,'one') - Replaces all values equal to 1 with 'one'

1. JOIN/COMBINE

df1.append(df2) - Adds the rows in df1 to the end of df2 (columns should be identical) pd.concat([df1, df2],axis=1) - Adds the columns in df1 to the end of df2 (rows should be identical)

1. SELECTION

df[col] - Returns column with label col as Series

df[[col1, col2]] - Returns Columns as a new DataFrame df.iloc[0] - Selection by position

df.loc[0] - Selection by index

df.iloc[0,:] - First row

df.iloc[0,0] - First element of first column

1. STATISTICS

These can all be applied to a series as well.

df.describe() - Summary statistics for numerical columns

df.mean() - Returns the mean of all columns

df.corr() - Returns the correlation between columns in a DataFrame

df.count() - Returns the number of non-null values in each DataFrame column

df.max() - Returns the highest value in each column

df.min() - Returns the lowest value in each column

df.median() - Returns the median of each column

df.std() - Returns the standard deviation of each column