Course Code: CA LAB-XI (B) LAB on AI Practice using Python

**Ass 2 Data Preparation using techniques like Data Cleansing**

**1. Importing Libraries**

Let’s get Pandas and NumPy up and running on your Python script.

**INPUT:**

*import pandas as pd*

*import numpy as np*

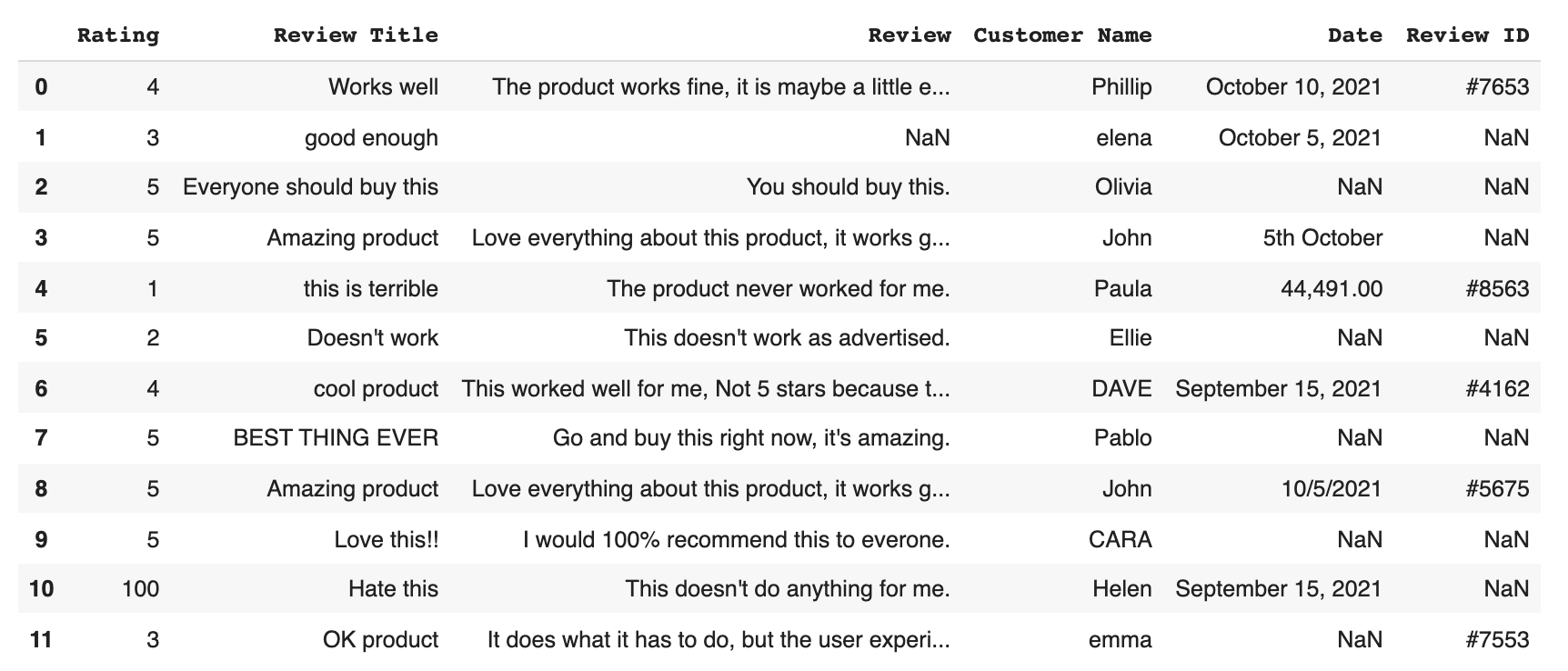
**2. Input Customer Feedback Dataset**

**INPUT:**

*data = pd.read\_csv('feedback.csv')*

*print(data)*

**OUTPUT:**



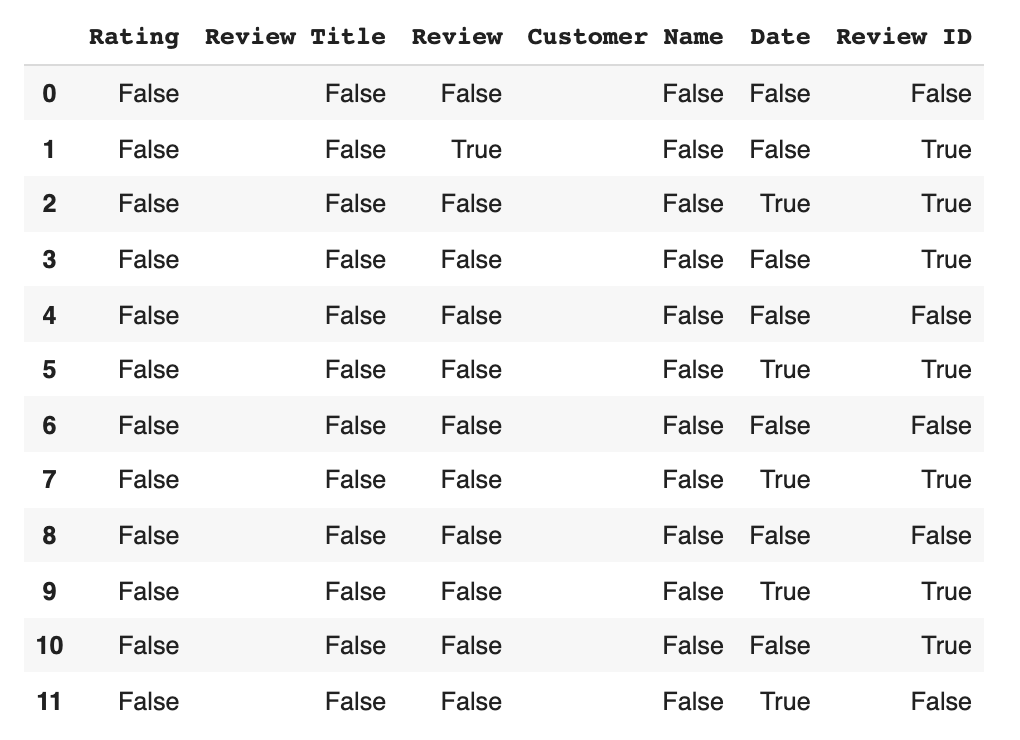
**3. Locate Missing Data**

Next, we are going to use a secret Python hack known as ‘isnull function’ to discover our data. Actually a common function, 'isnull' helps us find where in our dataset there are missing values. This is useful information as this is what we need to correct while data cleaning.

**INPUT:**

*print(data.isnull()*

**OUTPUT:**

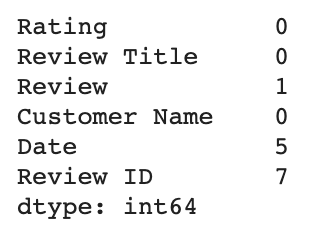


Our output result is a list of **boolean values**.

**INPUT:**

*print(data.isnull().sum())*

**OUTPUT:**



**We use code to actually clean the data.** This boils down to two basic options.

**a) Drop the data** :-

**b) Input missing data**. :-

**a). Drop the data :-**

You’ll have to make another decision – whether to drop only the missing values and keep the data in the set, or to eliminate the feature (the entire column) wholesale because there are so many missing datapoints that it isn’t fit for analysis.

**INPUT:**

*remove = ['Review ID','Date']*

*print(data.drop(remove, inplace =True, axis =1))*

**OUTPUT:**



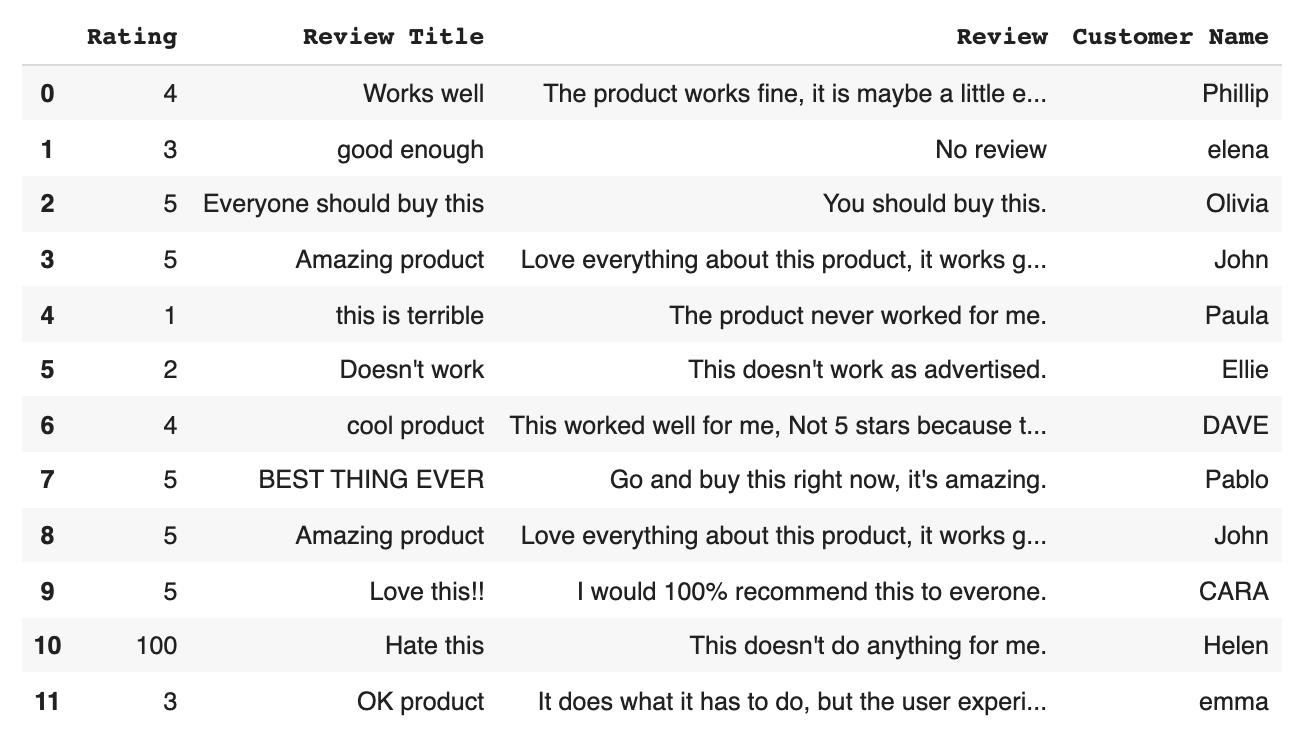
**b). Input missing data :-**

Technically, the method described above of filling in individual values with Pandas or NumBy standards is also a form of inputting missing data – we call it adding ‘No Review’. When it comes to inputting missing data you can either add ‘No Review’ using the code below, or manually fill in the correct data.

**INPUT:**

*print(data['Review'] = data['Review'].fillna('No review'))*

**OUTPUT:**



**4. Check for Duplicates :-**

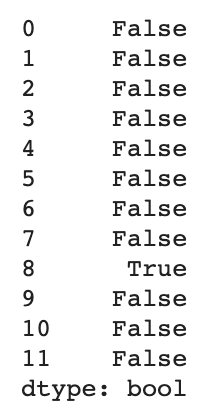
Duplicates, like missing data, cause problems and clog up analytics software. Let’s locate and eliminate them.

To locate duplicates we start out with:

**INPUT:**

*print(data.duplicated())*

**OUTPUT:**



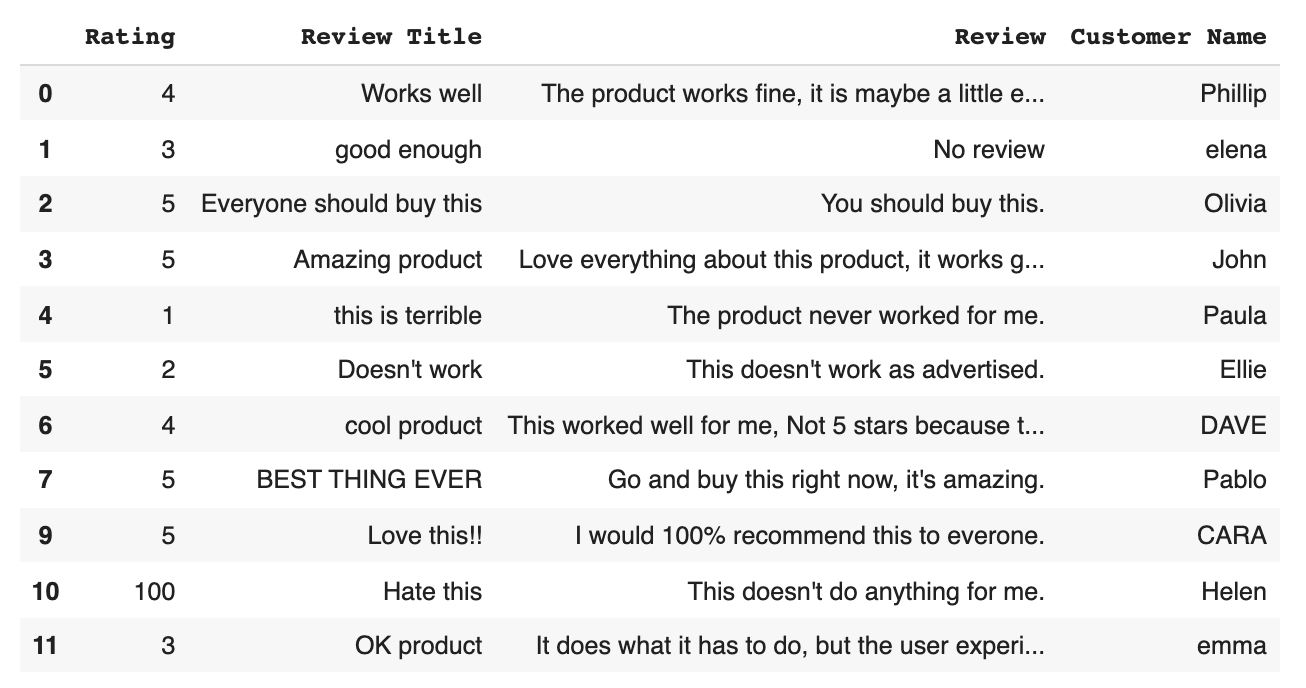
Aka a list of boolean values where a ‘True’ reading indicated duplicate values.

**To drop duplicate value :**

**INPUT:**

*print(data.drop\_duplicates())*

**OUTPUT:**



And there we have it, our dataset with our duplicate removed. Onwards.

**5. Detect Outliers :-**

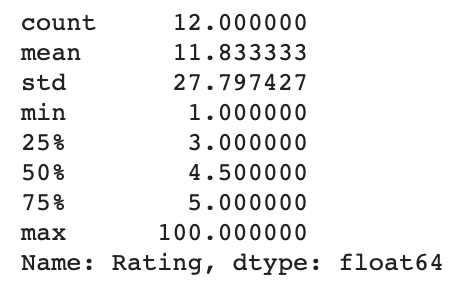
Outliers are numerical values that lie significantly outside of the statistical norm. Cutting that down from unnecessary science garble – they are data points that are so out of range they are likely misreads.

They, like duplicates, need to be removed. Let’s sniff out an outlier by first, pulling up our dataset.

**INPUT:**

*Print(data['Rating'].describe())*

**OUTPUT:**

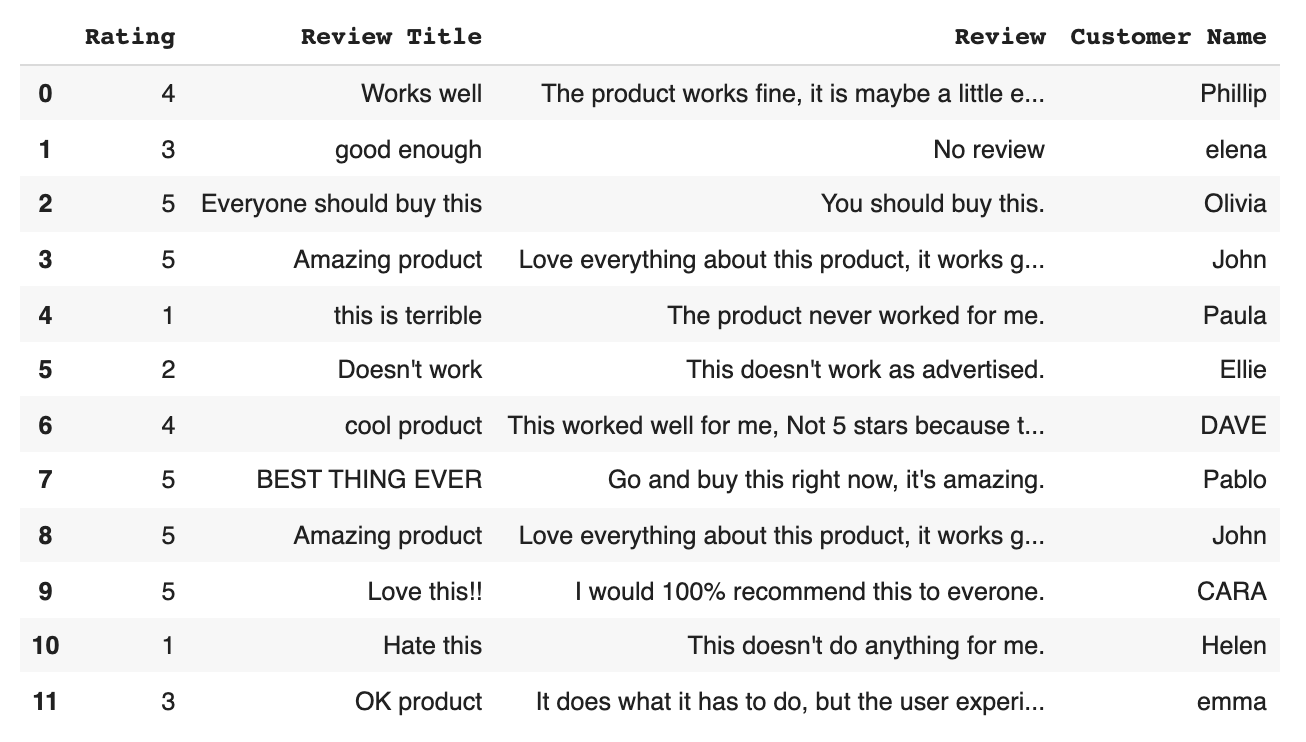


Take a look at that ‘max’ value - none of the other values are even close to 100, with the mean (the average) being 11. Now, your solution to outliers will depend on your knowledge of your dataset. In this case, the data scientists who input the knowledge know that they meant to put a value of 1 not 100. So, we can safely remove the outlier to fix our data.

**INPUT:**

*Print(data.loc[10,'Rating'] = 1)*

**OUTPUT:**



Now our dataset has ratings ranging from 1 to 5, which will save major skew from if there was a rogue 100 in there.

**6. Normalize Casinga:**

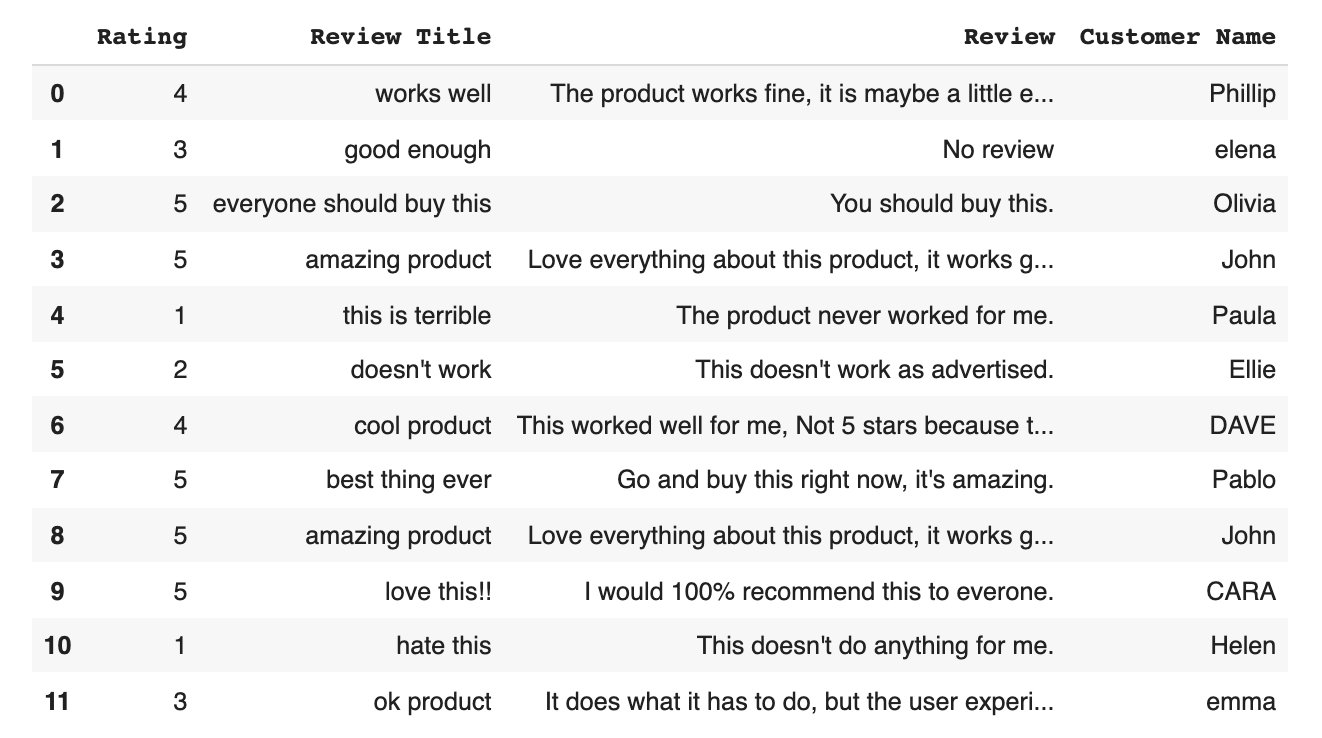
Last but not least we are going to dot our i’s and cross our t’s. Meaning we are going to standardize (lowercase) all review titles so as not to confuse our algorithms, and we are going to capitalize Customer Names, so that our algorithms know they are variables (you’ll see this in action below).

Here’s how to make every review title lowercase:

**INPUT:**

*print(data['Review Title'] = data['Review Title'].str.lower())*

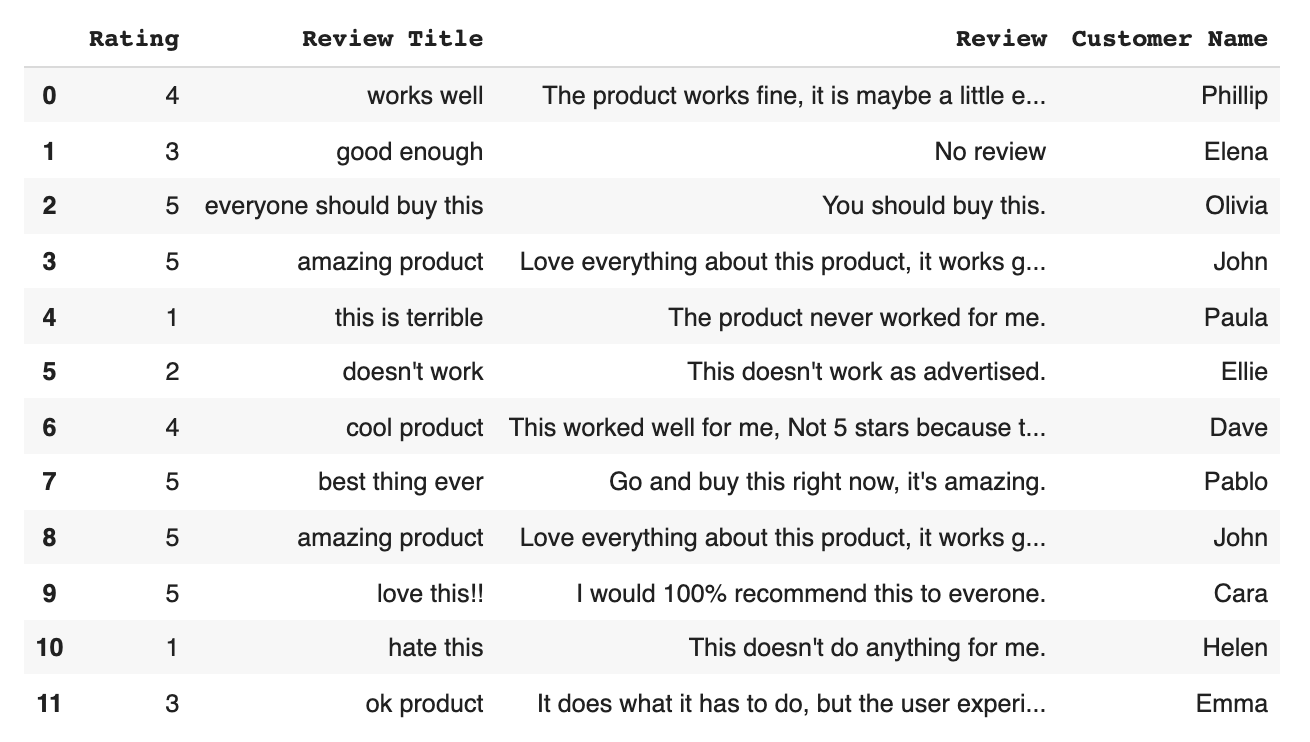
**OUTPUT:**



**INPUT:**

*print(data['Customer Name'] = data['Customer Name'].str.title())*

**OUTPUT:**



We’ve made good use of intuitive Python libraries to locate and eliminate bad data, and standardize the rest. We are now ready to make the most of it with our [machine learning data analysis software](https://monkeylearn.com/).