Option 2: How Python Represents Missing Data

Inevitably, data preparation is a very tedious task. Data “in the wild” is almost never ready to be analyzed right off the bat. This is due to several problems such as outlier presence, incorrect units, or missing data. Missing data can pose a problem for some programming languages; I will explain how it is represented in Python and some of the drawbacks of missing data.

As we know, Python has libraries such as NumPy and Pandas for manipulating data. If data is imported into Python as a data frame with some cells missing, however, Python will automatically label the missing value “NaN” for Not a Number. NaN values are a placeholder for missing values to prevent the script from throwing an error when importing and manipulating data. For example, if we have a CSV file with one column: [1, 2, blank, 2, 1], then the blank will be represented as NaN when it is imported into Python.

Generally, it’s not good to leave NaN values as they are due to many reasons:

1. NaN values are gaps in data, and too many means there could be some underlying data trends that get lost.
2. NaN values will always be skipped during arithmetic operations such as *sum()*, *count(),* etc. This reduces the statistical power of the data due to there being less samples to analyze.

We can choose to either remove the rows with NaN values completely, or we can replace them with something else. It is up to the data scientist to select what action to perform on NaN data.

References:

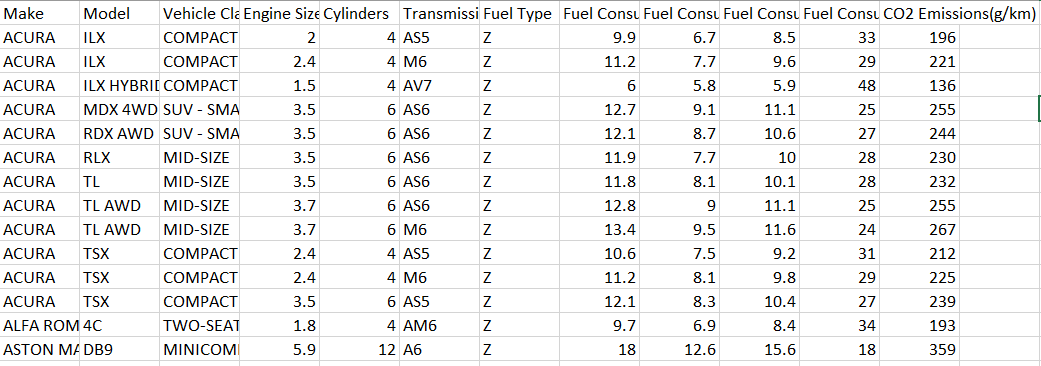
Brownlee, J. (2017). *How to Handle Missing Data with Python.* Machine Learning Mastery. Retrieved on January 26, 2021 from https://machinelearningmastery.com/handle-missing-data-python/#:~:text=In%20Python%2C%20specifically%20Pandas%2C%20NumPy,columns%20we%20are%20interested%20in.

Sullivan, J. (2018). *Data Cleaning with Python and Pandas: Detecting Missing Values.* Towards Data Science. Retrieved on January 26, 2021 from https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b.

Option 10: The Split-Apply-Combine Method

The Split-Apply-Combine method (SAC) is a useful data manipulation technique used in a variety of settings when group operations are required. As the name suggests, it allows for a function/operation to be applied to multiple groups in a data frame by first *splitting* the data into the desired groups using a key, then *applying* the desired function, then *combining* the new data back into a data frame. The key used to split the data could be a list, a column label, or even a dictionary mapping column names and values to be grouped. The SAC method is ultimately performed using the *groupby()* function.

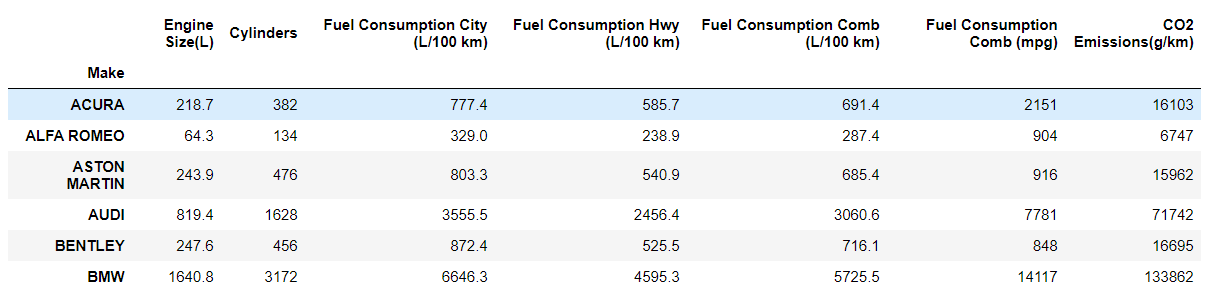
For example, let’s look at the vehicle dataset I am using for my term project. It is a data frame containing vehicle attributes that looks like this:



Let’s say I want to find the total of all numerical attributes from all cars in a make. First, I would split the data into group representing each car make, then I would apply the *sum()* function to each group before recombining the groups back into one data frame. The following line of code should accomplish all of this:

cars\_df.groupby(cars[‘Make’]).sum()

The *groupby()* function does our split based on our key (the make), then the *sum()* function is applied to each group. The returned data frame shows one line for each make and the aggregated attributes next to them, as seen below.



What if I only want to aggregate only the CO2 emissions column, though? For this, I will apply *groupby()* to the Pandas series rather than the entire data frame:

*cars[‘CO2 Emissions(g/km)]*.groupby(cars[‘Make’]).sum()

This will return the same results as the rightmost column above, but *only* that column.

The SAC method can come in handy for a variety of operations as well, not just sums.

References:

Jain, S. (2020). *Pandas dataframe.groupby().* GeeksforGeeks. Retrieved on January 26, 2021 from https://www.geeksforgeeks.org/python-pandas-dataframe-groupby/.

McKinney, W. (2017). *Python for Data Analysis.* O’Reilly Media.