Data Types & Data Wrangling

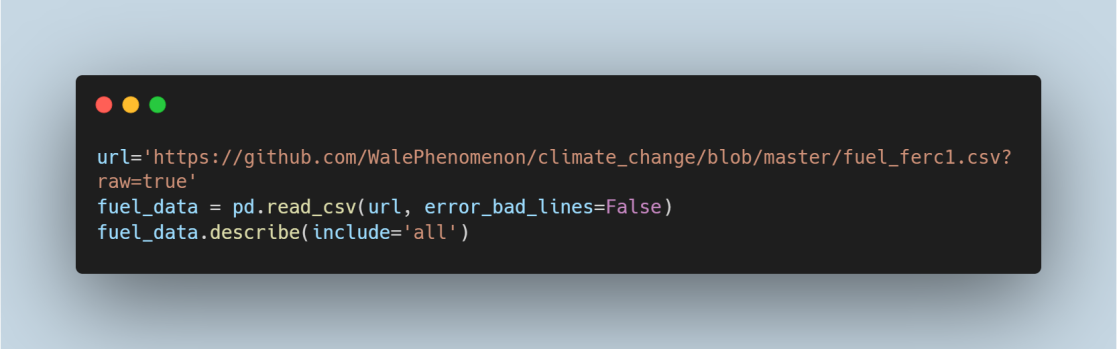
**-Working with different types of data: text files, CSV, JSON objects, HTML and databases**

The pandas library is vast enough to read data from and save to several file formats such as CSV, JSON, HTML and even databases.

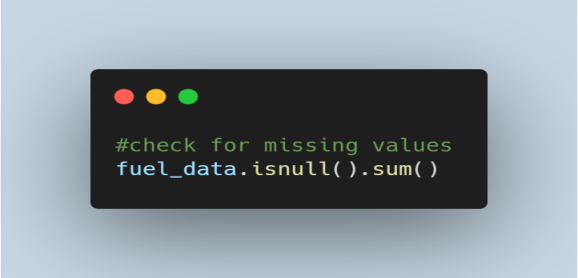


Pandas can connect to databases, get data with queries, and save in a dataframe.

In the later part of this section, we will use [this](https://github.com/WalePhenomenon/climate_change/blob/master/fuel_ferc1.csv?raw=true) fuel dataset to perform some data wrangling operations which can be found in the example notebook for this module. In our notebook, after reading the csv file, we proceed to get a summary of the dataset using the describe function.



Our analysis shows that there are 180 missing values in the fuel data column. We handle this by filling with the most common value in the column - mcf.



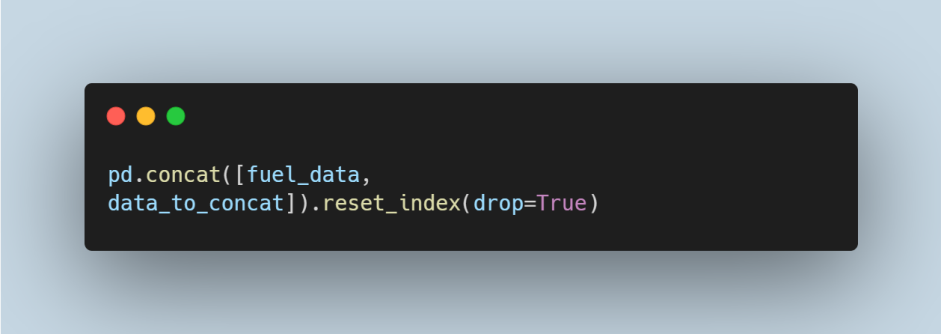
A dataframe can be easily categorised into different segments based on a given criteria using the **groupby()**function. This initially splits the dataframe into the groups then applies a function to the groups after which the results are combined.



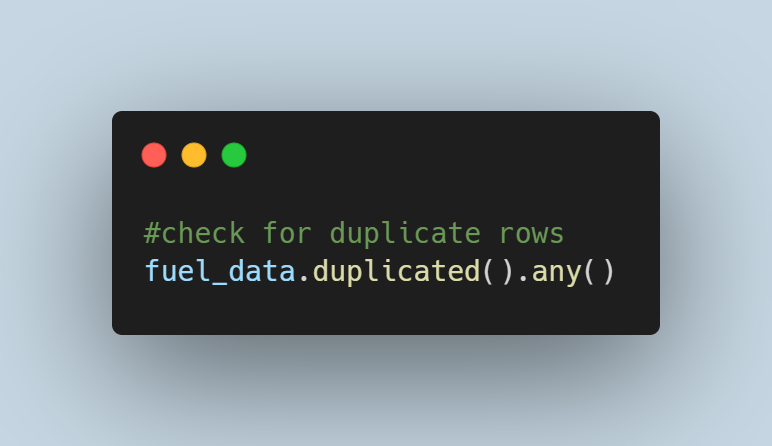
Merging in Pandas can be likened to join operations in relational databases like SQL. Left, inner, right and outer are the merging methods available to the merge() function. The left method can be likened to SQL left outer join,  inner to  SQL inner join, right to SQL right outer join and outer to SQL full outer join. In our analysis, we split the fuel data into two groups and merge using different methods.



Concatenation is performed with the concat() function by combining series or dataframes while keeping the indices of the individual unit irrespective of duplicate indices. In the notebook, we created a dummy dataframe **data\_to\_concat** which we concatenated to the fuel\_data as below:



Duplicates are a common occurrence in datasets which alter the results of data analysis. Hence, in practice, removing duplicate values is very important. The duplicated() function is used in Pandas to check for and handle duplicates.



#### Data Visualization and Representation in Python

Assuming we have collected data on energy consumption across different states in a  country, and how much people earn in these states; calculating some summary statistics can give quick insights to which state consumes the most energy; how much energy is used on average in the country, the correlation between people’s earnings, and energy consumed in their states and many others. It is essential to note that while these statistics are important and give a description of the dataset, it is not sufficient to use the results alone without plotting the data to obtain a holistic view of the overall distribution. Anscombe Quartet identifies that different datasets can have the same or very identical statistical properties such that they can be labelled the same, but when graphed, they are seen to have different distributions.

**Plotting: Area plots, Histograms, Bar charts, Pie Charts and Scatter Charts**

There are several types of charts used in data visualisation which are selected based on the data and the information aimed to be communicated. Area plots, histograms, bar charts, pie charts, and scatter charts are some of the simple and common graphs used in data analysis. Graphs assist in understanding data when performing EDA and in conveying insights easily.  A line graph is a basic plot that displays the relationship between two variables on each axis by connecting data points together with straight lines. To show magnitude, the segment between the line and the x-axis is filled which results in an area graph. Histograms and bar charts are completely different plots that can be mistaken as similar. The former are charts used to represent the distribution of a group and use adjacent rectangular bars to display the frequency of intervals while the latter are charts that represent categories using equally spaced rectangular bar.