**Phase-2**

**Data Pre-processing**

|  |  |
| --- | --- |
| Date | 8 October 2023 |
| Team ID | proj-212168-Team-2 |
| Project Name | Market Basket Insights |
| Maximum marks |  |

Data pre-processing refers to the process of cleaning, transforming, and organizing raw data into a format that is suitable for machine learning algorithms and models. Data pre-processing aims to make the data more understandable and valuable for the AI model by addressing issues such as noise, missing values, outliers, and inconsistencies. It is an important step in the data mining process.

**Program:**

#import packages:

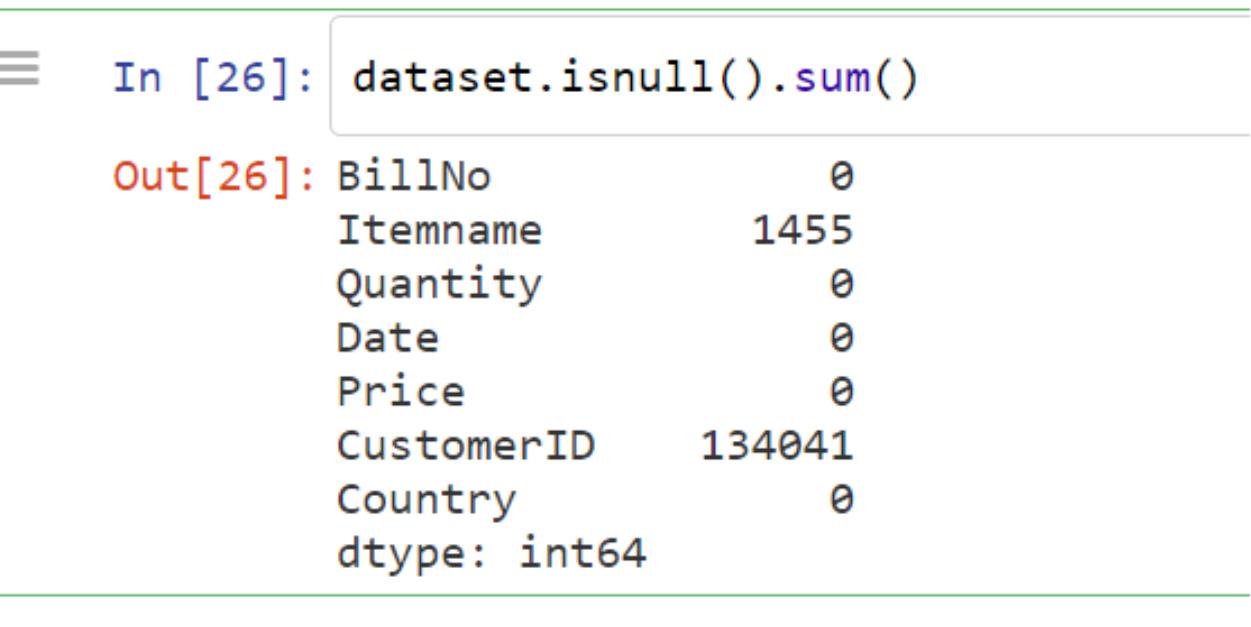
• Numpy :(import numpy as np) a library for mathematical operations and handling arrays.

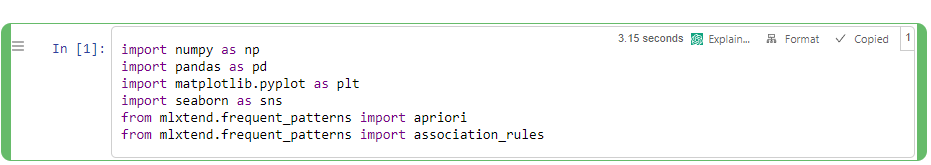
• Pandas :(import pandas as pd) a library for data manipulation and analysis.

• matplotlib.pyplot: (import as plt) a library for creating visualization.

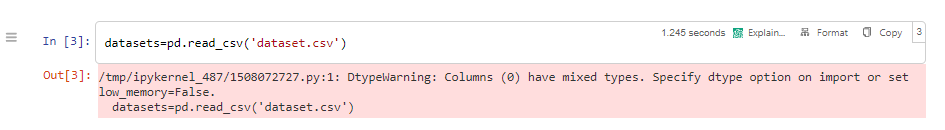
• Seaborn :as a library for creating additional data visualization.

• mlxtend.frequent\_patterns: a module for performing frequent itemset

mining and association rule learning.

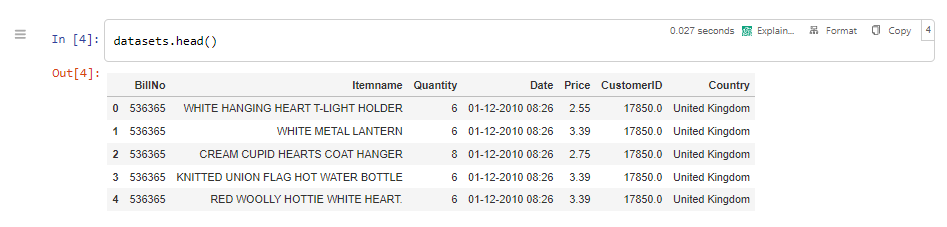


#Load the dataset:



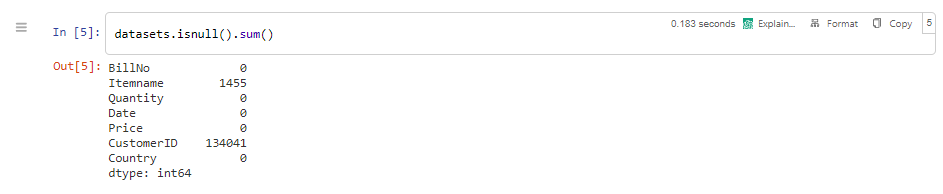
This code reads contents of a csv file called "dataset.csv" and saves it a

variable called "datasets".

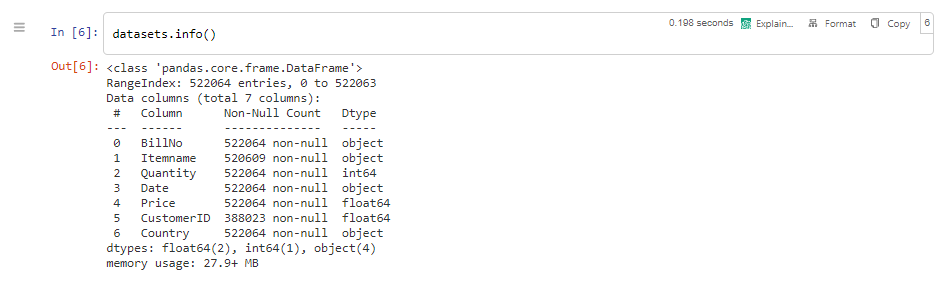


The code datasets.head() is calling the head() function on the dataset

is used display first few rows of a data set.

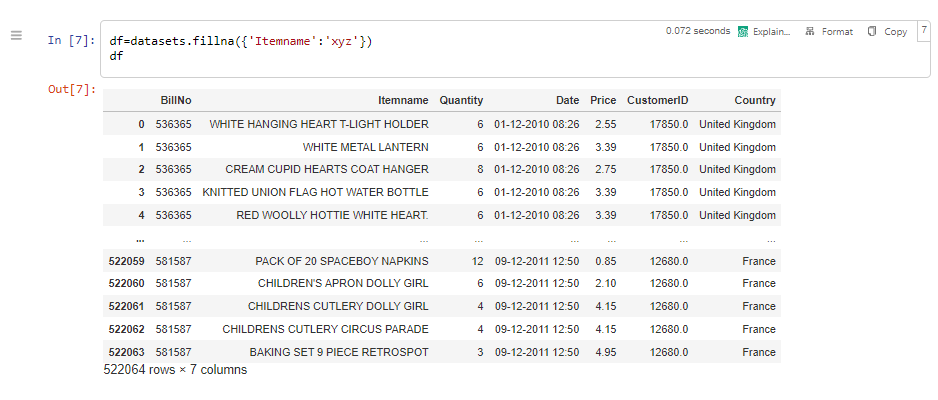


The isnull() function is used to find the number of missing values in column of a dataset. The sum() function is count the number of missing values.

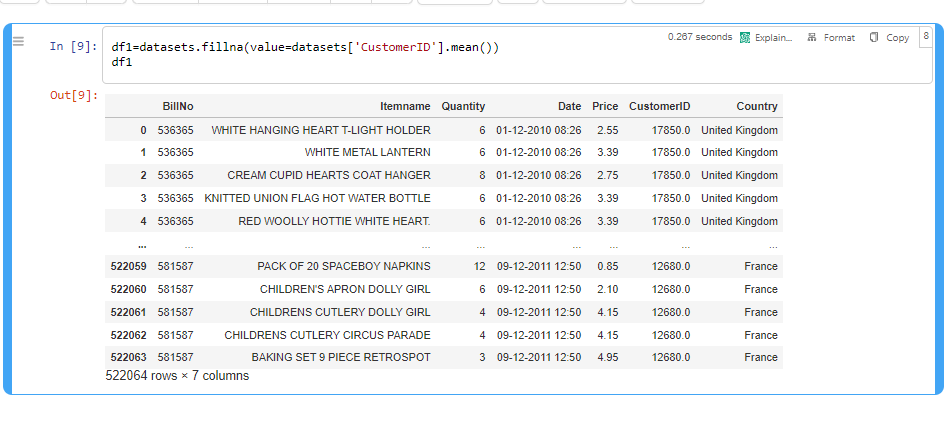


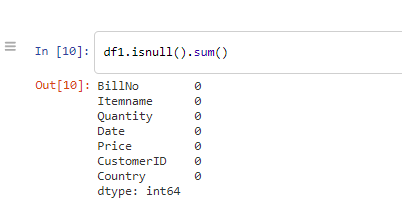
The code datasets.info() is a method call in python to display the

information about dataset. The info() method provides such as number of columns and rows datatypes of columns and memory usage of the dataset.

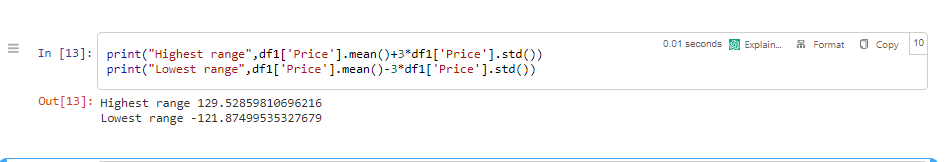


The fillna() is used to filling the missing values in the columns "Itemname" of the data frame "datasets" with the value "xyz". The filled data frame is then displayed.

This code is fills the missing values in a data frame called dataset, using the mean of the "CustomerID" column. The filled data frame than assigned variable df1and displayed.



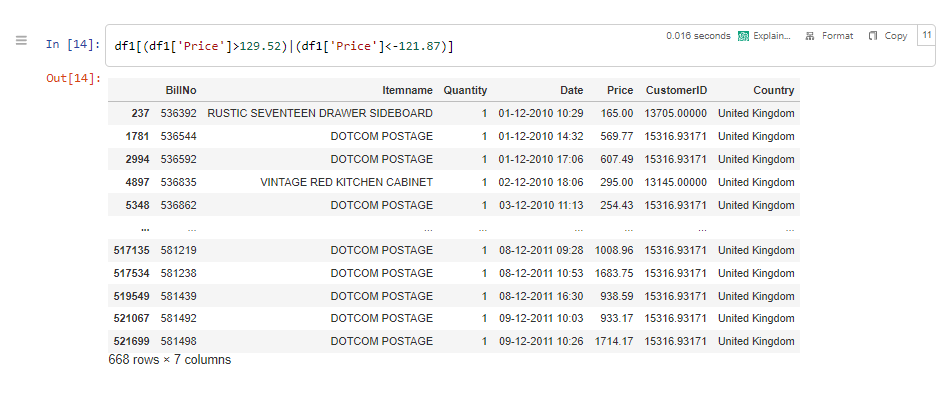
The isnull() function is used to find the number of missing values in column of a dataset. The sum() function is count the number of missing values.

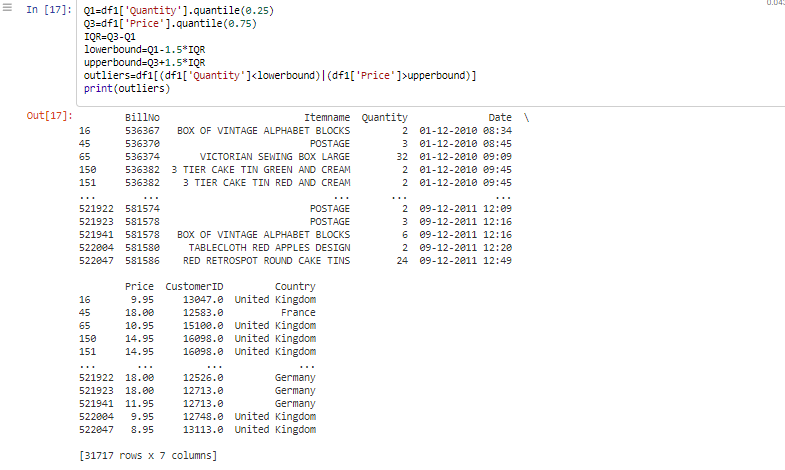


This code is printing the highest and lowest range based on

statistical calculation. It calculates the mean and standard deviation of

column called "Price" in data frame called df1.

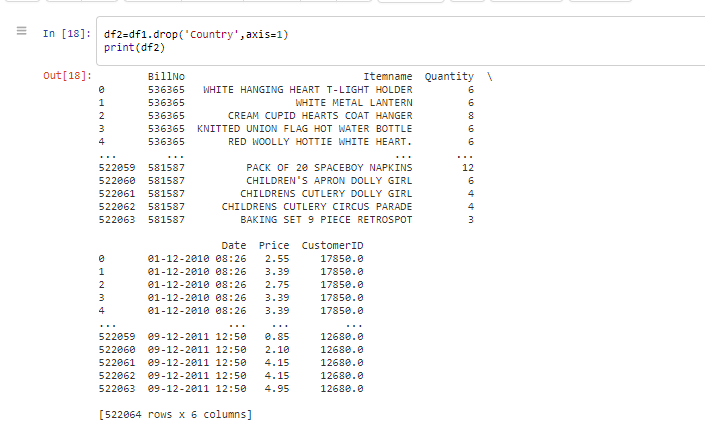
This code is used to filtering the data frame df1 based on the given condition.



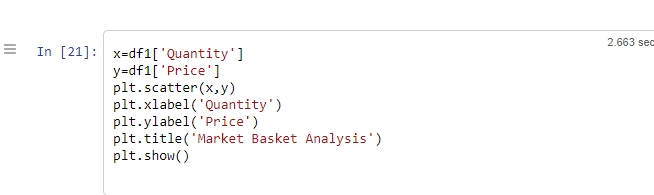
* Q1 and Q3 are the first and third quartiles of the 'Quantity' and 'Price' columns, respectively.
* IQR is the interquartile range, calculated as the difference between Q3 and Q1.
* lowerbound and upperbound are the lower and upper bounds,

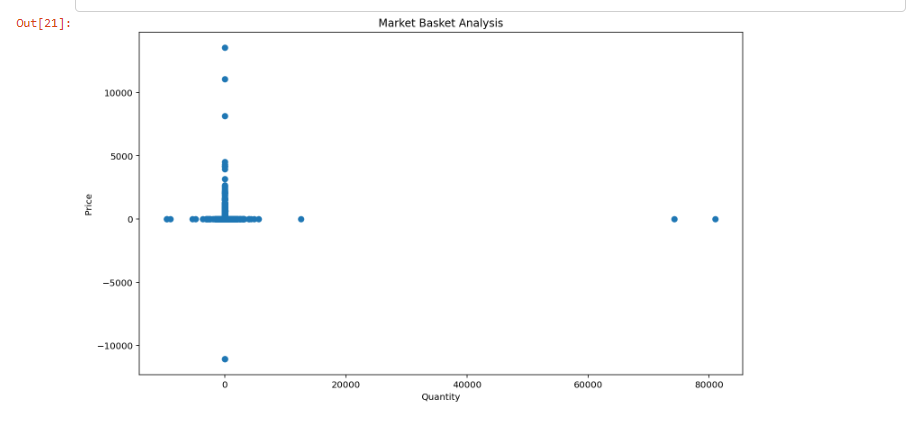
respectively, for identifying outliers. They are calculated as Q1 - 1.5 \* IQR and Q3 + 1.5 \* IQR.

* Outliers is a Data Frame containing the rows from df1 where either the 'Quantity' is less than lowerbound or the 'Price' is greater than upperbound.
* Finally, the code prints out the outliers Data Frame.



This code using the pandas library in Python to drop the 'Country' column from a Data Frame called df1. The 'axis=1' parameter specifies that the column is being dropped.





This code takes two column values from a data frame and assign them to the variables x and y, plots them as a scatterplot using the scatter() function from the pyplot module of the matplotlib library, adds labels to the x-axis using the xlabel() and y-axis using the ylabel(), sets a title to the plot using the title() and displays the plot using the show() .