# Phase-2

**Data Preprocessing Market Basket Analysis:**

|  |  |
| --- | --- |
| Date | 09 oct 2023 |
| Team ID |  |
| Project Name | Market Basket Insights |
| Maximum Mark |  |

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining taks.

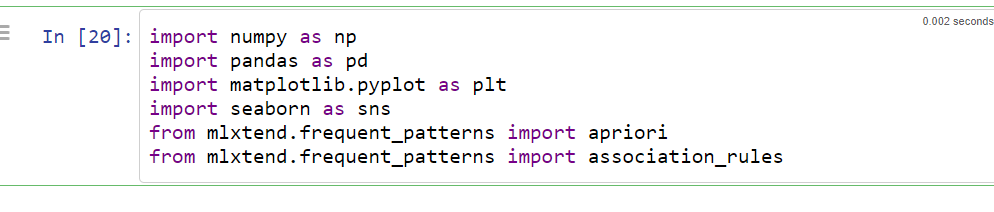
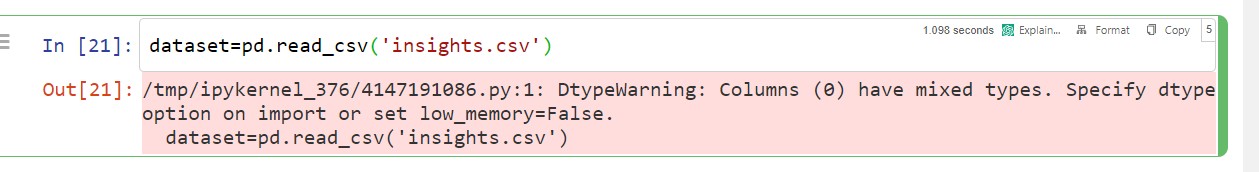
# Program:

**#Import package:**

**Explaination:**

* 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 visualiiation.
* Seaborn :as a library for creating additional data visualiiation.
* mlxtend.frquent\_patterns: a module for performing frequent itemset

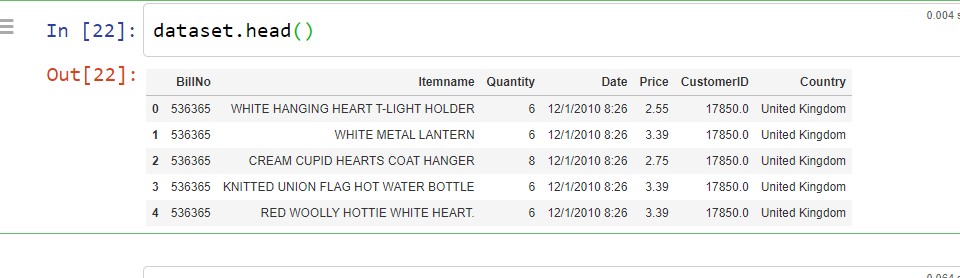
mining and association rule leaening.

•

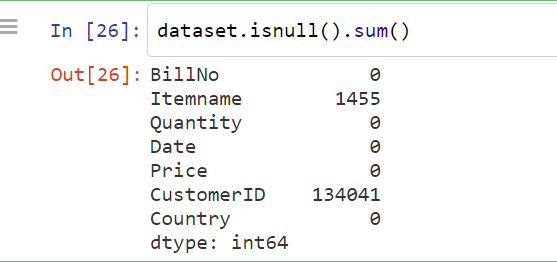
**#Load the Dataset:**

This code reads contents of a csv file called "insights.csv" and saves it a variable called"dataset".The "pd" modul is already imported.

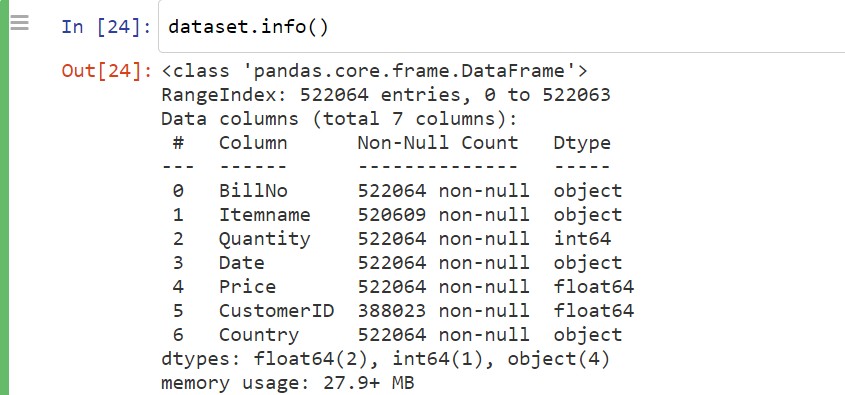
•



The code dataset.head() is calling th head() function on the dataset onjecr.the head() function is used display first few rows of a data set.



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

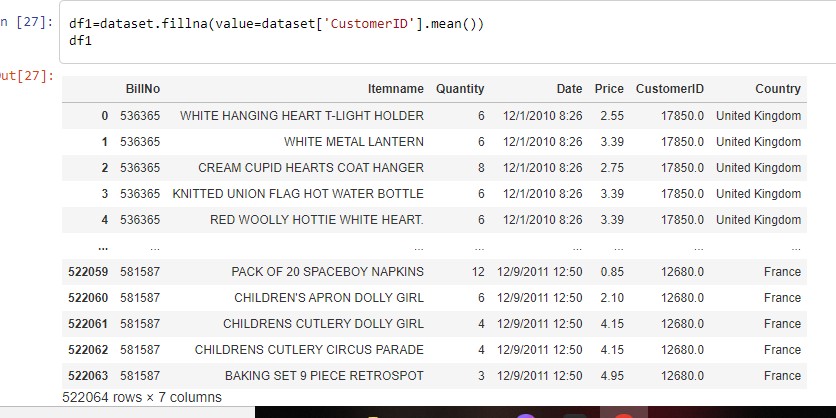


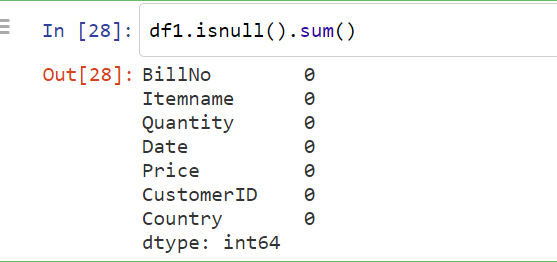
* The code dataset.info() is a method call in python to display the information about data set.The .info() method provides such as number of columns and rows datatypes of columns and memory usage of the dataset.



* This code is filling the missing values in the columns "itemname" of the dataframe "dataset" with the value "abcd".The filled dataframe is then displayed.

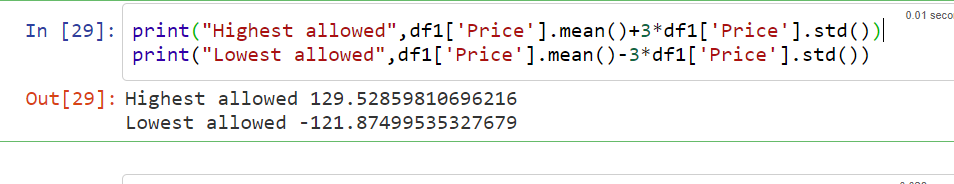
•

This code is fiils the missing values in a dataframe calles dataset,using the mean of the "CustomerID" column.The filled dataframe than assigned variable df1and displayed.

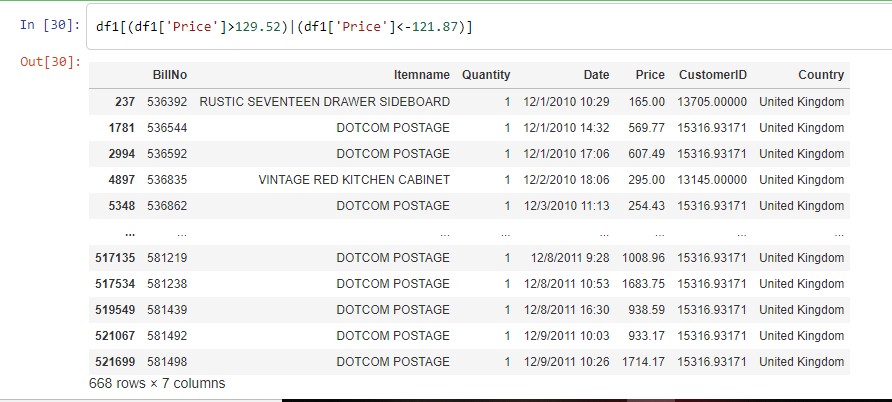


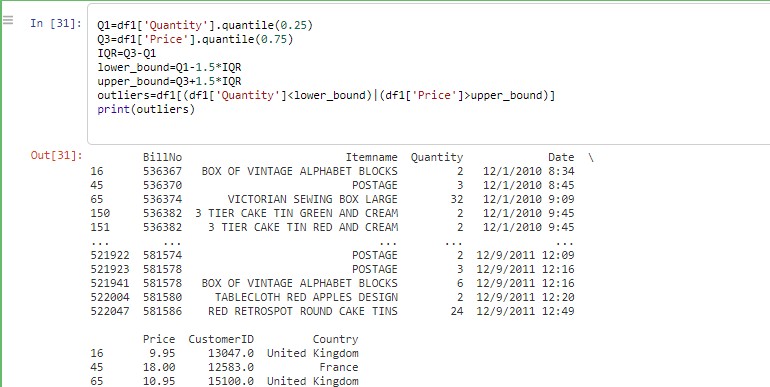
* The given code is used to find the number of missing values in column of a dataset.The sum() function is count the number of missing values.

•

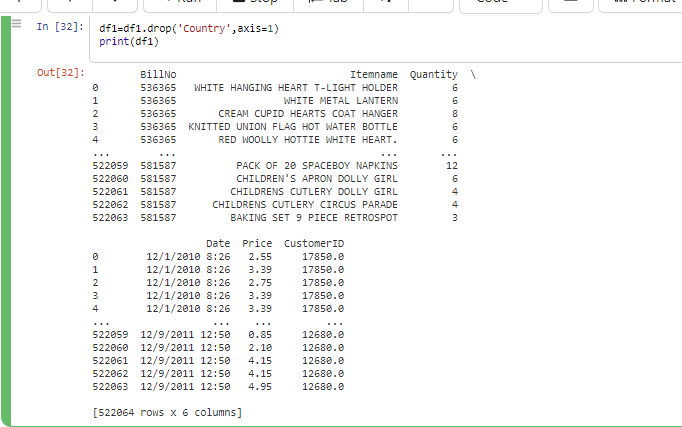
This code is printing the highest and lowest allowed values based on statistical calculation.It calculates the mean and standared diviation of column called "price" in dataframe calld df1.

•

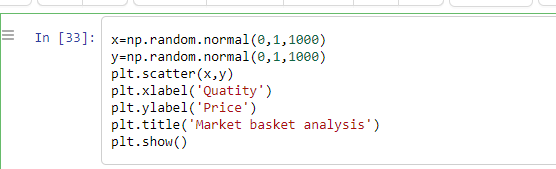
The code is filtering a dataframe df1 based on a condition.

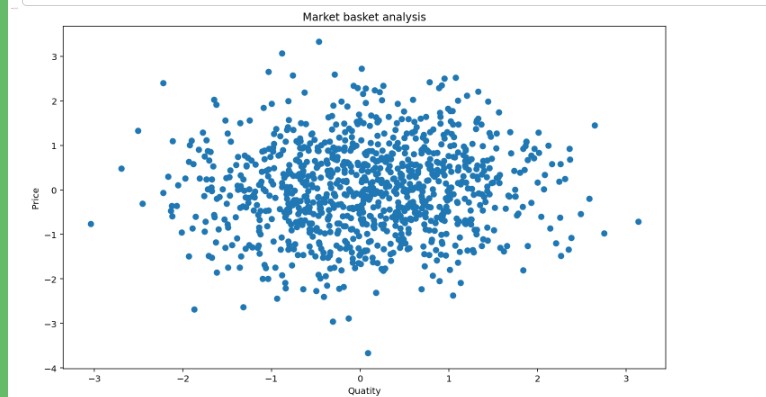


* Q1 and Q3 are the first and third quartiles of the 'Quantity' and 'Price' columns, respectively.
* IQR is the interquartile range, calculated as the diference between Q3 and Q1.
* lower\_bound and upper\_bound are the lower and upper bounds, respectively, for identifying outliers. They are calculated as Q1 - 1.5 \* IQR and Q3 + 1.5 \* IQR.
* outliers is a DataFrame containing the rows from df1 where either the 'Quantity' is less than lower\_bound or the 'Price' is greater than upper\_bound.
* Finally, the code prints out the outliers DataFrame.



* The code is using the pandas library in Python to drop the 'Country' column from a DataFrame called df1. The 'axis=1' parameter specifies that the column is being dropped.
* After dropping the column, the code then prints the updated DataFrame.





* This code generates two arrays of random numbers with a normal distribution, assigns 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 and y-axis, sets a title the

plot, and displays the plot.