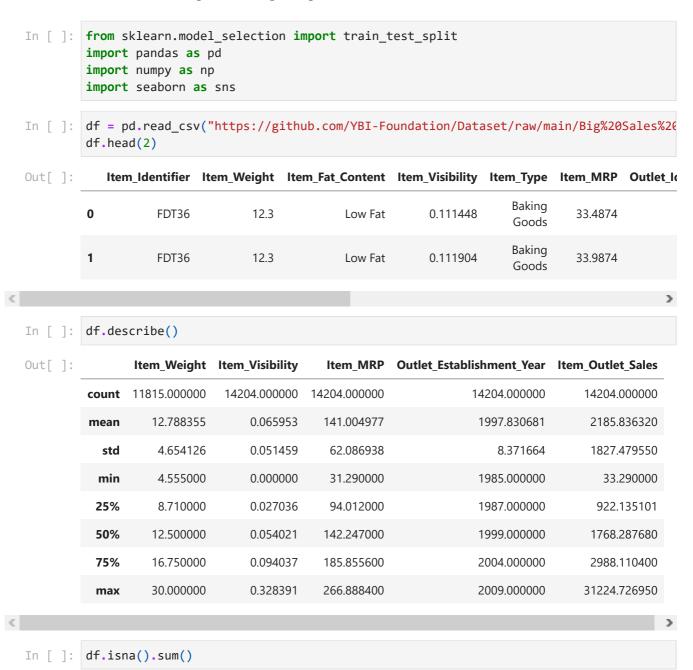
Collect the appropriate business dataset from any repository and analyze the summary of the dataset and further divide dataset for training and testing purpose.

- a. Download the official dataset and import it as input.
- b. Give the details of the dataset such as no of samples, no of features, target vector etc by using python command.
- c. By using python command, divided the dataset into training, testing dataset and target dataset for training and testing along with validation dataset.



```
Out[]: Item Identifier
                                       0
                                    2389
        Item Weight
        Item_Fat_Content
                                       a
        Item Visibility
                                       0
        Item_Type
                                       0
        Item_MRP
                                       0
        Outlet Identifier
                                       0
        Outlet_Establishment_Year
                                       0
                                       0
        Outlet Size
        Outlet_Location_Type
                                       0
        Outlet_Type
                                       0
        Item_Outlet_Sales
                                       0
        dtype: int64
In [ ]: df.columns
Out[ ]: Index(['Item_Identifier', 'Item_Weight', 'Item_Fat_Content', 'Item_Visibility',
               'Item_Type', 'Item_MRP', 'Outlet_Identifier',
               'Outlet_Establishment_Year', 'Outlet_Size', 'Outlet_Location_Type',
               'Outlet_Type', 'Item_Outlet_Sales'],
              dtype='object')
In [ ]: df.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 14204 entries, 0 to 14203
        Data columns (total 12 columns):
         # Column
                                       Non-Null Count Dtype
                                       -----
         0
            Item Identifier
                                       14204 non-null object
           Item_Weight
                                      11815 non-null float64
         1
         2 Item_Fat_Content
                                      14204 non-null object
            Item_Visibility
                                      14204 non-null float64
         3
            Item Type
                                      14204 non-null object
         5
            Item MRP
                                      14204 non-null float64
            Outlet_Identifier
                                      14204 non-null object
         6
             Outlet_Establishment_Year 14204 non-null int64
         7
           Outlet_Size
                                      14204 non-null object
         8
         9
             Outlet_Location_Type
                                     14204 non-null object
                                      14204 non-null object
         10 Outlet Type
         11 Item Outlet Sales
                                       14204 non-null float64
        dtypes: float64(4), int64(1), object(7)
        memory usage: 1.3+ MB
In [ ]:
       df.shape
Out[ ]: (14204, 12)
In [ ]: df.corr()['Item_Outlet_Sales']
Out[]: Item Weight
                                    0.228297
        Item Visibility
                                    -0.158813
        Item MRP
                                    0.532261
        Outlet_Establishment_Year
                                    -0.110786
        Item_Outlet_Sales
                                    1.000000
        Name: Item_Outlet_Sales, dtype: float64
In [ ]: X= df.drop(['Item_Identifier','Outlet_Establishment_Year','Item_Outlet_Sales'],axis
        Y = df['Item Outlet Sales'] # target variable
In [ ]: x_main,x_test,y_main,y_test = train_test_split(X,Y,random_state=42,test_size=0.2)
```

```
In [ ]: x_train,x_val,y_train,y_val = train_test_split(x_main,y_main,random_state=42,test_s
In [ ]: x_train.shape, y_train.shape, x_val.shape, y_val.shape, x_test.shape,y_test.shape
Out[ ]: ((9090, 9), (9090,), (2273, 9), (2273,), (2841, 9), (2841,))
In [ ]: print('training set: ', x_train.shape, y_train.shape)
    print('validation set: ', x_val.shape, y_val.shape)
    print('testing set: ',x_test.shape,y_test.shape)

    training set: (9090, 9) (9090,)
    validation set: (2273, 9) (2273,)
    testing set: (2841, 9) (2841,)
```