data preprocessing:

1. import modules

2. connect to drive1

from google.colab import drive

drive.mount('/content/drive')

3. read dataset

ml = pd.read\_csv('/content/drive/MyDrive/ml/archive.zip')

4.Removal of unnamed column (first column) and printing the dataset

5.IDENTIFICATION AND HANDLING OF MISSING VALUES

6.CATEGORICAL AND NUMERICAL FEATURES

7.Encoding categorical values into numerical values

8.Examining the spread of data and identifying which of the numerical features contains outliers and visualize it using Boxplots

a .counting the number of outliers present in each column

b.Removing:

=> example: AveragePrice has less no. of outliers. As removing them won't cause any information loss, outliers in this column are removed.

c.Visualization after handling outliers (boxplots)

9.Feature Scaling - Normalization

10.Splitting of Training data and Testing data

11. Distribution Plot

REGRESSION STEPS :

1.Connect to drive

2.import the modules

3.read the dataset which is already preprocessed in datapreprocessing exercise

4.linear regression

a)Fitting a linear model to the data to predict the prices of avocados using linear regression

b)Assess the strength of the relationship between independent and dependent variables through statsmodels.summary()

c)Interpret the strength of the relationship between independent and dependent variable via the correlation matrix.

d)Calculating positive threshold and negative threshold for average price correlation values.

e)Selecting features that are either above the positive threshold or below the negative threshold.

f)Produce a scatterplots for each of the independent variable against the dependent variable

g)Based on the scatterplots and correlation values, identify which variable(s) you will use for prediction

Clustering :

1)from google.colab import drive

2)drive.mount('/content/drive')

3)Preprocessing the dataset

ml = pd.read\_csv('/content/drive/MyDrive/ml /live.zip')

Removing null columns :

Unique values :

Categorical and Numerical features

Applying feature scaling

4)K MEANS CLUSTERING

Plotting the clusters and centroids

majority types

5)Hierarchical clustering

Dendrogram

Optimal K value

Majority status types