**INFO 5810 Assignment-1**

3) 1. Dataset Quality checking:

There is no bias in the dataset.

A screenshot of a computer

Description automatically generated

3) 2. There might be a change of having missing values and duplicates while taking large datasets and we need to handle them by removing duplicates and replacing the missing values. Those missing values can be replaced with average, mean and median, zero or any other values if they are defined.

A screenshot of a computer

Description automatically generated

3) 3. In the given Avocado price dataset, we have 18,254 and 14 Attributes (columns).

4) 1. After removing the first column or Column 1 we have only 13 Attributes in the dataset.

A screenshot of a computer

Description automatically generated

To remove duplicates we need to select the object of the dataset in the design tab.

A screenshot of a computer

Description automatically generated

After successfully running the dataset, we need to add select attributes object to get all attributes.

A screenshot of a computer

Description automatically generated

We have taken the Year attribute and changed it to nominal. By selecting Numerical to Polynomial.

A screenshot of a computer

Description automatically generated

After changing the Year attribute to nominal.

A screenshot of a computer

Description automatically generated

4) 2. To remove duplicates, I have selected the Remove duplicates object to check the duplicates and remove duplicates in the dataset.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

4) 3. To replace missing values, we have taken the replace missing values object and run it to the dataset.

A screenshot of a computer

Description automatically generated

We can observe below all the missing values rows have value 0.

A screenshot of a computer

Description automatically generated

4) 4. To create the correlation matrix for the dataset, I have selected the correlation matrix and creating a matrix using dataset.

A screenshot of a computer

Description automatically generated

We can observe the correlation matrix below:

A screenshot of a computer

Description automatically generated

5) Selecting the Continuous Attributes from the dataset.

A screenshot of a computer

Description automatically generated

We can see Min, Max and Average of all the attributes selected.

A screenshot of a computer

Description automatically generated

Selecting the Categorial Values:

A screenshot of a computer

Description automatically generated

The earliest date is Jan 21, 1904. The range of data is huge, and I guess removing the data is the best option as most of the data is after the year 2015.

A screenshot of a computer

Description automatically generated

I have sorted in highest to lowest and I found Average is 3.250 and in San Francisco region.

A screenshot of a computer

Description automatically generated

62505646.520 is the highest total volume of avocados in TotalUS.

A screenshot of a computer

Description automatically generated