Problem Set IV - Data Preprocessing

1. Select a **subset of relevant attributes** from the given dataset that are necessary to know about the total volume of avocados with product lookup codes (PLU) 4046, 4225, 4770) which are of organic type. (Use AVOCADO dataset)

2. **Discard** all **duplicate** entries in the given dataset and fill all the missing values in the attribute "AveragePrice" as 1.25. Also print the size of the dataset before and after removing duplicates. (Use Trail dataset)

3. **Binarize** the attribute "Year". Set the threshold above 2016 and print it without truncation. (Use AVOCADO dataset)

4. Transform all categorical attributes in the dataset AVOCADO using **Integer Encoding.**

5. Transform the attribute = "Region" in the given dataset AVOCADO using **One-Hot Encoding**.

6. **Ignore** the tuples that hold missing values and print the subset of data from AVOCADO dataset.

7. **Drop** the attribute that has **high nullity** as it facilitates efficient prediction. (Use AVOCADO dataset)

8. Study the entire dataset and report the complete **statistical summary** about the data (Use AVOCADO dataset)

- Dimension of the dataset
- Most frequently occurring value under every attribute.
- Datatype of every attribute
- Count
- Mean
- Standard Deviation
- Minimum Value
- Maximum value
- 25% (Lower Quartile)
- Median i.e. 50%
- 75% (Upper Quartile)
- Find whether the class distribution of dataset is imbalanced. (Note: Fix the class label as "Type" in the given dataset)
- Correlation matrix
- Skewness of every attribute.

(For the below exercises, you are free to choose an appropriate data set as merited by the problem statements)

9. Test drive the use of Gini Index, Information Gain, Entropy and other measures that are supported in your platform, performing the role of data selection.

10. Test drive the implementation support in your platform of choice for data preprocessing phases such as cleaning, selection, transformation, integration in addition to the earlier exercises.