Problem Set IV - Data Preprocessing

1. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Use ***min-max normalization*** to transform the values of age to the range[0:1].
(b) Use ***z-score normalization*** to transform the values of *age.*
(c) Use normalization by ***decimal scaling*** to transform the values of *age* such that the transformed value is less than 1.

2. Use the given dataset and perform the operations listed below.

Dataset Description

Attributes in the dataset:

- Date - The date of the observation
- Average Price - the average price of a single avocado
- type - conventional or organic
- year - the year
- Region - the city or region of the observation
- Total Volume - Total number of avocados sold
- 4046 - Total number of avocados with PLU* 4046 sold
- 4225 - Total number of avocados with PLU* 4225 sold
- 4770 - Total number of avocados with PLU* 4770 sold

(Product Lookup codes (PLU's)) *

a. Sort the attribute "Total Volume" in the given dataset and distribute the data into equal sized/frequency bins. Let the number of bins be 250. ***Smooth*** the sorted data by

(i)bin-means *(ii) bin-medians* *(iii) bin-boundaries*

b. The dataset represents weekly retail scan data for National retail volume (units) and price. However, the company is interested in knowing the *monthly (total per month) and annual sales (total per year)*, rather than the total per week. So, ***reduce*** the data accordingly.

c. **Summarize** the number of missing values for each attribute

d. Populate data for the ***missing values*** of the attribute= "Average Price" by averaging all the values of the "Avg Price" attribute that fall under the same "**REGION**" attribute value.

e. ***Discretize*** the attribute= "Date" using ***concept hierarchy*** into {Old, New, Recent} (Consider 2015,2016 : Old, 2017: New, 2018: Recent).