Problem Set III- 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 value 25 for *age* onto the range [0:0;1:0].
(b) Use ***z-score normalization*** to transform the value 25 for *age*, where the standard deviation of *age* is 12.94 years.
(c) Use normalization by ***decimal scaling*** to transform the value 25 for *age* such that transformed value is <1

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

Dataset Description

It is a well-known fact that Millenials LOVE Avocado Toast. It's also a well known fact that all Millenials live in their parents basements.

Clearly, they aren't buying home because they are buying too much Avocado Toast!

But maybe there's hope... if a Millenial could find a city with cheap avocados, they could live out the Millenial American Dream. Help them to filter out the clutter using some pre-processing techniques.

Some relevant columns in the dataset:

- Date - The date of the observation
- AveragePrice - 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 dataset given and distribute the data into equal sized/frequency bins of size 50 & 250.***Smooth*** the sorted data by

(i)bin-means
(ii) bin-medians
(iii) bin-boundaries (smooth using bin boundaries after trimming the data by 2%).

b. The dataset represents weekly retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers' cash registers based on actual retail sales of Hass avocados. However, the company is interestedin 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 their values that fall under the same region.

e. ***Discretize*** the attribute="Date" using ***concept hierarchy*** into {Old, New, Recent} {2015,2016 : Old, 2017: New, 2018: Recent} and plot in q-q plots