Data Preprocessing and Normalization Report

Analytics and Systems of Big Data

Thallapally Nimisha
CS22B1082
B.Tech in Computer Science and Engineering
IIITDM Kancheepuram

Contents

1 Question 1

Problem Statement:

Suppose that the data for analysis includes the attribute `age`. The age values for the data tuples are:

$$13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70$$

Perform the following:

(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.

1.1 Solution

1.1.1 (a) Min-Max Normalization

Formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, $\min(x) = 13$, $\max(x) = 70$.

Age	Normalized Age (0–1)
13	0.000
15	0.035
16	0.053
70	1.000

Table 1: Sample Min-Max Normalized Values

1.1.2 (b) Z-Score Normalization

Formula:

$$z = \frac{x - \mu}{\sigma}$$

where $\mu = 29.96$, $\sigma \approx 12.94$.

Age	Z-score
13	-1.32
20	-0.78
25	-0.39
35	+0.39
70	+3.15

Table 2: Sample Z-Score Normalized Values

1.1.3 (c) Decimal Scaling Normalization

Formula:

$$x' = \frac{x}{10^j}$$

Here, $\max(x) = 70$, so $j = 2$.

Age	Normalized Age
13	0.13
25	0.25
35	0.35
70	0.70

Table 3: Sample Decimal Scaling Normalized Values

1.1.4 Visualizations

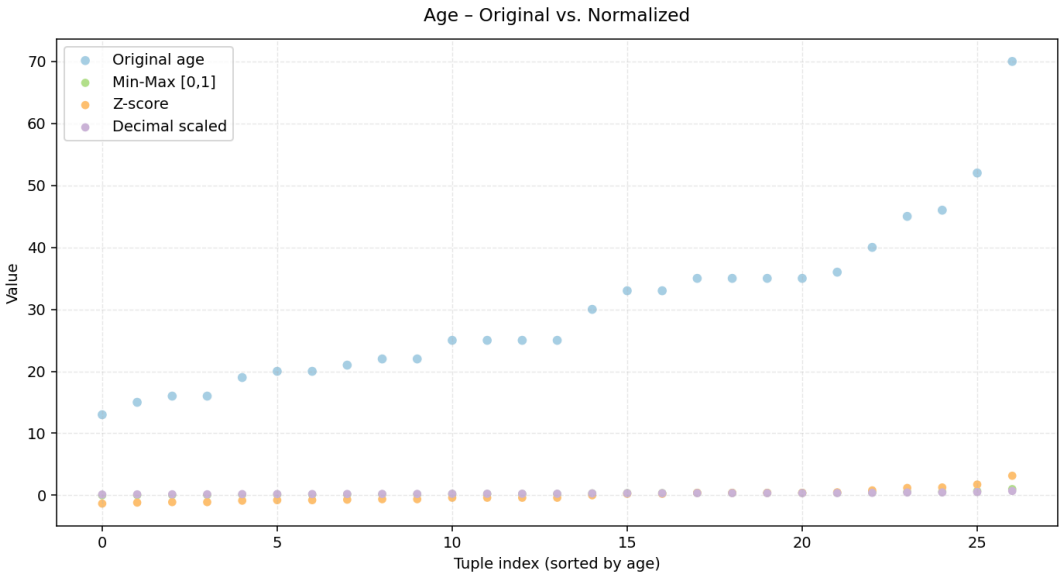

Figure 1: Scatter comparison of Original vs. Normalized values

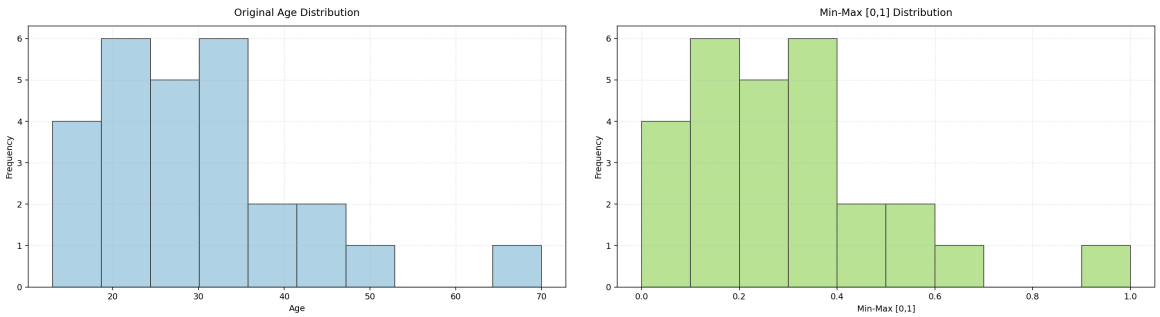

Figure 2: Distributions of Original Age and Min-Max Normalized Age

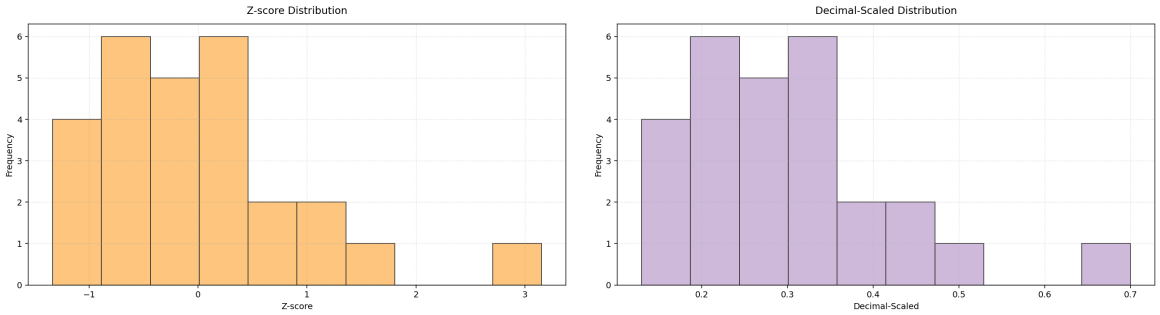

Figure 3: Distributions of Z-Score and Decimal Scaled Ages

2 Question 2

Problem Statement:

Use the given avocado dataset with the following attributes: Date, Average Price, Type, Year, Region, Total Volume, 4046, 4225, 4770. Perform the following operations:

(a) Sort "Total Volume" and distribute into 250 bins. Smooth the data by (i) bin-means (ii) bin-medians (iii) bin-boundaries.

(b) Convert weekly sales data into monthly and yearly aggregates.

(c) Summarize the number of missing values per attribute.

(d) Fill missing values of "Average Price" using region-wise averages.

(e) Discretize "Date" using concept hierarchy: {2015,2016: Old; 2017: New; 2018: Recent}.

2.1 Solution

2.1.1 (a) Binning and Smoothing

The `Total Volume` attribute was sorted and divided into 250 equal-frequency bins. Smoothing was then applied using bin-means, bin-medians, and bin-boundaries.

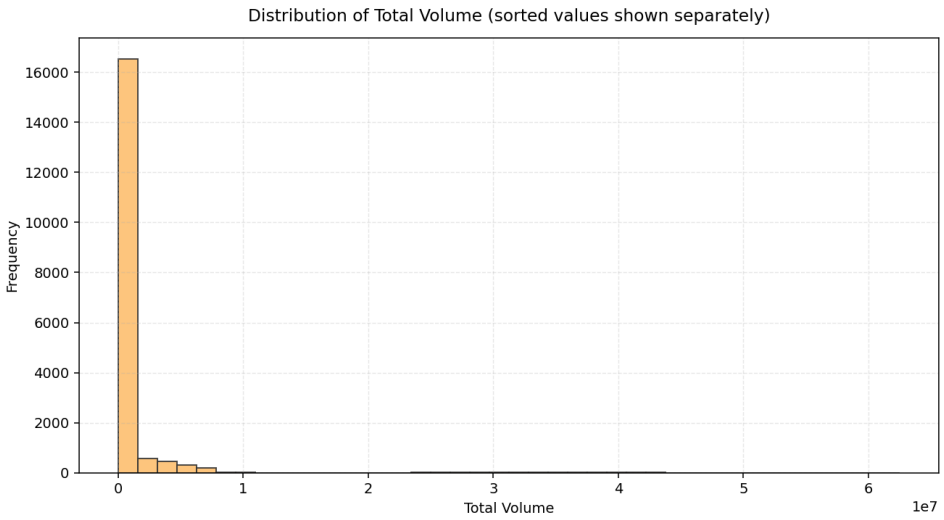

Figure 4: Distribution of Total Volume (sorted values)

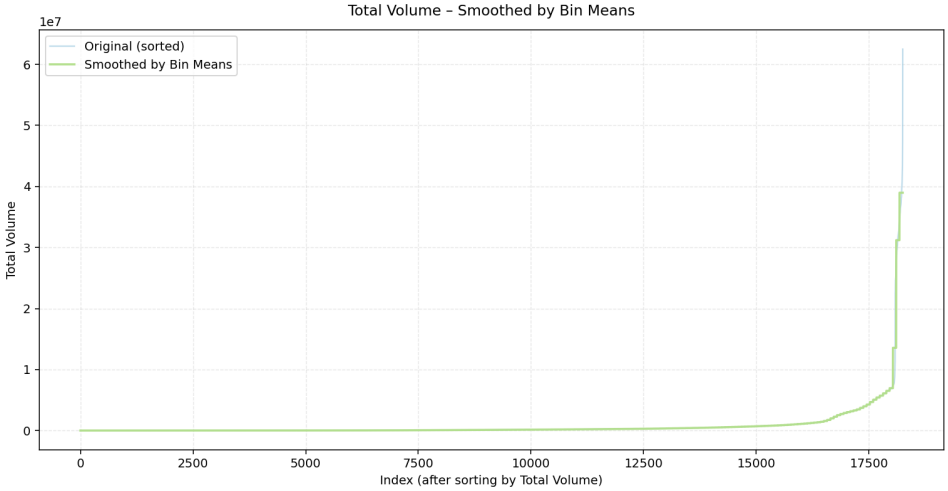

Figure 5: Total Volume smoothed using Bin Means

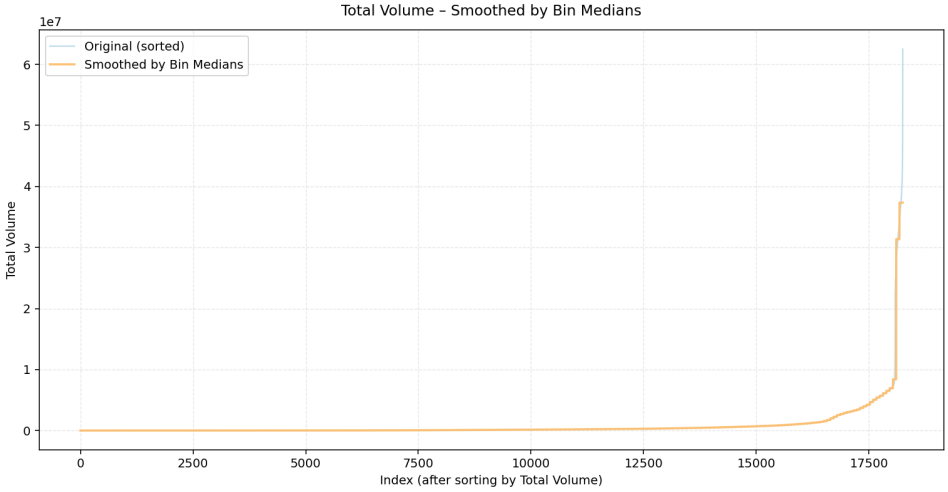

Figure 6: Total Volume smoothed using Bin Medians

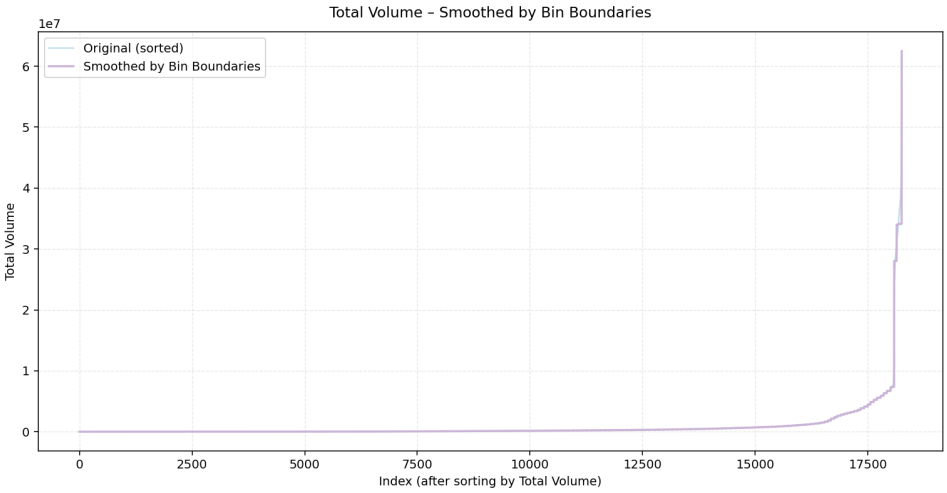

Figure 7: Total Volume smoothed using Bin Boundaries

2.1.2 (b) Data Reduction (Weekly → Monthly, Annual)

Weekly sales data were aggregated into monthly and annual totals.

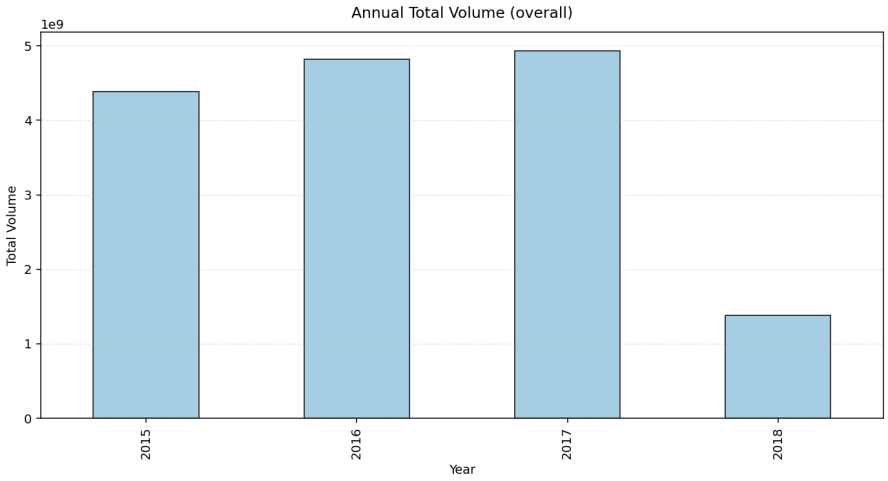

Figure 8: Annual Total Volume (overall)

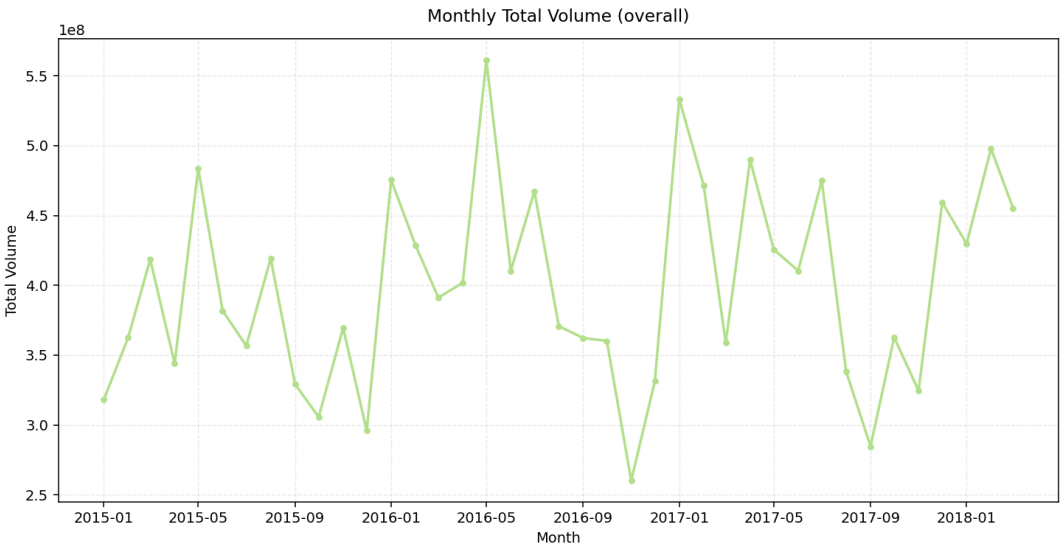

Figure 9: Monthly Total Volume (overall)

2.1.3 (c) Missing Value Summary

The number of missing values per attribute is summarized below.

Figure 10: Missing Values per Attribute

2.1.4 (d) Handling Missing Average Price

Missing `Average Price` values were imputed using the mean for the same `Region`.

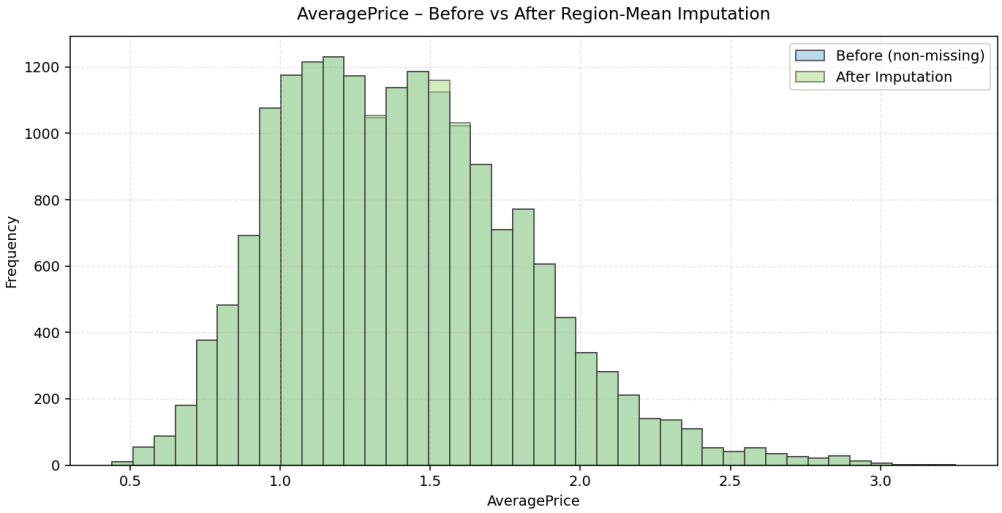

Figure 11: Distribution of Average Price before and after Region-wise Imputation

2.1.5 (e) Discretization of Dates

Dates were mapped to categories:

$$2015, 2016 \rightarrow Old, \quad 2017 \rightarrow New, \quad 2018 \rightarrow Recent$$

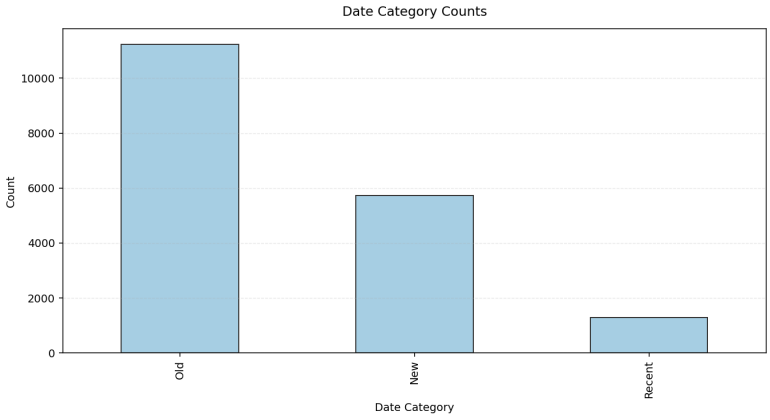

Figure 12: Date Category Counts after Concept Hierarchy Mapping