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: Attribute Selection for Organic Avocados

Problem Statement

Select relevant attributes to analyze the total volume of organic avocados with PLU codes 4046, 4225, and 4770.

Approach

The dataset was filtered to include only entries where the `type` attribute is `organic`. Relevant columns including Date, Region, PLU codes, and Total Volume were selected for analysis.

Results

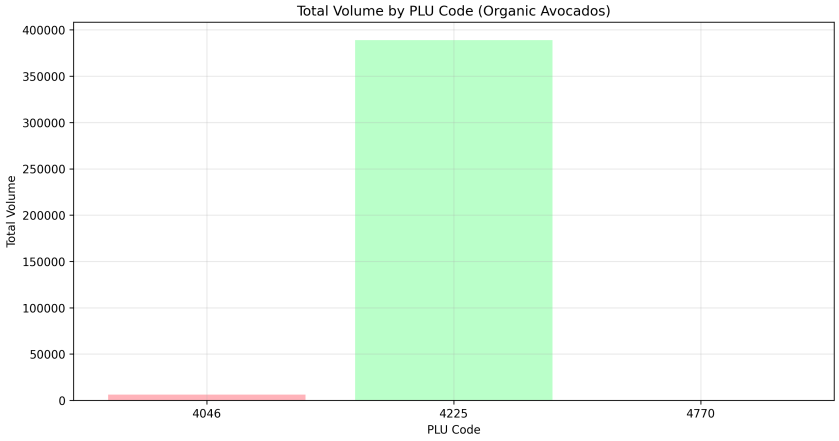

Figure 1: Total Volume by PLU Code for Organic Avocados

Detailed Analysis

Analysis of Organic Avocados - Detailed Calculations

Dataset Statistics:

- Total number of records: 18,250
- Number of organic records: 9,124
- Percentage of organic records: 49.99%

PLU Volumes Analysis:

- **PLU 4046:**
  - Total Volume: 66,702,877.39
  - Average Volume: 7,310.71
  - Maximum Volume: 361,996.84

  - Minimum Volume: 0.00
  - Percentage of Total: 31.80%

- **PLU 4225:**
  - Total Volume: 140,603,877.57
  - Average Volume: 15,410.33
  - Maximum Volume: 680,037.45
  - Minimum Volume: 0.00
  - Percentage of Total: 67.04%

- **PLU 4770:**
  - Total Volume: 2,429,040.55
  - Average Volume: 266.23
  - Maximum Volume: 26,765.78
  - Minimum Volume: 0.00
  - Percentage of Total: 1.16%

2 Question 2: Duplicate Handling and Missing Value Imputation

Problem Statement

Remove duplicate entries and fill missing values in `AveragePrice` with 1.25.

Approach

The dataset was processed by identifying and removing duplicate records, as well as filling missing values in the `AveragePrice` attribute with the value 1.25. The dataset size and price statistics were compared before and after the processing.

Results

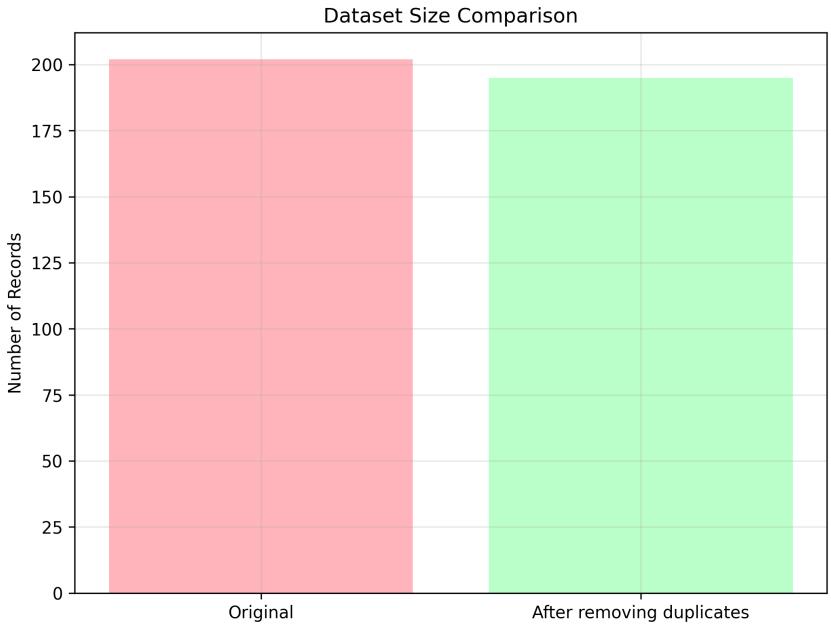

Figure 2: Dataset Size Comparison Before and After Removing Duplicates

Detailed Analysis

Trail Dataset Processing - Detailed Analysis

Dataset Statistics:

- Original number of records: 202
- Number of duplicates found: 7
- Final number of records: 195
- Percentage of duplicates: 3.47%

Missing Values Analysis:

- Original missing values in `AveragePrice`: 29
- Missing values after filling with 1.25: 0

Price Statistics:

- **Before Processing:**
  - Mean price: 1.12
  - Median price: 1.11
  - Standard deviation: 0.11
- **After Processing:**
  - Mean price: 1.11
  - Median price: 1.11
  - Standard deviation: 0.10

3 Question 3: Year Binarization

Problem Statement

Binarize the `Year` attribute using 2016 as the threshold.

Approach

Converted the year column to numeric and applied binarization.

Results

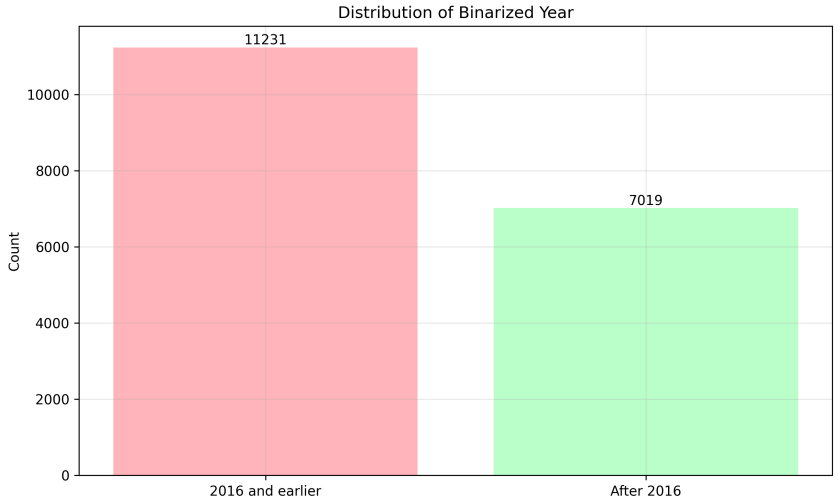

Figure 3: Distribution of Binarized Year

Distribution:
2016 and earlier: 11231 records (61.54%)
After 2016: 7019 records (38.46%)

4 Question 4: Integer Encoding of Categorical Attributes

Problem Statement

Apply integer encoding to categorical attributes.

Approach

Label encoding (also known as integer encoding) was applied to all categorical columns using scikit-learn's `LabelEncoder`. Each unique category was assigned an integer based on alphabetical order.

Results

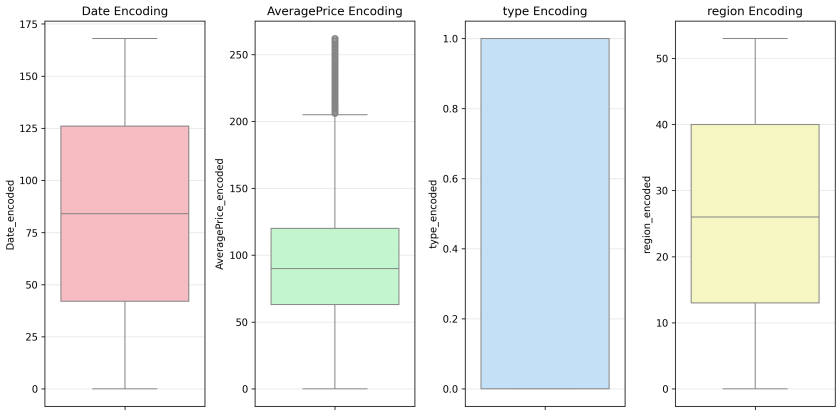

Figure 4: Integer Encoding of Categorical Variables

The mapping between original and encoded values is provided in the supplementary file `q4_encoding_mapping.csv` for reproducibility and interpretation. This encoding is useful for tree-based models but may not preserve ordinal relationships, so careful consideration is needed based on the downstream task.

—

5 Question 5: One-Hot Encoding of Region

Problem Statement

Transform the `region` attribute using one-hot encoding.

Approach

`pandas.get_dummies()` was used to one-hot encode the `region` attribute, producing a binary column for each unique region. This technique avoids introducing ordinal relationships into nominal categorical features and is especially suitable for linear models.

Results

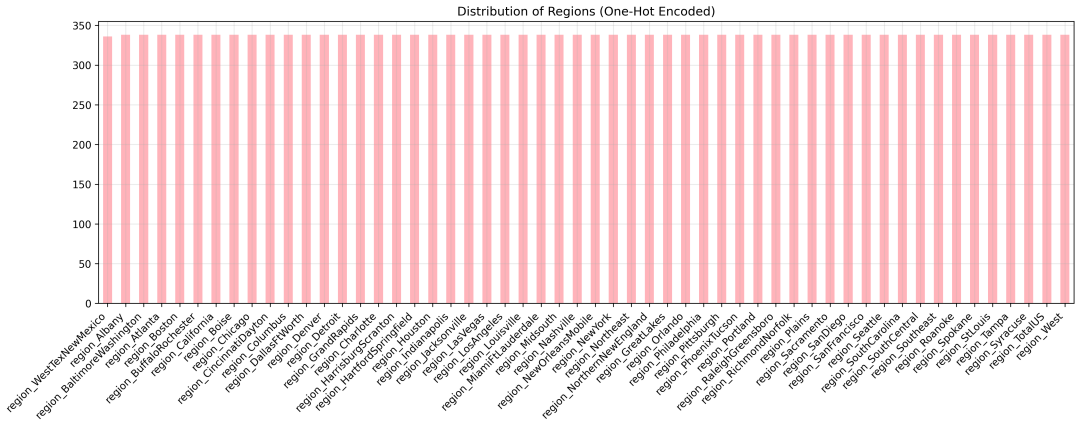

Figure 5: Distribution of Regions (One-Hot Encoded)

The plot shows the frequency distribution of records across all regions. This helps highlight regional imbalance, which may need to be addressed in modeling via sampling or regularization.

—

6 Question 6: Handling Missing Values

Problem Statement

Exclude records with missing values and analyze the resulting dataset.

Approach

The dataset was scanned for `NaN` values. Rows containing any missing values were dropped, and the dataset size was compared before and after cleaning to understand the impact. Additionally, missing value counts and percentages were computed per column.

Results

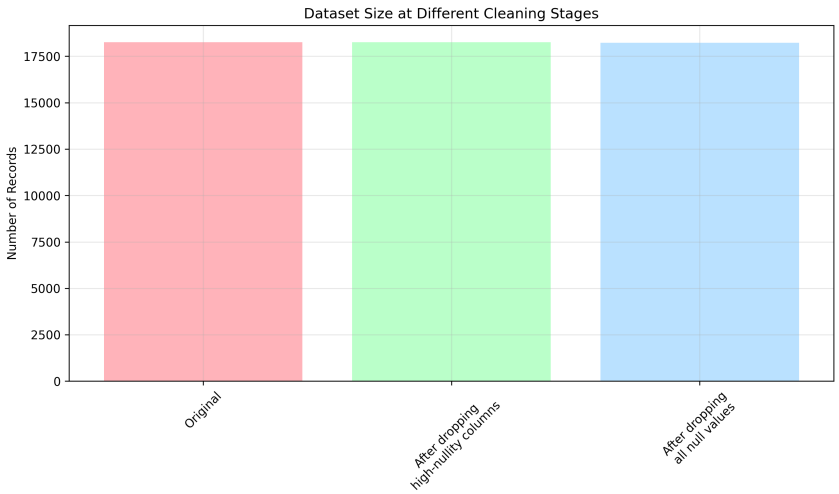

Figure 6: Dataset Size at Different Cleaning Stages

The dataset experienced a minor reduction in size after excluding rows with missing values. Since the percentage of missing entries was low, this approach preserved most of the data while ensuring cleaner input for modeling.

—

7 Question 7: Dropping Attributes with High Nullity

Problem Statement

Drop attributes with high nullity to facilitate efficient prediction.

Approach

The percentage of missing values was computed for each column. Columns with over 50% missing values were dropped. This threshold balances data preservation and model reliability by avoiding excessive imputation or information loss.

Results

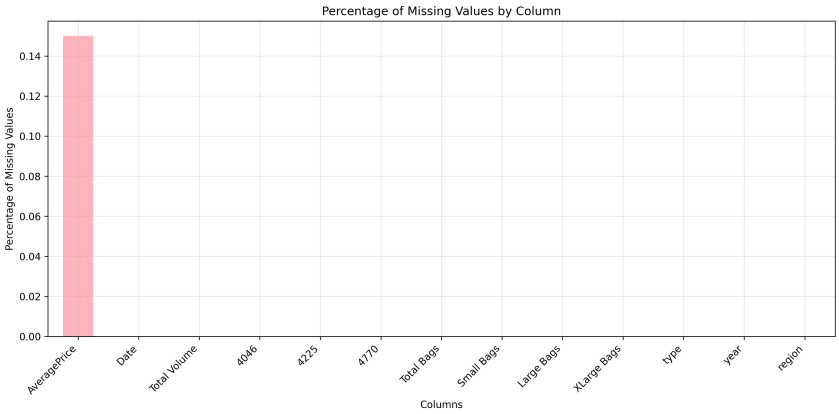

Figure 7: Percentage of Missing Values by Column

The nullity distribution plot reveals which columns had excessive missing values. Removing them simplifies the feature space, leading to faster and more stable model training. A cleaned dataset was saved as `q6_7_cleaned_data.csv`.

8 Question 8: Statistical Analysis of the Dataset

Problem Statement

Provide a complete statistical summary including dataset dimensions, most frequent values, datatype information, correlation matrix, skewness, etc.

Approach

Used descriptive statistics and visualization to explore data properties.

Results

Dataset Dimensions: (18250, 13)

Datatypes:

- Date: `object`
- AveragePrice: `object`
- Total Volume: `float64`
- 4046: `float64`
- 4225: `float64`
- 4770: `float64`
- Total Bags: `float64`

- Small Bags: `float64`
- Large Bags: `float64`
- XLarge Bags: `float64`
- type: `object`
- year: `int64`
- region: `object`

Most Frequent Values:

- Date: `18-03-2018`
- AveragePrice: `1.15`
- Total Volume: `2038.99`
- 4046: `0.0`
- 4225: `0.0`
- 4770: `0.0`
- Total Bags: `0.0`
- Small Bags: `0.0`
- Large Bags: `0.0`
- XLarge Bags: `0.0`
- type: `conventional`
- year: `2017`
- region: `Albany`

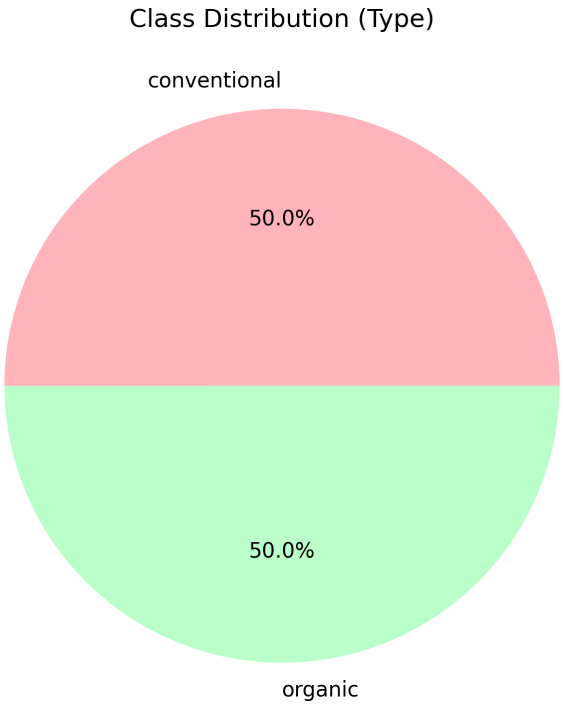

Figure 8: Class Distribution (Type)

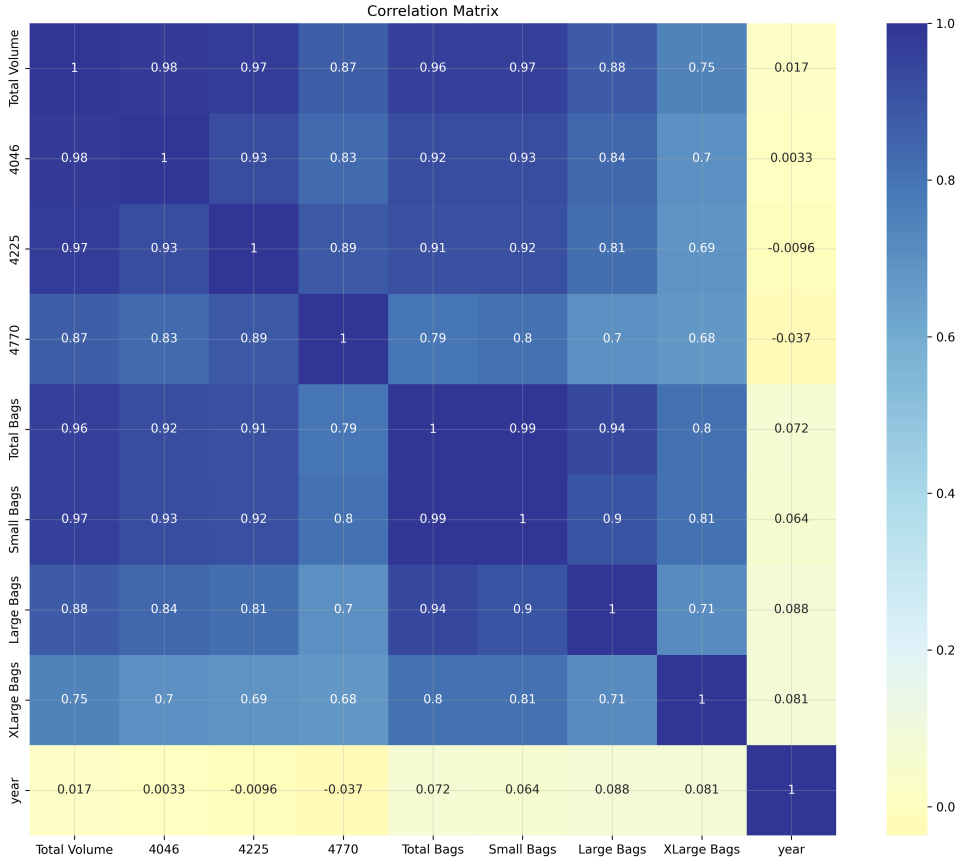

Figure 9: Correlation Matrix of Numerical Attributes

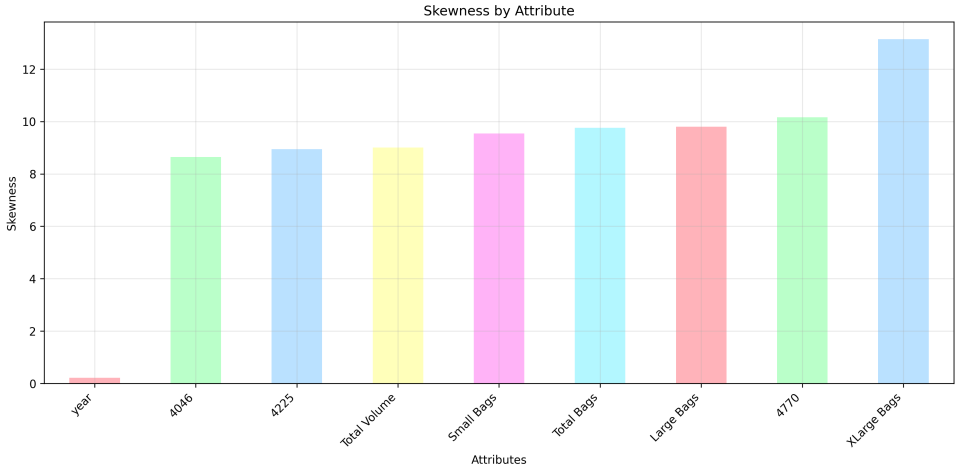

Figure 10: Skewness of Attributes

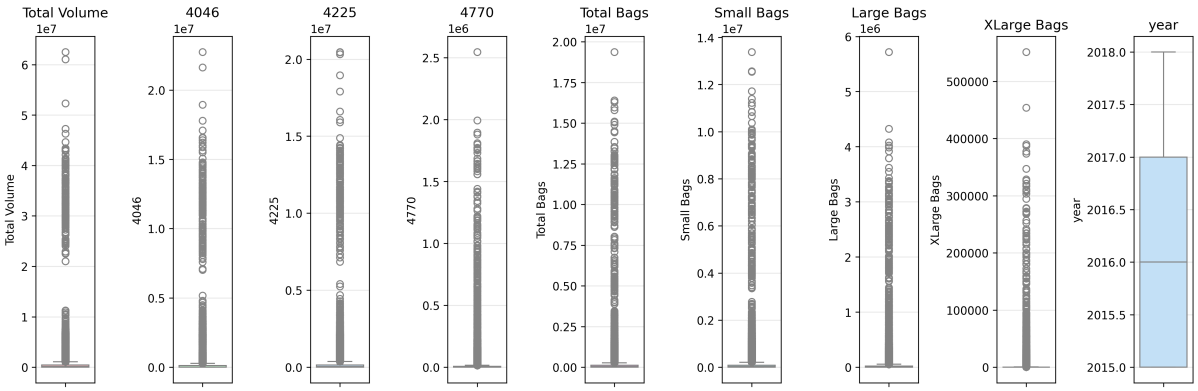

Figure 11: Box Plots for Numerical Attributes

9 Question 9: Feature Selection Measures

Problem Statement

Apply statistical and tree-based measures such as Gini Index, Entropy, and Information Gain for evaluating and selecting relevant features from the avocado dataset.

Approach

The feature selection process consisted of two parts:

- **Statistical Metrics:** Entropy, Gini Index, and Information Gain were calculated for each numeric feature based on discretized class bins (quintiles).
- **Tree-Based Importance:** A Decision Tree Classifier was trained on the data, and its intrinsic feature importance scores were analyzed.

Results

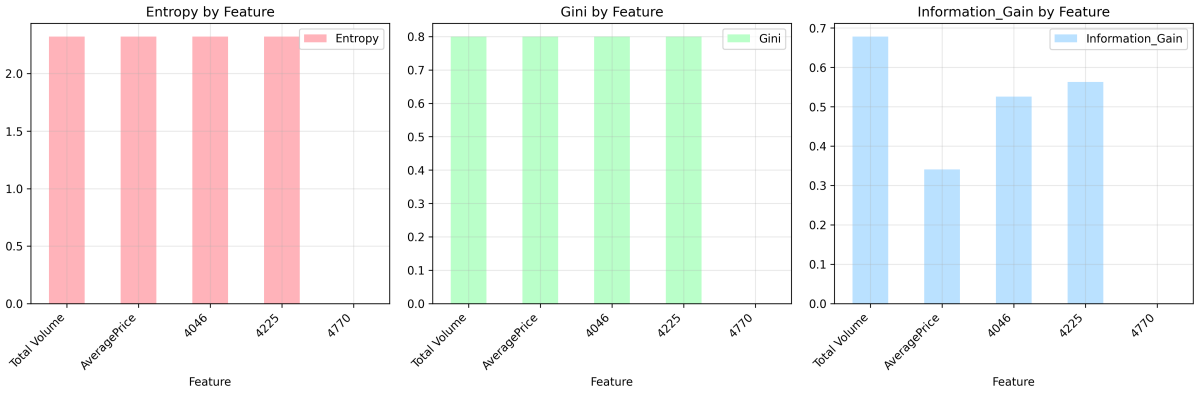

Figure 12: Feature Selection Measures (Entropy, Gini, Information Gain)

Decision Tree Feature Importance Analysis

- **Tree Parameters:**
  - `criterion`: gini `max_depth`: None `random_state`: 42
  - `min_samples_split`: 2 `min_samples_leaf`: 1 `ccp_alpha`: 0.0
- **Feature Importance Scores:**
  - `Total Volume`: 0.7395
  - `4770`: 0.1148
  - `4046`: 0.0550
  - `4225`: 0.0461
  - `AveragePrice`: 0.0447
- **Tree Properties:**
  - Number of nodes: 711
  - Tree depth: 20
- **Data Summary:**
  - Training samples: 18,250
  - Numeric features used: 5
  - Missing values (before imputation): 48 in `AveragePrice` (0.3%)

Feature Selection Measures – Detailed Calculations

Total Volume

- Entropy: 2.3219
- Gini Index: 0.8000

- Information Gain: 0.6783

AveragePrice

- Entropy: 2.3214
- Gini Index: 0.7999
- Information Gain: 0.3405

4046

- Entropy: 2.3219
- Gini Index: 0.8000
- Information Gain: 0.5260

4225

- Entropy: 2.3219
- Gini Index: 0.8000
- Information Gain: 0.5630

4770

- Entropy: 2.3219
- Gini Index: 0.8000
- Information Gain: 0.5630

10 Question 10: Comprehensive Preprocessing

Problem Statement

Implement and demonstrate a comprehensive preprocessing pipeline, including data cleaning, transformation, feature selection, and integration steps, applied to the avocado dataset.

Approach

The following phases were implemented:

- **Data Cleaning:** Removed duplicate entries and handled missing values using mean imputation for numerical columns and most frequent value imputation for categorical columns.
- **Data Transformation:** Applied standard scaling to numerical features and label encoding to categorical features to normalize the dataset for machine learning use.
- **Feature Selection:** Used `SelectKBest` with ANOVA F-score (`f_classif`) to identify the top 5 most relevant features in predicting the avocado type.

Results

- **Original Shape:** (18250, 13)
- **Cleaned Shape:** (18249, 13)
- **Selected Features:**
  - `AveragePrice`
  - `Total Volume`
  - `4046`
  - `4225`
  - `Small Bags`

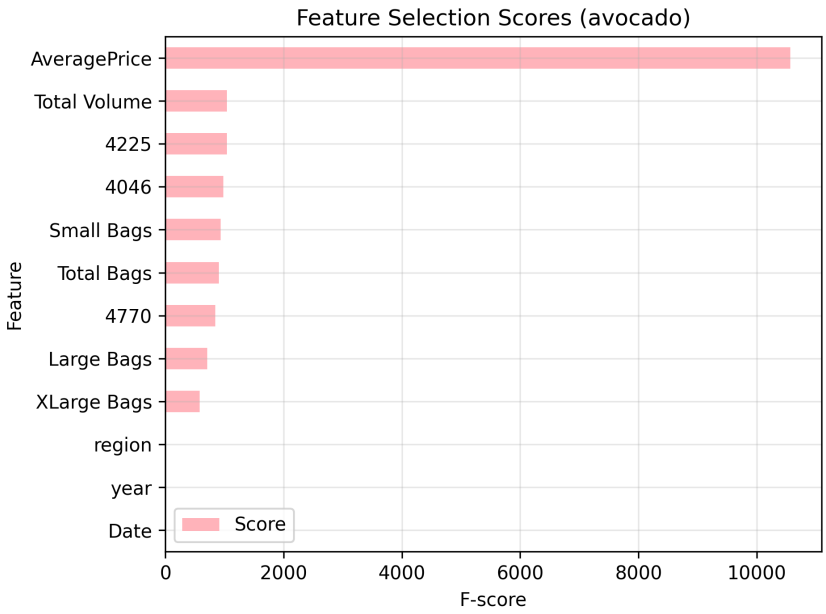

Figure 13: Feature Selection Scores for Avocado Dataset