========================================

步骤1: 数据加载与初始检查

========================================

原始数据加载成功! 尺寸: (1757, 26)

列名: City Name, Type, Package, Variety, Sub Variety, Grade, Date, Low Price, High Price, Mostly Low, Mostly High, Origin, Origin District, Item Size, Color, Environment, Unit of Sale, Quality, Condition, Appearance, Storage, Crop, Repack, Trans Mode, Unnamed: 24, Unnamed: 25

样例记录:

City Name Type Package Variety Sub Variety Grade Date Low Price High Price Mostly Low Mostly High Origin Origin District Item Size Color Environment Unit of Sale Quality Condition Appearance Storage Crop Repack Trans Mode Unnamed: 24 Unnamed: 25

BALTIMORE NaN 24 inch bins NaN NaN NaN 4/29/17 270.0 280.0 270.0 280.0 MARYLAND NaN lge NaN NaN NaN NaN NaN NaN NaN NaN E NaN NaN NaN

BALTIMORE NaN 24 inch bins NaN NaN NaN 5/6/17 270.0 280.0 270.0 280.0 MARYLAND NaN lge NaN NaN NaN NaN NaN NaN NaN NaN E NaN NaN NaN

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1757 entries, 0 to 1756

Data columns (total 26 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 City Name 1757 non-null object

1 Type 45 non-null object

2 Package 1757 non-null object

3 Variety 1752 non-null object

4 Sub Variety 296 non-null object

5 Grade 0 non-null float64

6 Date 1757 non-null object

7 Low Price 1757 non-null float64

8 High Price 1757 non-null float64

9 Mostly Low 1654 non-null float64

10 Mostly High 1654 non-null float64

11 Origin 1754 non-null object

12 Origin District 131 non-null object

13 Item Size 1478 non-null object

14 Color 1141 non-null object

15 Environment 0 non-null float64

16 Unit of Sale 162 non-null object

17 Quality 0 non-null float64

18 Condition 0 non-null float64

19 Appearance 0 non-null float64

20 Storage 0 non-null float64

21 Crop 0 non-null float64

22 Repack 1757 non-null object

23 Trans Mode 0 non-null float64

24 Unnamed: 24 0 non-null float64

25 Unnamed: 25 103 non-null object

dtypes: float64(13), object(13)

memory usage: 357.0+ KB

None

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1757 entries, 0 to 1756

Data columns (total 12 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 City Name 1757 non-null object

1 Package 1757 non-null object

2 Variety 1752 non-null object

3 Date 1757 non-null object

4 Low Price 1757 non-null float64

5 High Price 1757 non-null float64

6 Mostly Low 1654 non-null float64

7 Mostly High 1654 non-null float64

8 Origin 1754 non-null object

9 Item Size 1478 non-null object

10 Color 1141 non-null object

11 Repack 1757 non-null object

dtypes: float64(4), object(8)

memory usage: 164.8+ KB

None

========================================

步骤2: 数据预处理

========================================

检测到日期格式: 自动检测 | 样例: 4/29/17

日期处理完成 | 时间范围: 2014-11-29 - 2017-09-30 | 缺失率: 0.00%

价格处理完成 | 保留记录: 1757 | 平均价格: $128.77

========================================

改进的包装单位处理

========================================

包装标准化结果:

Standard\_Package

36 inch bins 663

24 inch bins 564

0.5 bu cartons 234

1.111 bu cartons 117

35 lb cartons 42

1.0 bu cartons 37

40 lb cartons 19

1.111 bu crates 17

each 17

50 lb sacks 11

50 lb cartons 10

1.0 bu baskets 10

22 lb cartons 10

20 lb cartons 6

Name: count, dtype: int64

蒲式耳当量分布:

mean 1.113300

std 0.353908

min 0.250000

max 1.500000

Name: Bushel\_Equivalent, dtype: float64

C:\Users\左百航\PycharmProjects\大三小学期\week4\Day3\data\_processing.py:73: UserWarning: Could not infer format, so each element will be parsed individually, falling back to `dateutil`. To ensure parsing is consistent and as-expected, please specify a format.

df['Date'] = pd.to\_datetime(df['Date'], format=date\_format, errors='coerce')

========================================

步骤3: 数据完整性验证

========================================

========================================

最终数据质量报告

========================================

数据集维度: (1757, 22)

时间范围: 0 2017-04-29

1 2017-05-06

2 2016-09-24

3 2016-09-24

4 2016-11-05

...

1752 2016-09-30

1753 2016-09-26

1754 2016-09-27

1755 2016-09-28

1756 2016-09-29

Name: Date, Length: 1757, dtype: datetime64[ns]

数值特征摘要:

Low Price High Price ... Bushel\_Equivalent Std\_Price

mean 124.571338 132.970939 ... 1.113300 112.213437

std 83.876591 89.524958 ... 0.353908 86.975122

[2 rows x 6 columns]

Item Size分布:

Item Size

M 971

S 363

L 295

Jumbo 128

Name: count, dtype: int64

City分布:

City

BOSTON 352

COLUMBIA 263

CHICAGO 248

FRANCISCO 166

BALTIMORE 153

DALLAS 137

YORK 112

LOUIS 103

ANGELES 62

ATLANTA 57

Name: count, dtype: int64

Origin Group分布:

Origin Group

Other 699

PENNSYLVANIA 250

MICHIGAN 237

CALIFORNIA 218

MASSACHUSETTS 197

ILLINOIS 156

Name: count, dtype: int64

Season分布:

Season

Fall 1683

Summer 30

Winter 29

Spring 15

Name: count, dtype: int64

Standard\_Package分布:

Standard\_Package

36 inch bins 663

24 inch bins 564

0.5 bu cartons 234

1.111 bu cartons 117

35 lb cartons 42

1.0 bu cartons 37

40 lb cartons 19

1.111 bu crates 17

each 17

50 lb sacks 11

Name: count, dtype: int64

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1757 entries, 0 to 1756

Data columns (total 22 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 City Name 1757 non-null object

1 Package 1757 non-null object

2 Variety 1752 non-null object

3 Date 1757 non-null datetime64[ns]

4 Low Price 1757 non-null float64

5 High Price 1757 non-null float64

6 Mostly Low 1654 non-null float64

7 Mostly High 1654 non-null float64

8 Origin 1754 non-null object

9 Item Size 1757 non-null object

10 Color 1141 non-null object

11 Repack 1757 non-null object

12 Month 1757 non-null int32

13 Year 1757 non-null int32

14 Day 1757 non-null int32

15 Season 1757 non-null object

16 Avg Price 1757 non-null float64

17 Standard\_Package 1757 non-null object

18 Bushel\_Equivalent 1757 non-null float64

19 Std\_Price 1757 non-null float64

20 City 1757 non-null object

21 Origin Group 1757 non-null object

dtypes: datetime64[ns](1), float64(7), int32(3), object(11)

memory usage: 281.5+ KB

None

【任务1】树结构：

|--- cat\_\_Variety\_MINIATURE <= 0.50

| |--- cat\_\_Standard\_Package\_1.111 bu cartons <= 0.50

| | |--- cat\_\_Standard\_Package\_35 lb cartons <= 0.50

| | | |--- cat\_\_Standard\_Package\_1.111 bu crates <= 0.50

| | | | |--- cat\_\_Origin Group\_MASSACHUSETTS <= 0.50

| | | | | |--- cat\_\_City\_ANGELES <= 0.50

| | | | | | |--- cat\_\_Variety\_HOWDEN TYPE <= 0.50

| | | | | | | |--- cat\_\_Standard\_Package\_1.0 bu cartons <= 0.50

| | | | | | | | |--- cat\_\_Variety\_BIG MACK TYPE <= 0.50

| | | | | | | | | |--- cat\_\_Standard\_Package\_50 lb sacks <= 0.50

| | | | | | | | | | |--- cat\_\_Standard\_Package\_50 lb cartons <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 20

| | | | | | | | | | |--- cat\_\_Standard\_Package\_50 lb cartons > 0.50

| | | | | | | | | | | |--- truncated branch of depth 4

| | | | | | | | | |--- cat\_\_Standard\_Package\_50 lb sacks > 0.50

| | | | | | | | | | |--- num\_\_Day <= 19.50

| | | | | | | | | | | |--- value: [15.00]

| | | | | | | | | | |--- num\_\_Day > 19.50

| | | | | | | | | | | |--- truncated branch of depth 3

| | | | | | | | |--- cat\_\_Variety\_BIG MACK TYPE > 0.50

| | | | | | | | | |--- cat\_\_Origin Group\_Other <= 0.50

| | | | | | | | | | |--- cat\_\_Standard\_Package\_36 inch bins <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | | |--- cat\_\_Standard\_Package\_36 inch bins > 0.50

| | | | | | | | | | | |--- value: [212.50]

| | | | | | | | | |--- cat\_\_Origin Group\_Other > 0.50

| | | | | | | | | | |--- cat\_\_City\_YORK <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 6

| | | | | | | | | | |--- cat\_\_City\_YORK > 0.50

| | | | | | | | | | | |--- value: [77.50]

| | | | | | | |--- cat\_\_Standard\_Package\_1.0 bu cartons > 0.50

| | | | | | | | |--- num\_\_Month <= 9.50

| | | | | | | | | |--- cat\_\_Origin Group\_Other <= 0.50

| | | | | | | | | | |--- value: [20.00]

| | | | | | | | | |--- cat\_\_Origin Group\_Other > 0.50

| | | | | | | | | | |--- value: [19.00]

| | | | | | | | |--- num\_\_Month > 9.50

| | | | | | | | | |--- value: [19.00]

| | | | | | |--- cat\_\_Variety\_HOWDEN TYPE > 0.50

| | | | | | | |--- cat\_\_City\_BOSTON <= 0.50

| | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | |--- cat\_\_City\_LOUIS <= 0.50

| | | | | | | | | | |--- cat\_\_City\_PHILADELPHIA <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 15

| | | | | | | | | | |--- cat\_\_City\_PHILADELPHIA > 0.50

| | | | | | | | | | | |--- truncated branch of depth 7

| | | | | | | | | |--- cat\_\_City\_LOUIS > 0.50

| | | | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 4

| | | | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins > 0.50

| | | | | | | | | | | |--- value: [105.00]

| | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | |--- cat\_\_Origin Group\_Other <= 0.50

| | | | | | | | | | |--- cat\_\_City\_ATLANTA <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 7

| | | | | | | | | | |--- cat\_\_City\_ATLANTA > 0.50

| | | | | | | | | | | |--- truncated branch of depth 3

| | | | | | | | | |--- cat\_\_Origin Group\_Other > 0.50

| | | | | | | | | | |--- cat\_\_City\_CHICAGO <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 7

| | | | | | | | | | |--- cat\_\_City\_CHICAGO > 0.50

| | | | | | | | | | | |--- value: [60.00]

| | | | | | | |--- cat\_\_City\_BOSTON > 0.50

| | | | | | | | |--- num\_\_Year <= 2016.50

| | | | | | | | | |--- cat\_\_Item Size\_S <= 0.50

| | | | | | | | | | |--- cat\_\_Origin Group\_PENNSYLVANIA <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 9

| | | | | | | | | | |--- cat\_\_Origin Group\_PENNSYLVANIA > 0.50

| | | | | | | | | | | |--- value: [150.00]

| | | | | | | | | |--- cat\_\_Item Size\_S > 0.50

| | | | | | | | | | |--- num\_\_Day <= 18.50

| | | | | | | | | | | |--- value: [150.00]

| | | | | | | | | | |--- num\_\_Day > 18.50

| | | | | | | | | | | |--- value: [145.00]

| | | | | | | | |--- num\_\_Year > 2016.50

| | | | | | | | | |--- cat\_\_Item Size\_Jumbo <= 0.50

| | | | | | | | | | |--- value: [187.50]

| | | | | | | | | |--- cat\_\_Item Size\_Jumbo > 0.50

| | | | | | | | | | |--- value: [212.50]

| | | | | |--- cat\_\_City\_ANGELES > 0.50

| | | | | | |--- cat\_\_Item Size\_L <= 0.50

| | | | | | | |--- cat\_\_Variety\_HOWDEN TYPE <= 0.50

| | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins <= 0.50

| | | | | | | | | |--- cat\_\_Variety\_CINDERELLA <= 0.50

| | | | | | | | | | |--- cat\_\_Variety\_BIG MACK TYPE <= 0.50

| | | | | | | | | | | |--- value: [0.50]

| | | | | | | | | | |--- cat\_\_Variety\_BIG MACK TYPE > 0.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | |--- cat\_\_Variety\_CINDERELLA > 0.50

| | | | | | | | | | |--- value: [0.30]

| | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins > 0.50

| | | | | | | | | |--- num\_\_Year <= 2016.50

| | | | | | | | | | |--- value: [0.24]

| | | | | | | | | |--- num\_\_Year > 2016.50

| | | | | | | | | | |--- value: [0.30]

| | | | | | | |--- cat\_\_Variety\_HOWDEN TYPE > 0.50

| | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins <= 0.50

| | | | | | | | | |--- value: [209.00]

| | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins > 0.50

| | | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | | |--- num\_\_Day <= 26.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | | |--- num\_\_Day > 26.50

| | | | | | | | | | | |--- value: [107.50]

| | | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | | |--- value: [85.00]

| | | | | | |--- cat\_\_Item Size\_L > 0.50

| | | | | | | |--- cat\_\_Standard\_Package\_each <= 0.50

| | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | |--- num\_\_Day <= 25.50

| | | | | | | | | | |--- value: [125.00]

| | | | | | | | | |--- num\_\_Day > 25.50

| | | | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins <= 0.50

| | | | | | | | | | | |--- value: [122.50]

| | | | | | | | | | |--- cat\_\_Standard\_Package\_24 inch bins > 0.50

| | | | | | | | | | | |--- value: [107.50]

| | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | |--- value: [85.00]

| | | | | | | |--- cat\_\_Standard\_Package\_each > 0.50

| | | | | | | | |--- value: [250.00]

| | | | |--- cat\_\_Origin Group\_MASSACHUSETTS > 0.50

| | | | | |--- cat\_\_Standard\_Package\_0.5 bu cartons <= 0.50

| | | | | | |--- cat\_\_Standard\_Package\_36 inch bins <= 0.50

| | | | | | | |--- cat\_\_Variety\_PIE TYPE <= 0.50

| | | | | | | | |--- value: [272.50]

| | | | | | | |--- cat\_\_Variety\_PIE TYPE > 0.50

| | | | | | | | |--- value: [162.50]

| | | | | | |--- cat\_\_Standard\_Package\_36 inch bins > 0.50

| | | | | | | |--- cat\_\_Variety\_PIE TYPE <= 0.50

| | | | | | | | |--- cat\_\_Item Size\_Jumbo <= 0.50

| | | | | | | | | |--- num\_\_Year <= 2016.50

| | | | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | | | |--- truncated branch of depth 3

| | | | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | |--- num\_\_Year > 2016.50

| | | | | | | | | | |--- num\_\_Day <= 12.50

| | | | | | | | | | | |--- value: [175.00]

| | | | | | | | | | |--- num\_\_Day > 12.50

| | | | | | | | | | | |--- value: [187.50]

| | | | | | | | |--- cat\_\_Item Size\_Jumbo > 0.50

| | | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | | |--- num\_\_Month <= 9.50

| | | | | | | | | | | |--- value: [212.50]

| | | | | | | | | | |--- num\_\_Month > 9.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | | |--- value: [202.50]

| | | | | | | |--- cat\_\_Variety\_PIE TYPE > 0.50

| | | | | | | | |--- cat\_\_Item Size\_S <= 0.50

| | | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | | |--- num\_\_Year <= 2016.50

| | | | | | | | | | | |--- value: [275.00]

| | | | | | | | | | |--- num\_\_Year > 2016.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | | |--- value: [210.00]

| | | | | | | | |--- cat\_\_Item Size\_S > 0.50

| | | | | | | | | |--- num\_\_Day <= 15.00

| | | | | | | | | | |--- value: [210.00]

| | | | | | | | | |--- num\_\_Day > 15.00

| | | | | | | | | | |--- num\_\_Month <= 9.50

| | | | | | | | | | | |--- value: [210.00]

| | | | | | | | | | |--- num\_\_Month > 9.50

| | | | | | | | | | | |--- value: [220.00]

| | | | | |--- cat\_\_Standard\_Package\_0.5 bu cartons > 0.50

| | | | | | |--- value: [15.50]

| | | |--- cat\_\_Standard\_Package\_1.111 bu crates > 0.50

| | | | |--- value: [15.50]

| | |--- cat\_\_Standard\_Package\_35 lb cartons > 0.50

| | | |--- num\_\_Day <= 2.50

| | | | |--- value: [36.00]

| | | |--- num\_\_Day > 2.50

| | | | |--- num\_\_Day <= 8.00

| | | | | |--- cat\_\_Season\_Fall <= 0.50

| | | | | | |--- num\_\_Year <= 2016.50

| | | | | | | |--- value: [24.00]

| | | | | | |--- num\_\_Year > 2016.50

| | | | | | | |--- value: [25.00]

| | | | | |--- cat\_\_Season\_Fall > 0.50

| | | | | | |--- value: [26.25]

| | | | |--- num\_\_Day > 8.00

| | | | | |--- num\_\_Day <= 12.00

| | | | | | |--- value: [36.00]

| | | | | |--- num\_\_Day > 12.00

| | | | | | |--- num\_\_Month <= 9.50

| | | | | | | |--- num\_\_Day <= 27.00

| | | | | | | | |--- num\_\_Day <= 23.50

| | | | | | | | | |--- value: [30.00]

| | | | | | | | |--- num\_\_Day > 23.50

| | | | | | | | | |--- value: [31.00]

| | | | | | | |--- num\_\_Day > 27.00

| | | | | | | | |--- value: [33.12]

| | | | | | |--- num\_\_Month > 9.50

| | | | | | | |--- num\_\_Day <= 24.00

| | | | | | | | |--- cat\_\_Origin Group\_CALIFORNIA <= 0.50

| | | | | | | | | |--- num\_\_Day <= 17.00

| | | | | | | | | | |--- value: [29.50]

| | | | | | | | | |--- num\_\_Day > 17.00

| | | | | | | | | | |--- value: [30.00]

| | | | | | | | |--- cat\_\_Origin Group\_CALIFORNIA > 0.50

| | | | | | | | | |--- num\_\_Day <= 18.50

| | | | | | | | | | |--- value: [29.75]

| | | | | | | | | |--- num\_\_Day > 18.50

| | | | | | | | | | |--- value: [28.00]

| | | | | | | |--- num\_\_Day > 24.00

| | | | | | | | |--- num\_\_Day <= 27.50

| | | | | | | | | |--- value: [24.00]

| | | | | | | | |--- num\_\_Day > 27.50

| | | | | | | | | |--- value: [28.00]

| |--- cat\_\_Standard\_Package\_1.111 bu cartons > 0.50

| | |--- cat\_\_City\_CHICAGO <= 0.50

| | | |--- num\_\_Year <= 2016.50

| | | | |--- cat\_\_City\_DETROIT <= 0.50

| | | | | |--- cat\_\_City\_PHILADELPHIA <= 0.50

| | | | | | |--- num\_\_Month <= 10.50

| | | | | | | |--- cat\_\_Origin Group\_PENNSYLVANIA <= 0.50

| | | | | | | | |--- num\_\_Day <= 11.50

| | | | | | | | | |--- cat\_\_Item Size\_S <= 0.50

| | | | | | | | | | |--- cat\_\_City\_BALTIMORE <= 0.50

| | | | | | | | | | | |--- value: [17.00]

| | | | | | | | | | |--- cat\_\_City\_BALTIMORE > 0.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | |--- cat\_\_Item Size\_S > 0.50

| | | | | | | | | | |--- value: [18.00]

| | | | | | | | |--- num\_\_Day > 11.50

| | | | | | | | | |--- cat\_\_City\_YORK <= 0.50

| | | | | | | | | | |--- num\_\_Month <= 9.50

| | | | | | | | | | | |--- value: [15.00]

| | | | | | | | | | |--- num\_\_Month > 9.50

| | | | | | | | | | | |--- truncated branch of depth 4

| | | | | | | | | |--- cat\_\_City\_YORK > 0.50

| | | | | | | | | | |--- value: [17.00]

| | | | | | | |--- cat\_\_Origin Group\_PENNSYLVANIA > 0.50

| | | | | | | | |--- num\_\_Day <= 25.50

| | | | | | | | | |--- value: [17.75]

| | | | | | | | |--- num\_\_Day > 25.50

| | | | | | | | | |--- value: [17.00]

| | | | | | |--- num\_\_Month > 10.50

| | | | | | | |--- cat\_\_Origin Group\_MICHIGAN <= 0.50

| | | | | | | | |--- value: [15.50]

| | | | | | | |--- cat\_\_Origin Group\_MICHIGAN > 0.50

| | | | | | | | |--- value: [14.75]

| | | | | |--- cat\_\_City\_PHILADELPHIA > 0.50

| | | | | | |--- num\_\_Day <= 18.50

| | | | | | | |--- value: [13.00]

| | | | | | |--- num\_\_Day > 18.50

| | | | | | | |--- value: [12.00]

| | | | |--- cat\_\_City\_DETROIT > 0.50

| | | | | |--- value: [12.00]

| | | |--- num\_\_Year > 2016.50

| | | | |--- cat\_\_Origin Group\_PENNSYLVANIA <= 0.50

| | | | | |--- num\_\_Day <= 19.50

| | | | | | |--- cat\_\_Item Size\_M <= 0.50

| | | | | | | |--- cat\_\_Origin Group\_MICHIGAN <= 0.50

| | | | | | | | |--- value: [22.00]

| | | | | | | |--- cat\_\_Origin Group\_MICHIGAN > 0.50

| | | | | | | | |--- value: [20.00]

| | | | | | |--- cat\_\_Item Size\_M > 0.50

| | | | | | | |--- value: [18.00]

| | | | | |--- num\_\_Day > 19.50

| | | | | | |--- cat\_\_Item Size\_S <= 0.50

| | | | | | | |--- cat\_\_City\_BALTIMORE <= 0.50

| | | | | | | | |--- cat\_\_Season\_Fall <= 0.50

| | | | | | | | | |--- value: [19.00]

| | | | | | | | |--- cat\_\_Season\_Fall > 0.50

| | | | | | | | | |--- value: [20.00]

| | | | | | | |--- cat\_\_City\_BALTIMORE > 0.50

| | | | | | | | |--- value: [18.00]

| | | | | | |--- cat\_\_Item Size\_S > 0.50

| | | | | | | |--- value: [15.00]

| | | | |--- cat\_\_Origin Group\_PENNSYLVANIA > 0.50

| | | | | |--- value: [16.00]

| | |--- cat\_\_City\_CHICAGO > 0.50

| | | |--- cat\_\_Origin Group\_Other <= 0.50

| | | | |--- num\_\_Year <= 2016.50

| | | | | |--- num\_\_Day <= 2.00

| | | | | | |--- value: [17.50]

| | | | | |--- num\_\_Day > 2.00

| | | | | | |--- num\_\_Month <= 10.50

| | | | | | | |--- num\_\_Day <= 18.50

| | | | | | | | |--- value: [17.50]

| | | | | | | |--- num\_\_Day > 18.50

| | | | | | | | |--- value: [17.00]

| | | | | | |--- num\_\_Month > 10.50

| | | | | | | |--- value: [17.00]

| | | | |--- num\_\_Year > 2016.50

| | | | | |--- cat\_\_Origin Group\_MICHIGAN <= 0.50

| | | | | | |--- cat\_\_Season\_Fall <= 0.50

| | | | | | | |--- value: [20.00]

| | | | | | |--- cat\_\_Season\_Fall > 0.50

| | | | | | | |--- num\_\_Day <= 12.50

| | | | | | | | |--- value: [20.00]

| | | | | | | |--- num\_\_Day > 12.50

| | | | | | | | |--- value: [19.00]

| | | | | |--- cat\_\_Origin Group\_MICHIGAN > 0.50

| | | | | | |--- value: [18.00]

| | | |--- cat\_\_Origin Group\_Other > 0.50

| | | | |--- num\_\_Day <= 20.50

| | | | | |--- num\_\_Day <= 10.00

| | | | | | |--- value: [18.38]

| | | | | |--- num\_\_Day > 10.00

| | | | | | |--- num\_\_Month <= 10.50

| | | | | | | |--- value: [20.00]

| | | | | | |--- num\_\_Month > 10.50

| | | | | | | |--- value: [19.75]

| | | | |--- num\_\_Day > 20.50

| | | | | |--- value: [22.50]

|--- cat\_\_Variety\_MINIATURE > 0.50

| |--- cat\_\_Standard\_Package\_36 inch bins <= 0.50

| | |--- cat\_\_Standard\_Package\_24 inch bins <= 0.50

| | | |--- cat\_\_Standard\_Package\_1.0 bu baskets <= 0.50

| | | | |--- cat\_\_Standard\_Package\_0.5 bu cartons <= 0.50

| | | | | |--- cat\_\_Standard\_Package\_22 lb cartons <= 0.50

| | | | | | |--- cat\_\_Standard\_Package\_1.111 bu crates <= 0.50

| | | | | | | |--- cat\_\_Standard\_Package\_35 lb cartons <= 0.50

| | | | | | | | |--- cat\_\_Item Size\_S <= 0.50

| | | | | | | | | |--- cat\_\_Origin Group\_CALIFORNIA <= 0.50

| | | | | | | | | | |--- value: [30.00]

| | | | | | | | | |--- cat\_\_Origin Group\_CALIFORNIA > 0.50

| | | | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | | | |--- value: [30.00]

| | | | | | | | |--- cat\_\_Item Size\_S > 0.50

| | | | | | | | | |--- value: [28.00]

| | | | | | | |--- cat\_\_Standard\_Package\_35 lb cartons > 0.50

| | | | | | | | |--- num\_\_Day <= 26.50

| | | | | | | | | |--- cat\_\_City\_FRANCISCO <= 0.50

| | | | | | | | | | |--- value: [24.50]

| | | | | | | | | |--- cat\_\_City\_FRANCISCO > 0.50

| | | | | | | | | | |--- value: [26.00]

| | | | | | | | |--- num\_\_Day > 26.50

| | | | | | | | | |--- cat\_\_City\_FRANCISCO <= 0.50

| | | | | | | | | | |--- value: [24.50]

| | | | | | | | | |--- cat\_\_City\_FRANCISCO > 0.50

| | | | | | | | | | |--- value: [24.05]

| | | | | | |--- cat\_\_Standard\_Package\_1.111 bu crates > 0.50

| | | | | | | |--- value: [18.00]

| | | | | |--- cat\_\_Standard\_Package\_22 lb cartons > 0.50

| | | | | | |--- num\_\_Day <= 28.50

| | | | | | | |--- num\_\_Day <= 27.50

| | | | | | | | |--- value: [16.75]

| | | | | | | |--- num\_\_Day > 27.50

| | | | | | | | |--- value: [14.75]

| | | | | | |--- num\_\_Day > 28.50

| | | | | | | |--- value: [12.08]

| | | | |--- cat\_\_Standard\_Package\_0.5 bu cartons > 0.50

| | | | | |--- cat\_\_City\_FRANCISCO <= 0.50

| | | | | | |--- cat\_\_City\_LOUIS <= 0.50

| | | | | | | |--- cat\_\_City\_BALTIMORE <= 0.50

| | | | | | | | |--- cat\_\_Origin Group\_MASSACHUSETTS <= 0.50

| | | | | | | | | |--- cat\_\_Item Size\_M <= 0.50

| | | | | | | | | | |--- num\_\_Year <= 2016.50

| | | | | | | | | | | |--- truncated branch of depth 7

| | | | | | | | | | |--- num\_\_Year > 2016.50

| | | | | | | | | | | |--- truncated branch of depth 5

| | | | | | | | | |--- cat\_\_Item Size\_M > 0.50

| | | | | | | | | | |--- cat\_\_Origin Group\_MICHIGAN <= 0.50

| | | | | | | | | | | |--- truncated branch of depth 6

| | | | | | | | | | |--- cat\_\_Origin Group\_MICHIGAN > 0.50

| | | | | | | | | | | |--- value: [21.00]

| | | | | | | | |--- cat\_\_Origin Group\_MASSACHUSETTS > 0.50

| | | | | | | | | |--- value: [16.50]

| | | | | | | |--- cat\_\_City\_BALTIMORE > 0.50

| | | | | | | | |--- num\_\_Day <= 23.00

| | | | | | | | | |--- num\_\_Day <= 18.50

| | | | | | | | | | |--- num\_\_Day <= 11.50

| | | | | | | | | | | |--- truncated branch of depth 2

| | | | | | | | | | |--- num\_\_Day > 11.50

| | | | | | | | | | | |--- value: [16.35]

| | | | | | | | | |--- num\_\_Day > 18.50

| | | | | | | | | | |--- value: [15.00]

| | | | | | | | |--- num\_\_Day > 23.00

| | | | | | | | | |--- num\_\_Month <= 9.50

| | | | | | | | | | |--- value: [18.00]

| | | | | | | | | |--- num\_\_Month > 9.50

| | | | | | | | | | |--- value: [17.00]

| | | | | | |--- cat\_\_City\_LOUIS > 0.50

| | | | | | | |--- cat\_\_Origin Group\_MICHIGAN <= 0.50

| | | | | | | | |--- cat\_\_Origin Group\_Other <= 0.50

| | | | | | | | | |--- num\_\_Day <= 29.50

| | | | | | | | | | |--- num\_\_Day <= 26.50

| | | | | | | | | | | |--- value: [12.50]

| | | | | | | | | | |--- num\_\_Day > 26.50

| | | | | | | | | | | |--- truncated branch of depth 3

| | | | | | | | | |--- num\_\_Day > 29.50

| | | | | | | | | | |--- value: [12.88]

| | | | | | | | |--- cat\_\_Origin Group\_Other > 0.50

| | | | | | | | | |--- value: [15.00]

| | | | | | | |--- cat\_\_Origin Group\_MICHIGAN > 0.50

| | | | | | | | |--- value: [12.00]

| | | | | |--- cat\_\_City\_FRANCISCO > 0.50

| | | | | | |--- num\_\_Month <= 9.50

| | | | | | | |--- value: [23.00]

| | | | | | |--- num\_\_Month > 9.50

| | | | | | | |--- num\_\_Day <= 6.50

| | | | | | | | |--- value: [20.00]

| | | | | | | |--- num\_\_Day > 6.50

| | | | | | | | |--- num\_\_Day <= 25.50

| | | | | | | | | |--- num\_\_Month <= 10.50

| | | | | | | | | | |--- value: [21.00]

| | | | | | | | | |--- num\_\_Month > 10.50

| | | | | | | | | | |--- value: [20.83]

| | | | | | | | |--- num\_\_Day > 25.50

| | | | | | | | | |--- value: [20.00]

| | | |--- cat\_\_Standard\_Package\_1.0 bu baskets > 0.50

| | | | |--- num\_\_Day <= 11.50

| | | | | |--- num\_\_Month <= 10.50

| | | | | | |--- value: [36.50]

| | | | | |--- num\_\_Month > 10.50

| | | | | | |--- value: [45.00]

| | | | |--- num\_\_Day > 11.50

| | | | | |--- num\_\_Day <= 22.50

| | | | | | |--- value: [46.50]

| | | | | |--- num\_\_Day > 22.50

| | | | | | |--- num\_\_Day <= 26.50

| | | | | | | |--- value: [49.17]

| | | | | | |--- num\_\_Day > 26.50

| | | | | | | |--- value: [47.00]

| | |--- cat\_\_Standard\_Package\_24 inch bins > 0.50

| | | |--- value: [400.00]

| |--- cat\_\_Standard\_Package\_36 inch bins > 0.50

| | |--- value: [275.00]

【任务2】分支解释：

当样本满足：

- feature\_2 ≤ 3.5，

- 并且 feature\_1 ≤ 2.1，

则该样本将被分配到此叶子节点，预测值为 10.5。

这个值是所有符合这两个条件的训练样本目标值的平均值。

【任务3】手动推导该棵树的生长过程所需信息如下：

原始数据集：

Variety Low Price High Price ... Std\_Price City Origin Group

2 HOWDEN TYPE 160.0 160.0 ... 160.0 BALTIMORE Other

3 HOWDEN TYPE 160.0 160.0 ... 160.0 BALTIMORE Other

4 HOWDEN TYPE 90.0 100.0 ... 95.0 BALTIMORE Other

5 HOWDEN TYPE 90.0 100.0 ... 95.0 BALTIMORE Other

6 HOWDEN TYPE 160.0 170.0 ... 110.0 BALTIMORE Other

[5 rows x 15 columns]

第一次分裂：cat\_\_City\_BALTIMORE <= 0.5

左子集样本数量（<= 0.5）: 1601

右子集样本数量（> 0.5）: 151

总方差：7455.15

左子集方差：7187.51

右子集方差：9582.98

加权方差：7393.97

方差减少量：61.19