Exploratory Data Analysis on Avocado data set

Exploratory Data Analysis on Avocado data set

The dataset contains 18,249 rows and 14 columns, including:

- **Date**: Date of observation.
- **average_price**: Price of a single avocado.
- **Total Volume**: Total volume of avocados sold.
- **4046, 4225, 4770**: PLU codes for different sizes of avocados.
- **Total Bags, Small Bags, Large Bags, XLarge Bags**: Number of avocados sold in bags of different sizes.
- **type**: Type of avocado (conventional or organic).
- **year**: Year of observation.
- **region**: Region where the data was recorded.

Load necessary libraries

Import Data

Understanding the Structure of the Data

```
tibble [18,249 × 13] (S3: tbl_df/tbl/data.frame)
 $ date         : Date[1:18249], format: "2015-12-27" "2015-12-20" ...
 $ average_price: num [1:18249] 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02
1.07 ...
 $ total_volume : num [1:18249] 64237 54877 118220 78992 51040 ...
 $ x4046        : num [1:18249] 1037 674 795 1132 941 ...
 $ x4225        : num [1:18249] 54455 44639 109150 71976 43838 ...
 $ x4770        : num [1:18249] 48.2 58.3 130.5 72.6 75.8 ...
 $ total_bags   : num [1:18249] 8697 9506 8145 5811 6184 ...
 $ small_bags   : num [1:18249] 8604 9408 8042 5677 5986 ...
 $ large_bags   : num [1:18249] 93.2 97.5 103.1 133.8 197.7 ...
 $ x_large_bags : num [1:18249] 0 0 0 0 0 0 0 0 0 0 ...
 $ type         : chr [1:18249] "conventional" "conventional" "conventional"
"conventional" ...
 $ year         : num [1:18249] 2015 2015 2015 2015 2015 ...
 $ region       : chr [1:18249] "Albany" "Albany" "Albany" "Albany" ...
```

Quick glimpse of data

```
Rows: 18,249
Columns: 13
$ date          <date> 2015-12-27, 2015-12-20, 2015-12-13, 2015-12-06, 2015-
11…
$ average_price <dbl> 1.33, 1.35, 0.93, 1.08, 1.28, 1.26, 0.99, 0.98, 1.02,
1.…
$ total_volume  <dbl> 64236.62, 54876.98, 118220.22, 78992.15, 51039.60,
55979…
$ x4046         <dbl> 1036.74, 674.28, 794.70, 1132.00, 941.48, 1184.27,
1368.…
$ x4225         <dbl> 54454.85, 44638.81, 109149.67, 71976.41, 43838.39,
48067…
$ x4770         <dbl> 48.16, 58.33, 130.50, 72.58, 75.78, 43.61, 93.26,
80.00,…
$ total_bags    <dbl> 8696.87, 9505.56, 8145.35, 5811.16, 6183.95, 6683.91,
83…
$ small_bags    <dbl> 8603.62, 9408.07, 8042.21, 5677.40, 5986.26, 6556.47,
81…
$ large_bags    <dbl> 93.25, 97.49, 103.14, 133.76, 197.69, 127.44, 122.05,
56…
$ x_large_bags  <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.…
$ type          <chr> "conventional", "conventional", "conventional",
"convent…
$ year          <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,
20…
$ region        <chr> "Albany", "Albany", "Albany", "Albany", "Albany",
"Alban…
```

Quick Summary Statistics

```
      date            average_price    total_volume         x4046
 Min.   :2015-01-04   Min.   :0.440   Min.   :      85   Min.   :       0
 1st Qu.:2015-10-25   1st Qu.:1.100   1st Qu.:   10839   1st Qu.:     854
 Median :2016-08-14   Median :1.370   Median :  107377   Median :    8645
 Mean   :2016-08-13   Mean   :1.406   Mean   :  850644   Mean   :  293008
 3rd Qu.:2017-06-04   3rd Qu.:1.660   3rd Qu.:  432962   3rd Qu.:  111020
 Max.   :2018-03-25   Max.   :3.250   Max.   :62505647   Max.   :22743616
     x4225              x4770           total_bags         small_bags
 Min.   :       0   Min.   :      0   Min.   :       0   Min.   :       0
 1st Qu.:    3009   1st Qu.:      0   1st Qu.:    5089   1st Qu.:    2849
 Median :   29061   Median :    185   Median :   39744   Median :   26363
 Mean   :  295155   Mean   :  22840   Mean   :  239639   Mean   :  182195
 3rd Qu.:  150207   3rd Qu.:   6243   3rd Qu.:  110783   3rd Qu.:   83338
 Max.   :20470573   Max.   :2546439   Max.   :19373134   Max.   :13384587
   large_bags       x_large_bags          type                   year
 Min.   :      0   Min.   :     0.0   Length:18249       Min.   :2015
 1st Qu.:    127   1st Qu.:     0.0   Class :character   1st Qu.:2015
 Median :   2648   Median :     0.0   Mode  :character   Median :2016
```

```
Mean   :  54338   Mean   :  3106.4                      Mean   :2016  
3rd Qu.:  22029   3rd Qu.:   132.5                      3rd Qu.:2017  
Max.   :5719097   Max.   :551693.7                      Max.   :2018  
   region         
Length:18249      
Class :character  
Mode  :character  
```

Skim a data frame, getting useful summary statistics

Data summary

Name	avocado_data
Number of rows	18249
Number of columns	13
_______________________	
Column type frequency:	
character	2
Date	1
numeric	10
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
type	0	1	7	12	0	2	0
region	0	1	4	19	0	54	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2015-01-04	2018-03-25	2016-08-14	169

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
average_price	0	1	1.41	0.40	0.44	1.10	1.37	1.66	3.25	▂▇▅▁▁
total_volume	0	1	850644.01	3453545.36	84.56	10838.58	107376.76	432962.29	62505646.52	▇▁▁▁▁
x4046	0	1	293008.42	1264989.08	0.00	854.07	8645.30	111020.20	22743616.17	▇▁▁▁▁
x4225	0	1	295154.57	1204120.40	0.00	3008.78	29061.02	150206.86	20470572.61	▇▁▁▁▁
x4770	0	1	22839.74	107464.07	0.00	0.00	184.99	6243.42	2546439.11	▇▁▁▁▁
total_bags	0	1	239639.20	986242.40	0.00	5088.64	39743.83	110783.37	19373134.37	▇▁▁▁▁
small_bags	0	1	182194.69	746178.51	0.00	2849.42	26362.82	83337.67	13384586.80	▇▁▁▁▁
large_bags	0	1	54338.09	243965.96	0.00	127.47	2647.71	22029.25	5719096.61	▇▁▁▁▁
x_large_bags	0	1	3106.43	17692.89	0.00	0.00	0.00	132.50	551693.65	▇▁▁▁▁
year	0	1	2016.15	0.94	2015.00	2015.00	2016.00	2017.00	2018.00	▇▇▁▇▂

Missing Values

```
        date average_price  total_volume         x4046         x4225 
           0             0             0             0             0 
       x4770    total_bags    small_bags    large_bags  x_large_bags 
           0             0             0             0             0 
        type          year        region 
           0             0             0 
```

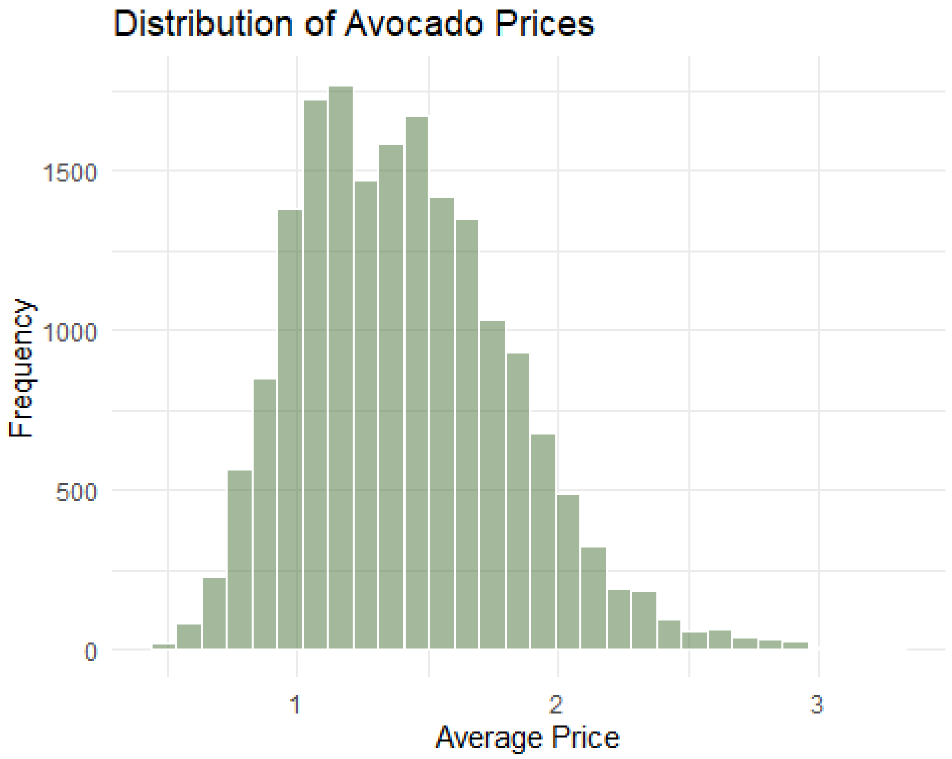
```
[1] 0
```

Exploratory Data Analysis Questions:

1. What is the distribution of avocado prices?

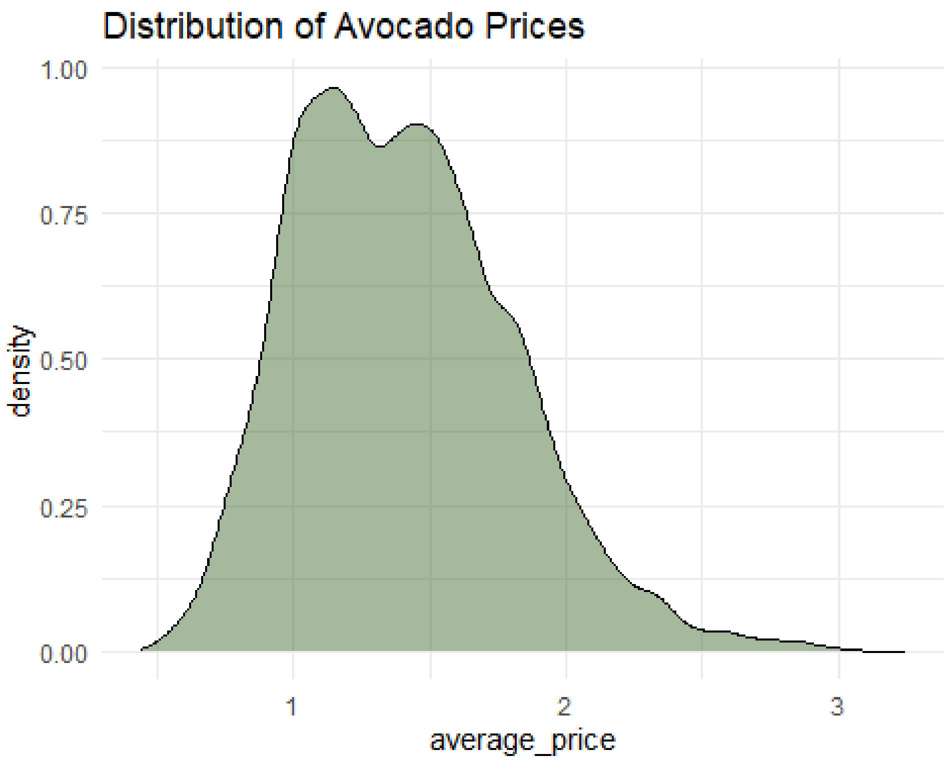
Option (a) - Histogram

- A histogram is a bar chart that groups data into bins, showing the frequency or count of values within each bin.

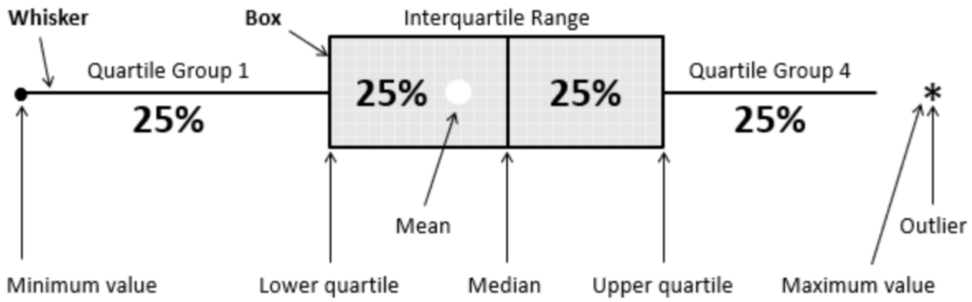


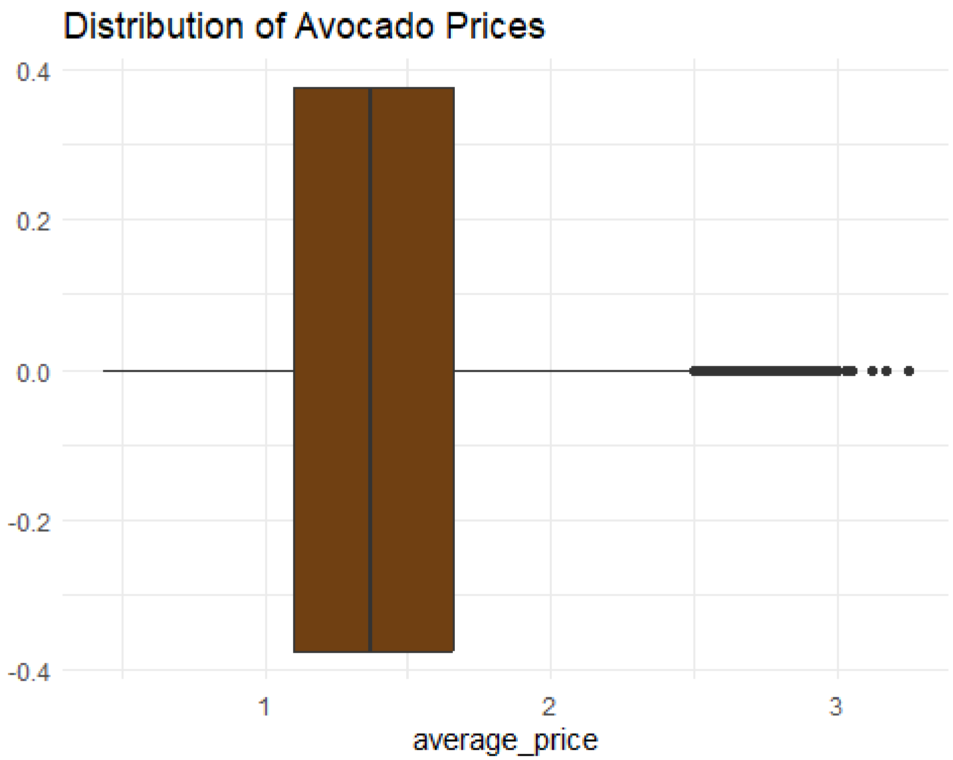
Option (b) - Density plot

- A density plot displays the proportion of data points within each range, providing a continuous and visually appealing estimate of the distribution, particularly useful for larger datasets. It uses a smooth curve to represent the data distribution.
- They are created using kernel density estimation (KDE), which smooths the data to show its underlying shape without the abrupt transitions seen in histograms.

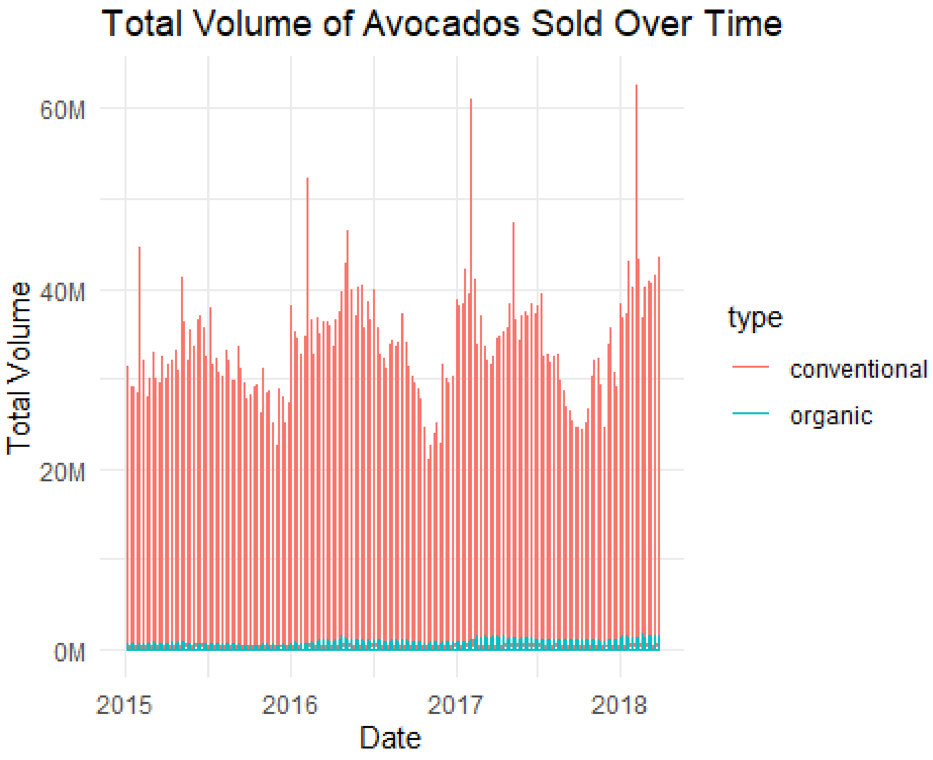

Distribution of Avocado Prices
density
1.00
0.75
0.50
0.25
0.00
1
2
3
average_price

Boxplot

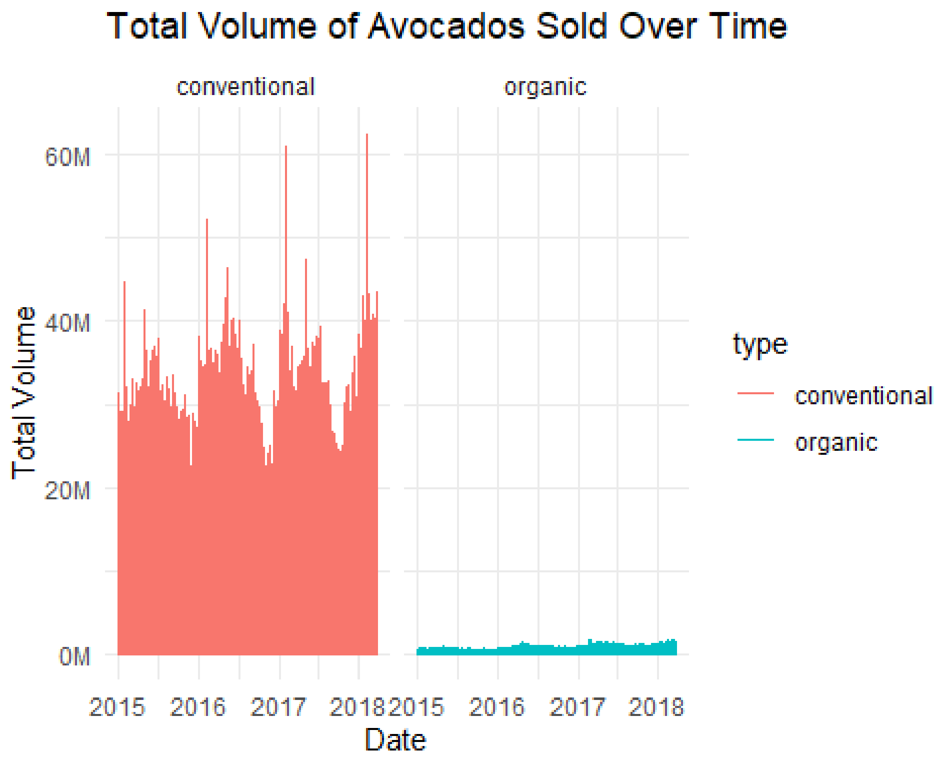

Whisker
Box
Interquartile Range
Quartile Group 1
25%
25%
25%
Quartile Group 4
25%
Mean
Outlier
Minimum value
Lower quartile
Median
Upper quartile
Maximum value



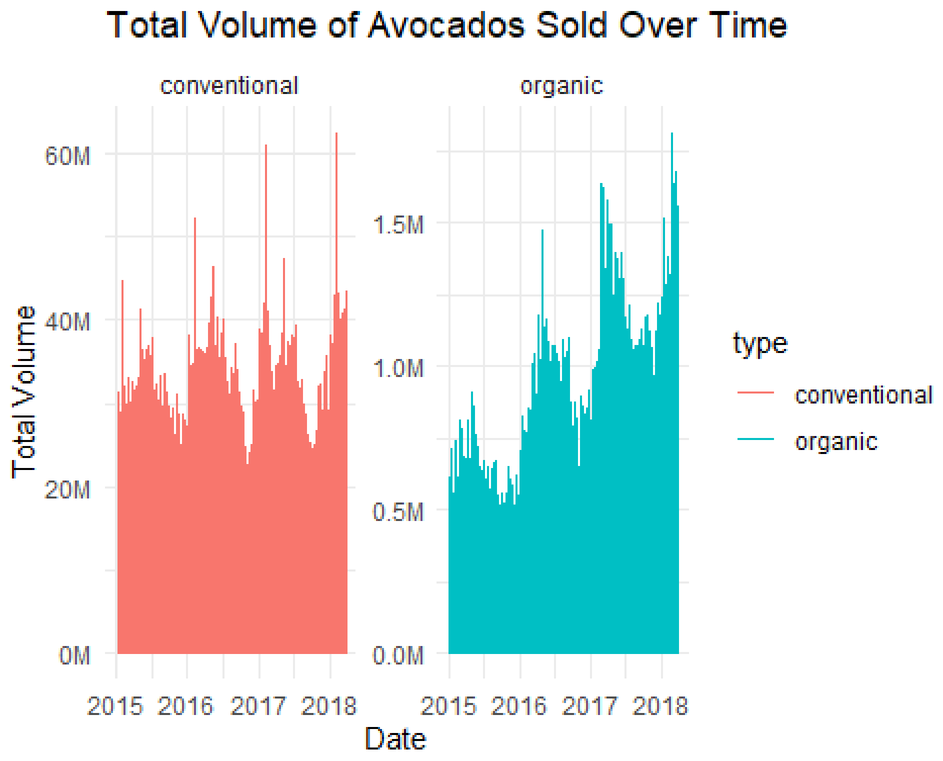
2. How does the total volume of avocados sold vary over time?



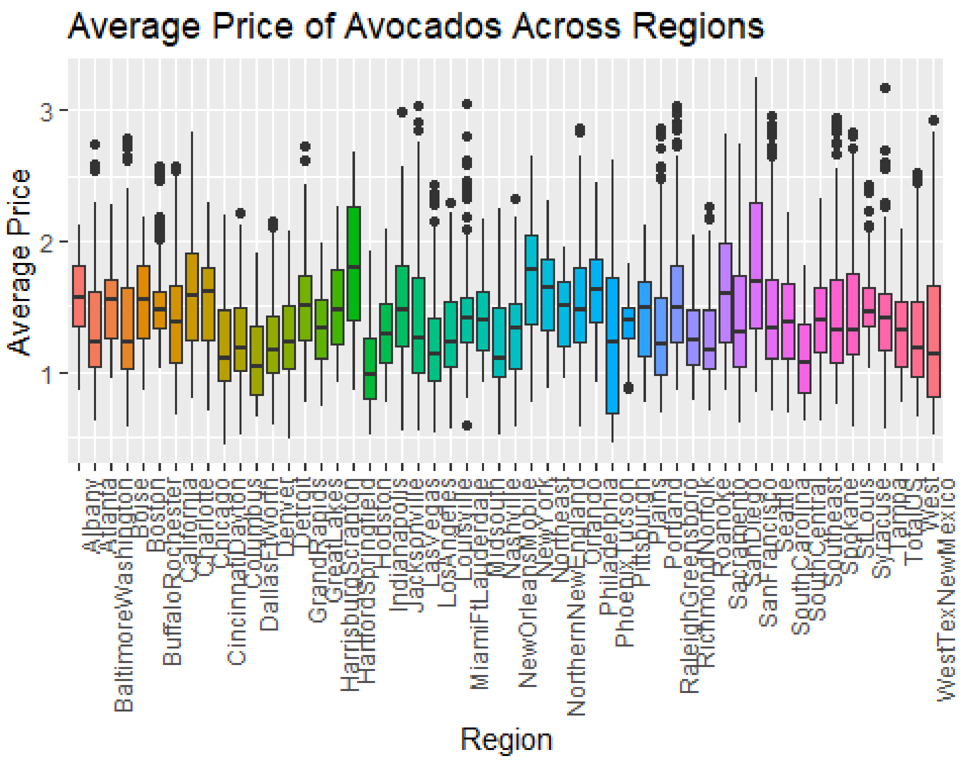
Break the display using facet_wrap



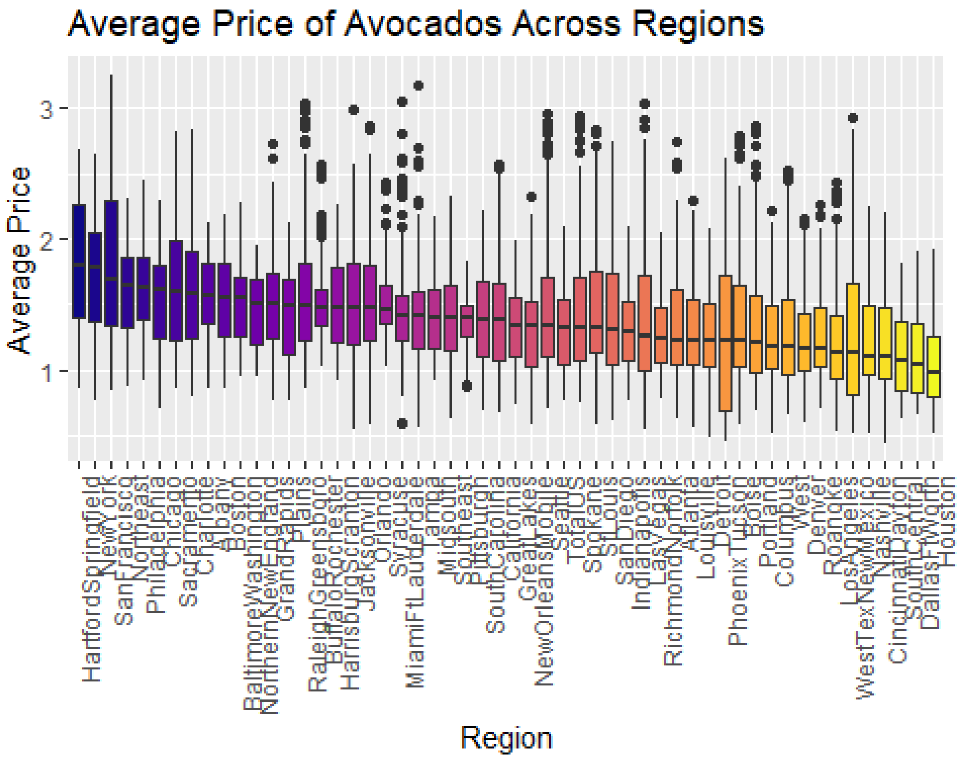
Free the y-axis for each facet



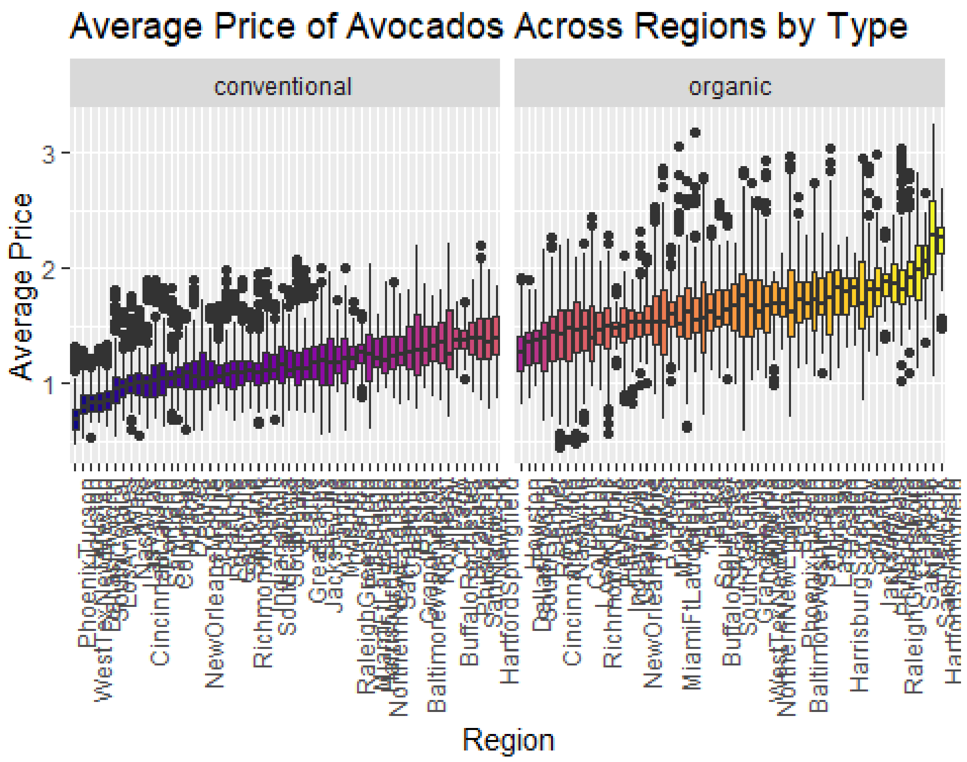
3. How do average prices vary across regions?



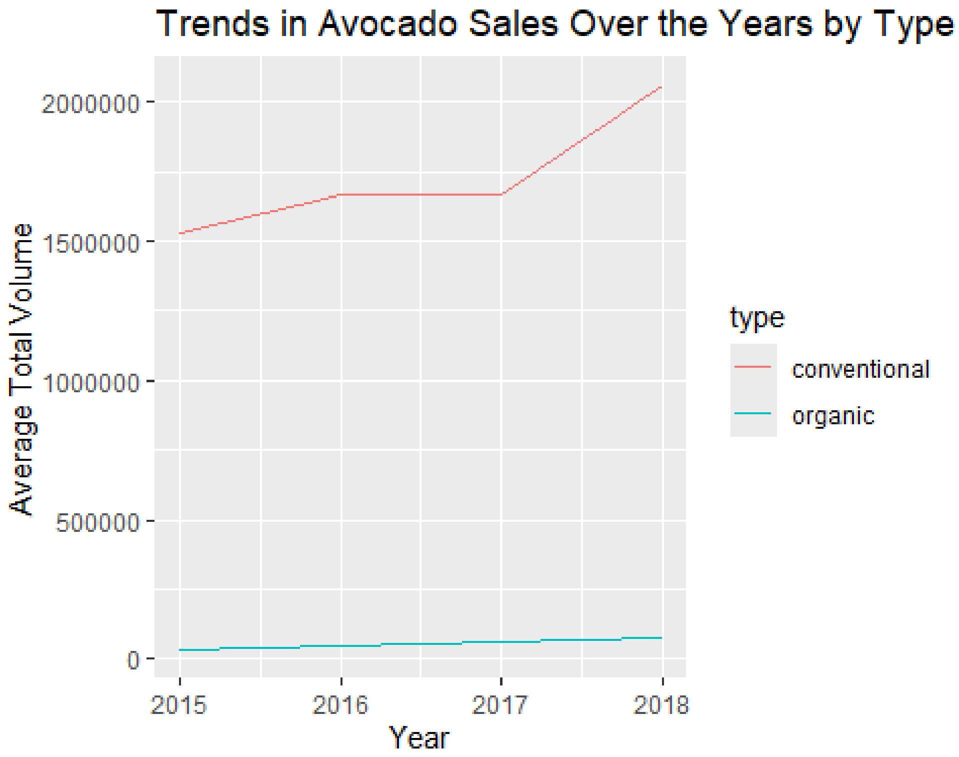
Arrange the box plots using the median of the average price for each region



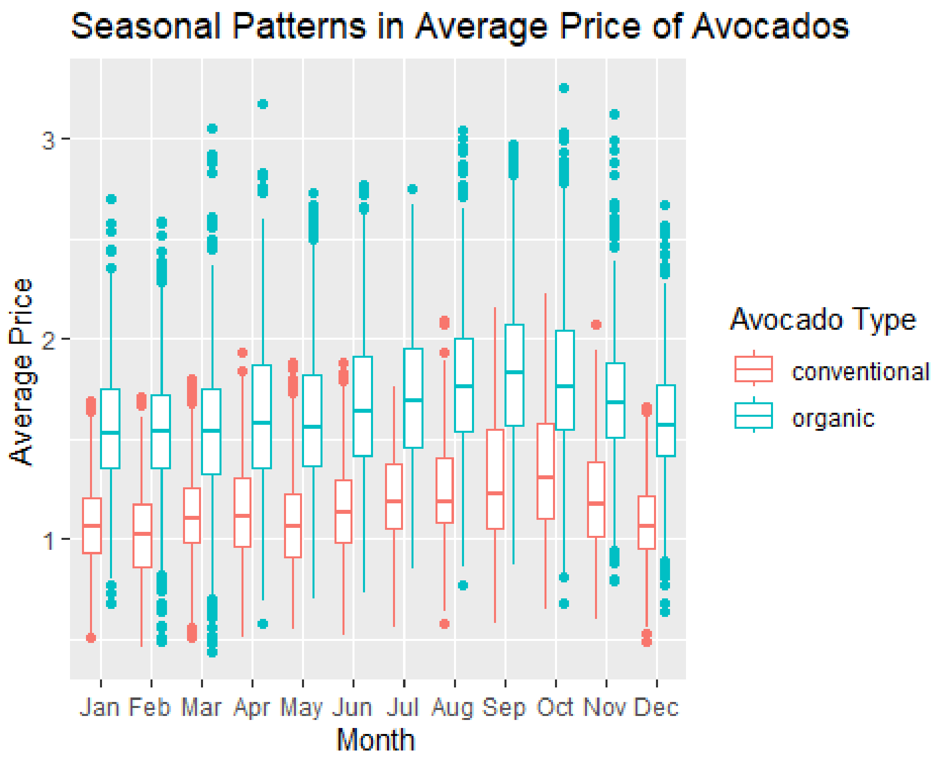
4. How do average prices vary across regions based on the type of avocado?



5. What are the trends in avocado sales over the years for each type?



6. Are there seasonal patterns in the average price of avocados?

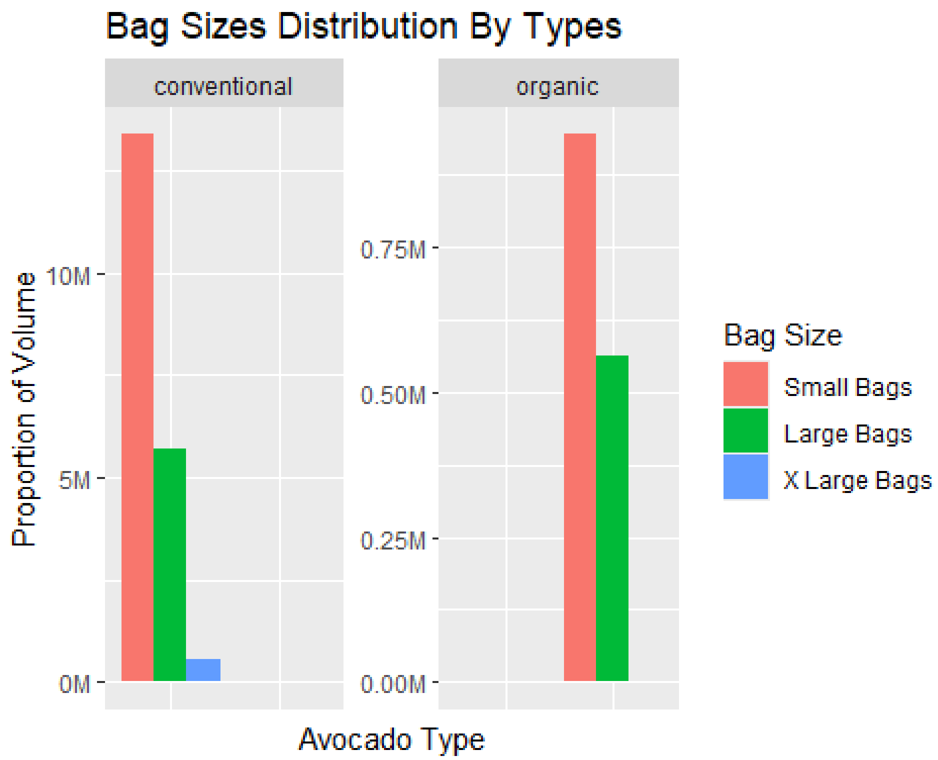


7. Which regions have the highest and lowest average avocado prices?

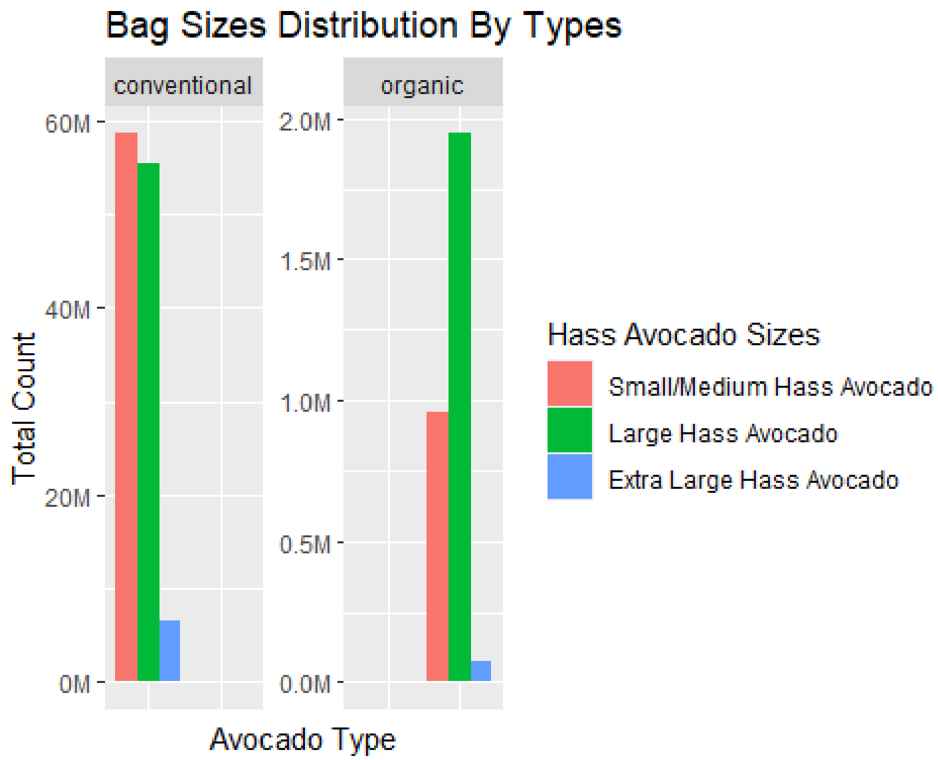
```
# A tibble: 5 × 2
  region                          avg_price
  <fct>                               <dbl>
1 HartfordSpringfield___organic        2.23
2 SanFrancisco___organic               2.21
3 NewYork___organic                    2.05
4 Sacramento___organic                 1.97
5 Charlotte___organic                  1.94

# A tibble: 5 × 2
  region                            avg_price
  <fct>                                 <dbl>
1 SouthCentral___conventional           0.869
2 DallasFtWorth___conventional          0.846
3 WestTexNewMexico___conventional       0.842
4 Houston___conventional                0.825
5 PhoenixTucson___conventional          0.728
```

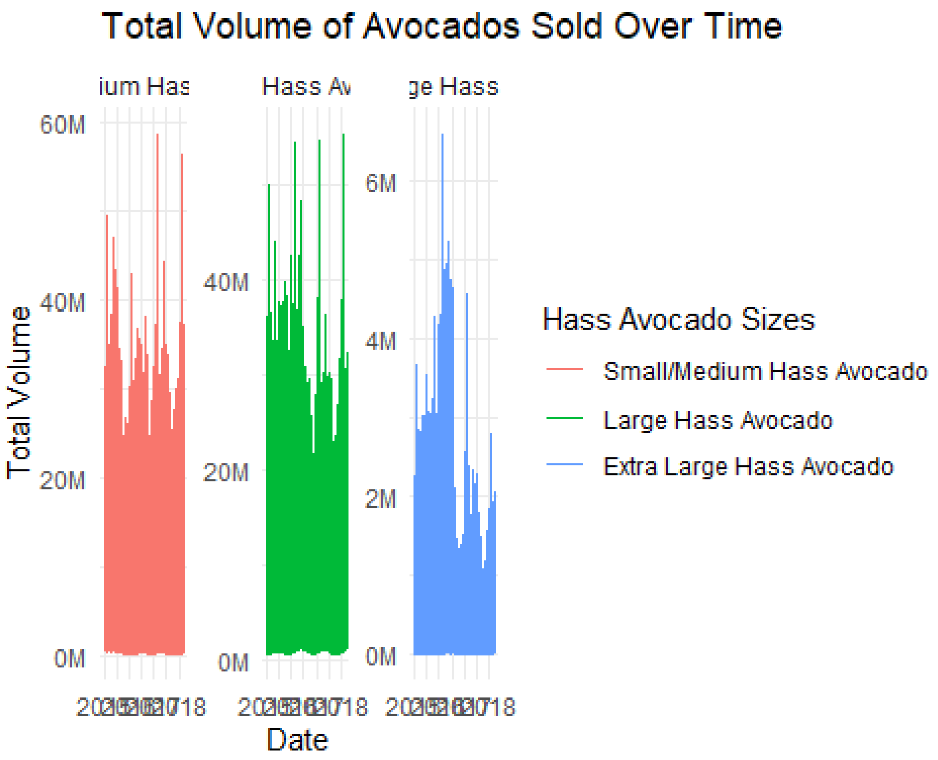
8. How are bag sizes (small, large, x-large) distributed by different type?



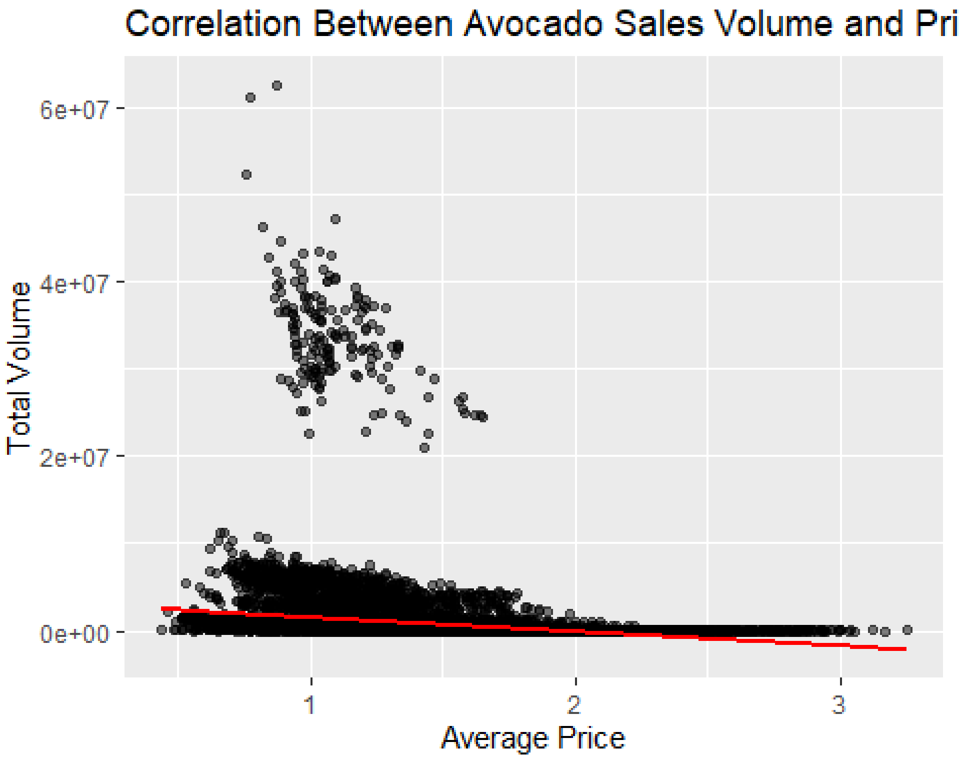
9. How are avocado sizes (x4046, x4225, x4770) distributed by different type?



10. How are total number of avocado sizes (x4046, x4225, x4770) sold over time?



11. Is there a correlation between avocado sales volumes and prices?


Correlation Between Avocado Sales Volume and Pri
6e+07
4e+07
2e+07
0e+00
Total Volume
Average Price