Avocados Price Report

Malay Raj

2023-05-20

## AVOCADOS PRICE ANALYSIS

: Historical data about avocado prices and sales in grocery stores throughout the United States. Avocado Prices (made available by Justin Kiggins under a CC0 license).

## Loading Required Packages.

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.1.8  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(lubridate)  
library(ggplot2)  
library(plyr)

## ------------------------------------------------------------------------------  
## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)  
## ------------------------------------------------------------------------------  
##   
## Attaching package: 'plyr'  
##   
## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize  
##   
## The following object is masked from 'package:purrr':  
##   
## compact

library(skimr)

## Warning: package 'skimr' was built under R version 4.2.3

## Displaying my current working directory.

getwd()

## [1] "C:/Avocado Report"

## Collecting the data.

Avocados = read.csv("C:/Avocado Report/avocado.csv")

## Displaying the first six rows of the dataset.

head(Avocados)

## X Date AveragePrice Total.Volume X4046 X4225 X4770 Total.Bags  
## 1 0 2015-12-27 1.33 64236.62 1036.74 54454.85 48.16 8696.87  
## 2 1 2015-12-20 1.35 54876.98 674.28 44638.81 58.33 9505.56  
## 3 2 2015-12-13 0.93 118220.22 794.70 109149.67 130.50 8145.35  
## 4 3 2015-12-06 1.08 78992.15 1132.00 71976.41 72.58 5811.16  
## 5 4 2015-11-29 1.28 51039.60 941.48 43838.39 75.78 6183.95  
## 6 5 2015-11-22 1.26 55979.78 1184.27 48067.99 43.61 6683.91  
## Small.Bags Large.Bags XLarge.Bags type year region  
## 1 8603.62 93.25 0 conventional 2015 Albany  
## 2 9408.07 97.49 0 conventional 2015 Albany  
## 3 8042.21 103.14 0 conventional 2015 Albany  
## 4 5677.40 133.76 0 conventional 2015 Albany  
## 5 5986.26 197.69 0 conventional 2015 Albany  
## 6 6556.47 127.44 0 conventional 2015 Albany

## Checking the number of rows and dimension of the dataset.

nrow(Avocados)

## [1] 18249

dim(Avocados)

## [1] 18249 14

## Renaming few columns.

Avocados=plyr::rename(Avocados, replace=c( "X4046" = "PLU4046",  
 "X4225" = "PLU4225",  
 "X4770" = "PLU4770"))

## Viewing the new column names.

colnames(Avocados)

## [1] "X" "Date" "AveragePrice" "Total.Volume" "PLU4046"   
## [6] "PLU4225" "PLU4770" "Total.Bags" "Small.Bags" "Large.Bags"   
## [11] "XLarge.Bags" "type" "year" "region"

## Inspecting the dataframe and look for inconguencies.

str(Avocados)

## 'data.frame': 18249 obs. of 14 variables:  
## $ X : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ Date : chr "2015-12-27" "2015-12-20" "2015-12-13" "2015-12-06" ...  
## $ AveragePrice: num 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...  
## $ Total.Volume: num 64237 54877 118220 78992 51040 ...  
## $ PLU4046 : num 1037 674 795 1132 941 ...  
## $ PLU4225 : num 54455 44639 109150 71976 43838 ...  
## $ PLU4770 : num 48.2 58.3 130.5 72.6 75.8 ...  
## $ Total.Bags : num 8697 9506 8145 5811 6184 ...  
## $ Small.Bags : num 8604 9408 8042 5677 5986 ...  
## $ Large.Bags : num 93.2 97.5 103.1 133.8 197.7 ...  
## $ XLarge.Bags : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ type : chr "conventional" "conventional" "conventional" "conventional" ...  
## $ year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ region : chr "Albany" "Albany" "Albany" "Albany" ...

# Convert Date and year to date and numeric datatypes so that they can stack correctly.

Avocados$Date = as.Date(Avocados$Date)  
Avocados$year = as.numeric(Avocados$year)

## Inspecting the dataframe after changing the datatypes.

str(Avocados)

## 'data.frame': 18249 obs. of 14 variables:  
## $ X : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ Date : Date, format: "2015-12-27" "2015-12-20" ...  
## $ AveragePrice: num 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...  
## $ Total.Volume: num 64237 54877 118220 78992 51040 ...  
## $ PLU4046 : num 1037 674 795 1132 941 ...  
## $ PLU4225 : num 54455 44639 109150 71976 43838 ...  
## $ PLU4770 : num 48.2 58.3 130.5 72.6 75.8 ...  
## $ Total.Bags : num 8697 9506 8145 5811 6184 ...  
## $ Small.Bags : num 8604 9408 8042 5677 5986 ...  
## $ Large.Bags : num 93.2 97.5 103.1 133.8 197.7 ...  
## $ XLarge.Bags : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ type : chr "conventional" "conventional" "conventional" "conventional" ...  
## $ year : num 2015 2015 2015 2015 2015 ...  
## $ region : chr "Albany" "Albany" "Albany" "Albany" ...

## Removing the ‘X’ column because we don’t need it in our analysis.

Avocados = subset(Avocados, select = -c(X))

## Viewing all the column left after removing the column ‘X’.

colnames(Avocados)

## [1] "Date" "AveragePrice" "Total.Volume" "PLU4046" "PLU4225"   
## [6] "PLU4770" "Total.Bags" "Small.Bags" "Large.Bags" "XLarge.Bags"   
## [11] "type" "year" "region"

## Checking for null values in our dataset.

sum(is.null(Avocados))

## [1] 0

## Checking for any duplicate values in our dataset.

anyDuplicated(Avocados)

## [1] 0

## Seeing how many observations fall under each type.

table(Avocados$type)

##   
## conventional organic   
## 9126 9123

## CONDUCTNG DESCRIPTIVE ANALYSIS:

Descriptive analysis on few columns.

summary(Avocados$AveragePrice)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.440 1.100 1.370 1.406 1.660 3.250

summary(Avocados$Total.Volume)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 85 10839 107377 850644 432962 62505647

summary(Avocados$PLU4046)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 854 8645 293008 111020 22743616

summary(Avocados$PLU4225)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 3009 29061 295155 150207 20470573

summary(Avocados$PLU4770)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 185 22840 6243 2546439

summary(Avocados$Total.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 5089 39744 239639 110783 19373134

summary(Avocados$Total.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 5089 39744 239639 110783 19373134

summary(Avocados$Small.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 2849 26363 182195 83338 13384587

summary(Avocados$Large.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 127 2648 54338 22029 5719097

summary(Avocados$XLarge.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 3106.4 132.5 551693.7

Comparing AveragePrice and different types.

aggregate(Avocados$AveragePrice ~ Avocados$type, FUN = mean)

## Avocados$type Avocados$AveragePrice  
## 1 conventional 1.158040  
## 2 organic 1.653999

aggregate(Avocados$AveragePrice ~ Avocados$type, FUN = median)

## Avocados$type Avocados$AveragePrice  
## 1 conventional 1.13  
## 2 organic 1.63

aggregate(Avocados$AveragePrice ~ Avocados$type, FUN = max)

## Avocados$type Avocados$AveragePrice  
## 1 conventional 2.22  
## 2 organic 3.25

aggregate(Avocados$AveragePrice ~ Avocados$type, FUN = min)

## Avocados$type Avocados$AveragePrice  
## 1 conventional 0.46  
## 2 organic 0.44

## SUMMARY STATISTICS.

skimr::skim\_without\_charts(Avocados)

Data summary

|  |  |
| --- | --- |
| Name | Avocados |
| 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**

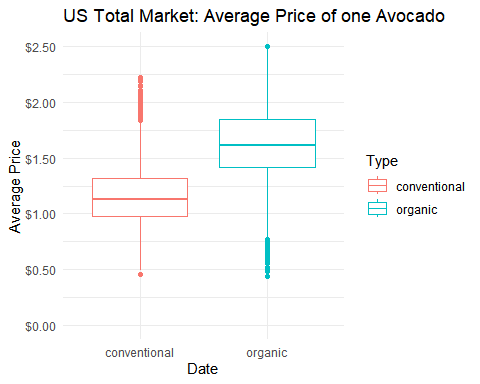
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AveragePrice | 0 | 1 | 1.41 | 0.40 | 0.44 | 1.10 | 1.37 | 1.66 | 3.25 |
| Total.Volume | 0 | 1 | 850644.01 | 3453545.36 | 84.56 | 10838.58 | 107376.76 | 432962.29 | 62505646.52 |
| PLU4046 | 0 | 1 | 293008.42 | 1264989.08 | 0.00 | 854.07 | 8645.30 | 111020.20 | 22743616.17 |
| PLU4225 | 0 | 1 | 295154.57 | 1204120.40 | 0.00 | 3008.78 | 29061.02 | 150206.86 | 20470572.61 |
| PLU4770 | 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 |
| XLarge.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 |

## DATA VISUALIZATION

Comparing the AveragePrices of different types.

ggplot(Avocados, aes(x = type,y = AveragePrice, color=type)) +   
 geom\_boxplot()+  
 scale\_y\_continuous(labels = scales::dollar,limits = c(0,2.5))+  
 labs(title="US Total Market: Average Price of one Avocado", x="Date",y="Average Price", color="Type") +   
 theme\_minimal()

## Warning: Removed 203 rows containing non-finite values (`stat\_boxplot()`).



Comparing the AveragePrice of different types of Avocados using histogram.

ggplot(Avocados, aes(x = AveragePrice, fill = type)) +   
 geom\_histogram(binwidth = 0.20, alpha=0.5)+  
 labs(title="Total US Market: Average Price of one Avocado", x="Average Price") +   
 theme\_minimal()

