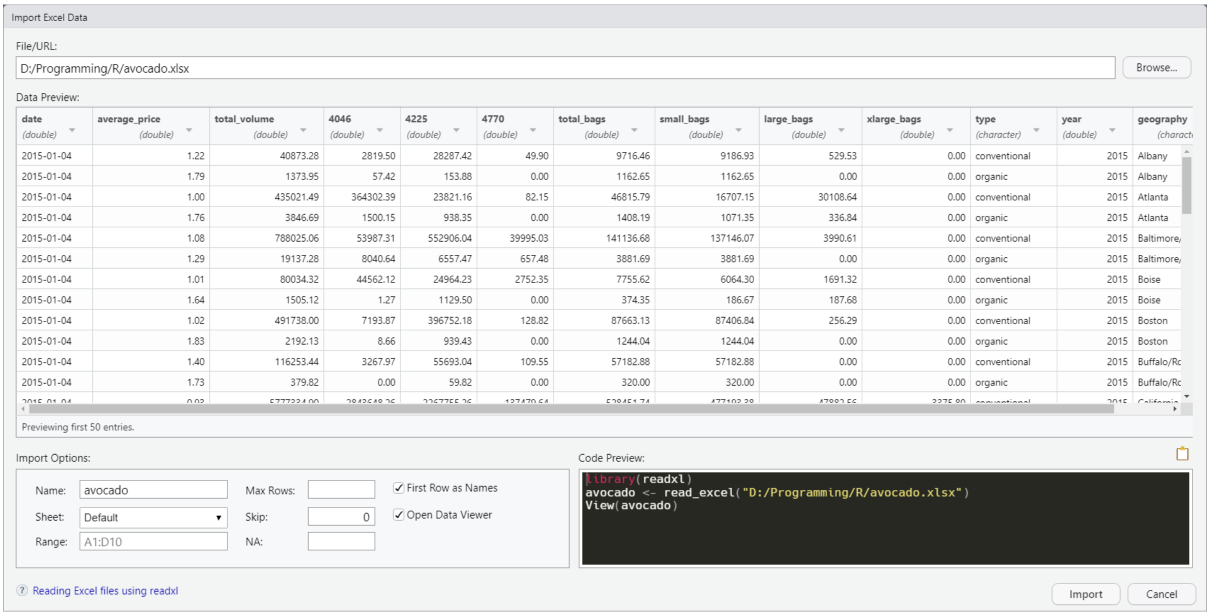
Avocado Prices Prediction

Approach:

a. Extracting data from a large Dataset

Data Cleansing: Data cleansing is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data is usually not necessary or helpful when it comes to analysing data because it may hinder the process or provide inaccurate results. This process identifies incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.



File/URL: D:/Programming/R/avocado.xlsx

Data Preview:

date (double)	average_price (double)	total_volume (double)	4046 (double)	4225 (double)	4770 (double)	total_bags (double)	small_bags (double)	large_bags (double)	xlarge_bags (double)	type (character)	year (double)	geography (charact
2015-01-04	1.22	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.00	conventional	2015	Albany
2015-01-04	1.79	1373.95	57.42	153.88	0.00	1162.65	1162.65	0.00	0.00	organic	2015	Albany
2015-01-04	1.00	435021.49	364302.39	23821.16	82.15	46815.79	16707.15	30108.64	0.00	conventional	2015	Atlanta
2015-01-04	1.76	3846.69	1500.15	938.35	0.00	1408.19	1071.35	336.84	0.00	organic	2015	Atlanta
2015-01-04	1.08	788025.06	53987.31	552906.04	39995.03	141136.68	137146.07	3990.61	0.00	conventional	2015	Baltimore/
2015-01-04	1.29	19137.28	8040.64	6557.47	657.48	3881.69	3881.69	0.00	0.00	organic	2015	Baltimore/
2015-01-04	1.01	80034.32	44562.12	24964.23	2752.35	7755.62	6064.30	1691.32	0.00	conventional	2015	Boise
2015-01-04	1.64	1505.12	1.27	1129.50	0.00	374.35	186.67	187.68	0.00	organic	2015	Boise
2015-01-04	1.02	491738.00	7193.87	396752.18	128.82	87663.13	87406.84	256.29	0.00	conventional	2015	Boston
2015-01-04	1.83	2192.13	8.66	939.43	0.00	1244.04	1244.04	0.00	0.00	organic	2015	Boston
2015-01-04	1.40	116253.44	3267.97	55693.04	109.55	57182.88	57182.88	0.00	0.00	conventional	2015	Buffalo/Rc
2015-01-04	1.73	379.82	0.00	59.82	0.00	320.00	320.00	0.00	0.00	organic	2015	Buffalo/Rc

Previewing first 50 entries.

Import Options:

Name: avocado

Sheet: Default

Range: A1:D10

Max Rows:

Skip: 0

NA:

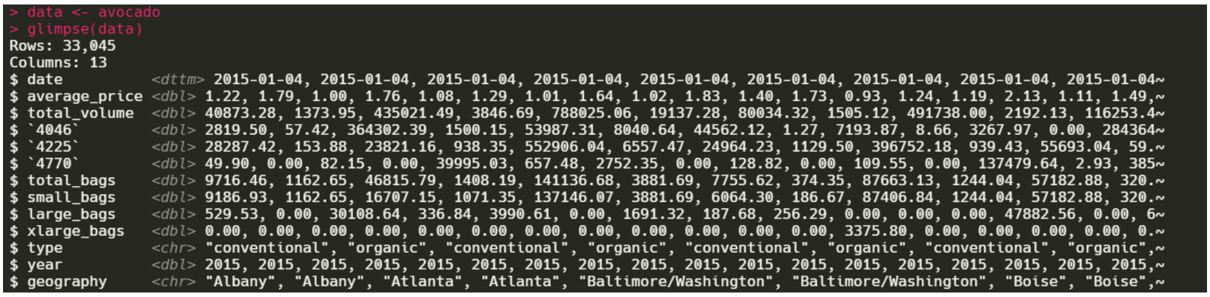
☑ First Row as Names

☑ Open Data Viewer

Code Preview:

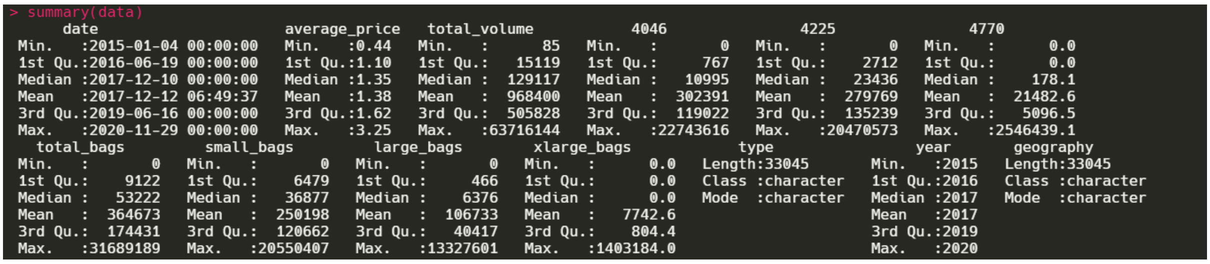
```
library(readxl)
avocado <- read_excel("D:/Programming/R/avocado.xlsx")
View(avocado)
```

Fig 1.1: Importing dataset in RStudio.



```
> data <- avocado
> glimpse(data)
Rows: 33,045
Columns: 13
$ date          <dttm> 2015-01-04, 2015-01-04, 2015-01-04, 2015-01-04, 2015-01-04, 2015-01-04, 2015-01-04, 2015-01-04, 2015-01-04~
$ average_price <dbl> 1.22, 1.79, 1.00, 1.76, 1.08, 1.29, 1.01, 1.64, 1.02, 1.83, 1.40, 1.73, 0.93, 1.24, 1.19, 2.13, 1.11, 1.49,~
$ total_volume  <dbl> 40873.28, 1373.95, 435021.49, 3846.69, 788025.06, 19137.28, 80034.32, 1505.12, 491738.00, 2192.13, 116253.4~
$ `4046`        <dbl> 2819.50, 57.42, 364302.39, 1500.15, 53987.31, 8040.64, 44562.12, 1.27, 7193.87, 8.66, 3267.97, 0.00, 284364~
$ `4225`        <dbl> 28287.42, 153.88, 23821.16, 938.35, 552906.04, 6557.47, 24964.23, 1129.50, 396752.18, 939.43, 55693.04, 59.~
$ `4770`        <dbl> 49.90, 0.00, 82.15, 0.00, 39995.03, 657.48, 2752.35, 0.00, 128.82, 0.00, 109.55, 0.00, 137479.64, 2.93, 385~
$ total_bags    <dbl> 9716.46, 1162.65, 46815.79, 1408.19, 141136.68, 3881.69, 7755.62, 374.35, 87663.13, 1244.04, 57182.88, 320.~
$ small_bags    <dbl> 9186.93, 1162.65, 16707.15, 1071.35, 137146.07, 3881.69, 6064.30, 186.67, 87406.84, 1244.04, 57182.88, 320.~
$ large_bags    <dbl> 529.53, 0.00, 30108.64, 336.84, 3990.61, 0.00, 1691.32, 187.68, 256.29, 0.00, 0.00, 0.00, 47882.56, 0.00, 6~
$ xlarge_bags   <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 3375.80, 0.00, 0.00, 0.00, 0.00, 0.~
$ type          <chr> "conventional", "organic", "conventional", "organic", "conventional", "organic", "conventional", "organic",~
$ year          <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,~
$ geography     <chr> "Albany", "Albany", "Atlanta", "Atlanta", "Baltimore/Washington", "Baltimore/Washington", "Boise", "Boise",~
```

Fig 1.2: Glimpse of the data.



```
> summary(data)
      date                         average_price   total_volume           4046               4225               4770          
 Min.   :2015-01-04 00:00:00   Min.   :0.44   Min.   :      85   Min.   :       0   Min.   :       0   Min.   :      0.0  
 1st Qu.:2016-06-19 00:00:00   1st Qu.:1.10   1st Qu.:   15119   1st Qu.:     767   1st Qu.:    2712   1st Qu.:      0.0  
 Median :2017-12-10 00:00:00   Median :1.35   Median :  129117   Median :   10995   Median :   23436   Median :    178.1  
 Mean   :2017-12-12 06:49:37   Mean   :1.38   Mean   :  968400   Mean   :  302391   Mean   :  279769   Mean   :  21482.6  
 3rd Qu.:2019-06-16 00:00:00   3rd Qu.:1.62   3rd Qu.:  505828   3rd Qu.:  119022   3rd Qu.:  135239   3rd Qu.:   5096.5  
 Max.   :2020-11-29 00:00:00   Max.   :3.25   Max.   :63716144   Max.   :22743616   Max.   :20470573   Max.   :2546439.1  
   total_bags         small_bags         large_bags        xlarge_bags             type                year        geography        
 Min.   :       0   Min.   :       0   Min.   :       0   Min.   :      0.0   Length:33045       Min.   :2015   Length:33045      
 1st Qu.:    9122   1st Qu.:    6479   1st Qu.:     466   1st Qu.:      0.0   Class :character   1st Qu.:2016   Class :character  
 Median :   53222   Median :   36877   Median :    6376   Median :      0.0   Mode  :character   Median :2017   Mode  :character  
 Mean   :  364673   Mean   :  250198   Mean   :  106733   Mean   :   7742.6                      Mean   :2017                     
 3rd Qu.:  174431   3rd Qu.:  120662   3rd Qu.:   40417   3rd Qu.:    804.4                      3rd Qu.:2019                     
 Max.   :31689189   Max.   :20550407   Max.   :13327601   Max.   :1403184.0                      Max.   :2020                     
```

Fig 1.3: Summary of the attributes.

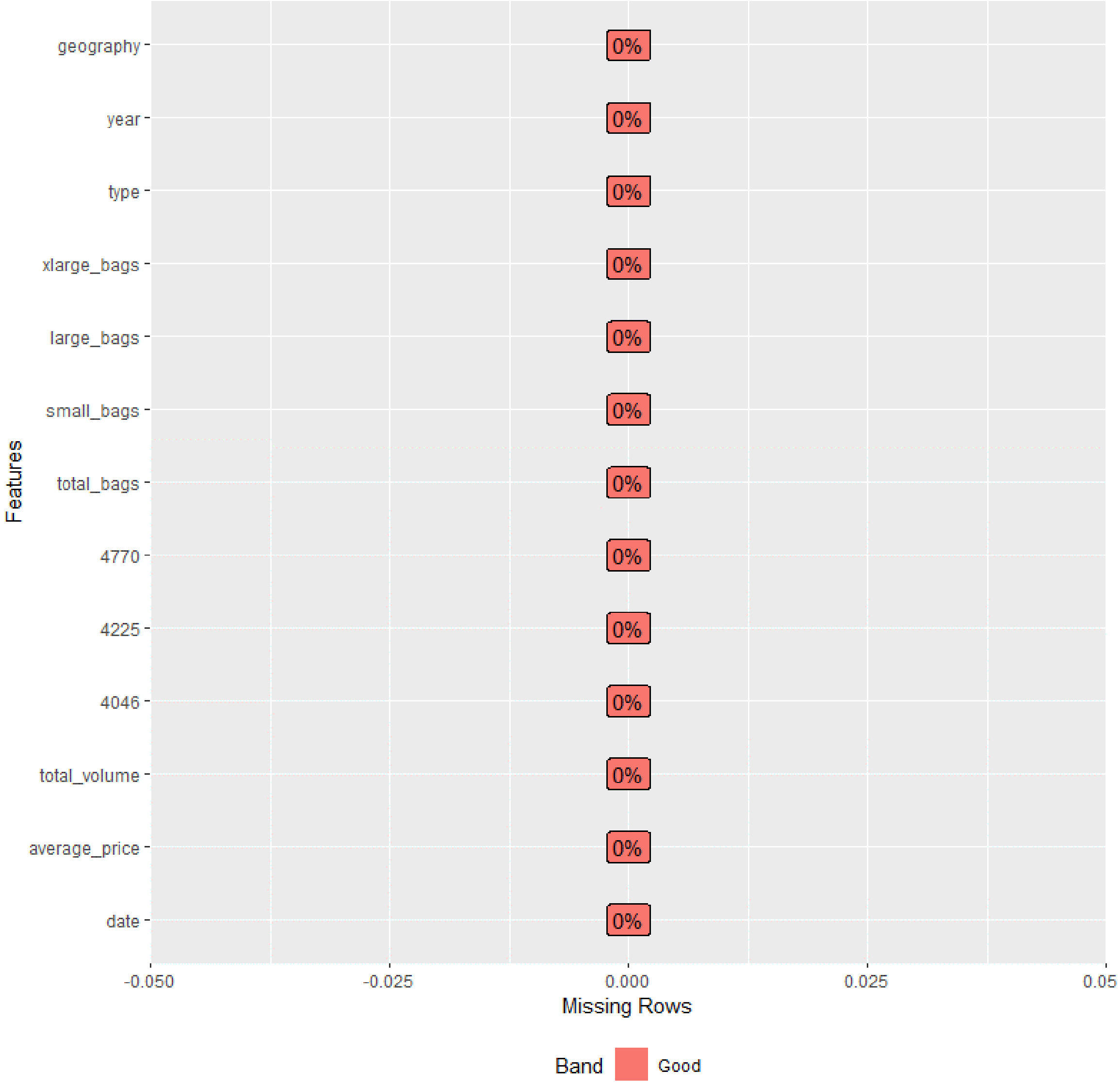
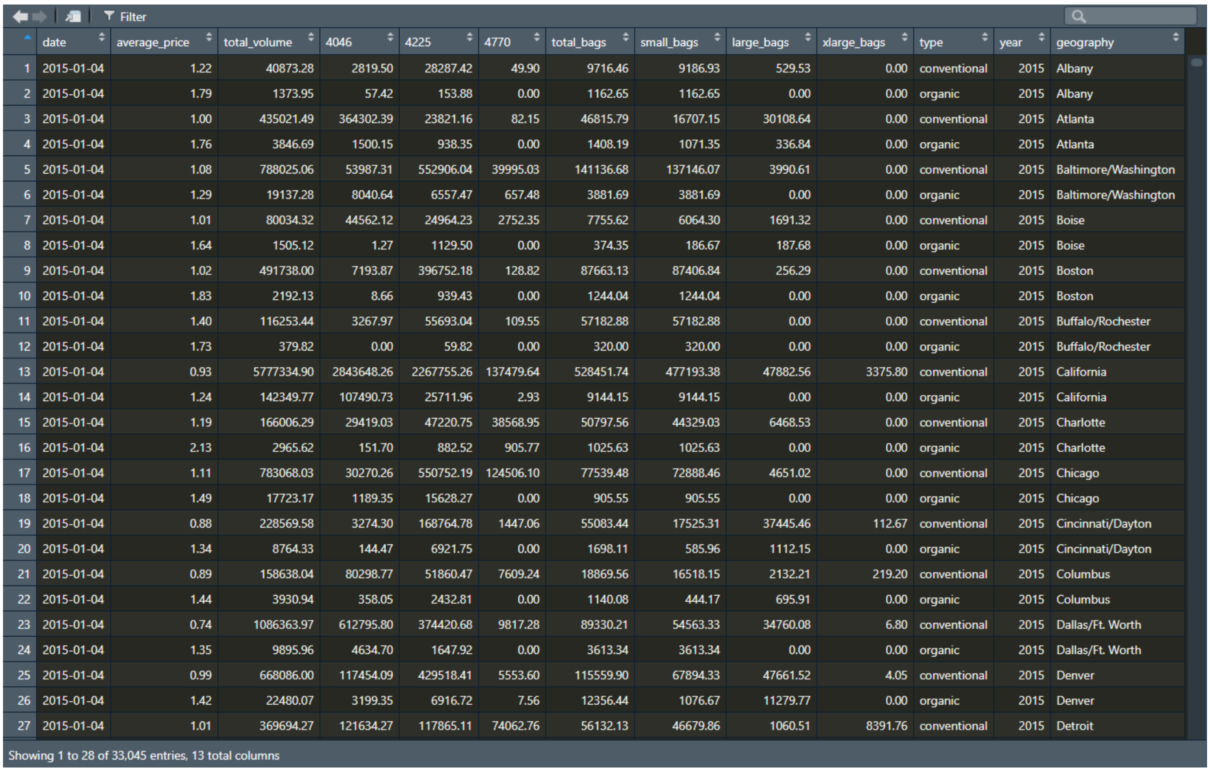


Fig 1.4: Missing values information using plot_missing() method.

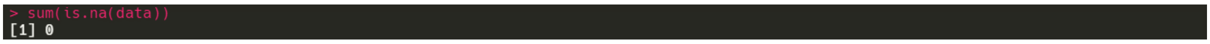


	date	average_price	total_volume	4046	4225	4770	total_bags	small_bags	large_bags	xlarge_bags	type	year	geography
1	2015-01-04	1.22	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.00	conventional	2015	Albany
2	2015-01-04	1.79	1373.95	57.42	153.88	0.00	1162.65	1162.65	0.00	0.00	organic	2015	Albany
3	2015-01-04	1.00	435021.49	364302.39	23821.16	82.15	46815.79	16707.15	30108.64	0.00	conventional	2015	Atlanta
4	2015-01-04	1.76	3846.69	1500.15	938.35	0.00	1408.19	1071.35	336.84	0.00	organic	2015	Atlanta
5	2015-01-04	1.08	788025.06	53987.31	552906.04	39995.03	141136.68	137146.07	3990.61	0.00	conventional	2015	Baltimore/Washington
6	2015-01-04	1.29	19137.28	8040.64	6557.47	657.48	3881.69	3881.69	0.00	0.00	organic	2015	Baltimore/Washington
7	2015-01-04	1.01	80034.32	44562.12	24964.23	2752.35	7755.62	6064.30	1691.32	0.00	conventional	2015	Boise
8	2015-01-04	1.64	1505.12	1.27	1129.50	0.00	374.35	186.67	187.68	0.00	organic	2015	Boise
9	2015-01-04	1.02	491738.00	7193.87	396752.18	128.82	87663.13	87406.84	256.29	0.00	conventional	2015	Boston
10	2015-01-04	1.83	2192.13	8.66	939.43	0.00	1244.04	1244.04	0.00	0.00	organic	2015	Boston
11	2015-01-04	1.40	116253.44	3267.97	55693.04	109.55	57182.88	57182.88	0.00	0.00	conventional	2015	Buffalo/Rochester
12	2015-01-04	1.73	379.82	0.00	59.82	0.00	320.00	320.00	0.00	0.00	organic	2015	Buffalo/Rochester
13	2015-01-04	0.93	5777334.90	2843648.26	2267755.26	137479.64	528451.74	477193.38	47882.56	3375.80	conventional	2015	California
14	2015-01-04	1.24	142349.77	107490.73	25711.96	2.93	9144.15	9144.15	0.00	0.00	organic	2015	California
15	2015-01-04	1.19	166006.29	29419.03	47220.75	38568.95	50797.56	44329.03	6468.53	0.00	conventional	2015	Charlotte
16	2015-01-04	2.13	2965.62	151.70	882.52	905.77	1025.63	1025.63	0.00	0.00	organic	2015	Charlotte
17	2015-01-04	1.11	783068.03	30270.26	550752.19	124506.10	77539.48	72888.46	4651.02	0.00	conventional	2015	Chicago
18	2015-01-04	1.49	17723.17	1189.35	15628.27	0.00	905.55	905.55	0.00	0.00	organic	2015	Chicago
19	2015-01-04	0.88	228569.58	3274.30	168764.78	1447.06	55083.44	17525.31	37445.46	112.67	conventional	2015	Cincinnati/Dayton
20	2015-01-04	1.34	8764.33	144.47	6921.75	0.00	1698.11	585.96	1112.15	0.00	organic	2015	Cincinnati/Dayton
21	2015-01-04	0.89	158638.04	80298.77	51860.47	7609.24	18869.56	16518.15	2132.21	219.20	conventional	2015	Columbus
22	2015-01-04	1.44	3930.94	358.05	2432.81	0.00	1140.08	444.17	695.91	0.00	organic	2015	Columbus
23	2015-01-04	0.74	1086363.97	612795.80	374420.68	9817.28	89330.21	54563.33	34760.08	6.80	conventional	2015	Dallas/Ft. Worth
24	2015-01-04	1.35	9895.96	4634.70	1647.92	0.00	3613.34	3613.34	0.00	0.00	organic	2015	Dallas/Ft. Worth
25	2015-01-04	0.99	668086.00	117454.09	429518.41	5553.60	115559.90	67894.33	47661.52	4.05	conventional	2015	Denver
26	2015-01-04	1.42	22480.07	3199.35	6916.72	7.56	12356.44	1076.67	11279.77	0.00	organic	2015	Denver
27	2015-01-04	1.01	369694.27	121634.27	117865.11	74062.76	56132.13	46679.86	1060.51	8391.76	conventional	2015	Detroit

Showing 1 to 28 of 33,045 entries, 13 total columns

Fig 1.5: Cleansed tabular view of the dataset.

Counting missing values in the dataset:



```
> sum(is.na(data))
[1] 0
```

b. Exploratory Analysis

In this phase of the project, we will concentrate mainly on the following aspect of the time series forecasting analysis which consists of:

Seasonal Patterns: In this section we will focus on constant patterns that occur frequently from year to year and from month to month in both types of avocados conventional and organic.

```
library(dplyr)
library(ggplot2)
library(DataExplorer)

options(repr.plot.width = 8, repr.plot.height = 4)
ggplot(
    data,
    aes(x = average_price, fill = type)
) +
geom_density() +
facet_wrap(~type) +
theme_minimal() +
theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "bottom"
) +
labs(title="Avocado Price by Type") +
scale_fill_brewer(palette = "Set2")
```

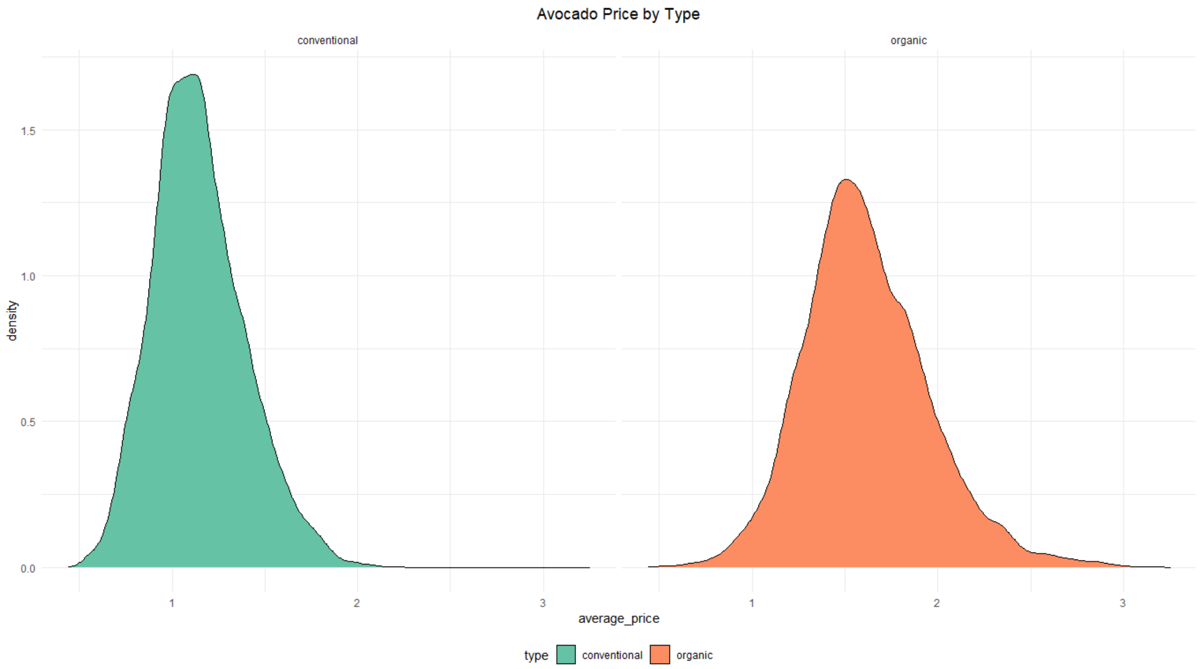
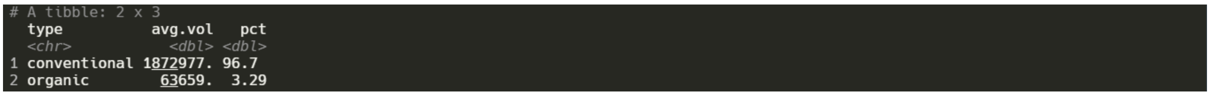


Fig 2.1: Density plots of the different type of avocadoes.

```
vol_type <- data %>%
            group_by(type) %>%
            summarise(avg.vol = mean(total_volume)) %>%
            mutate(pct=prop.table(avg.vol) * 100)
print(vol_type)
```

Output:



```
# A tibble: 2 x 3
  type          avg.vol   pct
  <chr>           <dbl> <dbl>
1 conventional 1872977. 96.7
2 organic        63659.  3.29
```

Types of Avocados:

In this section we will analyze the different types of avocados that we have in this dataset. Basically, we have two types of avocados: **Conventional and Organic.**

Summary:

- **Organic avocados:** Based on the price changes throughout time we can see that they are more expensive.

- **Conventional avocados:** Based on price changes throughout time we can see that they are less expensive.

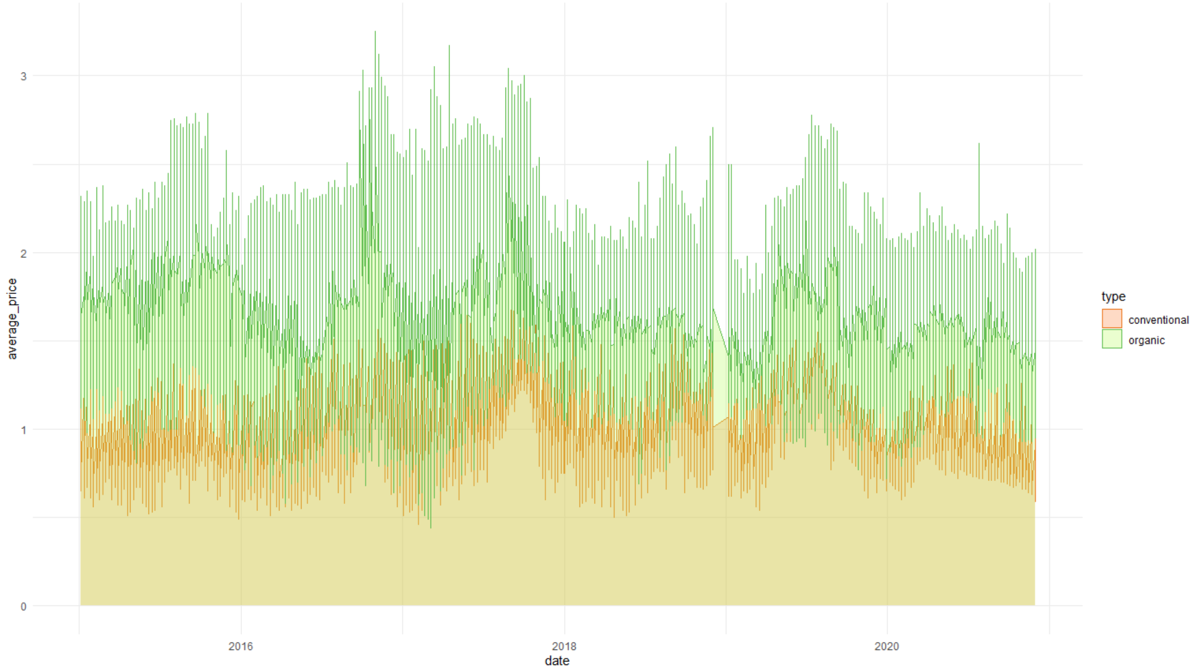


Fig 2.2: Range of average prices of avocados throughout the years.

Univariate Analysis:

```
plot_histogram(data)
```

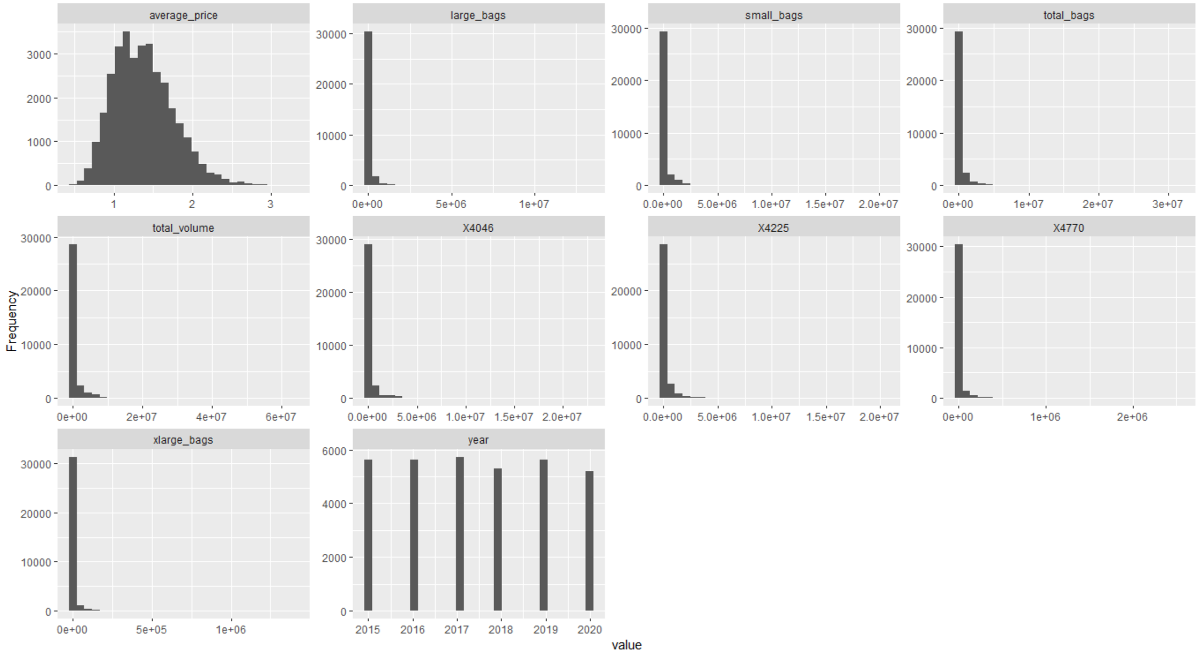


Fig 2.3: Histogram for every continuous feature.

```
plot_density(data)
```

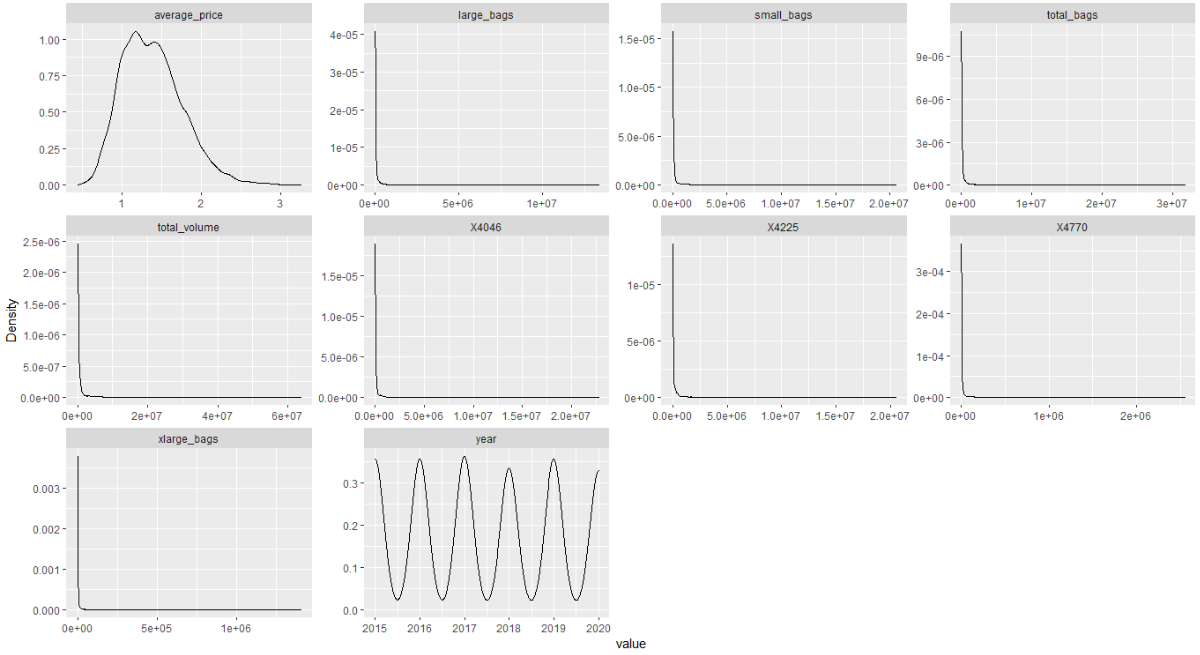


Fig 2.4: Density comparison plot.

The data is provided from 2015 till 2020. Next is to convert the year column to factor to treat it as categorical variable and also create a new column called month from date.

```
avocado$year = as.factor(avocado$year)
avocado$date = as.Date(avocado$date)
avocado$month = factor(months(avocado$date), levels = month.name)
```

Trend of Avocado Prices: The trend of avocado prices in last six years by avocado type.

```
options(repr.plot.width = 7, repr.plot.height = 5)
ggplot(avocado, aes(avocado$type, avocado$average_price)) +
geom_boxplot(aes(colour = avocado$year)) +
labs(
    colour = "Year",
    x = "Type",
    y ="Average Price",
    title = "Boxplot - Average price per year by type."
)
```

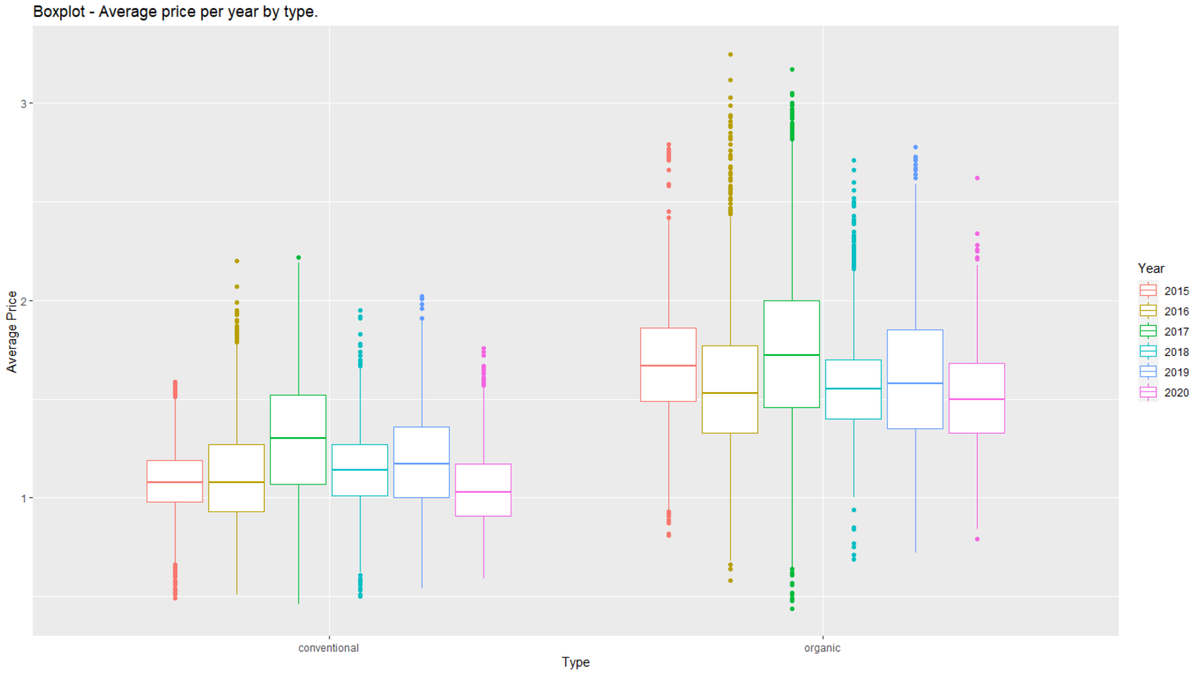


Fig 2.5: Average price of avocados per year by type of avocados.

Monthly trend of avocado prices by Avocado type in last years

Step 1: Group the dataset by Year, Month and Avocado Type. Calculate the monthly average price for Avocado respectively for each year.

```
grouped = avocado %>%
    group_by(year, month, type) %>%
    select(year, month, type, average_price) %>%
    summarise(average_price = mean(average_price))
```

Step 2: Plot a line chart showing the trend for both Conventional and Organic Avocado.

```
options(repr.plot.width = 12, repr.plot.height = 5)
ggplot(data = grouped, aes(x = month, y = average_price, colour = year, group
= year)) +
labs(
    colour = "Year",
    x = "Month",
    y ="Average Price",
    title = "Line Plot - Average monthly prices by type for every year."
) +
geom_line() +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
facet_grid(. ~grouped$type)
```

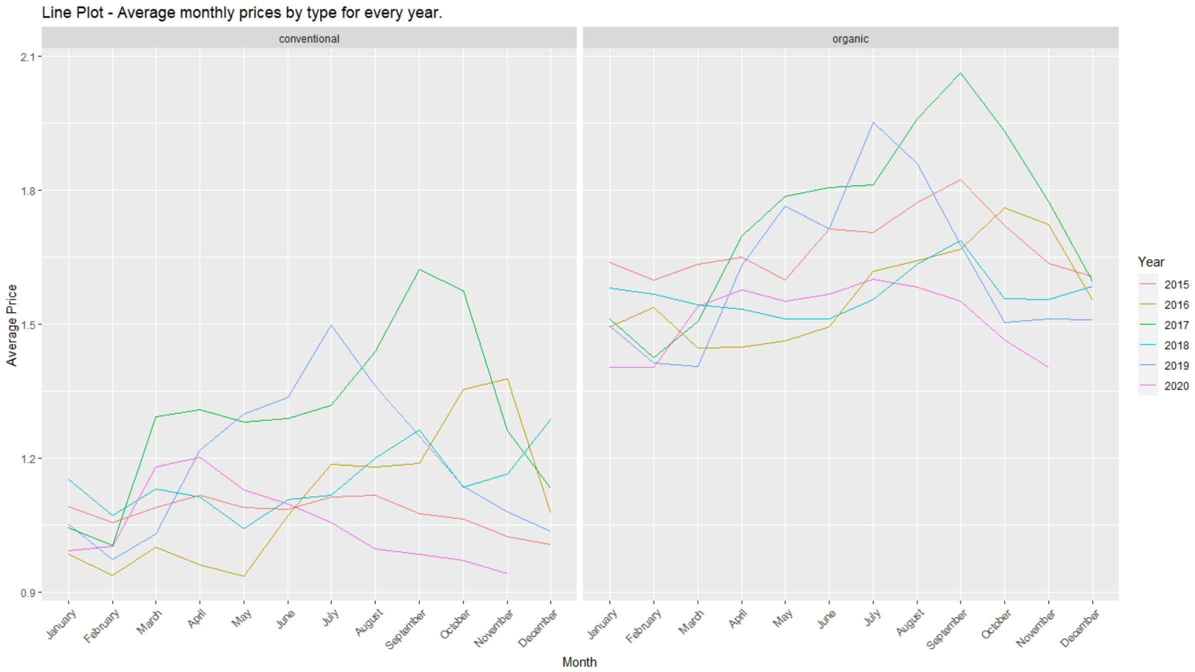


Fig 2.6: Average price of avocados per month by type of avocados for every year.

Group by geography:

```
grouped_geography_conv = avocado %>%
    select(year, geography, type, average_price) %>%
    filter(type == 'conventional')
min_con = round(min(grouped_geography_conv$average_price), 1) - 0.1
max_con = round(max(grouped_geography_conv$average_price), 1) + 0.1

grouped_geography_org = avocado %>%
    select(year, geography, type, average_price) %>%
    filter(type == 'organic')

min_org = round(min(grouped_geography_org$average_price), 1) - 0.1
max_org = round(max(grouped_geography_org$average_price), 1) + 0.1

options(repr.plot.width = 10, repr.plot.height = 12)
ggplot(grouped_geography_conv, aes(x = geography, y = average_price)) +
geom_tufteboxplot() +
facet_grid(.~grouped_geography_conv$year, scales = "free") +
labs(
    colour = "Year",
    x = "Geography",
    y ="Average Price",
    title = "Average prices of Conventional Avocados for each geography by yea
r"
) +
scale_y_continuous(breaks = c(seq(min_con, max_con, 0.2)), limits = c(min_con,
 max_con)) +
coord_flip() +
theme(axis.text.x = element_text(angle = 90, vjust = 0))
```

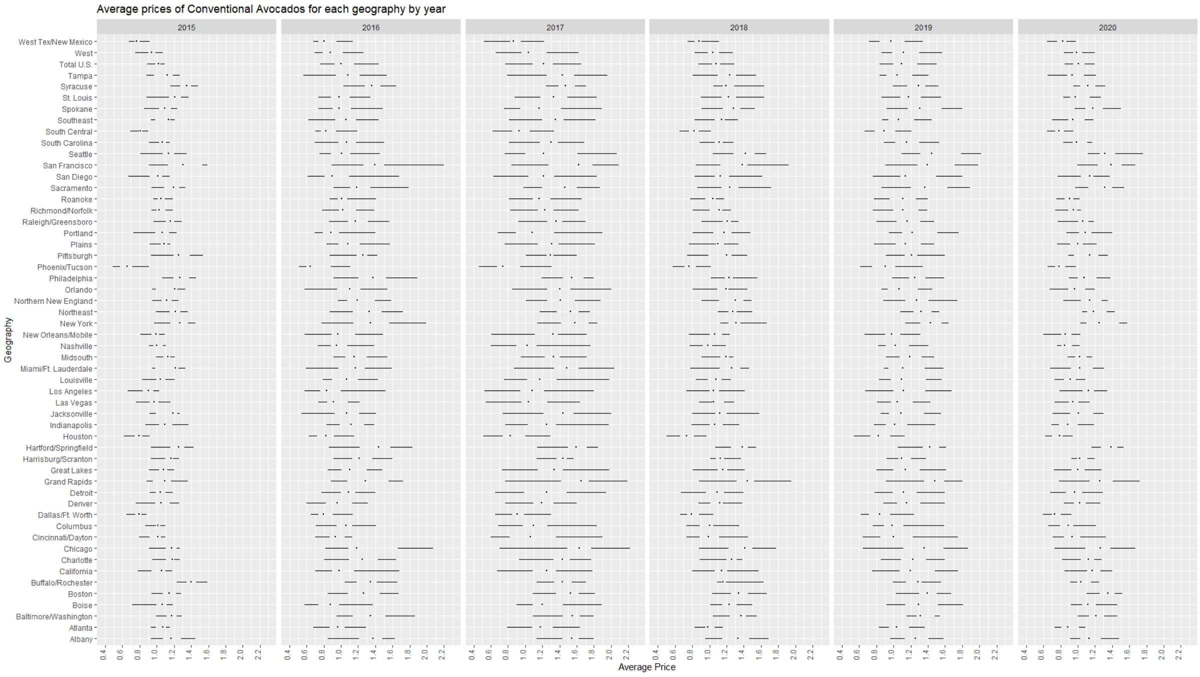


Fig 2.7: Average prices of conventional avocados for each region by year.

```
options(repr.plot.width = 12, repr.plot.height = 12)
ggplot(grouped_geography_org, aes(x = geography, y = average_price)) +
geom_tufteboxplot() +
facet_grid(.~grouped_geography_org$year, scales = "free") +
labs(
    colour = "Year",
    x = "Geography",
    y ="Average Price",
    title = "Average prices of Organic Avocados for each geography by year"
) +
scale_y_continuous(breaks = c(seq(min_org, max_org, 0.2)), limits = c(min_org,
 max_org)) +
coord_flip() +
theme(axis.text.x = element_text(angle = 90, vjust = 0))
```

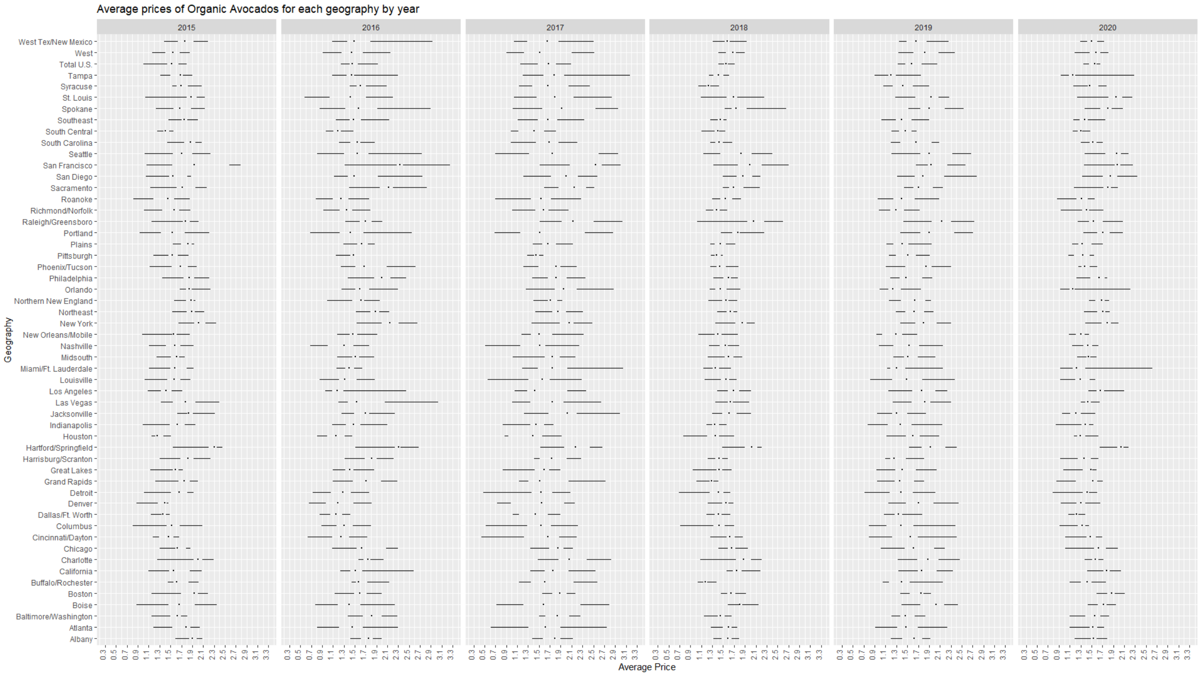


Fig 2.8: Average prices of organic avocados for each region by year.

Multivariate Analysis:

```
plot_correlation(data, type = 'continuous', 'quality')
```

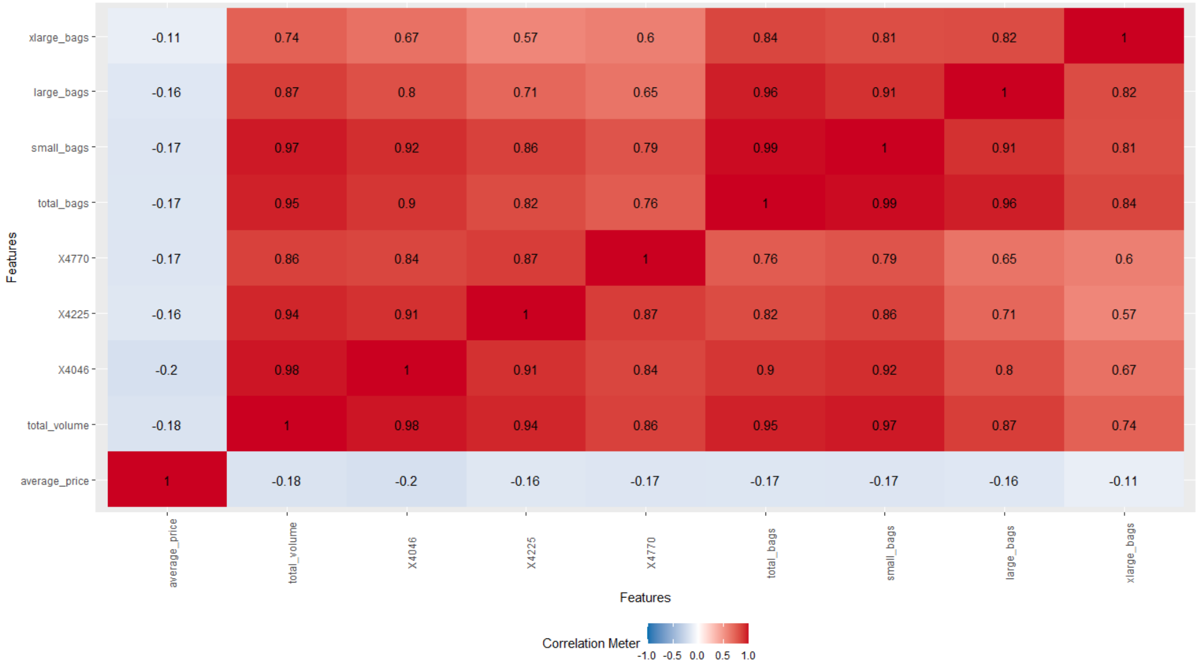
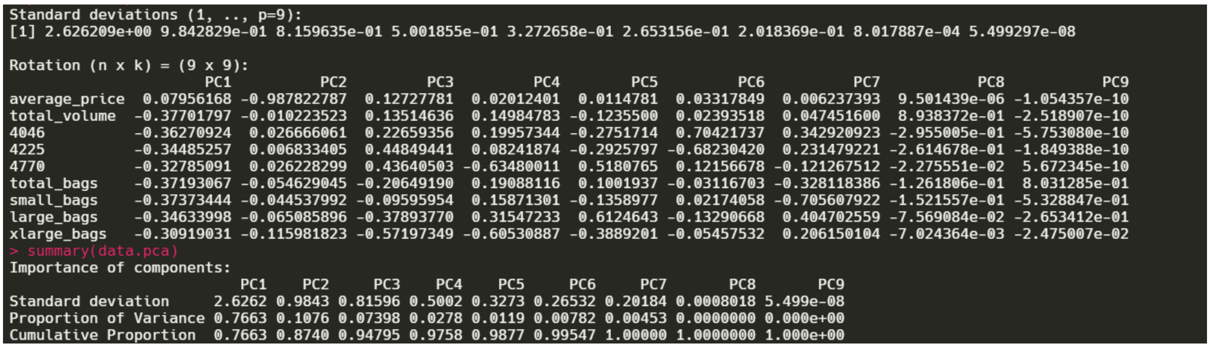


Fig 2.9: Correlation heatmap for all the continuous categories.

Principal Component Analysis

```
data.pca <- prcomp(
  select(data, 2, 3, 4, 5, 6, 7, 8, 9, 10),
  center = TRUE,
  scale = TRUE
)
print(data.pca)
summary(data.pca)
```



```
Standard deviations (1, .., p=9):
[1] 2.626209e+00 9.842829e-01 8.159635e-01 5.001855e-01 3.272658e-01 2.653156e-01 2.018369e-01 8.017887e-04 5.499297e-08

Rotation (n x k) = (9 x 9):
                     PC1          PC2         PC3         PC4        PC5         PC6          PC7           PC8           PC9
average_price  0.07956168 -0.987822787  0.12727781  0.02012401  0.0114781  0.03317849  0.006237393  9.501439e-06 -1.054357e-10
total_volume  -0.37701797 -0.010223523  0.13514636  0.14984783 -0.1235500  0.02393518  0.047451600  8.938372e-01 -2.518907e-10
4046          -0.36270924  0.026666061  0.22659356  0.19957344 -0.2751714  0.70421737  0.342920923 -2.955005e-01 -5.753080e-10
4225          -0.34485257  0.006833405  0.44849441  0.08241874 -0.2925797 -0.68230420  0.231479221 -2.614678e-01 -1.849388e-10
4770          -0.32785091  0.026228299  0.43640503 -0.63480011  0.5180765  0.12156678 -0.121267512 -2.275551e-02  5.672345e-10
total_bags    -0.37193067 -0.054629045 -0.20649190  0.19088116  0.1001937 -0.03116703 -0.328118386 -1.261806e-01  8.031285e-01
small_bags    -0.37373444 -0.044537992 -0.09595954  0.15871301 -0.1358977  0.02174058 -0.705607922 -1.521557e-01 -5.328847e-01
large_bags    -0.34633998 -0.065085896 -0.37893770  0.31547233  0.6124643 -0.13290668  0.404702559 -7.569084e-02 -2.653412e-01
xlarge_bags   -0.30919031 -0.115981823 -0.57197349 -0.60530887 -0.3889201 -0.05457532  0.206150104 -7.024364e-03 -2.475007e-02
> summary(data.pca)
Importance of components:
                          PC1    PC2     PC3    PC4    PC5     PC6     PC7       PC8       PC9
Standard deviation     2.6262 0.9843 0.81596 0.5002 0.3273 0.26532 0.20184 0.0008018 5.499e-08
Proportion of Variance 0.7663 0.1076 0.07398 0.0278 0.0119 0.00782 0.00453 0.0000000 0.000e+00
Cumulative Proportion  0.7663 0.8740 0.94795 0.9758 0.9877 0.99547 1.00000 1.0000000 1.000e+00
```

c. Applying DMBI/ML algorithms

1. Linear Regression

Linear regression is one of the most commonly used predictive modelling techniques to predict the value of a continuous variable Y based on one or more input predictor variables X. The aim is to establish a mathematical formula between the response variable (Y) and the predictor variables (X's). Linear regression model can be used to learn which features are important by examining coefficients. If a coefficient is close to zero, the corresponding feature is considered to be less important than if the coefficient was a large positive or negative value. That's how the linear regression model generates the output. Coefficients are multiplied with corresponding input variables, and in the end, the bias (intercept) term is added.

Step 1: It is observed in the dataset that the price of avocados goes down as volume sold goes up. Test if this relationship is statistically significant.

```
library(sjPlot)
library(ggfortify)
library(readxl)

# Import the dataset.
avocado <- read_excel("avocado.xlsx")
data = avocado

conmod <- data %>%
  filter(type == "conventional") %>%
  mutate(Volume = log(`total_volume`)) %>%
  lm(average_price ~ Volume, data = .)

orgmod <- data %>%
  filter(type == "organic") %>%
  mutate(Volume = log(`total_volume`)) %>%
  lm(average_price ~ Volume, data = .)
```

Save the model in order to display the results in a table. Use the tab_model function from the sjPlot package for this purpose.

```
# Printing the output in a table rather than in the console.
tab_model(conmod, orgmod)
```

Step 2: Regression Diagnostics

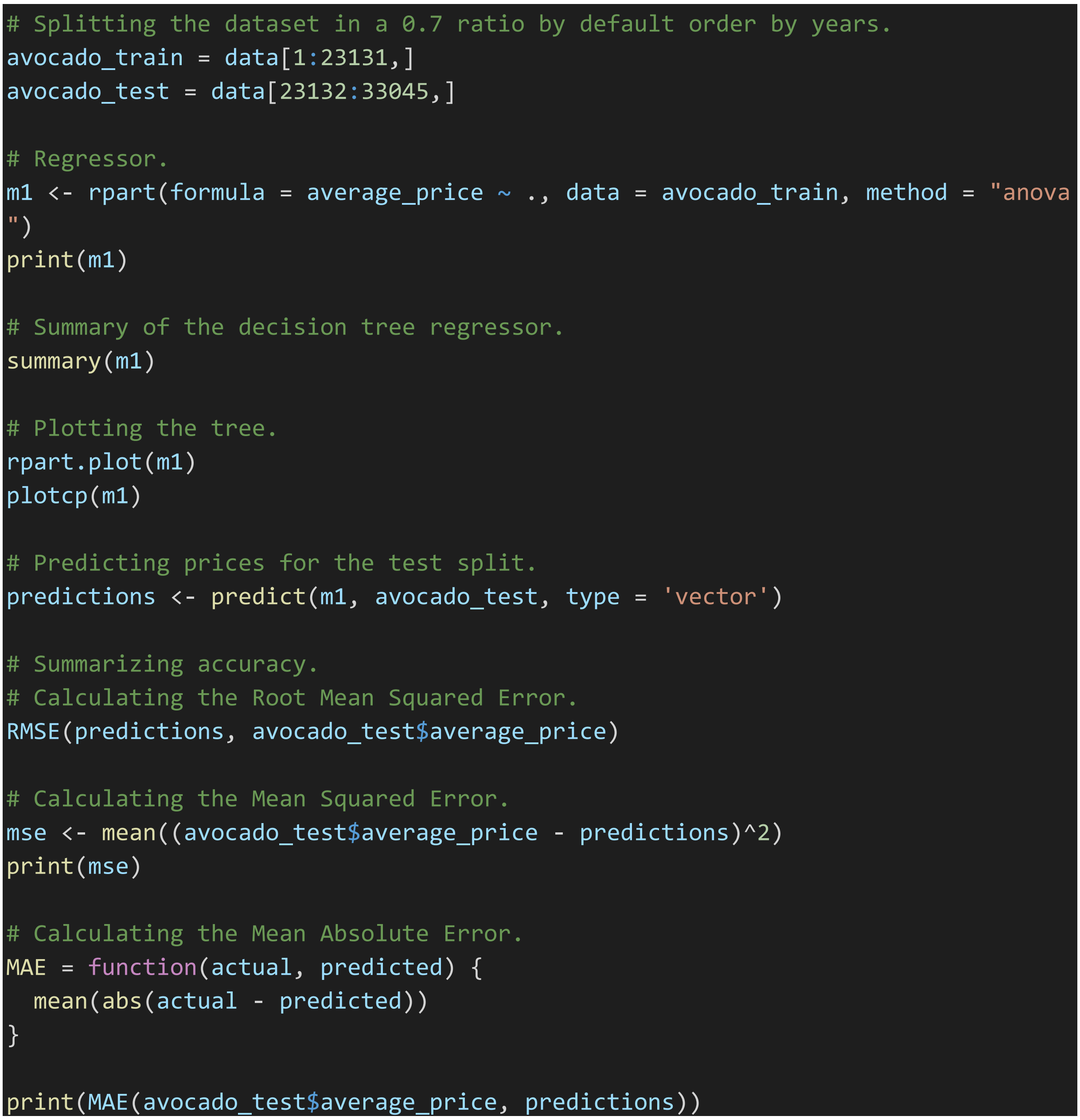
Examine if the fitted model is any good with autoplot and ggfortify. This will give a ggplot2 version of the regression diagnostics plot from base R.

```
#Regression diagnostics.
autoplot(conmod)
autoplot(orgmod)
```

2. Decision Trees

Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables. It builds classification models in the form of a tree-like structure, just like its name. It is also used to create data models that will predict class labels or values for the decision-making process. The models are built from the training dataset fed to the system (supervised learning). Using a decision tree, we can visualize the decisions that make it easy to understand and thus it is a popular data mining technique.

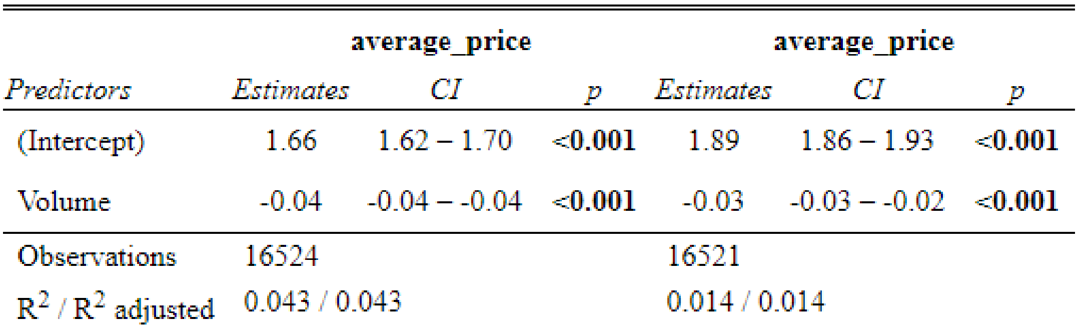
```
library(rsample)     # Data splitting.
library(dplyr)       # Data wrangling.
library(rpart)       # Performing regression trees.
library(rpart.plot)  # Plotting regression trees.
library(readxl)

# Import the dataset.
avocado <- read_excel("avocado.xlsx")

# Drop the geography column.
data = avocado[, 1:12]
```



```
# Splitting the dataset in a 0.7 ratio by default order by years.
avocado_train = data[1:23131,]
avocado_test = data[23132:33045,]

# Regressor.
m1 <- rpart(formula = average_price ~ ., data = avocado_train, method = "anova")
print(m1)

# Summary of the decision tree regressor.
summary(m1)

# Plotting the tree.
rpart.plot(m1)
plotcp(m1)

# Predicting prices for the test split.
predictions <- predict(m1, avocado_test, type = 'vector')

# Summarizing accuracy.
# Calculating the Root Mean Squared Error.
RMSE(predictions, avocado_test$average_price)

# Calculating the Mean Squared Error.
mse <- mean((avocado_test$average_price - predictions)^2)
print(mse)

# Calculating the Mean Absolute Error.
MAE = function(actual, predicted) {
  mean(abs(actual - predicted))
}

print(MAE(avocado_test$average_price, predictions))
```

d. Visualization and Interpretation of results.

1. Linear Regression



	average_price			average_price		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	1.66	1.62 – 1.70	**<0.001**	1.89	1.86 – 1.93	**<0.001**
Volume	-0.04	-0.04 – -0.04	**<0.001**	-0.03	-0.03 – -0.02	**<0.001**
Observations	16524			16521		
R^2 / R^2 adjusted	0.043 / 0.043			0.014 / 0.014		

Fig 4.1: Tabular data showing the relation between average avocado prices and volume.

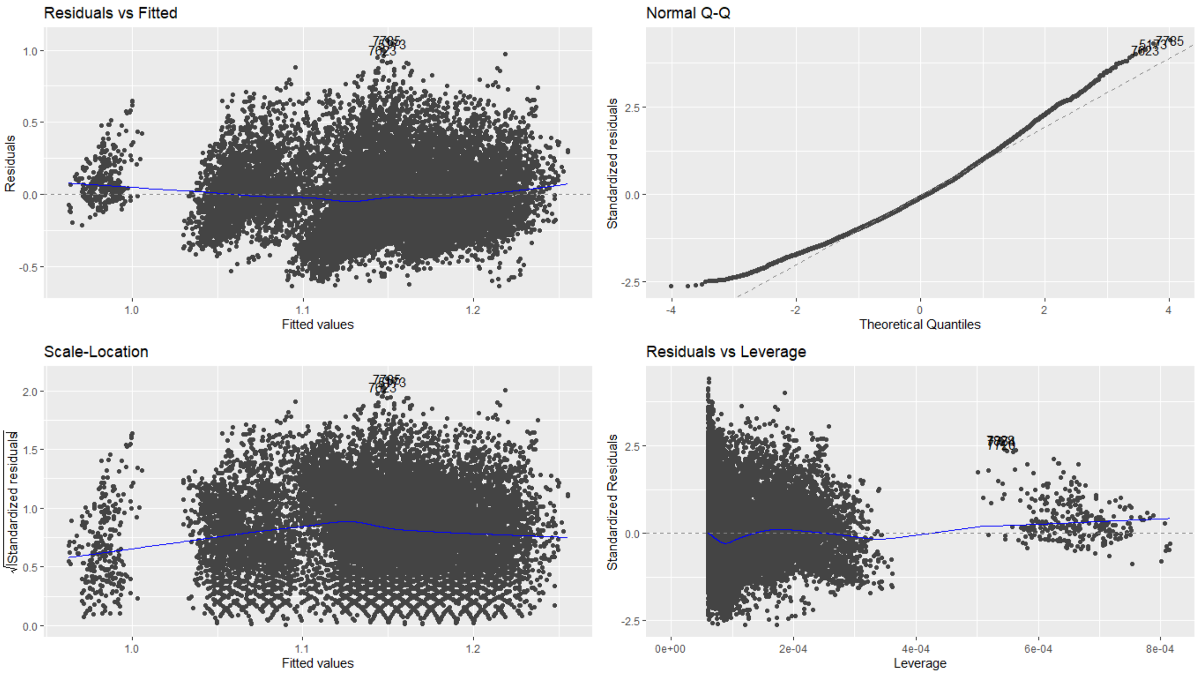


Fig 4.2: Regression diagnostic plots for the conventional avocados.

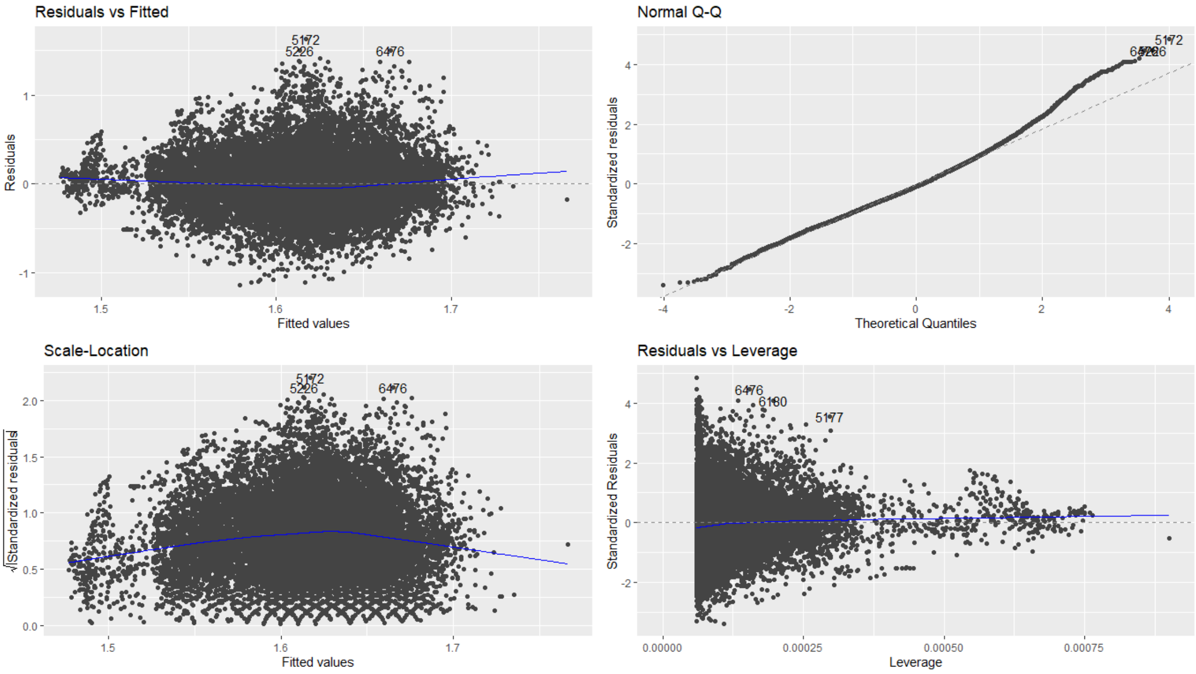
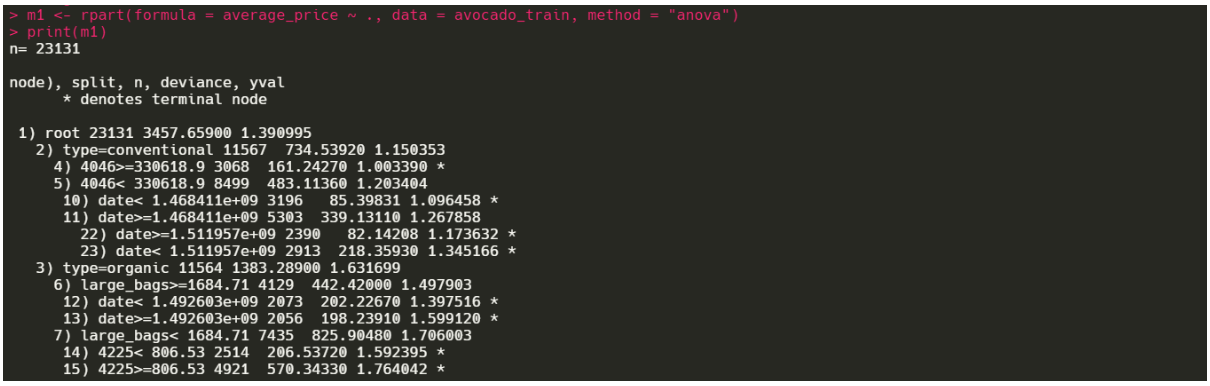


Fig 4.3: Regression diagnostic plots for the organic avocados

2. Decision Tree Regressor



```
> m1 <- rpart(formula = average_price ~ ., data = avocado_train, method = "anova")
> print(m1)
n= 23131

node), split, n, deviance, yval
      * denotes terminal node

 1) root 23131 3457.65900 1.390995
   2) type=conventional 11567  734.53920 1.150353
     4) 4046>=330618.9 3068  161.24270 1.003390 *
     5) 4046< 330618.9 8499  483.11360 1.203404
      10) date< 1.468411e+09 3196   85.39831 1.096458 *
      11) date>=1.468411e+09 5303  339.13110 1.267858
        22) date>=1.511957e+09 2390   82.14208 1.173632 *
        23) date< 1.511957e+09 2913  218.35930 1.345166 *
   3) type=organic 11564 1383.28900 1.631699
     6) large_bags>=1684.71 4129  442.42000 1.497903
      12) date< 1.492603e+09 2073  202.22670 1.397516 *
      13) date>=1.492603e+09 2056  198.23910 1.599120 *
     7) large_bags< 1684.71 7435  825.90480 1.706003
      14) 4225< 806.53 2514  206.53720 1.592395 *
      15) 4225>=806.53 4921  570.34330 1.764042 *
```

Fig 4.4: Decision tree regressor model.

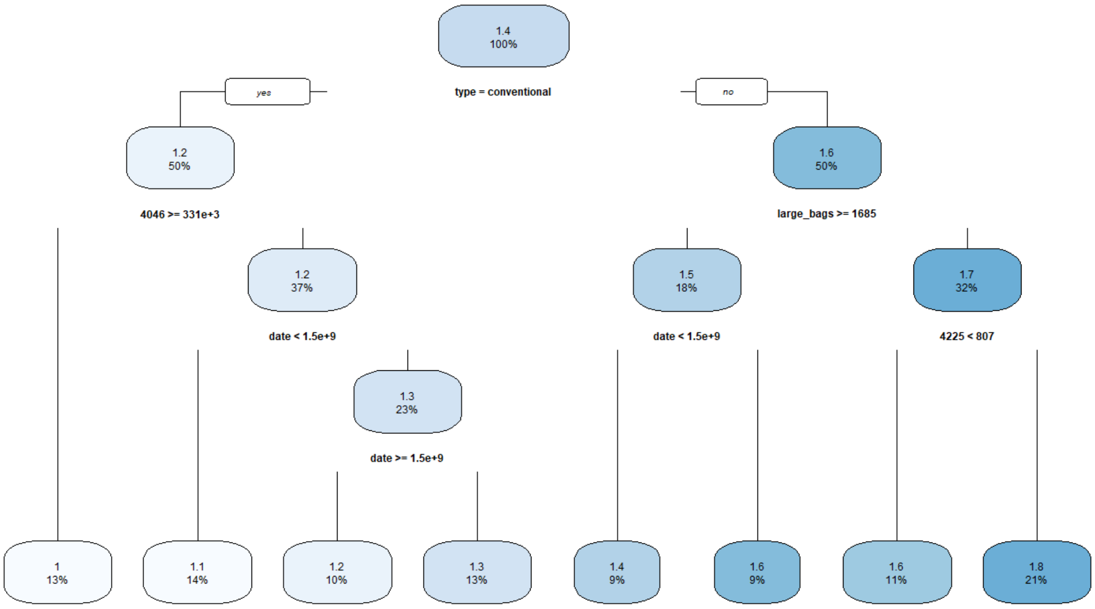


Fig 4.5: Visualizing the decision tree regressor model.

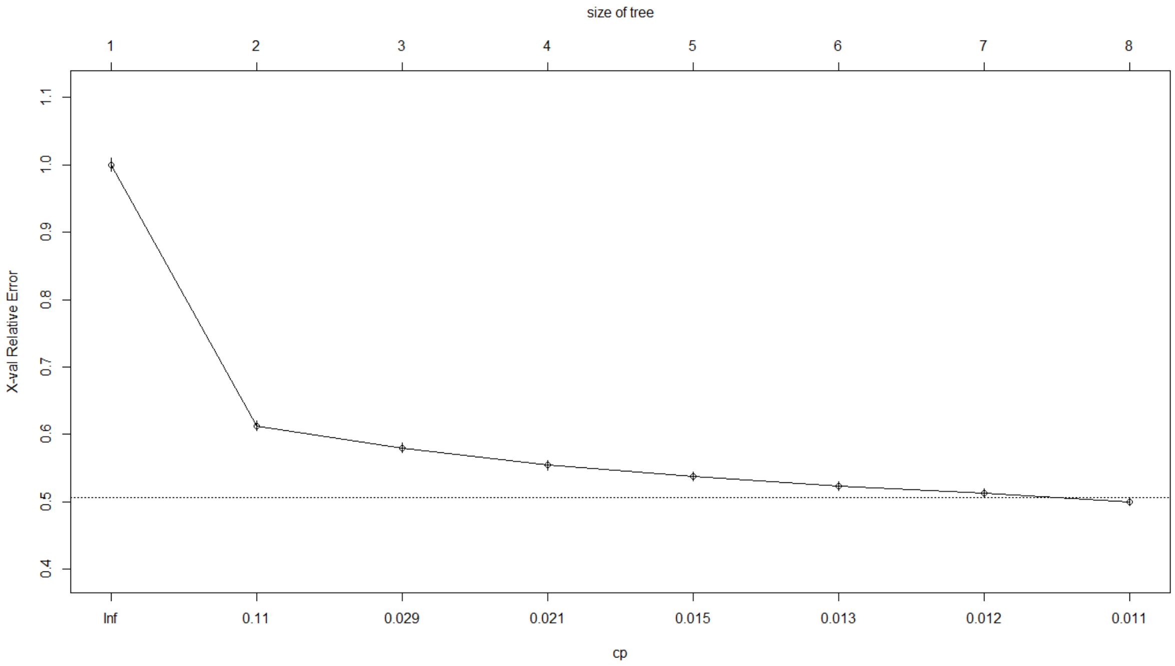
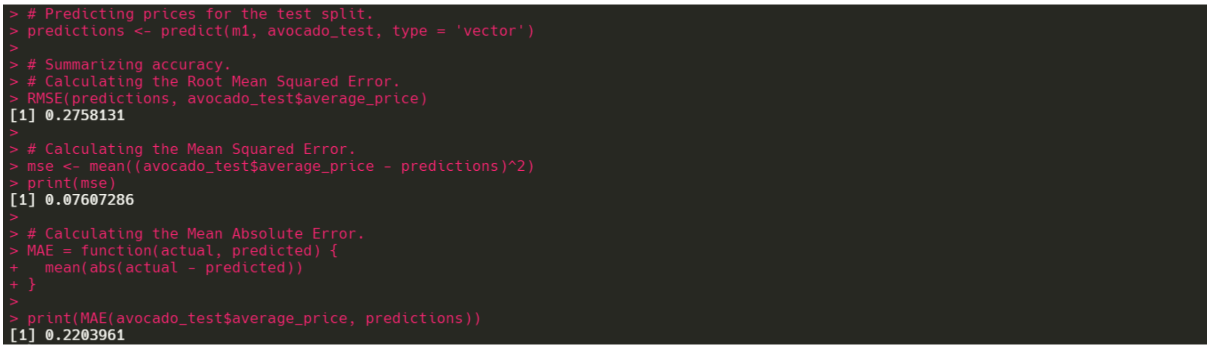


Fig 4.6: Visual representation of the cross-validation results in the model.



```
> # Predicting prices for the test split.
> predictions <- predict(m1, avocado_test, type = 'vector')
>
> # Summarizing accuracy.
> # Calculating the Root Mean Squared Error.
> RMSE(predictions, avocado_test$average_price)
[1] 0.2758131
>
> # Calculating the Mean Squared Error.
> mse <- mean((avocado_test$average_price - predictions)^2)
> print(mse)
[1] 0.07607286
>
> # Calculating the Mean Absolute Error.
> MAE = function(actual, predicted) {
+   mean(abs(actual - predicted))
+ }
>
> print(MAE(avocado_test$average_price, predictions))
[1] 0.2203961
```

Fig 4.7: Summarizing accuracy with RMSE, MSE and MAE.

e. Results and Discussions

1. Linear Regression

Observations claiming the relationship between avocado prices and volumes where prices go down with increase in volumes sold were confirmed by the results. The p-value is shows that the negative coefficient is statistically significant.

It is observed from the Linear Regression plots, for both type of avocados (conventional and organic), that the model roughly follows a normal distribution as the deviations are not too steep. Furthermore, the standard deviation of the residuals does not exceed 3 and are mostly below 2. We can therefore conclude that the price of avocado drops as volume increases.

2. Decision Tree Regressor

A decision tree is one of the most common classification algorithms that are implemented. The avocado dataset was first split in the ratio 0.7 for training and testing sets. The decision tree implemented in this project was plot on the training dataset. From the testing set, predictions were calculated. To summarize the accuracy of the model, different errors like Root Mean Squared Error, Mean Squared Error, Mean Absolute Error were calculated. MSE resulted in a 0.076 which is much closer to 0, indicating that the predictions were significantly accurate and closer to the actual values in the testing set. Lower values of MSE indicate good predictions, the closer the MSE is to 0, the accurate the prediction is.

Conclusion: The two algorithms that were used to create models were Linear Regression and Decision Tree Regressor.

From the Linear Regression plots, for both type of avocados (conventional and organic), we conclude that the model roughly follows a normal distribution as the deviations are not too steep (less than 2.5). We therefore conclude that the price of avocado drops as volume increases.

Decision Tree Regression model was implemented to predict the average prices of the avocados. Different errors were calculated to test the accuracy. They are:

- Root Mean Squared Error = 0.2758
- Mean Squared Error = 0.0760
- Mean Absolute Error = 0.2209

Thus, we can use R in order to analyse and design various algorithms on any dataset. Other functions can also be performed using R such as exploratory data analysis and functions in pre-processing of data like cleaning data and replacing missing values. Larger datasets can also be modified to work on smaller subsets of data to avoid problems like overfitting. Using R for carrying out such operations is very beneficial.