Avocados Pricing Engine: Capstone Project

Data Science Professional Certificate Program (HarvardX - EdX Platform)

Luis Sousa

December 2021

Contents

1. INTRODUCTION **2**

1.1 Goals & Challenges 2

1.2 The Avocado Dataset 2

2. DATA ANALYSIS **3**

2.1 Downloading Data 3

2.2 Data Transformation & Cleaning 4

2.3 Basic Data Exploration: R Basic Functions 5

2.4 Advanced Data Exploration: Data Visualization 7

3. METHODS: MODEL ENGINEERING **12**

3.1 Model 1: Average Model 12

3.2 Model 2: Linear Regression 13

3.3 Model 3: Multi Linear Regression 15

3.4 Model 4: Lasso Regression 16

3.5 Model 5: Decision Tree (Regression) 17

3.5 Discussion 19

4. CONCLUSION **19**

4.1 Discussion 20

4.2 Future work and considerations 20

4.3 References 20

1. INTRODUCTION

This report is part of the last assignment of the Professional Data Science Certificate program, run by HarvardX in the edX platform. In this final assignment, the students are given the chance to choose freely a topic to apply some of the techniques studied in previous courses of the professional certificate series. To fulfill the requirements of the assignment, a topic with plentiful of data is required.

The **avocado dataset** from the Kaggle platform was chosen for this project, because it is a topic that is simple enough for the majority of people to understand it, but it also allows for the develop advanced data visualization techniques and also advanced machine learning models.

This report analysis in detail the avocado dataset in four steps:

1. **Data preparation**: Download and cleaning the data
2. **Data analysis**: Basic data exploration and advanced data visualization
3. **Model development**: Multiple distinct approaches for model engineering
4. **Model selection**: Based on model performance KPIs: $RMSE$, and R^2
5. **Conclusion**: Final discussion/comments, including future work

Please note that the complete project comprises of a PDF file, containing the report, and the respective R and RMD files. All omitted code in this report can be found in detail in the R and RMD files.

1.1 Goals & Challenges

This capstone project has two main goals, described below:

- Apply to the **avocado** dataset the data science concepts acquired throughout the Professional Certificate Program (HarvardX) in a sinful and innovative way
- Train and validate a model that predicts avocado prices, using different regression approaches

Two KPIs will be used to assess model performance, the $RMSE$ and the R^2 (or adjusted R^2 for multi regression). The goal of this project is not to set a target for the $RMSE$ or R^2, but to analyse how different predictors work on the same basic but yet challenging task of predicting the prices of avocados. As a reminder, the lower $RMSE$ value is, the better the model is. R-squared explains to what extent the variance of one variable explains the variance of the second variable. For instance, a R2 of a model of 0.50, explains approximately half of the observed variation of the model's inputs. Typically, a higher R^2 is better and in many finance applications, a R^2 close to .7 is deemed very good.

1.2 The Avocado Dataset

The avocado dataset was obtained in the Kaggle data base. The avocado dataset consists of data collected from the Hass Avocado Board website in May of 2018.

The data represents weekly 2015-2018 retail scan data for US National retail volume (in units) and price. The data aggregates several sales channels e.g. grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags.

Section 2.3, depicts in more detail the most relevant columns of the avocado dataset for this report.

2. DATA ANALYSIS

In this chapter, all steps related to data preparation and analysis will be performed. The first step is to download the data. Then the data will be analysed and explored to derive hypothesis on which models could be used for the rating engine. Additionally, the advanced visualization of the avocado dataset is going to unveil the first report insights.

2.1 Downloading Data

As explained previously, data for this project was downloaded in the Kaggle website - link to the avocado dataset. The dataset was replicated on GitHub, to allow all the project files to be saved in the same hosting server.

The first section of the R script downloads the avocado dataset in *.cvs* format from GitHub and loads it into a dataframe of raw data. Data will only be downloaded if not present in the current R working directory, to save the script´s running time. The R code is depicted below:

```r
# Original Kaggle link with the Avocado dataset
Kaggle_data_link <-
    "https://www.kaggle.com/neuromusic/avocado-prices/download/"

# Mirrow Avocado dataset on gitHub
gitHub_data_link <-
  "https://github.com/luisfdsousa/EdX_Capstone_Avocado/raw/main/avocado.csv"

# Download link
data_weblink <- gitHub_data_link

# File name containing the desired dataset
data_file_name <- "avocado.csv"

# Create an empty avocados data frame
avocados <- NULL

#Check if data exists within the working directory
if (file.exists(data_file_name) == F) {

  #Info message
  cat("Avocado data not found.",
      "Next step: Download data from Kaggle")

  # Create temporary file to store the data
  temp_file <- tempfile()

  # Download the data from the web
  download.file(data_weblink, temp_file)

  # Open and load .cvs data
  avocados <- read.csv(temp_file,
                       header = T)

  # Free memory, remove temp file
  rm(temp_file)
} else {

  # Open and load .cvs data
  avocados <- read.csv(data_file_name,
                       header = T)
}
```

The R packages required for this project were loaded before downloading the data. This step is omitted in this report, but can be analysed in detail in the R script.

2.2 Data Transformation & Cleaning

The raw data stored in the **avocados** data frame needs to be transformed to be fit for the next phase of data exploration and visualization. The list of data transformation performed in the **avocados** data frame is as follows:

- **Filter data**: release from memory the following columns X (data row ID), Total.Bags, Small.Bags, Large.Bags, XLarge.Bags. The number of bags sold is going to be ignored as it is highly correlated with the total volume sold, as it is going to be observed later.
- **Change column names**: Make the **avocados** data set more readable by changing the column names e.g. X4225 is the column of total number of large avocados sold (T4225L).
- **Categorical variables**: Transform strings/characters into categorical variables via factorisation.
- **Dates**: convert dates into POSIX format and create new columns for month and year.
- **Non-Haas avocados**: Create a new column with the total number of non-Haas avocados sold
- **Invalid data points**: Remove all rows containing invalid data points, as these can create substantial challenges during the more advanced sections e.g. advanced data visualisation and model engineering.

Below are the detail steps in R for the initial data transformation phase.

```
# Drop the first column (gibberish)
# drop also the bags column (description not clear)
avocados <-
  avocados %>%            # pipe the avocados data set
  subset(select = -c(X,
                     Total.Bags,
                     Small.Bags,
                     Large.Bags,
                     XLarge.Bags))

# Rename columns to simplify reading
names(avocados)[names(avocados) == "Total.Volume"] <- "TotalVolume"
names(avocados)[names(avocados) == "X4046"] <- "T4046S"
names(avocados)[names(avocados) == "X4225"] <- "T4225L"
names(avocados)[names(avocados) == "X4770"] <- "T4770XL"
names(avocados)[names(avocados) == "type"] <- "Type"
names(avocados)[names(avocados) == "year"] <- "Year"
names(avocados)[names(avocados) == "region"] <- "Region"

#avocados$Year = as.factor(avocados$Year)
avocados$Type = as.factor(avocados$Type)
avocados$Region = as.factor(avocados$Region)
```

```
avocados$Year = as.factor(avocados$Year)
#avocados$Date = as.Date(avocados$Date)
avocados$Date = ymd(avocados$Date)
avocados <- add_column(avocados, Month = factor(months(avocados$Date), levels = month.name))
avocados <- add_column(avocados, Day = factor(format(avocados$Date, format = "%d")))
avocados <- avocados %>% mutate(NonHaasVolume = TotalVolume - T4046S - T4225L - T4770XL)

# Remove NA entries
avocados <- avocados %>% drop_na()
```

Next, the seed is going to be set to achieve repeatable results and the **avocados** data set will be 90/10 split. This means 90% of the avocados data frame is going to be used for the training dataset, **train_data**, and 10% to the testing dataset, **test_data**. The 90/10 split ratio is in this project necessary as the whole dataset is not deemed very large. The creation of the testing and training datasets are done as follows:

```
# Initiate seed for repeated results
# if using R 3.5 or earlier, use `set.seed(1)`
set.seed(1, sample.kind = "Rounding")

# 90/10 % ratio for test/validation datasets
training_samples <-
  avocados$AveragePrice %>%
  createDataPartition(p = 0.9,
                      list = FALSE)

# Create respective train and test datasets
train_data  <- avocados[training_samples, ]
test_data <- avocados[-training_samples, ]
```

Additional steps to transform the data for some of the models in this report will be required and covered in Chapter 3, as these are specific to the needs of the regression functions in R.

2.3 Basic Data Exploration: R Basic Functions

In sections 2.1 and 2.2, the avocado dataset was downloaded and prepared into the training, **train_data**, and testing, **test_data**, datasets. In a typical data science project, the data analysis phase comes before the data preparation phase. In this report the phases are showed in reversed order to showcase the exact same cleaned data version that is used in the model engineering phase.

This section is fundamental to build the general knowledge about the **avocados** dataset. Bellow, the basic characteristics of the cleaned dataset is showed. These next basic exploration steps are fundamental to get a glimpse on the data structure and to obtain insights about potential model to use for the next chapter.

There are **18.000+ data rows** and **12 columns**, which is far from being considered a large dataset. This may significantly influence the $RMSE$ and R^2 of the models. Moreover, the data only spans over a limited amount of time (2015-2018), which may be enough to predict avocado prices into 2019 but not enough to go beyond it, especially if the dataset depicts a recession or a growth phase of the market. There are **54 regions** where avocados are produced in this dataset – 54 regions for 18.000 datapoints means that there are only a few data points region (300 datapoints per region on average). It is confirmed that the dataset has zero invalid data entries, as during the preparation phase all invalid entries were removed. There are two types of avocados: the *organic* and *conventional*.

The next table depicts some of the project´s most important columns in the **avocados** dataset.

Table 1: Basic Data Exploration: Avocado Dataset

Metric	Result
Number of Rows	18249.00
Number of Columns	12.00
Number of Invalid Data Points	0.00
Average avocado price (whole dataset)	1.41
SD avocado price (whole dataset)	0.40
Distinct Types of Avocados	2.00
Distinct Produce Regions	54.00
Data Start Date	2015.00
Data End Date	2018.00

Table 2: Avocado dataset: Columns description

Columns	Description
Date	The date of the observation
AveragePrice	The average price of a single avocado
TotalVolume	Total number of avocados sold
T4046S	Total number of avocados: Small/medium Haas
T4225L	Total number of avocados: Large Haas
T4770XL	Total number of avocados: Extra large Haas
Type	Type of avocado
Year	Produce year
Region	Produce region

Last but not least, a summary of the **avocados** dataset is showed for all the columns. Please remember that the **avocados**, **test_data** and **train_data** data frames share the same data structure.

```
# Print data summary
summary(avocados)
```

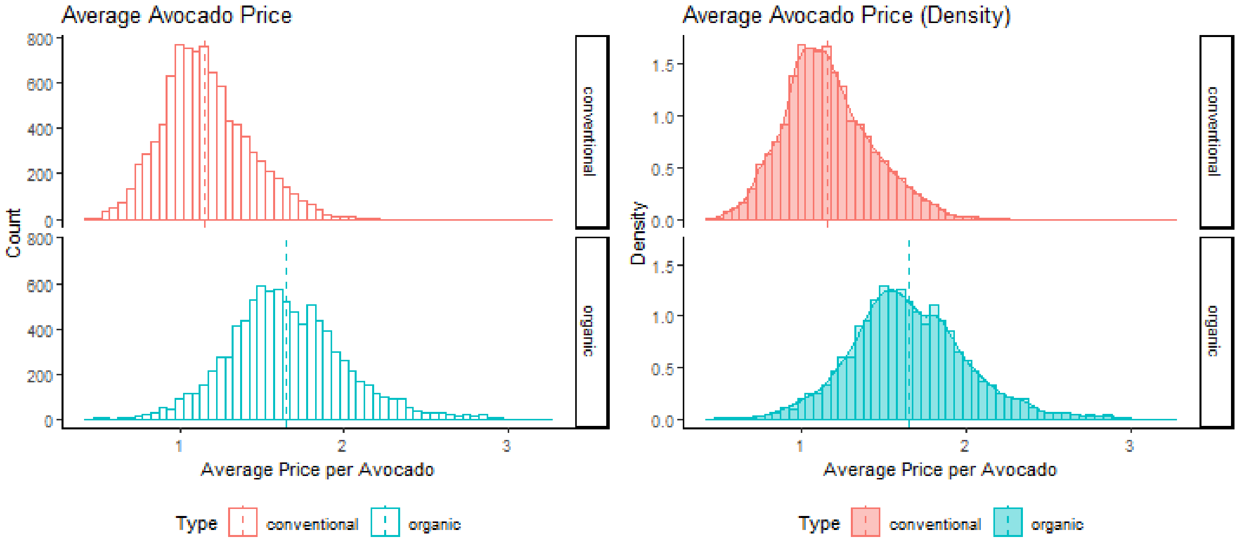
```
##       Date             AveragePrice    TotalVolume         T4046S       
##  Min.   :2015-01-04   Min.   :0.440   Min.   :      85   Min.   :       0  
##  1st Qu.:2015-10-25   1st Qu.:1.100   1st Qu.:   10839   1st Qu.:     854  
##  Median :2016-08-14   Median :1.370   Median :  107377   Median :    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  
##                                                                           
##      T4225L           T4770XL                    Type        Year     
##  Min.   :       0   Min.   :      0   conventional:9126   2015:5615  
##  1st Qu.:    3009   1st Qu.:      0   organic     :9123   2016:5616  
##  Median :   29061   Median :    185                       2017:5722  
##  Mean   :  295155   Mean   :  22840                       2018:1296  
##  3rd Qu.:  150207   3rd Qu.:   6243                                  
##  Max.   :20470573   Max.   :2546439                                  
##                                                                      
##                  Region           Month           Day          NonHaasVolume    
##  Albany             :  338   January :1944   04     :  756   Min.   :       0  
##  Atlanta            :  338   March   :1836   11     :  756   1st Qu.:    5089  
##  BaltimoreWashington:  338   February:1728   18     :  755   Median :   39744  
##  Boise               :  338   May      :1512   25      :  755   Mean   :  239641
##  Boston              :  338   July     :1512   01      :  648   3rd Qu.:  110783
##  BuffaloRochester    :  338   October  :1512   03      :  648   Max.   :19373134
##  (Other)             :16221   (Other)  :8205   (Other):13931
```

2.4 Advanced Data Exploration: Data Visualization

Section 2.3 was focused on quick data exploration facts and remarks. This section focusses on advanced data visualization. Since the main goal is to predict the avocado's price, most of the plots will be based around the *AveragePrice* column and how it relates to data contained in other columns.

2.4.1 Plot 1: Histogram of average prices by avocado type

The first plot shows avocados average price distribution and density for the **avocados** dataset. The mean values are depicted in as the dashed vertical lines. The plots are split by avocado type (organic vs. conventional).

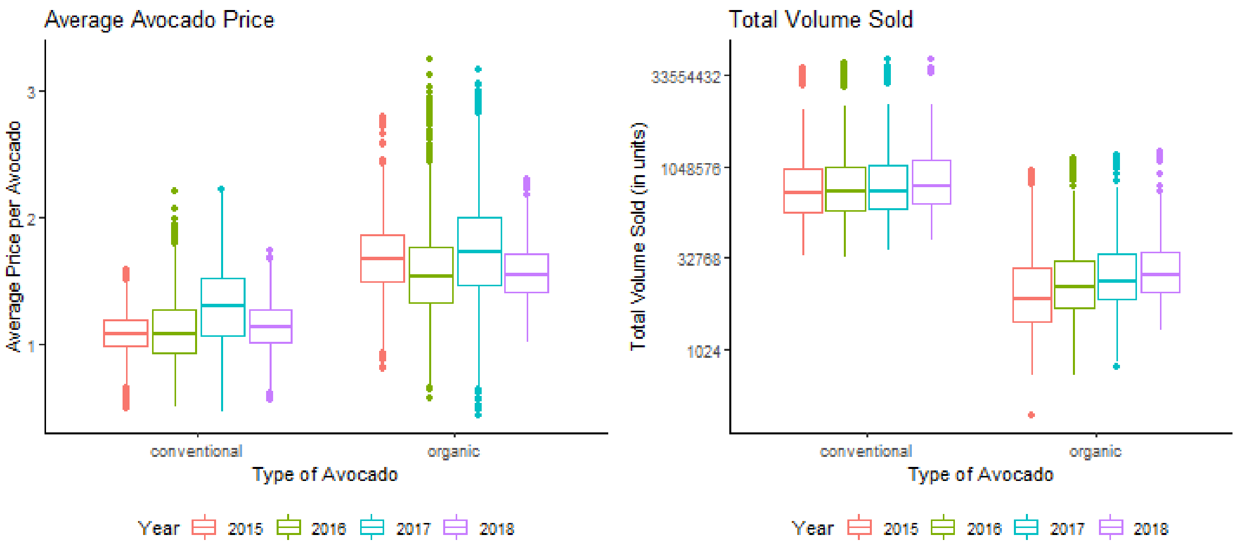


The key takeaways of Plot 1 are as follows:

- The histograms show that the average prices follow approximately a normal distribution centered around the mean prices for each avocado type.
- It is clear that organic avocados are on average more expensive than conventional avocados.
- The type of avocado is a fundamental indicator of the average price and must be taken into consideration during the model engineering phase.
- **Average avocado price**: *$1.16* (*conventional*), *$1.65* (*organic*)

2.4.2 Plot 2: Boxplot of average prices by volume and type

The avocado prices follow a normal distribution, but little is known about the range of average prices. Next, the average avocados are plotted by type and produce year. The goal is to evaluate price trends over the years.

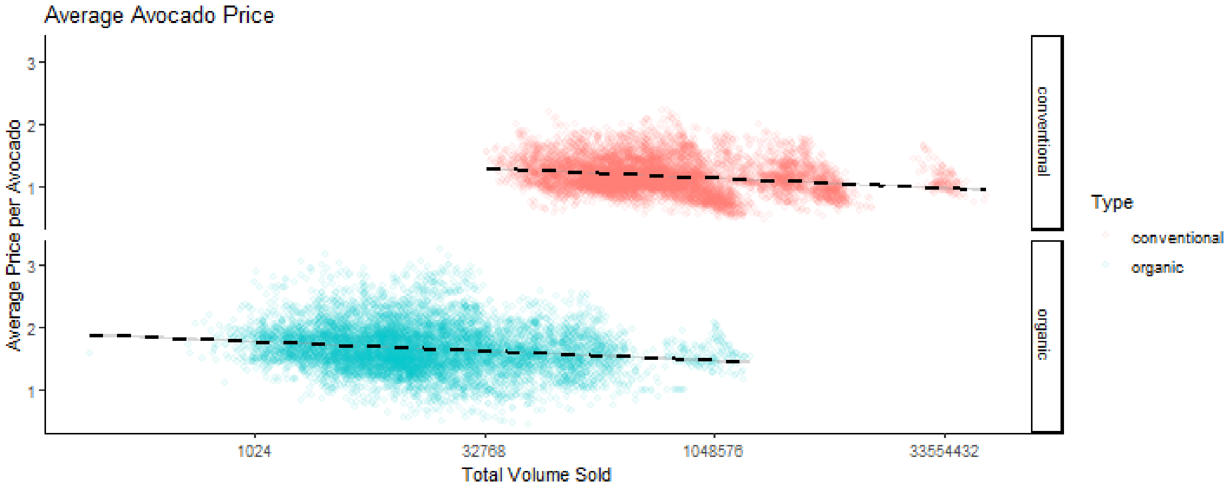


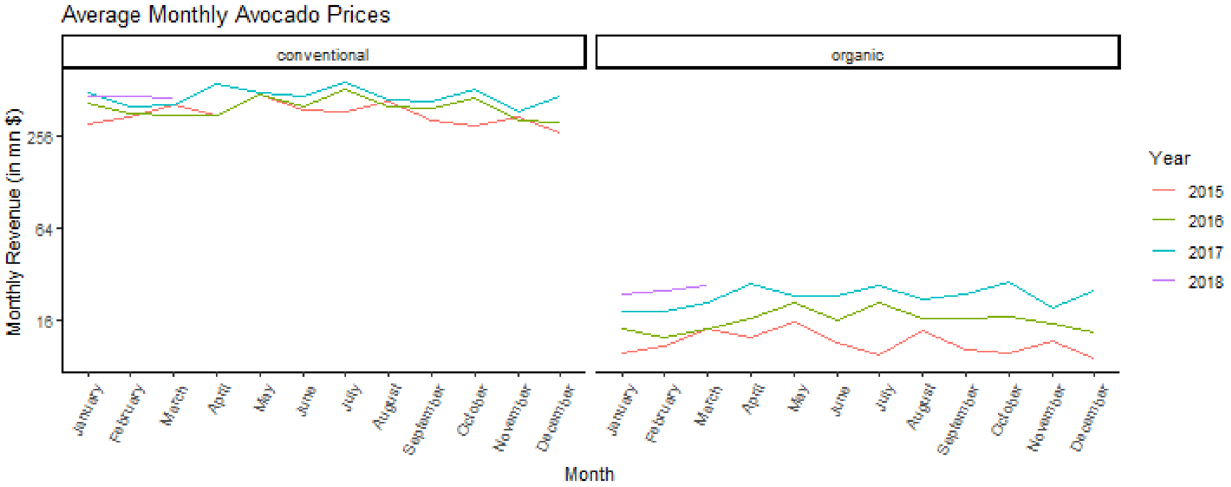
The key takeaways of Plot 2 are as follows:

- The total volume sold for avocados sold have seen an upwards trend over the observed years. The total volume sold of organic avocados have been growing at a steeper rate than that of conventional ones.
- There is significant lower volume of organic avocados available in the market when compared to the conventional ones. This explains partially the difference in average prices of both types of avocados.
- The year of the produce influences greatly the average prices – 2016 and 2018 have seen a drop in the average price. However, it is not straight forward that this represents a typical business cycle: every second year the prices either increase or decrease.
- Outliers are more frequent in the total volume sold of conventional avocados and in the organic average prices. These values can lead to less effective predictive models and must be taken into account.

2.4.3 Plot 3: Average prices by volume sold, type, and date

The next plot has two goals, the first one is to understand the business cyclicality over the months and years. The second goal analyses, as a first model hypothesis, whether the total volume sold can be a good proxy for the average avocado prices.



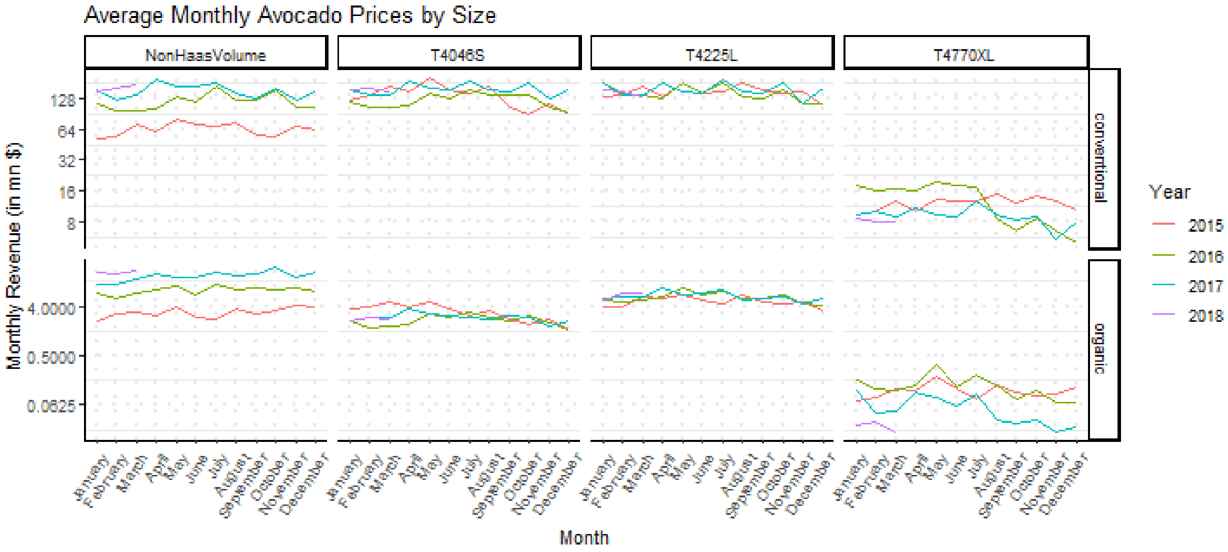


The key takeaways of Plot 3 are as follows:

- The function that estimates the average price with the volume sold captures lightly the overall slope trend of prices, but is far from being a good model. Prices seem to get lower with the total volume in an aggregated basis.
- Data for average prices for 2018 is only available up to March. The implications of the lack of data for the remaining months of 2018 is that the models will fail to yet observe a full 4th business cycle.
- Prices towards the end of each year seem to plummet, while they trend to peak in Summer and Fall.
- The month of the year influences significantly the average price as it captures the typical avocado business cycle in the US
- It seems to be a good time to invest in the organic avocado produce, as the monthly revenue YoY has been increasing every year.
- 2017 was the best business year overall. It is not possible to explain why 2017 was the best year, but it could be due to hidden market factors that are not shown in the data.

2.4.4 Plot 4: Total Rev sold by Size, Year & Type

It is clear now that the volume, type and data influence average prices, but little is known about how the different sizes of avocados influence the prices. The next plot attempts at answering this question.

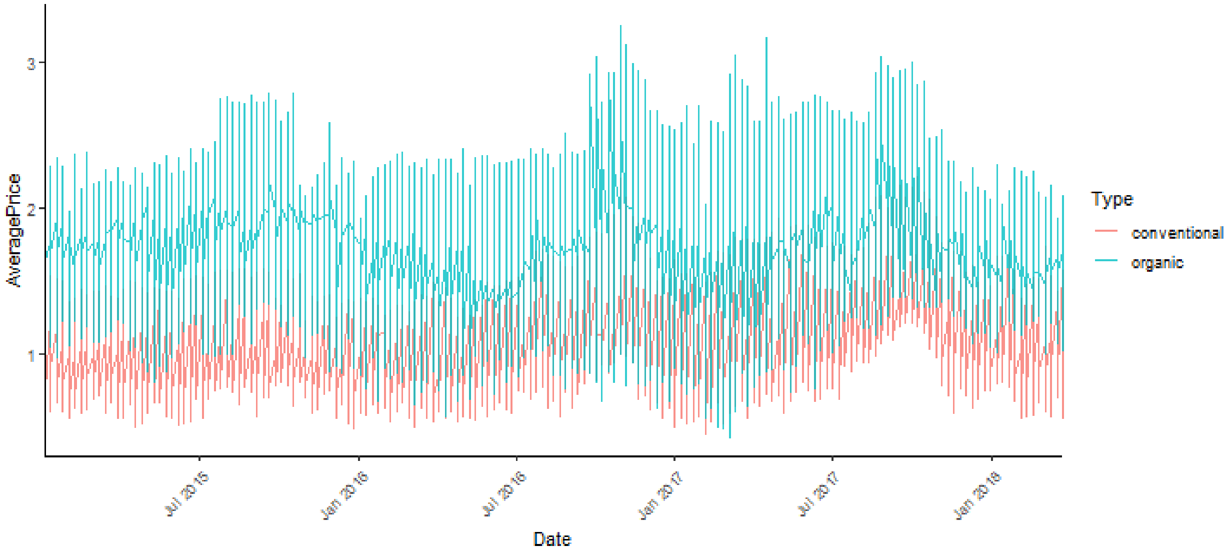


The key takeaways of Plot 4 are as follows:

- The monthly revenue for non-Haas avocados have been growing YoY, for both types (organic and conventional).
- The monthly revenue for organic avocados appears to be constant and with little growth for and large avocados.
- Large avocados have the most unstable average prices of all sizes and also these are the avocados that are produced and sold the least.

2.4.5 Plot 5: Average Price over time by type

One more plot is required to fully visualize the business cycle of the avocados over the period between 2015 and 2018. This next plot will be the benchmark by which the models in Chapter 3 will be compared.



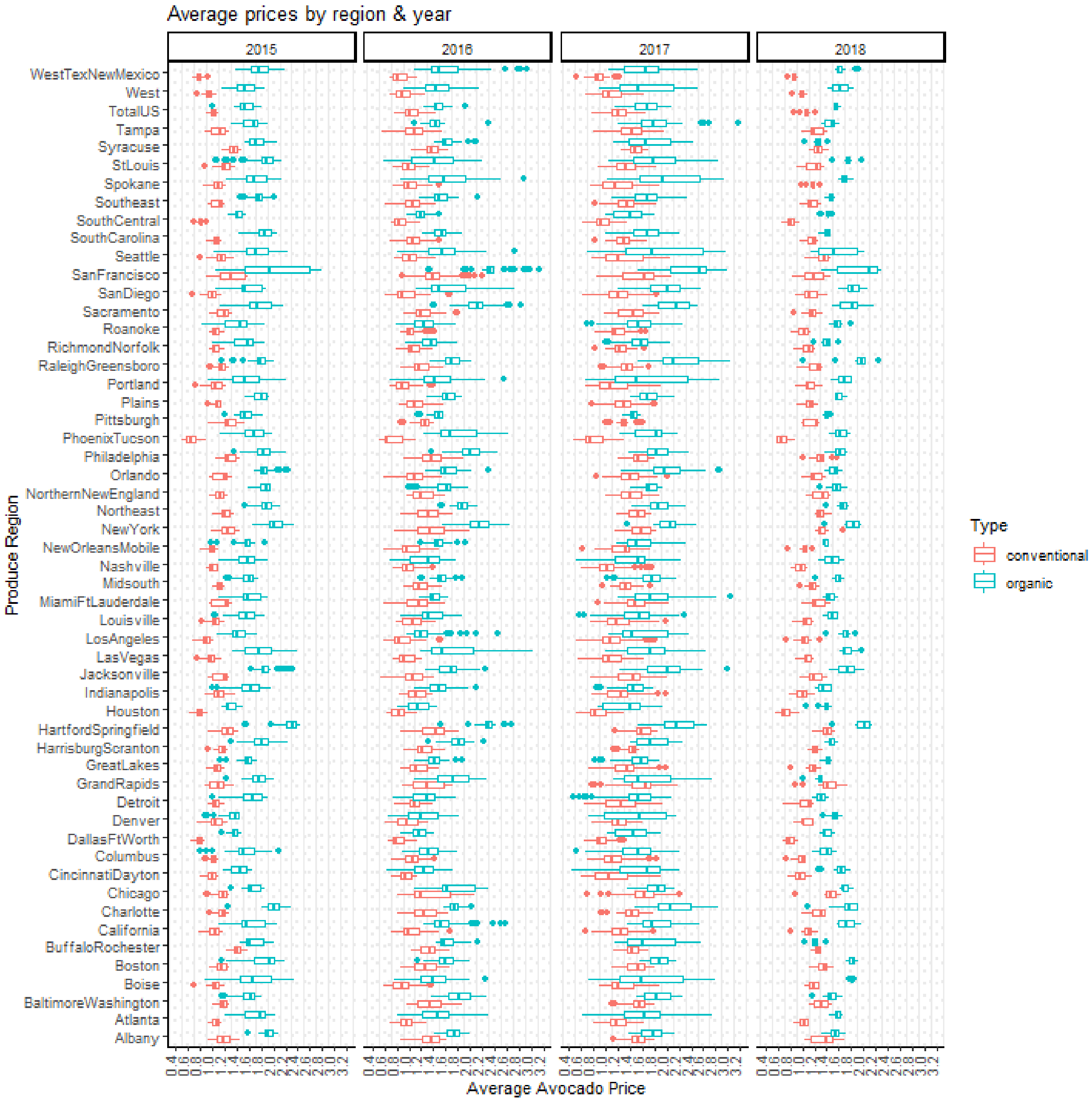
The key takeaways of Plot 5 are as follows:

- Prices vary with high amplitude over the observed period. The reason for this behaviour is the number of outlier data points that the initial dataset contains.

- Prices rise on average during summer, but there is no apparent reason why prices fall so steeply in the beginning of 2018.
- The price range for organic avocados is clearly larger that that of conventional avocados.

2.4.6 Plot 6: Average price ranges for each region

The last dimension to be analysed in this section is the price per produce region. The next plot highlights how competitive regions are and how they compare with one another.



The key takeaways of Plot 6 are as follows:

- There are 54 regions in the dataset and no one is the same in terms of average price behavior.
- The reason why average prices seem to fluctuate so heavily is due to the distinct behaviors of each US state. There are extreme states, where the average avocado prices vary very aggressively over the year

e.g. Portland.
- It will be challenging to perform a regression model over 54 regions, however it is expected that not all will have the same statistical relevance.

3. METHODS: MODEL ENGINEERING

Several predictive models are built in this chapter to estimate the average price of avocados. Firstly, a benchmark model is built, the average model, then a multi-regression model is designed and finally a decision tree model will be tested. The judges of the model performances are the $RMSE$ and the R^2. Both mathematical expressions can be seen below.

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\ \hat{y}_{u,i} - y_{u,i})}$$

$$R^2 = 1 - \frac{\sum(y_i - \bar{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

The main goal of this section is to explore different methods to predict the average price of avocados and with this exercise observe how the $RMSE$ and R^2 evolve.

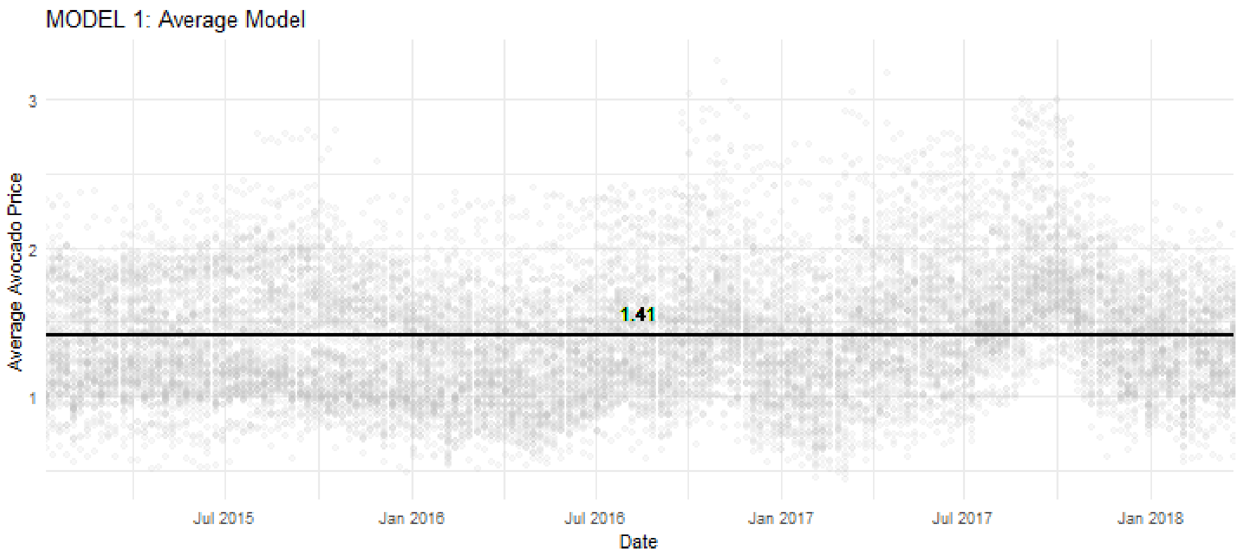
3.1 Model 1: Average Model

The first model is the simplest, but also one of the most important as it will set the benchmark KPI values to look forward to beat. This model is the average model in which it is assume the same average avocado price to all the samples. The average avocado price is computed using the **train_data** set.

```
# MODEL 1: Train model
mean_avocado_price <- mean(train_data$AveragePrice)

# Validate model 1
model1.RMSE <- RMSE(test_data$AveragePrice, mean_avocado_price)
cat("Assuming the same average avocado price of: ",
    round(mean_avocado_price, 2),
    " the RMSE (Model 1: Average) = ",
    model1.RMSE)
```

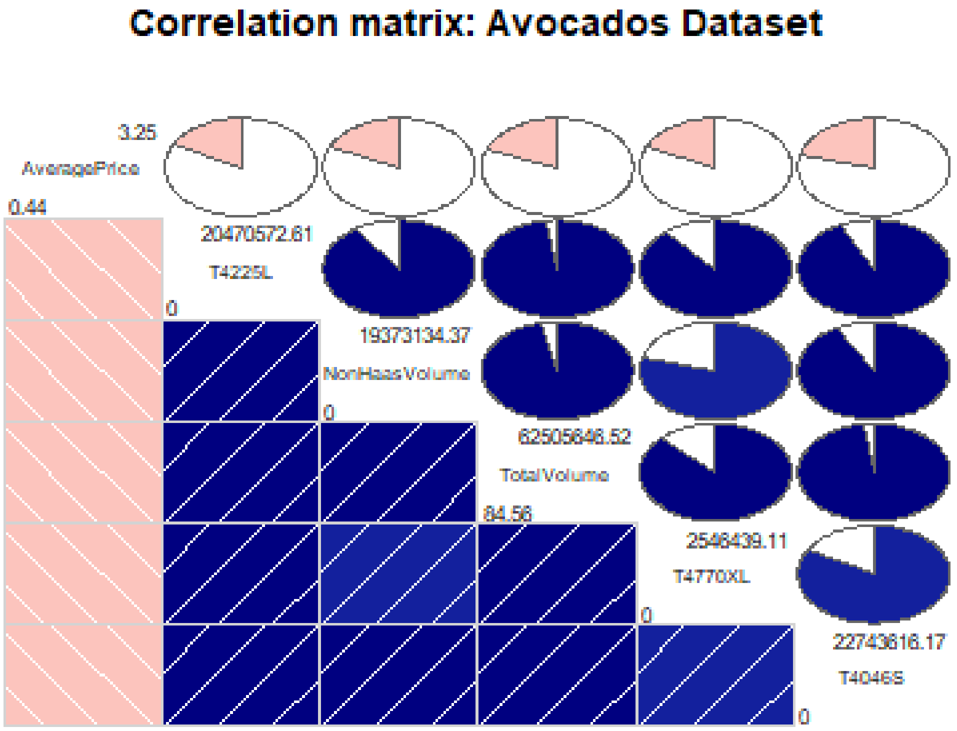
```
## Assuming the same average avocado price of:  1.41  the RMSE (Model 1: Average) =  0.3989559
```

From now on the **benchmark** for **RMSE is 0.3989559**. The benchmark R^2 will be calculated in the next multi-linear model. Below is the plot that depicts how well assuming the same average price describes the price fluctuations over time.



3.2 Model 2: Linear Regression

There are a few useful steps that should be done before starting developing regression models. The first step is to observe how the different variables correlate among one another on the **avocados** dataset. This plot is shown below.



The **corregram** function plots how all columns/variables of the **avocados** dataset are correlated. Here, it is possible to observe that total volume and individual small, large, XL and non-Haas volumes are highly correlated. However, there is no information about the categorical variables. These will be evaluated later during the multi-linear regression model.

```r
# Prepare data
new_train_data <- train_data %>% select(-c(Date, T4046S, T4225L, T4770XL, Day, NonHaasVolume))
new_train_data <- new_train_data %>% filter(!Region %in% c("RaleighGreensboro",
                                                           "Chicago",
                                                           "Boston",
                                                           "BaltimoreWashington"))
new_train_data.all <- model.matrix(~., new_train_data)[,-1]
new_train_data.x <- model.matrix(AveragePrice~., new_train_data)[,-1]
new_train_data.y <- new_train_data$AveragePrice
new_test_data <- test_data %>% select(-c(T4046S, T4225L, T4770XL, Day, NonHaasVolume))
new_test_data <- new_test_data %>% filter(!Region %in% c("RaleighGreensboro",
                                                         "Chicago",
                                                         "Boston",
                                                         "BaltimoreWashington"))
new_test_data.all <- new_test_data
new_test_data <- new_test_data.all %>% select(-c(Date))
new_test_data.xy <- model.matrix(~., new_test_data)[,-1]
new_test_data.x <- model.matrix(AveragePrice~., new_test_data)[,-1]
new_test_data.y <- new_test_data$AveragePrice
```

Before delving deeper into the regression models, further data preparation is needed. These steps were based on previous small data exploration steps and also based on the evaluation of p-values. The complete list of data transformation steps are as follows:

- Create new training, **new__train__data**, and testing data, **new__test__data**, datasets without the Date, T4046S, T4225L, T4770XL, Day, and NonHaasVolume columns. This step is necessary because some of these variables are highly correlated and won´t add much value to the training of the models
- Remove the regions: *RaleighGreensboro*, *Chicago*, *Boston*, and *BaltimoreWashington* as the p-value in the multi-linear model have shown very high p-values.
- Transform categorical variables such as *Type* and *Region* into a matrix of columns to ease the injection of data for the model training phase.
- Repeat all the above steps for the **new__test__data**

Now that all the data preparation is finalized, it is possible to derive a simple linear model as a function of the *TotalVolume* to determine the R^2 benchmark.

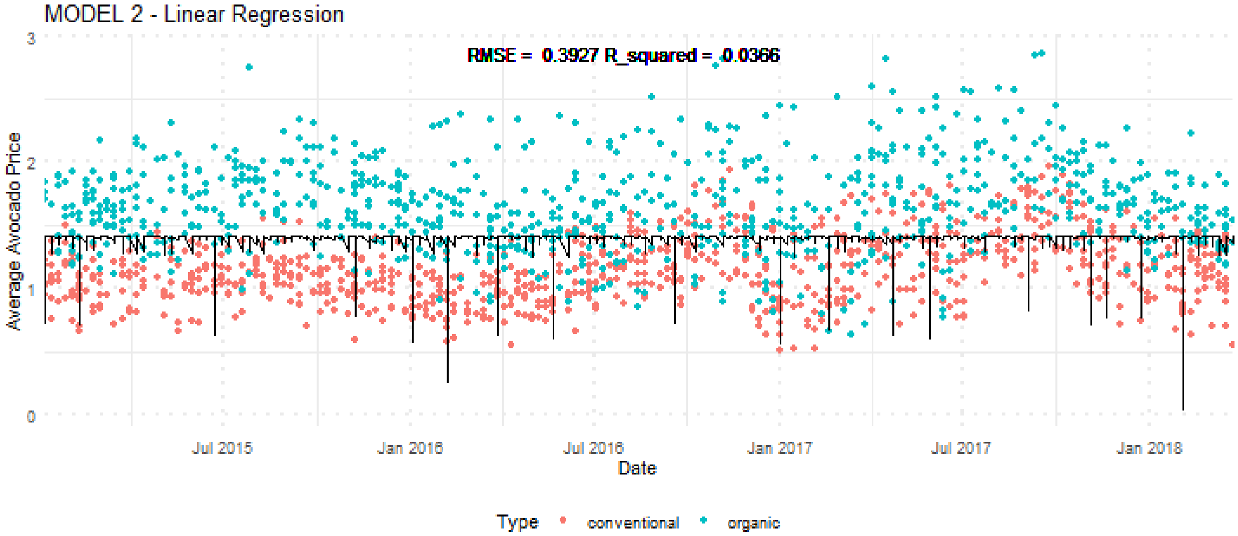
```r
# MODEL 2: Train model
model2 = lm(AveragePrice ~ TotalVolume, data = new_train_data)
model2.summary <- summary(model2)

# Validate model
model2.validation <- model2 %>% predict(new_test_data) %>% as.vector()
model2.RMSE <- RMSE(new_test_data$AveragePrice, model2.validation)
model2.R2 <- model2.summary$r.squared
cat("Model 2 - Linear Model: RMSE =",
    model2.RMSE,
    "Model 2 - Linear Model: R-squared =",
    model2.R2)
```

```
## Model 2 - Linear Model: RMSE = 0.3927159 Model 2 - Linear Model: R-squared = 0.03660494
```

The benchmark for the R^2 is **0.03660494**. Notice that the $RMSE$ is not significantly better than the average model $RMSE$:

- Model 1 - Average model: $RMSE = 0.3989559$
- Model 2 - Linear model: $RMSE = 0.3927159, R^2 = 0.03660494$

The plot with the results of model 2 – linear model are depicted below in black.



3.3 Model 3: Multi Linear Regression

The model 3 encompasses all variables in the **new__train__data** data frame. The code below represents the refined model after removing all $p - values > .05$. It is expected that this model is significantly better than the previous 2 models. This model is also significantly more complex and harder to compute because of the sheer number of variables.

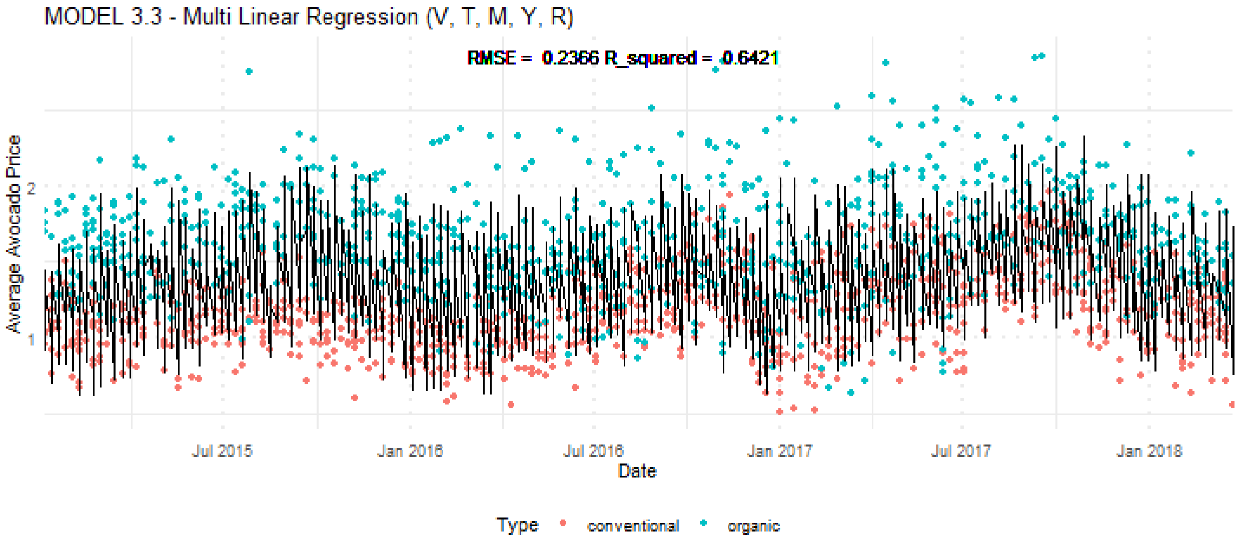
```
# MODEL 3.3: T model: (V, T, M, Y, R)
model3 = lm(AveragePrice ~ TotalVolume + Type + Month + Year + Region, data = new_train_data)
model3.summary <- summary(model3)

# MODEL 3.3: Validate model: (V, T, M, Y, R)
model3.validation <- model3 %>% predict(new_test_data) %>% as.vector()
model3.RMSE <- RMSE(new_test_data$AveragePrice, model3.validation)
model3.R2 <- model3.summary$r.squared
cat("Model 3.3 - Multi Linear Model (V, T, M, Y, R): RMSE =",
    model3.RMSE,
    "Model 3.3 - Multi Linear Model (V, T, M, Y, R): R-squared =",
    model3.R2)
```

```
## Model 3.3 - Multi Linear Model (V, T, M, Y, R): RMSE = 0.2366201 Model 3.3 - Multi Linear Model (V,
```

The first striking point in this model is its ability to follow more closely the business cycles (see plot below). Five variables were considered for this model: the *Total volume sold*, *Type of avocado*, *Produce month*, *Produce year* and *Region*.

- Model 1 - Average model: $RMSE = 0.3989559$
- Model 2 - Linear model: $RMSE = 0.3927159, R^2 = 0.03660494$
- Model 3 – Multi-linear model: $RMSE = 0.2366201, R^2 = 0.6420727$

Both the $RMSE$ and R^2 values improved significantly and are the new benchmark to beat.



Note: Several combinations of the predictors used for this model were tried but will not be depicted here because the model results were not as good as this one.

3.4 Model 4: Lasso Regression

Next, the lasso regression model is built. This regression method shrinks values towards a central point, like the mean. This regression performs regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. Some data points may become zero and eliminated from the model.

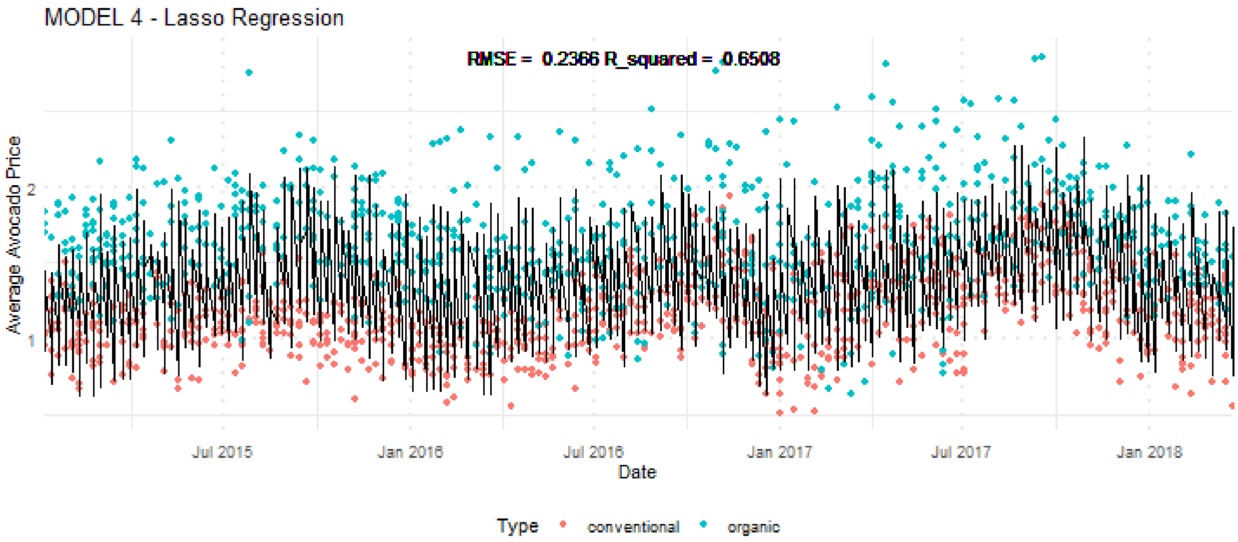
```r
########################################
# MODEL 4.2: Lasso regression (V, T, M, Y, R)
# Evaluate best lambda using cross-validation
model4.2.cv <- cv.glmnet(new_train_data.x,
                         new_train_data.y,
                         alpha = 1) #alpha = 1 for lasso regression

# Fit the final model on the training data w/ min lambda
model4.2 <- glmnet(new_train_data.x,
                   new_train_data.y,
                   alpha = 1, #alpha = 1 for lasso regression
                   lambda = model4.2.cv$lambda.min)

# MODEL 4.2: Validate model: (V, T, M, Y, R)
model4.2.validation <- model4.2 %>% predict(new_test_data.x) %>% as.vector()
model4.2.RMSE <- RMSE(new_test_data$AveragePrice, model4.2.validation)
model4.2.R2 <- R2(model4.2.validation, new_test_data$AveragePrice)

cat("Model 4 - Lasso Model: RMSE =",
    model4.2.RMSE,
    "Model 4 - Lasso Model: R-squared =",
    model4.2.R2)
```

```
## Model 4 - Lasso Model: RMSE = 0.236575 Model 4 - Lasso Model: R-squared = 0.650799
```

This model also considers the *Total volume sold*, *Type of avocado*, *Produce month*, *Produce year* and *Region* as variables. The results are not much better that those of the multi-linear model, nevertheless there was a slight improvement in the model performance vis-à-vie the previous models:

- Model 1 - Average model: $RMSE = 0.3989559$
- Model 2 - Linear model: $RMSE = 0.3927159, R^2 = 0.03660494$
- Model 3 – Multi-linear model: $RMSE = 0.2366201, R^2 = 0.6420727$
- Model 4 – Lasso model: $RMSE = 0.236575, R^2 = 0.650799$

There is therefore a new benchmark to beat. Several variants of linear regressions were tried: Lasso, Ridge, and Elastic Net Regression. However, the lasso model performed the best among these – For more details, please refer to the R script. The results of the lasso model are plotted below.



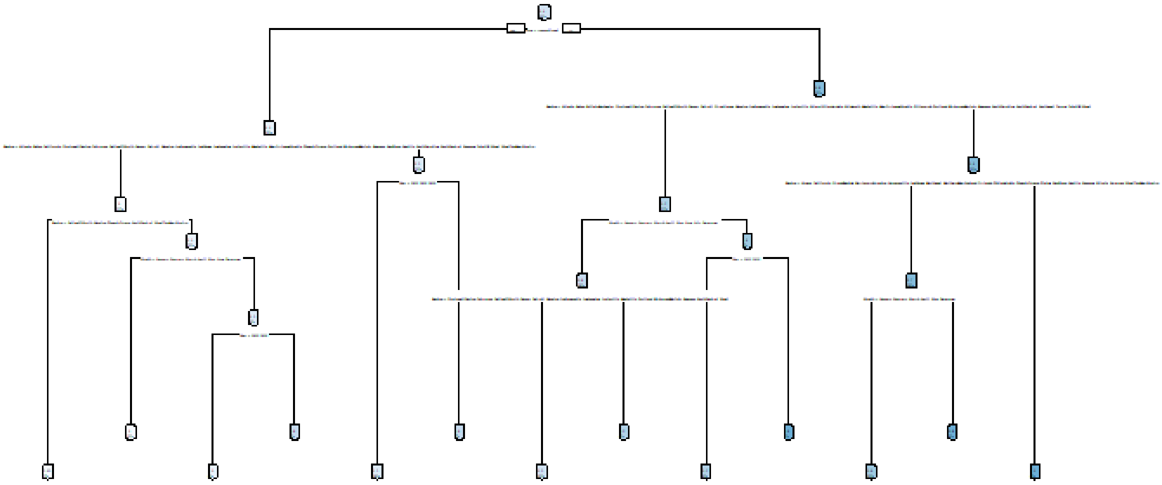
3.5 Model 5: Decision Tree (Regression)

Decision trees can be either used for building a classification or a regression model. In this case the decision tree regression model is explored. The goal is to verify whether a decision tree can beat the benchmark set by the Model 4 – Lasso regression. One of the key advantages to use a decision model is the fact that it is generally quicker to execute than a multi-linear model. This is important as this project uses many predictors.

```
########################################
# MODEL 5: Decision Tree (Regression) (V, T, M, Y, R)
# MODEL 5: Train model: (V, T, M, Y, R)
model5 <- rpart(AveragePrice ~ .,
                data = new_train_data,
                method = "anova")
```

```
## Predictor importance
##      Type TotalVolume      Region       Month        Year
##      39          33          19           5           4
```

In the decision tree trained with the **new__train__data** data frame, the order of the predictors importance is: 1) Type, 2) TotalVolume, 3) Region, 4) Month and 5) Year. Below is the decision tree plotted. Due to the complexity of the categorical variables, in particular the complexity of the Region predictor, It is very challenging to read the tree.



The trained model has **12 nodes**, which maximized the model accuracy. The model performance in terms of $RMSE$ and R^2 is below analyzed.

```
# MODEL 3: Validate model: (V, T, M, Y, R)
model5.validation <- model5 %>% predict(new_test_data) %>% as.vector()
model5.RMSE <- RMSE(new_test_data$AveragePrice, model5.validation)
model5.R2 <- R2(model5.validation, new_test_data$AveragePrice)

cat("Model 5 - Decision Tree (Regression): RMSE =",
    model5.RMSE,
    "Model 5 - Decision Tree (Regression): R-squared =",
    model5.R2)
```

```
## Model 5 - Decision Tree (Regression): RMSE = 0.2395234 Model 5 - Decision Tree (Regression): R-square
```

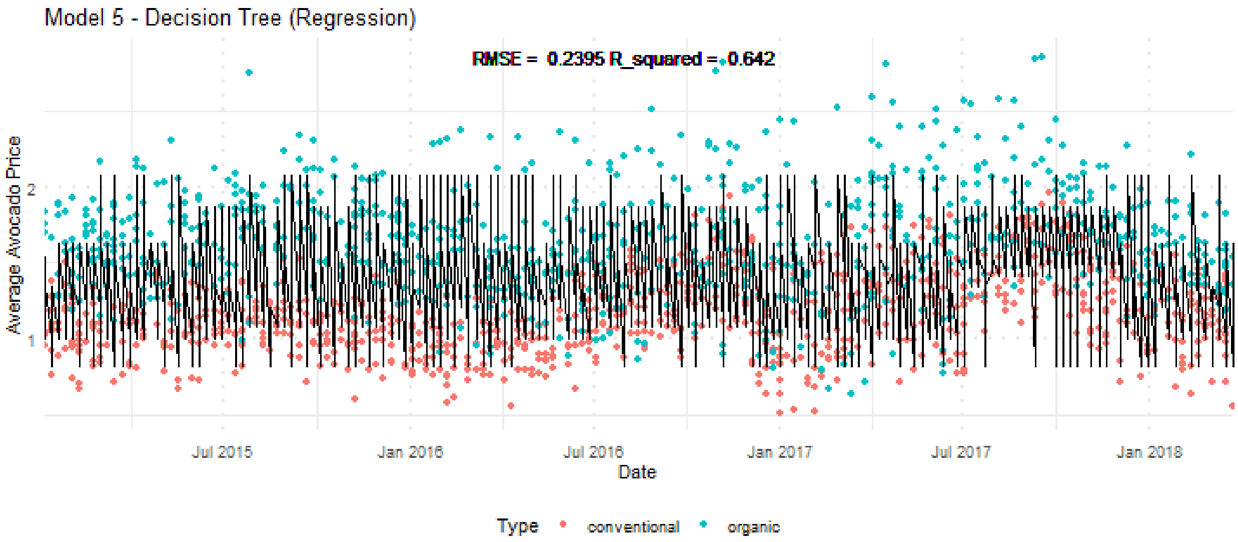
The decision tree did not improve the Lasso regression model of the previous section. The KPIs are as follows:

- Model 1 - Average model: $RMSE = 0.3989559$
- Model 2 - Linear model: $RMSE = 0.3927159, R^2 = 0.03660494$
- Model 3 – Multi-linear model: $RMSE = 0.2366201, R^2 = 0.6420727$
- Model 4 – Lasso model: $RMSE = 0.236575, R^2 = 0.650799$
- Model 5 – Decision Tree (Regression): $RMSE = 0.2395234, R^2 = 0.6420457$

This model concludes the model engineering phase of this project. The results of this model are plotted below:

Table 3: Model Results: Summary

Models	RMSE Value	R-squared
MODEL 1 - Mean Model	0.3989559	N/A
MODEL 2 - Linear Regression	0.3927159	0.0366049
MODEL 3 - Multi Linear Regression	0.2366201	0.6420727
MODEL 4 - Lasso Regression	0.2365750	0.650799
MODEL 5 - Decision Tree (Regression)	0.2395234	0.6420457



3.5 Discussion

This chapter presented 5 different predictive models to estimate the average avocado price in a given time. The summary of results is:

The **best model** is the **Model 4 – Lasso Regression**, because it has the lowest $RMSE$ and the highest R^2 value. Model 4 – Lasso model: $RMSE = 0.236575, R^2 = 0.650799$ It was unexpected that the Model 5 – Decision Tree did not perform better than the Lasso regression, because Decision Trees bisect the data into fine regions, whereas Logistic Regression fits a single line to divide the data.

There are a few hypotheses on the reasons why none of the models came close to a R^2 of *0.8*. Firstly, the **avocados** dataset is not necessarily large, which influence the ability of models to see patterns. Secondly, the year of 2018 is not complete and does not show a proper business cycle. Last but not least, there are other economic factors that are not capture in the dataset e.g. CPI, consumer behaviour indexes, etc.

4. CONCLUSION

The goals of this report were to apply data science methods learned in the previous courses of the Data Science Certificate (HarvardX), and to build an acceptable regression model given the available avocado data. The data used for this project is available for free on Kaggle.

The report covers basic and advanced data exploration/visualization, and also, once the first data insights are known, explores 5 different regression models to predict the average avocado price.

This chapter discusses the model results, and what can be done in the future to improve the models.

4.1 Discussion

Data science models are quite powerful even when used with limited data. For finance modelling, an R^2 close to .7 is a respected result, due to the psychological nature of markets. The high quality of the data provided a solid base to all the models. It was not possible to determine whether the models would have performed better with a larger volume of data. The computing power and memory levels of the hardware used for the project was adequate and there was no need to resource to other platforms. This fact could be different with a substantial larger dataset.

It is often very optimistic to predict financial variables, such as the average price of avocados, when looking only at the avocados market. Financial markets are entangled and definitely the available data could have been supplemented with further with other economic indicators. This project would not have been possible without the data visualization capabilities of R. Arguably, data visualization is the most important part of the project as it provides hints on which models/predictors are the most important to predict the prices.

The goals of these project were achieved in the sense that several types of regression models were tested. It is remarkable how easily regressions models can be derived in R. Additionally, many of the techniques and methods learned in the previous data science modules were applied successfully.

4.2 Future work and considerations

The following points should be considered in the future to further improve the average avocado pricing engine:

- Evaluate whether the unused columns could be used to improve the models.
- Supplement the avocados dataset with other economic KPIs such as CPI, S&P500, water prices, etc.
- Increase the time span of the data to include 10 years of data.
- Explore non-linear models e.g. Michaelis-Menten, Weibull, Ricker curve, Biexponential, etc.

4.3 References

- Irizarry, Rafael A., "Introduction to Data Science: Data Analysis and Prediction Algorithms in R", webpage:https://rafalab.github.io/dsbook/
- LaTeX Equation Builder, webpage:https://latex.codecogs.com/eqneditor/editor.php
- knitr::kable and kableExtra Tutorial, webpage:https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html