HarvardX: PH125.9x Data Science: Capstone - CYO Avocados Project

Carlos Dominguez Monferrer

September 4th, 2020

Contents

1 Executive summary

The key idea is to create a system to predict prices of avocados in U.S. Avocado per capita consumption grew at 405.8% from 1990-1991 to 2016-2017, and the overall fruit category in the United States grew just 28.5% over that same time period. Rapid growth of U.S. demand for fresh avocados has increased the fruit's prominence in retail sales and consumer diets. This growth is largely due to California producer and importer-funded research and promotion programs that have changed avocados image to that of a healthy superfood. Total California production has decreased slightly over time with the growth in consumption satisfied by imports, primarily from Mexico. U.S. consumers now enjoy year-round availability of avocados with more stable month-to-month prices than previously observed.

The purpose of this project is to know the influence of different variables such as total number of avocados sold, type of avocado (conventional or organic) or region on prices and predict prices of avocados in U.S. to avoid an inflation in a certain region and to help to find a city with cheap avocados.

The data that will be used has been downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. More information at https://www.kaggle.com/neuromusic/avocado-prices?select=avocado.csv and https://hassavocadoboard.com/

2 Methods/Analysis

Before creating and optimizing the algorithm, an analysis of Avocados dataset is needed to know the type of data we will work with and the influence of the different variables on average price. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost.

In order to make the code easier to understand, the Analysis section has been divided in two parts:

Data exploration:

- Number of rows and columns
- Name of the variables
- Summary of Development and Validations sets
- Number of different types, years and regions in both datasets

Data cleaning and Influence of variables on average price:

- Convert Date variable (factor) to a date.
- Relation between Date and average price.
- Relation between the type of avocado (conventional or organic) and average price.
- Relation between the city or region of the observation (region variable) and average price.
- Relation between total number of avocados sold (Total Volume) and average price.

Other variables like Product Lookup codes (PLU's) (X4046, 4225, 4770) and bags have not been used in this project.

Note: Development and Validation sets will be 80% and 20% respectively of Avocados data. Train and test sets will be 70% and 30% respectively of Development data. These percentages are based on the paper *Shahin, M. A., Maier, H. R., and Jaksa, M. B. (2004). "Data division for developing neural networks applied to geotechnical engineering." Journal of Computing in Civil Engineering,ASCE, 18(2), [105-114]*. However, it is important to know that these values depend on the size of the database.

2.1 Data exploration

2.1.1 Number of rows & columns

- Development dataset

Number of rows

```
## [1] 14601
```

Number of columns

```
## [1] 14
```

- Validation dataset

Number of rows

```
## [1] 3648
```

Number of columns

```
## [1] 14
```

2.1.2 Name of the variables

There are 14 different variables in both datasets:

```
##  [1] "X"            "Date"         "AveragePrice" "Total.Volume" "X4046"
##  [6] "X4225"        "X4770"        "Total.Bags"   "Small.Bags"   "Large.Bags"
## [11] "XLarge.Bags"  "type"         "year"         "region"
```

2.1.3 Summary stadistics

- Development dataset

```
##        X                  Date        AveragePrice    Total.Volume
##  Min.   : 0.00   2017-03-26:  100   Min.   :0.440   Min.   :      85
##  1st Qu.:10.00   2018-03-11:   97   1st Qu.:1.100   1st Qu.:   10785
##  Median :23.00   2015-11-01:   95   Median :1.370   Median :  106346
##  Mean   :24.14   2017-12-24:   95   Mean   :1.406   Mean   :  836744
##  3rd Qu.:38.00   2015-10-18:   94   3rd Qu.:1.660   3rd Qu.:  431791
##  Max.   :52.00   2016-06-19:   94   Max.   :3.250   Max.   :61034457
##                  (Other)   :14026
##      X4046              X4225              X4770             Total.Bags
##  Min.   :       0   Min.   :       0   Min.   :      0.0   Min.   :       0
##  1st Qu.:     850   1st Qu.:    2962   1st Qu.:      0.0   1st Qu.:    5054
##  Median :    8631   Median :   28665   Median :    183.3   Median :   39438
##  Mean   :  289900   Mean   :  289760   Mean   :  22873.6   Mean   :  234209
##  3rd Qu.:  111615   3rd Qu.:  147995   3rd Qu.:   6188.6   3rd Qu.:  110392
##  Max.   :22743616   Max.   :20328162   Max.   :1993645.4   Max.   :16394524
##
##    Small.Bags         Large.Bags       XLarge.Bags                type
##  Min.   :       0   Min.   :      0   Min.   :     0.0   conventional:7276
##  1st Qu.:    2823   1st Qu.:    128   1st Qu.:     0.0   organic     :7325
##  Median :   26313   Median :   2663   Median :     0.0
##  Mean   :  178392   Mean   :  52819   Mean   :  2997.9
##  3rd Qu.:   82906   3rd Qu.:  21877   3rd Qu.:   137.5
##  Max.   :12567156   Max.   :4324231   Max.   :551693.7
##
##       year                     region
##  Min.   :2015   BuffaloRochester :  282
##  1st Qu.:2015   LosAngeles       :  281
##  Median :2016   Louisville       :  281
##  Mean   :2016   SouthCarolina    :  281
##  3rd Qu.:2017   MiamiFtLauderdale:  279
##  Max.   :2018   Orlando          :  278
##                 (Other)          :12919
```

- Validation dataset

```
##        X                  Date       AveragePrice    Total.Volume
##  Min.   : 0.00   2016-11-13:  34   Min.   :0.510   Min.   :     380
##  1st Qu.:10.00   2017-01-01:  32   1st Qu.:1.100   1st Qu.:   11202
##  Median :24.00   2016-05-08:  31   Median :1.370   Median :  110821
##  Mean   :24.59   2016-06-12:  30   Mean   :1.407   Mean   :  906280
##  3rd Qu.:38.00   2017-05-14:  30   3rd Qu.:1.660   3rd Qu.:  442454
##  Max.   :52.00   2018-03-04:  30   Max.   :2.920   Max.   :62505647
##                  (Other)   :3461
##      X4046              X4225              X4770             Total.Bags
##  Min.   :       0   Min.   :       0   Min.   :      0.0   Min.   :       3
##  1st Qu.:     878   1st Qu.:    3160   1st Qu.:      0.0   1st Qu.:    5144
##  Median :    8796   Median :   30362   Median :    192.2   Median :   41356
##  Mean   :  305451   Mean   :  316747   Mean   :  22704.1   Mean   :  261375
##  3rd Qu.:  109957   3rd Qu.:  159008   3rd Qu.:   6450.6   3rd Qu.:  114307
##  Max.   :21620181   Max.   :20470573   Max.   :2546439.1   Max.   :19373134
##
##    Small.Bags          Large.Bags        XLarge.Bags               type
##  Min.   :       0   Min.   :      0   Min.   :     0.0   conventional:1850
##  1st Qu.:    2960   1st Qu.:    127   1st Qu.:     0.0   organic     :1798
##  Median :   26617   Median :   2536   Median :     0.0
##  Mean   :  197416   Mean   :  60419   Mean   :  3540.7
##  3rd Qu.:   85699   3rd Qu.:  22499   3rd Qu.:   104.5
##  Max.   :13384587   Max.   :5719097   Max.   :454343.7
##
##       year                     region
##  Min.   :2015   Pittsburgh        :  84
##  1st Qu.:2015   DallasFtWorth     :  81
##  Median :2016   LasVegas          :  81
##  Mean   :2016   Northeast         :  80
##  3rd Qu.:2017   Denver            :  78
##  Max.   :2018   HarrisburgScranton:  75
##                 (Other)           :3169
```

2.1.4 How many different types, years and regions are in both datasets

- Development dataset

Different types

```
## [1] 2
```

Different years

```
## [1] 4
```

Different regions

```
## [1] 54
```

- Validation dataset

Different types

```
## [1] 2
```

Different years

```
## [1] 4
```

Different regions

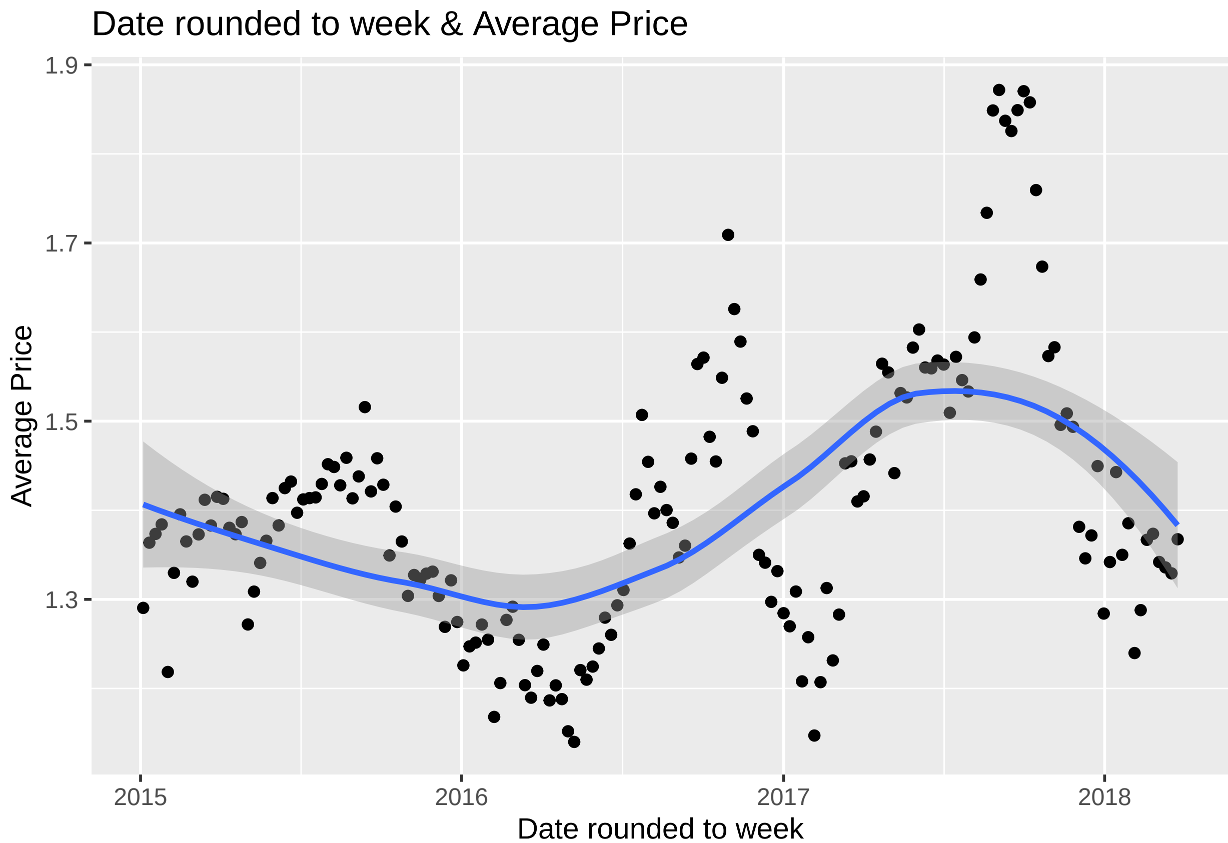
```
## [1] 54
```

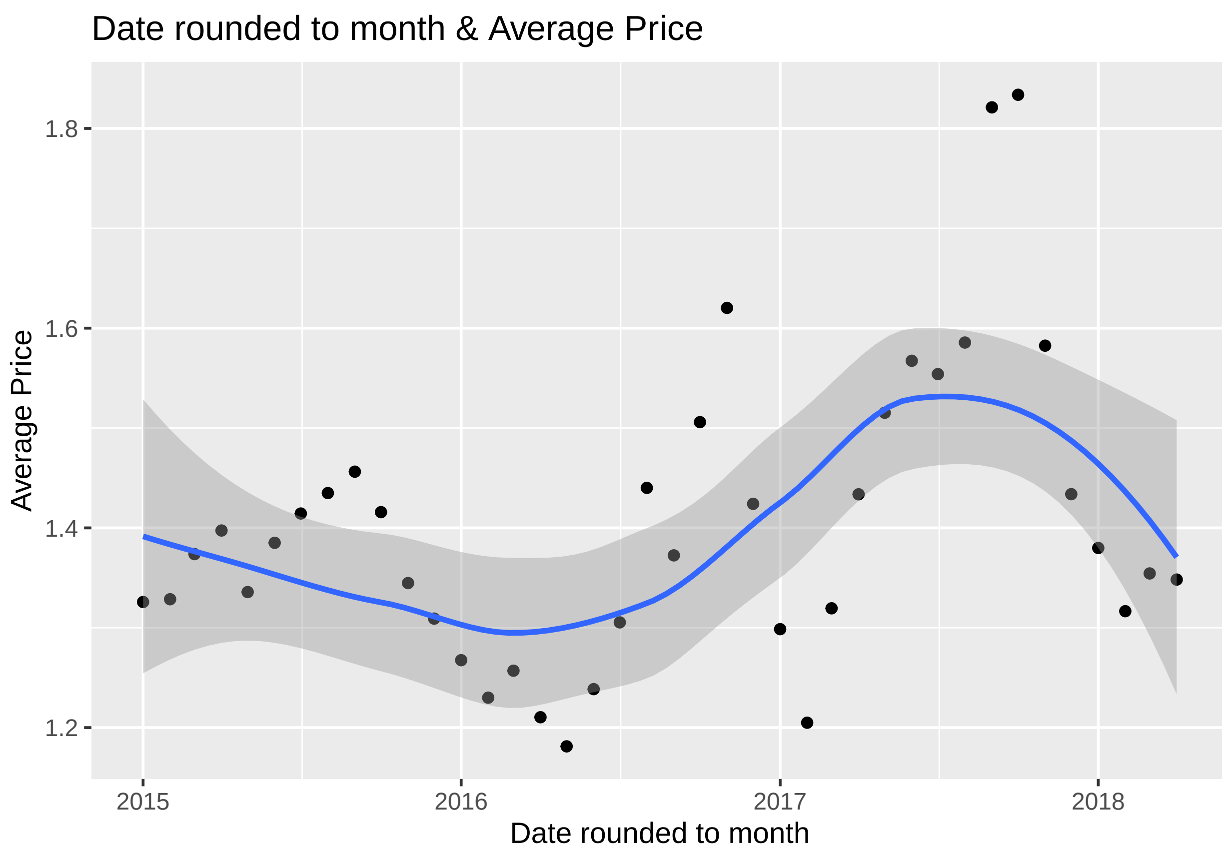
2.2 Data cleaning and Influence of variables on rating

2.2.1 Date & Average Price

In order to do a complete analysis of the influence of Date on average price, 2 graphs are plotted:

- Date rounded to week.
- Date rounded to month.

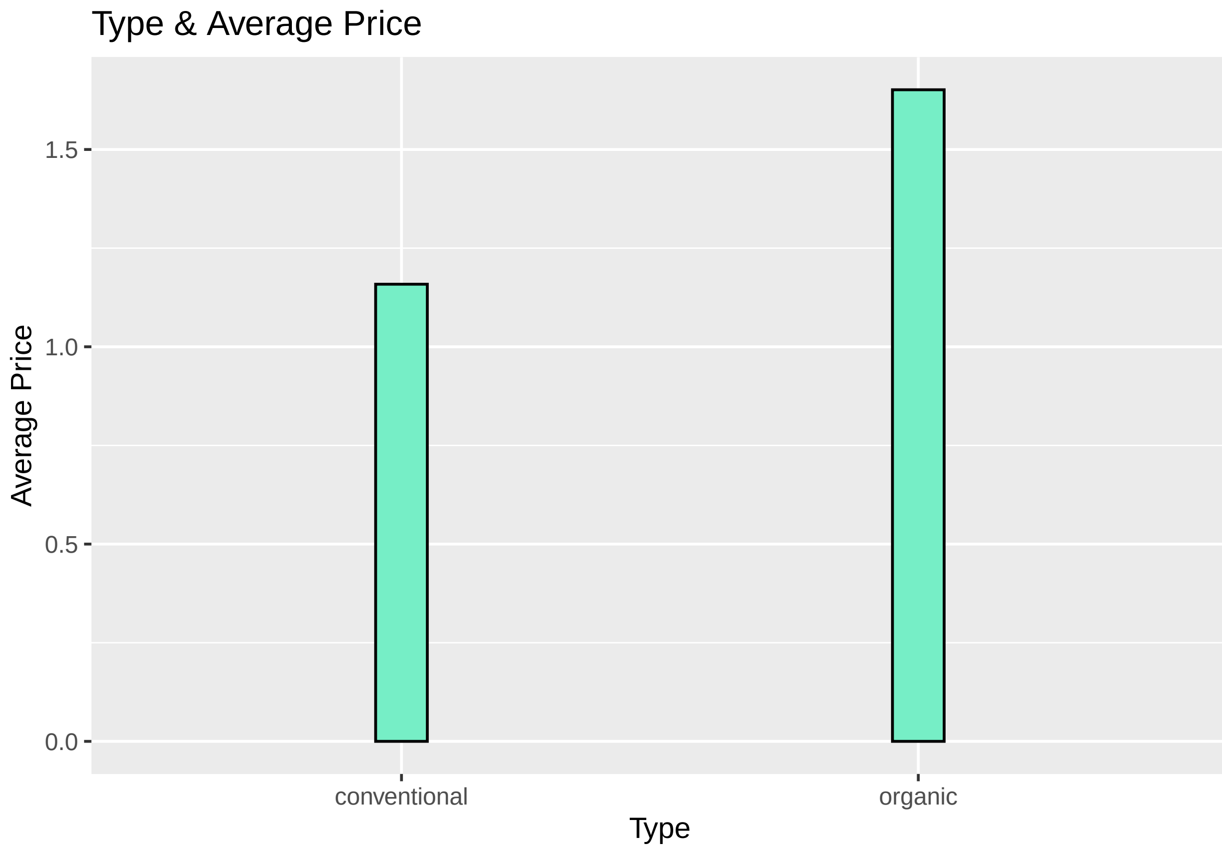




Conclusion 1.-: There is strong evidence of a date effect on average price.

2.2.2 Type & Average Price

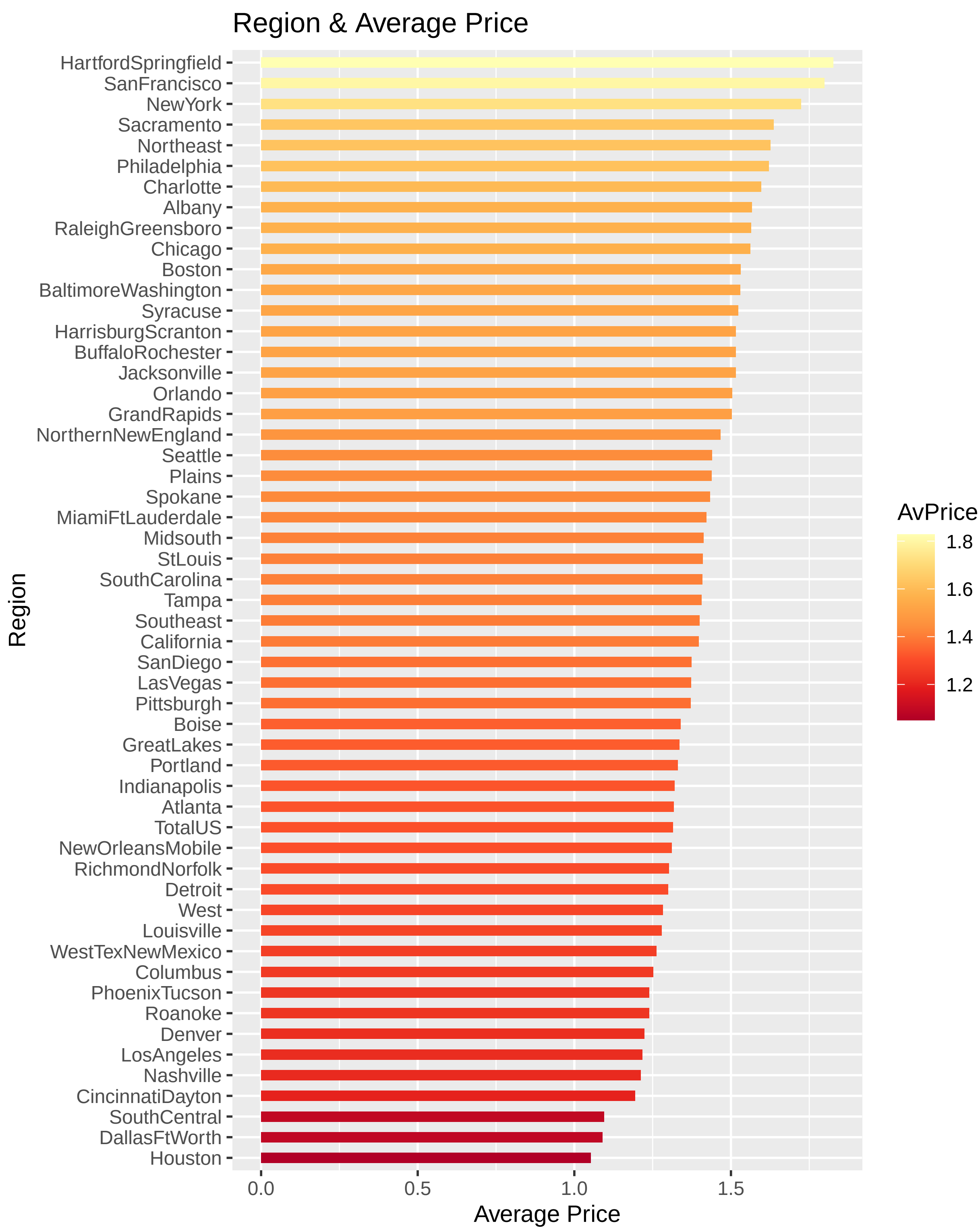
Relation between the type of avocado (conventional or organic) and Average Price.



Conclusion 2.-: There is strong evidence of a type effect on average price.

2.2.3 Region & Average Price

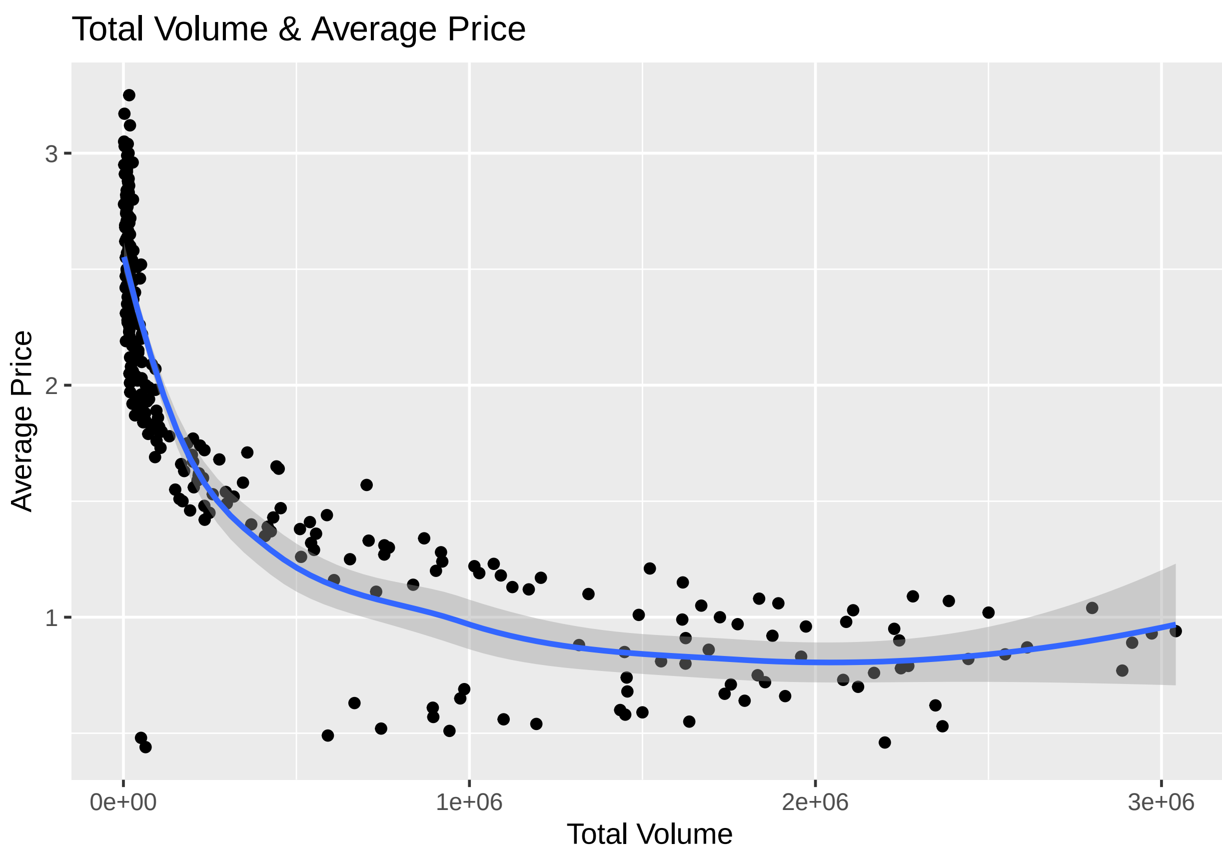
Relation between the city or region of the observation (region variable) and Average Price.



Conclusion 3.-: There is strong evidence of a region effect on average price. Hartford–Springfield is the most expensive region with an average price of 1.8 dollars per unit and Houston is the cheapest city with an average price of 1.2 dollars per unit. A difference of 66%!

2.2.4 Total volume & Average Price

Relation between total number of avocados sold (Total Volume) and Average Price.



Conclusion 4.-: There is strong evidence of a Total Volume effect on average price.

3 Results

3.1 Training process

To train our algotithm, we will calculate first RMSE without regularization technique.

3.1.1 Just the average

```
## # A tibble: 1 x 2
##   Model             RMSE
##   <chr>            <dbl>
## 1 Just the average 0.39861
```

3.1.2 Date effect

```
## # A tibble: 2 x 2
##   Model             RMSE
##   <chr>            <dbl>
## 1 Just the average 0.39861
## 2 Date Effect      0.37223
```

3.1.3 Type effect

```
## # A tibble: 3 x 2
##   Model             RMSE
##   <chr>            <dbl>
## 1 Just the average 0.39861
## 2 Date Effect      0.37223
## 3 Type Effect      0.31506
```

3.1.4 Region effect

```
## # A tibble: 4 x 2
##   Model             RMSE
##   <chr>            <dbl>
## 1 Just the average 0.39861
## 2 Date Effect      0.37223
## 3 Type Effect      0.31506
## 4 Region Effect    0.36834
```

3.1.5 Total Volume effect

```
## # A tibble: 5 x 2
##   Model                RMSE
##   <chr>               <dbl>
## 1 Just the average    0.39861
## 2 Date Effect         0.37223
## 3 Type Effect         0.31506
## 4 Region Effect       0.36834
## 5 Total Volume Effect 0.42521
```

Due to Type and Region variables got the smallest RMSE values, we will combine them in order to check if we can reduce the Root Mean Squared Error.

3.1.6 Type + Region effect

```
## # A tibble: 6 x 2
```

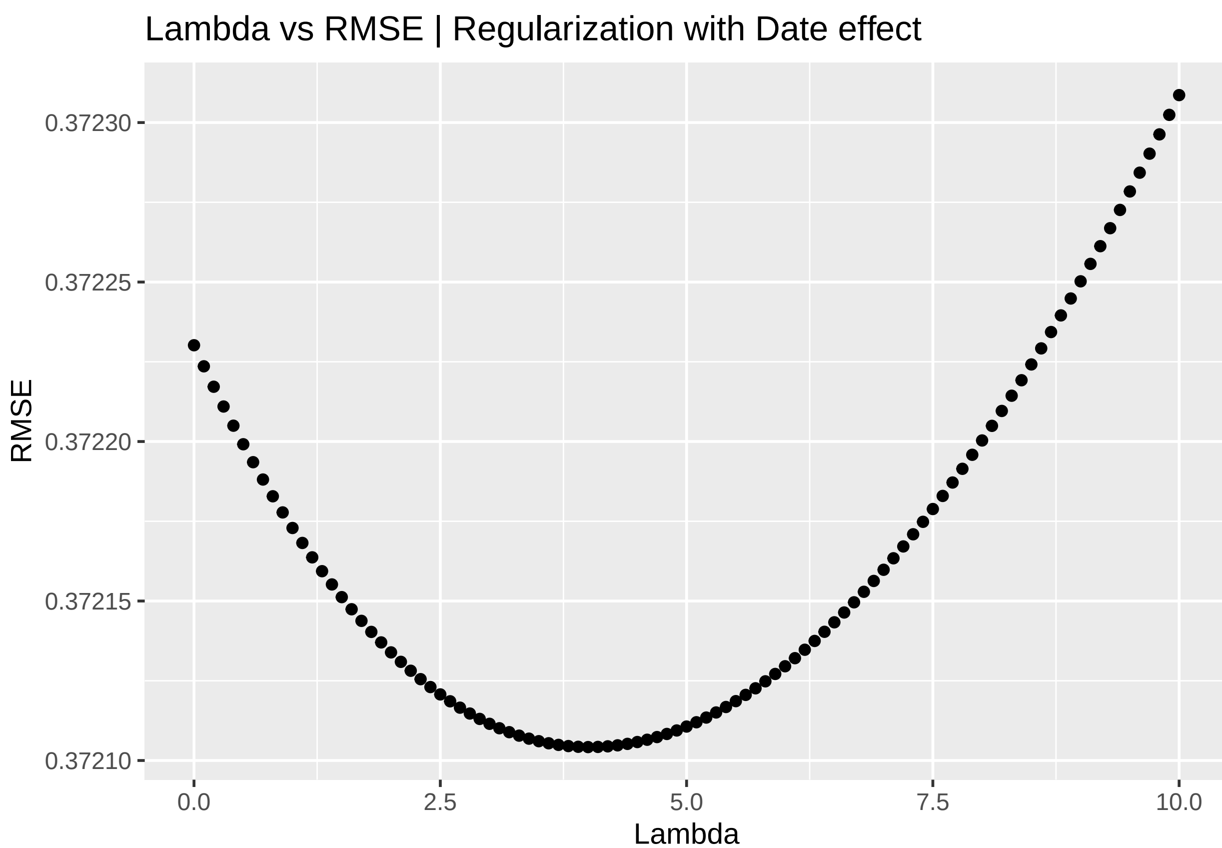
```
##   Model                    RMSE
##   <chr>                   <dbl>
## 1 Just the average      0.39861
## 2 Date Effect           0.37223
## 3 Type Effect           0.31506
## 4 Region Effect         0.36834
## 5 Total Volume Effect   0.42521
## 6 Type + Region Effects 0.27161
```

Know, we will calculate RMSE with regularization technique.

3.1.7 Regularization with Date effect

Lambda value:

```
## [1] 4
```

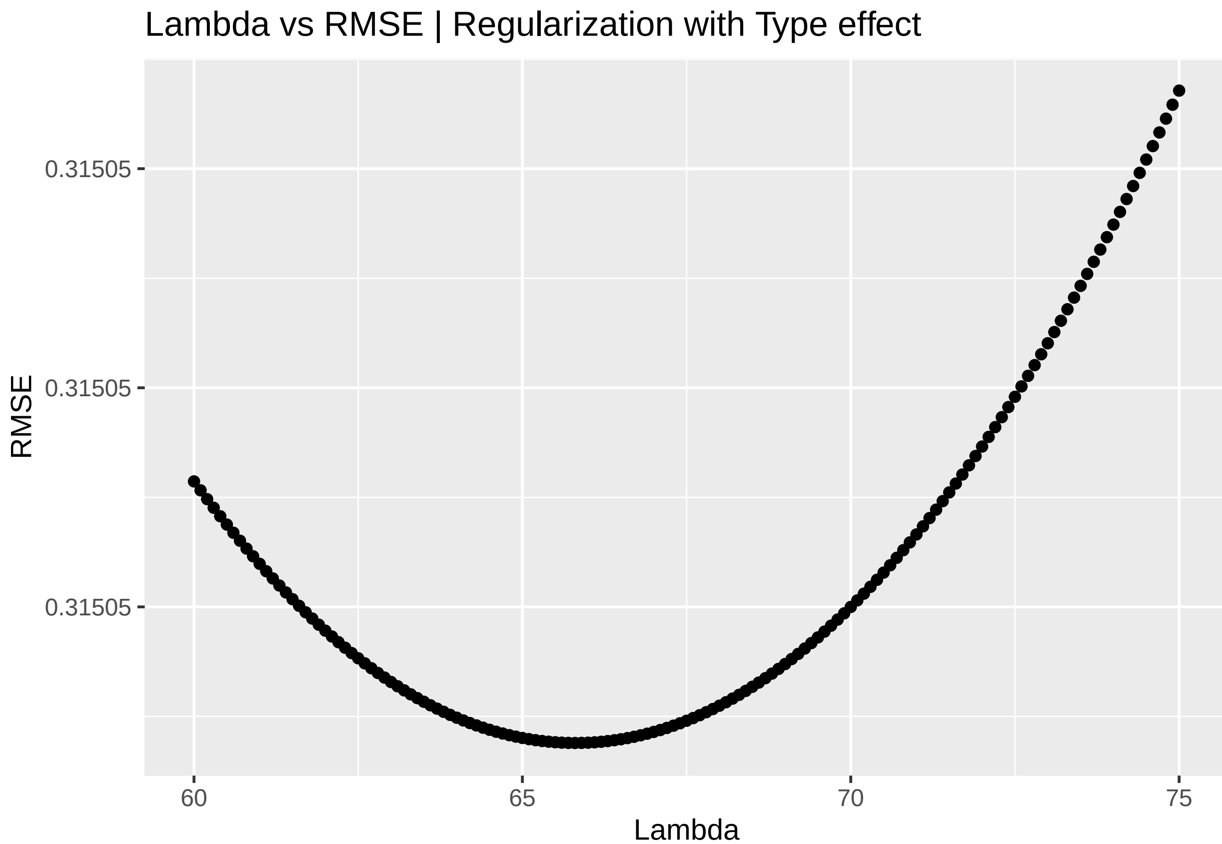



```
## # A tibble: 7 x 2
##   Model                      RMSE
##   <chr>                     <dbl>
## 1 Just the average        0.39861
## 2 Date Effect             0.37223
## 3 Type Effect             0.31506
## 4 Region Effect           0.36834
## 5 Total Volume Effect     0.42521
## 6 Type + Region Effects   0.27161
## 7 Regularized Date Effect 0.37210
```

3.1.8 Regularization with Type effect

Lambda value:

```
## [1] 65.8
```

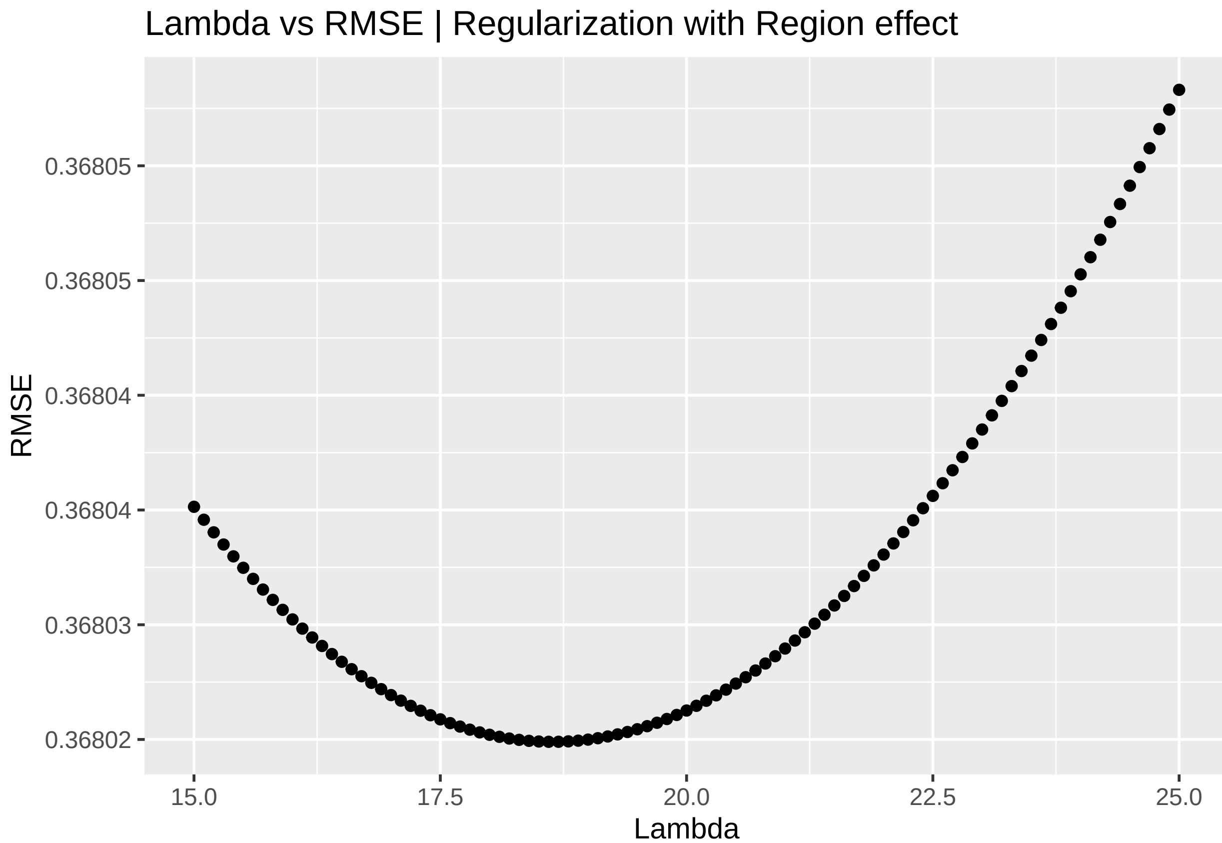



```
## # A tibble: 8 x 2
##   Model                       RMSE
##   <chr>                      <dbl>
## 1 Just the average         0.39861
## 2 Date Effect              0.37223
## 3 Type Effect              0.31506
## 4 Region Effect            0.36834
## 5 Total Volume Effect      0.42521
## 6 Type + Region Effects    0.27161
## 7 Regularized Date Effect 0.37210
## 8 Regularized Type Effect 0.31505
```

3.1.9 Regularization with Region effect

Lambda value:

```
## [1] 18.6
```

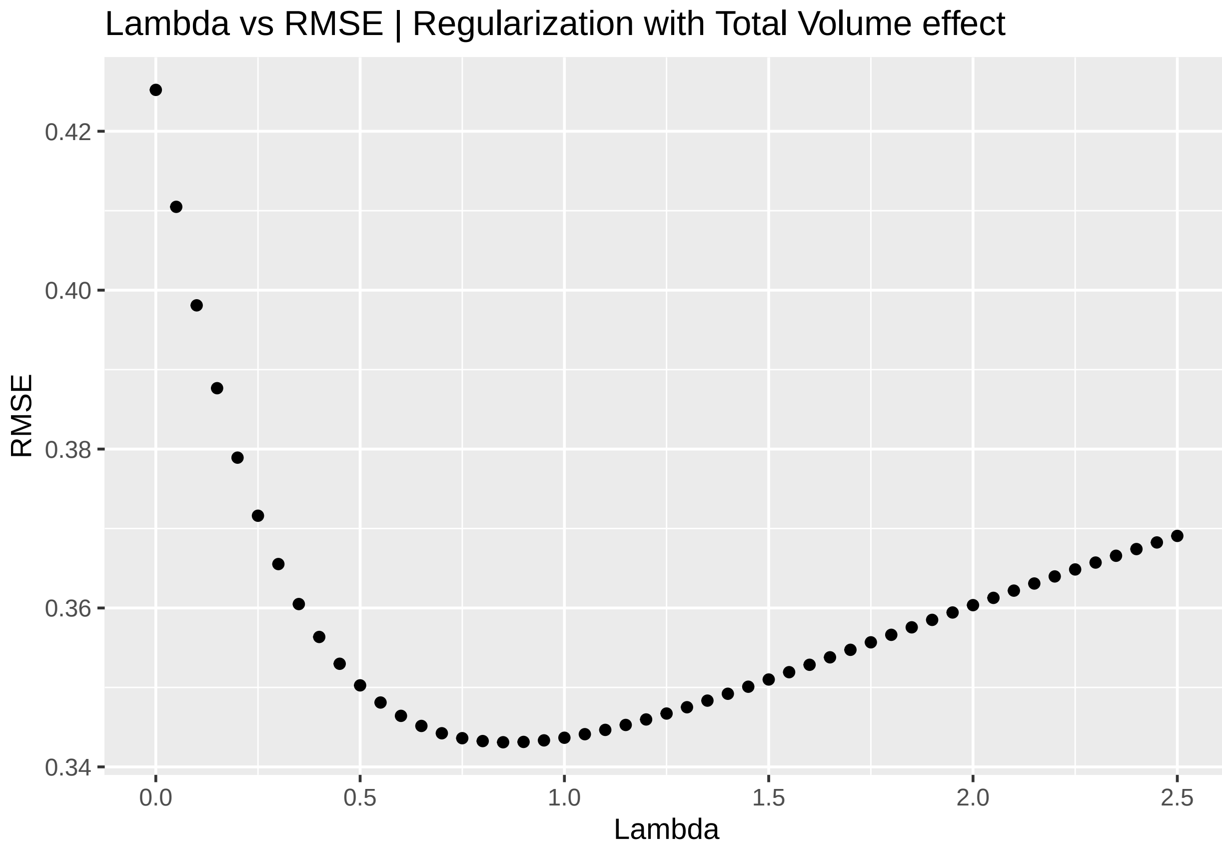



```
## # A tibble: 9 x 2
##   Model                         RMSE
##   <chr>                        <dbl>
## 1 Just the average           0.39861
## 2 Date Effect                0.37223
## 3 Type Effect                0.31506
## 4 Region Effect              0.36834
## 5 Total Volume Effect        0.42521
## 6 Type + Region Effects      0.27161
## 7 Regularized Date Effect    0.37210
## 8 Regularized Type Effect    0.31505
## 9 Regularized Region Effect 0.36802
```

3.1.10 Regularization with Total Volume effect

Lambda value:

```
## [1] 0.85
```



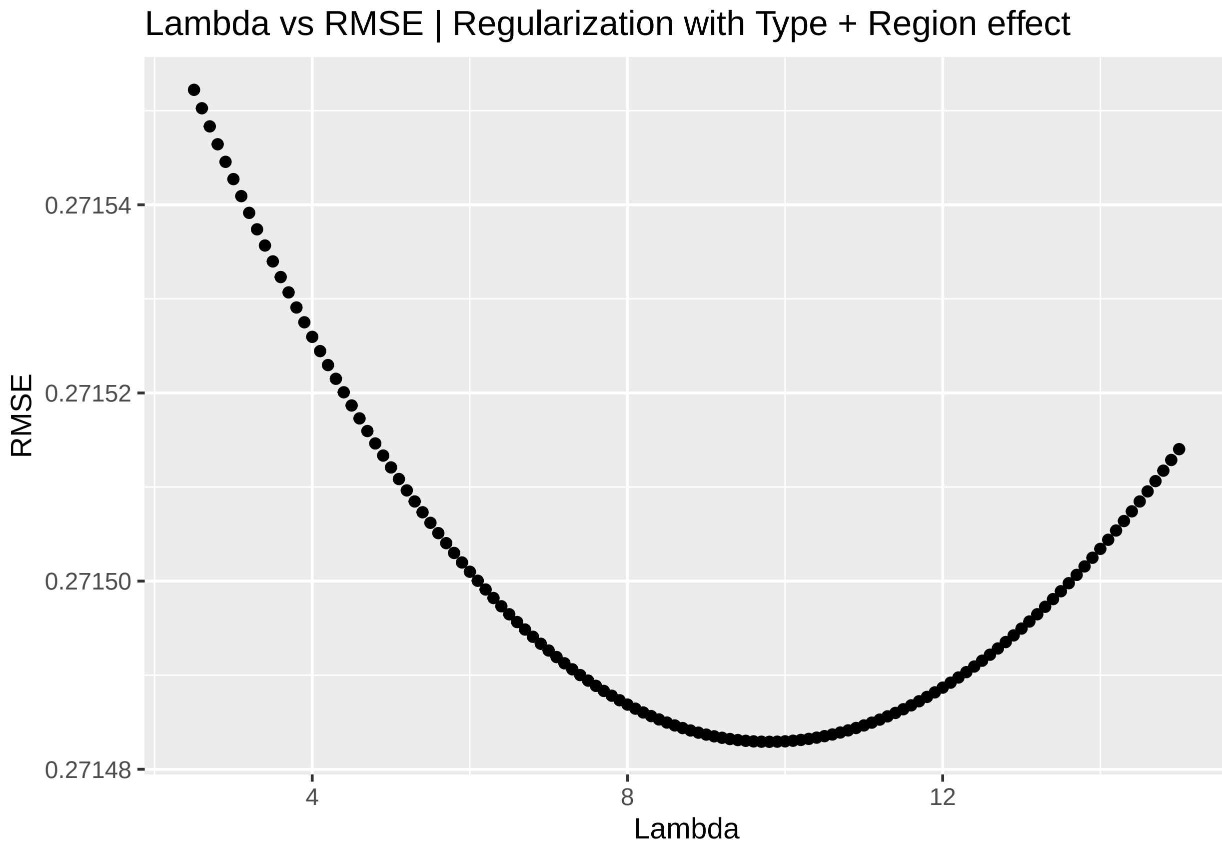

```
## # A tibble: 10 x 2
##    Model                             RMSE
##    <chr>                            <dbl>
##  1 Just the average               0.39861
##  2 Date Effect                    0.37223
##  3 Type Effect                    0.31506
##  4 Region Effect                  0.36834
##  5 Total Volume Effect            0.42521
##  6 Type + Region Effects          0.27161
##  7 Regularized Date Effect        0.37210
##  8 Regularized Type Effect        0.31505
##  9 Regularized Region Effect      0.36802
## 10 Regularized Total Volume Effect 0.34310
```

Due to Type and Region variables got the smallest RMSE values, we will combine them in order to check if we can reduce the Root Mean Squared Error with regularization technique.

3.1.11 Regularization with Type + Region effect

Lambda value:

```
## [1] 9.8
```




```
## # A tibble: 11 x 2
##    Model                               RMSE
##    <chr>                              <dbl>
##  1 Just the average                 0.39861
##  2 Date Effect                      0.37223
##  3 Type Effect                      0.31506
##  4 Region Effect                    0.36834
##  5 Total Volume Effect              0.42521
##  6 Type + Region Effects            0.27161
##  7 Regularized Date Effect          0.37210
##  8 Regularized Type Effect          0.31505
##  9 Regularized Region Effect        0.36802
## 10 Regularized Total Volume Effect   0.34310
## 11 Regularized Type + Region Effects 0.27148
```

For this project, we have to apply machine learning techniques that go beyond standard linear regression so glm, RandomForest and knn techniques are also tested to try to reduce RMSE value. Other techniques such as lda, qda or Naive Bayes have not been finally used because they have generated errors whose solution has not been found.

3.1.12 Generalized Linear Models (Glm)

```
## # A tibble: 12 x 2
##    Model                               RMSE
##    <chr>                              <dbl>
##  1 Just the average                 0.39861
##  2 Date Effect                      0.37223
```

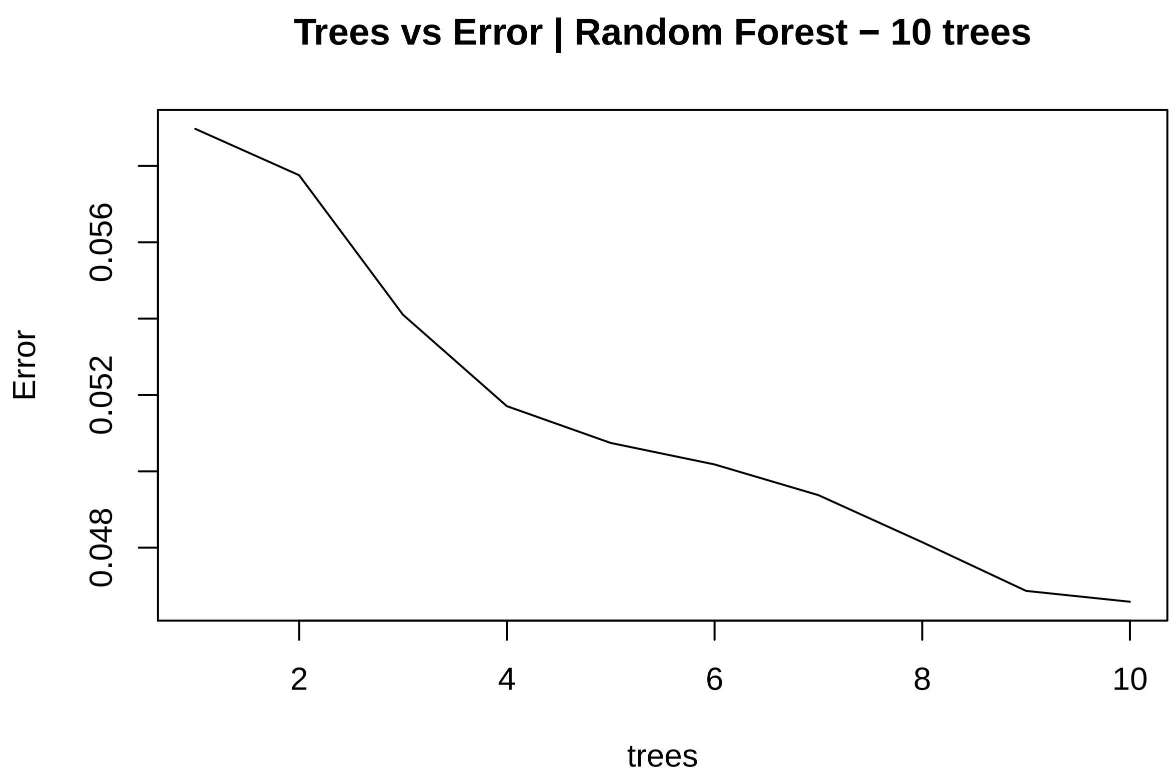
```
##   3 Type Effect                        0.31506
##   4 Region Effect                      0.36834
##   5 Total Volume Effect                0.42521
##   6 Type + Region Effects              0.27161
##   7 Regularized Date Effect            0.37210
##   8 Regularized Type Effect            0.31505
##   9 Regularized Region Effect          0.36802
##  10 Regularized Total Volume Effect   0.34310
##  11 Regularized Type + Region Effects 0.27148
##  12 Glm                                0.22719
```

3.1.13 Random Forest

Because with random forest the fitting is the slowest part of the procedure rather than the predicting (as with kNN), we will use only three ntrees values: 10, 30 and 50. It is recommend it to use more than 100 trees but the time of computation is too hight.

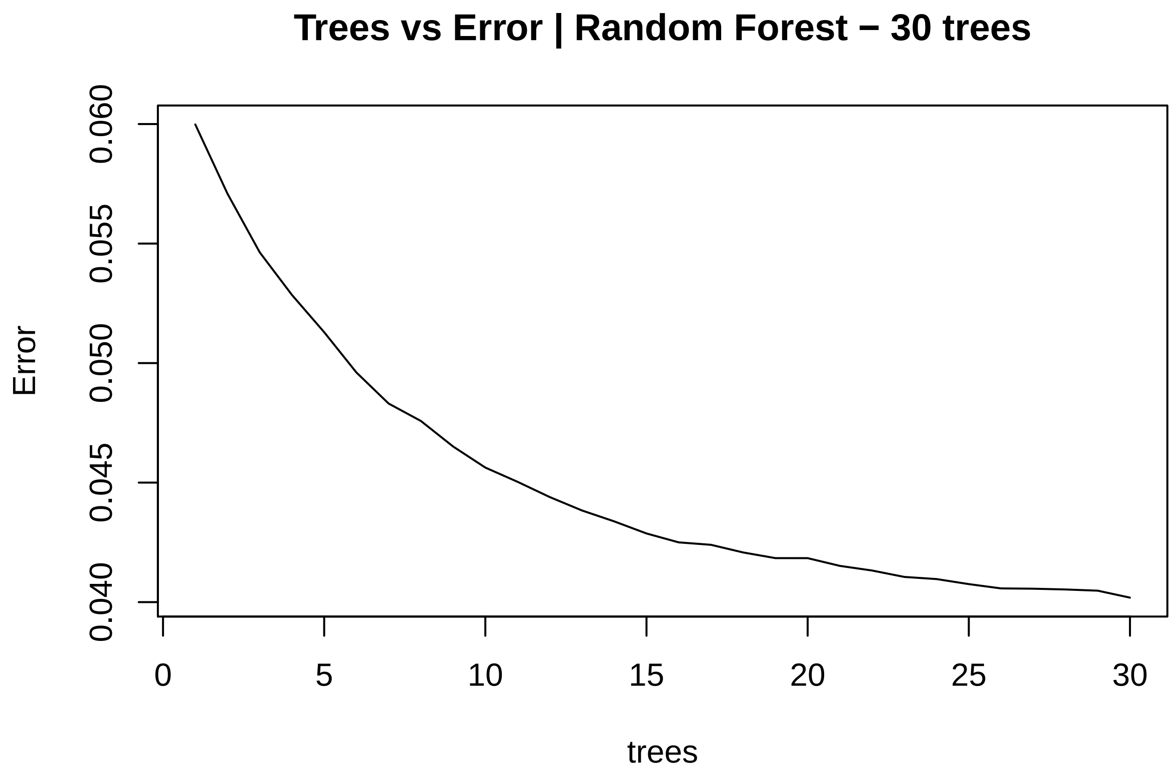
3.1.13.1 10 trees

```
## # A tibble: 13 x 2
##    Model                               RMSE
##    <chr>                              <dbl>
##   1 Just the average                  0.39861
##   2 Date Effect                       0.37223
##   3 Type Effect                       0.31506
##   4 Region Effect                     0.36834
##   5 Total Volume Effect               0.42521
##   6 Type + Region Effects             0.27161
##   7 Regularized Date Effect           0.37210
##   8 Regularized Type Effect           0.31505
##   9 Regularized Region Effect         0.36802
##  10 Regularized Total Volume Effect   0.34310
##  11 Regularized Type + Region Effects 0.27148
##  12 Glm                               0.22719
##  13 Random Forest - 10 trees          0.19826
```



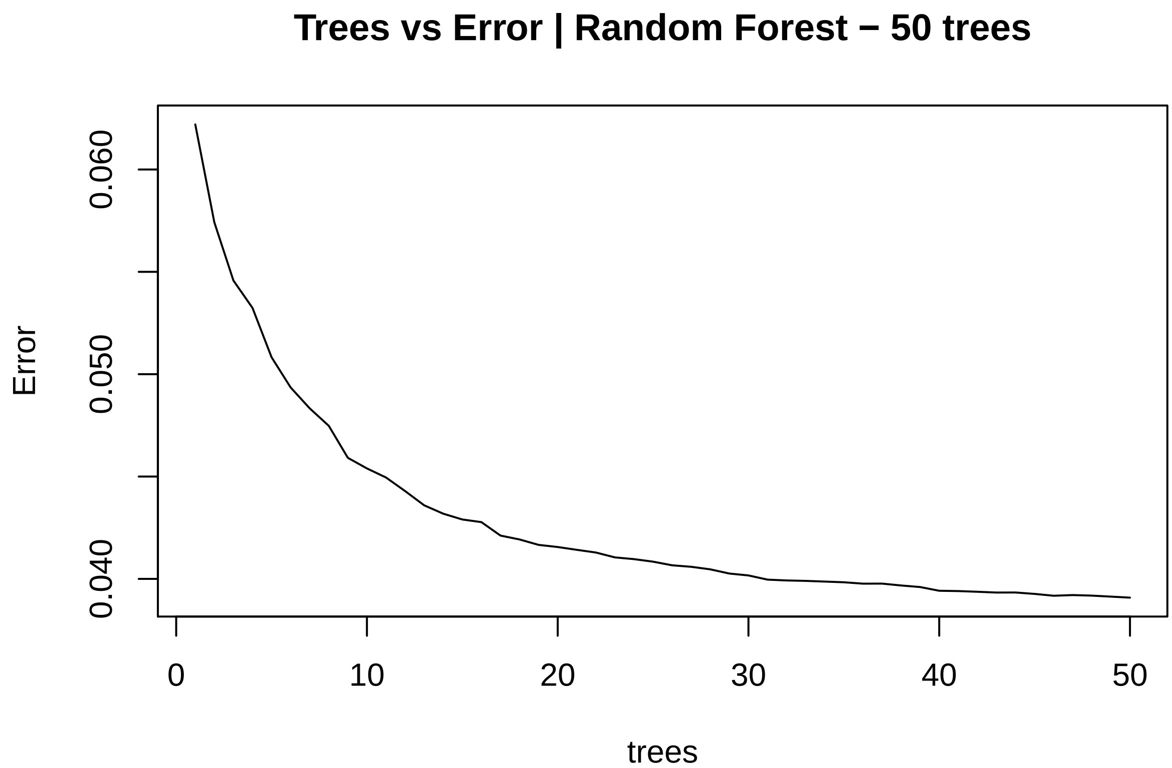
3.1.13.2 30 trees

```
## # A tibble: 14 x 2
##    Model                              RMSE
##    <chr>                             <dbl>
##  1 Just the average                0.39861
##  2 Date Effect                     0.37223
##  3 Type Effect                     0.31506
##  4 Region Effect                   0.36834
##  5 Total Volume Effect             0.42521
##  6 Type + Region Effects           0.27161
##  7 Regularized Date Effect         0.37210
##  8 Regularized Type Effect         0.31505
##  9 Regularized Region Effect       0.36802
## 10 Regularized Total Volume Effect   0.34310
## 11 Regularized Type + Region Effects 0.27148
## 12 Glm                             0.22719
## 13 Random Forest - 10 trees        0.19826
## 14 Random Forest - 30 trees        0.19390
```



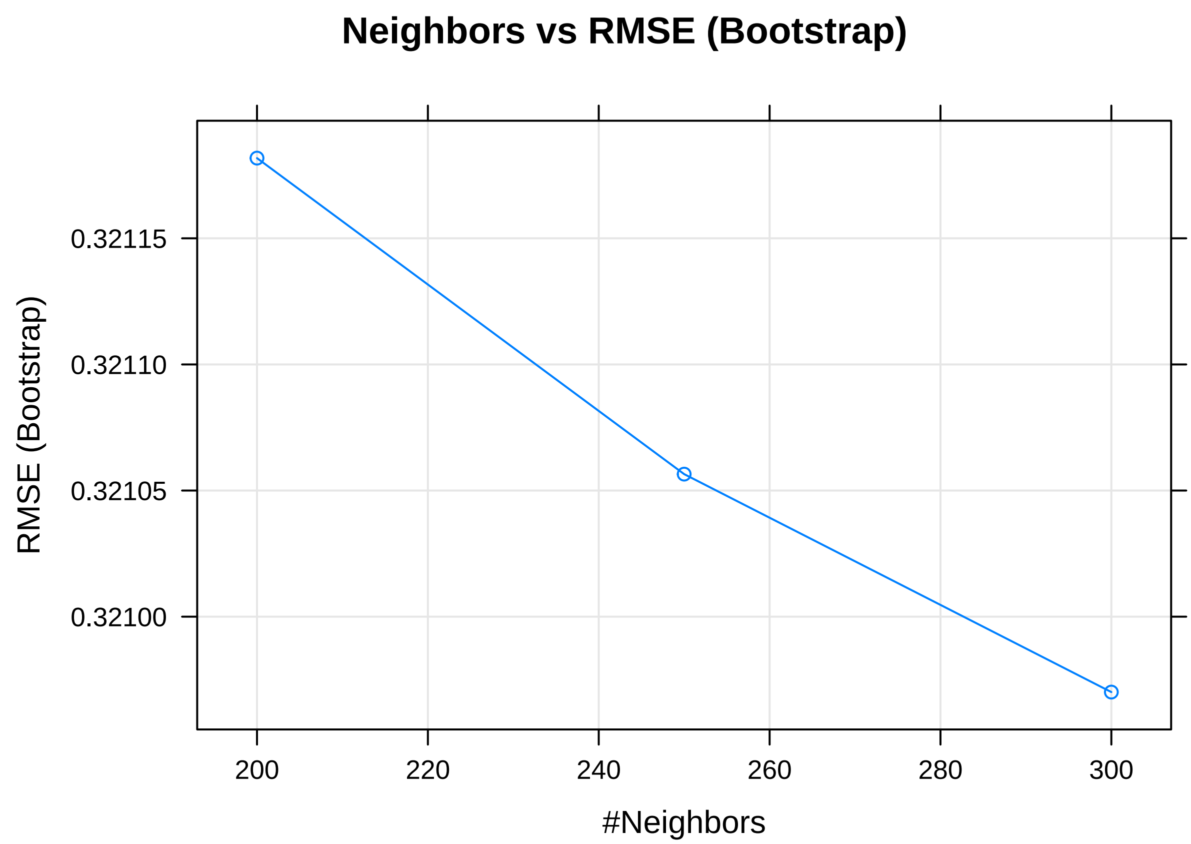
3.1.13.3 50 trees

```
## # A tibble: 15 x 2
##    Model                              RMSE
##    <chr>                             <dbl>
##  1 Just the average                0.39861
##  2 Date Effect                     0.37223
##  3 Type Effect                     0.31506
##  4 Region Effect                   0.36834
##  5 Total Volume Effect             0.42521
##  6 Type + Region Effects           0.27161
##  7 Regularized Date Effect         0.37210
##  8 Regularized Type Effect         0.31505
##  9 Regularized Region Effect       0.36802
## 10 Regularized Total Volume Effect 0.34310
## 11 Regularized Type + Region Effects 0.27148
## 12 Glm                             0.22719
## 13 Random Forest - 10 trees        0.19826
## 14 Random Forest - 30 trees        0.19390
## 15 Random Forest - 50 trees        0.19239
```



3.1.14 Knn

As Random Forest model, in Knn the fitting is the slowest part of the procedure rather than the predicting. We will use only three-fold cross validation: 200, 250 and 300. Other values have been tested (1,7,50,100, etc) but the trend of the error curve was decreasing for higher values of k.

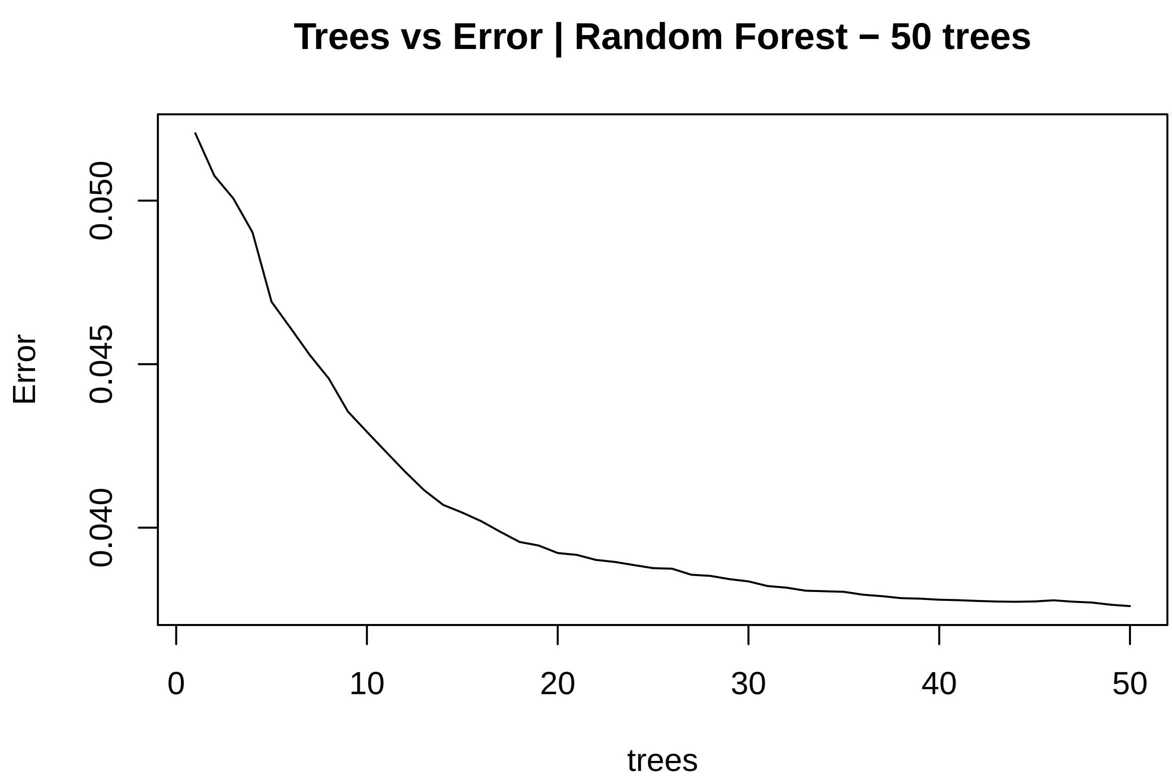



```
## # A tibble: 16 x 2
##    Model                              RMSE
##    <chr>                             <dbl>
##  1 Just the average                0.39861
##  2 Date Effect                     0.37223
##  3 Type Effect                     0.31506
##  4 Region Effect                   0.36834
##  5 Total Volume Effect             0.42521
##  6 Type + Region Effects           0.27161
##  7 Regularized Date Effect         0.37210
##  8 Regularized Type Effect         0.31505
##  9 Regularized Region Effect       0.36802
## 10 Regularized Total Volume Effect 0.34310
## 11 Regularized Type + Region Effects 0.27148
## 12 Glm                             0.22719
## 13 Random Forest - 10 trees        0.19826
## 14 Random Forest - 30 trees        0.19390
## 15 Random Forest - 50 trees        0.19239
## 16 Knn                             0.31769
```

Analyzing the results, we notice that Random Forest with 50 trees model give us the smallest RMSE.

```
##                                Model    RMSE
## 1           Random Forest - 50 trees 0.19239
## 2           Random Forest - 30 trees 0.19390
## 3           Random Forest - 10 trees 0.19826
## 4                                Glm 0.22719
## 5  Regularized Type + Region Effects 0.27148
## 6              Type + Region Effects 0.27161
## 7            Regularized Type Effect 0.31505
## 8                        Type Effect 0.31506
## 9                                Knn 0.31769
## 10   Regularized Total Volume Effect 0.34310
## 11         Regularized Region Effect 0.36802
## 12                     Region Effect 0.36834
## 13           Regularized Date Effect 0.37210
## 14                       Date Effect 0.37223
## 15                  Just the average 0.39861
## 16               Total Volume Effect 0.42521
```

3.2 Validations process



Validation Root Mean Squared Error

```
## [1] 0.19218
```

This RMSE value seems reasonable to achieve our objetive: to avoid an inflation in a certain region and to help to find a city with cheap avocados.

4 Conclusion

The Methods/Analysis section has been necessary to know the type of data we were going to work with.

With the analysis of the influence of variables on average price, we have seen that Type and Region were the most important variables. However, other variables such as the date of observation or year were also important.

In the beginning, RMSEs with basic models, like Just the Average, have been obtained. Then, regularization techniques have been used in order to reduce de Root Mean Squared Error but the results were not very good:

- Type + Region Effects, RMSE = 0.27161
- Regularized Type + Region Effects, RMSE = 0.27148

For this project, we have to apply machine learning techniques that go beyond standard linear regression so glm, RandomForest and knn techniques are also tested to try to reduce RMSE value. Other techniques such as lda, qda or Naive Bayes have not been finally used because they have generated errors whose solution has not been found.

It has been concluded that Random Forest with 50 trees has been optimal for the lower RSME value. However, because with this technique the fitting is the slowest part of the procedure rather than the predicting, only three ntrees values have been tested. Other biggers ntrees values could have been used to get lower RMSE but the time of computation would be too hight.

5 Appendix - Enviroment

```
## [1] "Operating System:"
```

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          3
## minor          6.3
## year           2020
## month          02
## day            29
## svn rev        77875
## language       R
## version.string R version 3.6.3 (2020-02-29)
## nickname       Holding the Windsock
```