**ENVS615 Computational Essay on Avocado Prices**

by Amy Jungmin Seo

16th March 2021(Extended Deadline)

*Title: Analysis of Avocado Prices in the US*

**Introduction**

I came across an interesting dataset about avocado prices while exploring Kaggle.com. Specifically, it is historical data on avocado prices and sales volume in multiple US markets. This dataset caught my eyes instantly because I love avocados, and avocados are things that I don’t forget to buy when I go grocery shopping. Having lived in different cities, I have been able to observe different avocado prices in different seasons, different cities and regions. For example, when I was living in Blacksburg, Virginia, avocados were 2 dollars each during winter while avocados were 99 cents in Atlanta, Georgia year around. I wanted to see if the avocado data can reveal any trend about avocado pricing in the US.

This avocado dataset was shared by Justin Kiggins on Kaggle.com. He mentioned that the data was originally downloaded from the Hass Avocado Board website in May of 2018. Using this dataset, I want to see if I can group some regions that are similar in avocado consumption and sales behaviors using different clustering techniques. K-means with raw data, K-means with standardized data, and Principal Component Analysis (PCA) techniques were performed. Next, I want to test different regression methods to find the best performing model to predict avocado prices. Linear regression and Random Forest method were performed.

The dataset and the .ipynd file can be located in this GitHub location: **https://github.com/amyseoj1/ENVS615-Data-Science-Studio-**

**Description of Data**

The original creator of this dataset, Hass Avocado Board, describes that the dataset represents weekly 2018 retail scan data for US national retail volume (units) and price. The retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. The Average Price of avocados in the dataset reflects a per avocado cost, even when multiple avocados are sold in bags. PLU stands for the product lookup code.

The dataset contains following data fields:

* Date - The date of the observation
* AveragePrice - the average price of a single avocado
* type - conventional or organic
* year - the year
* region - the city or region of the observation
* Total Volume - Total number of avocados sold
* 4046 - Total number of avocados with PLU 4046 sold (Small/Medium Hass Avocado)
* 4225 - Total number of avocados with PLU 4225 sold (Large Hass Avocado)
* 4770 - Total number of avocados with PLU 4770 sold (Extra Large Hass Avocado)

**Data Wrangling**

In this analysis, the first step I took was data wrangling. The avocado dataset was a relatively clean dataset without any missing or null values. Therefore, there were only a few minor steps to prepare the data for the analysis.

First of all, I renamed field names to remove empty spaces in the names. I also renamed the field names that do not have intuitive meanings. Second of all, during the preliminary inspection of the data, I noticed that the field “region” contained Total US aggregated data. This piece of data was significantly larger than other observations, and thus I excluded the rows that had Total US as the “region”. Thirdly, I had to create dummies for the categorical fields, “type” and “region”. The field “type” contains either “conventional” or “organic” as the value. The field “region” contains names of different regions of the observation. The problem with these categorical values is that many machine learning algorithms cannot operate on categorical data directly. One resolution to this is converting the categorical data to a numerical format. Where there is no ordinal relationship exists in the categories, one-hot coding can be applied to the integer representation. One-hot encoding creates one binary variable for each category (Brownlee 2020). This means one category is defined by a set of binary variables, 0 and 1. For example, the conventional type can be expressed as [1,0] and the organic type can be expressed as [0,1].

**Data Visualization**

As a preliminary visualization of the data, I plotted a box plot to compare the avocado average prices of conventional and organic. According to the plot, organic avocados are more expensive than conventional avocados.

Another plot I created was a strip plot. This was done to see the volume of total bags sold in different regions. In terms of a single city, Los Angeles has the highest consumption of avocados. In terms of the consumption of organic avocados, West region recorded highest number of organic bags sold.

Chart, box and whisker chart

Description automatically generated

Figure 1. A box plot of average price of conventional and organic avocados (a) and a strip plot of total bags sold in different US regions.

**Methodology: Unsupervised Learning**

For the unsupervised learning, several clustering techniques were operated on the avocado prices data. K means raw, K means standardized, and PCA methods were performed. I assumed five clusters to be sufficient for this analysis. Chart, scatter chart

Description automatically generated

Figure 2. Results of different clustering techniques: PCA, K-means Raw, and K-means Standardized.

The plots of the clustering results shows that PCA provides the most clear-cut division of clusters. Other two methods seem like the boundaries of clusters are not clear. In order to measure the quality of clustering, Calinski and Harabasz Score (CHS) and Silhouette Score were computed.

Clustering validation can be categorized into two classes, external and internal clustering validation (Wei, 2020). Both CHS and Silhouette score are internal clustering measures. Other internal clustering methods are Root-mean-square std dev, R-squared, Modified Hubert Γ Statistics, *I* index, Dunn’s Indices, Davies-Bouldin index, Xie-Beni index, SD validity index, S\_Dbw validity index (Liu et al. cited in Wei 2020).

In terms of the CHS score, the higher, the better. A higher CHS score means that the clusters are dense and well separated (Wei 2020).

The Silhouette score ranges from -1 for incorrect clustering and +1 for highly dense clustering. Scores near zero indicate overlapping clusters (Wei 2020). This means a higher Silhouette score (close to +1) relates to a model with better defined clusters.

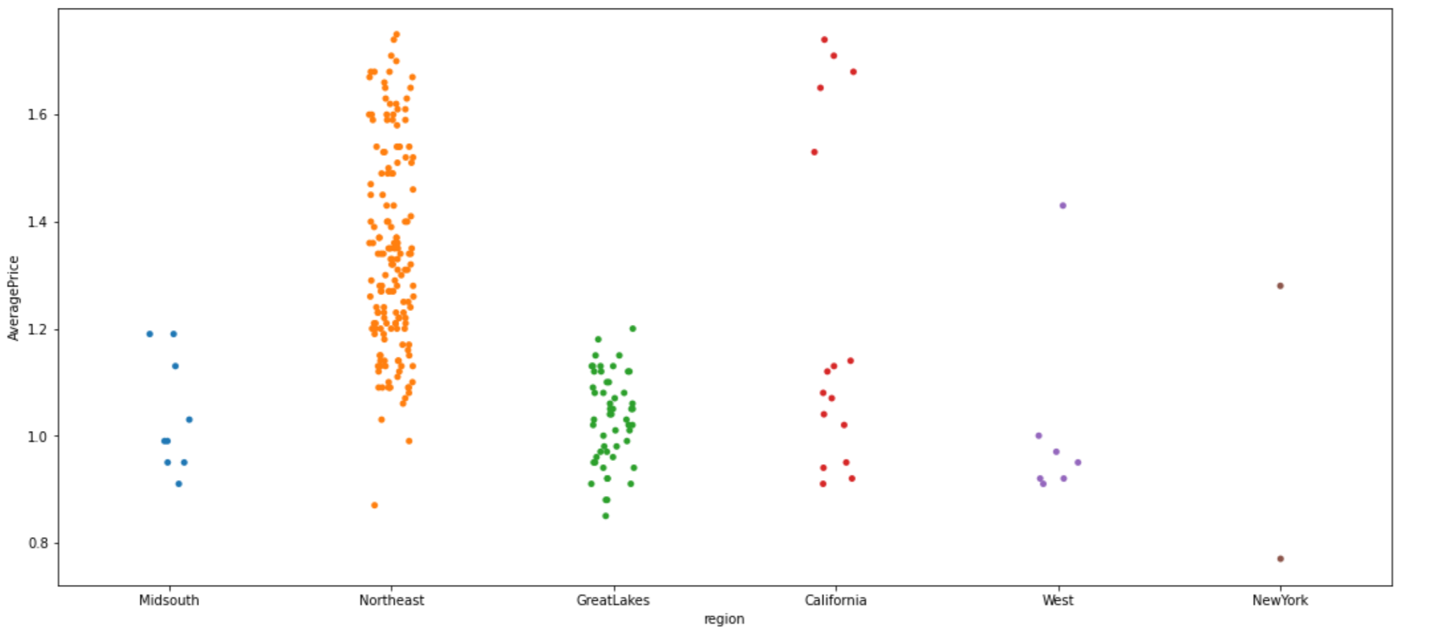
Davies-Bouldin Index (DBI) signifies the average ‘similarity’ between clusters, where the similarity is the ratio of within-cluster distances to between-cluster distances (Davies & Bouldin 1979). A lower DBI relates to a model with better separation between clusters.

Table 1. Summary of metrics on clustering quality.

|  |  |  |  |
| --- | --- | --- | --- |
|  | K5\_raw | K5\_std | PCA |
| Calinski and Harabasz Score | **44469.95** | 23172.54 | 24218.10 |
| Silhouette Score | **0.700041** | 0.683492 | 0.675465 |
| Davies-Bouldin Index | **0.810743** | 2.211983 | 1.753309 |

Based on the CHS scores obtained from three clustering methods, we can conclude that K5\_raw works the best. Similarly, the Silhouette score shows that K5\_raw method works the best; K5\_raw has the highest score, which means that the clusters are dense and well separated. DBI results are in alignment with the results from other two methods; K5\_raw scores the lowest which means the clusters are well separated.

After assessing the quality of clustering, one out of five clusters (n = 2) was plotted. By doing so, it was easier to observe which regions belong to the same cluster. According to the plot, we can conclude that Great Lakes, California, and New York are similar in terms of avocado sales and consumption behavior.

****Figure 3. One of the five K-means raw clusters (n=2).

**Methodology: Supervised Learning**

Linear regression and the Random Forest method were used as two selected supervised learning method. I split the data into 75% train and 25% test sets. The train data set will be used to train the model, and the remaining 25% will be used to compare the actual results to the predicted values using the model. Using the model trained with 75% of data (X\_train), the remaining 25% of data (X\_test) was used predict the avocado price (y\_pred). y\_pred is compared to y\_test actual values.

The results from Linear Regression – Raw, Linear Regression­­ ­– Low, Random Forest – Raw, and Random Forest – Log were summarized in a single table comparing against the actual value.

Table 2. Summary of predicted avocado prices using different models and the actual values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LR-Raw | LR-Log | RF-Raw | RF-Log | Truth |
| 0 | 1.639168 | 0.954593 | 1.9016 | 1.055229 | 2.07 |
| 1 | 1.675619 | 0.973647 | 1.3258 | 0.826529 | 1.26 |
| 2 | 0.928097 | 0.641551 | 0.6051 | 0.476180 | 0.58 |
| 3 | 1.559651 | 0.928342 | 1.4909 | 0.892424 | 1.43 |
| 4 | 1.166316 | 0.760254 | 1.4834 | 0.902393 | 1.54 |

Chart, histogram

Description automatically generated

Table 3. Sumary of inference metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LR-Raw | LR-Log | RF-Raw | RF-Log |
| R-Squared | 0.576 | -1.471 | 0.895 | -1.281 |
| MAE | 0.197 | 0.548 | 0.090 | 0.544 |
| MSE | 0.068 | 0.397 | 0.017 | 0.367 |
| RMSE | 0.261 | 0.630 | 0.130 | 0.605 |

The R-squared scores were 0.576 for the linear regression raw model and 0.895 for the random forest-raw model respectively. R-squared score closer to 1 represents higher accuracy of the model, which means that random forest model yields more accurate result. Both log models had negative R-squared scores, meaning that the chosen model does not follow the trend of the data. Therefore, taking the log of AveragePrice may not be ideal. Mean Absolute Error (MAE) measures the absolute average distance between the actual and the predicted data, but it fails to penalize large prediction errors. Mean Squared Error (MSE) measures the squared average distance between the actual and the predicted data. Root Mean Squared Error (RMSE) is the square root of Mean Squared error. It measures the standard deviation of residuals. RMSE is more widely used than MSE to evaluate the performance of the regression model. The lower value of MAE, MSE, and RMSE implies higher accuracy of the model (Chugh 2020). Similar to the R-squared results, random forest-raw had the lowest MAE, MSE, and RMSE values. In conclusion, random forest – raw model seems to be the best model.

**References**

Kiggins, J 2018, *Avocado Prices: Historical data on avocado prices and sales volume in multiple US markets*, Kaggle, viewed 1 March 2021, <https://www.kaggle.com/neuromusic/avocado-prices>

Brownlee, J 2017, *Why One-Hot Encode Data in Machine Learning,* Machine Learning Mastery, viewed 3 March 2021, <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>

Brownlee, J 2020, *Ordinal and One-Hot Encodings for Categorical Data,* Machine Learning Mastery, viewed 3 March 2021, < https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/ >

Wei, H 2020, *How to measure clustering performances when there are no ground truth,* Medium, viewed 5 March 2021, < https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c>

Davies, David L.; Bouldin, Donald W. (1979). [“A Cluster Separation Measure”](https://ieeexplore.ieee.org/document/4766909). IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227

Chugh, A 2020, *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R squared — Which Metric is better?,* Medium, viewed 8 March 2021, < https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>

**Appendix I. Jupyter Notebook**

https://github.com/amyseoj1/ENVS615-Data-Science-Studio-

Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

A picture containing chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated