Avocado Trends – A Visual Report

Coursework Report for Module INM433 "Visual Analytics"

Karin Falconer-Bailey

Abstract — This paper focuses on data originally sourced from the Hass Avocado Board (HAB). We combine two sets of data: one file with dates from 2015 to 2018 and the second source of the year 2021. This approach has been implemented to further knowledge of prior trends to generate more reliable insight into current and future trends of the avocado industry. Upon merging the datasets in Jupyter Notebook and undergoing data preparation, we first analyse the data using the filter method to identify correlations and factors that have influenced avocado prices. We then build multiple linear regression model(s) to develop visualisations and identify how significant each feature is concerning the model's predictor. Next, using Jupyter Notebook's plotly extension, we further visualise geographical and time series aspects relating to the data. The final segment exhibits an autoregressive integrated moving average model (ARIMA) to conduct a time series analysis and to predict future prices of the Hass avocado.

1 Problem Statement

Centred with a seed of "usually more than five centimetres long", and subject to variation in its hue due to ripeness, the avocado has become one of the most popular consumption trends of the decade [1].

Since 2004, as a search term, the avocado has grown tremendously worldwide—with exponential growth in numerous cities of the US [2]. Over the current year, the United States has increased avocado importations by \$0.36 billion, from its former expenditure of approximately \$0.90 billion [3]. Avocado imports accounted for 15.06% of the country's \$8.03 billion fruit imports in the first quarter of 2021, and despite its high demand, the United States Food and Drug Administration (FDA) reported the nation's food waste to be "between 30–40 per cent of its food supply" [4].

This report explores and analyses the time-series trends and geospatial components of avocado prices and consumption throughout the US. Under this analysis, the paper aims to address the following research questions:

- Is there a pattern of avocado consumption?
- How significant is the change in avocado consumption?
- How does avocado consumption alter between states?

This report would prove beneficial to the United States Food and Drug Administration and the International Trade Administration (ITA), with the intent of providing insight into avocado preferences, reducing the country's overall output of food waste, and creating a better-placed distribution of the country's wealth.

2 State of the Art

A study conducted by Jones et al. (2021), utilizes an almost identical dataset for analysis. Although the paper's data lacks the results of the year 2021, they aim to provide insight into consumers' buying habits in the United States and recent health trends from 2015-2020. Jones et al.'s (2021) study has derived data from Kaggle, a domain that promotes diverse studies within the field of data, originally sourced via the HAB website. The data is comprised of a total of 3.37 megabytes across its 3 main features: 4046 – Small/Medium Hass Avocado, 4225 – Large Hass Avocado, and 4770 – Extra Large Hass Avocado, which is slightly larger than the 2.54 megabytes used for this paper, despite its numerous additional features [5]. The paper uses SAP Analytics Cloud to undergo data preparation during which the feature '*total U.S.*', comprising of the total United States' demand for each Hass avocado, is removed from the dataset due to "skewing the results when used in certain models" [5]. Time-series analysis and regression analysis are used to display the temporal components of the data, and to predict and forecast future values.

Ricon-Patino, Lasso, and Corrales's (2018) paper '*Estimating Avocado Sales Using Machine Learning*

Algorithms and Weather Data' aims to compose an accurate machine learning model which can be applied to allow "avocado producers and sellers to plan sales through the estimation of the profits" [6]. The paper's data had likewise been derived from the HAB website, consisting of monthly statistics from January 2013 to June 2017. The paper uses Weather Underground to attain time-series weather information on cities in the United States to increase the study's complexity. The final model is comprised of the following attributes: weather, units-py, units-cy, weather and sales-py, weather and ln(units-py), weather and ln(sales-py), weather and units-py/population, and weather and sales py/population—with ln(units-py) representing the "natural logarithm of the number of avocados sold" and ln(sales-py) signifying the "natural logarithm of the total sales" [6]. Ricon-Patino, Lasso and Corrales's study utilizes four alternative machine learning algorithms for predictive modelling: linear regression, multilayer perceptron (MLP), support vector machine for regression (SVM) and multivariate regression. To evaluate each model a 10 k-fold cross-validation method was implemented, alongside assessing each model's accuracy via the model's summary (MAE, RMSE and RAE).

Jones et al.'s (2021) paper presents promise, however, during the process of visualising the data, this could have been further explored. Limited visualisations are provided regarding spatial information, nevertheless, the visualisations which are presented for exploratory analysis and time-series analysis will too be included in this paper's illustrations. Overall Jones's study does not provide considerable insight into its applied methods; regression models have not been stated, nor have its results or its validation techniques.

Conversely, Ricon-Patino, Lasso, and Corrales's (2018) study allow for much insight into the paper's applied techniques. Despite the lack of varying visualisations, this paper will similarly apply multivariate regression to analyse and visualise the significance of the model's features.

3 Properties of the Data

Source and structure

The main dataset (1894 KB) which this paper relies upon has been collated from Kaggle, however, its exact source differs from Jones et al.'s (2021) aforementioned study [7]. A total of 18,249 observations over 13 attributes are present within the data which details the "historical data on avocado prices and sales volume in multiple US markets" from 2015-2020 [7]. The 13 features include the date, average avocado prices, region, year, type of avocado, 4046 units, 4225 units, 4770 units, small bags, large bags and extra-large bags, the total volume of avocados sold, and the total number of bags sold.

The second source of data (648 KB) was extracted directly from the Hass Avocado Board's website. The data, which also details avocado prices and volume of avocado sales from 2021, contains 5,192 observations across 14 features. Like the larger dataset, it features the following variable names: geography, timeframe, current year week ending, type, ASP current year, total bulk and bag units, 4046 units, 4225 units, 4770 units, total bagged units, small bagged units, large bagged units, x-large bagged units and bulk GTIN.

Units

Type — Conventional or organic avocados.

Region / Geography – 53 cities in the United States

AveragePrice / ASP current year — Measured in U.S. dollars.

Date / Current Year Week Ending — Weekly DD/MM/YYYY formatted data.

4046 units, 4225 units, 4770 units, total bulk and bag units, total bagged units / total bags — Measured in millions.

4046 units — non-organic small/medium Hass Avocado (~3-5 oz)

4225 units — non-organic large Hass Avocados (~8-10 oz)

4770 units — non-organic extra-large Hass Avocado (~10-15 oz)

Data inconsistencies

Since this paper relies on two sources of data, zero inconsistencies must be displayed throughout the files. Several inconsistencies were detected in the *region* and *type* observations for instance the type of avocados in the 2015-2020 file was written as *conventional* and *organic* whereas, in the 2021 file, it was written as *Conventional* and *Organic.* Inconsistent values across this dataset were identified and replaced through Jupyter Notebook's pandas *dataframe.replace()* function.

Figure 01: Correction of data inconsistencies

	Date	AvgPrice	TotalVol	4046	4225	4770	TotalBags	SBags	LBags	XLBags	Type	Region
0	27/12/2015	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	Albany
1	20/12/2015	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	Albany
2	13/12/2015	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	Albany
3	06/12/2015	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	Albany
4	29/11/2015	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	Albany
...	...	...	...	...	...	...	...	...	...	...	...	...
18244	04/02/2018	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	organic	WestTexNewMexico
18245	28/01/2018	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	organic	WestTexNewMexico
18246	21/01/2018	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	organic	WestTexNewMexico
18247	14/01/2018	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	organic	WestTexNewMexico
18248	07/01/2018	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	organic	WestTexNewMexico

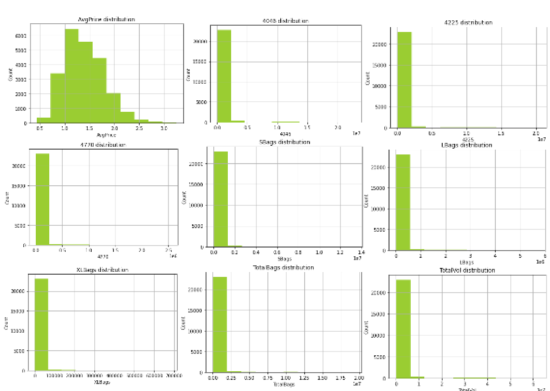
	Date	AvgPrice	TotalVol	4046	4225	4770	TotalBags	SBags	LBags	XLBags	Type	Region
0	27/12/2015	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	Conventional	Albany
1	20/12/2015	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	Conventional	Albany
2	13/12/2015	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	Conventional	Albany
3	06/12/2015	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	Conventional	Albany
4	29/11/2015	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	Conventional	Albany
...	...	...	...	...	...	...	...	...	...	...	...	...
18244	04/02/2018	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	Organic	West Tex/New Mexico
18245	28/01/2018	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	Organic	West Tex/New Mexico
18246	21/01/2018	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	Organic	West Tex/New Mexico
18247	14/01/2018	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	Organic	West Tex/New Mexico
18248	07/01/2018	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	Organic	West Tex/New Mexico

As displayed in figure 01, we can see how the data inconsistencies have been corrected. The *type* features have been capitalised, and we can identify that *WestTexNewMexico* has been altered to *West Tex/New Mexico.* Other features have similar alterations such as *LosAngeles* to *Los Angeles* and so forth.

Data distribution

To evaluate the distribution of the dataset's features, using Jupyter Notebook's dataframe.plot(kind='hist') function, each numeric variable's distribution is exhibited in figure 02.

Figure 02: Distribution of avocado features



From the figure, it can be noticed that only *AvgPrice* displays a clear distribution with an almost normal distribution, however, it appears to be slightly skewed to the right and thus, "the mean value is higher than the median (median closer to the first quartile" [9].

It is clear that zero outliers are present in the dataset, and thus, we do not have to determine whether they should remain or should be dropped from the dataset.

4 Analysis

4.1. Approach

Through each stage, we specify where human reasoning, human judgement or computation methods must be applied.

Data preparation

First, we cleanse the data of any errors, inconsistencies, and missing values. These initial stages of analysis are vital, it ensures that errors are minimised when calculating the accuracy of the model(s). During this, human judgement is the preferred means. This guarantees the observations are handled in a manner deemed fit by the analyst, and per the study's needs.

EDA

We conduct EDA to provide visualisations of the data i.e., central tendencies, variable distributions, and feature exploration. During this stage, human reasoning and judgement will determine which features should be visualised, alongside which methods of visualisation are appropriate. To conduct such visualisations, computational methods are also necessary.

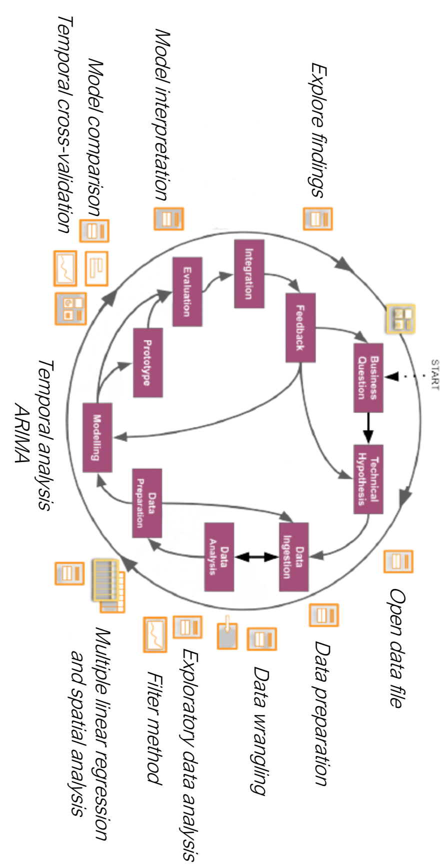


Figure 03: Analysis workflow diagram (Reinterpretation from Serrano's (2019) study)

Multiple linear regression

Building a multiple linear regression model will be used similarly as a correlation matrix would. Through multiple linear regression, we aim to provide further insight into the significance of each feature within the study, and exactly how it responds to changes in the study's predictor. To construct a reliable regression model, we must ensure that the model meets the assumptions of MLR and thus, a correlation matrix is too needed under the filter method. No methods of cross-validation will be implemented. Regression relies upon computational methods to compute the model and human reasoning to ensure the assumptions are adhered to.

Spatial analysis

Through spatial analysis, we aim to explore and visualise geographic trends in the dataset using Jupyter Notebook's *plotly* extension. Here, we will create mapped geographic distributions of all avocado trends across all featured cities of the dataset and total U.S. avocado trends. While computational methods will be implemented to construct the outputs of the plotly visualisations, human reasoning is required to provide context to the visualisations.

Temporal analysis

To conduct time-series analysis, the data will be split into training and testing where an autoregressive integrated moving average model will be implemented via Jupyter Notebook. ARIMA models rely on stationarity and thus the data will be differenced until stationarity is existent. If the data becomes stationarity after its first differencing, we can say the series is integrated of order one I(1) [11].

$$I(1) — Z_t = \Delta Y_t = Y_t - Y_{t-1}$$
$$I(2) — \Delta Z_t = Z_t - Z_{t-1}.$$

Under Box-Jenkins's identification process, we will analyse patterns of autocorrelation functions and partial autocorrelations functions. The use of the two functions will be observed to identify whether an AR(p), MA(q) or ARMA(p,d,q) model should be implemented. Again, computational methods are essential to computing ARIMA models, yet, human reasoning is necessary to interpret results and draw conclusions.

Findings

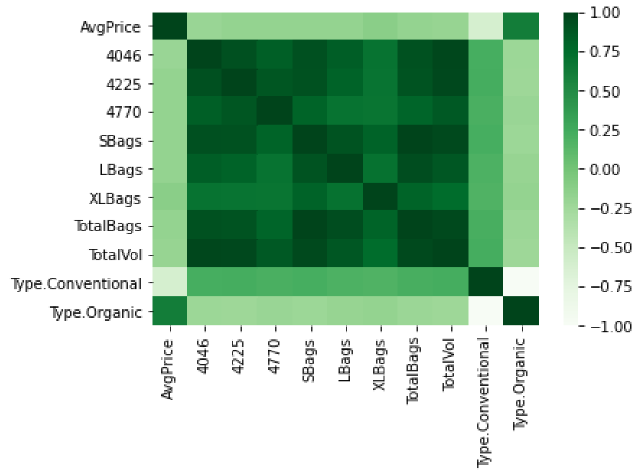
Across all stages of the analysis process, results will be interpreted. The main interpretation for the model relies on time-series analysis where we will not only analyse temporal components but will aim to forecast future avocado trends. Upon collating the model's findings, we will conclude the study and answer the paper's research questions.

4.1. Process

Multiple linear regression (mlr)

From figure 04 we can see that numerous variables hold strong correlations amongst one another. Many of the observations possess a correlation greater than 0.5. While this would be problematic for predictive modelling and would not hold to the assumptions of multiple linear regression, this correlation matrix is for visualisation and interpretation purposes only and has not been relied upon for predictive or forecasting techniques.

Figure 04: Avocado correlation matrix



With the study's predictor being the total volume of avocados consumed, we can see that 4046, 4225 and small bags have the strongest positive correlation—such that we can expect an increase in total volume to see an increase in these particular attributes. To see how significant this change is, we will construct a multiple linear regression model and interpret the model's coefficients.

Interpreting the coefficients

Figure 05 exhibits the coefficients of the study's independent variables using ordinary least squares regression analysis.

Figure 05: OLS regression results

	coef	std err	t	P>|t|	[0.025	0.975]
AvgPrice	-0.7480	1.168	-0.641	0.522	-3.037	1.541
4046	1.0000	9.15e-07	1.09e+06	0.000	1.000	1.000
4225	1.0000	1.11e-06	8.99e+05	0.000	1.000	1.000
4770	1.0000	7.87e-06	1.27e+05	0.000	1.000	1.000
SBags	-0.1413	5.692	-0.025	0.980	-11.298	11.015
LBags	-0.1413	5.692	-0.025	0.980	-11.298	11.015
XLBags	-0.1415	5.692	-0.025	0.980	-11.298	11.015
TotalBags	1.1413	5.692	0.201	0.841	-10.015	12.298
Type.Conventional	0.6246	1.467	0.426	0.670	-2.250	3.500
Type.Organic	5.2297	1.999	2.616	0.009	1.312	9.147

Holding other variables constant…

- *AvgPrice*: Total volume will decrease by 0.7480 units, for every unit ($) increase in the average price.
- *4046*: Total volume will increase by 1.00 units for each unit increase in 4046 plu.
- *4225*: Total volume will increase by 1.00 units for each unit increase in 4225.
- *4770*: Total volume will increase by 1.00unitst for each unit increase in 4770.
- *SBags*: Total volume will decrease by 0.1413 units for each unit increase in small bags.
- *LBags*: Total volume will decrease by 0.1413 units for each unit increase in large bags.
- *XLBags:* Total volume will increase by 0.1415 units for each unit increase in extra-large bags.
- *TotalBags*: Total volume will increase by 1.1413 units for each unit increase total bags.
- *Type.Conventional*: Total volume will increase by 0.6246 units, for every unit ($) increase in the avocado type being conventional.
- *Type.Organic*: Total volume will increase by 5.2297 units, for every unit ($) increase in the avocado type being organic.

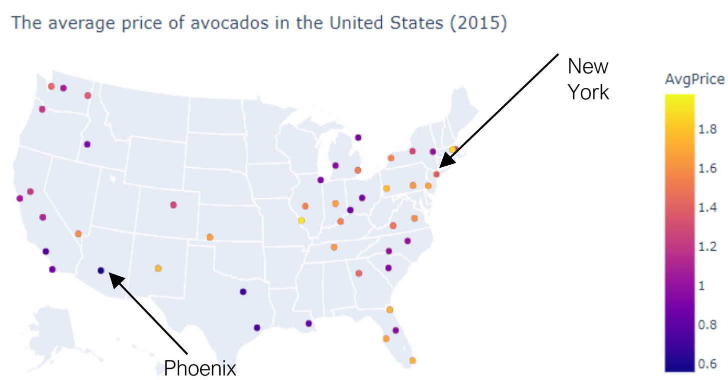
Spatial Analysis

In order to construct a spatial analysis of the dataset, we use Plotly Express to utilize the function *px.scatter_geo()* an extension package from Python, which enables the output of high-level data visualisation plots. We first began composing our spatial graphs by adding two additional columns to the dataset: longitude and latitude. To conduct our spatial analysis, we required coordinates of the cities and thus, the region column was transformed into longitude and latitude coordinates. We then filtered and grouped the dataset by their dates. This way, we could build plots of the dataset's features across alternative years. We begin with a spatial exploration of the average price variable followed by total volume.

AvgPrice

Figure 06 below displays the average price of avocados in the cities of the United States. From the first glances of the graph, we can see that numerous cities hold an average price of less than $1.00, with several other cities avocado prices being greater than $1.00. Take New York, for example, the average avocado price is $1.36 whereas, in Phoenix, the average price of avocados is $0.56.

Figure 06: AvgPrice of avocados in 2015/2017

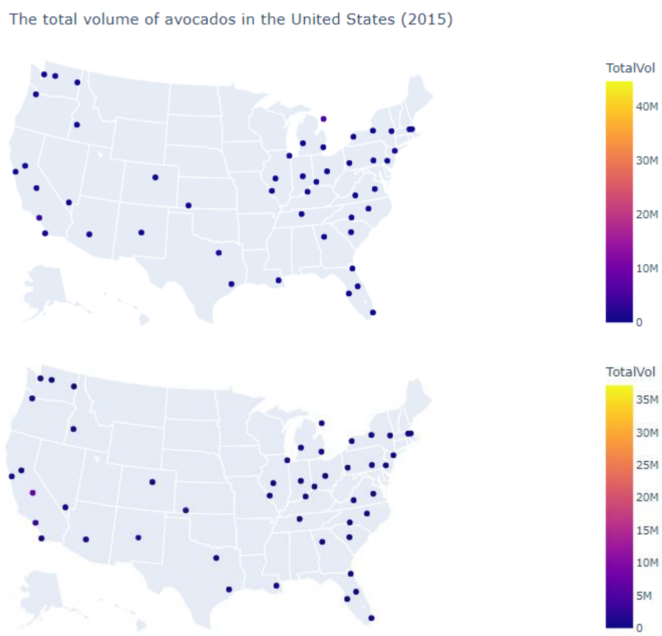


At the beginning of 2017, fewer sightings of high average prices are existent, and it is too evident that low avocado prices have been disregarded and instead higher prices have been opted for. In the second portion of figure 06 (the graph of average prices in 2017), if we observe the figure's colour legend, it is too apparent that the average prices of avocados are no longer $1.20, $0.6, $1.8 but instead either $1.00, $1.50, and $2.00.

TotalVol

Figure 07 displays the total volume of avocados in the cities of the United States in 2015 and 2017. Conducting basic analysis of figure 07, it is noticeable that both 2015 and 2017 total volume appears to be around 5 million units. While some cities exhibit values of hundreds of thousands in 2015, and others present values of approximately 41 million in the year of 2018.

Figure 07: TotalVol of avocados in 2015/2017



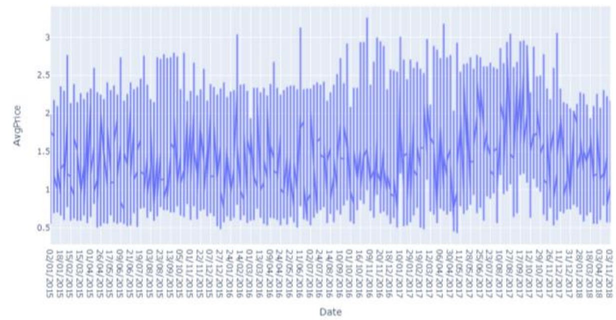
Temporal analysis

For the paper's temporal analysis, we decided to construct time-series plots of the key numeric attributes within the dataset. To construct these plots, we again use Plotly. We first begin by exploring the feature *AvgPrice* against a series of dates.

AvgPrice

The figure below illustrates the time-series plot of the *AvgPrice*. From the graph, there is a visible fluctuation in the *AvgPrice* of avocados across each year of the dataset. From the graph, we can conclude that there is no evidence of seasonality in the changes in *AvgPrice* and instead appears to fluctuate on a month-to-month basis.

Figure 09: AvgPrice time-series plot



At the beginning of the time-series plot, *AvgPrice* reaches a high of \$1.98 and a low of \$0.56, thus a \$1.42 difference across alternative regions of the United States. Towards the end of the plot, we can see that *AvgPrice* begins to curve upwards, however, begins to depreciate reaching a high of \$2.1 and a low of \$0.58.

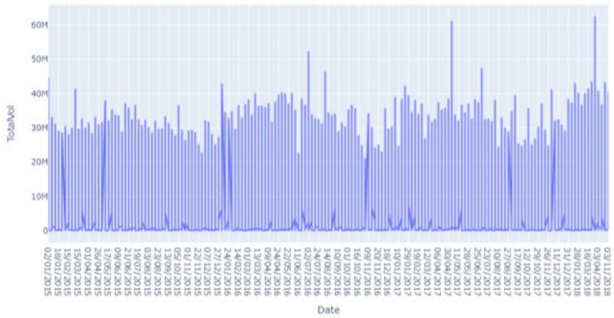
In accordance with the law of supply and demand, price and demand possess an inverse relationship. An increase in price will drive down demand and vice versa. It is assumed that we can attribute the changes in *AvgPrice* down to consumer demand. This will be later explored in the analysis of the *TotalVol* variable.

TotalVol

Figure 10 displays a time-series plot of *TotalVol*. Similar to the plot of *AvgPrice,* we can see there is constant fluctuation in the total volume of avocados consumed from the beginning to the end of the dataset. It is too visible that the fluctuations of TotalVol act in a similar manner, in regard to monthly appreciation and depreciation in the attribute's value.

The TotalVol in the beginning months of the dataset presents values as high as 44.65 million and as low as 769.05 accumulating a difference of approximately 44.649 million between alternate cities of the United States. Towards the end of the time-series plot, TotalVol amasses a high of 40.449 million and a low of 2,570.02. From this, though we cannot say that the demand has significantly increased over the years, it is evident that it has improved over time. While there have been reports of TotalVol accruing a value < 10,000, demand has not reached values as low as <1,000.

Figure 10: TotalVol time-series plot



From the aforementioned assumption of the law of supply and demand being present within the plot, at an average price of $1.98, the total volume equates to 44.65 million. At an average price of $2.34, total volume equals 21 million and at a price of $0.62, total volume equals 313,757 thousand. From these extractions from the dataset, we can say that price does not necessarily affect the amount of demand for avocados, however, this will further be explored upon constructing an ARIMA model.

RESULTS

According to Smith (2021) "provides a test of seasonality for including seasonal terms in your ARIMA models"[12]. Using the *auto_arima()* function from the *pmdarima* package, we input the model's dependent variable and retrieve the best ARIMA model to be ARIMA(1,0,0). From this, we can conclude that the model requires 1 lag, to be integrated of order, or I(0), and to have a moving average window of 0. With an I(0), it can be said that the data is stationary, and does not exhibit a significant increasing or decreasing trend. Using sklearn.model_selection's *train_test_split()* function. The data was split into training and testing. With 80% of the data aside for training and the remaining 20% held for testing.

There is no significant nor notable trend in the consumption of avocados. While it is existent their prices alter between different cities in the United States, there is no exact provided reason as to why this occurs. Though no significant trend has been detected, the consumption of avocados continues to grow and as does the conversation around the topic of avocados.

5 CRITICAL REFLECTION

Upon conducting a multiple linear regression model, we were able to gain valuable insight into the correlations of the variables through using the filter method and obtaining how each feature behaved amongst one another. Our multiple linear regression model could have further been explored, to provide a basis for model prediction or forecasting future values. Multiple linear regression models are reliable in machine learning and a relatively easy model to implement.

Our spatial analysis could have proven more useful if another method of geospatial analysis was implemented. The models utilized in the paper were relatively easy to construct and thus, this was chosen as opposed to a more complex model which would have harnessed greater visualizations as well as results.

The temporal analysis which this paper relied on provided valuable insight for the study. While again, more complex methods for visualization could have been composed, from the model's output, we were able to make good assumptions and make appropriate conclusions about the dataset. Within the study's temporal section, further explanation could have been conducted about the ARIMA model used to predict future values of avocado prices and total volumes. Reflecting onto the literature review of Jones et al's. (2021) study, this paper does not significantly differentiate on the amount of information provided about predictive modelling. In future studies, this lack of information will be improved upon to avoid an absence of conclusion and context about the study.

The ARIMA model constructed for the study did not provide enough information and was not successfully composed adequately to make appropriate predictions about the future trends or provide insights regarding the future.

Table of word counts

Abstract	149/200
Problem statement	230/250
State of the art	487/500
Properties of the data	494/500
Analysis: Approach	487/500
Analysis: Process	1148/1500
Analysis: Results	177/200
Critical reflection	269/500

References

[1] Soletrust (2021) Definition of avocado | What is, meaning and concept, *Soletrust.* Available at: https://soetrust.org/misc/definition-of-avocado-what-is-meaning-and-concept/ (Accessed 15th December 2021)

[2] Google Trends (2021) Avocado, *Google Trends*. Available at: https://trends.google.co.uk/trends/explore?date=2004-01-01%202021-01-01&geo=US&q=%2Fm%2F01645p (Accessed 15th December 2021)

[3] Karst, T. (2021) U.S. fruit imports surge so far in 2021, vegetables mostly steady, *The Packer.* Available at: https://www.thepacker.com/news/produce-crops/us-fruit-imports-surge-so-far-in-2021-vegetables-mostly-steady (Accessed 16th December 2021)

[4] FDA (2020) Food loss and waste, *U.S Food And Drug Administration.* Available at: https://www.fda.gov/food/consumers/food-loss-and-waste (Accessed 16th December 2021)

[5] Jones, V., Keyse, K., Melgoza, A., Perez, K., Qamar, T., Villalpando, J., Woo, J. (2021) Avocado Buying Trends in the United States Using SAC, *Introduction to Business Analytics.* Available at: https://arxiv.org/abs/2104.04649v1 (Accessed 17th December 2021)

[6] Ricon-Patino, J., Lasso, E., and Corrales, J. (2018) Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data, Sustainability, 10(10). Available at: https://doi.org/10.3390/su10103498 (Accessed 20th December 2021)

[7] Kiggins, J. (2017) Avocado Prices, *Kaggle.* Available at: https://www.kaggle.com/neuromusic/avocado-prices (Accessed 12th December 2021)

[8] Hass Avocado Board (2021) Category Data, *Hass Avocado Board.* Available at: https://hassavocadoboard.com/category-data/ (Accessed 13th December 2021)

[9] Tan, C. and Sulong, G., Shafry, M., Rahim, M., Chaw, J. and Tan, G. (2020) High Imperceptibility Medical Image Watermarking Scheme Based on Slantlet Transform by using Dynamic Visibility Threshold. DOI: 10.1109/ICIDM51048.2020.9339651 (Accessed 17th December 2021)

[10] Serrano, J. (2019) Accelerating Data Science Workflows, *BBVA AI Factory.* Available at: https://www.bbvaaifactory.com/accelerating-data-science-workflows/ (Accessed 20th December 2021)

[11] Baumöhl, E., and Lyócsa, Š., (2009) Stationarity of time series and the problem of spurious regression, Faculty of Business Economics in Košice. Available at: https://mpra.ub.uni-muenchen.de/id/eprint/27926 (Accessed 20th December 2021)

[12] Smith, T. (2021) pmdarima, *Alkaline ML.* Available at: https://alkaline-ml.com/pmdarima/tips_and_tricks.html (Accessed 23rd December 2021)