Avocado Prices by Region in the U.S.

BIOS 7400 Spring 2025 Final Project

Murphy John

2025-04-30

Dataset

The avocado, also known as the alligator pear, is an evergreen tree native to the Americas and was first domesticated in Mesoamerica over 5,000 years ago [1]. It likely originated in the highlands of south-central Mexico and Guatemala [2] [3] [4]. The fruit, commonly referred to simply as “avocado”, is a large berry containing a single seed or “pit” [3]. Today, avocados are cultivated in tropical and Mediterranean climates across many countries [2]. As of 2023, Mexico is the world’s leading producer, accounting for 29% of the global harvest of 10.4 million tonnes [5]. In the United States, California dominates production with 88%, followed by Florida (12%) and Hawaii (less than 1%) [5].

Avocados have become a global dietary staple, prized for their creamy texture and nutritional value. Global production has more than tripled since 2000, rising from 6 billion pounds to 19 billion pounds in 2021 [5]. U.S. consumption has followed this upward trend [6]. As health-conscious eating grows in popularity, tracking avocado pricing trends has become important for both consumers and industry stakeholders.

This report analyzes weekly Hass avocado prices from 2015 to mid-2018 across various U.S. regions. It also incorporates global temperature change data from NASA and includes the political party of the U.S. president during this period.

Data Management

All data processing and analysis for this report was generated using SAS software, Version 9.4 of the SAS OnDemand for Academics System. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Avocado Data

The primary dataset used in this project was sourced from Kaggle and originates from the Hass Avocado Board [7]. It includes weekly average prices for Hass avocados across U.S. regions from January 2015 to March 2018, with distinctions between conventional and organic types. In this analysis, average price is treated as the outcome variable, while month, year, region, and avocado type serve as predictors.

The raw data was imported into SAS using the PROC IMPORT procedure, creating a dataset named raw_avo. The GUESSINGROWS=MAX option ensures accurate detection of data types by examining all rows.

Data preprocessing occurs in the clean_avo data step. Selected variables are renamed for clarity, and the date field is parsed to extract month and year, which are reformatted to align with monthly temperature data. A new month variable is created with the STRIP and PROPCASE functions, and the original date is standardized to the first of the month using MDY. Using the IF THEN statement, the data is filtered to retain only relevant U.S. regions. An unused variable (VAR1) is removed using DROP.

Next, PROC SQL aggregates the data by year, month, region, and avocado type using GROUP BY, calculating average prices and summing total volume, sizes, and bag counts. Finally, PROC SORT removes duplicate observations with NODUPKEY and BY based on date, region, and type, producing the final cleaned sorted dataset.

Temperature Data

Global climate change can impact the environmental conditions essential for agriculture, including temperature, which may influence crop yields and market prices. Based on this, we hypothesize that temperature is a meaningful predictor of average avocado prices. To explore this relationship, we use Northern Hemisphere monthly temperature anomaly data from NASA [8], which reports how much each month's average temperature deviates from a 1951–1980 average baseline. These anomalies, rather than absolute temperatures, provide insight into warming trends over time.

The accompanying SAS code reads and processes this raw temperature data for analysis. A FILENAME reference (raw_temp) is assigned to the text file containing the anomaly values. Using absolute pointer control and IF THEN, the code inspects each line and filters out any non-numeric rows. For valid entries, it extracts the year and monthly anomaly values using named INPUT.

The dataset is then restricted to the years 2015 through 2018 with IF THEN and OR. Using an ARRAY and a DO loop, temperature anomalies are converted to standard degrees Celsius by dividing by 100 and rounding to two decimal places and the data is reshaped from a wide to a long format, where each row represents a year-month pair with its corresponding anomaly in a single temp column. Month names are standardized to proper case using STRIP and PROPCASE, and only the year, month, and temp variables are retained in the final dataset.

Presidential Data

We create a reference dataset containing the U.S. president and their political party for each year from 2015 to 2018. We hypothesize that presidential political party affiliation may influence avocado prices, making it a relevant variable for analysis. The SAS code defines a dataset named dat_pres with three variables: year (numeric), president

(character), and pres_party (character). The INFILE DATALINES statement reads inline text, using spaces as delimiters and allowing quoted values, as established by the DSD and DLM statements. The dataset assigns Barack Obama (Democratic Party) to 2015–2016 and Donald Trump (Republican Party) to 2017–2018.

Final Processed SAS Dataset

Finally, we must merge all three data sets. Using PROC SQL, the SAS code performs two INNER JOIN statements: first, it joins the avocado and temperature datasets on year and month; then, it joins the result with the presidential dataset using year as the key. This ensures only records with matching entries across all datasets are retained, appropriate for a many-to-one merge, where multiple avocado records may correspond to a single presidential year.

The merged dataset is temporarily stored in the work library, then saved permanently as dat in a subsequent data step. During this step, LABEL is used to add descriptions to each variable. The final dataset is clean, labeled, and ready for exploration and statistical modeling.

Analysis

Exploratory

The analysis begins with the setup of necessary libraries, enabling of graphics output using ODS, and sorting of the merged dataset by avocado type, date, and region with PROC SORT. We use PROC SGPANEL to visualize trends in avocado pricing over time and across various factors. Figure 1 shows a time-series plot of average price by avocado type and region. The temporal patterns suggest that both factors influence pricing trends. Boxplots stratified by political party (pres_party) and region indicated some variation in average prices under different presidential administrations, while scatterplots of price versus temperature revealed a weak to moderate relationship depending on the avocado type and region.

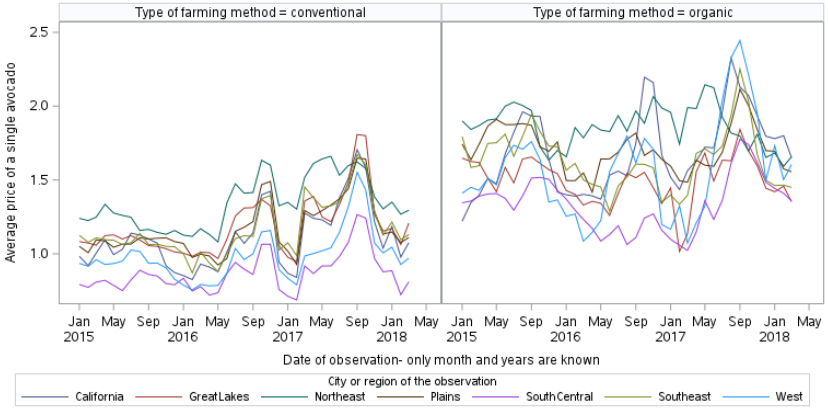

Figure 1: Time series of avocado prices per month by type and region.

We use PROC UNIVARIATE to compute summary statistics on mean prices by avocado type. PROC CORR PEASRON reveals a negative Pearson correlation between average price and temperature, with little variation by type. Using a MACRO and PROC MEANS, additional mean comparisons were performed for categorical variables including year, month, region, and presidential party. Table 1 shows the resulting descriptive statistics of avocado prices by region, stratified by type. In all regions, the mean prices for organic avocados are higher than those for conventional avocados. Further, the SouthCentral region has the lowest mean across both avocado types and the Northest has the highest.

The MEANS Procedure

Type of farming method=conventional

Analysis Variable : avgprice Average price of a single avocado						
City or region of the observation	N Obs	N	Mean	Std Dev	Minimum	Maximum
California	39	39	1.1050769	0.2130309	0.8250000	1.7050000
GreatLakes	39	39	1.1806282	0.2003116	0.9475000	1.8075000
Northeast	39	39	1.3442821	0.1775740	1.0800000	1.6600000
Plains	39	39	1.1639615	0.1826189	0.9240000	1.6475000
SouthCentral	39	39	0.8679103	0.1325447	0.6875000	1.2650000
Southeast	39	39	1.1620128	0.1870763	0.8700000	1.6725000
West	39	39	0.9838974	0.1689162	0.7550000	1.5525000

Type of farming method=organic

Analysis Variable : avgprice Average price of a single avocado						
City or region of the observation	N Obs	N	Mean	Std Dev	Minimum	Maximum
California	39	39	1.6847821	0.2718256	1.2200000	2.3250000
GreatLakes	39	39	1.4931667	0.1586735	1.0150000	1.8425000
Northeast	39	39	1.8608333	0.1397678	1.5925000	2.1450000
Plains	39	39	1.7066282	0.1540389	1.4175000	2.1150000
SouthCentral	39	39	1.3325513	0.1840713	1.0225000	1.7750000
Southeast	39	39	1.6325897	0.2027917	1.2875000	2.2525000
West	39	39	1.5609615	0.3276098	1.0725000	2.4450000

Table 1: Descriptive statistics of avocado prices by region and type.

Model Fits

The modeling phase began with data splitting. Using PROC SURVEYSELECT, IF THEN, ELSE, and OUTPUT, 80% of the dataset was allocated to training and the remaining 20% reserved for testing. PROC GLM fit a simple linear regression using avocado type as a predictor. This baseline model had an RMSE of 0.251 and R-squared of 0.486, indicating moderate explanatory power. Subsequent bivariable models were fit with MACRO and incorporate both main effects and two-way interaction terms between type and the additional covariate. Among these, the model including region as a covariate yielded the best performance (RMSE = 0.203, R-squared = 0.673), suggesting strong regional influences on price. Other covariates considered in the bivariable analysis included temperature, year, presidential party, and month. Table 2 reports all RMSE and R-squared metrics of the baseline and bivariable models. All bivariable models improved the fit as measured by RMSE and R-squared compared to the “Type only” univariable baseline.

Covariate	RMSE	R_Square
Type only	0.25115	0.48568
Temperature	0.23479	0.55258
Region	0.20299	0.67329
Presidential Party	0.24441	0.51518
Year	0.23964	0.53820
Month	0.23113	0.58645

Table 2: RMSE and R-Squared values for bivariable regression models by second predictor

Next, a full model incorporating all covariates and their two-way interactions with type was fit with `PROC GLM`. This model showed substantial improvement in fit (RMSE = 0.145, R-squared = 0.845), but it included several interaction terms with weak statistical support, as assessed by interaction plots. A reduced model excluding the unnecessary interactions was fit. This model achieved a very similar fit (RMSE = 0.152, R-squared = 0.823), but with improved parsimony, making it the preferred final model. The final model includes type, temperature, region, presidential party, month, year, and type-region interaction.

Model evaluation on the withheld test data confirmed the reduced model's robustness. Predictions were generated on the test data using `PROC PLM`, and residuals were calculated. Figure 2 shows the model predictions versus the observed values in the test data. Final performance metrics showed the reduced model maintained high explanatory power on unseen data, with a test RMSE of 0.133 and R-squared of 0.854. These results support the conclusion that avocado prices are strongly influenced by type, region, time (year and month), temperature, and political context.

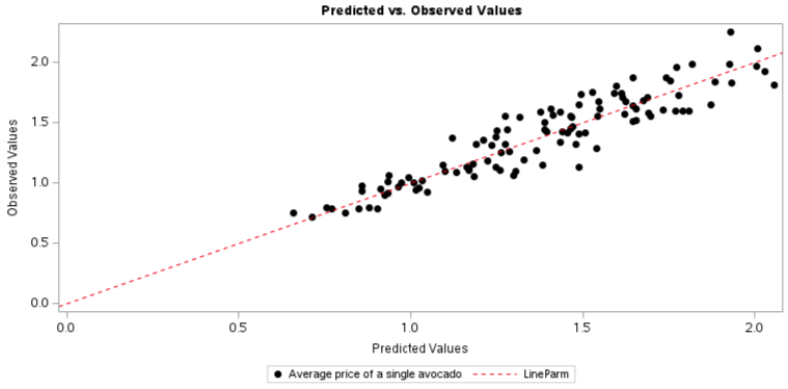

Figure 2: Model predictions versus observed test data

Conclusions

The generalized linear model demonstrated strong performance in predicting avocado prices based on type, temperature, region, time, and presidential party. Notably, the region variable emerged as the most influential predictor, as evidenced by the bivariable analysis. These findings provide insights into the factors that drive avocado pricing. Future work could include forecasting methods, such as ARIMA models and exponential smoothing for predicting future avocado prices. By utilizing such methods, this research could offer a more comprehensive understanding of price trends over time, benefiting both consumers and industry stakeholders.

References

[1] "Avocado," *Wikipedia*. Apr. 2025. Accessed: Apr. 30, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Avocado&oldid=1287604678#cite_note-morton-4

[2] J. Morton, "Avocado." Accessed: Apr. 30, 2025. [Online]. Available: https://hort.purdue.edu/newcrop/morton/avocado_ars.html

[3] UC Riverside, "Avocado Variety Collection College of Natural & Agricultural Sciences." Oct. 2023. Accessed: Apr. 30, 2025. [Online]. Available: https://avocado.ucr.edu/

[4] H. Chen, P. L. Morrell, V. E. T. M. Ashworth, M. de la Cruz, and M. T. Clegg, "Tracing the Geographic Origins of Major Avocado Cultivars," *Journal of Heredity*, vol. 100, no. 1, pp. 56–65, Jan. 2009, doi: 10.1093/jhered/esn068.

[5] "FAOSTAT." Accessed: Apr. 30, 2025. [Online]. Available: https://www.fao.org/faostat/en/#data/QCL

[6] "FE1150/FE1150: An Overview of the Avocado Market in the United States," *Ask IFAS - Powered by EDIS*. Accessed: Apr. 30, 2025. [Online]. Available: https://edis.ifas.ufl.edu/publication/FE1150

[7] "Avocado Prices." Accessed: Apr. 30, 2025. [Online]. Available: https://www.kaggle.com/datasets/neuromusic/avocado-prices

[8] GISTEMP Team, "GISS Surface Temperature Analysis (GISTEMP), version 4." NASA Goddard Institute for Space Studies. Available: https://data.giss.nasa.gov/gistemp/