Avocado Price Forecasting with Gaussian Processes

Anna Billiard
Data Science
University of Michigan
Ann Arbor, MI 48104
annabill@umich.edu

Rohan Janakiraman
Economics
University of Michigan
Ann Arbor, MI 48104
rohanja@umich.edu

Alex Brace
Computer Science
University of Michigan
Ann Arbor, MI 48109
braceal@umich.edu

Yi Zhang
Computer Science and Data Science
University of Michigan
Ann Arbor, MI 48104
zackLt@umich.edu

Abstract

Time series data often pose difficult, non-linear modeling problems, and predictions tend to suffer from the inability to quantify uncertainty. We investigate alternatives to the traditional ARIMA methods of time series analysis by considering a Bayesian Gaussian process with a composite kernel function as the prior. The model fitted on historic avocado prices from 2015 to 2018 contained 94.7% of the sample/observed data and 94.1% of the out-of-sample/unobserved true data, despite having a fairly minimal feature space. This paper attempts to establish that for price prediction problems, Gaussian processes are an effective and interpretable model choice.

1 Introduction

Gaussian processes (GP) present an opportunity to fit heavily non-linear models using Bayesian inference in a way that makes quantifying uncertainty extremely easy. For instance, as a direct result of its Bayesian nature, GPs can generate posterior predicate check (PPC) intervals for credibility ranges enclosing the predictions. The uncertainty quantification distinguishes GPs from many other machine learning approaches.

As a demonstration, we fit a Gaussian process model to historic avocado prices from 2015 to 2018. We used several optimization techniques such as Broyden–Fletcher–Goldfarb–Shanno (BFGS) and random search to tune model hyperparameters. The resulting model performed satisfactorily despite having data limited in quantity and quality. The 95% PPC intervals we generated encompassed most of the true data points in our test dataset, validating our model and the use of GPs in price forecasting.

2 Data

Hass avocados are one of the primary types of avocados sold in the United States. The non-profit organization Hass Avocado Board (HAB) collected the avocado price data through Information Resources' channels with "multi-outlet stores... based on actual retail sales of Hass avocados" [1]. HAB published the dataset including weekly average prices for individual Hass avocados sold in grocery stores segmented by organic and conventional and by state, as well as nationally and regionally aggregated averages [2].

With 161 weeks of data and 18,249 individual observations (averages for different places and types of avocado) from January 2015 through March 2018, the dataset spans a wide range of areas. Figure 1

STATS 451 - Bayesian Data Analysis (University of Michigan, Fall 2019), Ann Arbor, MI.

shows a histogram of observations. The bi-modality results from the difference between organic and conventional avocado prices. Note the right tail. Our hypothesis is that the tail could be attributed to two separate things: the year 2016 and 2017 having high outliers (see figure 2), and the fact that some states have higher avocado prices in general (see California and Connecticut in Figure 6, appendix A).

Figure 2 shows a slight upward shift in prices in 2017 (by approximately $0.15). We attribute this uptick to a growers' strike in Mexico and a concurrent drought in California [3].

For our model, we chose to only use the prices of organic avocado in West Texas and New Mexico, a predefined region in the dataset. There are three reasons for the choice of location. First, although it may seem that the U.S. average prices would be the best choice, it turns out to contain some quite abnormal price points with wide fluctuations that would make modeling unnecessarily challenging for the demonstration. Second, West Texas and New Mexico, in particular, have easy access to avocados while not having a lot of avocado production [4]. Additionally, the choice to work only on organic avocados lies in the fact that its larger variance would better highlight the pattern fitting ability of GP (see Figure 3 and Table 3 in Appendix A).

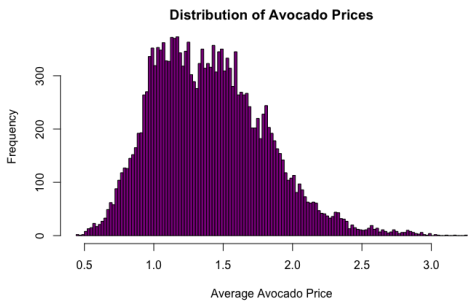

Figure 1: Right-skewed bi-modal distribution of avocado prices.

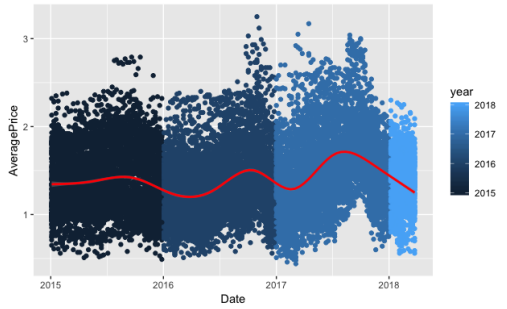

Figure 2: Spike in summer 2017 prices due to drought and growers strike.

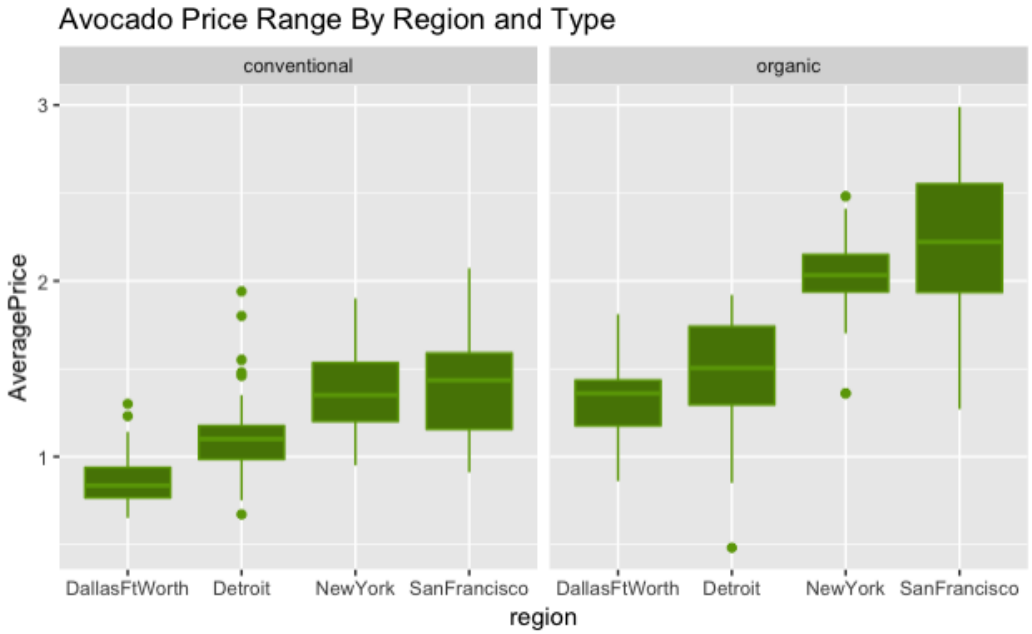

Figure 3: Price distribution of organic avocados vs. conventional avocados in various areas.

3 Methods

3.1 Definition of Gaussian Process

A Gaussian process fits a non-parametric model that assumes observations in the data set follow a multivariate-normal (MVN) distribution with a mean function and a kernel function to define the covariance matrix. The initial kernel and mean function define the MVN prior, while our training data are used to model the data likelihood function. The prior and likelihood functions are then used to make Bayesian updates to arrive at a conjugate posterior model. We define the GP model below as adapted from *Gaussian processes for machine learning* [5].

Let $\mathcal{X}$ be the set of weekly enumerated time intervals $\mathbf{X}$, where $\mathbf{X} = (x_1, ..., x_n)$ is the indices of weekly avocado price observations.

Let $(f(x_1), ..., f(x_n))$ s.t. $f(x_i) \in \mathbb{R}$ be the time series of avocado price observations.

Define mean function $\mu : \mathcal{X} \to \mathbb{R}^n$ where $\mu(\mathbf{X}) = \mathbb{E}[f(\mathbf{X})]$. It is often the mean of the training set if the prior belief holds the variable mean(s) to be stationary, as in our case.

Define covariance function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{n \times n}$ where $\kappa(\mathbf{X}, \mathbf{X}') = \mathbb{E}[(f(\mathbf{X}) - \mu(\mathbf{X}))(f(\mathbf{X}') - \mu(\mathbf{X}'))]$.

We define a Gaussian process as: $(f(x_1), ..., f(x_n)) \sim MVN(\mu(\mathbf{X}), \kappa(\mathbf{X}, \mathbf{X}'))$.

3.2 Kernel Function Selection

We found that the most important part of specifying our model to be kernel function selection. The kernel function encodes prior knowledge about the data. From exploratory data analysis, we concluded that the avocado price series has a cyclic pattern over a one year period. To account for both the seasonal trend and the autoregressive correlation of neighboring data points, we decided to use exponential sine squared (ESS) kernel function, a sinusoidal variant of the common radial basis function (RBF) kernel, equation (1), implemented by scikit-learn [10]. This kernel has two hyperparameters, $p > 0$ which accounts for the periodicity and $l > 0$ which accounts for the decay of importance between far away points. To initialize the hyperparameters we selected $p = 358$ and $l = 20$ to account for a yearly periodicity.

$$\kappa(x_i, x_j)_{ESS} = exp\left\{ \frac{-2}{l^2} sin^2(\frac{\pi}{p} \cdot \|x_i - x_j\|_2) \right\} \tag{1}$$

For robustness, we also added the rational quadratic (RQ) kernel function, equation (2). This kernel accounts for medium term irregularities according to scikit-learns documentation [6]. Introducing this kernel helps account for the irregular price spikes such as the one in 2017. This kernel also has two hyperparameters, $l > 0$ which has the same effect as above and $\alpha > 0$ which accounts for relative weighting of small and large scale variations [7]. To initialize the hyperparameters we selected $\alpha = 20$ and $l = 80$ to account for a short term effects.

$$\kappa(x_i, x_j)_{RQ} = \left(1 + \frac{\|x_i - x_j\|_2^2}{2\alpha l^2}\right)^{-\alpha} \tag{2}$$

Lastly, we define a white noise (WN) kernel function which acts as a regularizer in our model, equation (3). This kernel works by adding a constant value ϵ to the diagonal of the covariance matrix i.e. Tikhonov regularization [9]. Here epsilon is learned from the global noise in the training data.

$$\kappa(x_i, x_j)_{WN} = \mathbb{I}(x_i = x_j) \cdot \epsilon \tag{3}$$

Combining equations (1-3), we define a composite kernel function, equation (4), that we used to model and forecast the avocado prices. Note, the summation of kernel functions is still a valid differentiable kernel function since this operation effectively makes the covariance matrix in the MVN a summed matrix with non-negative diagonal values. For this model, the fitted/posterior hyperparameters where determined through optimization to be $l = 4.57$ and $p = 337$ for $\kappa(x_i, x_j)_{ESS}$, $\alpha = 100000.0$ and $l = 21.7$ for $\kappa(x_i, x_j)_{RQ}$, and $\epsilon = 0.0158$ for $\kappa(x_i, x_j)_{WN}$. The coefficient on $\kappa(x_i, x_j)_{RQ}$ was determined from the training data during fitting.

$$\kappa(x_i, x_j) = \kappa(x_i, x_j)_{ESS} + 0.0497 \cdot \kappa(x_i, x_j)_{RQ} + \kappa(x_i, x_j)_{WN} \tag{4}$$

3.3 Model Implementation

To implement our model, we used scikit-learn, a Python library, which has several built-in kernel functions that we arranged in a linear combination. To optimize kernel hyperparameters we employed the limited memory BFGS optimizer implemented by scikit-learn [6], which optimizes kernel hyperparameters during model fitting by maximizing the log-marginal likelihood using a gradient ascent iteration [8]. To avoid getting trapped in local optima in the hyperparameter space (such as a high noise model), we restart the optimizer a fixed number of times (typically 10). The first run of the optimizer is initialized with a best guess prior estimate and subsequent optimizations are initialized via a random sample from the hyperparameter space.

To reduce over-fitting to the training data, we regularize our model using Tikhonov regularization. This is achieved by the addition of a white noise kernel to our kernel function which models the global noise in the training data and adds it to the MVNs covariance matrix diagonal.

4 Results

To assess model fit, we split our data into training and testing sets with an 80-20 split into each category respectively. The training data runs from January 1, 2015 to July 31, 2017 while the testing data runs from August 06, 2017 to March 18, 2018.

Figure 4 shows our model against the training and test data. The dark blue band represents the GP uncertainty and the light blue band represents the 95% PPC interval. Our model showed excellent results with 94.1% of testing data and 94.7% of the training data within the 95% PPC interval. The fact that the training and testing sets showed very similar prediction accuracy is evidence that our model generalized well to off-sample data, proving the success of the regularization. As a comparison, we have included Figure 5, which shows the original data and a regression done using splines. Evidently, the GP model fits the volatility in the data much better.

The 95% PPC interval also demonstrates the utility of a GP model. Not only can we forecast a price on a particular week but we can also provide our uncertainty of the estimate with mathematically sound credibility.

Gaussian process model performance			
Data set	RMSE	R^2	Accuracy
Train	0.109	0.874	94.7%
Test	0.295	-0.796	94.1%

Table 1: Reported statistics for GP model performance in the West Texas/New Mexico region.

Table 1 above has relevant statistics. The quantity R^2 is defined as $1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$, or in other words one minus the variance of our model divided by the variance of the data itself. Note that this can be negative if the model variance is higher than the data variance or because of systematic bias i.e. overfitting. Our R^2 value for the testing set is indeed negative, but it is important to point out that this does not make it a bad model and is likely due to the large posterior variance rather than overfitting. Root mean squared error (RMSE) was calculated by comparing our models predicted price of organic avocados in West Texas/New Mexico to the true values. An RMSE of 0.295 in the test set reveals that our model did fairly well at explaining the variance of the test set, though this model leaves lots of room for improvement (see discussion).

Our model predicted mean for the test period was 1.655 and the true mean over this time was 1.887. The optimized hyperparameter for the kernel function that described the periodicity ("ExpSineSquared" in Sklearn) was 337, which is about the same as the number of days in a year. This is of course not coincidental, as plants are heavily dependent on the yearly cycle of seasons.

Our white noise parameter, which was responsible for describing random fluctuations in the data, was 0.0158. The interpretation of this is that the price could randomly fluctuate by 1 or 2 cents, which is quite reasonable for produce prices. For our rational quadratic function, our α value was 100,000,

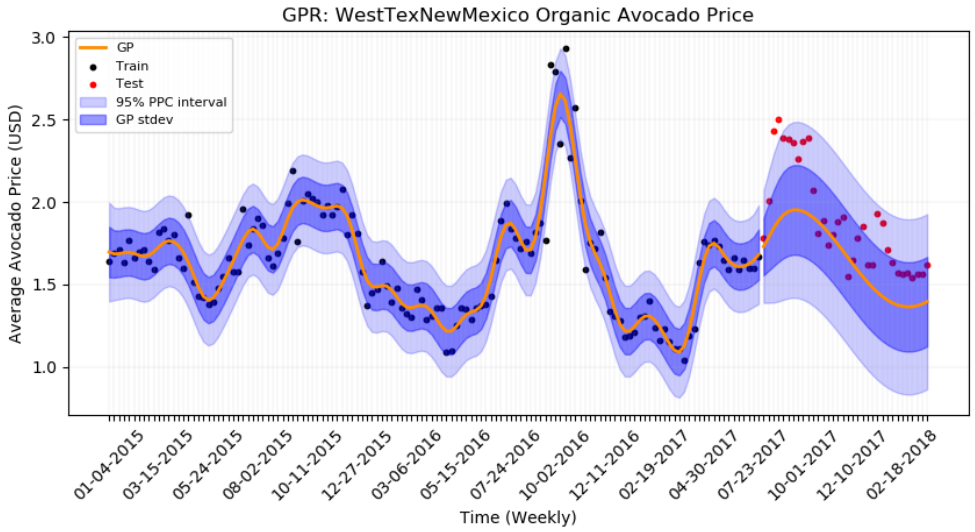

Figure 4: Gaussian process regression results for organic avocado prices in the West Texas/New Mexico region. The training and testing data are shown in black and red respectively. The regression line is shown in orange and the GP uncertainty (one standard deviation) and 95% PPC interval are depicted by the dark blue band and light blue band respectively. The 95% PPC interval accounts for 94.1% of testing data and 94.7% of training data.

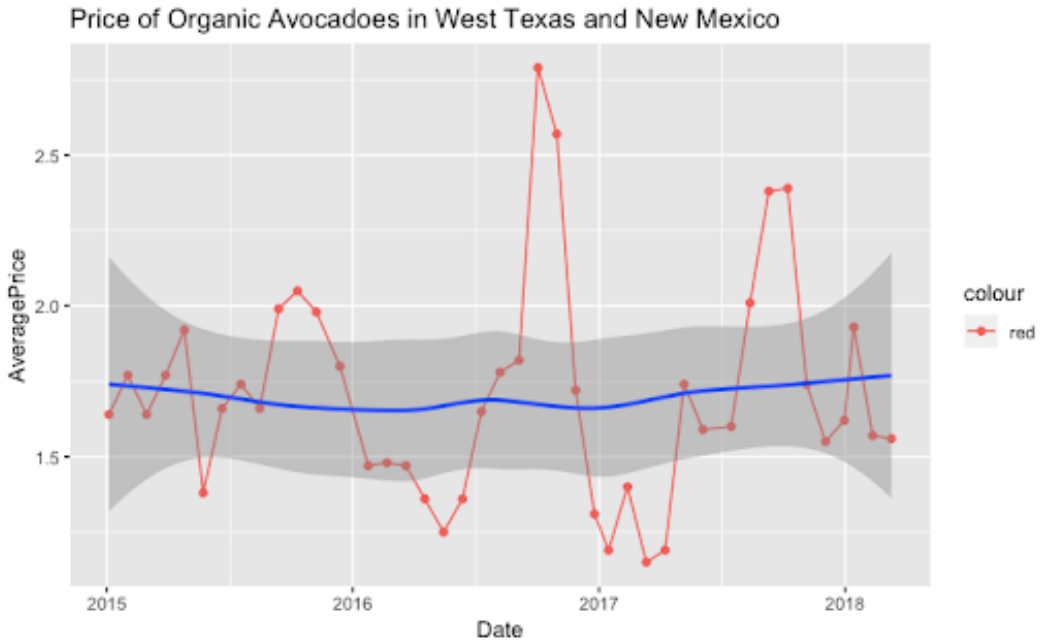

Figure 5: Standard regression results for organic avocado prices in the West Texas/New Mexico region. The regression line is shown in blue and the 95% confidence interval is depicted by the grey bands.

which suggests that the rational quadratic kernel is quite close to a squared exponential type kernel [7]. Our l value was just over 21, or three weeks, which means that the window of points that this model considered relevant to the point it was attempting to fit was approximately 3 weeks on either side of the point itself. Thus, with the combination of the rational quadratic and the ExpSineSquared

kernels, we are able to account for both long and short term periodic trends which explains why this model performed so well off-sample.

5 Discussion

Some drawbacks to our analyses should be noted. There are many other factors that drive avocado prices other than the price alone. Factors such as tariffs, farm subsidies, additional growers' strikes (possibly with their own cyclical patterns), persecution of immigrants who disproportionately make up the agricultural workers in the U.S., and the climate crisis all affect the market for avocados but are not included in our data [11]. To illustrate the need to include such factors, consider the aforementioned Mexican growers' strike, which contributed to the price spike that we observed in the 2017 data. After doing more research, we found that these strikes are a semi-regular event, as there was also a strike in November 2018 which disrupted shipments from Mexico [12]. The regularity of these strikes is likely tied to the length of the union contracts the avocado growers have with APEAM, an avocado industry group that manages the "Avocados from Mexico" brand. Future research could be done to ascertain the magnitude of the price shocks such strikes cause as well as their seasonality.

Future work along these lines would ideally culminate in a higher-dimensional GP model. This model would have several more ExpSineSquared kernels to model more types of periodicity in the dataset. This would allow us to model for several of those other factors discussed above in addition to the year, which could be useful when we procure the aforementioned additional data.

Despite the low dimensionality of our data, Gaussian processes have important positives. Gaussian process regression has an advantage over loss-function based models, since it learns a generative model of the data that can be used to provide credible intervals for predictions. In loss-function based models, we do not have any knowledge of the stability or uncertainty around predictions. Thus, GP models are very interpretable and provide robust predictions for off-sample data.

In addition, for low dimensional data, as in the case of Avocado prices, optimizing hyperparameters for GPs can be done very efficiently by taking advantage of gradient information implicit in the kernel function. This is a significant advantage over loss-function based black box models such as neural networks, which resort to grid search, Bayesian optimization, etc. for optimizing hyperparameters.

6 Conclusion

Overall, our Gaussian process model performed quite well. The component kernel functions ended up with parameters that reflected concrete aspects of the data, such as the periodicity of approximately one year. Though our RMSE for the test set was nearly 0.3, we attribute this to the low-dimensionality of our data set. Additionally, the accuracy of the 95% PPC indicates that our model works quite well on the data as a whole, though perhaps less well at individual points.

We are pleased with the results of using Gaussian processes to fit a Bayesian model to the avocado price dataset. Learning a model that did not overfit to the data and had predictive power is a clear indication that Gaussian processes were appropriate in this case.

7 Contributions

Anna Billiard - Research into Gaussian processes, doing Gaussian processes with R, kernel functions; the presentation slides and presentation to the class; abstract, introduction, data, conclusion, the bulk of discussion, and parts of methods and results sections of paper; editing of the overall paper.

Alex Brace - Research into Gaussian processes and how they work; the data segmentation in Python; fitting, tuning and visualizing Gaussian process models in Python; built avocado Python package for rapidly testing GP models; the methods section and part of the discussion section in the report.

Rohan Janakiraman - Exploratory data analysis in R, calculated summary statistics, most of the figures, tables 2 and 3 in appendix A, all of appendix B, references section, some of the discussion section, the research behind the price spike in 2017, organized project materials, some editing.

Yi Zhang - Research and coding Gaussian processes in R and Python, kernel functions; studying Gaussian process papers and reports by Hass Avocado Board; code repository management; evaluating the model results; writing parts of the Discussion of the paper; overall editing of the paper.

References

[1] Carman, Hoy F., Tina L. Saitone, and Richard J. Sexton. (2013). Five-year evaluation of the Hass Avocado Board's promotional programs: 2008-2012. *Hass Avocado Board*.

[2] Kiggins, J. (2017). Avocado Prices: Historical data on avocado prices and sales volume in multiple US markets. *Kaggle*.

[3] Perez, M. G., & Durisin, M. (2017). Avocado Prices Are Skyrocketing. *Bloomberg*.

[4] Nesbitt, M., Stein, L., & Kamas, J. (2015). Avocados. *Texas AM AgriLife Extension Service*.

[5] Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning. *MIT press*.

[6] Pedregosa, F. et. al. (2011). Scikit-learn: Machine Learning in Python. Gaussian Process Regressor. *Journal of Machine Learning Research, 12*, 2825 - 2830.

[7] Duvenaud, D. (2014). Automatic model construction with Gaussian processes (Doctoral dissertation, *University of Cambridge*).

[8] Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. Mathematical programming, 45(1-3), 503-528.

[9] Golub, G. H., Hansen, P. C., & O'Leary, D. P. (1999). Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications, 21*(1), 185-194.

[10] Pedregosa, F. et. al. (2011). Scikit-learn: Machine Learning in Python. Exp Sine Squared. *Journal of Machine Learning Research, 12*, 2825 - 2830.

[11] Rooney, K. (2019). Avocados will probably get more expensive under Trump's Mexico tariffs. *CNBC*.

[12] Chiwaya, N. (2018). Can't get your hands on avocados lately? Here's why. *NBC News*.

A Avocado Data Summary Statistics & State-by-State Price Breakdown

Data Summary Statistics					
Variable	Num of Obs.	Mean	Std Dev	Range	IQR
AveragePrice	18249	1.406	0.402	[0.44, 3.25]	[1.10, 1.66]
Total Volume	18249	850644	3453545	[85, 62505647]	[10839, 432962]
4046	18249	293008	1264989	[0, 22743616]	[854, 111020]
4225	18249	295155	1204120	[0, 20470573]	[3009, 150207]
4770	18249	22840	107464.1	[0, 2546439]	[0, 6243]
Total Bags	18249	239639	986242.4	[0, 19373134]	[5089, 110783]
Small Bags	18249	182195	746178.5	[0, 13384587]	[2849, 83338]
Large Bags	18249	54338	243966	[0, 5719097]	[127, 22029]
XLarge Bags	18249	3106.4	17692.89	[0.0, 551693.7]	[0.0, 132.5]

Table 2: The variables *4046*, *4225*, *4770* denote the total number of avocados with PLU 4046, PLU 4225 and PLU 4770 sold respectively. PLU means 'Price Look Up' code and they have been used by supermarkets to make inventory control easier.

Conventional Avocado Prices vs. Organic Avocado Prices				
Type	Num of Obs.	Mean	Variance	IQR
Organic	9123	1.654	0.132	[1.42, 1.87]
Conventional	9126	1.158	0.069	[0.98, 1.32]

Table 3: Comparison of average avocado prices for both conventional and organic avocados using the variable *AveragePrice*.

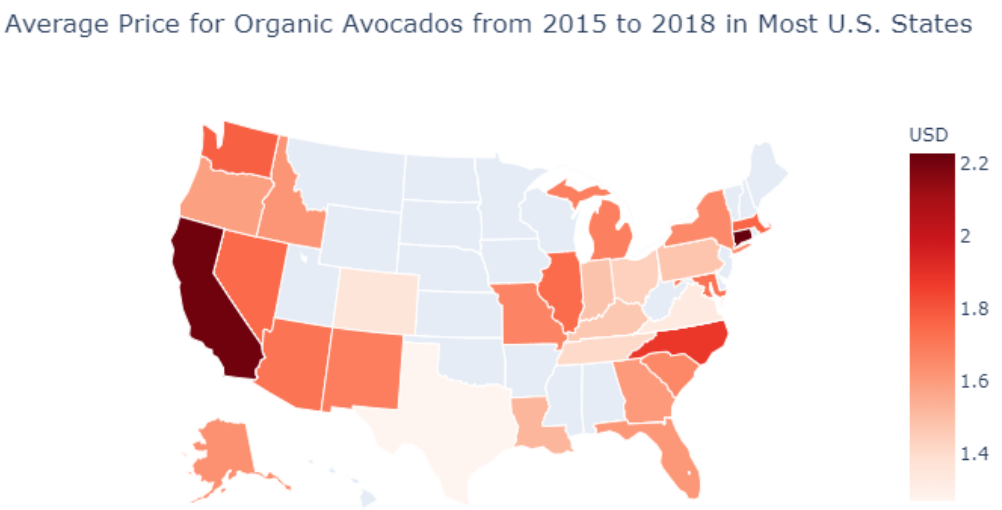

Figure 6: Average prices in most U.S.

B Avocado Price Distribution Plots, By Year