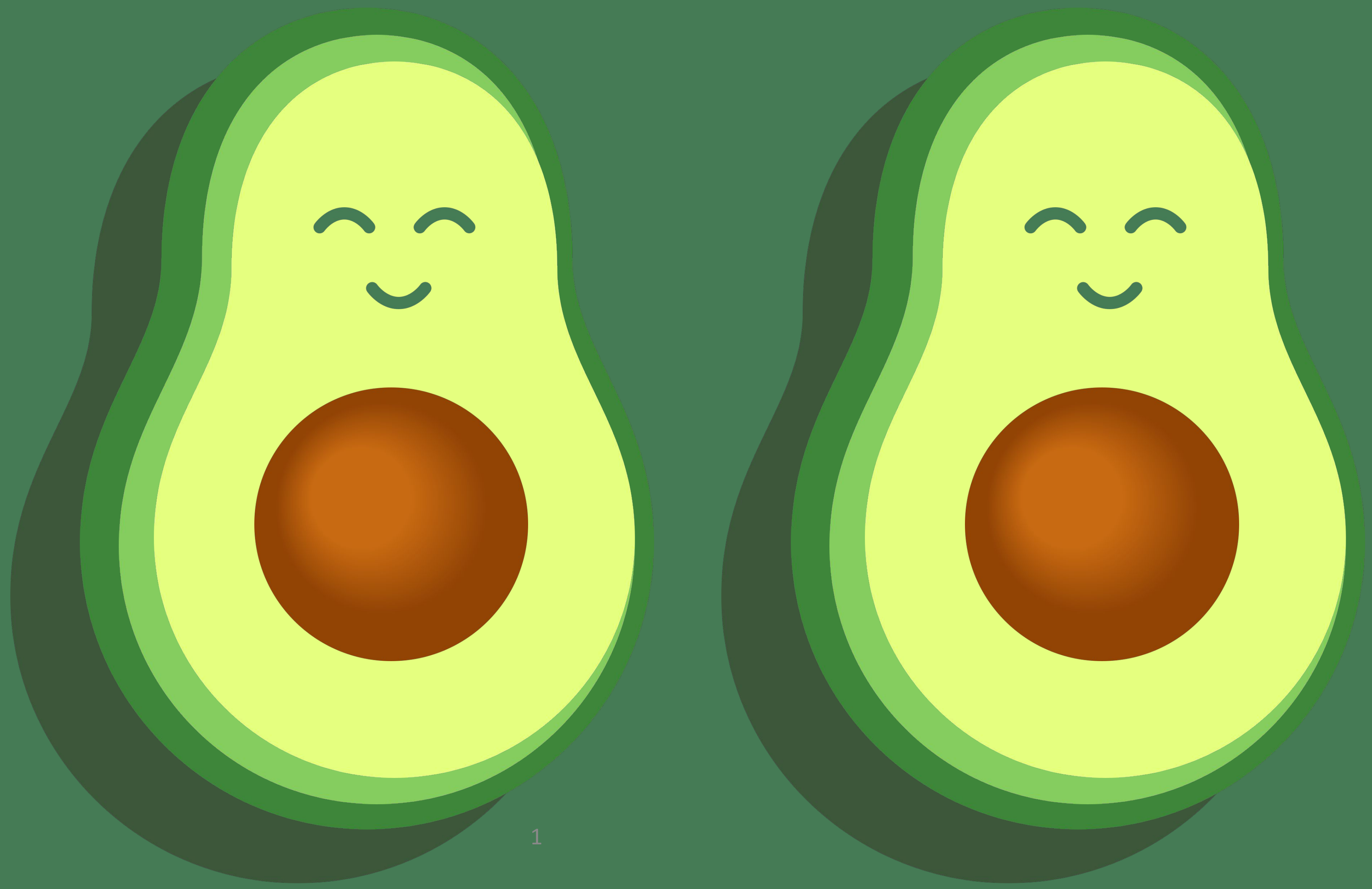
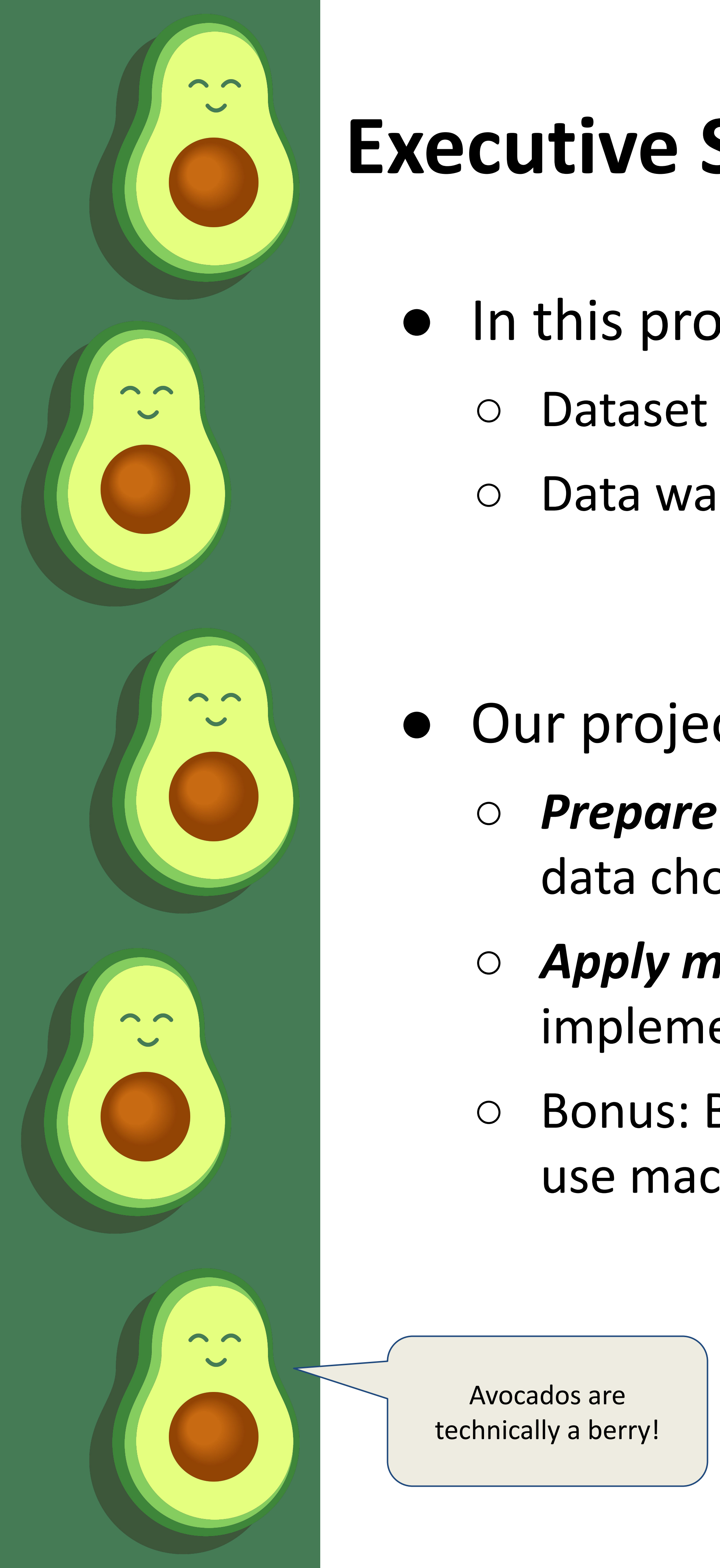


USING MACHINE LEARNING TO PREDICT AVOCADO SALES

PRESENTED BY: THE AVOCADO HARVESTERS

PROJECT 2: MARCH 19, 2025



Executive Summary

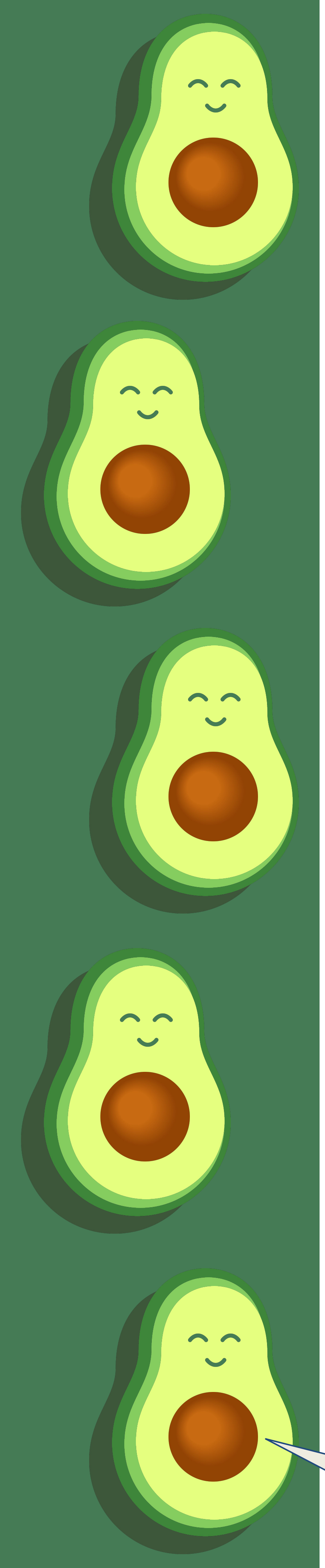
- In this project, we leveraged several machine learning tools to analyze Avocado Sales
  - Dataset covered 2015-2017 and Jan - March of 2018
  - Data was sourced from the Haas Avocado Board and Kaggle

- Our project goals:
  - ***Prepare and analyze the data in parallel***, to better understand the impact of prepping and data choices on model effectiveness
  - ***Apply machine learning models also in parallel,*** to also compare our preparation and implementing of similar models can generate different accuracy and results
  - Bonus: Because avocados are considered to be a luxury item, we wanted to see if we could use machine learning to determine the ***price elasticity of demand***

Data Prep and Estimates

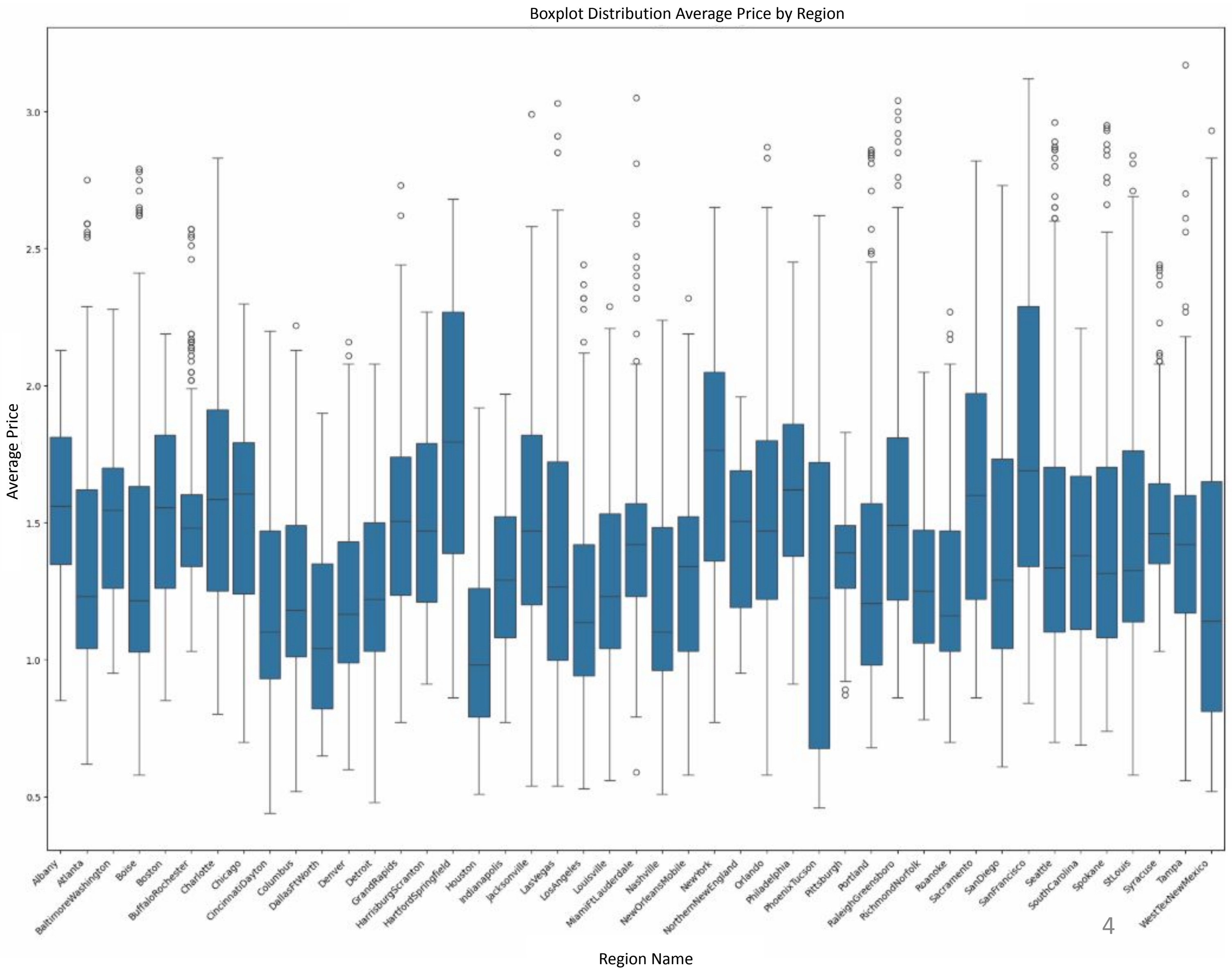
- We aligned on overall data prep approach
  - **Temporal Feature Engineering:** Created explicit YEAR, SEASON, and MONTH columns to facilitate granular time-series analysis and seasonal pattern detection
  - **Geographic Granularity:** Maintained CITY-level analysis over REGION-level to preserve statistical power and capture local market dynamics
  - **Data Quality Control:** Removed ~5,000 records (XL Avocados) due to data integrity issues
  - **Parallel Analysis Protocol:** To compare findings and identify potential methodological biases

- Project Estimates
  - This involved 2 Data Analyst-level roles for an estimated ~40 hours
  - Estimated Base project cost: $3740 ($47 hourly rate for Analyst role)
  - The benefit to the client is in the ***price elasticity insights***, which enables retailers to increase prices on avocados to increase their profit margins.
    - Because of this, we charged $12K for the project, which is small enough to get easily approved, but large enough to deliver a 3.20 ROI (to our company).

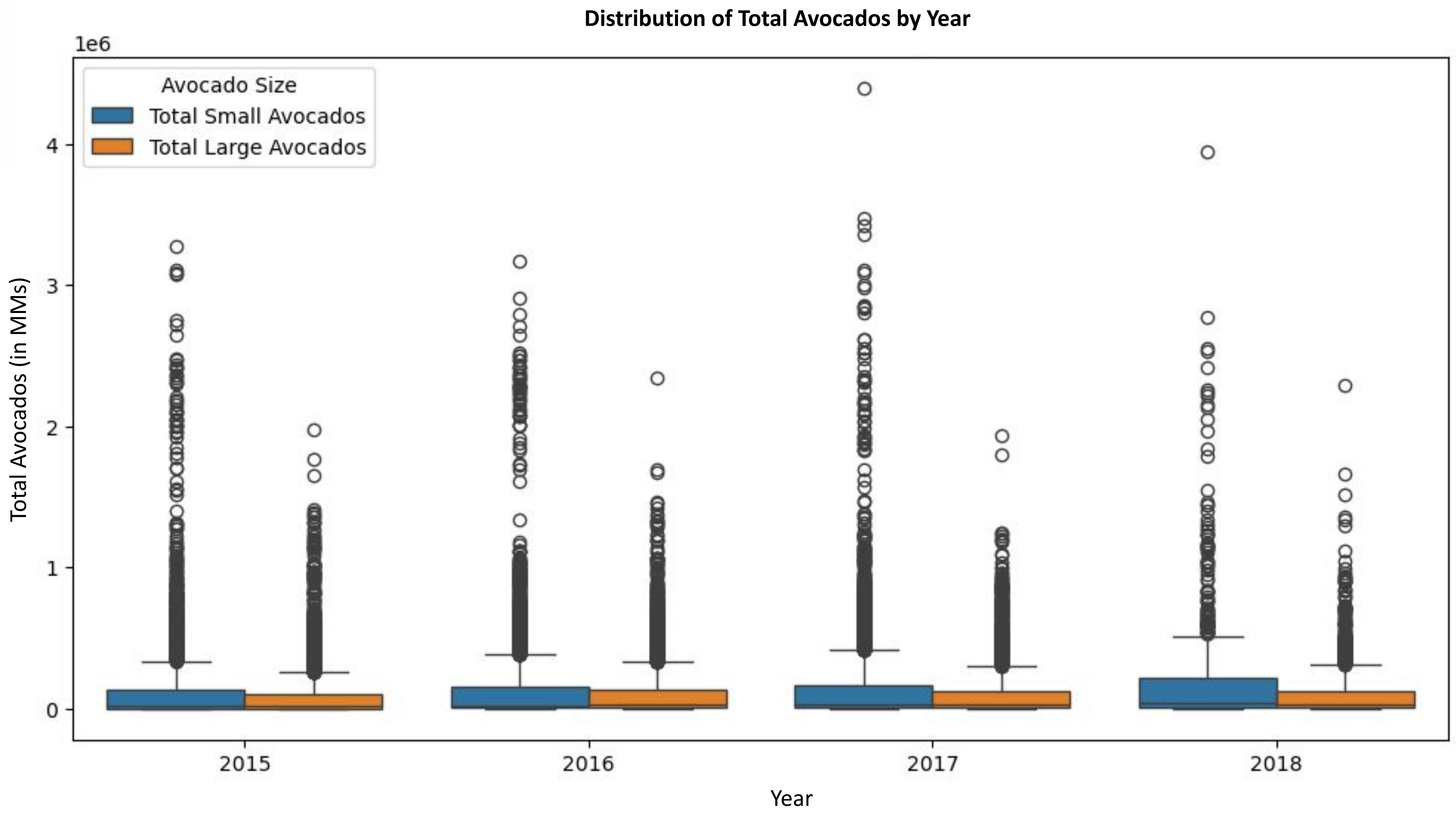


Over 10K years ago, Avocados were the favorite foods of Giant Sloths and Mammoths, who helped spread their seeds

Initial Visualization Differences



- Raymond's Box Plot evaluated Price of Avocados by City
- Kristin's Box Plot mapped Avocado Volume by Year
- Both visualizations show lots of outliers, but Kristin's shows the most!



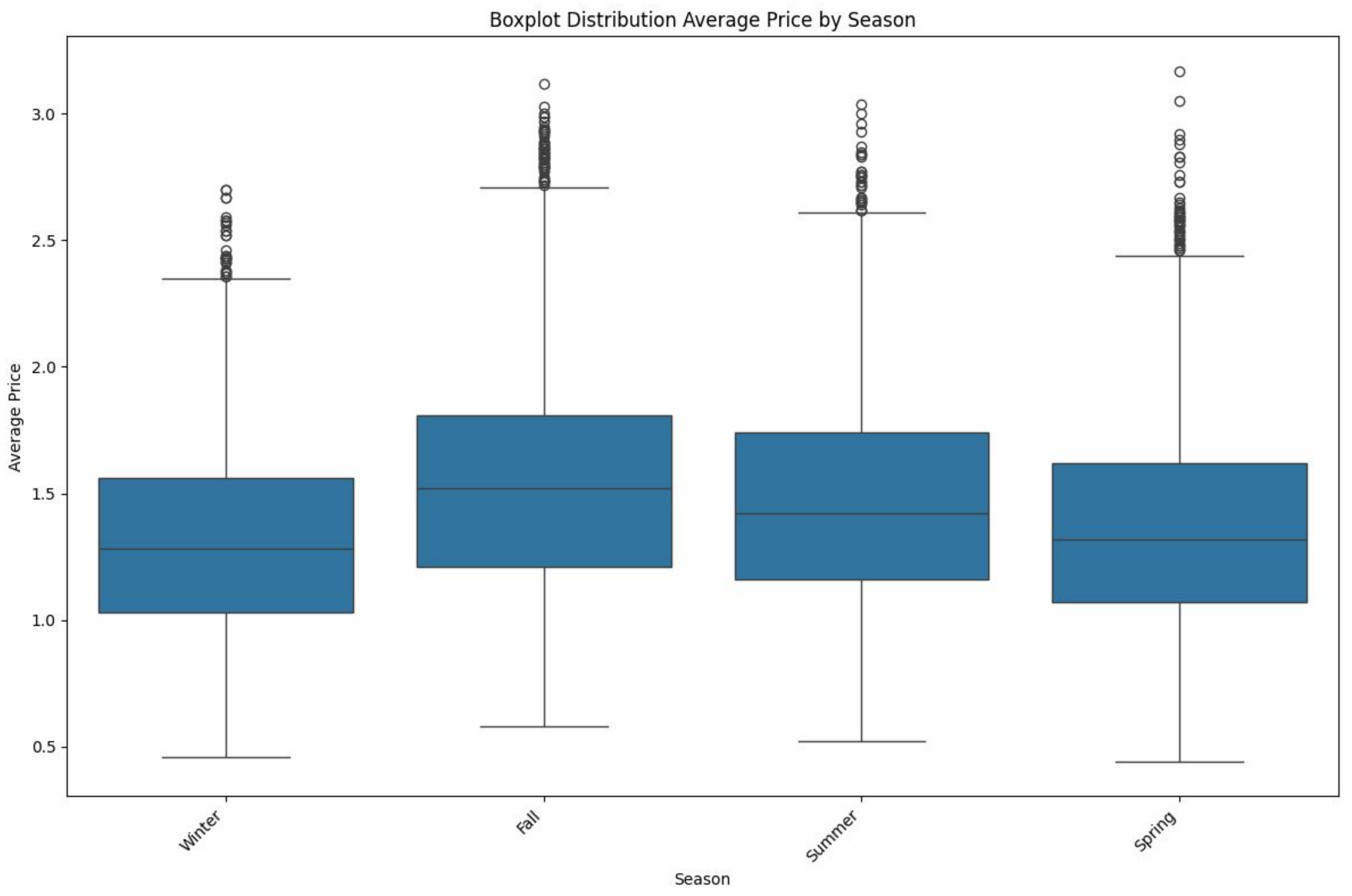
- The key takeaway from both charts is the similar: Avocados have a highly variable sales pattern, both per year and per region.

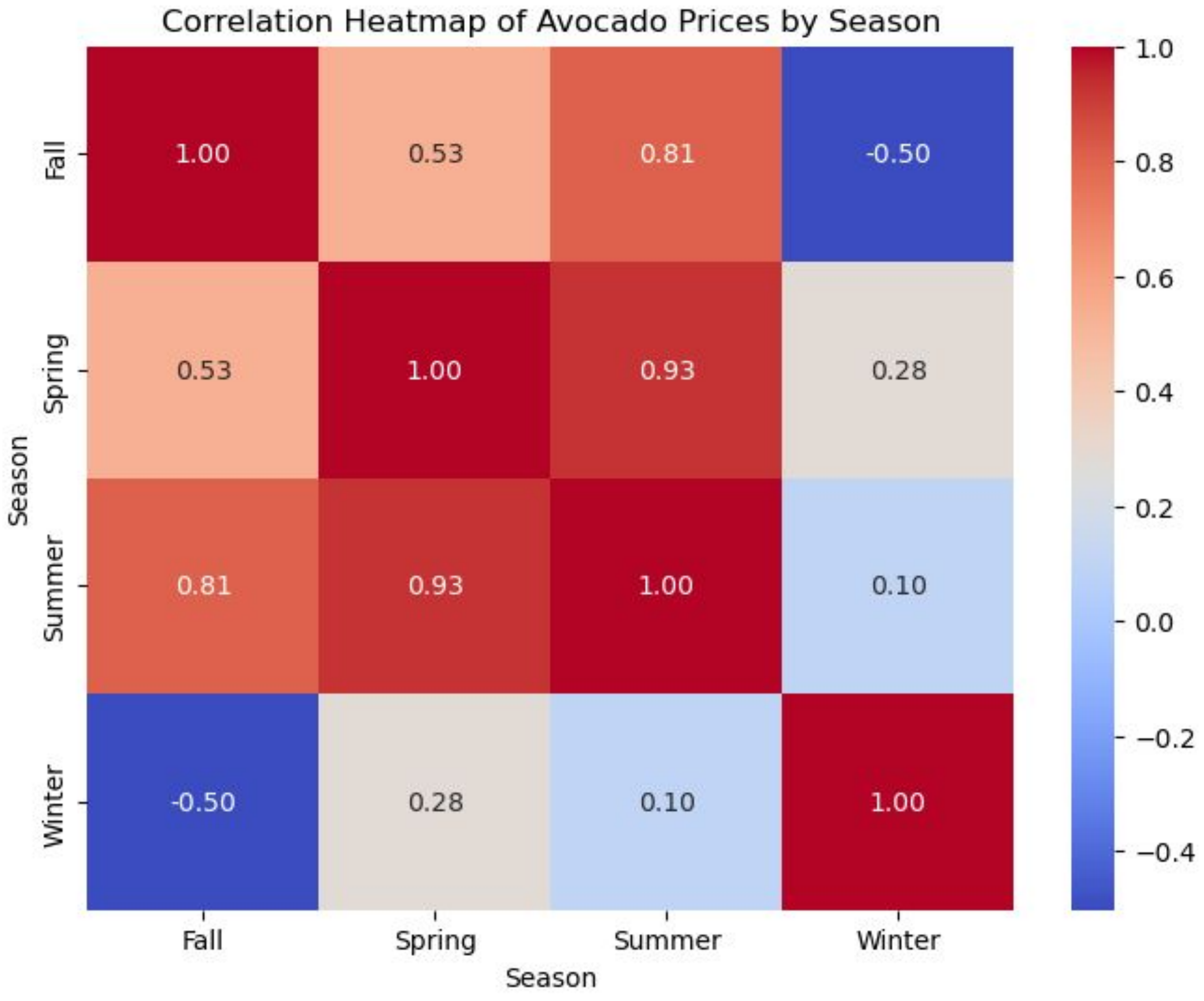
Avocados produce ethylene gas, which helps them ripen. Put them in a bag with apples and bananas to speed the ripening process!

Seasonality Analysis

- Raymond created another boxplot that looked at Avocado Price by Season
  - Fall has the highest median price
  - Winter has the lowest median price

- Kristin created a heatmap that correlated Avocado Prices by Season
  - Fall and Summer Prices are highly correlated (0.81); Spring closely correlates to Summer as well (0.93)
  - Winter does not have a strong correlation to any season
  - Fall and Winter have a negative correlation (-0.50), so Winter moves in the opposite direction to Fall
- Each chart points to similar takeaways:
  - If prices are high in Fall, they will likely also be high in Spring and Summer
  - Winter prices will move in an ***opposite*** pattern





Avocados were once called “Alligator Pears”

ML Model 1: Linear Regression

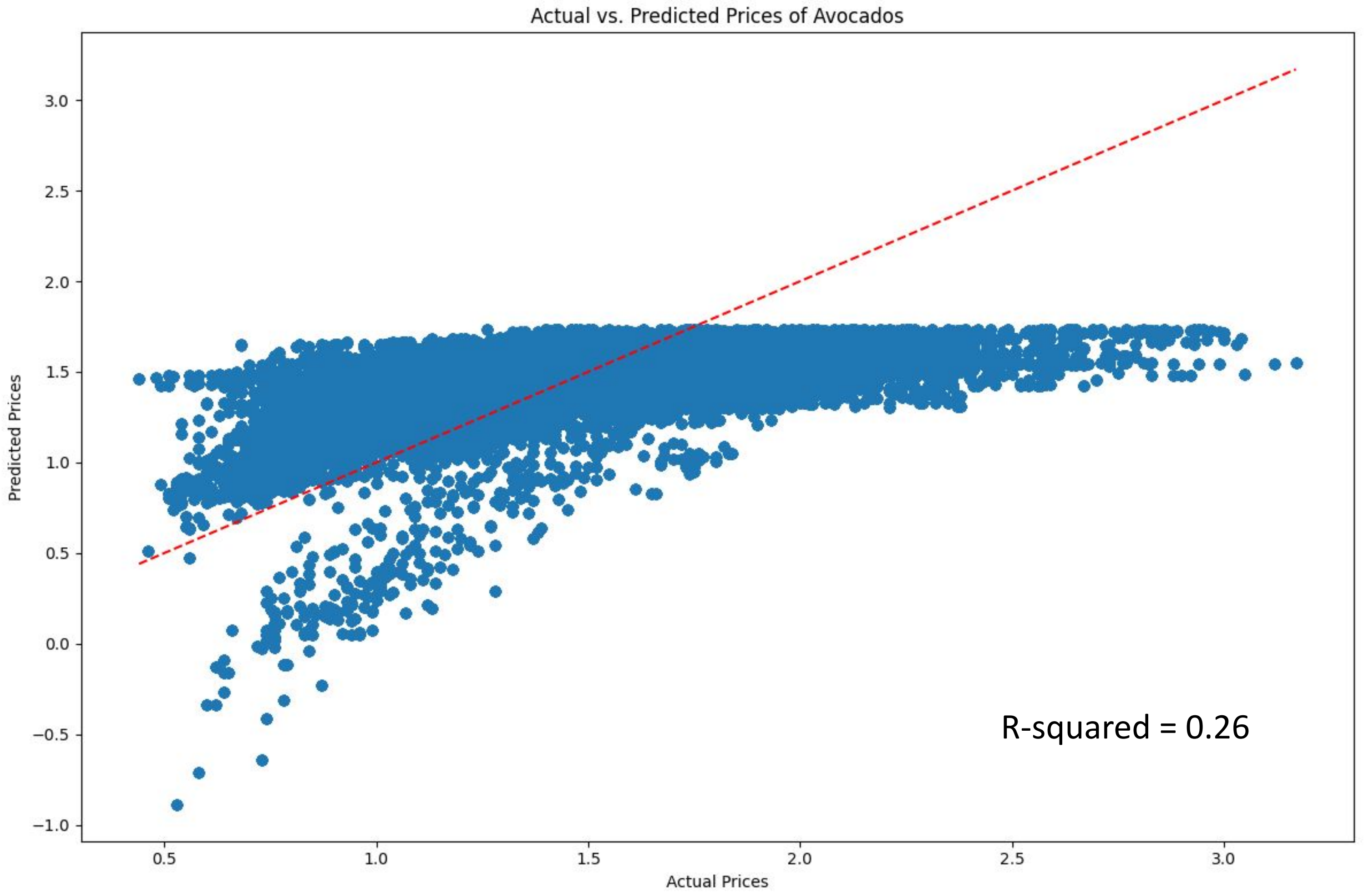
- Raymond decided to base his regression on Price, while Kristin decided to focus on Volume

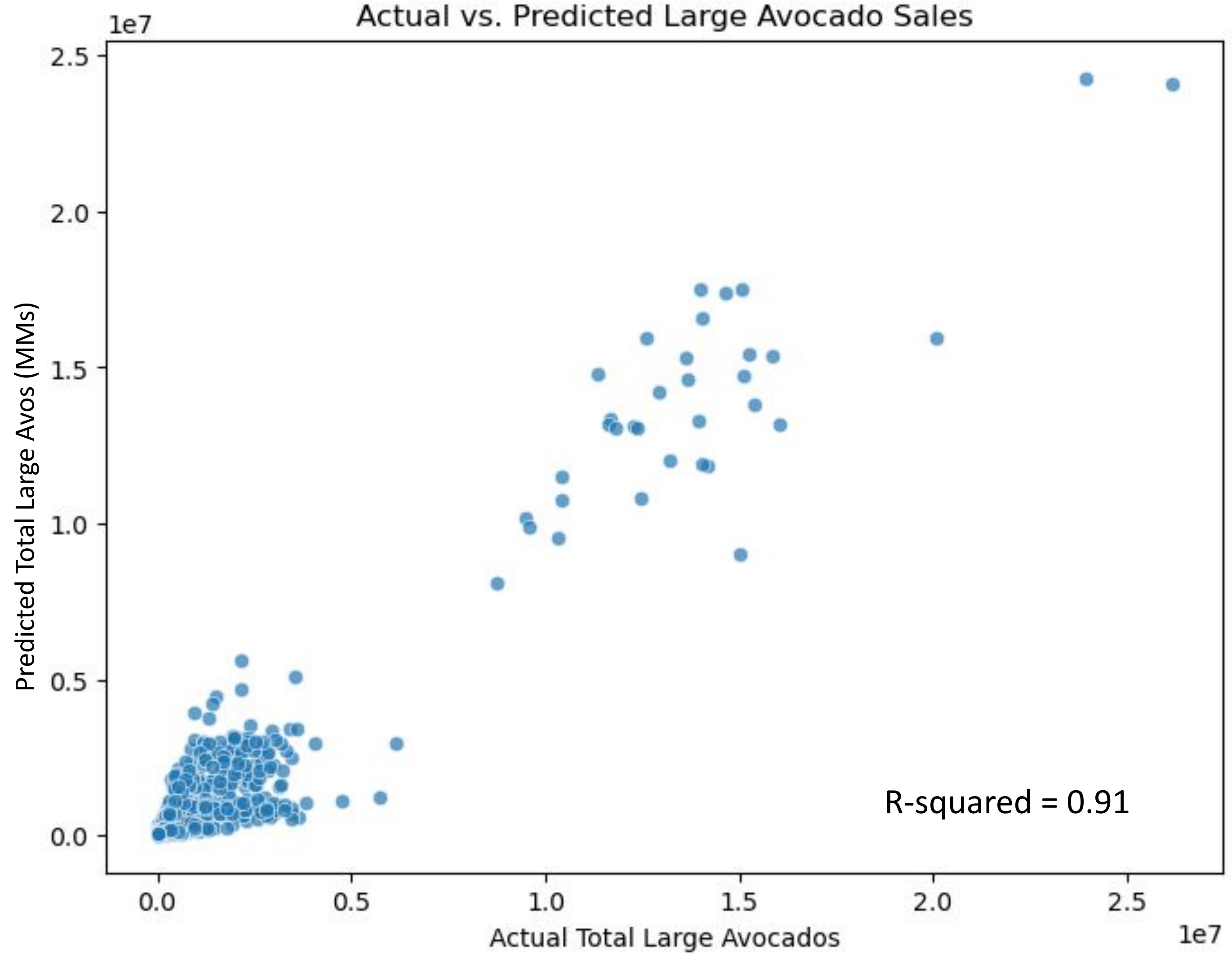
- In Raymond's model, see a high volume of datapoints being analyzed
  - OneHotEncoder 10x-d the number of columns in the dataset!
  - Introduced a lot of duplicate data into the model
  - Data is a poor fit to the the predicted price
  - Model runs the risk of underpredicting as prices get higher

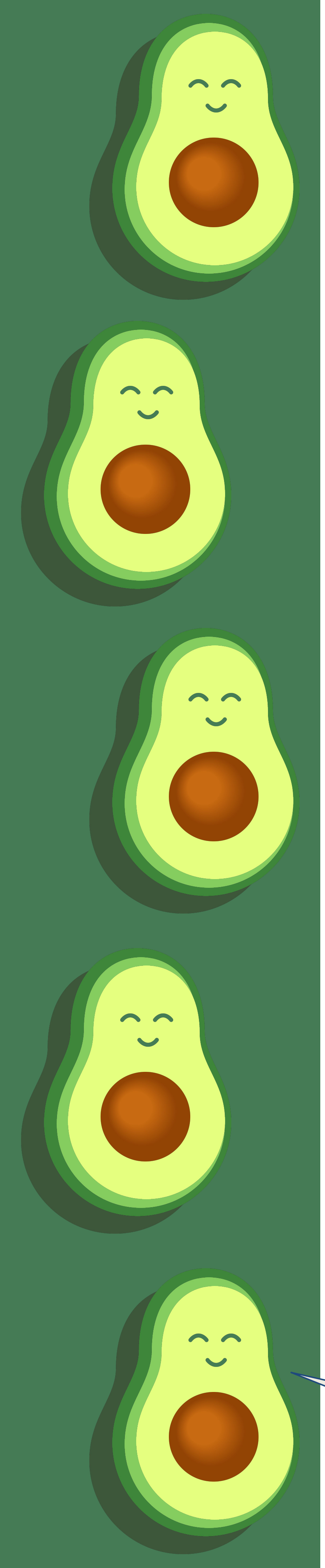
- In Kristin's model, the scatterplot has more of 45* angle shape
  - However, not all points fall closely within the prediction line, which also means that this model will struggle to predict higher prices
  - The model is overfitting - it works well for certain ranges, but struggles to generalize *across the entire range.*
    - You can see this in the clustering of the datapoints

6

Mexico produces over 40% of the world's Avocados







ML Model 1a: RandomForest Regressions

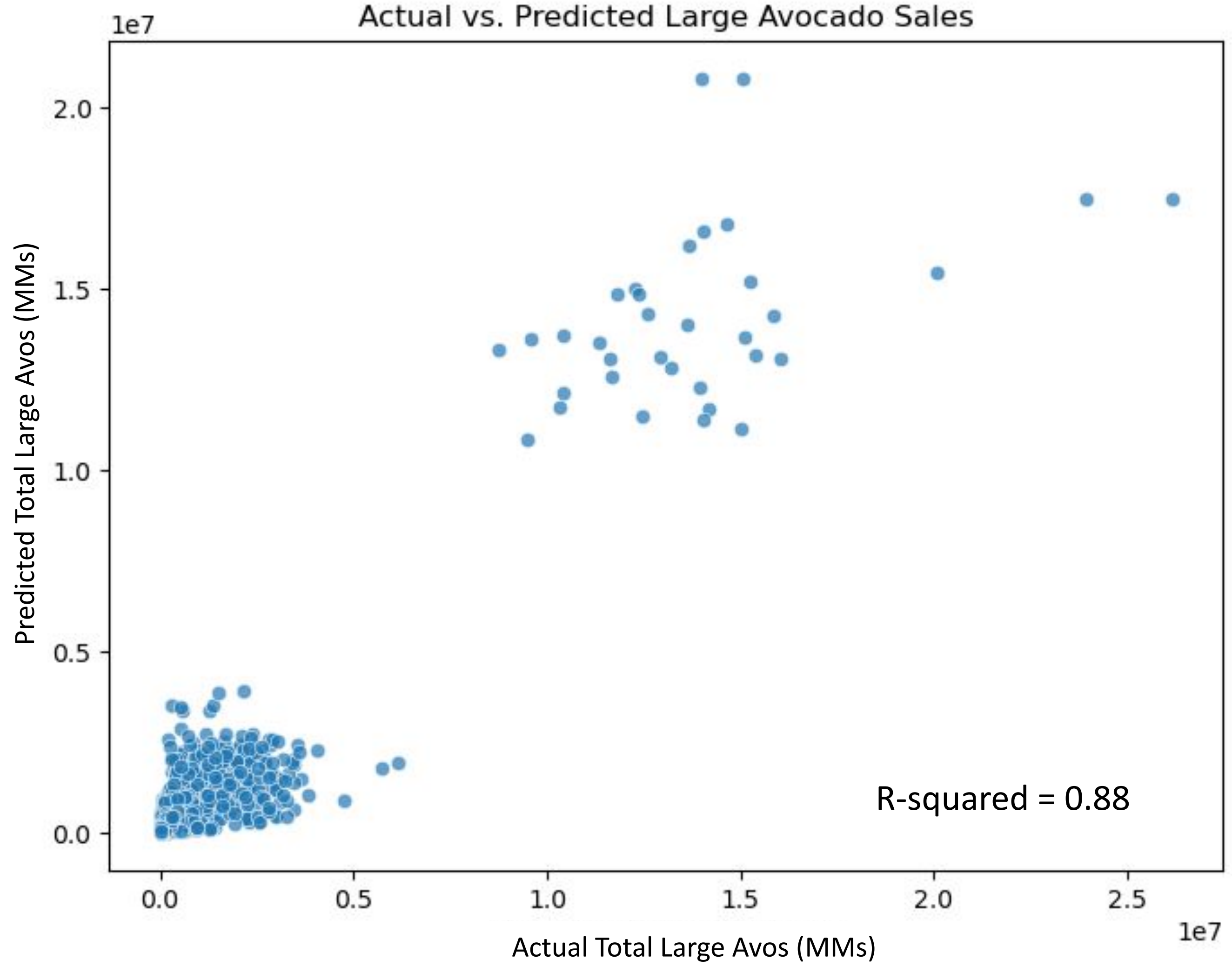
- Kristin then applied a RandomForest Regression in 2 ways:
  - First on the full dataset
  - Second, testing on full years within the dataset (2015-17) and training on the partial year (2018)

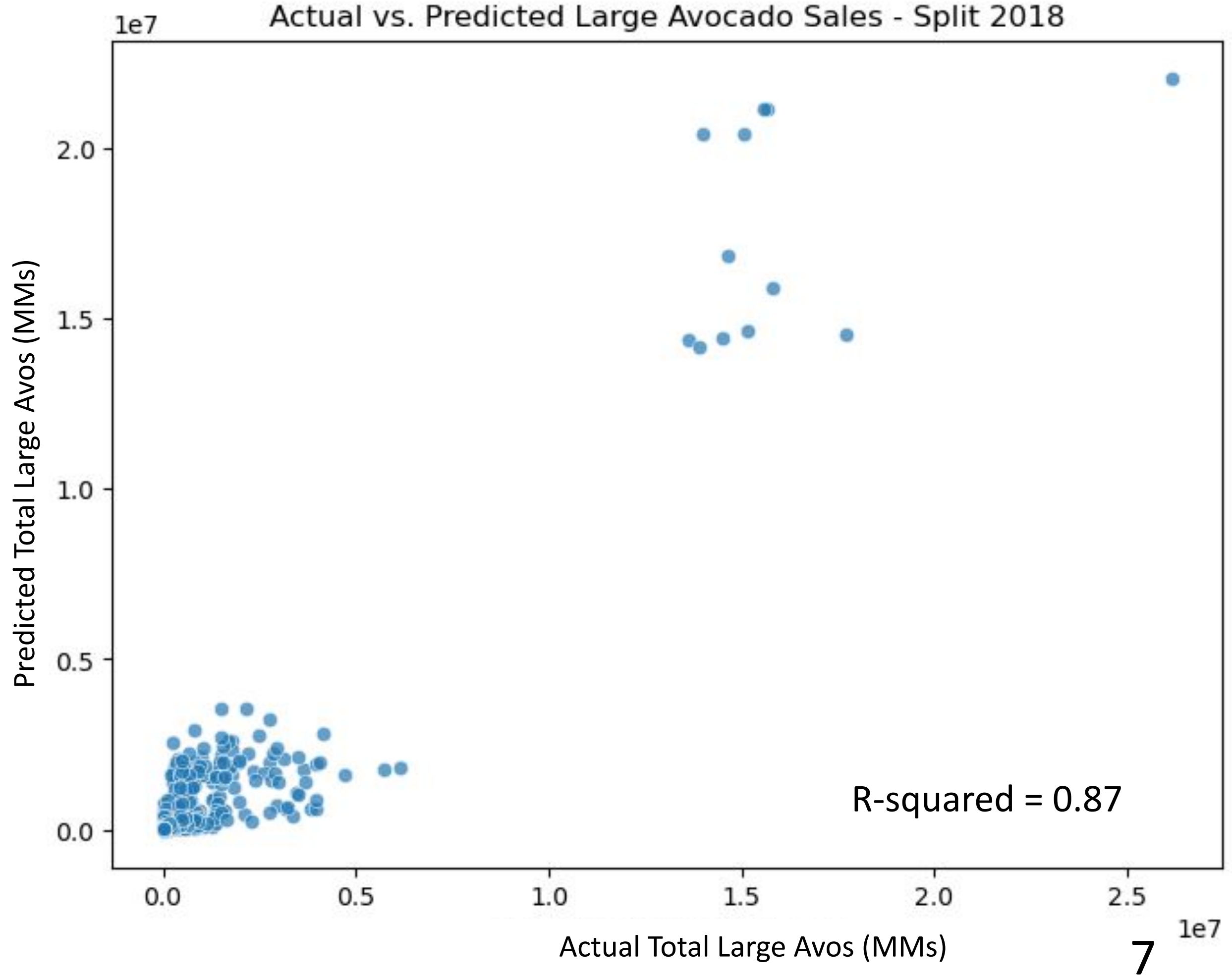
- Both models *roughly* predict on the 45* line
  - Although neither model has a very ‘tight’, linear shape

- You can see that both models suffer from underfitting
  - Lots of clustering, very few predictions as prices increase

- What could make these models stronger?
  - A much larger dataset (2015 - 2025 and/or Price by Type data)
  - Better data leakage controls (removing outliers, etc.)
  - Using the model to predict on Conventional Avocados (majority of the dataset)
  - Possibly using synthetic data to fill out the dataset

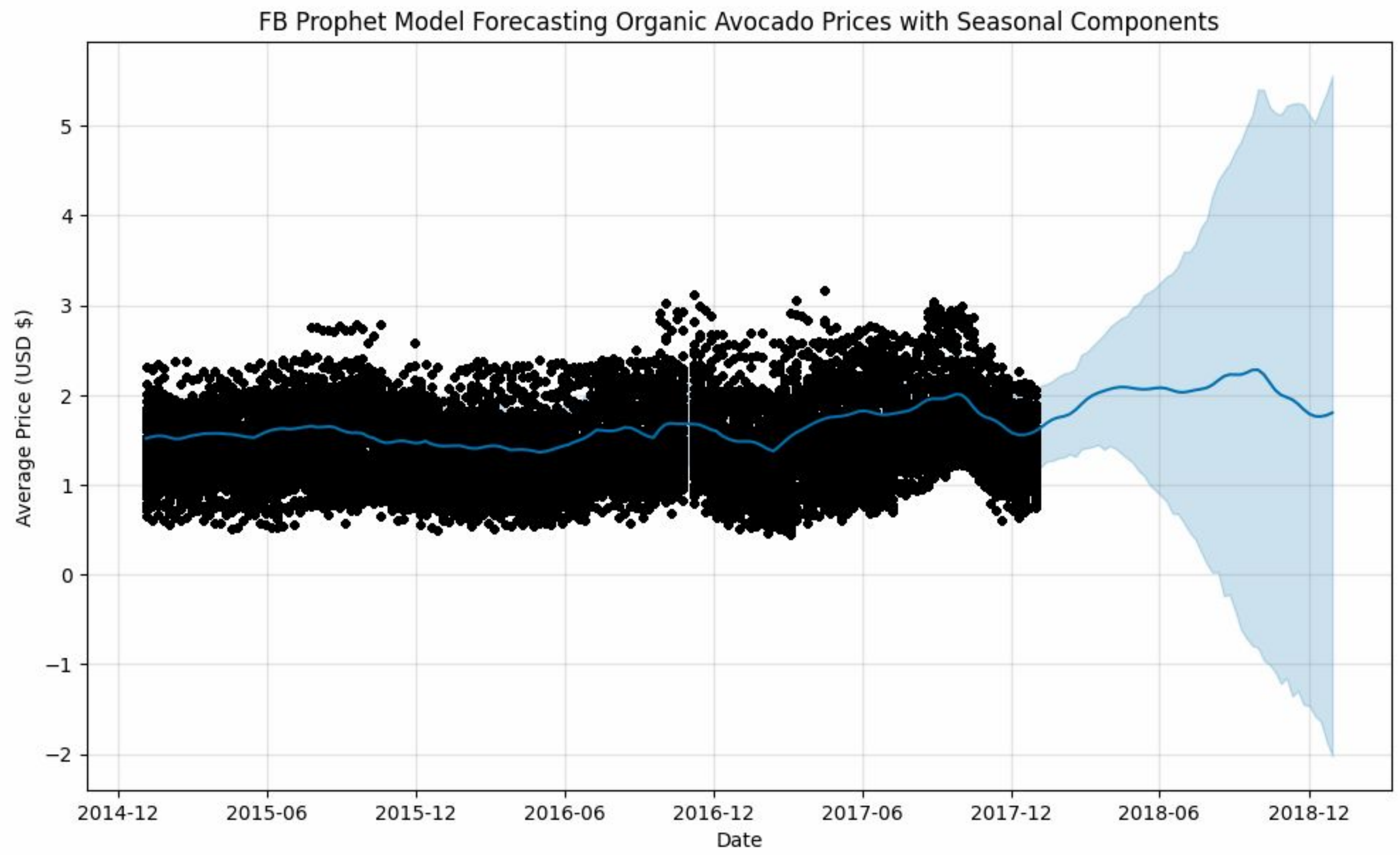
Avocado consumption has grown +400% since the early 2000s!

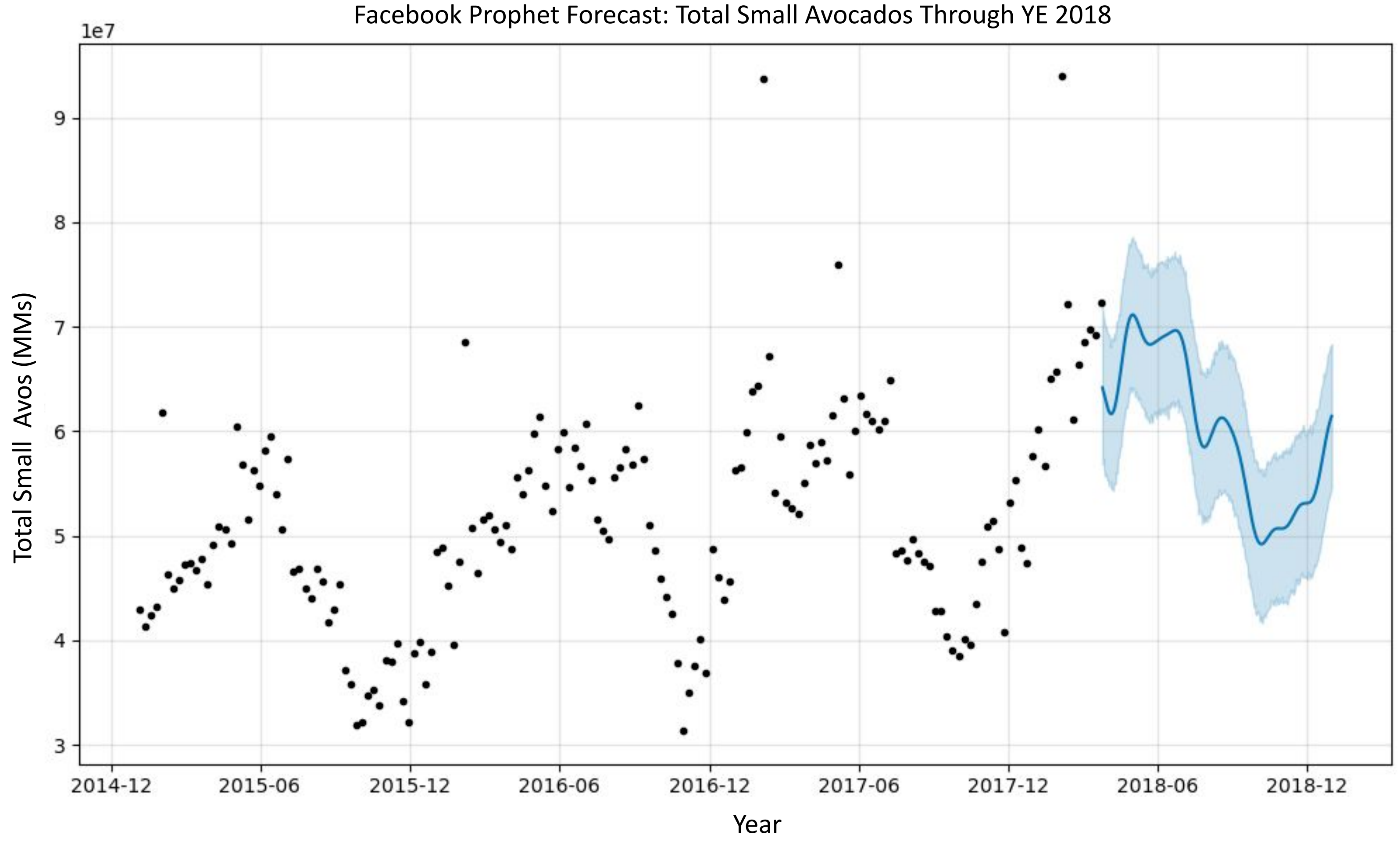




Model 2: Facebook Prophet

- Raymond decided to forecast Price, while Kristin focused on forecasting Small Avocado Volume
- Raymond's Facebook Prophet model was calculated on the 10x-d data
  - Resulting in a dense but ultimately imprecise forecast
  - You can see a high degree of variance as the date gets farther
  - It took over 10 minutes to run!
- Kristin's Facebook Prophet model had a tighter prediction line, with smaller variance
  - Some dramatic outliers!
- Some ways to improve:
  - Remove outliers and add Seasonality settings
  - Test how well the model generalizes instead of just train-test
  - Possibly using synthetic data to fill out the dataset





Avocados were discovered by accident by Rudolph Haas in the 1920s

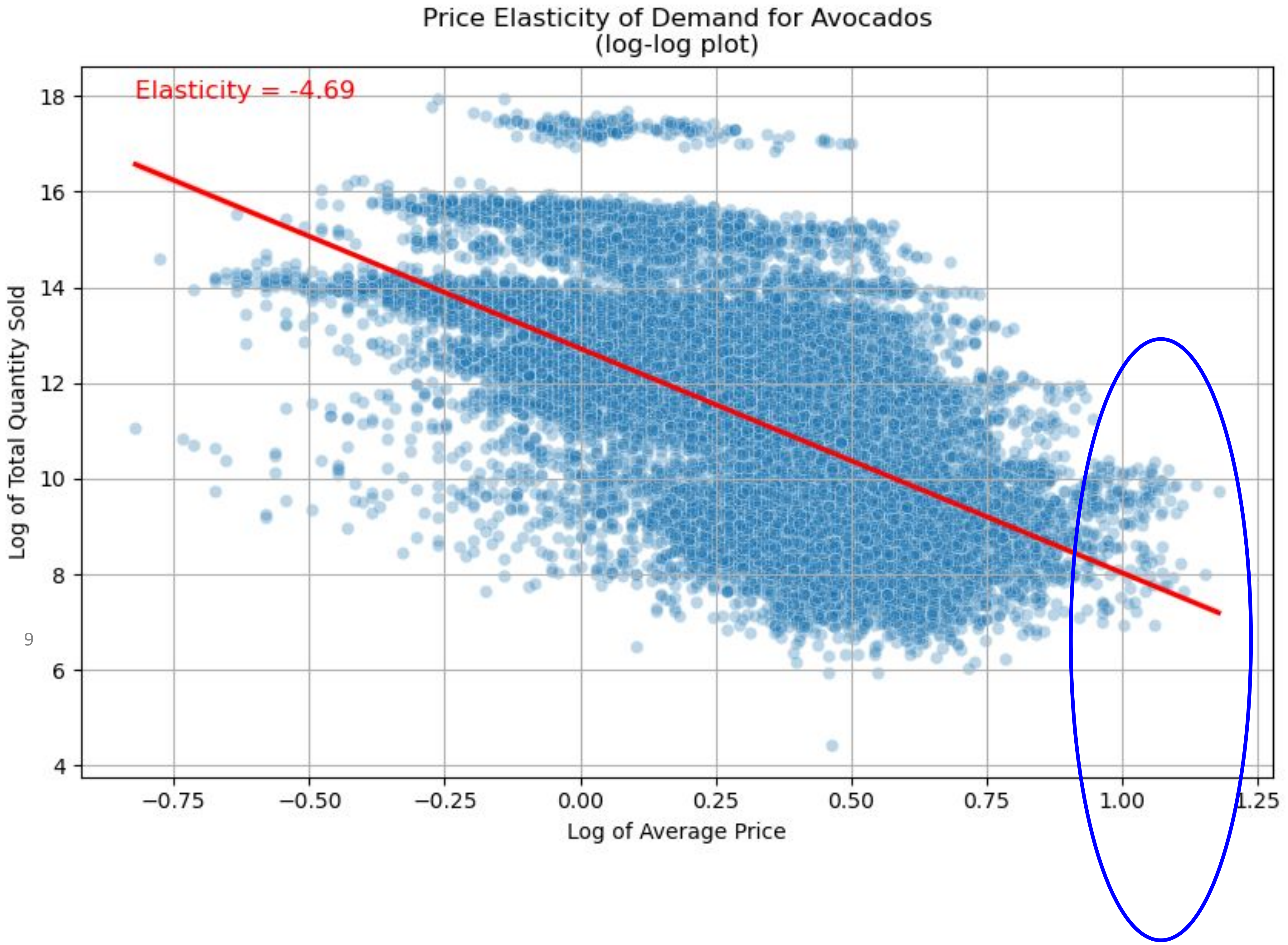
Model 3: Price Elasticity of Demand

- Elasticity determines the "price sensitivity", and can tell us a lot about consumer behavior

- We used a Log-Log regression to calculate the elasticity value

- Overall Price Elasticity of Demand for the period is -4.69
    - This means that for every 1% increase in price, there will be a corresponding 4.69% *decrease* in volume sold
    - You can see this on the red line in the chart

- Avocados are *highly elastic*, which means they are very price sensitive



People are generally not willing to pay >$0.90 for their avocados!

Avocados are loaded with 20 vitamins and minerals, including fiber, folate, and vitamin K

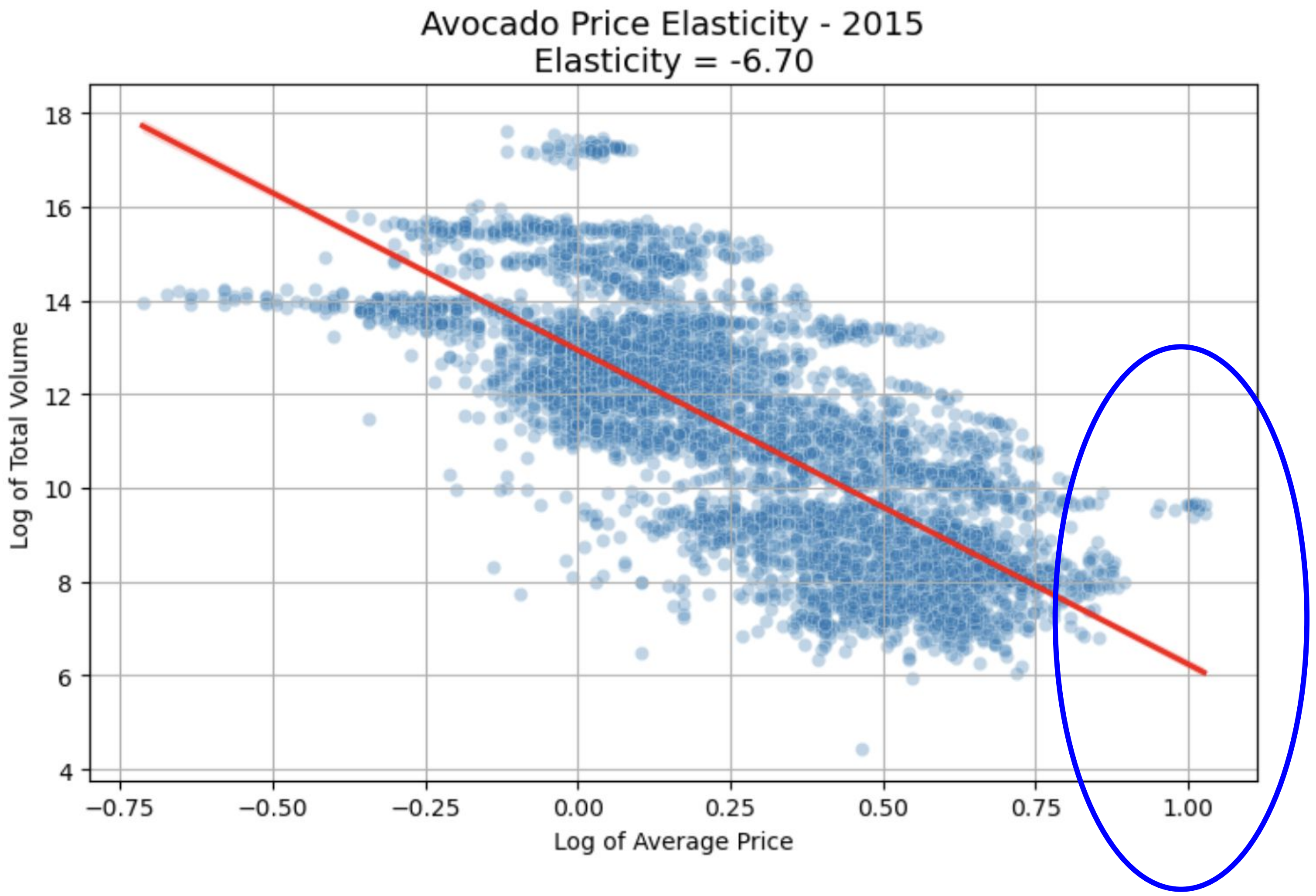
Model 3: Comparing Elasticities

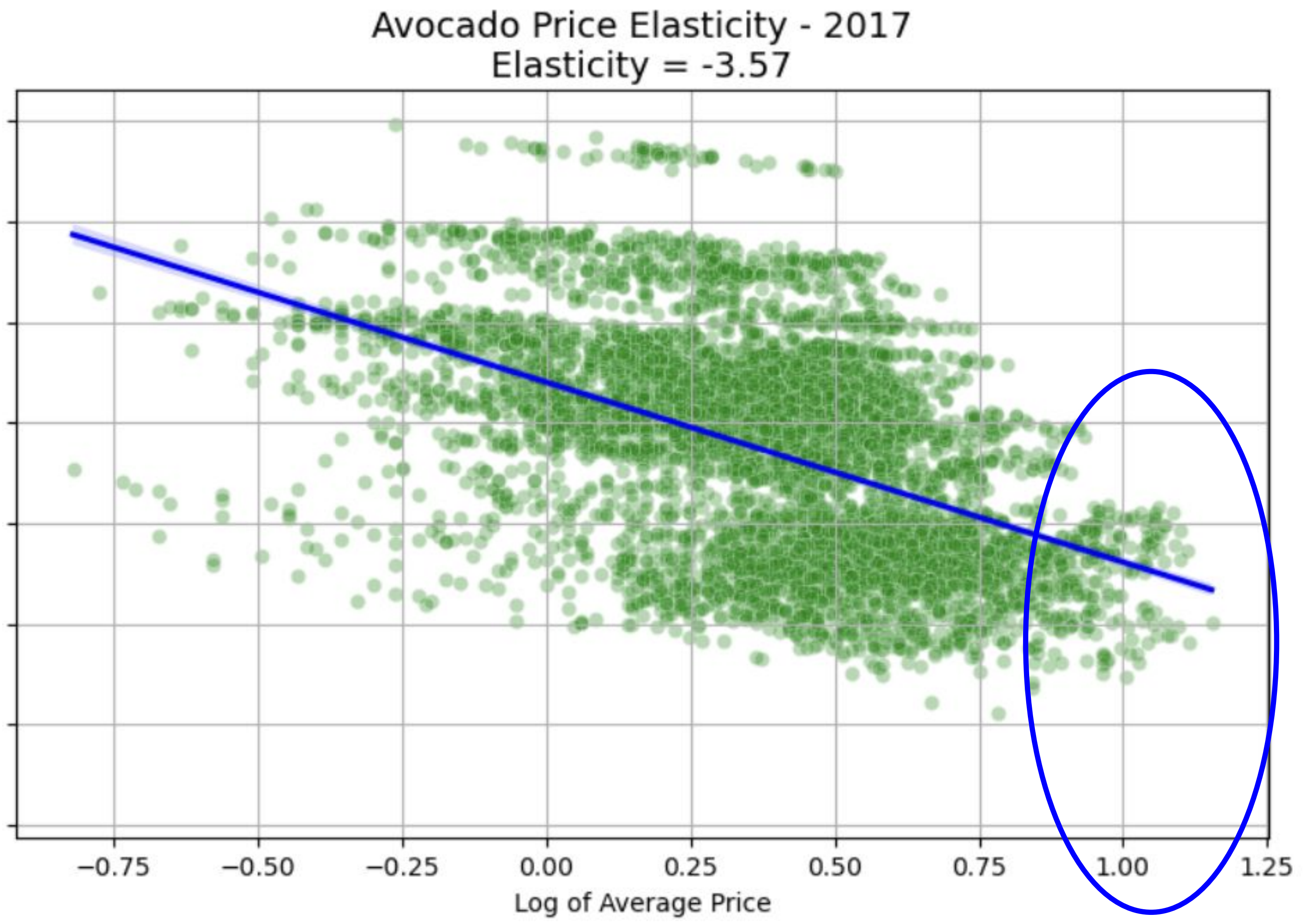
We analyzed the elasticities between two different years: 2015 and 2017

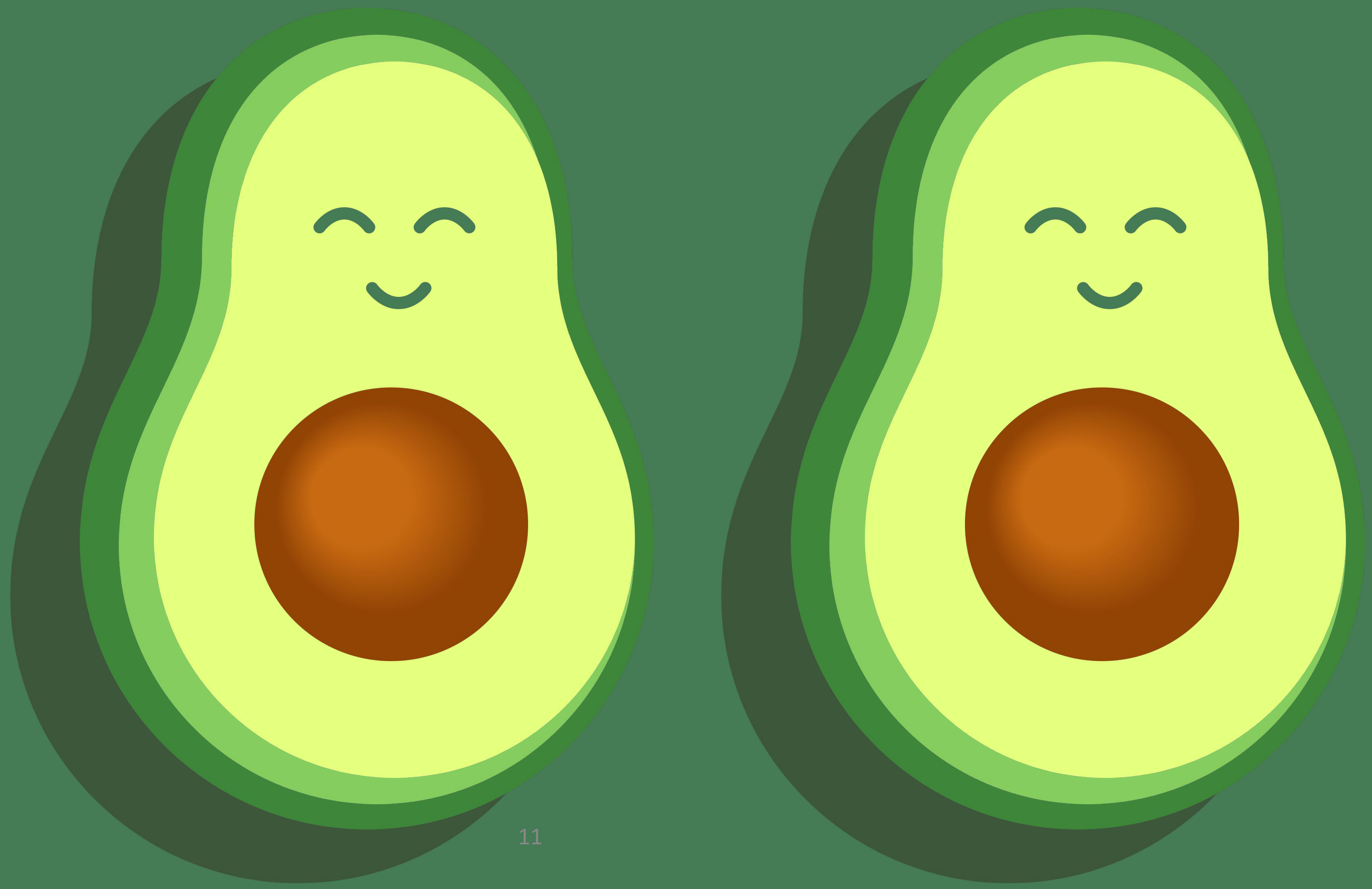
- Some interesting things happened in 2017:
  - ***Demand* increased:**
    - 2017 was the **peak** of the **avocado toast trend**, marked by frequent appearances in Insta/FB
    - China entered the market
  - **While *Supply* also decreased:**
    - Drought & poor weather severely impacted crop availability
    - Labor strikes in Mexico impacted production

- 2015 turned out to be a high elasticity year (-6.70)
  - Avocado consumption declined dramatically as the price went up
  - Volume really drops off once the price hits $0.75
- 2017 showed much lower elasticity (-3.57)
  - Because fewer Avocados were available, people were willing to pay more for them
  - More avocados were sold at >$0.50, and LOTS more avocados were sold at >$1.00!

The name 'Avocado' comes from the Aztec '*ahuacatl*', which means (ahem) *testicle.* Makes sense, because avocados grow in pairs!







THANK YOU!

KRISTIN PETERS AND RAYMOND STOVER