# CASE STUDY

# Avocado Market

Prophet Time Series

Author: Sameer Hate

Date: 27-05-2025

# Table of Contents

# Introduction

## General

The avocado has grown in popularity worldwide over the last decade, becoming a staple in both health-conscious diets and mainstream culinary trends. As a result, understanding and predicting avocado market behavior has become increasingly valuable for producers, retailers, and supply chain analysts. The price of avocados is influenced by various factors including seasonality, region, type (conventional or organic), and overall market demand.

In this project, we apply time series forecasting techniques to predict future avocado prices using historical sales data. Specifically, we employ Facebook Prophet, a robust and flexible forecasting tool developed to handle time series data with strong seasonal effects and multiple influencing components. By forecasting future price trends, this model can provide meaningful insights to support better planning and decision-making across the avocado supply chain.

The report outlines the process of data preparation, exploratory analysis, model building with Prophet, and evaluation of forecast accuracy, ultimately aiming to uncover patterns and trends that can enhance market understanding and pricing strategy.

## Author’s Note

This case study presented in this work has been independently completed by me. This project serves as a medium to deepen my understanding of Data Science concepts and enhance my practical skills in applying machine learning and data analysis techniques to real-world problems.It may be prone to errors but I try my best to eliminate them as much as I can. In case you identify any such error or have a more optimal technique to solve this problem, your feedback and suggestions are always welcome as this will eventually help me strengthen my concepts.

# Problem Statement & Expected Results

## Objective

The objective of this case study is to predict the future price of avocados sold in the United States by analyzing historical sales data. By leveraging time series forecasting techniques, specifically the Facebook Prophet model, this project aims to identify trends, seasonality, and other temporal patterns in avocado pricing.

## Goals

- To understand and implement the Facebook Prophet model to forecast and predict average avocado prices over time.
- To preprocess and structure the data in a format suitable for time series forecasting.
- To visualize trend, seasonality, and forecast components of avocado pricing.
- To generate actionable insights that can support decision-making in pricing, marketing, and supply chain planning.

## Expected Outcomes

The expected outcome of this project is the development of a reliable and interpretable time series forecasting model capable of predicting future avocado prices in the United States with reasonable accuracy. By leveraging Prophet's ability to capture trend and seasonality components, the model should provide valuable insights into price fluctuations over time, including anticipated increases or decreases in average prices.

# Dataset

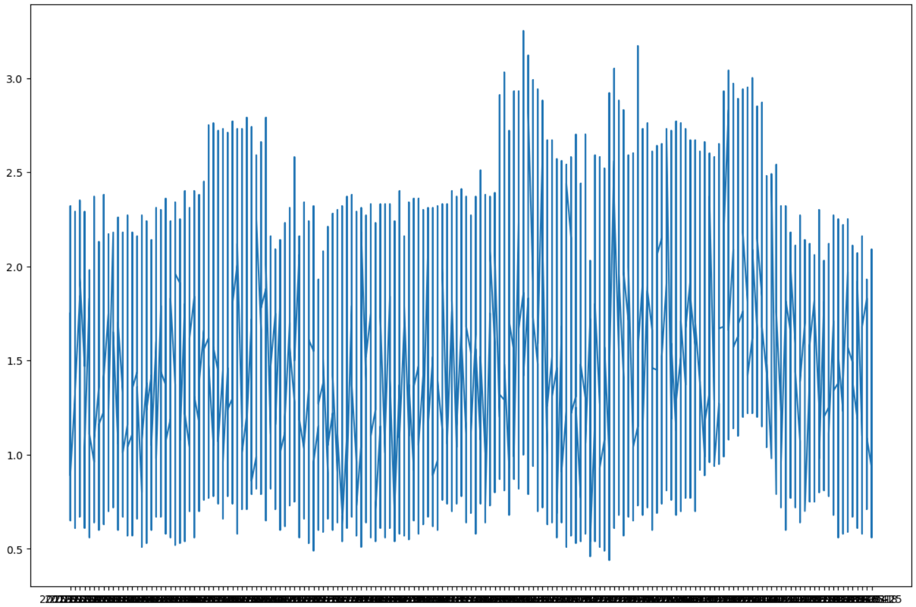
## Description

The dataset used in this project contains historical data on avocado sales and pricing in the United States, spanning multiple years. Each row in the dataset represents the weekly sales statistics for avocados in a specific region and product type. The key columns include:

- Date: The timestamp representing the week of the observation.
- AveragePrice: The average selling price of a single avocado.
- Total Volume: The total number of avocados sold.
- 4046, 4225, 4770: PLU (Price Look-Up) codes representing different avocado sizes or types.
- Total Bags, Small Bags, Large Bags, XLarge Bags: Quantity of avocados sold in different bag sizes.
- Type: Indicates whether the avocado was conventionally grown or organic.
- Year: The year in which the data was recorded.
- Region: The U.S. region where the sales occurred.

This rich dataset enables detailed time series analysis by region and product type, and serves as a robust foundation for modeling and forecasting avocado prices using Prophet.

Dataset: [Click Here]



*Average Price vs Date(Crude Representation)*

# Approach

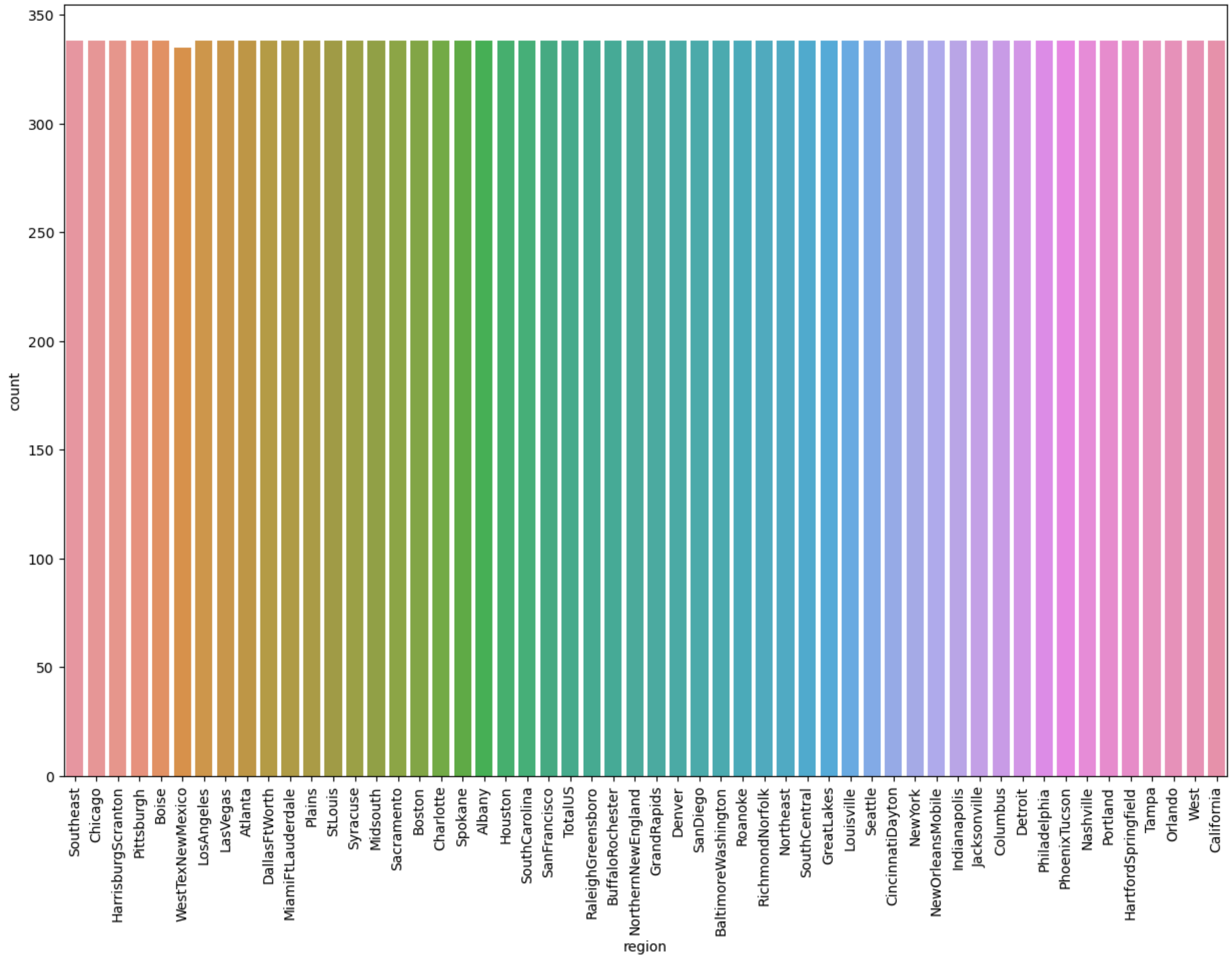
- #IMPORTING LIBRARIES : To implement the forecasting model and perform exploratory data analysis, a set of essential Python libraries were imported at the beginning of the project. Pandas and NumPy were used for data manipulation, cleaning, and numerical operations. Matplotlib and Seaborn facilitated visualizations to better understand the trends, seasonality, and regional variations within the avocado dataset. The random module was included for any randomized operations required during analysis or testing. Most importantly, the Prophet library, developed by Facebook, was imported to build the core time series forecasting model. Prophet is particularly well-suited for capturing seasonality, trend changes, and holiday effects in time-stamped data, making it ideal for modeling avocado price behavior over time.

- #LOADING THE DATASET : The avocado market dataset was loaded into the environment using Pandas' read_csv() function, which reads the data from a CSV file into a structured DataFrame. This operation allows for efficient manipulation and analysis of the data using Python's data science tools. To confirm successful loading and understand the structure of the dataset, the first 17 rows were displayed using the head() function. This preview provided an initial look at the dataset's columns, including average price, sales volume, PLU codes, bag sizes, type (conventional or organic), region, and date—all of which are critical for forecasting avocado prices over time.

- #SORTING THE DATASET ACCORDING TO THE DATE : To prepare the data for time series forecasting, the dataset was sorted in chronological order based on the 'Date' column using the sort_values() function. This step ensures that all subsequent analyses and model training processes follow the natural temporal sequence of events, which is crucial for accurate forecasting. After sorting, the first 10 records were displayed to confirm the correct order of the dates, laying the foundation for clean and structured time series modeling.

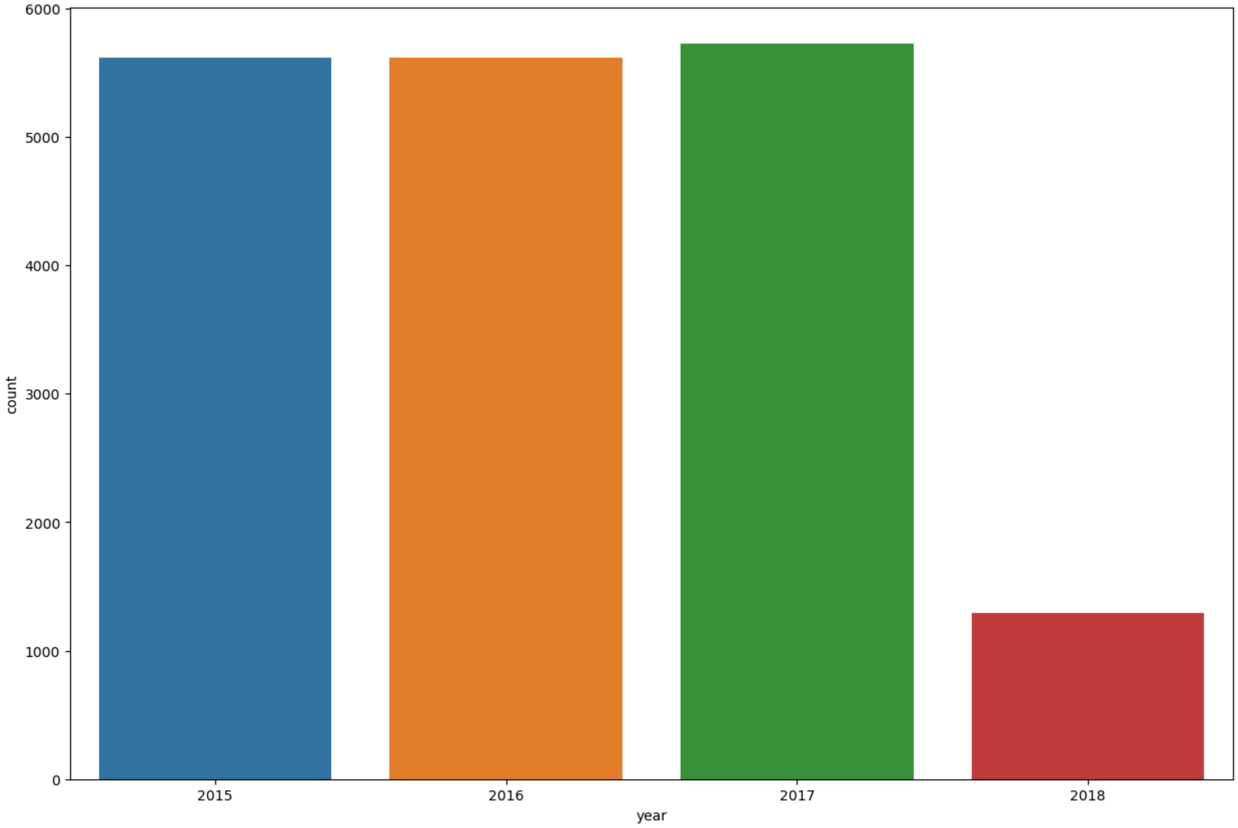
- #VISUALIZING THE DATASET - DATE VS AVERAGE PRICE : To gain an initial understanding of how avocado prices have changed over time, a line plot was created using Matplotlib with 'Date' on the x-axis and 'AveragePrice' on the y-axis. The figure size was adjusted for clarity using figsize=(15, 10), allowing for a more detailed view of the time series. This visualization helps reveal important trends, fluctuations, and potential seasonality in avocado pricing, which are essential components for effective time series forecasting. The plot serves as a foundational step in exploring temporal dynamics prior to applying the Prophet model.

- #VISUALIZING THE DATASET - COUNT OF AVOCADOS PER REGION : To explore the geographical distribution of avocado sales, a count plot was generated using Seaborn with ‘region’ on the x-axis. This visualization displays the number of records (weeks of data) available for each U.S. region in the dataset, providing insights into regional data coverage and representation. The figure was expanded using figsize=(15,10) for better readability, and the x-axis labels were rotated 90 degrees to prevent overlap. This analysis helps identify regions with more extensive data, which may lead to more reliable forecasts, and also reveals any potential data imbalance across regions.



*Count of Avocados per region*

- #VISUALIZING THE DATASET - COUNT OF AVOCADOS BATCHES PER YEAR : To understand the temporal distribution of the dataset, a count plot was created for the 'year' column using Seaborn. This visualization shows the number of weekly avocado sales records available for each year in the dataset. The figure size was set to (15,10) to ensure clear visibility of the data. This step helps in identifying whether the data is evenly distributed across years or if certain years have more or fewer entries, which is important for ensuring consistency in training and validating the time series forecasting model.



*Count of Avocado Batches per Year*

- #CLEANING THE DATASET : To prepare the data for use with the Prophet forecasting model, the dataset was filtered to include only the two essential columns: 'Date' and 'AveragePrice'. These were extracted into a new DataFrame named avocado_prophet_df, which serves as the primary input for Prophet. This step simplifies the dataset and ensures it aligns with Prophet's required format, where the date column (to be renamed ds) serves as the time index and the target column (to be renamed y) represents the value to forecast. Displaying the first 17 rows of the new DataFrame confirmed the successful extraction of the relevant fields.

- #RENAMING THE COLUMNS : To align the dataset with Prophet’s input requirements, the columns were renamed appropriately: the ‘Date’ column was renamed to ‘ds’ (denoting datestamp), and the ‘AveragePrice’ column was renamed to ‘y’ (the target variable to be forecasted). This naming convention is mandatory for Prophet to interpret the input data correctly. After renaming, the updated DataFrame was previewed to confirm the structure, marking the dataset ready for time series modeling.

- #FITTING THE PROPHET MODEL : With the dataset prepared and properly formatted, the next step involved initializing and fitting the Prophet time series forecasting model. A new instance of the Prophet model was created using its default parameters, which are well-suited for capturing trend and yearly seasonality in time series data. The model was then trained on the cleaned avocado pricing dataset (avocado_prophet_df) using the .fit() method. This step enables Prophet to learn from historical price patterns and underlying temporal structures, forming the foundation for generating accurate future forecasts.

- #MAKING A PREDICTION : After fitting the Prophet model on historical avocado prices, the next step was to generate future forecasts. This was achieved by using the make_future_dataframe() method, which extends the date range by 365 days beyond the original dataset, allowing the model to forecast for a full year into the future. The resulting future DataFrame was then passed into the predict() method, which returned a comprehensive forecast including the predicted average price (yhat), as well as the lower and upper uncertainty intervals (yhat_lower and yhat_upper). Displaying the first 17 rows of the forecast confirmed the structure and content of the prediction results.
    - ds : The date corresponding to each prediction (stands for datestamp)
    - yhat : The predicted value of the target variable (in your case, avocado price). This is the main forecast.This is the main prediction — the value Prophet thinks will happen (like the predicted avocado price).
    - yhat_lower : The lower bound of the uncertainty interval (usually 80% by default).This is the lowest value Prophet thinks might happen — it’s like saying “the price might go as low as this.”
    - yhat_upper : The upper bound of the uncertainty interval.This is the highest value Prophet thinks might happen — it’s like saying “the price might go up to this.”
    - trend : The non-periodic part of the forecast — the overall trend learned from the data, if its going up or down.
    - trend_lower : Lower bound of the trend’s uncertainty interval.
    - trend_upper : Upper bound of the trend’s uncertainty interval.
    - additive_terms : The sum of all seasonality and holiday components

- #VISUALIZING THE PREDICTION : To visually interpret the forecasting performance of the Prophet model, the plot() function was used to generate a time series graph displaying the predicted values alongside the historical data.
  - The black dots represent the actual observed data points (historical average prices).
  - The blue line shows the Prophet model’s predicted trend over time (yhat).
  - The shaded light blue area around the forecast indicates the uncertainty intervals (by default, 80% confidence interval), which reflect the model’s confidence in its predictions.

- #EVALUATING THE MODEL : To assess the performance of the Prophet time series forecasting model, a set of standard regression evaluation metrics was calculated by comparing the predicted values against the actual observed prices. The actual and predicted data were merged based on the common 'ds' (date) column. The evaluation focused on three key metrics:
  - Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values, providing an easily interpretable measure of prediction error in the same unit as the target variable.
  - Root Mean Squared Error (RMSE): Emphasizes larger errors by squaring the residuals before averaging, offering a more sensitive measure for evaluating model performance.
  - $R^2$ Score (Coefficient of Determination): Indicates the proportion of variance in the actual data that is explained by the model’s predictions. A higher $R^2$ value signifies better predictive power.

These metrics collectively help evaluate how accurately the model is forecasting avocado prices over time and guide further tuning or model comparison.

- #PLOTTING THE COMPONENTS : To gain a deeper understanding of the underlying patterns captured by the Prophet model, the plot_components() function was used to visualize the key components of the forecast.
  - Trend: Displays the long-term movement in avocado prices over time. This helps identify any consistent upward or downward shifts in the data.
  - Yearly Seasonality : Highlights recurring annual fluctuations in the data, often influenced by market cycles, harvesting seasons, or consumer behavior.

These plots allow for visual validation of the model’s ability to detect and represent temporal patterns, aiding in interpretability and further refinement of the forecasting strategy.

- #VISUALIZING THE ACTUAL VS PREDICTED PRICES : To evaluate the Prophet model's predictive performance visually, a line plot was generated comparing the actual average avocado prices to the predicted values over the same date range.
  - The black line represents the actual observed prices from the dataset.
  - The blue line represents the prices predicted by the Prophet model (yhat values).

  This visualization is critical for identifying how well the model captures the temporal structure of the data. It highlights any discrepancies between actual and forecasted values, allowing for qualitative analysis of underfitting, overfitting, or lagging predictions. The alignment of the two curves over time is a strong indicator of the model's effectiveness.

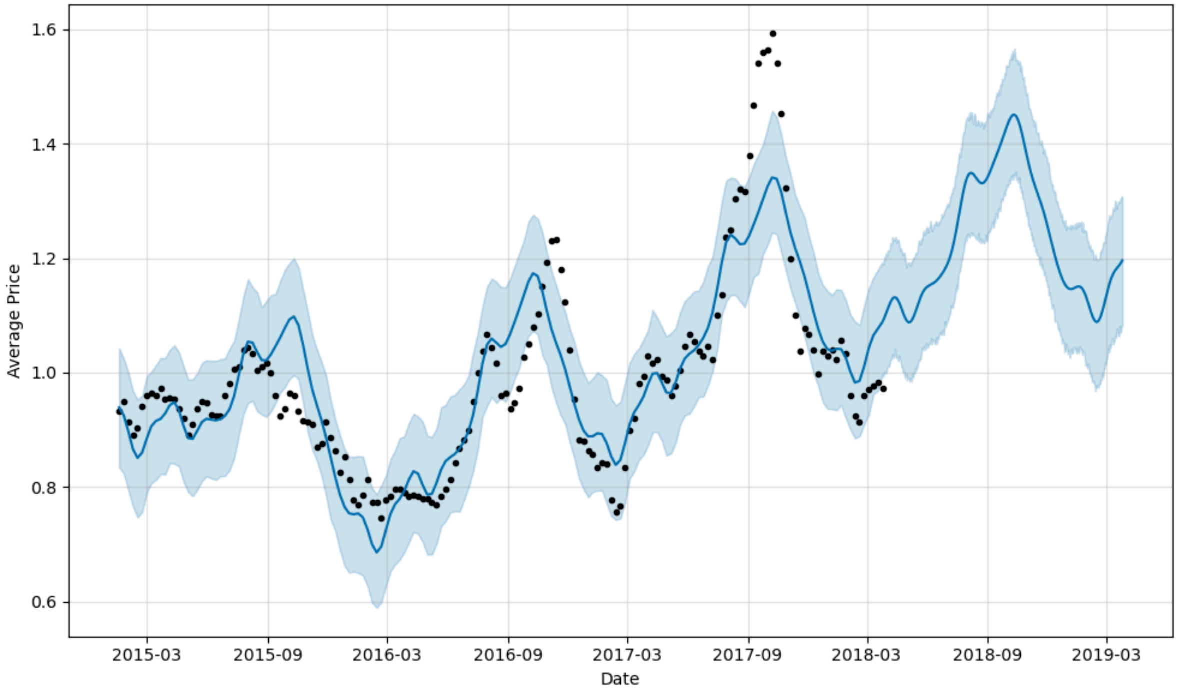
# Results

## Overall

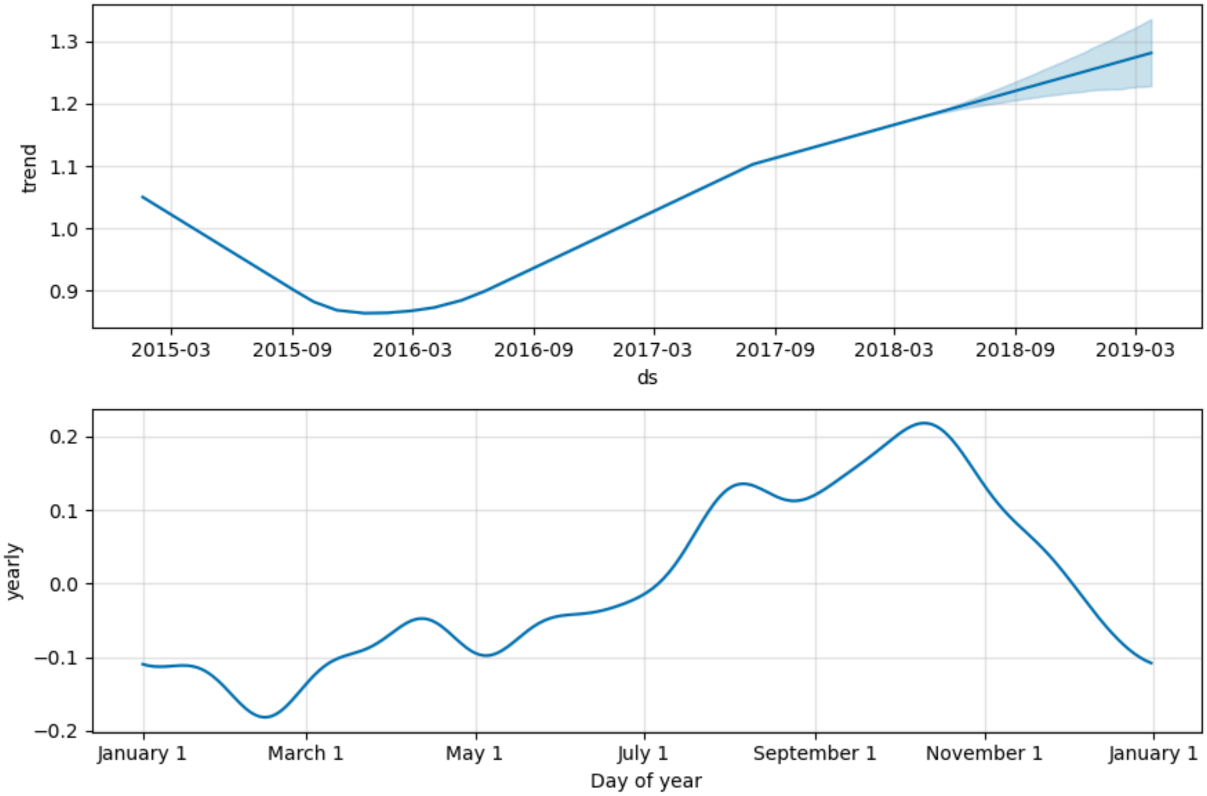
The Prophet model achieved strong predictive performance on the filtered dataset, with a Mean Absolute Error (MAE) of **0.0569**, a Root Mean Squared Error (RMSE) of **0.0783**, and an $R^2$ Score of **0.7861**. These metrics indicate that the model was able to explain approximately **78.6%** of the variance in average avocado prices within the selected test period, while maintaining a relatively low average error.

- $R^2$ Score: 0.7861 (78.61%) : The model successfully explains 78.61% of the variance in avocado price fluctuations, indicating that approximately four-fifths of price movements can be accurately predicted using the selected features. This represents a robust predictive capability, with only 21.39% of price variation attributed to external factors or market noise not captured by the model.
- Mean Absolute Error (MAE): $0.0569, On average, the model's price predictions deviate from actual prices by approximately 5.7 cents. This low error margin demonstrates high precision in forecasting, making the model suitable for operational planning and inventory management decisions.
- Root Mean Squared Error (RMSE): $0.0783, The RMSE of 7.8 cents indicates that the model handles both typical and outlier predictions well. The relatively small difference between MAE and RMSE (2.1 cents) suggests consistent prediction accuracy without significant bias toward large errors.

To enhance model accuracy and stability, the data was deliberately filtered to include only the ‘West’ region and ‘conventional’ avocados. This selective approach reduces variability introduced by regional differences and organic vs. conventional market segments. By narrowing the scope, the model was able to learn more coherent temporal patterns and produce more reliable forecasts. Using the entire dataset without such filtering would have introduced greater heterogeneity in pricing behaviors, seasonal trends, and volumes, potentially diluting the model’s ability to generalize effectively across such varied conditions.



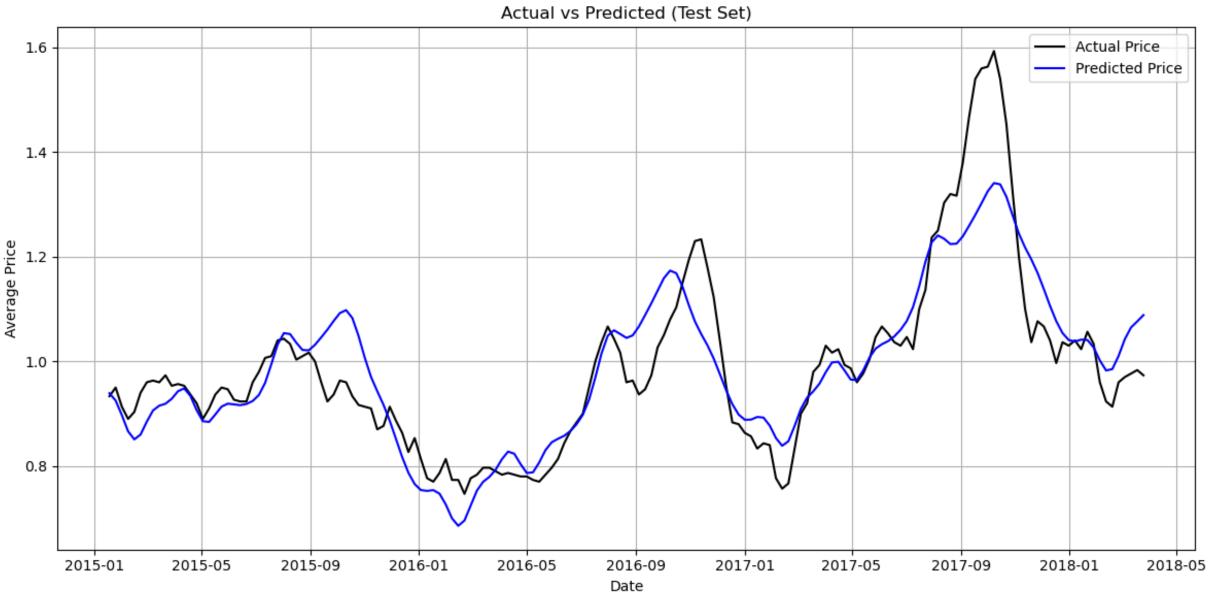
*Average Price vs Date*



*Trends*

Key Observations:

- U-shaped pattern: Prices started around $1.05 in early 2015, dropped to a low of approximately $0.85 in early 2016, then steadily increased to $1.30 by 2019
- Sharp recovery: After hitting bottom in 2016, prices showed consistent upward momentum
- Acceleration: The price increase accelerated significantly after 2017, suggesting growing demand or supply constraints
- Winter low: Prices typically drop in February-March (lowest point around -0.15 to -0.20)
- Summer surge: Strong price increases begin in July and peak in October-November (+0.20)
- Holiday premium: The October-November peak likely reflects increased demand during Thanksgiving/holiday season
- Consistent pattern: This seasonal cycle repeats predictably each year



*Actual vs Predicted Average Price*

# References

- GitHub Repository - Click Here
- A similar Case Study of the Crime Rate of Chicago was also developed. The GitHub Repository for it is present here - Click Here

*******