INFO 208: Mini-project

Figure 1 (Retrieved from https://pixabay.com)

Analysis and Prediction

of Avocado Prices

Megha Rajam Rao | MS DA, SJSU | Id: 013709488

**Abstract**

Renowned statistician and management guru late W. Edwards Deming rightly said - "If you do not know how to ask the right question, you discover nothing." Therefore, it is essential to define the problem and carefully interrogate the data with the right questions, before proceeding to perform the necessary pre-processing of the sample as per the research design. Following are the details of the project undertaken.

**Topic:** Analysis and Prediction of Avocado prices (specifically Hass Avocados)

**Industry:** Retail

**Data:** The dataset comprises of attributes such as the Average sales price per unit, Date of observation, Total volume, Total bags and volume breakup for each bag size and product code of Hass avocados. Here is the link to the dataset homepage - <http://www.hassavocadoboard.com/retail/volume-and-price-data> and <https://www.kaggle.com/neuromusic/avocado-prices>.

**Problem Statement:** Find the following insights from the data:

* Effect of avocado type (conventional or organic) on pricing. Furthermore, explore if we can reverse engineer and build a model to predict the type of avocado in front of us, based on price.
* Understand the price fluctuation across different regions. Where was the maximum price?
* Study the price distribution and fluctuation in the past few years. Find the most expensive year.
* Predict and forecast of future prices. Will it increase or decrease?

**Acknowledgement**

First and foremost, many thanks to Hass Avocado board for the dataset and author Brain Godsey for his book ‘Think Like a Data Scientist: Tackle the data science process step-by-step’. I express my gratitude towards my professor Dr. Glen Mules for the useful links, encouragement and guidance. Furthermore, I thank all the publishers of useful links and codes online.

**Table of Contents**

[Abstract](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 1

[Acknowledgement](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 1

[Table of Contents](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 2

[Brief Overview](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 3

[Dataset Studied](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 3

[Description & Structure of the data](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 4

[Notes on Exploration of the dataset](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 5

* [Explore and document the data and data structures manually](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964285) 7

[Questions to Ask of the Dataset](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964297) 8

[Develop and document a processing strategy](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964283) 10

* [Pre-processing the data](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964285) 10
* [Identify how this dataset could be studied](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964285) 10

[Summary](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964298) 11

* [What have you learned about this dataset?](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964285) 11
* [Few more Questions and Answers about the mini-project topic](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964285) 11
* [What have you learned from this Mini-Project?](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964285) 12
* [Reflection on the process](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964285) 12

[Insights](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964298) 13

[References 1](file:///C:\Users\megha\Downloads\INFO_208_MiniProject-Rajput.docx#_Toc530964300)4

**Brief Overview**

* **First things first. What is a Hass Avocado?**

A delicious variant of avocado, also known as Bilse avocado, with dark green-colored, bumpy skin.

* **Why this topic?**

Predictive analytics was routinely performed during my undergraduate study in Bioinformatics engineering. At work, we used to generate reports to enable users to analyze and predict fuel and commodity prices. Hence the interest.

Figure 2 (Retrieved from https://pixabay.com)

* **Why is subject matter important?**

I consider Avocado price as a unit of study. Price analysis and prediction can be done with any product or commodities irrespective of field. The questions we ask may slightly differ, but the basic process remains the same. Hence, it has countless applications that makes it important. On a lighter note, Avocados are tasty and nutritious, and worthy of study.

* **What are you hoping to achieve?**

Reiterating the problem statement, the goal is to unearth valuable insights from the data. Below are some specific goals.

* Effect of avocado type (conventional or organic) on pricing. Furthermore, explore if we can reverse engineer and build a model to predict the type of avocado in front of us, based on price.
* Understand the price fluctuation across different regions. Where was the maximum price?
* Study the price distribution and fluctuation in the past few years. Find the most expensive year?
* Predict and forecast of future prices. Will it increase or decrease?

Upon completion of this study, I hope to learn how to approach and handle data in the initial step of data preparation.

**Dataset Studied**

* Source URL: <http://www.hassavocadoboard.com/retail/volume-and-price-data> and <https://www.kaggle.com/neuromusic/avocado-prices>
* The dataset comprises of retail volume and price data of conventional and organic Hass avocados in the United States. We are excluding global data from this study. Nevertheless, the Hass Avocado Board website carries several insightful datasets for future research.

**Description & Structure of the Data**

* + - **Layout**
  + Quoting directly from the data owners, below is an excerpt from the website of [Hass Avocado Board describing the data on their website](http://www.hassavocadoboard.com/retail/volume-and-price-data).

The table represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU) in the table are only for Hass avocados. Other varieties of avocados (e.g. greenskins) are not included in this table.

* + **Domains of attribute values:** The columns in the initial and final table table are as follows.
* **Date**- The date of the observation
* **Average Price** - the average price of a single avocado
* **Total Volume** - Total number of avocados sold
* **4046**- Total number of avocados with PLU 4046 sold
* **4225** - Total number of avocados with PLU 4225 sold
* **4770** - Total number of avocados with PLU 4770 sold
* **Total Bags –** Total number of avocado bags sold
* **Small Bags -** Total number of small avocado bags sold
* **Large Bags -** Total number of large avocado bags sold
* **XLarge Bags -** Total number of extra-large avocado bags sold

Consolidated table from Kaggle.com has the following additional entries.

* **Unnamed field:** probably meant with Index number
* **Year:** Year of sale
* **Region:** Region of sale
* **Type:** Specifies if it is Conventional or Organic Avocado
  + **Any metadata analysis**
    - Markets in the United states (from Hass Avocado board website) *Table No.1*

|  |  |  |  |
| --- | --- | --- | --- |
| CALIFORNIA    Los Angeles   Sacramento   San Diego   San Francisco | WEST    Denver   Phoenix   Portland   Salt Lake City   Seattle   West Texas/New Mexico | PLAINS    St. Louis | SOUTH CENTRAL    Dallas   Houston   New Orleans |
| GREAT LAKES    Chicago   Cincinnati   Columbus   Detroit   Grand Rapids   Indianapolis | MIDSOUTH    Baltimore   Charlotte   Louisville   Nashville   Raleigh   Richmond   Roanoke | SOUTHEAST    Atlanta   Jacksonville   Miami   Orlando   South Carolina   Tampa | NORTHEAST    Albany, Boston, Buffalo   Harrisburg/Scranton   Hartford/Springfield   New England   New York   Philadelphia, Pittsburgh   Syracuse |

* **Regions:** IRI Retail Sales Regions in addition to the places listed above (in Table 1).

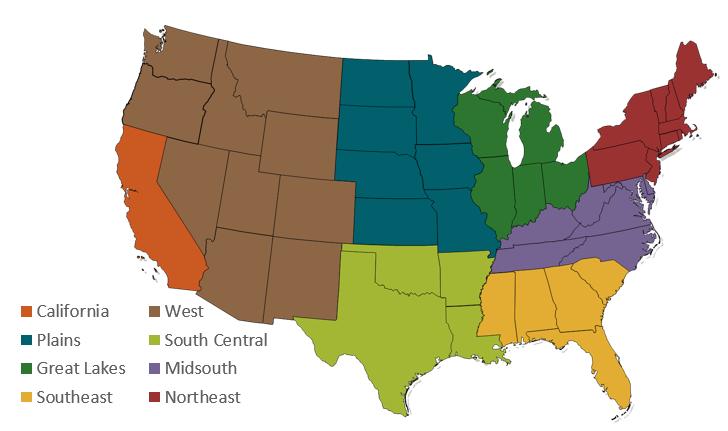


Figure 3 - Retrieved from <http://www.hassavocadoboard.com/retail/volume-and-price-data>

* **Data updation:** Every four weeks.

**Notes on Exploration of the dataset**

* There are 14 columns and 18,249 entries in the consolidated ‘Avocado.csv’ file from Kaggle dataset.
* Dataset has the records from January 2015 to March 2018.
* The initial dataset from Hass Avocado board is in excel format with fewer columns. From visual examination, the additional columns in the consolidated sheet are Type (specifying if conventional or organic), Region and Year. Hence, we can append data in the future by filtering out latest data, adding the aforementioned appending to existing consolidated dataset ‘avocado.csv’.
* A quick study using Tableau software revealed that all regions are well-represented and there is sufficient yearly data (*refer Figure 4 & 5).*

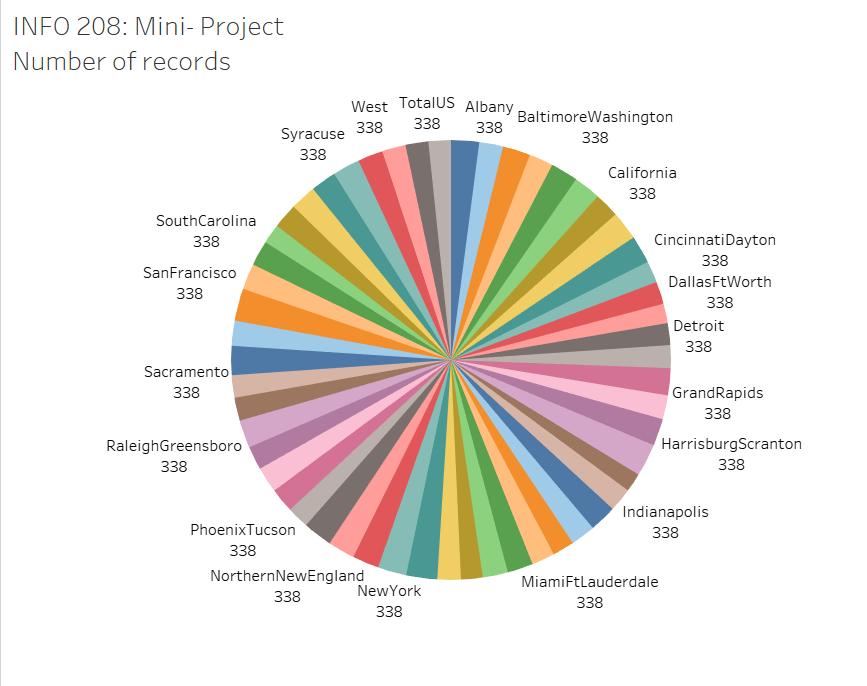


Figure 4 – Number of records per region

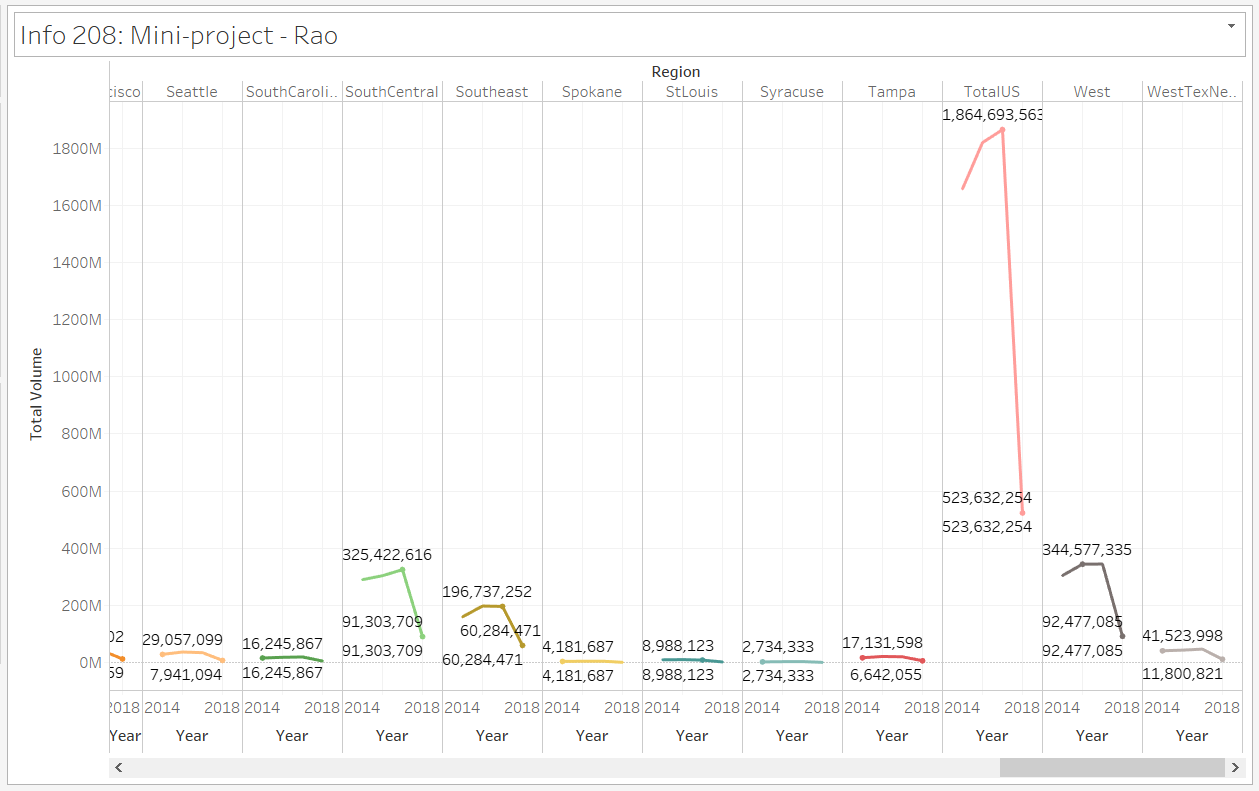


Figure 5 - Yearly data

* As expected, the volume of entire United States is higher than the individual regions and the average price is lower than some of the regions (*refer Figure 6).*
* Data looks consistent and complete post the initial exploration.

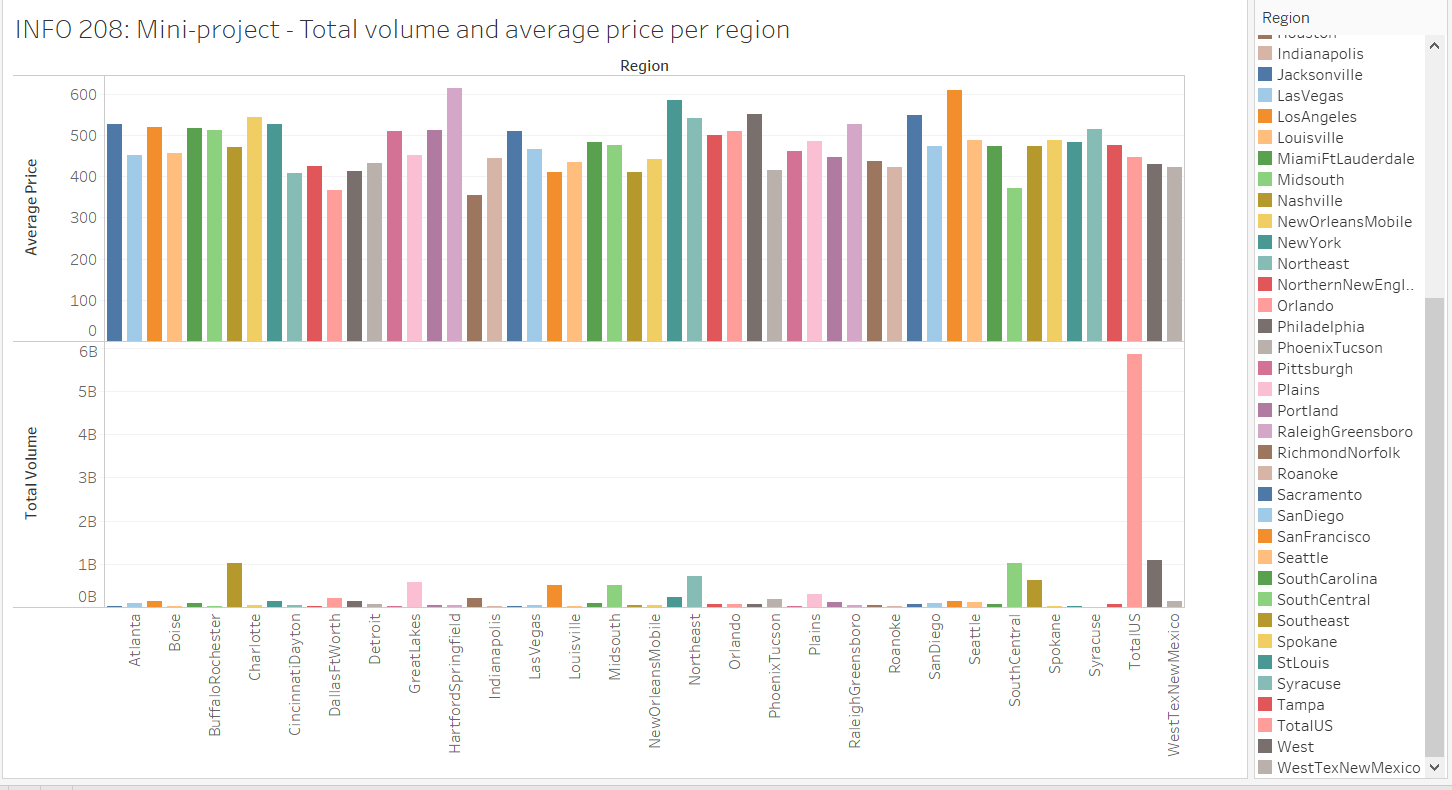


Figure 6 – Total volume and Average price for each region using Tableau

* + - **Explore and document the data and data structures manually**
  + The product lookup codes (PLU) codes were studied. Below are the details.
    - 4046 – Hass Avocado - small
    - 4225 – Hass Avocado - large
    - 4770 – Hass Avocado - Extra Large
  + Index column in the consolidated sheet from Kaggle can be eliminated as there are many inconsistent entries. It is also unnamed in the csv file.
  + As per the python pre-processing (using Jupyter notebook), there are no null value rows in the dataset.
  + Certain columns do have zero (for example No. of Extra-large bags) but it looks genuine and correct.
  + Certain columns representing number of avocados is in float format which is unlikely to be true as number of items must ideally have integer format.

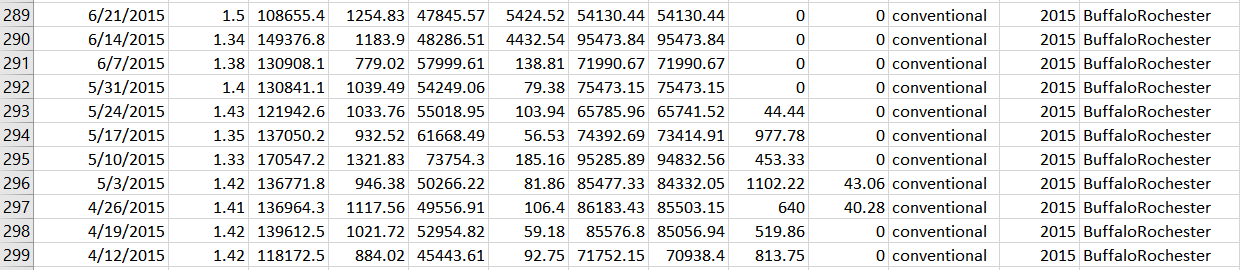


Figure 7 - Blank values and float values

**Questions to Ask of the Dataset**

* + **Why was the data collected?**

Data was collected for research purpose by Hass Avocado Board. They routinely perform various projections, market analysis and forecast.

* + **How was the data collected?**

As per their (Hass Avocado board or HAB) website, data is collected from IRI/FreshLook Marketing Multi-Outlet (MULO) retail scans. The reporting reflects retail sales scans across the following channels: grocery, mass, club, drug, dollar and military. HAB's calculation is based in part on data reported by Information Resources, Inc. through its Freshlook Service.

**Do we have the required license, credentials or permission to access the data?** (Castle, 2018) Yes, it is an open dataset and I verified using the website’s Terms of use section.

* + **Is the data reliable and accurate? Is there any missing data?**

Again, as per the HAB website, the information was believed to be reliable at the time supplied by IRI even though it is not legally guaranteed. Without limiting the generality of the foregoing, specific data points may vary from other information sources. Generally, it is deemed to be reliable and fairly accurate. There is no missing data as per initial analysis.

* + **Is the data up to date?**

Yes, it is updated every four weeks. For the purpose of study, we are using entries from January 2015 to March 2018.

* + **Do we have tools and network connections to implement this experiment?**

Yes, we do.

* + **What is the data format of the datasets? Is there any disparity that requires standardization of data (to convert to a single workable format)?**

The consolidated sheet from Kaggle Open dataset is in csv (column separated) format and data from HAB website is in excel format. It is very easy converting between these two formats. Hence no issues.

* + **What is the size of the dataset? Do we need to create a subset for granular analysis?** (Castle, 2018) **Furthermore, Is there data for future study?**

Dataset has 14 columns and 18,249 entries. We have ample data, thanks to the HAB website. There is sufficient data for current as well as future studies. The website itself provides smaller subsets of data that can be easily downloaded and used for any granular analysis. If needed, we can also work on creating a smaller subset using Python, R or SQL.

* + **Is there a need to join tables or create a summary table? Can it be done manually or using computational tools?**

The HAB website has smaller subsets of data for each year, region and type. Fortunately, we can use the consolidated dataset named ‘avocado.csv’ in Kaggle open dataset. There is no need to manually modify the table.

* + **Do we need more data to perform the current study? Is there a need to merge and consolidate data? Or to download data from outside the organization?** (Castle, 2018)

At the expense of sounding redundant, yes, there is sufficient data for current as well as future studies. There are various types of historical data in HAB’s website. However, we are relying on Kaggle’s consolidated dataset as they have painstakingly merged regional information of over 330 entries each, totaling to over 18,000 entries, and generously uploaded in their repository for academic purposes. To perform certain tasks during the research execution phase, we may need to use computational tools (aka python in this case) to segment and segregate the data.

* + **Do we have too much data to process? Will it be time-consuming?** (Godsey, 2017)

No, we have entries 18,249 entries that can be easily processed without much delay.

* + **Is the data accessed or modified in the production environment? Are there business users to be notified? Is there a downtime?**

No to all three. We are working with an open dataset downloadable in csv and excel format. Hence, there is no need to notify users or plan a downtime. We are not toying with sensitive data or environment.

* + **Any assumption?**

For the purpose of this study, we are assuming a single supplier for the avocado stock, in order to avoid undue complications.

* + **Are there inconsistent entries or redundant that require cleaning of the data? Can it be cleaned in the current environment with pre-existing tools and software? Is it possible to clean it manually?** (Castle, 2018)

Yes, there are couple of minor tweaks, discovered and documented in the data exploration step, to be performed. Even though manual cleaning sounds easy, python is more efficient in handling this task. Hence, we are planning to employ Jupyter notebook for the same.

* **Can the data at hand address our goals?**

Yes, we have the necessary data to answer the problems statements / goals and provide business insights.

* + **Is the proposed methodology compatible with the data format? Is it efficient?**

Yes, data format is compatible, and the methods proposed are efficient.

* + **What happens if a test fails? Is there a spot-check? If it shows an error, how do we proceed?** (Godsey, 2017)

We are not planning to rely on a single method. There are multiple ways and we can verify solutions from time-to-time using alternate and manual methods. If solution is wrong, we will substitute the main method with the alternate method.

* + **What inferences can we make? What truth is revealed or predicted**? (questions suggested by Professor Glen Mules)

Whether organic is more expensive compared to conventional Hass Avocados (we know the truth but let’s allow data to verify!). Furthermore, based on the price, experiment if we can build a model to predict the type of avocado in front of us. The average price distribution and yearly fluctuation across regions educates us about the market trends, and the forecast tells us what to expect in the future (whether the prices will increase or decrease in 2019 and 2020). Please refer Jupyter notebook for the complete results. Few screenshots below.

**Develop and document a processing strategy**

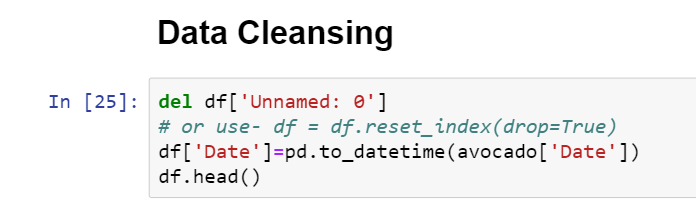
* + **Pre-processing the data**
    - **Data wrangling or munging** **step**: can be bypassed as consolidated sheet has all the necessary information in csv format. In the dataset exploration, the possible steps are explained if we are planning to start from scratch.
    - **Data cleansing:** Delete the unnamed column and convert date to datetime format. Search and delete any null rows. Look for missing values and check if it is genuinely missing.

Figure 8 & 9

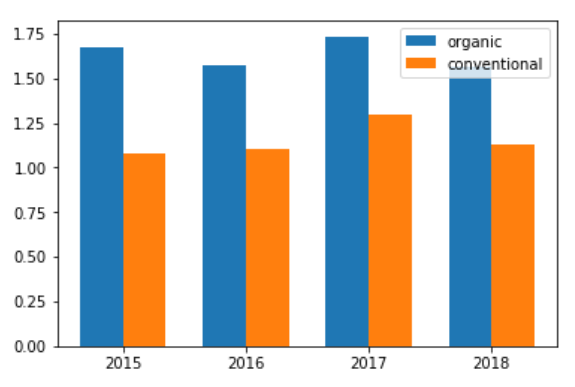
* + - Study the dataframe attributes and information to check for any irregularity.
  + **Note**: Please refer the attached python notebook for the complete code and output. Few screenshots (Figure 8 – 11 & Figure 13-17) are provided here for illustration purpose.
  + **Identify how this dataset could be studied.** 
    - For the analysis, we could be investigate using **basic statistical functions, data visualization tools such as boxplots, histograms, violin plot, heatplot, factorplot, distanceplot,** and so on.
    - The first three goals which are analyzing the effect of organic or conventional type on pricing, maximum and minimum average prices and Price distribution and fluctuation study can be completed using the aforementioned tools and methods.
    - For **timeseries** **forecast** the prices, an efficient method is using **fbprophet module** and **prophet () function**. Please refer the Jupyter notebook for the code and results. A little sneak peek – the prices may go down in 2019 and 2020. Good news for Avocado lovers!

Figure 10 above & Figure 11 below

* + - Other methods explored are **RandomForestClassifier,** which is supposed to have one of the highest accuracy in predicting if an avocado is organic or conventional based on its price. But **logistic regression** and **KNeighborsClassifier** will also help to know what type of avocado we are dealing with (based on its price). I have written the code of linear regression in markdown cell for future exploration.

* + **Identify the type of environment that would be good to process this dataset and how should use for processing to obtain the insight you would like to achieve.**
    - Suggested environment is Jupyter notebook for easy visualization of results. Tableau and excel sheet can be used for initial pre-processing and result confirmation.
    - We can code using Python, R, Julia and so on. At present, Python is recommended due to greater familiarity.

**Summary**

* + **What have you learned about this dataset?** 
    - I learnt how to approach and interrogate data with questions.
    - I learnt how crucial the data preparation step is. It is like the foundation of our analysis.
    - I found numerous ways to find free datasets. The search was enlightening as I came across some of the best codes. I created a word document to store important links and code.

Figure 12 - (Retrieved from https://pixabay.com)

* + - It was interesting finding questions that challenge my dataset and hypothesis. Apart from the suggested book and class inputs, I found solace in a couple of websites that guided me though this process.
    - I also learnt how hard it is to find the right data. It took me weeks to settle on a dataset but quickly I realized there was a better dataset. Now I understand why it takes almost 80 percent of the time to complete the data preparation step.
    - I tried my hand at data wrangling and it was time-consuming. Fortunately, I chanced upon a consolidated dataset saving me days of time. Again, learnt the importance of choosing the right dataset and evaluating it to examine its credibility and feasibility.
    - I learnt how to find refer and understand pre-existing code, adapt to our circumstance and troubleshoot issues.
    - I understood how to clean and modify the dataset as per the research needs.
    - I learned to reverse engineer from data to hypothesis to strategy to execution.

* + **Few more Questions and Answers about the mini-project topic –**
* **Does the overall strategy make sense on a general level?** (Castle, 2018) Yes, it makes sense. The strategy seems sound and logically correct. Even the outcome makes sense. For example, organic avocados are more expensive, and hence it falls short in sales volume.
* **Are the measures I’m seeing in line with what I already know about the business?** (Castle, 2018)

Yes, measures are similar to the ones that are used in a business scenario.

* + - **Do calculations in my analytical environment return the same results as the same calculations performed manually on the original data?** (Castle, 2018) Yes, the results can be verified using Tableau or statistical functions in excel sheet. All methods seem to agree in this case.
  + **What have you learned from this Mini-Project?**
    - **About being a Data Scientist** – It is not a cakewalk for sure. Apart from building the necessary skills, it is important to learn how to correctly approach the data and the problem in order to gain the right insights. Now, the class lecture on ‘right data’ and ‘right questions’ makes more sense. The book by Dr. Brian Godsey not only educated us on the methodology and necessary thought process but gave a taste of everyday life as a Data practitioner. The personal anecdotes in the book helped me understand some of the quintessential scenarios we may tackle in the future.
    - **About what skills you might need for the future** – Programming prowess, mathematical and statistical skills, along with proficiency in the subject matter is the most political answer, even though it is as rare as a unicorn. The patience to think, wild curiosity to dissect the problem at hand and the willingness to constantly learn and upgrade are some of the other skills that are essential to thrive in this field. Furthermore, I learnt that we need to be able to communicate insights in business language.
  + **Reflection on the process**
    - **Mistakes made** – Apart from constantly changing my topic, I was tempting to manipulate the data to fit my hypothesis. I admit I had weak moments when I pondered doing the same.

*“Torture the data, and it will confess to anything.” – Ronald Coase*

The above saying by British Economist and Nobel Laureate Ronald Coase beautifully highlights the key risk in any data science project. I learnt that often times, practitioners and especially gullible beginners are set out to prove the hypothesis right and try to fit the hypothesis on the data at hand. This approach is inherently wrong and sabotages our outcome.

* I understood that we have set goals, interrogated the data and scenarios with the right questions before reverse engineering with the help of this data. The dataset can be comparatively small, but it must be capable to providing the right insights.

**Insights**

Couple of interesting insights from the execution part.

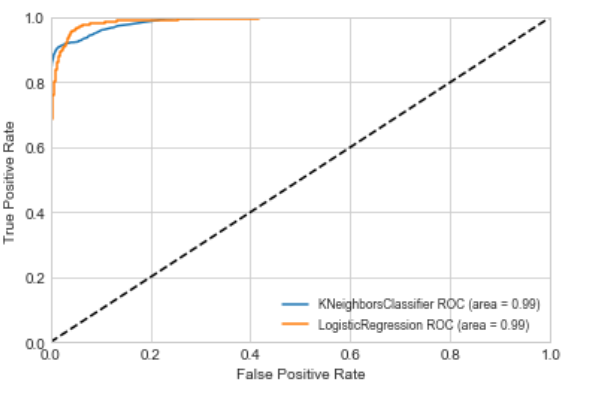
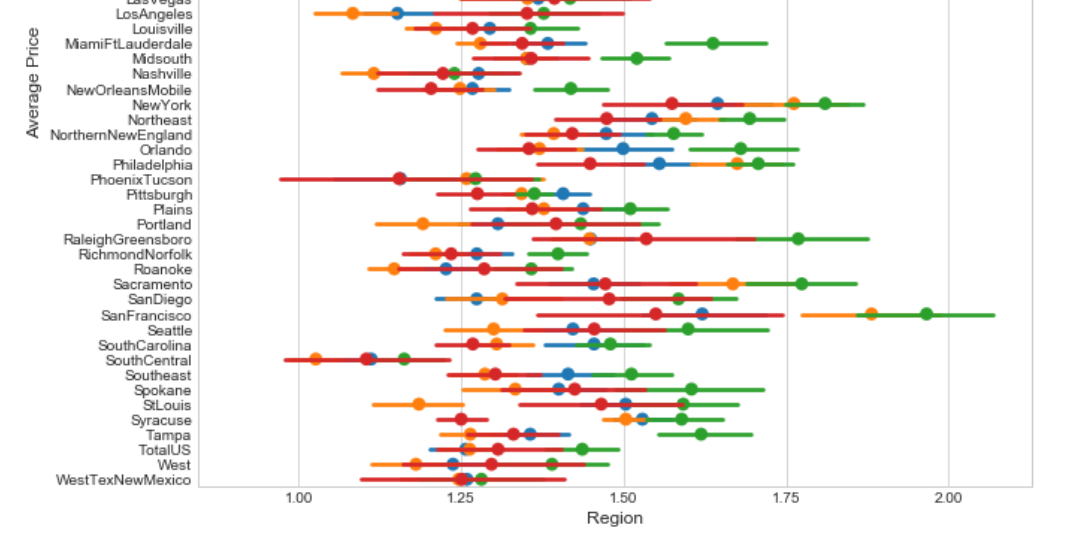
* + - As per the forecast, the prices are most likely to drop in the year 2019 and 2020 (*Figure 13).*
    - It is possible to predict the type of avocado placed in front of us, based on its price *(Figure 14*).

Figure 13 above

* + - San Francisco had the highest Avocado price in the year 2016. The type was organic Avocados. It comes as no surprise that Organic are pricier but sells in lower volumes

Figure 14 towards the right & Figure 15 below

*(Figure 15).*



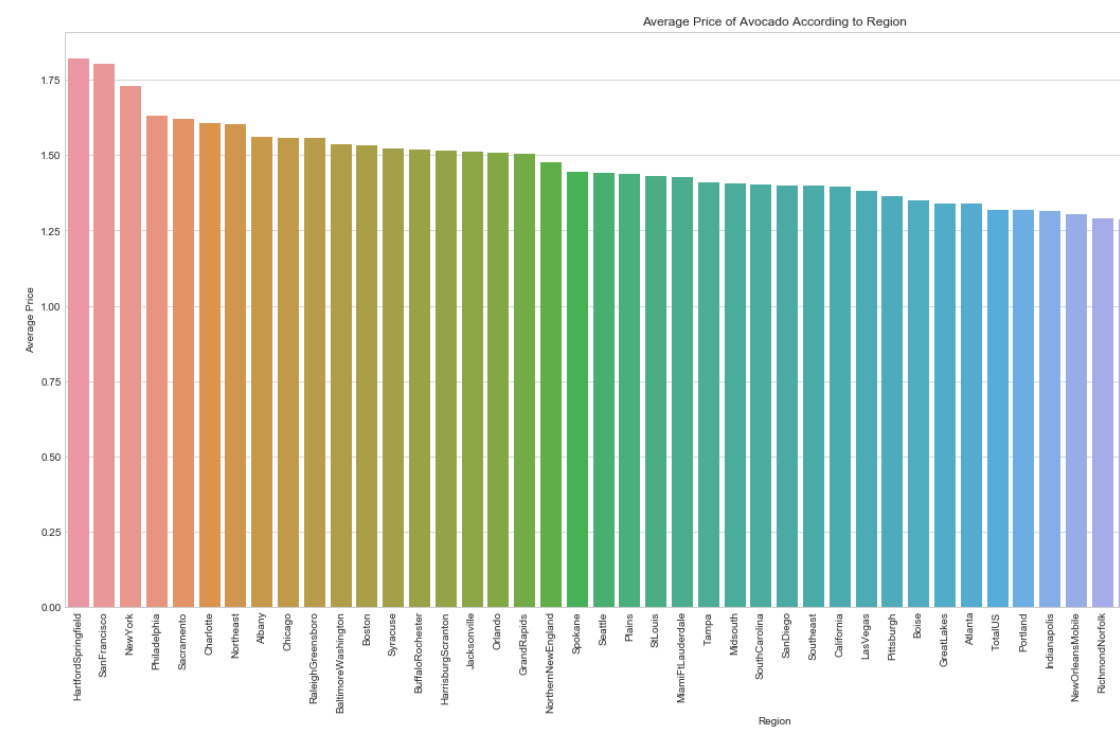
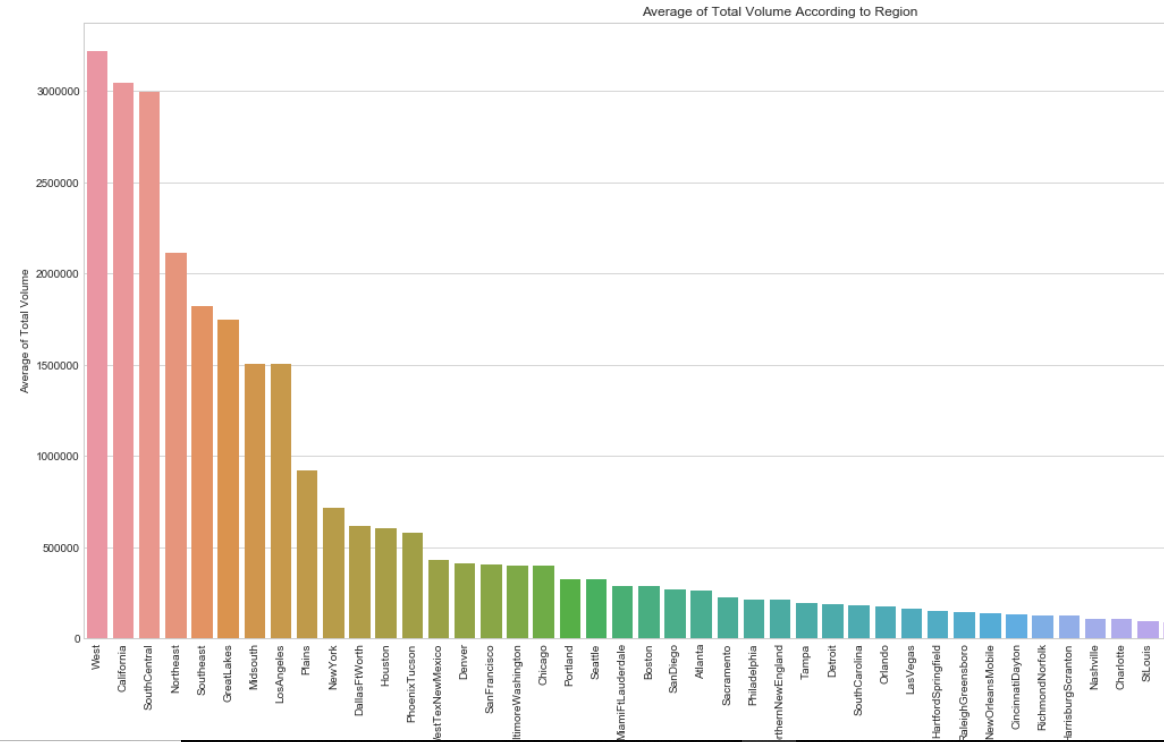
* + - Nevertheless, Hartford Springfield had the highest Average Price (followed by San Francisco and New York) *(Figure 17).*
    - West region is the highest consumer of Hass Avocados. State of California is the frontrunner among the states in US *(Figure 16).*

Figure 16 below Figure 17 below

**References**

Castle, E. (2018, August 29). 6-questions-to-ask-when-preparing-data-analysis [Web log post]. Retrieved November 11,2108 from /[https://www.sisense.com/blog/](https://www.sisense.com/blog/6-questions-to-ask-when-preparing-data-analysis/)

Godsey, B. (2017). *Think Like a Data Scientist: Tackle the Data Science Process Step-By-Step*. Manning Publications.

OpenClipArt-Vectors [Web log post - Animation]. <https://pixabay.com/en/avocado-food-green-half-nutrition-161822/>

Tookapic [Web log post - Photograph]. <https://pixabay.com/en/photos/avocado/>?

FoodieFactor [Web log post - Photograph]. <https://pixabay.com/en/avocado-avocados-food-healthy-food-2644150/>

<http://www.hassavocadoboard.com/retail/volume-and-price-data>