Data Mining Project Report

The Avocadoers

Evan Lee
University of Colorado - Boulder
CSCI 5502
Evan.N.Lee@colorado.edu

Hector Sanchez
University of Colorado - Boulder
CSCI 4502
Hector.Sanchez@colorado.edu

Madelaine Struwe
University of Colorado - Boulder
CSCI 4502
Madelaine.Struwe@colorado.edu

ABSTRACT

One of the biggest concerns in certain geographical regions of the United States is the relative market price and strength of the avocado, both as a crop and as a market product. The purpose of this project proposal is to determine a relationship between the weather and climate metrics of an avocado-producing region and the subsequent seasonal market share and market price in various geographical regions of the United States in order to produce a predictive model. We will accomplish this by analyzing and mining data sets from government agencies as well as surveys and other data from the private sector in order to determine the effects of individual weather and climate metrics (among which include standard metrics such as temperature, pressure, and precipitation) and using these calculations to produce a weighted model.

KEYWORDS

Avocados, Data Mining, Avocado Prices, Avocado Sales, Weather

ACM Reference format:

Evan Lee, Hector Sanchez, and Madelaine Struwe. 2018. Data Mining Project Report. In *Proceedings of Data Mining, Boulder, CO USA, Fall 2018 (CU BOULDER),* 7 pages.

https://doi.org/10.475/123_4

1 INTRODUCTION

For the project, we used data sets from Hass Avocado Board for data on avocado consumption, and pricing. We used the National Operational Model Archive and Distribution System to look at archived weather patterns where avocados are grown. We used the National Agricultural Statistics Service under the United States Department of Agriculture to determine the yield of crops of avocados. Using this data, we looked into how weather patterns can effect avocado harvesting, exports, and retail pricing.

CU BOULDER, Fall 2018, Boulder, CO USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

We used data from previous years to mine the relationship between these various metrics and data sets.

1.1 Motivation

The proportion of time the avocado spends in the thoughts of the average modern millennial has increased year after year since the advent and rise in popularity of avocado-based meals and foods. Menu items such as guacamole, avocado toast, and an adaptation on the popular BLT, the BLTA have fully integrated themselves into popular culture.

As such, a general drive exists in the food and crop industries for research into the trends, relationships, and effects of various metrics on the modern market. With our project research, not only can major industry players and consumer resource groups have insight into the predictive and projective capabilities of weather and climate data, but now, with our research and models, so can end-consumers with a vested interest into the market trends of the avocado.

We aim to give users at all levels of the avocado industry the power to predict with a high degree of confidence the future market trends to allow these users to make informed decisions, no matter the scope and impact of such.

1.2 Literature Survey

Research into this particular topic has been mostly top-down, starting at market trends/analysis data and has also been focused on comparative metrics against the competitors of the avocado. The premier source of market data collection and analysis have been consumer groups such as the Hass Avocado Board and the California Avocado Commission. With publications like *the Yearly AvoScore Card* and *State of the Category*, the majority of work by these consumer groups has been driven by increasing market share.

Groups that have performed research and analysis into the avocado industry have tended to be single-focus and small scoped, with each portion isolated and performed by a distinct organization. There has not been research into related trends and the modeling of the relationship as such.

Our team has identified an opportunity to tackle the bigger picture and link these discrete works together by approaching this problem from the bottom-up. Starting at the core of the avocado industry (weather and climate metrics), we can then link the relationship

between the base layer metrics to the layer up (avocado production and market volume) and finally to the most visible layer (avocado market trends and analysis).

2 METHODOLOGY

The following section details the data we used and the process to mine the data.

2.1 Data

We obtained data sets from:

- National Oceanic and Atmospheric Administration
  From this we gathered weather data from the area where avocados are grown and distributed to stores across the United States.
- United States Department of Agriculture
  From this we found data concerning the yield of avocado farms in the United States.
- Hass Avocado Board
  From this we obtained data on retail prices and sales. This information is divided by conventional and organic avocados, as well as by region in the United States.

2.1.1 Climate. Data was collected from the following major land-based stations in California:

- Camarillo Airport in Ventura County
- Los Angeles International Airport in Los Angeles County
- Monterey Peninsul Airport in Monterey County
- Paso Robles Municipal Airport in San Luis Obispo County
- Porterville Municipal Airport in Tulare County
- San Diego International Airport in San Diego County
- Santa Barbara Municipal Airport in Santa Barbara County
- Stockton Metropolitan Airport in San Joaquin County
- Watsonville Municipal Airport in Santa Cruz County

These sets were chosen due to these counties being the primary avocado-producing regions in California (**Figure 3**) as well as being the land-based stations with the highest availability & quality of data.

The scope of the data collected were climate metrics recorded at an hourly interval, for the years 2008-2017. In terms of raw data, these number of total rows across these data sets added up to around 1.2 million rows in total for this time period, with only 3 significant columns.

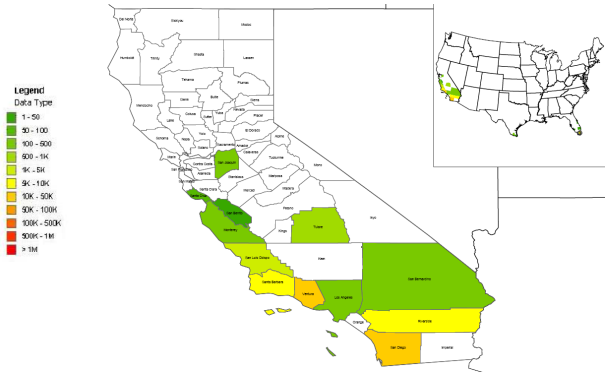

Figure 1: Number of harvested acres for avocados, 2002. *Source: USDA Census of Agriculture for California and conterminous United States (inset)*

2.1.2 Hass Avocado Board. The Hass Avocado Board Shipment Arrival Data sets go back as far as 2004 to 2018 and are available for free consumption online. These data sets include the total shipment volume per week-in-year keyed off of the source region. Most of the other data (such as that from other regions like Mexico) were pruned for this purpose of this project.

In terms of retail data, Hass Avocado board collects data from around the United States and is based on different store from different cities (note does not explicitly include county's but stores are based on county's from different states):

- California
- Dallas, TX
- Denver, CO
- Sacramento
- San Diego
- Los Angeles
- Seattle
- Total Us
- West
- Plains
- And a lot more

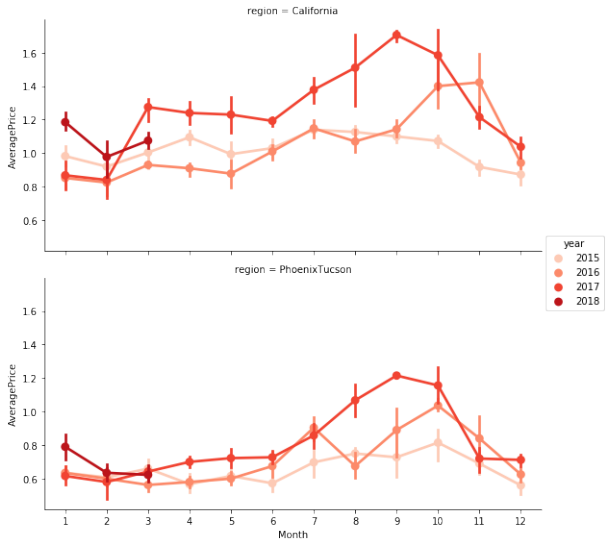

Figure 2: Averge price for region of Tuscon and California from 2015 - 2017 ***We can find interesting things with this data. Graph created by python pandas***

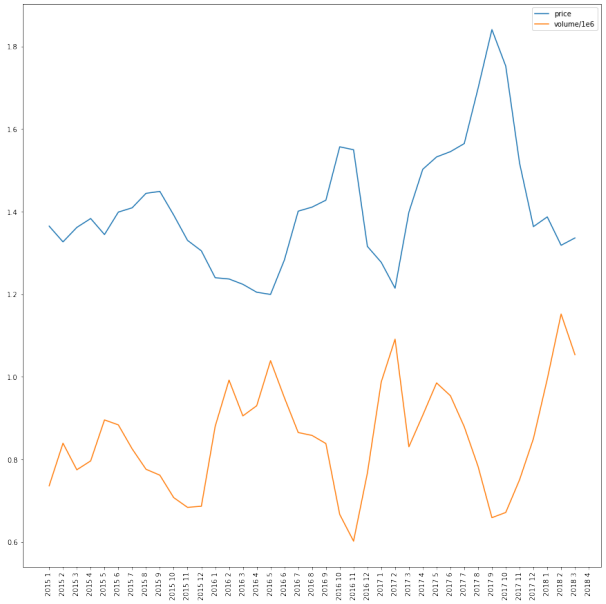

Figure 3: Average price for the entire US ***Graph created by python pandas***

With this data we have a good start in terms of what we can do for our project. Since the information mined from the data gave us information such as average price for the entire US, and also average price based on different cities and regions through out the US. And with different tools the python pandas has to offer and different python packages we can determine some very interesting things. Thus again making this data set quite rich in information.

2.2 Sub-tasks

The following major sub tasks were performed for this project:

(1) Obtain all relevant data sets.

(2) Process and clean all relevant data sets.

(3) Research and discover basic information about avocados & avocado processing.

(4) Mine relationships between neighbor data.

2.3 Obtaining Data

Luckily, most of the high-level data sets were conveniently available online in an easily-consumable format (.csv) and an intuitive scope.

The most detailed process came into effect when gathering NOAA weather and climate data. NOAA maintains an online portal, the *Climate Data Online* (CDO) tool, to allow citizens to request and certify data sets from their backend repositories. This was the tool employed for gathering the climate data, the process for which is described below:

(1) Locate a significant base station in the significant county.

(2) Determine the base-station ID for the station you are interested in, the ID format of which is *WLAN:XXXX*

(3) Query the CDO with the inputs of base-station ID, time span, format, and requested metrics.

(4) Wait for the CDO to report that the data has been retrieved and processed.

(5) Download the data in your designated format (usually .csv)

The Hass Avocado Board provided easily downloadable datasets in a consumable format (.xls) that was easily converted to a .csv file extension for our use in python parsing.

2.4 Processing & Cleaning

This task entailed clearly identifying and compiling data at the scope that we wished to accomplish. At this stage, we ran into our first issue of mismatches in availability and granularity of data. For example, the data sets from the National Oceanic and Atmospheric Administration (NOAA) are very extensive, ranging as far back as 1945, depending on the operational date of a particular land station. NOAA data sets are also extremely granular, with certain metrics recorded at a fifteen minute interval, others at an hour interval, and yet others only having daily summaries, averages, or ranges. In contrast, a large amount of the market volume/price data is aggregated at the weekly level.

To clean the data, we needed to identify which metrics we required and then drop the unnecessary data in the data sets we downloaded.

2.4.1 Climate. Some care was taken to normalize the climate data. In particular, two commonly recorded weather metrics are the *Bulb Temperature, Wet (F)* and *Bulb Temperature, Dry (F)*, which semantically entail the measurement of temperature using a wet and dry bulb thermometer. These recordings are very different, and represent temperature data at 100% and 0% humidity, respectively. However, normal conditions are typically a more mild humidity and are almost never at one polarity or another, so a calculated Normalized Temperature (F) metric was calculated and used for this project. The formula used to calculate this metric as as follows:

$$Distance_{Dry} = (Bulb_{Dry} - Bulb_{Wet}) * \frac{Humidity}{100}$$

$$Temp_{Normalized} = Bulb_{Dry} - Distance_{Dry}$$

These calculations also served to take some degree of humidity weighting into account while keeping the primary focus on temperature as a deciding metric.

Each row returned from NOAA was aggregated to an hourly level, normalized using the calculations above, aggregated again to the week-in-yearly level, then aggregated again to combine data across all distinct stations to form a representative week-in-year level data set for the state of California across major avocado-producing regions. This final product was used to relate against the other data from the other sources.

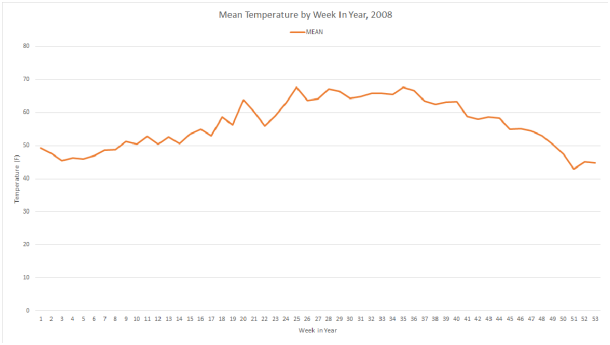

Figure 4: Sample graph representation of normalized data by week, 2008. *Generated using Microsoft Excel.*

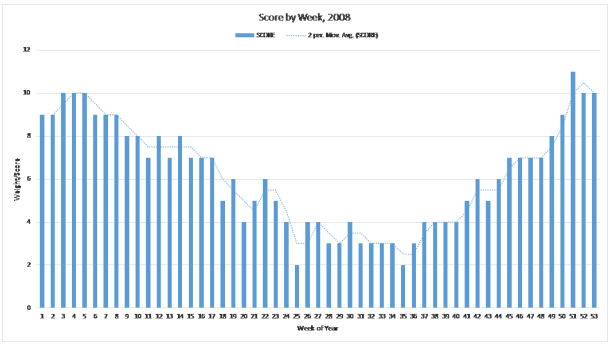

Figure 5: Sample graph representation of weekly score, 2008. *Generated using Microsoft Excel.* Two week running average included.

Referencing **Figure 5**, one can notice that the weights decrease during the middle of the year, which is generally during the more mild season.

2.4.2 Shipment Volume. Eight years' worth of shipment volume data was sourced from the Hass Avocado Board, from 2010 to 2018. This time range was selected to conform to the time range of the NOAA climate data, minus the two years needed to complete the avocado fruit growth/harvest cycle. Conveniently, this data was already scaled at the week-in-year to value format, there was just one simple script used to clean up the data to conform the formatting of the climate data (namely the Year:Week-In-Year to Value mappings described in the below section).

Unfortunately, some of the Hass Avocado Board's Shipment data was incomplete (which was evidenced by the presence of the string "n/a" in their data), so these holes were filled by taking the difference between the bounding values, dividing that by the number of rows in the hole, and either de- or incrementing the initial boundary by that amount to represent a linear progression from initial boundary to final boundary in order to avoid errors when processing data.

Note: this operation was only performed on holes in data, not if the data reported 0 pounds shipped that week.

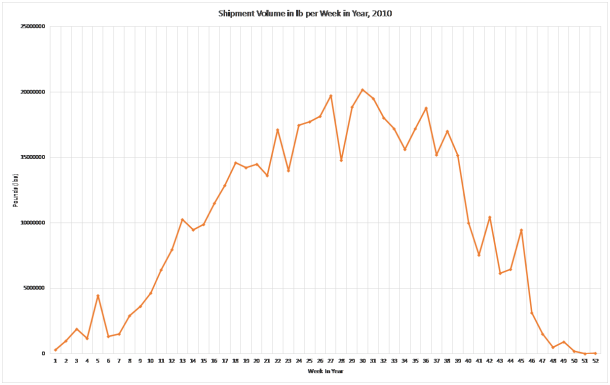

Figure 6: Sample graph representation of shipment volume by week in year, 2010. *Generated using Microsoft Excel.*

2.4.3 Retail. Similar to the above, the Retail Data sets that we selected for our use also spans eight years (2010 - 2018) and was also sourced from the Hass Avocado Board. These data sets were also scaled to week-in-year granularity, except there were two different formats of data, pre-2015 versus post-2015, which used differing date formats and differing column names. Additionally, pre-2015 data sets were malformed, being affected by column misalignment due to the presence of commas in the numerical values which caused initial reporting to be off-target. An additional script was written to correct and normalize the formats and column misalignment.

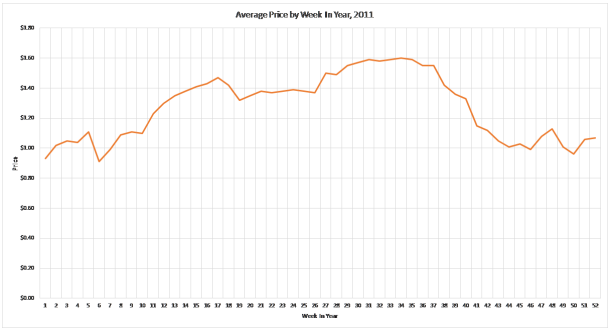

Figure 7: Sample graph representation of average price by week in year, 2011. *Generated using Microsoft Excel.*

2.5 Programming

The primary scripting language used in this context was *python*, version 3.6 and above. Scripts were written to normalize the data and aggregate the data to the desired weekly level. The normalization was coded to match the calculations described above.

The aggregation mechanism was to parse each reported time-stamp of the format *Year-Day-Month Hour:Minute* into *Year::WeekOfYear* tuple mappings (each distinct entry of which represented a discrete bucket), assign each row into the appropriate bucket, and then run aggregations based on the data inside that bucket; in this case, the mean and standard deviation for that bucket. This was performed using the standard *collections* and *statistics* modules available in python 3.

A simple script was also written to align the Hass Avocado Board Shipment Volume data to match this same format.

As mentioned above, an additional script was used to format, correct gaps, and normalized Hass Avocado Board Market Price retail data.

Another simple script was written to clean data from the USDA to determine the avocado yield per year, for the entire state of California.

2.6 Assigning Weights

The two greatest climate-based metrics that have protracted effects on the avocado production for a region are temperature and humidity, both metrics of which we have taken care to include in some way.

To assign a weight to each week, we have decided to implement a naive custom scale, that will assign weights to each week-in-year on a positive integer scale. This number represents the absolute value of weight of the effect of the weather and climate conditions for that week. In layman's terms, the greater the number, the worse effect that week is represented to have on that season.

2.7 Calculation

According to the California Avocado Commission, the effects of weather data are offset by two years, due to the avocado fruit harvest and growth cycle:

"Because of this two-year bud-break-to-harvest cycle, the amount of avocado fruit harvested in the current year is a result of cumulative weather and avocado grove cultural management from the preceding two years."

Therefore, as an example, for the 2015 Hass Avocado Board Retail Volume and Price Data, we will be evaluating the March 2013 - March 2015 NOAA climate data.

3 RESULTS & CONCLUSIONS

3.1 Assumptions

In order to simplify the analysis process, the following assumptions were made:

- The avocado growth/harvest cycle takes exactly two years to complete, ending on the same week-in-year as inception. Therefore, all avocado shipments consist of 104 weighted weeks of climate data.
- The first few weeks of avocado growth impact the fruit harvest to the greatest weight.

- Shipment effects on market price manifest the week after that shipment.

3.2 Conclusions

Loosely, when inspecting sample graphs for climate data (**Figure 5**) and shipment volume (**Figure 6**), it can be observed that a loose but intuitive relationship between poor weather conditions at inception and raw number of higher weighted weeks during an avocado's lifespan. This leads to the following conclusion:

Week-in-years with a higher average score yield fewer avocados.

However, when comparing trends between price and shipment volume or between price and rolling climate weight, the results look very different:

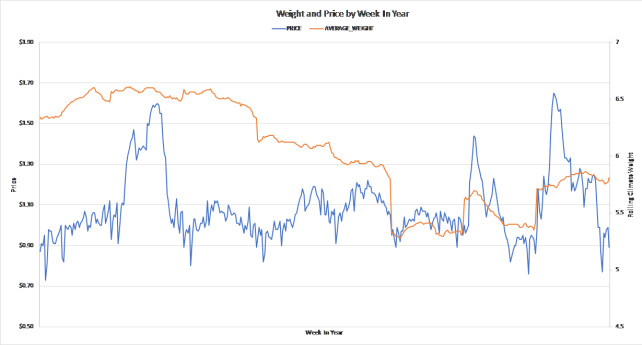

Figure 8: Comparisons of Market Price and Rolling Effective Climate Weight, all data. ***Generated using Microsoft Excel.***

Figure 8 displays **an apparent lack of relationship between rolling 104-week climate weights and its effects on average price** ***until around the post-2014 period.***

This can be attributed to the modern explosion of trending towards organic products and the consumer desire for organic produce.

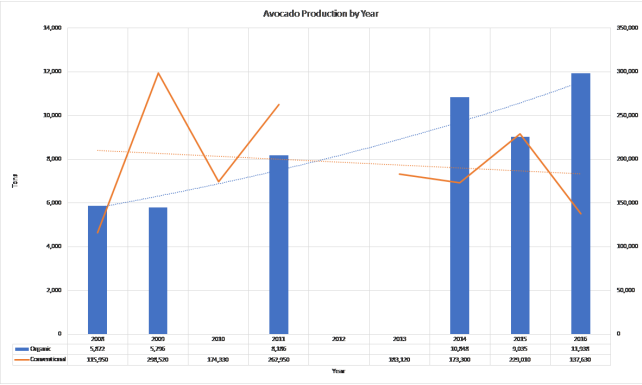

Figure 9: Avocado Production in Tons by Year, Organic and Conventional ***Generated using Microsoft Excel.***

Figure 9 displays an trending increase in the organic avocado production of the United States and a relatively steady-state production level of conventional avocado production. The timeline of this trend appears to correlate with the shift in pattern displayed in **Figure 8** whereupon climate metrics appear to weight more heavily on downstream market price post-2014.

The market share of organic avocados is vastly outshadowed by the market share of conventional avocados (tens of thousands of pounds vs hundreds of thousands of pounds). However, in common markets, the organic avocado is seem as as a premium product and is therefore generally priced higher than conventional avocados. This quality of the organic avocado means that fluctuations in average pricing of organic avocados entail a greater proprotion of change, and so might influence the total average market price at a greater proportion if there is no weighting in calculation of average market price data by the Hass Avocado Board.

This assertion is backed up by the raw market data reported by the Hass Avocado Board whereupon it appears that on average, organic avocados command a price of around one dollar greater than their conventional counterparts.

We can therefore conclude:

Weather and climate do not have as pronounced of an effect on market price for the conventional avocado, but is theorized to have more of an effect on the market price for the organic avocado.

4 DISCUSSIONS

4.1 Techniques & Concepts

During the course of this project, the following data mining techniques & concepts were used:

- Data aggregation & dimensionality reduction. High granularity data (e.g. hourly) was aggregated upwards multiple times to a lower granularity (weekly).
- Data normalization. In order to interpret the relationships between data, all data sets were normalized in formatting, granularity, and scope.
- Missing-data imputation. Missing data in Hass Avocado Board sets was filled in using a linear projection.

4.2 Challenges

One of the biggest challenges encountered was simply the overwhelming amount of data with the original scope of the project. Our original intended scope of the project is too large to accomplish in the remaining time allotted. We intended to accomplish the mining of the relationships of avocado market share with multiple of the following weather attributes:

- Sunshine & Visibility, Cloud Cover
- Precipitation

- Temperature
- Humidity

This would have increased the dimension of our data four-fold versus what we settled on (i.e. single-dimension scale of temperature only), which directly affected the amount of time used to gather and process the data.

Additionally, our original scope was intended to encompass the data collected from multiple major weather station locations distributed across avocado-producing regions in California.

However, the combination of the multiple dimensions plus multiple additional data sets would have unacceptably increased the amount of time required for processing & cleaning.

Related to these issues, we also found that the process for gathering climate data to be significantly tedious and time-consuming. The queries for data from the NOAA CDO were required to be very specific and had enforced limits on scope. CDO enforces a maximum query limit of 10 station years, which is defined by NOAA to be 10 years worth of data from a single land-based station. Therefore, queries are limited to 10 years' worth when querying for a single station or one year's worth when querying for 10 stations. Additionally, the amount of time required for CDO to generate a data report scaled directly with the amount of data queried for (i.e. introducing additional metrics increased the query time by that dimension's worth of data). Lastly, CDO queries could not be performed in parallel and had to take place in serial, which was a limitation of the CDO tool itself. All of these factors contributed to a high time investment in the initial base climate data.

4.3 Extending the Project

Given more time & resources, this project could be extended to a broader scope or more in-depth applications, such as:

- Expanding data sets to all major avocado-producing regions versus just the avocado-producing counties in California.
- Increasing time scope from 10 years to as large as possible given the availability of data (2002 - current).
- Increasing the dimensionality of climate data metrics by introducing additional attributes such as sunshine, humidity, pressure, or visibility.
- Tracking two parallel trends, separating conventional and organic avocado trends and data in order to isolate that attribute as an impactful variable.
- Combining consumer factors and attributing market volume/price reactions to such as compared to weather and climate factors.

ACKNOWLEDGMENTS

The authors would like to thank avocados everywhere. Let's keep up the good work. And let us provide great avocados for people everywhere.

5 REFERENCES

[1] Hass Avocado Board. 2018.
Retrieved from http://www.hassavocadoboard.com/retail/volume-and-price-data

[2] NOAA. National Centers for environmental information.
Retrieved from https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

[3] United States Department of Agriculture. National Agricultural Statistics Service.
Retrieved from https://quickstats.nass.usda.gov/

[4] California Avocados. 2018. Fresh California Avocados.
Retrieved from https://www.californiaavocado.com/

[5] EPA. 2018. CALIFORNIA AVOCADOS
Retrieved from https://www.epa.gov/sites/production/files/2015-09/ca_avocado.doc

[6] California Avocado Commission Retrieved from
https://www.californiaavocadogrowers.com/growing/how-california-avocado-tree-grows/avocado-fruit-growth