Github Repo: https://github.com/UC-Berkeley-I-School/Project2_Steward_Nhan_Pradjinata

Veggie Tales for Fruitful Discussion: Exploratory Analysis of Produce Prices for 2015-2019

by Amy Steward, Justin Nhan, Kevin Pradjinata

Introduction:

Fruits and vegetable crops are scattered across the U.S. regions, largely based on the growth requirements of each product. As a result, source location, purchase location, and seasonality can impact the price of the produce you purchase at the grocery store. To better understand these produce price trends for end consumers, our data exploration aims to evaluate the following questions.

Questions:

Markup Ratio - Deep Dive into specific products

- Has the markup ratio for produce risen since 2015, resulting in higher retail store profits while increasing the financial burden on consumers?
- How does the Markup ratio vary by product and location?

Which produce had the highest price spread across cities each year, and why?

- Is it the same product each year?
- Is it the same city with the highest & lowest price each year?

Data Cleaning / Sanity Checks

To address these questions, we utilized a historic dataset covering produce farm prices and retail prices across 4 cities: Atlanta, Chicago, New York and Los Angeles.

Our Data cleaning strategy consisted of 4 steps:

1. View the data: Understand what is there, where the nulls are and what a complete dataset would be for each product.
2. Clean the data: Ensure the numbers, dates and data are in the proper format for analysis
3. Sanity Check
4. Format the data: Ensure the table is in the proper format for analysis e.g. the correct data are in columns and the values are properly aggregated
5. Repeat 2 and 3 as needed

Step 1: View the data

Here we viewed the data in the data frame. This included bringing the data in using pandas and looking at the first 5 rows of the data. This indicated a few items that needed to be cleaned up including the format of the farm and retail prices. Additionally we identified the data types using

the info function and found they were all objects, meaning we would need to update the date to datetime and the numbers to strings. Finally, we identified any blanks in the data to replace them with NaNs so we could continue our data analysis.

Step 2: Cleaning the data

From our data viewing, we found that we needed to:

- Update the price fields to remove the $ sign and change the format to float
- Update the date field using to_datetime changing it to a datetime field
- Replace any '' with NaNs
- After this we trimmed the dates to just look at data from 2015 - 2019 in order to hone in on the most recent data.

It is also important to note we found an extreme outlier in the Broccoli dataset for the retail price in March in New York of $11.74, while otherwise the price for Broccoli in New York is generally less than $2.00

Step 3: Sanity Checks

Once we could aggregate the values of the data and we had our data frame focused on the timeframe we wanted to analyze (2015-2019). We checked the value counts for the products to understand what products had the largest data to work with.

```
17]: Broccoli Crowns        227
     Iceberg Lettuce        227
     Green Leaf Lettuce     227
     Cauliflower            226
     Red Leaf Lettuce       226
     Romaine Lettuce        225
     Carrots                225
     Strawberries           225
     Celery                 224
     Oranges                216
     Broccoli Bunches       183
     Potatoes               150
     Avocados               150
     Flame Grapes           121
     Cantaloupe             108
     Honeydews              105
     Tomatoes                68
     Thompson Grapes         54
     Nectarines              50
     Peaches                 47
     Plums                   36
     Asparagus               28
```

Figure 1: Each product available for analysis with the available count of data points. Our group focused on full datasets of >200 data points for the city retail analysis and 5 focus products to deep dive for the Markup Analysis.

Figure 1 shows the data points available for each product. Our group focused on the most complete datasets with more than 200 data points. For the city retail analysis, we analyzed all products with more than 200. However, for the Markup Analysis deep dive, we honed in on 5 random products (Broccoli, Romaine, Strawberries, Carrots and Cauliflower) having the fullest datasets between 225-227 values.

Markup Ratio Investigation and Trend Identification

This analysis focuses on differences in Markup Ratio Price across the 4 regions provided in the analysis. The Markup Price takes the Retail Price for each city and divides it by the Farm Cost.

This tells us the Markup Ratio, or how much higher the Retail Price for each city is over the Farm Price.

Here, we grouped segmented product data by month year and unpivoted it so each city was in a single column. This was completed through the creation of a function and allowed us to create a function to quickly group and pivot the data. Another function was created to build each chart.

Two examples of the Markup Ratio by city for a product are provided below. Figure 2, shows an example of the Market Price Analysis for Romaine showing a seasonal trend with highs around the winter months and lows in the summer. This was typical for all analyzed products except carrots where there were no obvious trends.

Romaine Markup Price Analysis

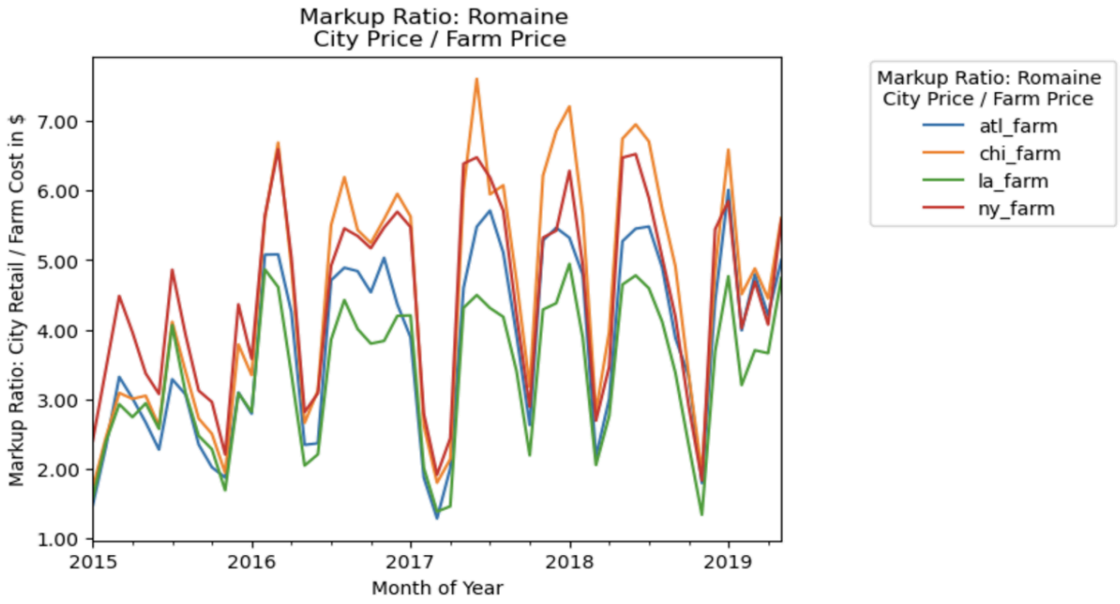


Figure 2: Romaine Markup Ratio. An example of the markup ratio showing the yearly trends of higher markups in the winter and lower markups in the summer with a variance between each city.

Markup Ratio Analysis: Intra month variation

Intra month variation looks at the Markup Ratio of a single product across all 4 locations for a specific month. This data tells us how the Markup Ratio varies among the cities.

The intra month variation was calculated by taking the Max and subtracting Min Markup Ratio for each month and product. e.g. If the highest price for Romaine in October 2016 in Chicago that was subtracted by the lowest price for October 2016 from Los Angeles. Figure 3 below shows this example of the Max - Min Markup Ratio difference in pink.

From here, the mean and standard deviation of the Max - Min Markup Ratio was calculated on a per product basis to understand how this range varies for each product.

The standard deviation was divided by the mean to gauge the relative size between the standard deviation as compared to the mean. E.g. with a Standard Deviation/Mean of 1, that means the standard deviation was as great as the average.

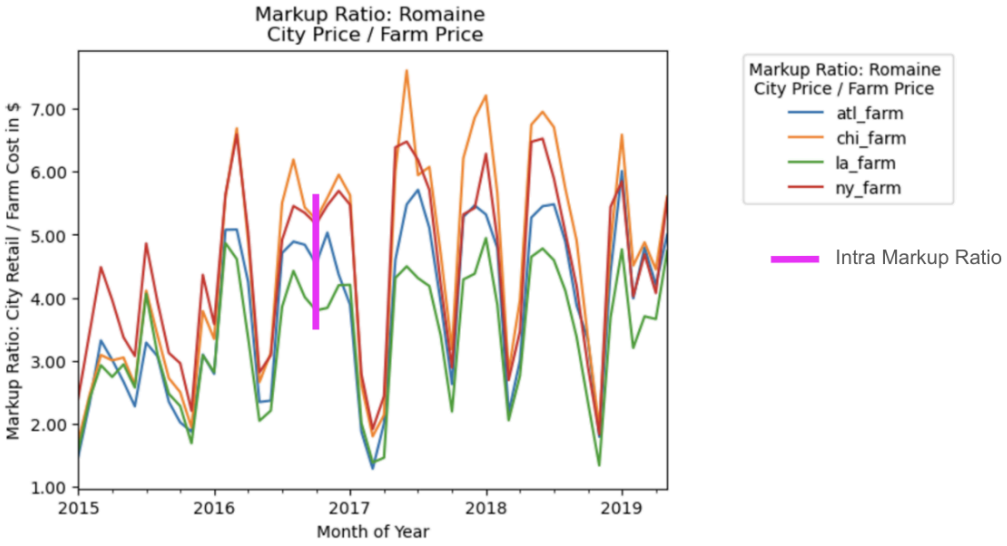


Figure 3: Romaine Markup Ratio is marked in pink, here showing the highest and lowest Markup Ratio for October 2016.

Markup Ratio: Overall mean, standard deviation and standard deviation / mean between cities, by Product

Product	Market Ratio: Mean	Market Ratio: Standard Deviation	Market Ratio: Standard Deviation / Mean
broccoli	0.634311	0.633945	0.999423
strawberry	0.695944	0.361854	0.519948
cauliflower	0.763134	0.286323	0.375193
carrots	0.883369	0.266731	0.301947
romaine	1.371330	0.500457	0.364943

Figure 4: Final output of the Market Ratio analysis showing the overall mean, standard deviation and standard deviation / mean by product.

City Retail Price Spread Analysis

Which produce had the highest price spread across cities each year, and why?

- Is it the same product each year?

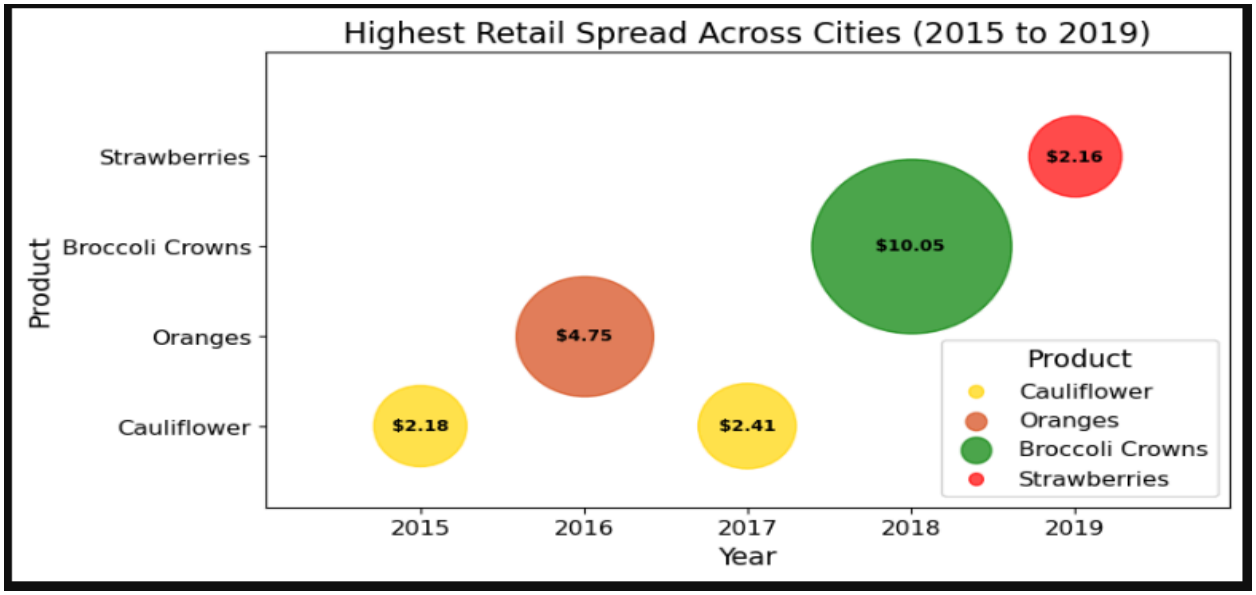
- Is it the same city with the highest & lowest price each year?

Data Exploration:
To understand which product had the largest retail price spread across all the cities (Atlanta, Chicago, Los Angeles, New York), we determined the highest price and lowest price for each product data entry. Then, the retail price spread was calculated by taking the difference between the maximum and minimum prices.

	year	month_year	productname	atlantaretail	chicagoretail	losangelesretail	newyorkretail	max_retail	min_retail	retail_spread
0	2019	2019-05-01	Strawberries	2.23	1.70	1.99	2.54	2.54	1.70	0.84
1	2019	2019-05-01	Romaine Lettuce	1.72	2.00	1.69	1.99	2.00	1.69	0.31
2	2019	2019-05-01	Red Leaf Lettuce	1.84	1.84	1.69	1.89	1.89	1.69	0.20
4	2019	2019-05-01	Oranges	1.42	1.45	1.34	2.05	2.05	1.34	0.71
5	2019	2019-05-01	Iceberg Lettuce	1.39	1.46	1.69	1.56	1.69	1.39	0.30

Table 1: Output snapshot of our retail price spread calculation. As seen on the first data row, the lowest retail price for strawberries was in Chicago, while the highest was in New York, with a total spread of $0.84.

After grouping and sorting, the following bubble chart shows which products had the highest retail price spread each year, from 2015 to 2019, and the relative sizes:



Graph 1: Products with highest retail price spread, per year (2015-2019).

Analysis:

Across the years, the product with the highest retail price spread typically shifted. The highest retail price spreads interestingly had a wide range, from $2.16 to $10.05. For each product, there was no trend or consistency for which city had the highest or lowest price contributing to the retail price spread.

Focusing on the top 2 retail price spreads found, we decided to further investigate the validity and reasoning behind the 2018 broccoli crowns price spread and 2016 oranges price spread.

2018 broccoli crowns price spread:

productname	date	farmprice	atlantaretail	chicagoretail	losangelesretail	newyorkretail
Broccoli Crowns	2018-03-18	0.41	1.69	1.9	1.99	11.74

Table 2: Depicts the only datapoint in the dataset for a product price < $10.

After a quick analysis of our dataset, we realized this datapoint for broccoli crowns price in New York from March 18, 2018, was the only datapoint for a product price that was greater than $10 in our entire dataframe. As this is also the datapoint reflecting the highest retail price spread found, we decided to further investigate the validity of this price point.

	productname	date	atlantaretail	chicagoretail	losangelesretail	newyorkretail
921	Broccoli Crowns	2018-03-25	1.69	1.90	1.99	1.61
935	Broccoli Crowns	2018-03-18	1.69	1.90	1.99	11.74
948	Broccoli Crowns	2018-03-11	1.69	1.95	1.99	1.91

Table 3: Broccoli crowns data for March 2018.

Observing the March 2018 broccoli crown data available, we found that the New York price for broccoli crowns in the surrounding weeks was relatively constant and low, below $2.

```
---BROCCOLI:  MARKET ABOUT STEADY.  cartons CA bchd 14s 14.00-16.00 occas 13.00
fineappear 17.00-18.00 bchd 18s 16.00-18.00 occas 13.00 Crown Cut Short Trim
14.00-16.00 FL bchd 14s 14.00 MX Crown Cut Short Trim 8.00-10.00 No Ice
12.00-14.00 mostly 12.00 fr appear fr cond 5.00-6.00 Baby Hybrid Type bchd 18s
offerings insufficient to quote
```

Figure 5: Excerpt from the New York Terminal Prices markup report, published via the USDA: https://mymarkupnews.ams.usda.gov/filerepo/sites/default/files/2315/2018-03-19/316300/NX_FV02020180319.TXT

Furthermore, the New York Terminal Prices markup report from March 19, 2018, stated that the broccoli markup was "about steady". The prices for broccoli cartons (14-count) from California range from $14.00 to $16.00, which coincides with the unit carton prices we see in our dataset that are in the $1-2 range. This further strengthens the concern we had over the $11.74 datapoint. Additionally, we were not able to find any news articles that highlighted an abnormality in broccoli crowns markup activity during this timeframe. With the contextual evidence of this datapoint for $11.74 broccoli in New York on March 18, 2018, **we suspect this outlier is a potential result of a data entry error. Nevertheless, we will not exclude this datapoint** until we can confirm it is an error with the USDA. We suspect this datapoint may have likely been entered wrongly as $11.74, rather than $1.74, which would be more consistent with the other broccoli prices.

2016 oranges price spread:

	productname	date	farmprice	atlantaretail	chicagoretail	losangelesretail	newyorkretail
2472	Oranges	2016-02-28	0.33	1.40	1.07	1.24	1.79
2482	Oranges	2016-02-21	0.33	5.84	1.61	1.09	1.79
2493	Oranges	2016-02-14	0.33	1.62	1.40	1.28	0.96
2503	Oranges	2016-02-07	0.36	1.71	1.18	1.09	1.70

Table 4: Oranges data for February 2016.

The retail price spread for Oranges on February 21, 2016, was considerably high at $4.75, with a high of $5.84 in Atlanta and low of $1.09 in Los Angeles. Compared to other weeks in February, the $5.84 price point in Atlanta was approx. three to four times higher. We further investigated this "spike".

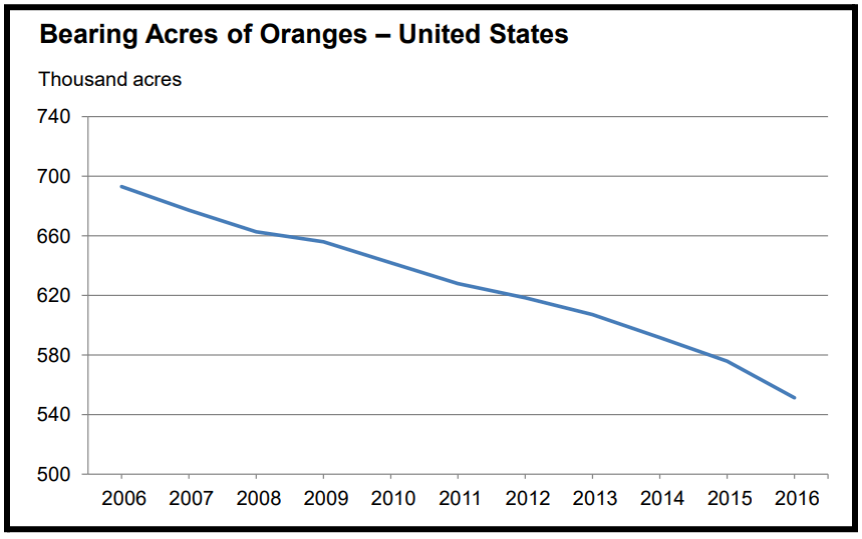
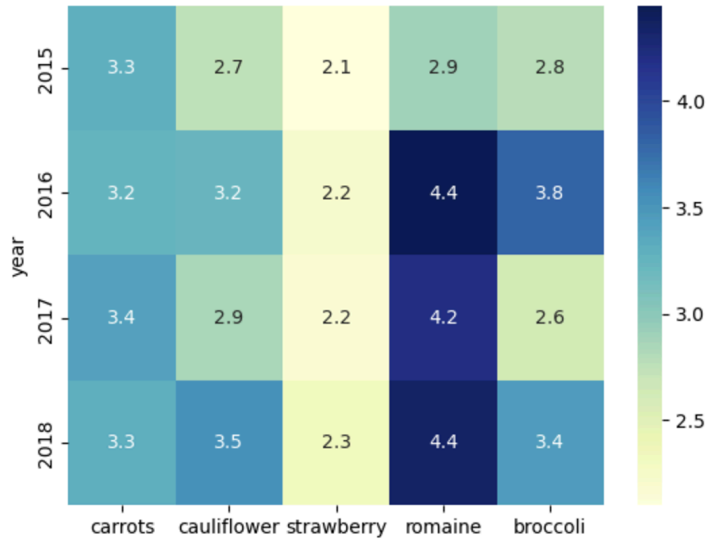


Chart 1: Trend of U.S. acreage for oranges decreasing over the 2010's, from the USDA Citrus Fruits 2016 Summary report: https://downloads.usda.library.cornell.edu/usda-esmis/files/j9602060k/7h149r77n/k643b339c/CitrFrui-09-12-2016.pdf

Although the Atlanta Terminal Market Fruit Prices report was not available prior to 2017, we were able to find Citrus Fruits 2016 Summary report from the USDA (see reference above). A key insight from this report is that **Florida orange production was down 16%, while California orange production was up 12%**, compared to the previous season. This supply constraint could be a potential contributing factor towards the high price in Atlanta (sources mainly Florida oranges due to regional proximity), compared to Los Angeles (sources mainly California oranges due to regional proximity).

Another key insight from the report is the gradual decrease of orange bearing acreage across the U.S., over the past decade (2006-2016). A University of Florida study from 2016 depicts the impact of a bacterial disease known as citrus greening or Huanglongbing (HLB) on the global citrus markup. The study shows that "since HLB was first found in 2005, orange acreage and yield in Florida have decreased by 26% and 42%, respectively." Given this 2016 report was the first "growers'-survey-based estimates of both the level of HLB infection in Florida and the impact of HLB on citrus operations in Florida", this could be another contributing factor towards the orange retail price spreads during this timeframe. **While there is no direct evidence or historic news pointing towards the Atlanta price elevation for oranges during that week in February 2016, we speculate the spread of HLB bacterial disease and constraint of Florida production may have contributed.**



Conclusions

Figure 6 highlights the mean Markup Ratio for each product over the 4 years analyzed. The highest Markups (3 of the 5 products) are seen in 2018, illustrating that the markup ratio has slowly increased from 2015, indicating a greater focus on profitability for produce producers.

Figure 6: Heatmap of the Mean Markup Ratio for each product across 2015-2018

Markup Ratio Conclusion: How does the Markup ratio vary by product and location?

The Markup Ratio analysis looks at the consistency of markups between cities for each different product. Carrots have the most predictable markup between cities, of just 30% of the mean, while broccoli is the least predictable with markup between cities being 100% of the mean.

The average markup ratio for romaine is the highest, meaning that romaine has the highest upcharge in price compared to the other 5 vegetables analyzed. The mean to standard deviation ratio is the smallest, around 0.36 overall, also making the price between cities more consistent.

Carrots had the 2nd highest average markup ratio, but with a much lower mean to standard deviation ratio, showing more the most consistent markup pricing of all the products analyzed - around 0.30. This indicates the carrots have a consistent and predictable markup between cities

Markup Ratio: Overall mean, standard deviation and standard deviation / mean between cities, by Product

Product	Market Ratio: Mean	Market Ratio: Standard Deviation	Market Ratio: Standard Deviation / Mean
broccoli	0.634311	0.633945	0.999423
strawberry	0.695944	0.361854	0.519948
cauliflower	0.763134	0.286323	0.375193
carrots	0.883369	0.266731	0.301947
romaine	1.371330	0.500457	0.364943

Figure 7: Final output of the Market Ratio analysis showing the overall mean, standard deviation and standard deviation / mean by product.

Finally, Broccoli had the highest mean to standard deviation ratio, but the lowest average markup ratio indicating that while the markup for broccoli is low, the price for broccoli varies almost 100% making it challenging to predict what the final price will be for each city. Further research indicates that Broccoli did have an outlier of $11.74 price, and therefore this data could be skewed. Unfortunately, we don't have enough data at this time to exclude this datapoint.

City Retail Price Spreads:

Based on the dataset we explored, there can be a large spread in retail prices for produce across different cities and regions. As studied in the case of 2016 orange prices, this can be the effect of source location, supply chain constraints, seasonality, or other factors. Taking these variables into account could aid in forecasting future produce prices and market effects.

Another key takeaway from this analysis is the importance of continually evaluating the data quality for the dataset of focus. Based on our investigation of the 2018 broccoli crowns price

data, we concluded there is likely a data entry error that may lead to potential misportrayal of produce retail price spreads. From the learnings above, we are able to draw the following conclusions:

- Oranges had the highest price spread in February 2016, potentially due to HLB bacterial disease negatively impacting orange acreage and production in Florida. But the highest price spread product varied by year
- There is no pattern or trend for which city contributed the most towards the highest or lowest price. Rather, it is moreso based on the product and the product's market factors instead.

References:

Impact of Citrus Greening on Citrus Operations in Florida: FE983/FE983, 2/2016. https://journals.flvc.org/edis/article/view/127756?articlesBySameAuthorPage=5

Produce Price Index Dataset, USDA Market News Reports for Farm Gate Prices & U.S. Marketing Services for Retail Prices. https://www.kaggle.com/datasets/everydaycodings/produce-prices-dataset

Veggie Tales for Fruitful Discussion

Exploratory Analysis of Produce Prices for 2015-2019

Amy Steward, Justin Nhan, Kevin Pradjinata

Overall Question

How does purchase location and farm price influence purchase cost for produce items?

Sub-Questions

Markup Ratio - Deep Dive into specific products:

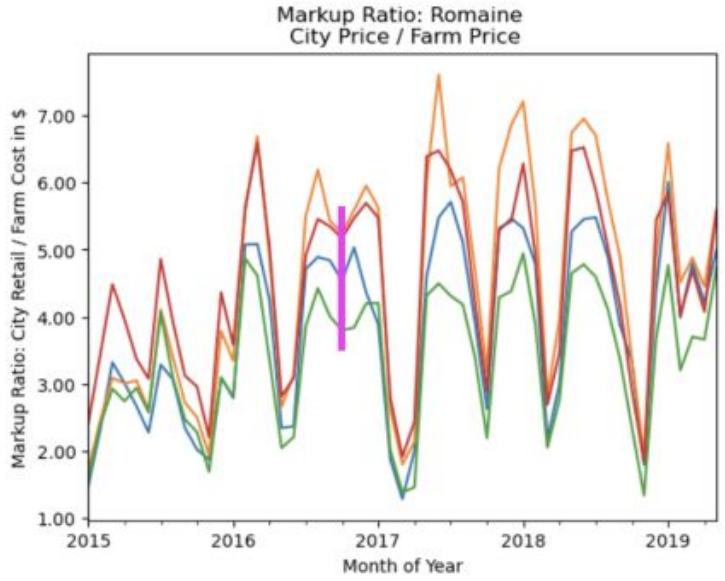
- Has the markup ratio for produce risen since 2015, resulting in higher retail store profits while increasing the financial burden on consumers?
- How does the Markup ratio vary by product and location?

Which produce had the highest price spread across cities each year, and why?

- Is it the same product each year?
- Is it the same city with the highest & lowest price each year?

Assumptions

- Data is accurate
- 2015-2019 data is recent enough to answer our questions
- Prices are properly unit-based, per USDA's definitions

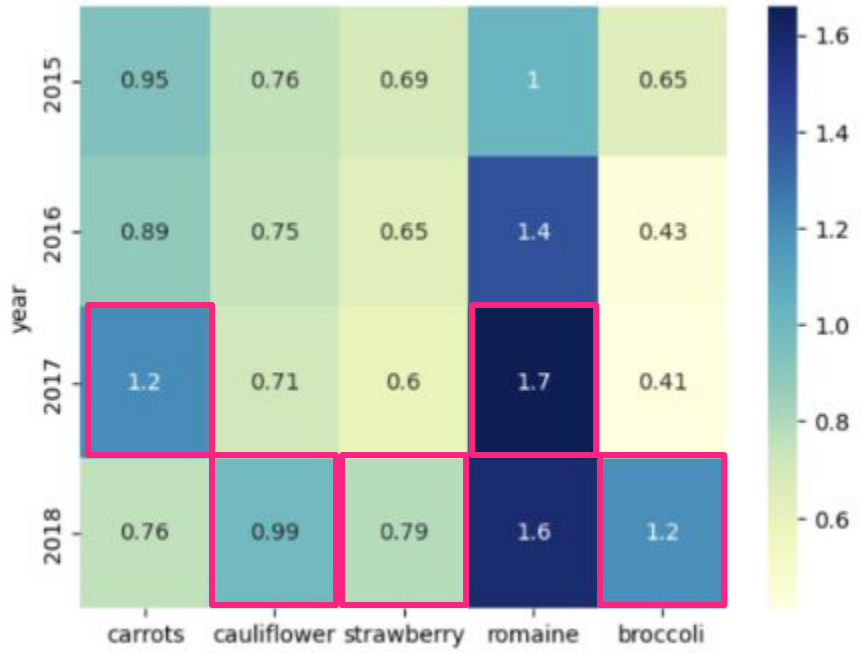


Steps to Answer Question: Markup Ratio

1) Clean the Data
2) High Level Data Analysis “Getting acquainted”
3) Calculated the Markup Ratio (Retail / Cost) and created charts comparing the Markup Ratio for each product
4) Calculated the Intra Month Markup Ratio to look at variance between cities

Overall Markup Ratio Analysis: Conclusions

Has the markup ratio for produce risen since 2015?



The markup ratio has slowly increased from 2015 and peaking towards 2018, resulting in higher retail store profits and increasing the financial burden on consumers

greater

Overall Markup Ratio Analysis: Conclusions

How does the Markup ratio vary by product and location?

Markup Ratio: Overall mean, standard deviation and standard deviation / mean between cities, by Product

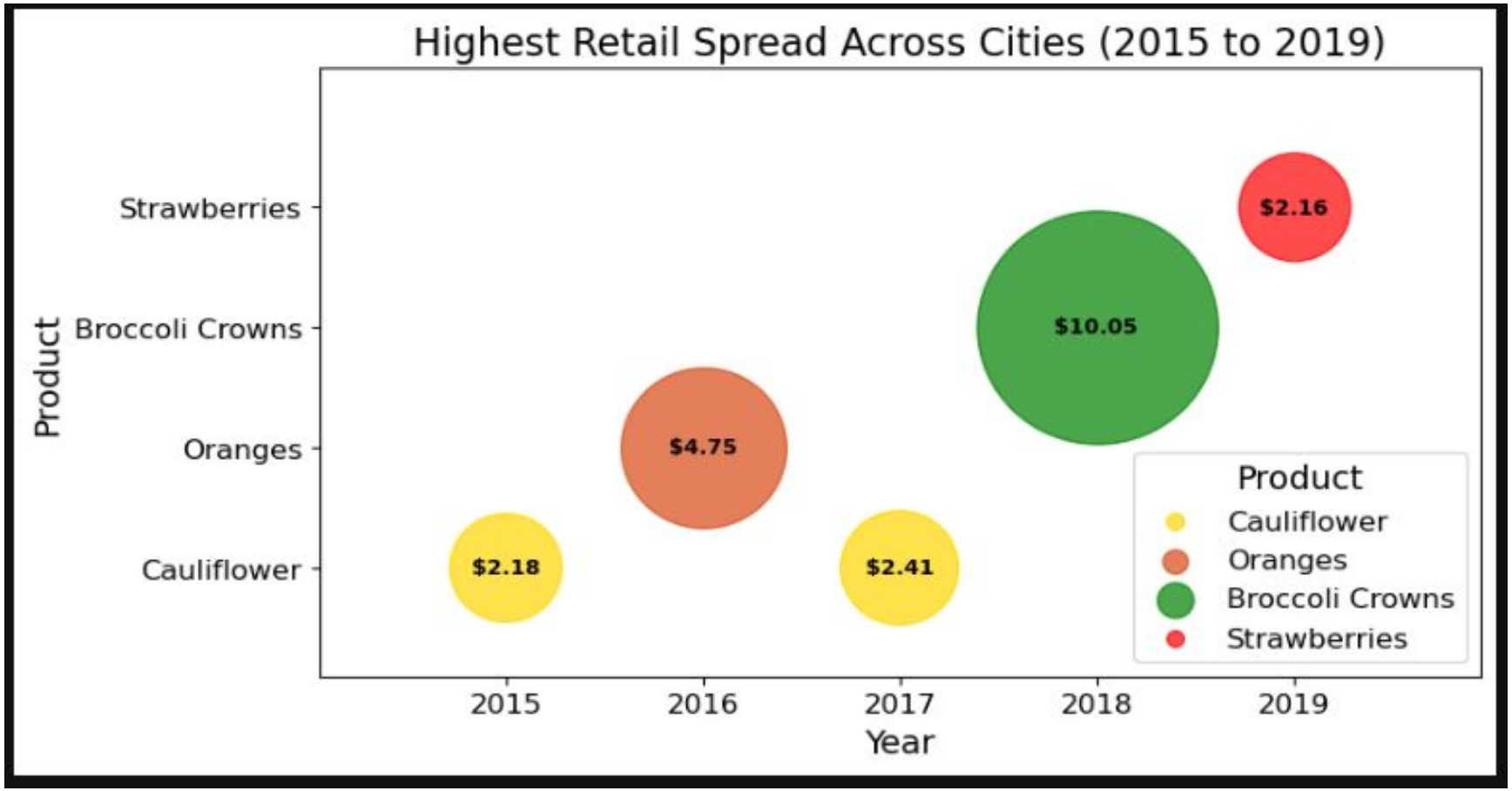
Product	Market Ratio: Mean	Market Ratio: Standard Deviation	Market Ratio: Standard Deviation / Mean
broccoli	0.634311	0.633945	0.999423
strawberry	0.695944	0.361854	0.519948
cauliflower	0.763134	0.286323	0.375193
carrots	0.883369	0.266731	0.301947
romaine	1.371330	0.500457	0.364943

Least Predictable Price: Broccoli

Most Predictable Price: Carrots

Highest markup: Romaine

City Retail Price Spread Analysis



1. Determine highest and lowest prices for each product, across all city prices

2. Calculate retail price spread:
   - The difference between highest and lowest prices

3. Bubble Chart:

   Products with highest retail price spread (2015-2019)

Let's further look into 2018 Broccoli crowns and 2016 Oranges…

Analysis: 2018 Broccoli Crowns Price Spread

date	atlantaretail	chicagoretail	losangelesretail	newyorkretail
2018-03-25	1.69	1.90	1.99	1.61
2018-03-18	1.69	1.90	1.99	11.74
2018-03-11	1.69	1.95	1.99	1.91

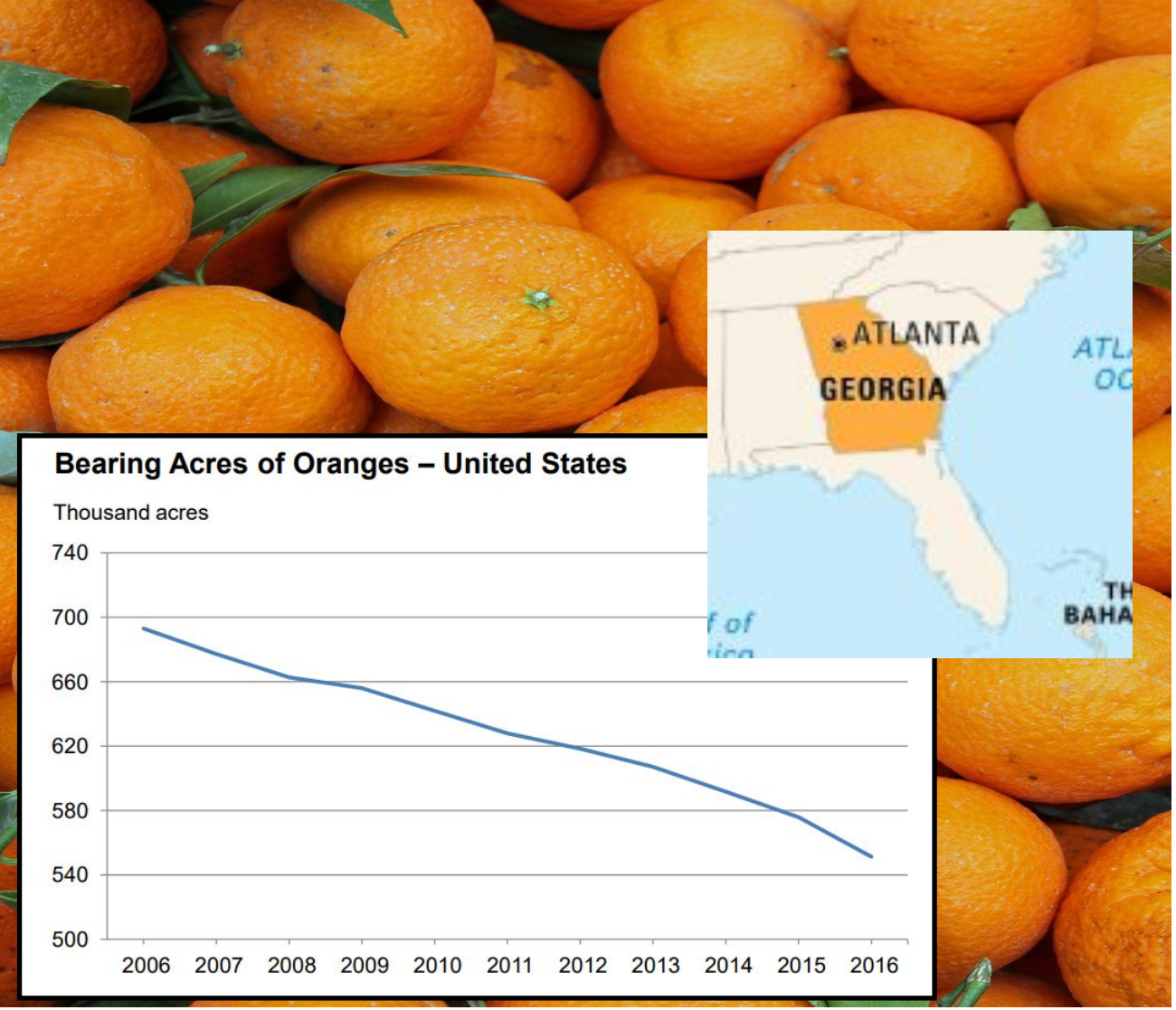
---BROCCOLI: MARKET ABOUT STEADY. cartons CA bchd 14s 14.00-16.00 occas 13.00 fineappear 17.00-18.00 bchd 18s 16.00-18.00 occas 13.00 Crown Cut Short Trim 14.00-16.00 FL bchd 14s 14.00 MX Crown Cut Short Trim 8.00-10.00 No Ice 12.00-14.00 mostly 12.00 fr appear fr cond 5.00-6.00 Baby Hybrid Type bchd 18s offerings insufficient to quote

- Price Spread was $10.05 across cities…..too good to be true?
- Evaluating the $11.74 price in New York
- Data Accuracy Investigation:

  Reports and March 2018 data show broccoli market and prices were stable.

  Potential data entry error identified.

Analysis: 2016 Oranges Price Spread



+ Price Point Analysis

  Price Spread of $4.75 across sites in Feb. 2016

  Inspection of the $5.84 price point in Atlanta

+ Supply Constraints

  Investigation of factors limiting Florida orange production, overall 16% decline (YoY)

+ HLB Bacterial Disease

  Impact of HLB on Florida orange acreage:

"since HLB was first found in 2005, orange acreage and yield in Florida have **decreased by 26% and 42%**, respectively."
- Impact of Citrus Greening on Citrus Operations in Florida

Conclusions

Markup Ratio Trends

- Gradual increase in markup ratio since 2015
- Indicates a greater focus on profitability

Predictability across cities

- Least Predictable Price: Broccoli
- Most Predictable Price: Carrots
- Highest markup: Romaine

- Oranges had the highest price spread in February 2016,
  - potentially due to HLB bacterial disease impacting orange acreage and production in Florida.
- The highest price spread product varied by year
- There is no pattern or trend for which city contributed the most towards the highest or lowest price.
- Rather, it is moreso based on each product's market factors (source location, supply chain constraints, seasonality)
- Lastly, data quality should always be continually assessed….

Q&A Preparation

References

Impact of Citrus Greening on Citrus Operations in Florida: FE983/FE983, 2/2016. https://journals.flvc.org/edis/article/view/127756?articlesBySameAuthorPage=5

Produce Price Index Dataset, USDA Market News Reports for Farm Gate Prices & U.S. Marketing Services for Retail Prices. https://www.kaggle.com/datasets/everydaycodings/produce-prices-dataset