Millennials and their Avocado Toast!

CSC466 Final Project Report

Sarah Bae | Wesley Benica | Roxanne Miller | Isabel Rico

shbae | wbenica | rmille60 | irico

Abstract

Our goal with this project was to see if we could find relationships between avocado prices and popularity with various traits about the area, such as homeownership rates, rainfall, or incomes. We used a variety of models from Random Forest for classification and Kmeans for clustering, and found that some factors which we expected to have less relationship, such as housing ownership and google trends, had stronger correlations than ones we expected to show strong correlations, such as rainfall and prices. Some of these correlations are likely for reasons beyond avocado trends, but they were interesting to find regardless and can be further investigated in the future.

Introduction/Approach

In 2017, Australian billionaire Tim Gurner went on the show *60 Minutes*, where he claimed that the millennial generation would be more able to afford houses if they bought less avocado toast. This inspired us to explore data to find information about if there may, as Gurner believes, exist any connections between avocado-related spending habits and the average person's wealth in that area, and about what factors, if any, can be used to predict avocado price or popularity trends.

Research Questions

Our goal was to see if we could find a relationship between avocado prices and popularity with various economic trends. The trends we decided to focus on looking into were if there was a relationship between avocado prices and monthly rainfall in the central/southern parts of california, avocado prices and state minimum wages, the popularity of avocado toast base and housing prices, and the popularity of avocados and their prices.

Datasets

Average Avocado Prices in US Regions

https://www.kaggle.com/neuromusic/avocado-prices
Three years (2015-2018) of data on sales of Hass avocados in the United States, including average price, number, and volume sold, as well as the metropolitan region the data was taken from. Each row is one observation (there are approximately 4 a month per region) with data about the observation's type (organic versus conventional), average price(per avocado), region(region in the US, later to be mapped to state abbreviations), and total volume purchased per recorded observation.

US Census Bureau Housing Vacancies and Homeownership

https://www.census.gov/housing/hvs/data/rates/tab2_state05_2019_hvr.xlsx
https://www.census.gov/housing/hvs/data/rates/tab2_state05_2019_hvr.xlsx
The Current Population Survey/Housing Vacancy Survey (CPS/HVS) is a survey of around 72,000 housing units in all 50 states plus Washington, D.C. For this study, we focused on two of the datasets collected from this survey, *Homeowner Vacancy Rates by State: 2005-present* and *Homeownership Rates by State: 2005-present*. The vacancy dataset refers only to home vacancies--there is a separate dataset (not used) that covers rental property vacancies.

Modified Avocado Sales Information and Minimum Wage

This dataset uses: https://www.kaggle.com/neuromusic/avocado-prices
Dataset name: avoWagesFixed.csv

This dataset was created by us using the avocado price dataset in conjunction with a dataset outlining US Minimum Wages by state from 2015 - 2018. We grouped the minimum wages into numerical categories from 7.25 - 11.25 in 35 cent increments. We then appended the minimum wage for a certain year in a certain state to the row in the avocado csv with the same date and the correct reason (i.e. Minimum wage in 2018 in California was appended to all rows in the avocado dataset with year 2018 and region Los Angeles, San Francisco, San Diego, and Sacramento).

In this dataset, date represents the date the rest of the data in that row was collected. Average Price represents the average price of a single avocado at the date data was collected. Total Volume represents the total number of avocados sold. There are three columns with numerical names. 4046, 4225, and 4770. The data in these columns represents the number of avocados with that PLU sold. Total bags represents the total number of avocado bags sold. This number is broken down to the number of small, large, and xl bags of avocados sold. These numbers are in the Small Bags, Large Bags, and XLargeBags columns. The values in the type column represent if the avocado was organic or not. Year represents the year in which the data was collected and region represents the region the data in that row applies to. The new column, wage, represents the minimum wage in that region in that year.

When analyzing this dataset we removed the XLargeBags column because very few of the data points had a nonzero values in this column and we though including it might cause overfitting for the decision trees. We also removed the date column because the minimum wage data was only granular enough to reflect the year, not specific months or days.

Avocado Prices vs Rainfall ()

Google Trends Data

Google Trends provides data on the relative popularity of search terms by region (down to the city-level) and over time (from 2004-present). The data is normalized to account for variations in population and variations in search volume over time then scaled from 0 to 100.

Methods

Avocado prices versus rainfall

Data Files

apvrainfall.csv

Month	Numerical month from 1-12 representing when the avocado's price was recorded
Year	Year that the price was recorded
Type	'Organic' or 'conventional' avocados
PriceRange	'v_cheap 'cheap 'moderate 'expensive 'v_expensive' as categorical options V_cheap and v_expensive are lowest and highest 15% of prices, cheap and expensive are avocados priced in the 15-30% and 60-85% quartile, and moderate is avocados in the middle 30%-60% of prices.
AvgRainfall(1, 3, and 6 mo)	Average rainfall recorded *n* months before **Month**, when the price is actually recorded

To see if we could find a relationship between rainfall and avocado prices, we used the "California Nevada River Forecast Center"'s monthly rainfall data from 2015-2018 with Kaggle's avocado price dataset.. Because many of the US's avocados are grown in the range from Monterey to San Diego, we selected only rainfall data from Central to Southern California and found the average for each month. Next, we wanted to be able to see if results were improved by using rainfall data from a few months prior to the actual pricing so we added various columns for different 'offsets' of how long ago the rain data was for. We also replaced the continuous price column with a categorical price column, using categories from 'cheapest' to 'most expensive' depending on where the price landed in the data. This resulted in our final dataset used, *'apvrainfall'*.

We decided to build a categorical random forest model because we've found that is a pretty robust model for limited amounts of data in the past, and we are able to use a large number of trees and run a lot of different hyperparameters because the runtime is pretty quick. To parameterize it we use 5 fold cross validation with scikit's 'RandomizedSearchCV' function, which creates multiple random forests using a random set of parameters from a range that we provided.

Avocado prices versus minimum wage

Data Files

avoWagesFixed.csv
To find a relationship between the data in the avocado prices dataset and minimum wage data we decided to use both a single decision tree as a classifier and a random forest classifier. We were interested in finding out if prices of avocados in a certain region, numbers of avocado sales, and type of avocado being sold could indicate what a region's minimum wage was. To create the training data, we shuffled the data points and selected an 80/20 split of the original data. We started with a single decision tree using bootstrapping with replacement for the training data. To get the best model we used cross-validated grid-search on ranges of hyperparameters.

The hyperparameters for the decision tree were tree depth, split function, minimum number of samples required to split an internal node, and maximum number of features considered when looking for the best split. For tree depth we tried values from 10 - 40, for max features we tried values from 2 - 9, and for min samples to split on we tried values from 3 - 50.

The random forest classifier was trained in the same way as the single decision tree. We tried all of the same hyperparameters as we tried with the single decision tree with the addition of number of trees in the forest.
For the number of trees in the forest we tried values from 5 - 25.

Homeownership versus Google Trends ("Avocado" and "Avocado Toast" Searches)

Data Files

trends_avocado.csv, trends_avocado_toast.csv
Each row of these datasets contains the monthly search popularity of the terms 'avocado' and 'avocado toast' for the state listed in the first columns. As part of the preprocessing of the data, columns were aggregated into quarters to match the homes datasets. The datapoints are between 0 and 100, which is a scaled value representing the relative popularity of the search terms, and which has been normalized to account for differences in population and search volume over time.
We collected this data from Google Trends using the Python module pytrends, which is licensed for use under the Apache 2.0 license.

homeowner_vacancy_rates.csv, *homeownership_rates_by_state.csv*
These datasets are laid out similarly to the *trends* datasets, though the columns represent quarterly data, rather than monthly. The datapoints are percentages of housing vacancies and homeownership respectively.

K-Means Clustering

We used k-means clustering to determine whether states that have higher interest in avocados and avocado toast have lower homeownership rates and/or higher homeowner vacancies. To see if there was a correlation, we used k-means clustering to group states quarter-by-quarter based on their searches for either the term “avocado” or “avocado toast” and then created clusters of states based on either their homeownership or vacancy rates. For each quarter, we then matched the clusters based on search terms with the clusters based on housing data that had the greatest overlap and calculated the percent overlap between the clusters.

To generate the clusters, we used scipy’s kmeans2 function. The function uses Euclidean distance to measure distances from centroids; of the initial centroid selection methods available, we chose the ‘random’ option, which “generate[s] k centroids from a Gaussian with mean and variance estimated from the data.” Through experimentation, we chose a k-value of 4; higher values of k resulted in many empty clusters, while lower values created lopsided clusters (one large cluster, and one or two very small clusters).

Avocado prices versus Google trends

Data Files

Trends_avocado.csv

See above in “Homeownership versus Google Trends”

avoPricesNumerical.csv

To use the avocado prices dataset, we one-hot encoded the “type” attribute and stripped the dates of each recording to the month, as the year is also recorded. Then, we grouped the dates of each datapoint by month, making the dataset completely numerical, save for the class variable, region. We also mapped regions to their respective states to make the data match up with the data per state in the google search dataset. With this dataset, we use all attributes, month, year, average price, total volume, numbers of avocados sold with PLU’s 4046, 4225, and 4770, total bags sold, small bags sold, large bags sold, and the one-hot encoded type: conventional or organic, to be part of the euclidean distance function in our clustering.

Clustering

To see if there might be an association between internet searches for “avocado” and prices of avocados by state, we clustered avocado prices, clustered internet searches for avocados, and evaluated the clusters to determine if the same states appear in clusters together in both datasets. We used both hierarchical agglomerative clustering and k-means clustering to see if one clustering method’s results will match the other’s. With 4134 datapoints in avocado prices and 51 datapoints in google trends, we pick a common k of 10 clusters. We calculate the distance between two points or clusters with the euclidean form, because we are interested in the degree of difference in magnitude. We used clustering to explore this question because we were interested in whether or not same groups of regions would appear in similar avocado prices and similar avocado searches.

Results

Avocado prices versus rainfall

Avocado Price Predictions (No Rainfall Data)

Confusion Matrix	cheap	expensive	moderate	v_cheap	v_expensive
cheap	240	4	223	176	18
expensive	1	143	356	1	211
moderate	153	123	1266	93	193
v_cheap	144	4	123	428	9
v_expensive	0	87	185	0	382

Statistics	Precision (%)	Recall (%)	F-Measure (%)
cheap	44.61	36.31	40.033
expensive	39.61	20.08	26.654
moderate	58.8	69.26	63.602
v_cheap	61.32	60.45	60.882
v_expensive	46.99	58.41	52.079
avg	**50.27**	**48.9**	**48.65**

Gini index:

Month	0.355751
Year	0.288920
TypeOrganic	0.191782
TypeConventional	0.163547

Avocado Price Predictions (1st, 3rd, and 6th month previous Rainfall Data)

Confusion Matrix	cheap	expensive	moderate	v_cheap	v_expensive
cheap	336	9	151	120	24
expensive	3	99	365	0	201
moderate	167	74	1170	47	187
v_cheap	186	1	94	359	10
v_expensive	0	51	166	0	392

Statistics	Precision	Recall	F-Measure
cheap	48.55	52.5	50.45
expensive	42.31	14.82	21.951
moderate	60.12	71.12	65.163
v_cheap	68.25	55.23	61.054
v_expensive	48.16	64.37	55.095
avg	**53.48**	**51.61**	**50.743**

Gini index:

Month	0.095530
Year	0.140782
TypeOrganic	0.211110
TypeConventional	0.194976
AvgRainfall(-6mo)	0.104398
AvgRainfall(-3mo)	0.112395
AvgRainfall(-1mo)	0.140808

What we found was that using previous rainfall data did improve performance slightly, by about 3% on the f-measure, this is a small enough change that we are not sure if it's attributed to chance or if it is an actual improvement due to the rainfall data being included. The model did find that of the rainfall info, the data from the last month was the most helpful according to the Gini index. Adding rain data also decreased the importance of the Month variable by a significant amount, likely because the month can be predicted as well by comparing rainfall data. The model also significantly struggled on predicting 'expensive' in all cases, we think because it had the least data to begin with.

Avocado prices versus minimum wage

We found that regardless of how we tuned the hyperparameters, there was an upper bound of 80% accuracy for both the single decision tree and the random forest classifier. Further, the random forest classifier did not perform any better than the single decision tree on average. One concern with decision trees and random forest is overfitting. However, when given a range of hyperparameters to choose, the best model was always trained on as much of the training data as possible. For example with the maximum number of features being with 2 - 9 the best model was trained using 8 features. Because of this, it didn't seem like the model was overfitting, rather that there was not enough data. Another reason for this upper bound on accuracy could be that avocado sales data is not the best predictor of minimum wage. We also found that models trained using Gini measure as a measure of split performed slightly better than those trained using entropy as a measure of split. Besides accuracy, the main difference in the models trained using each of those split measures was that models trained using Gini measure required more trees in the forest but models trained using entropy required greater tree depth. Because the classes for the data points were wages and therefore floats (dollars and cents) we had to use the scikit learn fit_transform function on our training and testing labels. This made them numerical labels, so the actual wages are not available as row and column labels for the confusion matrix.

Gini Impurity:

Decision Tree:

Accuracy: 0.7809481015502134
Best Training Score: 0.7862841519709461
Best Parameters: {'max_depth': 28}

Random Forest:

Accuracy: 0.7667939788811503
Best Training Score: 0.7845996106383449
Best Parameters: {'max_depth': 28, 'n_estimators': 10}

Impurity:

Decision Tree:

Accuracy: 0.7582565715569535
Best Training Score: 0.7750508157978601
Best Parameters: {'max_depth': 32}

Random Forest:

Accuracy: 0.7753313862053471
Best Training Score: 0.780723490119238
Best Parameters: {'max_depth': 24, 'n_estimators': 6}

Decision Tree Confusion Matrix:

Predicted	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	All
True																	
0	1292	0	1	8	1	4	0	2	0	0	0	2	0	0	0	0	1310
1	0	160	0	0	0	1	0	2	0	0	0	0	0	0	0	0	163
2	0	0	256	1	0	0	0	0	0	0	0	0	0	0	0	0	257
3	11	0	0	501	89	41	0	3	0	0	0	3	0	0	0	0	648

4	7	1	0	105	278	1	0	0	0	0	0	0	0	0	0	392
5	3	0	0	43	1	73	0	0	0	0	0	1	0	0	0	121
6	5	0	0	1	0	0	79	5	0	0	0	0	0	0	0	90
7	2	3	0	3	0	0	2	250	40	0	0	43	0	0	0	343
8	1	0	0	0	0	0	0	66	19	0	1	0	0	0	0	87
9	0	0	0	0	0	0	0	0	0	134	0	16	0	0	4	154
10	1	0	0	0	0	0	0	0	0	0	30	0	67	16	0	114
11	3	0	0	3	1	3	0	45	0	25	0	134	23	0	108	345
12	2	0	0	0	0	0	0	0	0	0	88	13	4	0	0	107
13	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	18
14	2	0	0	0	1	0	0	1	0	0	0	122	0	0	66	43
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	1
All	1329	164	257	665	371	123	81	374	59	159	137	334	94	16	244	44

Random Forest Classifier Confusion Matrix:

Predicted	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	All
True																	
0	1328	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1330
1	1	176	0	0	0	0	0	0	0	0	0	0	0	0	0	0	177
2	7	0	238	1	0	0	0	0	0	0	0	0	0	0	0	0	246
3	15	0	0	521	87	29	0	0	0	0	0	0	0	0	0	0	652
4	3	0	0	73	275	1	0	0	0	0	0	0	0	0	0	0	352
5	5	0	0	29	3	94	0	0	0	1	0	0	0	0	0	0	132
6	0	0	0	0	0	0	79	2	0	0	0	0	0	0	0	0	81
7	4	0	0	5	0	0	2	263	50	0	0	37	0	0	0	0	361
8	1	0	0	0	0	0	0	50	32	0	1	0	0	0	0	0	84
9	1	0	0	0	0	1	0	0	0	115	0	13	0	0	0	0	130
10	0	0	0	0	0	0	0	0	1	0	34	0	90	18	0	0	143
11	13	0	0	0	0	1	0	35	0	21	0	152	22	0	94	0	338
12	0	0	0	0	0	0	0	0	0	0	66	15	19	0	0	0	100
13	0	0	0	0	0	0	0	0	0	0	19	0	0	5	0	0	24
14	6	0	0	0	0	0	0	0	0	0	0	113	0	0	67	57	243
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	11	58
All	1384	176	238	631	365	126	81	350	83	137	120	330	131	23	208	68	4451

Homeownership versus Google Searches for ‘Avocado’ and ‘Avocado Toast’

There appears to be a slight correlation between searches for ‘avocado’ and ‘avocado toast’ and lower homeownership/higher vacancy rates, although there is no evidence for causation in either direction. The data is also skewed, because avocado toast didn’t start gaining popularity until around 2014-15, which resulted in a large cluster where the search interest was zero for each state, which would then have high overlap with one of the homeownership clusters, even though there was no actual correlation. Even once ‘avocado toast’ rose in popularity, correlations varied widely, from as low as 0.15 up to 0.52.
It is most likely that any perceived correlation between the popularity of avocado toast and homeownership rates can either be explained by other factors (weather, location, existing wealth) or is

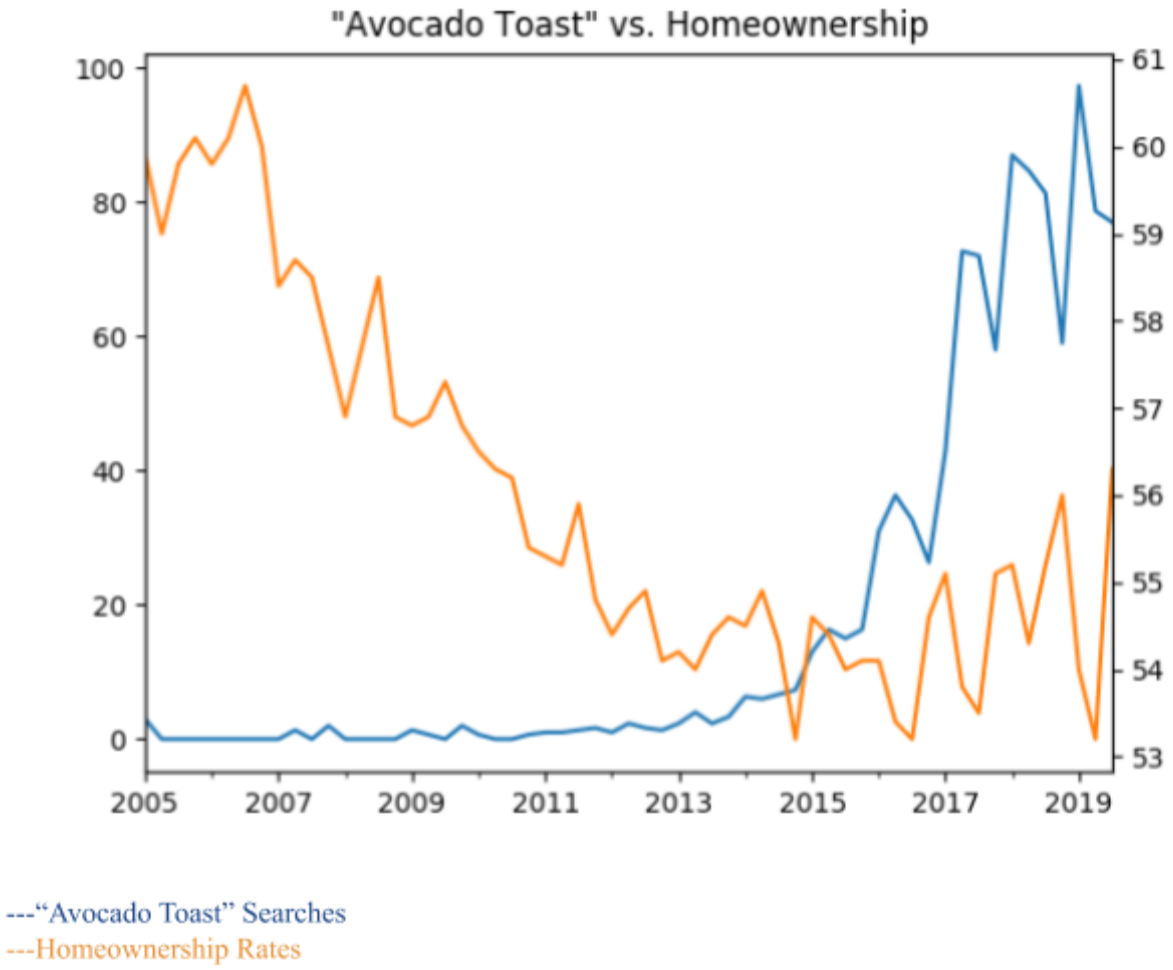
merely coincidental. Additionally, homeownership rates were already declining before “avocado toast” even became “a thing.”

Overall Results

Search Term	Housing Data	Accuracy
Avocado	Ownership	0.328
	Vacancies	0.322
Avocado Toast	Ownership	0.381
	Vacancies	0.334

Searches vs Homeownership Rates

California



Quarterly Accuracies

	2005Q1	2005Q2	2005Q3	2005Q4	2006Q1	2006Q2	2006Q3	2006Q4
avo_own	0.235	0.353	0.196	0.353	0.157	0.412	0.373	0.373
avo_vac	0.255	0.314	0.275	0.353	0.333	0.294	0.392	0.333
tst_own	0.549	0.431	0.569	0.51	0.529	0.51	0.451	0.451
tst_vac	0.314	0.353	0.353	0.294	0.275	0.294	0.392	0.314
	2007Q1	**2007Q2**	**2007Q3**	**2007Q4**	**2008Q1**	**2008Q2**	**2008Q3**	**2008Q4**
avo_own	0.451	0.294	0.275	0.275	0.471	0.392	0.412	0.196
avo_vac	0.392	0.314	0.373	0.294	0.255	0.471	0.294	0.431
tst_own	0.294	0.471	0.333	0.333	0.471	0.275	0.392	0.431
tst_vac	0.373	0.451	0.49	0.373	0.353	0.333	0.451	0.412
	2009Q1	**2009Q2**	**2009Q3**	**2009Q4**	**2010Q1**	**2010Q2**	**2010Q3**	**2010Q4**
avo_own	0.353	0.333	0.294	0.255	0.275	0.333	0.373	0.471
avo_vac	0.314	0.216	0.255	0.353	0.392	0.333	0.255	0.314
tst_own	0.451	0.431	0.353	0.51	0.373	0.49	0.431	0.431
tst_vac	0.275	0.412	0.275	0.333	0.333	0.471	0.49	0.196
	2011Q1	**2011Q2**	**2011Q3**	**2011Q4**	**2012Q1**	**2012Q2**	**2012Q3**	**2012Q4**
avo_own	0.353	0.333	0.353	0.294	0.294	0.333	0.373	0.196
avo_vac	0.412	0.392	0.373	0.373	0.392	0.353	0.294	0.392
tst_own	0.412	0.451	0.196	0.373	0.353	0.49	0.294	0.431
tst_vac	0.471	0.353	0.314	0.373	0.275	0.373	0.294	0.275
	2013Q1	**2013Q2**	**2013Q3**	**2013Q4**	**2014Q1**	**2014Q2**	**2014Q3**	**2014Q4**
avo_own	0.275	0.314	0.275	0.353	0.314	0.275	0.333	0.373
avo_vac	0.353	0.294	0.333	0.294	0.294	0.333	0.255	0.235
tst_own	0.294	0.255	0.412	0.412	0.353	0.333	0.333	0.275
tst_vac	0.353	0.333	0.275	0.275	0.314	0.314	0.333	0.255
	2015Q1	**2015Q2**	**2015Q3**	**2015Q4**	**2016Q1**	**2016Q2**	**2016Q3**	**2016Q4**
avo_own	0.353	0.235	0.255	0.392	0.333	0.373	0.294	0.275
avo_vac	0.373	0.314	0.412	0.216	0.275	0.235	0.216	0.314
tst_own	0.216	0.353	0.392	0.314	0.333	0.412	0.353	0.471

tst_vac	0.294	0.373	0.275	0.412	0.275	0.373	0.392	0.275
	2017Q1	**2017Q2**	**2017Q3**	**2017Q4**	**2018Q1**	**2018Q2**	**2018Q3**	**2018Q4**
avo_own	0.333	0.333	0.392	0.529	0.275	0.373	0.196	0.392
avo_vac	0.314	0.314	0.333	0.333	0.275	0.275	0.412	0.235
tst_own	0.314	0.157	0.373	0.412	0.294	0.373	0.353	0.275
tst_vac	0.294	0.314	0.294	0.392	0.235	0.255	0.294	0.216
	2019Q1	**2019Q2**	**2019Q3**					
avo_own	0.373	0.471	0.255					
avo_vac	0.314	0.353	0.314					
tst_own	0.392	0.294	0.235					
tst_vac	0.431	0.235	0.314					

Clusters 2019Q3 (Sample)

'Avocado' Search (S) vs Homeownership (H)

Cluster 1

```
S:                FL    HI
H: AK AZ AR CO CT FL GA HI IL LA MA MT NJ NM NC ND OK OR RI TX WA WI
```

Cluster 2

```
S: AK       ID          KY       MS MT NE ND          SD       VT    WV WY
H:    AL DE    IN IA KS KY MD MO       NE    OH PA SC SD TN UT VT VA    WY
```

Cluster 3

```
S: CA CO    GA LA MA MN NV NJ NM NY OH OR PA TX UT VA WA
H: CA    DC             NV       NY
```

Cluster 4

```
S: AL AZ AR CT DE DC    IL IN IA KS ME MD MI    MO    NH NC OK RI SC TN WI
H:                  ID             ME    MI MN    MS NH                   WV
```

'Avocado' Search (S) vs. Vacancy Rate (V)

Cluster 1

```
S:             FL    HI
V: AL AZ CT DE FL GA    IN ME MD MO NJ NM NC SC TX
```

Cluster 2

```
S: AK       ID       KY    MS MT NE ND          SD    VT    WV WY
V:    AR CA ID IL KS KY LA MS MT       NY OH PA    UT VT VA WV WY
```

Cluster 3

S: CA CO GA LA MA MN NV NJ NM NY OH OR PA TX UT VA WA

V: CO DC HI IA MA MI MN NE NV NH OR RI SD WA WI

Cluster 4

S: AL AZ AR CT DE DC IL IN IA KS ME MD MI MO NH NC OK RI SC TN WI

V: AK ND OK TN

'Avocado Toast' (T) vs Homeownership (H)

Cluster 1

T: AL ND WV

H: AL DE IN IA KS KY MD MO NE OH PA SC SD TN UT VT VA WY

Cluster 2

T: CT DE HI IL KS KY ME MI MN MS MO MT NE NV NJ NY NC

H: AK AZ AR CO CT FL GA HI IL LA MA MT NJ NM NC

T: OH PA TN VA WA WI

H: ND OK OR RI TX WA WI

Cluster 3

T: AK AR CO DC ID IN IA LA NH NM OK RI SD VT

H: ID ME MI MN MS NH WV

Cluster 4

T: AZ CA FL GA MD MA OR SC TX UT

H: CA DC NV NY

'Avocado Toast' (T) vs Vacancies (V)

Cluster 1

T: AL ND WV

V: AL AZ CT DE FL GA IN ME MD MO NJ NM NC SC TX

Cluster 2

T: CT DE HI IL KS KY ME MI MN MS MO MT NE NV NJ NY NC OH PA TN

V: AR CA ID IL KS KY LA MS MT NY OH PA UT VT

T: VA WA WI

V: VA WV WY

Cluster 3

T: AK AR CO DC ID IN IA LA NH NM OK RI SD VT

V: CO DC HI IA MA MI MN NE NV NH OR RI SD WA WI

Avocado prices versus google trends

K-means clustering (*k = 10)*

Avocado Prices

Cluster 0:['CA' 'TX' 'OR' 'IL' 'CO' 'PA' 'OH' 'MD' 'NY' 'MI' 'WA' 'MA' 'NM' 'ME' 'GA' 'NE' 'FL' 'CT' 'SC' 'AZ' 'VA' 'NC' 'NV' 'TN']

Cluster 4:['FL' 'CA' 'PA' 'NY' 'NC' 'VA' 'MI' 'OH' 'WA' 'ID' 'IN' 'KY' 'LA' 'MO' 'TN' 'NV' 'AZ' 'SC' 'CT' 'TX' 'GA' 'ME' 'MA' 'NM' 'MD' 'IL']

Cluster 2:['FL' 'NC' 'CA' 'PA' 'MI' 'NV' 'OH' 'LA' 'SC' 'CT' 'ME' 'VA' 'TN' 'GA' 'MO' 'IN' 'WA' 'MA' 'OR']

Cluster 1:['CA' 'FL' 'MA' 'OR' 'MD' 'WA' 'IL' 'CO' 'GA' 'NM' 'PA' 'ME' 'SC' 'MI' 'TX' 'AZ' 'CT' 'LA' 'NY']

Cluster 6:['NY' 'CA' 'WA' 'TX' 'ID' 'KY' 'NE' 'FL' 'PA' 'ME' 'MI' 'CO' 'IL' 'MD' 'OR' 'VA' 'NM']

Cluster 7:['MI' 'CA' 'NY' 'IN' 'NC' 'PA' 'TX' 'OH' 'VA' 'KY' 'CT' 'ME']

Cluster 8:['CA' 'TX' 'FL' 'ME']

Cluster 9:['FL' 'VA' 'TX' 'MO' 'OH' 'CA' 'TN' 'PA' 'WA' 'ME' 'ID' 'NC' 'NY' 'LA' 'MI' 'KY' 'NV' 'IN']

Cluster 5: ['TX' 'NY' 'AZ' 'NE' 'NM' 'CA' 'CO' 'IL' 'MD' 'OR' 'WA' 'FL']

Cluster 3:['CA' 'MI' 'TX' 'ME' 'FL' 'NE' 'NY']

Google Searches for “avocado”

Cluster 1:['AL 'CA 'CO 'CT 'DE 'FL 'GA 'LA 'MD 'MA 'MN 'NV 'NJ 'NM 'NY 'TN 'TX 'UT 'WA']

Cluster 7:['AZ 'AR 'IL 'IN 'IA 'KS 'KY 'MO 'NC 'OR 'RI 'SC 'VA 'WI']

Cluster 9:['ID 'ME 'NE 'NH 'SD']

Cluster 6:['MI 'OH 'OK 'PA']

Cluster 0:['DC']

Cluster 8:['HI']

Cluster 3:['MS 'MT 'VT 'WY']

Cluster 2:['ND']

Cluster 4:['AK']

Cluster 5:['WV']

Avocado Prices

Cluster 0:['CA' 'TX' 'OR' 'IL' 'CO' 'PA' 'OH' 'MD' 'NY' 'MI' 'WA' 'MA' 'NM' 'ME' 'GA' 'NE' 'FL' 'CT' 'SC' 'AZ' 'VA' 'NC' 'NV' 'TN']

Cluster 4:['FL' 'CA' 'PA' 'NY' 'NC' 'VA' 'MI' 'OH' 'WA' 'ID' 'IN' 'KY' 'LA' 'MO' 'TN' 'NV' 'AZ' 'SC' 'CT' 'TX' 'GA' 'ME' 'MA' 'NM' 'MD' 'IL']

Cluster 2:['FL' 'NC' 'CA' 'PA' 'MI' 'NV' 'OH' 'LA' 'SC' 'CT' 'ME' 'VA' 'TN' 'GA' 'MO' 'IN' 'WA' 'MA' 'OR']

Cluster 1:['CA' 'FL' 'MA' 'OR' 'MD' 'WA' 'IL' 'CO' 'GA' 'NM' 'PA' 'ME' 'SC' 'MI' 'TX' 'AZ' 'CT' 'LA' 'NY']

Cluster 6:['NY' 'CA' 'WA' 'TX' 'ID' 'KY' 'NE' 'FL' 'PA' 'ME' 'MI' 'CO' 'IL' 'MD' 'OR' 'VA' 'NM']

Cluster 7:['MI' 'CA' 'NY' 'IN' 'NC' 'PA' 'TX' 'OH' 'VA' 'KY' 'CT' 'ME']

Cluster 8:['CA' 'TX' 'FL' 'ME']

Cluster 9:['FL' 'VA' 'TX' 'MO' 'OH' 'CA' 'TN' 'PA' 'WA' 'ME' 'ID' 'NC' 'NY' 'LA' 'MI' 'KY' 'NV' 'IN']

Cluster 5: ['TX' 'NY' 'AZ' 'NE' 'NM' 'CA' 'CO' 'IL' 'MD' 'OR' 'WA' 'FL']

Cluster 3:['CA' 'MI' 'TX' 'ME' 'FL' 'NE' 'NY']

Google Searches for "avocado toast"

Cluster 9:['AL', 'DC', 'WV']

Cluster 4:['AK', 'AR', 'HI', 'IA', 'NV', 'NH', 'NM']

Cluster 0:['AZ', 'DE', 'FL', 'GA', 'KS', 'MD', 'MI', 'MN', 'MS', 'MO', 'NJ', 'NC', 'OH', 'OR', 'PA', 'SC', 'TX', 'UT', 'VA', 'WA']

Cluster 5:['CA', 'IL', 'MA']

Cluster 7:['CO', 'IN', 'ME', 'MT', 'ND', 'OK', 'RI', 'SD', 'TN', 'WI']

Cluster 1:['CT', 'NY']

Cluster 2:['ID', 'LA']

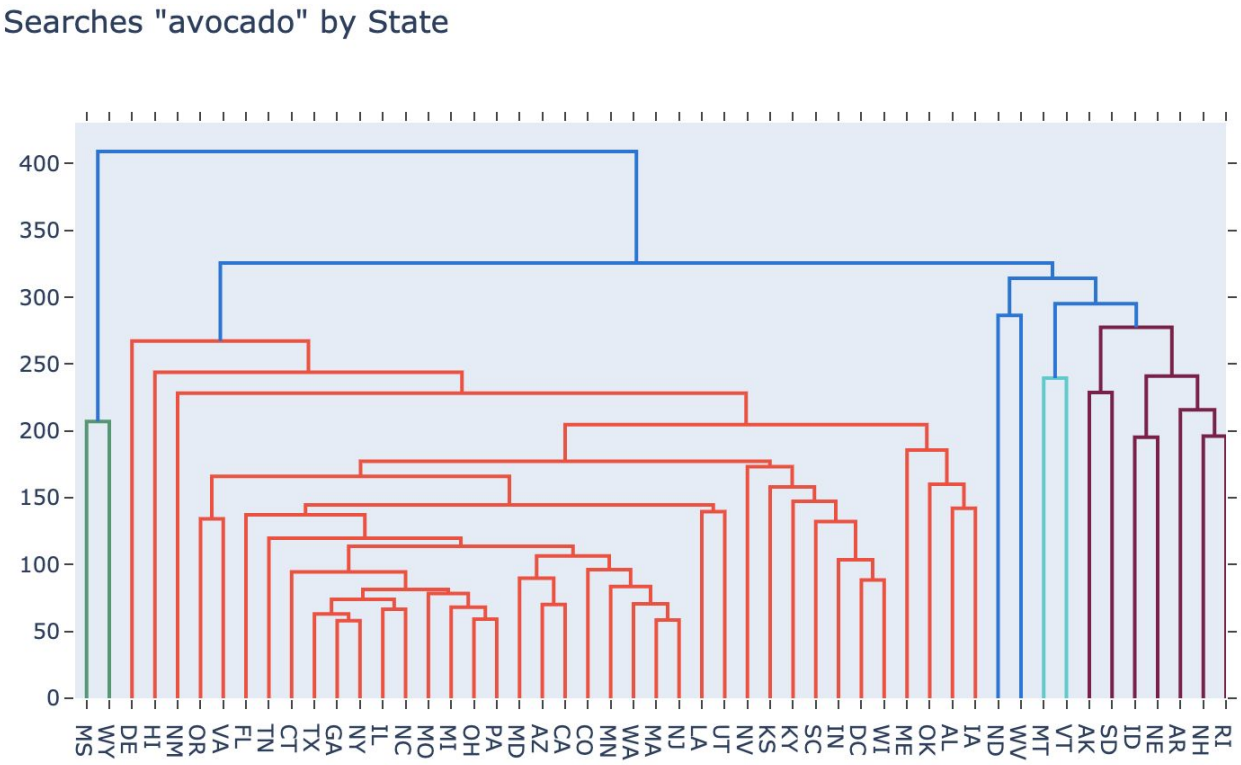
Cluster 6:['KY']

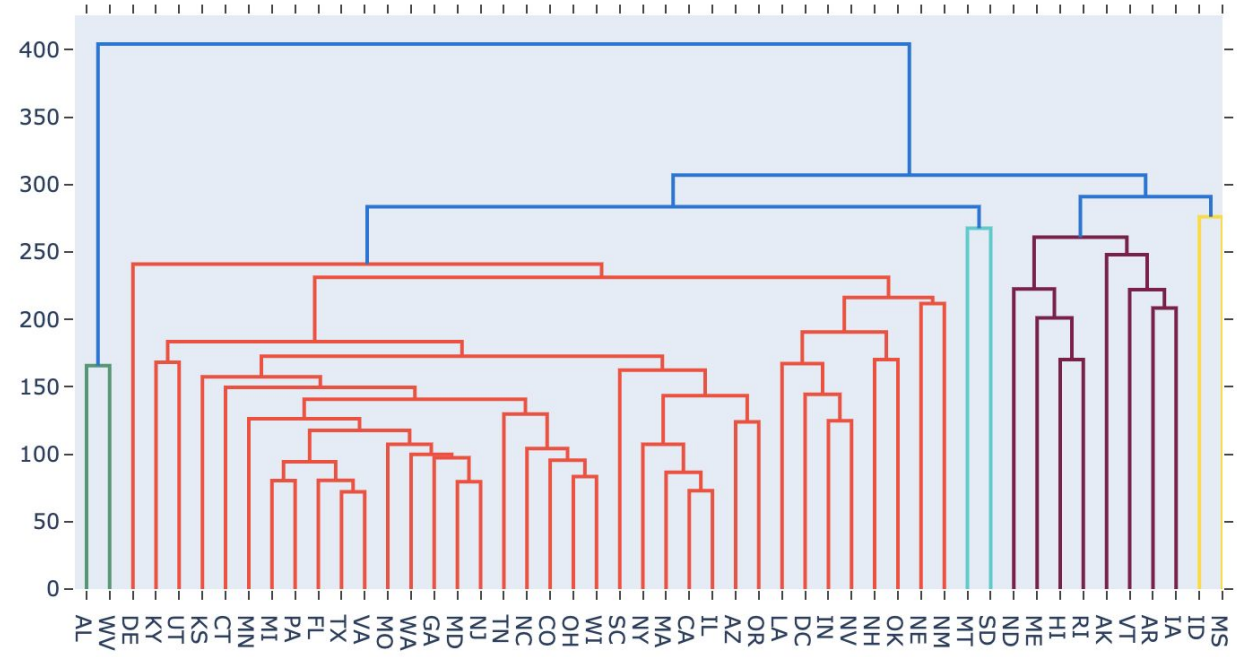
Cluster 8:['NE']

Cluster 3:['VT']

We see that the k means clustering yielded fairly uninteresting results, as we cannot see distinct cluster matches between clustered states by average avocado price data and clustered states by searches for "avocado" and "avocado toast". This is likely due to the differing natures of the datasets. While the avocado prices datapoints are many recordings per state, the google searches trends datapoints are per state with recordings throughout time. For this reason, states by avocado prices may appear in multiple clusters, but states by searches will appear only once. But still we can observe that states with little data about avocados are typically clustered together. To see if these clusters appear even with a different clustering method, we run hierarchical clustering for a second evaluation.

Hierarchical Agglomerative Clustering

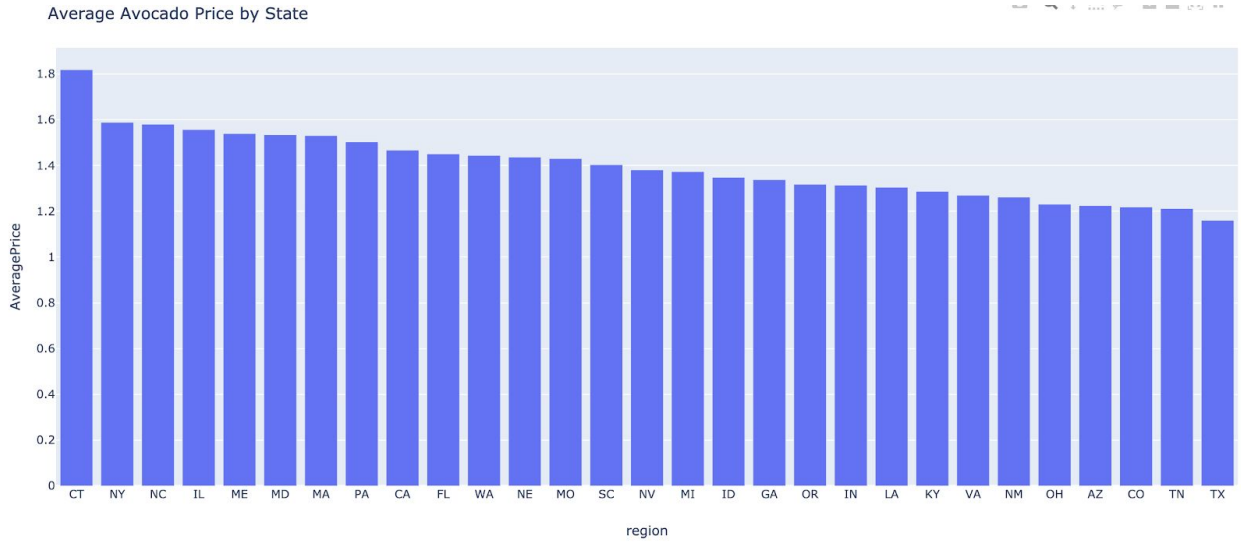


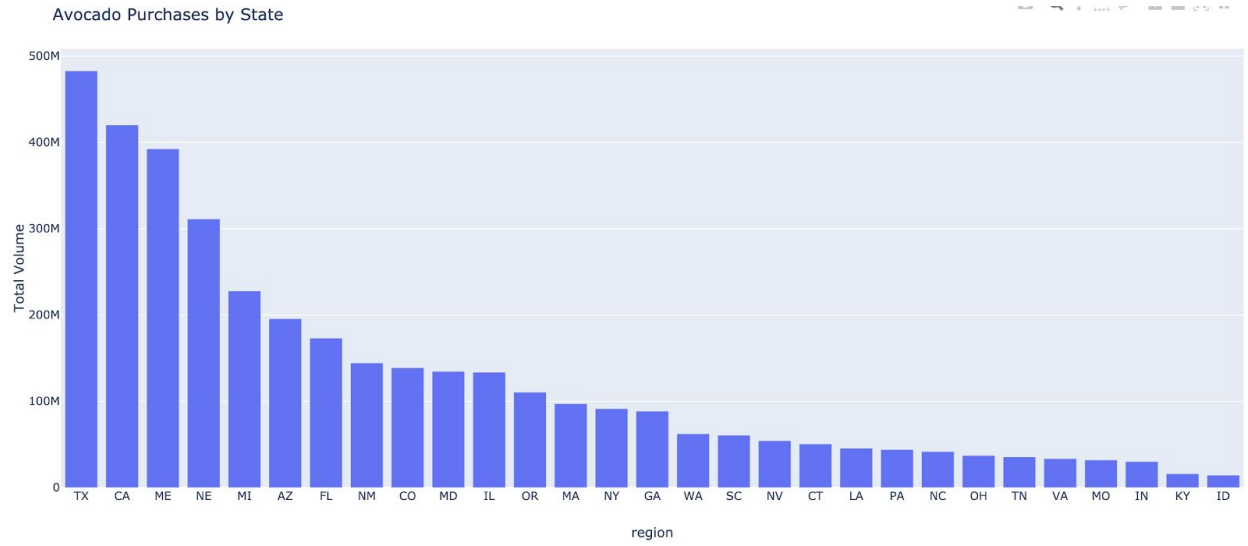


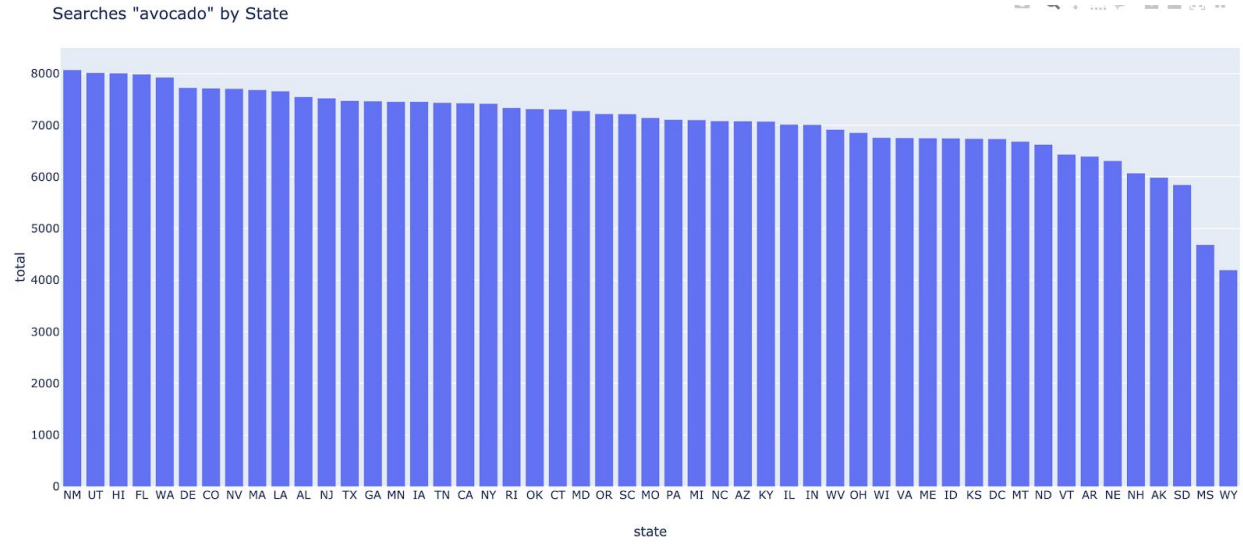
Because the clustering did not yield much interesting information about the association between avocado

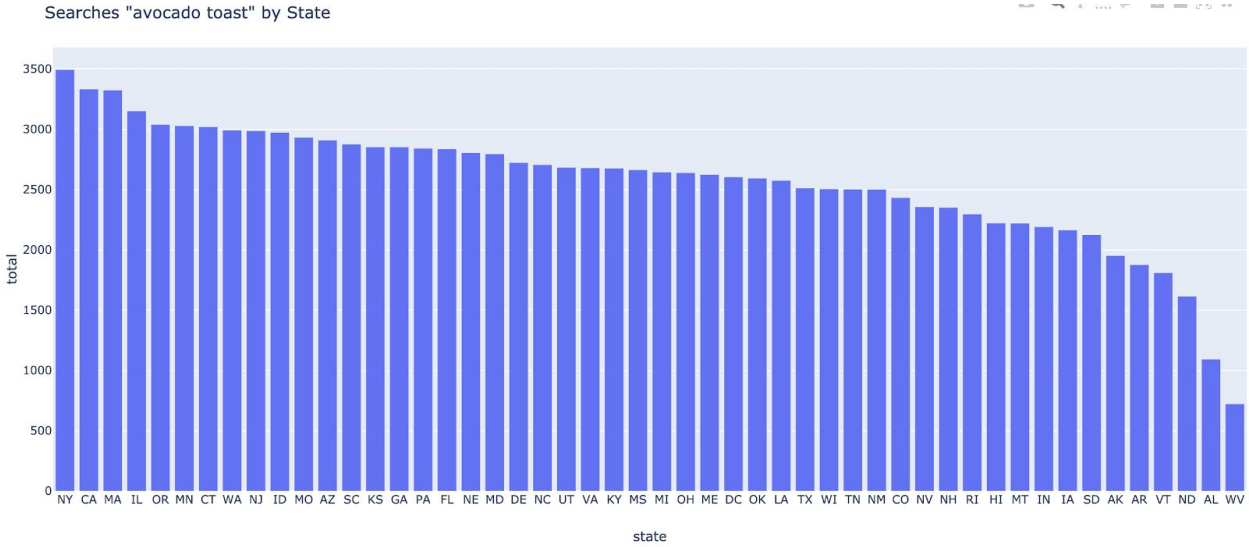
prices in states and google searches for “avocado” and “avocado toast”, we do a simple graph observation. To do this, we prepare the datasets again, aggregating data from the avocado prices to get the average price of avocados per state and the summed total volume of avocados sold per state. We also take the sum of the searches for “avocado” and “avocado toast” per state.

Graphed Data









Now, we see more interesting associations we can extrapolate from this data.
First, let's look at the graphs *Searches "avocado" by State* and *Searches "avocado toast" by State*. Though the top halves of the bar graphs and bottom halves match up and contain the same states, we see that states with the most searches for "avocado" are not the states with the most searches for "avocado toast". The top states with searches for "avocado" have high agricultural exports, and the top states with searches for "avocado toast" have more metropolitan areas. Of course, this is likely one of many confounding factors in the difference between these graphs.
Next, we observe the differences between the graphs *Average Avocado Price by State* and *Avocado Purchases by State*. The top four states of highest average price are fairly low on the list of avocado purchases by state. Now, the top states of most avocado purchases are likely due to population and size of the states. Interestingly, we don't see this same effect of population and state size in the google searches for "avocado" bar graphs.
Finally, we observe possible associations across these four graphs and see that there is a close association between the *Searches "avocado toast" by State* and *Average Avocado Price by State* graphs. Upon further investigation, we see that per-capita income of the top states in both graphs are high.

Conclusion

The strongest correlation we found was avocado price as a predictor of minimum wage. We were able to predict the minimum wage for a state with ~80% accuracy based on avocado prices. We would like to see whether avocado prices are specifically a good indicator of minimum wage, or whether it is simply a case of wages reflecting prices in general.
For the other questions we looked at, we found slight correlations, at best, but not enough to draw conclusions. For the comparisons between Google searches and homeownership, and Google searches and avocado prices, we may have been able to obtain better results by using different models. The search-price comparison might fit better to a linear regression model; the search-homeownership could have attempted to normalize for other factors known to affect homeownership rates, such as whether a region is more urban or rural.

While we weren't able to prove Tim Gurner wrong, we did learn a lot about the differences between regions where avocados are popular and regions where they are less so. We also learned that getting meaningful results can take several attempts using different approaches.