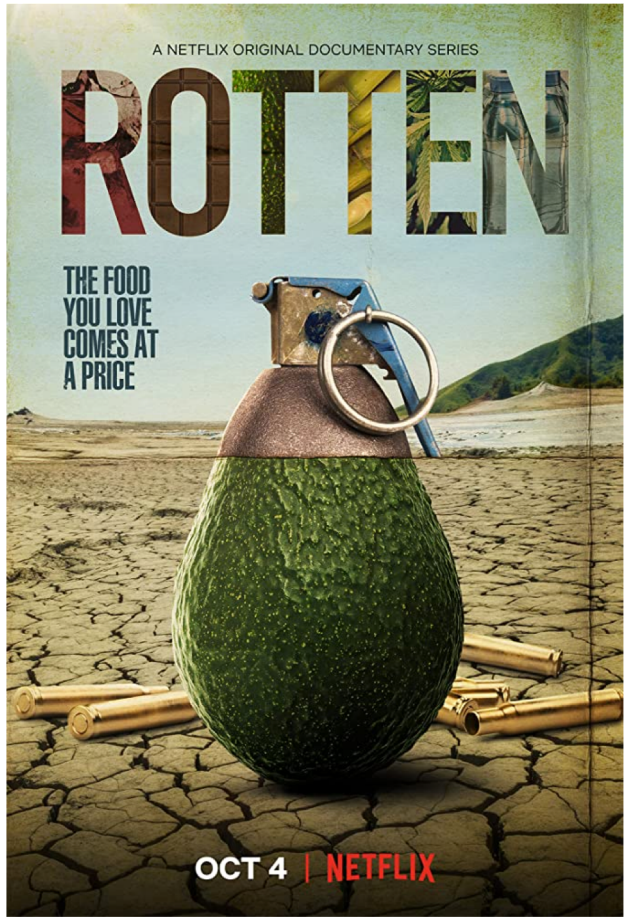
UTILIZE MACHINE LEARNING AND FORECASTING TO PREDICT AVOCADO PRICES

QUEENIE HU

JUNE 10 2020

INTRODUCTION



- Why avocados are the "Trendiest food"?
  - KOLs promote healthier eating diet
  - "Superfood" – nutritious and antioxidant
- Current problems with avocados:
  - Demand increases, over farming causes deforestation and destroying ecosystem
  - Cartels run avocado trade in Mexico, therefore enforcing extortion fee from framers
- Consequence:
  - Prices went up by 129% in the past few years
  - Average single Hass US$2.10 in 2019 (compare to US$1.17 from July 2018)

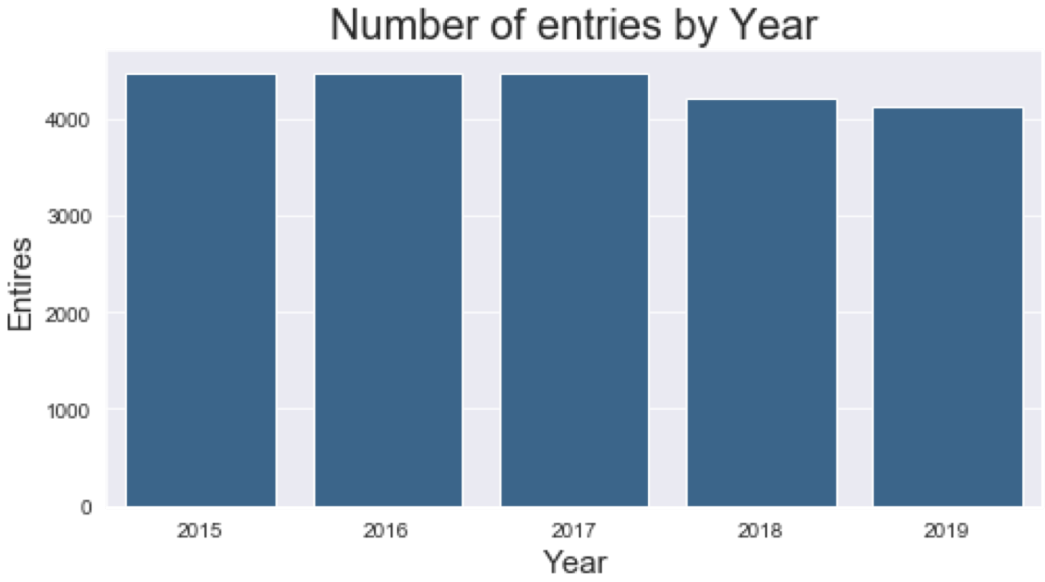
RESEARCH QUESTION AND APPROACH

- Research question
  - How would avocado prices look like in 2020?
  - What are factors that influence avocado prices?
  - Determine which model works best to predict avocado prices

- Approaches:
  - Data set was taken from Kaggle, which comprised of avocado prices in US from 2015 – 2019
  - Use python and various packages for data processing and graphing
    - Facebook Prophet for Time Series analysis
    - Regression and Machine Learning Algorithm

DATA PROCESSING

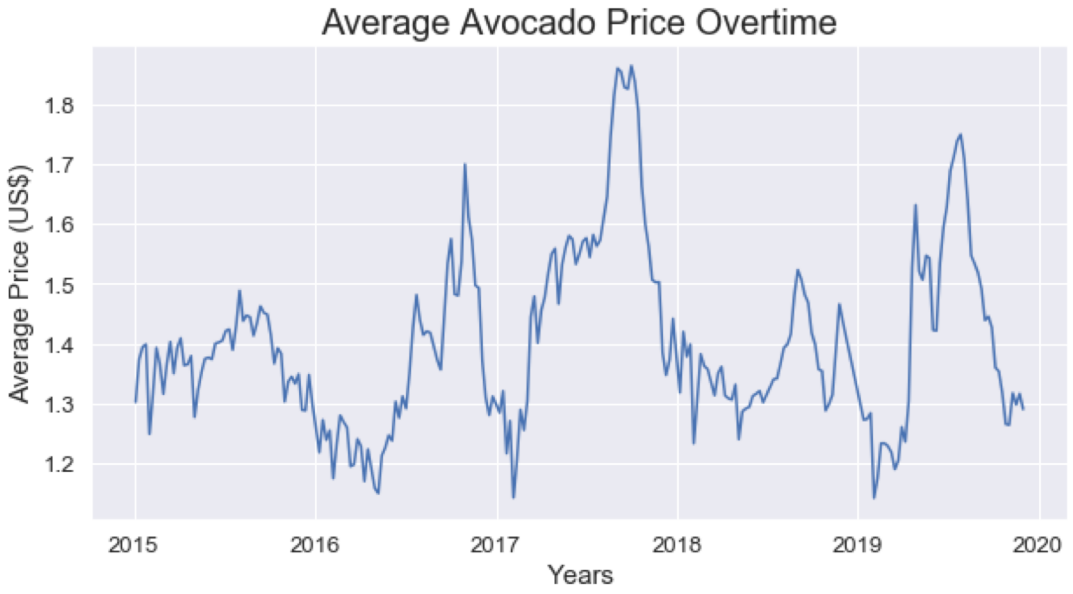
	Date	AveragePrice	Total Volume	Small Hass	Large Hass	XLarge Hass	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-01-04	1.22	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.0	conventional	2015	Albany
1	2015-01-11	1.24	41195.08	1002.85	31640.34	127.12	8424.77	8036.04	388.73	0.0	conventional	2015	Albany
2	2015-01-18	1.17	44511.28	914.14	31540.32	135.77	11921.05	11651.09	269.96	0.0	conventional	2015	Albany
3	2015-01-25	1.06	45147.50	941.38	33196.16	164.14	10845.82	10103.35	742.47	0.0	conventional	2015	Albany
4	2015-02-01	0.99	70873.60	1353.90	60017.20	179.32	9323.18	9170.82	152.36	0.0	conventional	2015	Albany

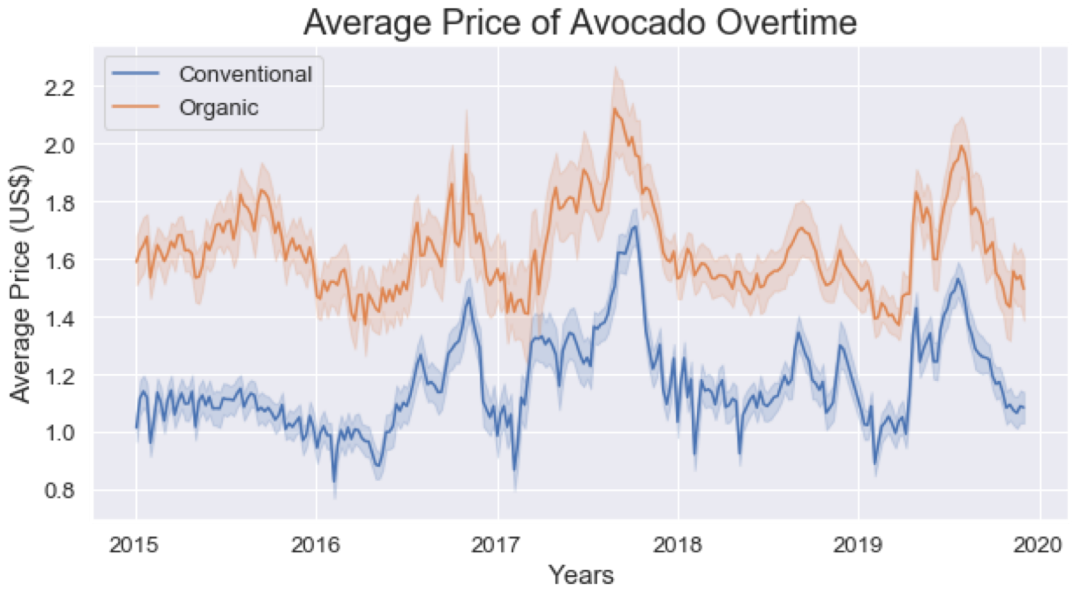


- Data source was taken from the Hass Avocado Board website
- There were 27323 entries with 13 columns
- Used ”dtype” and “describe” to ensure data were transformed in the right format
- Data did not consist of nulls
- ”Date” was changed from ‘object’ to ‘datetime’ for time series analysis
- Number of entries were consistent among years

DATA ANALYSIS – OVERVIEW OF AVOCADO PRICE

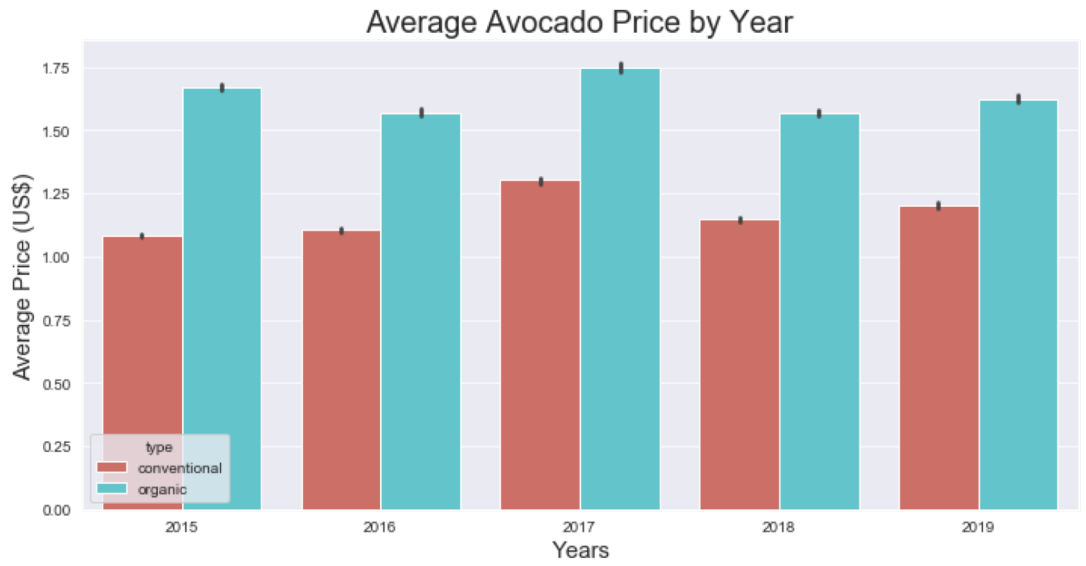
- With the emergence of social media, KOLs found a medium to deliver their key ideas to the public efficiently
- In mid 2016 avocado was introduce as ”Superfood”, demand increased and supply has to keep up
- The average organic avocado price was ~1.5x of the conventional avocado

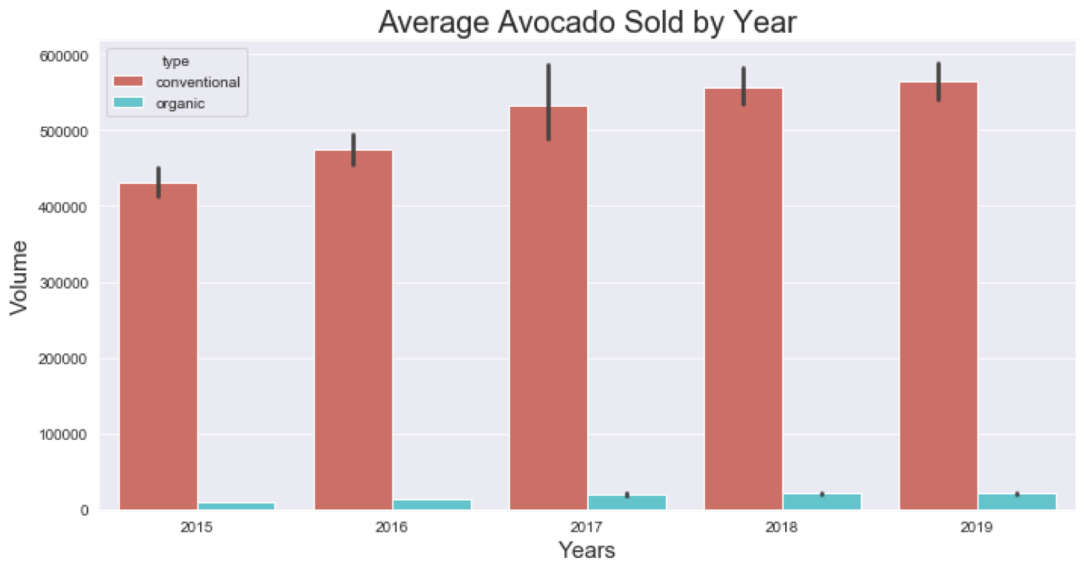




DATA ANALYSIS – AVOCADO PRICE AND VOLUME SOLD BY YEAR

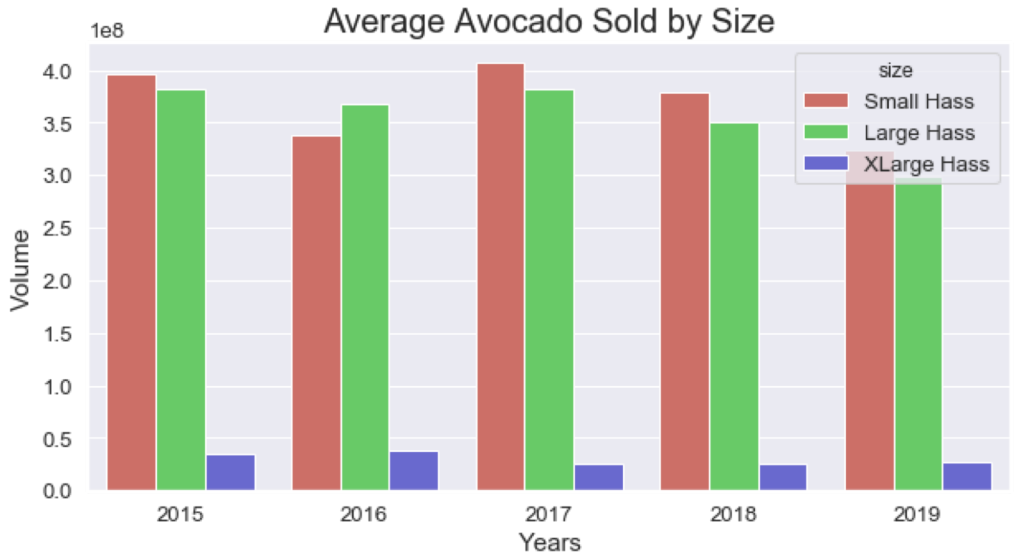
- As supply decreased in mid 2016, it pushed both conventional and organic avocado prices up
- Due to high price for organic avocados, Americans favour towards conventional than organic
- As expected demand continued to grow in 2017 – 2019

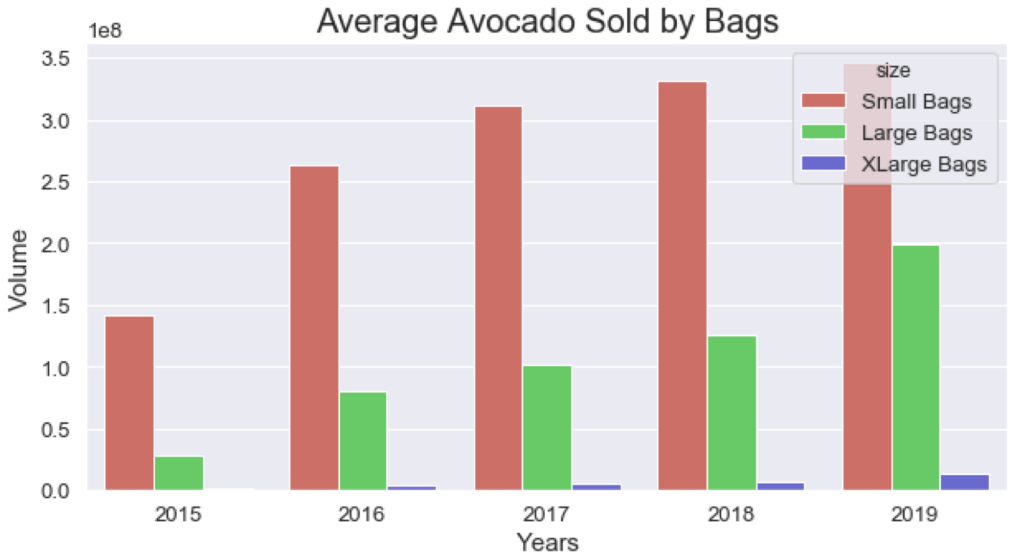




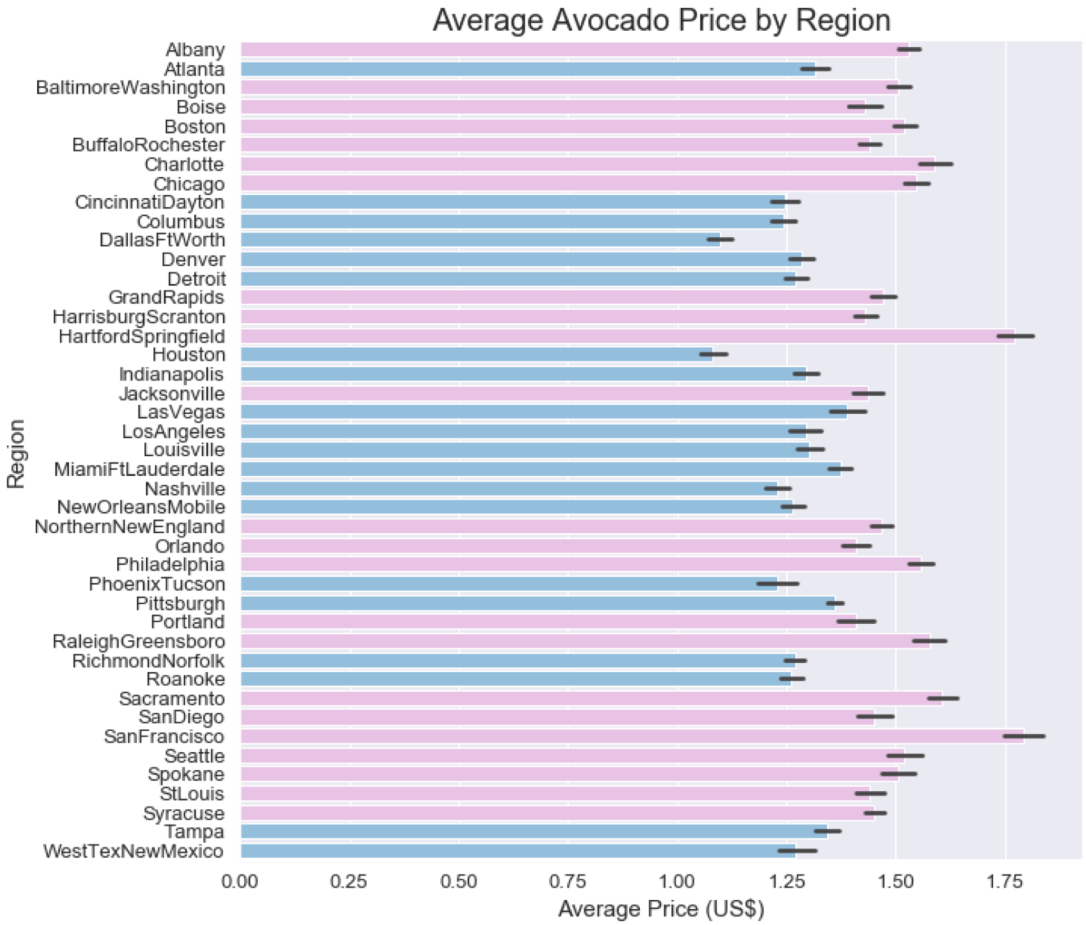
DATA ANALYSIS – AVOCADO PRICE BY SIZE

- Small Hass is the more preferred type to be sold in the States – both by unit or in bags
- Large Hass is also popular when sold by unit





DATA ANALYSIS – AVOCADO PRICE BY REGION

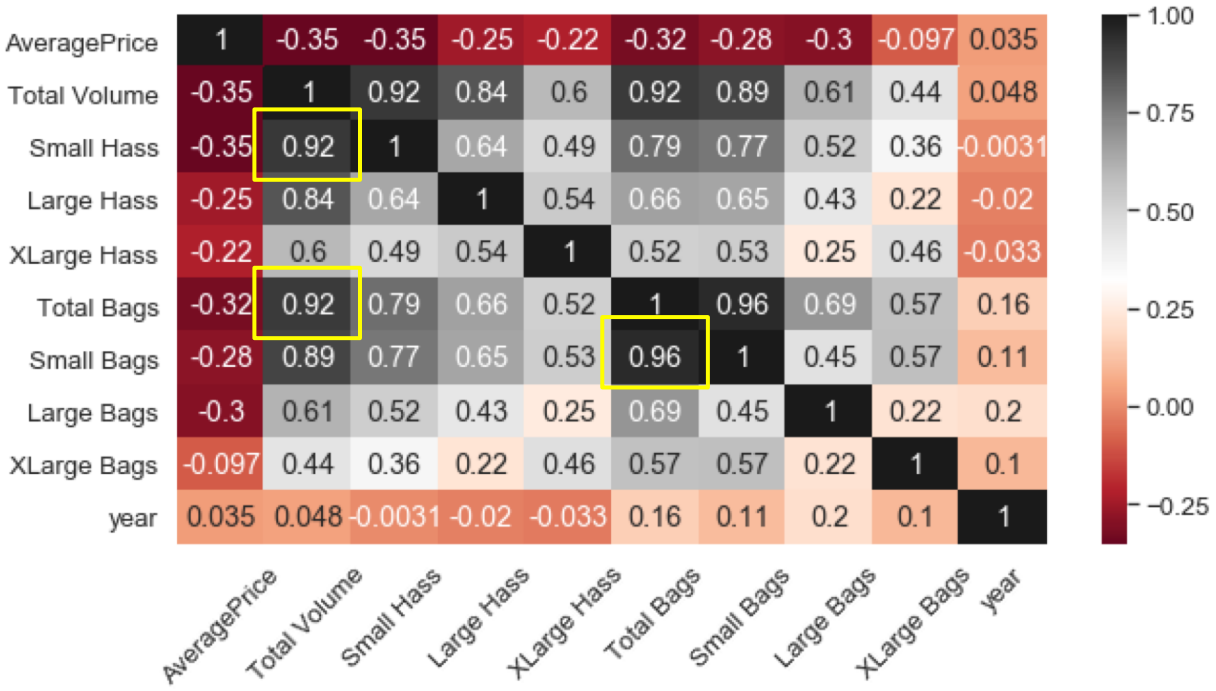


- Highlighted regions that have average price > than the overall average
  - Pink> $1.402
  - Light Blue < $1.402

```
custom_palette = {}
for q in set(df2.region):
    avg = (np.average(df2[df2.region ==q].AveragePrice))
    if avg > 1.402:
        custom_palette[q] = '#F0BBE9'
    else:
        custom_palette[q] = '#89C0E7'
```

- 23/43 regions surpass the the overall average price (that is 50%!)

DATA ANALYSIS - CORRELATION



- Analysis shows there is a high correlation between these pairs:
  - Small Hass & Total Volume (0.92)
  - Total Bags & Total Volume (0.92)
  - Small Bags & Total Bags (0.96)
- What this means?
  - Small Hass avocados are the most preferred/sold type in the US and customers tend to buy those avocados as bulk, not bag
  - Retailers want to increase the sales of bagged avocados instead of bulks.
  - Total Bags variable has a very high correlation with Total Volume and Small Bags, so we can say that most of the bagged sales comes from the small bags

TIME SERIES ANALYSIS USING FACEBOOK PROPHET

- What is FACEBOOK PROPHET?
  - An open source software released by Facebook's Core Data Science team
  - A procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects
  - Works best with time series that have strong seasonal effects and several seasons of historical data
  - Syntax follows sklearn model API – create an instance, then call its fit and predict methods

FACEBOOK PROPHET - SYNTAX

```
from fbprophet import Prophet
from fbprophet.plot import add_changepoints_to_plot

df_prophet = df_prophet.rename(columns={'Date':'ds', 'AveragePrice':'y'})
p.fit(df_prophet)
future = p.make_future_dataframe(freq='M', periods=12)

forecast = p.predict(df_prophet)
forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].head()

forecast = p.predict(future)
fig = p.plot(forecast)
```

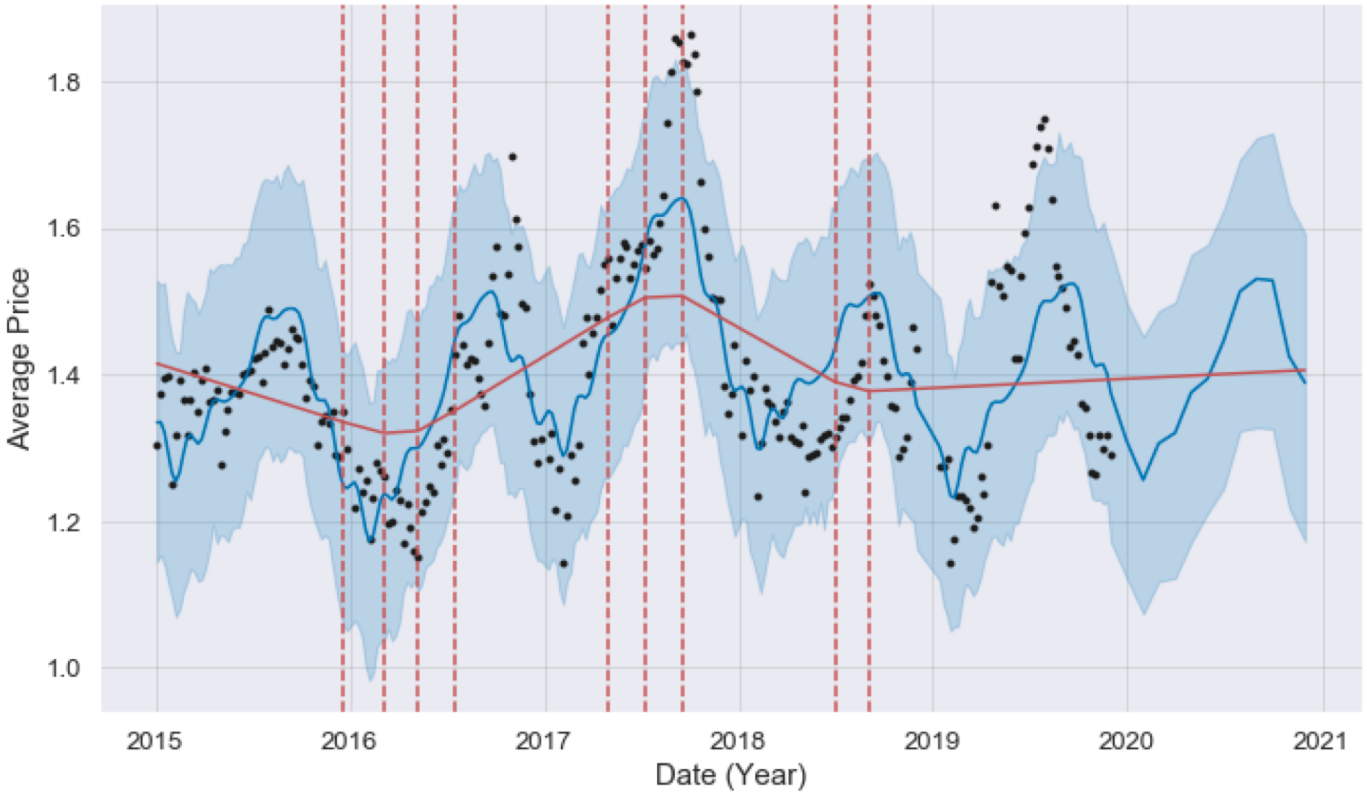
Create an instance

Fit into the model

Predict

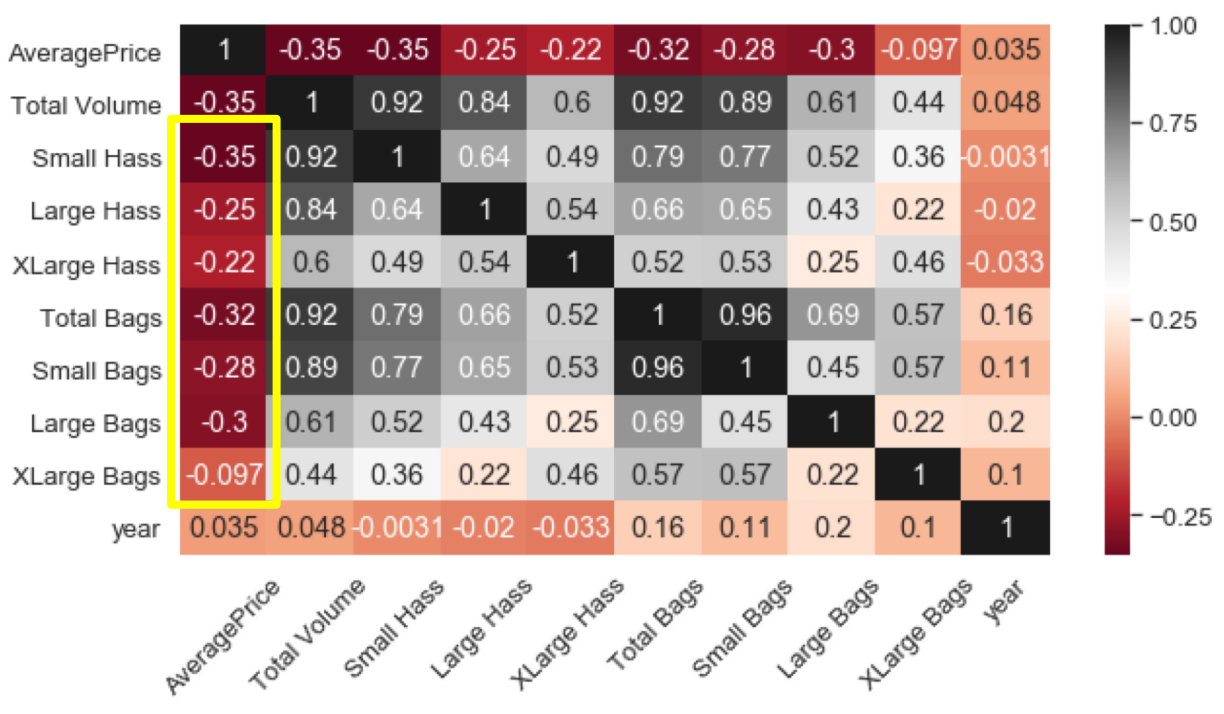
Create graph

TIME SERIES ANALYSIS USING FACEBOOK PROPHET



- Red dotted line shows data change point
  - 2016 – Avocado trend boomed and causes shortage
  - 2017 – Demand> Supply
- Predicted avocado prices in 2020 will remain constant

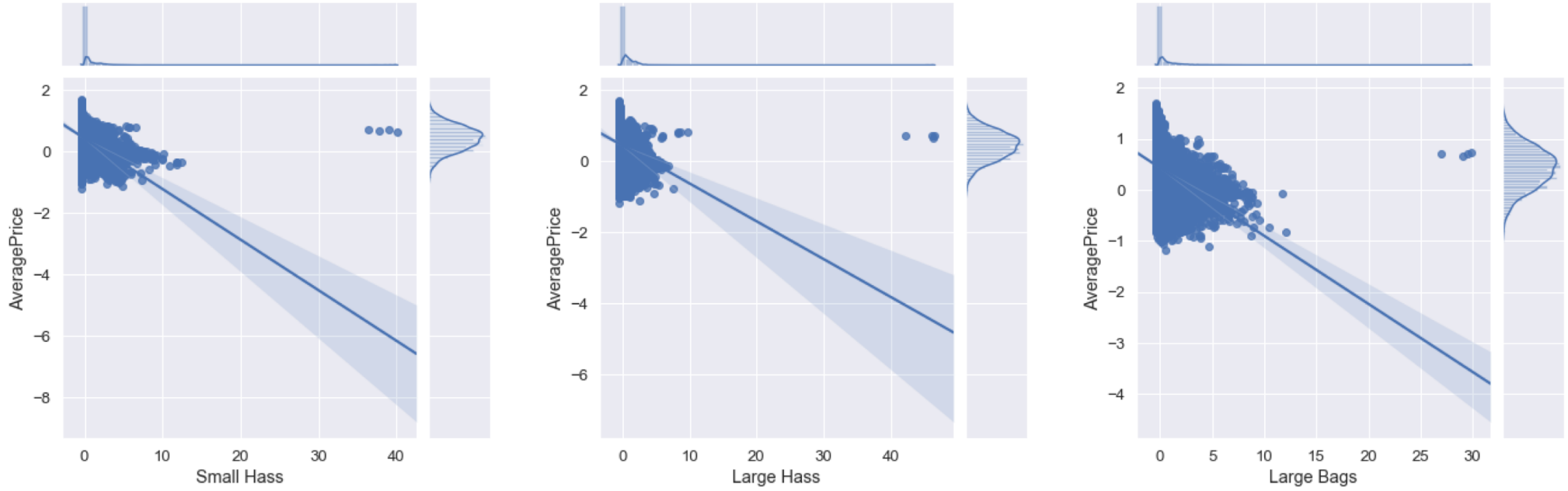
DATA MODELING – RECAP CORRELATION TABLE



- Analysis showed a moderate negative correlation between price and unit sold
- Indicates quantity demand increases as price decreases
- Highlight data will be used for the data modelling:
  - Multiple linear Regression
  - Decision Tree Regression
  - Random Forest Regression

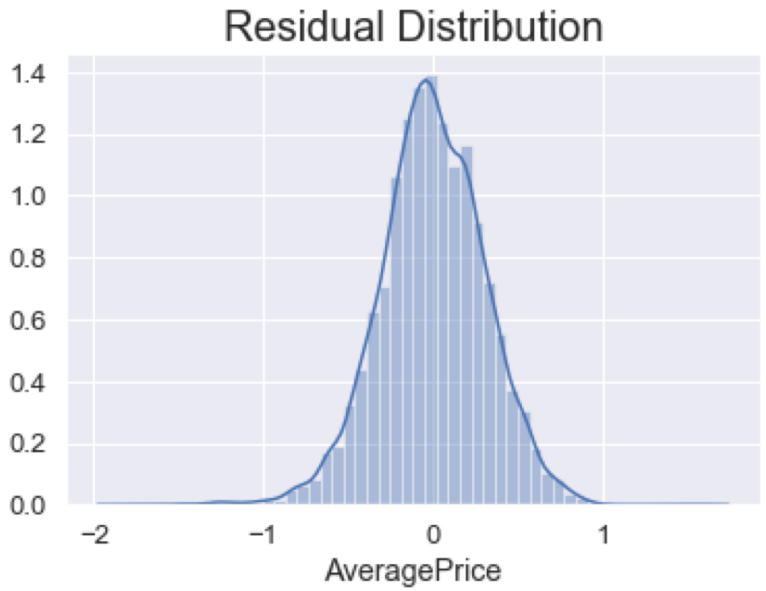
DATA MODELING – MULTIPLE LINEAR REGRESSION

- Visualization of highly correlated variables with the average prices before regression analysis



- Shaded area represents 95% confidence

DATA MODELING – MULTIPLE LINEAR REGRESSION



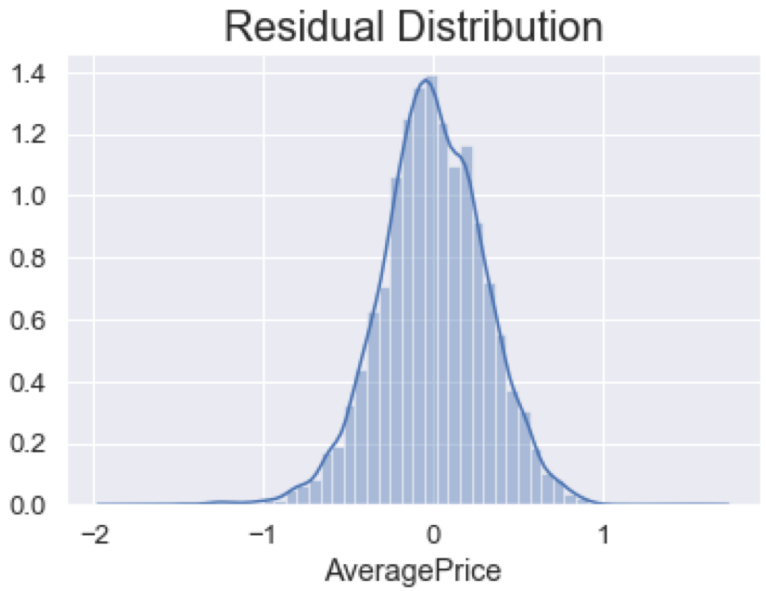

```
Mean Absolute Error (MAE):  0.241
Mean Squared Error (MSE):  0.095
Root Mean Squared Error (RMSE): 0.308


Intercept:  0.203
Coefficient:  [ 0.459 -0.123  0.098 -0.02   0.002 -0.034  0.035]
Train Score:  0.425
Test Score:  0.429
```

** Coefficients represent Type_organic, Small Hass, Large Hass, Xlarge Hass, Small Bags, Large Bags, Xlarge Bags, respectively

DATA MODELING – MULTIPLE LINEAR REGRESSION

- Also confirmed the r-squared value using statsmodels



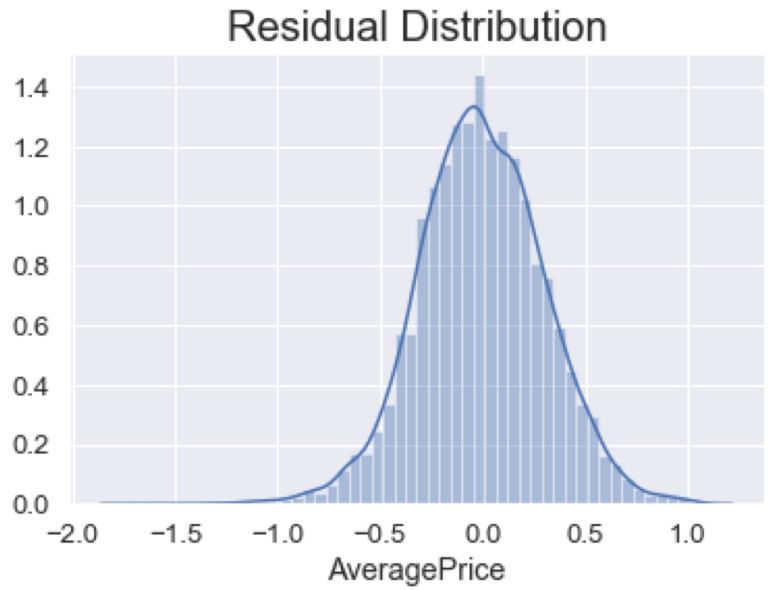
OLS Regression Results

Dep. Variable:	AveragePrice	R-squared:	0.428
Model:	OLS	Adj. R-squared:	0.428
Method:	Least Squares	F-statistic:	1394.
Date:	Tue, 09 Jun 2020	Prob (F-statistic):	0.00
Time:	09:21:20	Log-Likelihood:	-3135.3
No. Observations:	13054	AIC:	6287.
Df Residuals:	13046	BIC:	6346.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2040	0.004	50.027	0.000	0.196	0.212
type_organic	0.4557	0.006	75.018	0.000	0.444	0.468
Small Hass	-0.1234	0.005	-26.439	0.000	-0.133	-0.114
Large Hass	0.0970	0.004	23.687	0.000	0.089	0.105
XLarge Hass	-0.0193	0.004	-5.399	0.000	-0.026	-0.012
Small Bags	-0.0054	0.005	-1.050	0.294	-0.016	0.005
Large Bags	-0.0332	0.003	-10.263	0.000	-0.040	-0.027
XLarge Bags	0.0391	0.003	11.894	0.000	0.033	0.045

Omnibus:	354.803	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	796.089
Skew:	-0.130	Prob(JB):	1.35e-173
Kurtosis:	4.182	Cond. No.	5.04

DATA MODELING – DECISION TREE REGRESSION



- Optimize max_depth

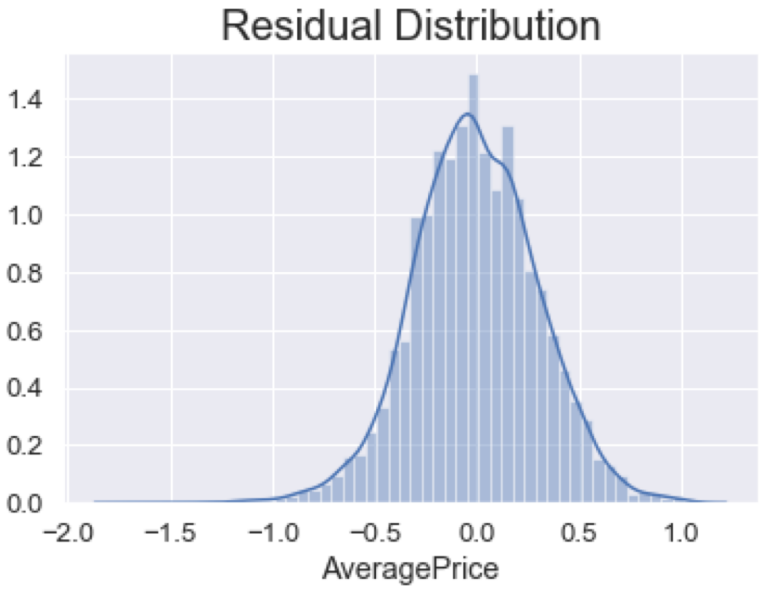
```
Depth: 2 , MSE: 0.094
Depth: 2 ,RMSE: 0.307
Depth: 3 , MSE: 0.089
Depth: 3 ,RMSE: 0.298
Depth: 4 , MSE: 0.083
Depth: 4 ,RMSE: 0.288
Depth: 5 , MSE: 0.078
Depth: 5 ,RMSE: 0.279
MinDepth: 5
MinRMSE: 0.279
```

- Scores when max_depth=5

```
Mean Absolute Error (MAE):  0.219
Mean Squared Error (MSE):  0.078
Root Mean Square Error (RMSE): 0.279

Train Score:  0.556
Test Score:  0.538
```

DATA MODELING – RANDOM FOREST REGRESSION



- Scores when max_depth=5

```
Mean Absolute Error (MAE):  0.222
Mean Squared Error (MSE):  0.079
Root Mean Square Error (RMSE): 0.281

Train Score:  0.538
Test Score:  0.529
```

CONCLUSION

- Data Analysis
  - Avocado supply had caught up demand and therefore prices will remain staggered in 2020
  - Conventional Small Hass will continue to be the most popular avocado among the all Hass
- Predictive Modeling

```
Linear Regression RMSE: 0.31

Decision Tree Regression RMSE: 0.279

Random Forest Regression RMSE: 0.281
```

  - Decision Tree could be the best fit among 3 models being used