AVOCADO PRICES DATASET

Bahattin AKSOY 171180004
Gamze AKSU 171180005
Oğuzhan AKBAL 171180001

Abstract

In this paper, the Avocado Prices dataset is examined.

Before Starting the Report

We decided to work on "Avocado Prices" which we found at Kaggle. Before we start, why did we choose that data? It seems like interesting to us then we want to decide inorganic and organic avocados' prices depends on its price or not. After that we are going to look years for planting inorganic avocados provides more money according to planting organic avocados.

Our Goal

We are going to answer the questions what we sent in our proposal. In this report, we are going to answer and explain that questions. We are going to explain how we worked, what we learn from this project, how the stages of the project is and how we finish the project.

Abstract of the Report

Firstly, we examine the dataset to understand whole data. It means, we were trying to decide;

- How many countries are producing avocados and which countries are them.
- The average price of avocados.
- Total volume of avocados. (Total number of avocados sold)
- Date of plantings for avocados.
- What year were avocados planted.

We analyzed this datas to start the project. After that we decided to where we start. We used "Pandas, Numpy, Matplotlib and SciKit-Learn and some other ready-to-use libraries" for that project. We explained our codes with comment lines. But we found it suitable to explain in the report too. So that we wrote our codes for visualization to explain our data with tables. We printed whole datas with our codes to show the data. At the end of the project, we used nearly whole classifications to show like: "Logistic Regression, KNN, Support Vector Machine, Naive Bayes, Decision Tree and Random Forest".

Dataset for the Project

Dataset Name: Avocado Prices

Dataset Link

https://www.kaggle.com/neuromusic/avocado-prices

Columns in the Dataset

1. Index

2. Date - The date of the observation-

3. AveragePrice - the average price of a single avocado -

4. Total Volume - Total number of avocados sold

5. #4046 - Total number of avocados with PLU 4046 sold -

6. #4225 - Total number of avocados with PLU 4225 sold -

7. #4770 - Total number of avocados with PLU 4770 sold -

8. Total Bags

9. Small Bags

10. Large Bags

11. XLarge Bags

12. Type - conventional or organic –

13. Year

14. Region - the city or region of the observation

Some details about columns

• The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The

Product Lookup codes (PLU's) in the table are only for Hass avocados. Other varieties of avocados (e.g. greenskins) are not included in this table.

- Index values between 0-52
- Date values between 4.06.2015 – 25.03.2018
- Type values conventional or organic
- Region column has 54 different region
    1. Albany
    2. Atlanta
    3. BaltimoreWashington
    4. Boise
    5. Boston
    6. ….
- There are 338 data from each region.
- #4046, 4225 and 4770 are cods of PLU (Price Look-Up). The PLU code identifies produce items based upon the commodity, variety and size group.
- There is no null data values.
- There are totally 18249 rows.

Motivation

Our motivation to work with data is data's be interesting. We would like to work on a extraordinary data. While we were looking for datas at Kaggle, we found this data interesting and we want to work with that data. We wanted to see is it depend on money and when look at the years inorganic planting provide farmers to earn more than planting organic. We were also new to work with team. We wish to improve our communication skills and we wish to learn work with team. It is almost a first work experience with team which we don't know each other. We were excited for that anyway.

The Questions We Were Working On

Q1. Which type of avocado prices increase over the years?

The mean of organic and conventional avocado sales will be plotted annually and the amount of increasing will be observed.

First, we take the necessary columns from data. (Avocado prices with year and type)

Second, we calculate the mean of an avocado price over the years for organic types.

Third, we calculate the mean of an avocado price over the years for conventional types.

Finally, most useful graph for this question is line graph. We create a line graph to follow changes avocado prices over the years.

-Red-: Conventional Avocados

-Blue-: Organic Avocados

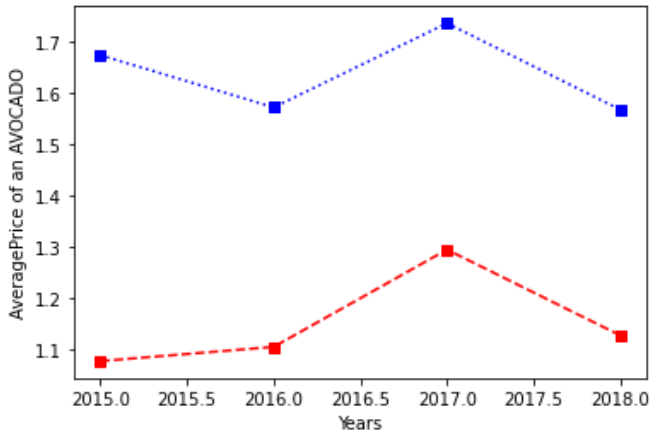

Figure 1: Visualization of question 1

As we can see in the table, organic avocados always have more price than conventional avocados all the years. Strangely, conventional avocado prices increase between 2015 and 2016 unlike organic avocados. Climate problems might be affect this table that year. It means, if the temperature value is appropriate for the mature, it affects the amount of the organic avocados. Because of that, organic avocados' values might be decrease. Or there is other way: farmers might planted less conventional avocados, so that conventional avocados' price increase.

At 2016-2017; both of their prices increase and at 2017-2018; both of their prices decrease. It is normal than the first example. It is normal to increase prices together and decrease prices together.

Q2. Does organic or inorganic planting effect to avocado prices at all region?

The unit amount of inorganic and organic avocados for each region will be shown on a column chart. In this way, it will be determined whether inorganic sales cost cheaper for each region.

First, we take the necessary columns for data. (Avocado types, prices and region)

Second, we group the taken data by type and region. And then we choose avocados type just for type = conventional first, then we did the same thing for the avocado type = organic.

Finally, we use plot bar to visualize the data. We can see at the Figure 2. If we examine the column chart, we can say organic avocado prices more expensive than conventional avocado prices at all regions. So that the question's answer is 'Yes'. Avocado prices depends on its type. Actually, it can be change by regions. (If we looking for group by group.)

Q3. In which region are avocado prices more expensive or cheaper for average price of an avocado?

We will analyze data for each region and see average of avocado prices in selected region and compare them.

First, we take the necessary columns for data. (Avocado prices and region)

Second, we group our taken data by region.

End of all, we use a plot bar to visualize the datas. It is the easiest way to answer the question. We can see at the Figure 3.

There is not more thing to say. If we look the bar chart, HartfordSpringfield sells the most expensive avocado. Houston sells the cheapest avocado. (For this dataset)

Q4. In which region are avocado sales volume more?

We will analyze data for each region and see sum of total avocado sales and compare them.

First, we take the necessary columns for data. (Avocado total volume and region)

Second, we group taken data by region and calculated the total volumes in this data for each region.

Finally, we used plot bar to visualize the taken data. We can see at the Figure 4.

As we can see in the table, TotalUS region is going with lead. Its value is too high for others. So, if we want to decide min top 3 and max top 3, we should remove the TotalUS region from the table. We can compare others easily. In this table we can see the max top 1, but we can't decide min top 1 by looking the table. Figure 5 is with the TotalUS removed from Figure 4. This table seems like a little bit clear than the first table. By looking that table we can decide other things easily. We can say that, Min 3 regions are: 'Syracuse', 'Boise', 'Spokane' and Max 3 regions are: 'SouthCentral', 'California' 'West'.

Q5. Which region would be cheaper to buy avocados from next year?

Prices of avocados sold by regions and years will be plotted and the price will be predicted for the next year.

First, we take the necessary columns for data. (Avocado prices and region)

Second, we group taken data by region and year.

Third, we calculated mean all prices in data.

Now, we get the data for 4 years: 2015, 2016, 2017 and 2018. In this question, we need to predict for the next year 2019. So that, we are going to create regression function for each region depends on year and then predict the avocado prices for each region in 2019.

Finally, we got the values for 2019. We used plot bar to visualize the data. We can see at the Figure 6. We can decide which region is going to be the cheapest value and which region is going to be the most expensive value in 2019.

By looking the Figure 6, we can answer our question, we can see Houston's avocados will be the cheapest in 2019.

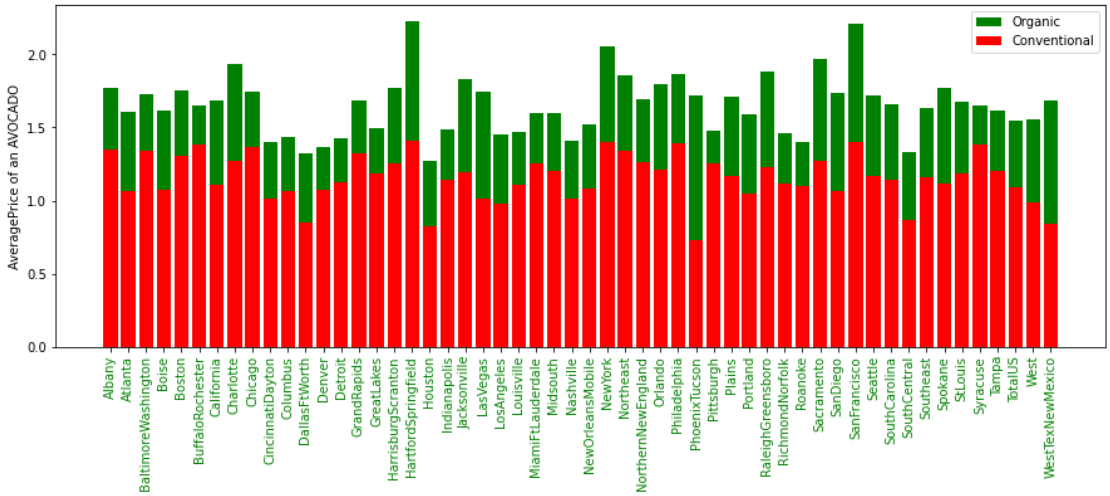

Figure 2: Visualization of question 2

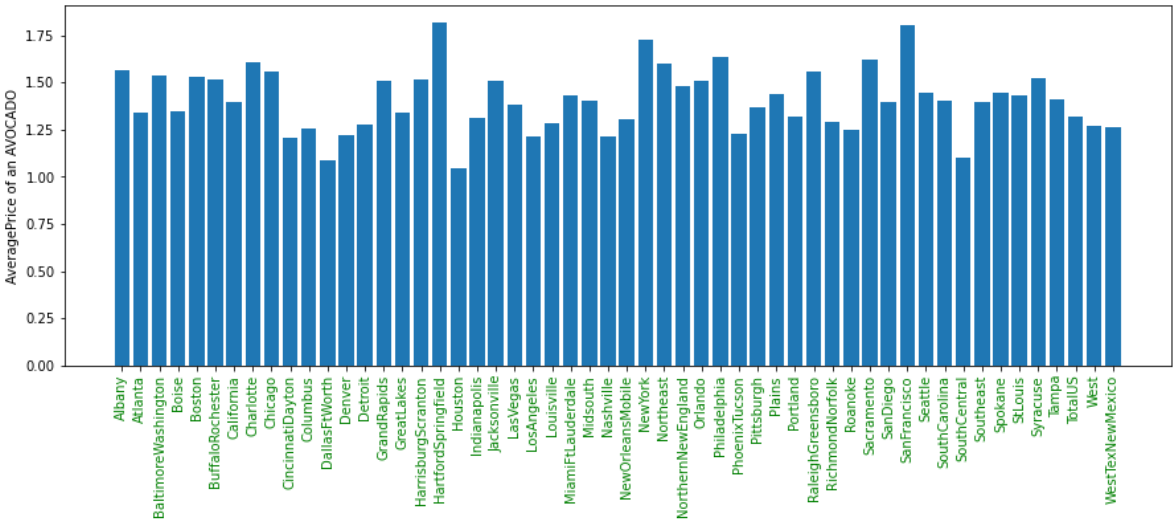

Figure 3: Visualization of question 3

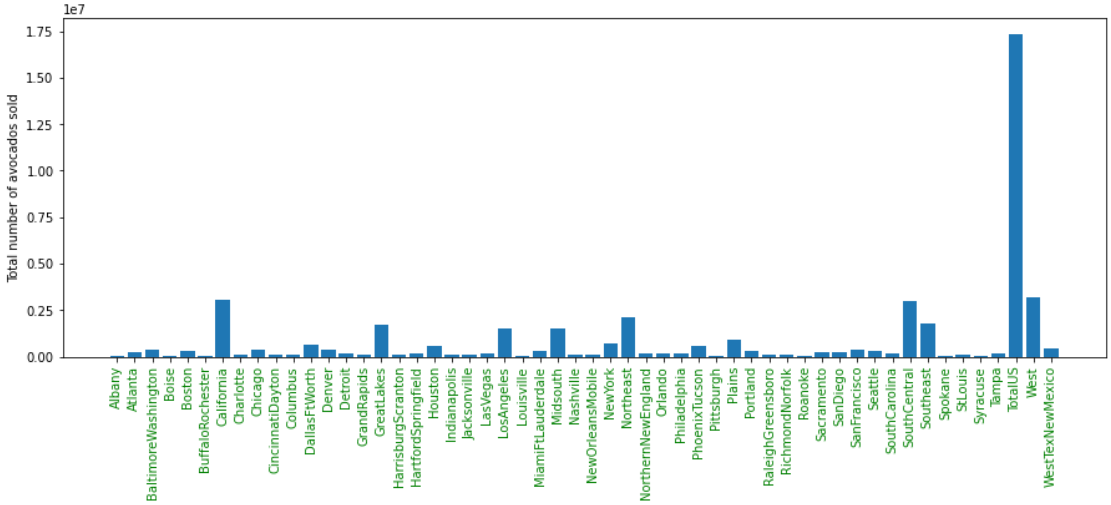

Figure 4: Visualization of question 4

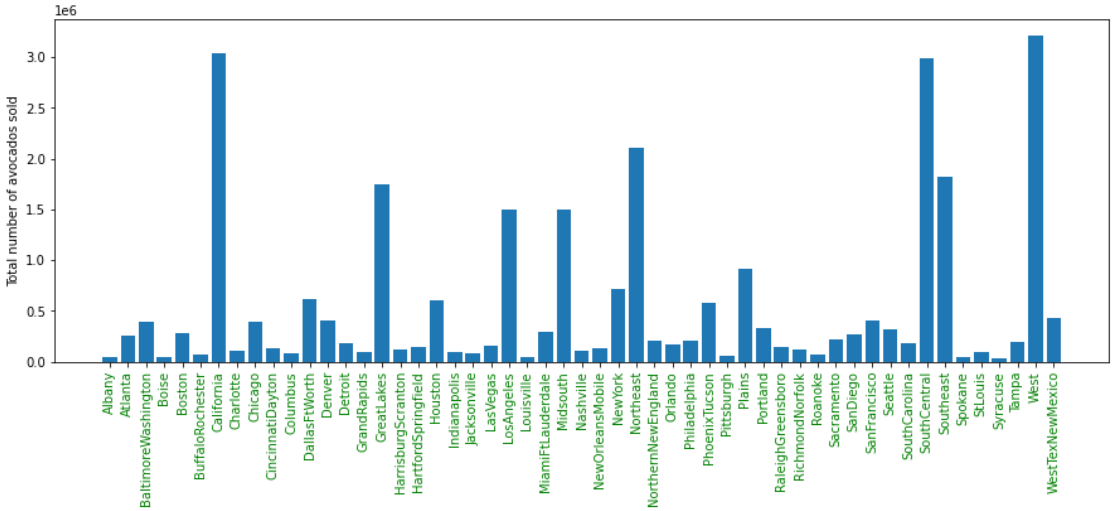

Figure 5: TotalUS removed from Figure 4.

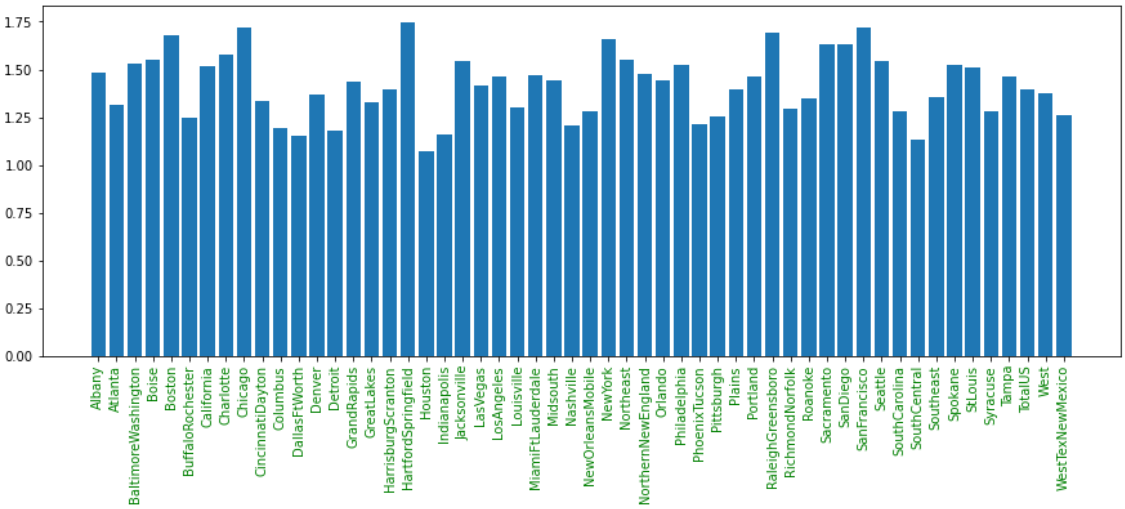

Figure 6: Visualization of question 5

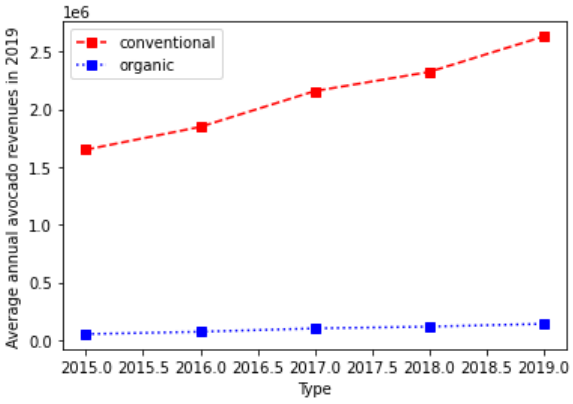

Figure 7: Visualization of question 6

Q6. Will organic or inorganic planting next year bring more income?

Organic and inorganic avocado prices and avocado sales averages will be calculated annually. And sales prices will be predicted for the next year. And the sales volumes for the next year will be predicted.

First, we take the necessary columns for data. (We need 2 dataframe here;

1. Average prices grouped by year and type

2. Total volume grouped by year and type.

We need avocado prices, type and also years to work with regression function)

Second, we group taken data by year. And calculate mean all prices in data.

Third, we are going to predict avocado prices in 2019. We have 2015, 2016, 2017 and 2018 values. We wrote code to predict the value with regression function. Also we did the same thing for total volume.

Forth, we grouped taken data by year and type of avocados. And we calculated mean all volumes in the data. Now we got 2015, 2016, 2017 and 2018 values. We are going to predict 2019 values by regression function.

Fifth, we added 2019 total volume values and average prices values to the table (dataset). Then we splitted types and have organic, conventional data frames for price and volume then multiple averagePrice x totalVolume values to calculate incomes for each year.

Finally, we visualized our last question as line graph. We can see at the Figure 7. As we can see in the table, if farmers going to farm conventional avocados, they will get more income.

Classifications

You can see in the Figure 8 which classification algorithms and parameters we used to predict the avocado type. At the table, they sorted by score and first 5 are random forest classifier. The lowest score is in the Support Vector Machine algorithm where the Sigmoid Function is used. Conclusion for this, we can say Random Forest Algorithm is the most suitable classifier algorithm for avocado data.

What We Learnt?

- We learned how to get part of a team project and how to work together.
- We learned how we can predict some values at python.
- We learned and practiced how to visualize data.
- We learned to take certain columns of data.
- We learned to a group over the new data received.
- We learned how we can predict some values at python.
- We learned classifications like: “Logistic Regression, KNN, Support Vector Machine, Naive Bayes, Decision Tree and Random Forest”.
- We learned to show the accuracy value of data classified with a confusion matrix.

Presentation Link

https://www.youtube.com/watch?v=qXA1luBRAQw

	Name	Score	Criterion	n_estimator or k value	Metric	Kernel
0	RandomForestClassifier	0.996679	gini	100		
1	RandomForestClassifier	0.996181	gini	150		
2	RandomForestClassifier	0.996181	gini	200		
3	RandomForestClassifier	0.996015	entropy	150		
4	RandomForestClassifier	0.995849	entropy	200		
5	RandomForestClassifier	0.995683	entropy	100		
6	DecisionTreeClassifier	0.993525	entropy			
7	DecisionTreeClassifier	0.990536	gini			
8	KNeighborsClassifier	0.955338		1	euclidean	
9	KNeighborsClassifier	0.955338		3	euclidean	
10	KNeighborsClassifier	0.955338		5	euclidean	
11	KNeighborsClassifier	0.955338		10	euclidean	
12	KNeighborsClassifier	0.955338		1	minkowski	
13	KNeighborsClassifier	0.955338		3	minkowski	
14	KNeighborsClassifier	0.955338		5	minkowski	
15	KNeighborsClassifier	0.955338		10	minkowski	
16	SupportVectorMachine	0.874979				rbf
17	GaussianNaiveBayes	0.871991				
18	SupportVectorMachine	0.846920				poly
19	SupportVectorMachine	0.835796				linear
20	LogisticRegression	0.825668				
21	BernoulliNaiveBayes	0.813382				
22	MultinomialNaiveBayes	0.783995				
23	SupportVectorMachine	0.613814				sigmoid

Figure 8: A table of classification algorithms

REFERENCES

1. https://www.kaggle.com/neuromusic/avocado-prices
2. https://realpython.com/linear-regression-in-python/
3. https://towardsdatascience.com/machine-learning-with-python-regression-complete-tutorial-47268e546cea
4. https://monkeylearn.com/blog/classification-algorithms/
5. https://analyticsindiamag.com/7-types-classification-algorithms/