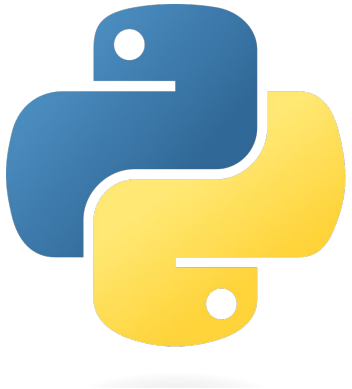
Study of the avocado market in the US between 2015-2018

Mid-bootcamp project
DA - Ironhack
Clara Balcells

1st step: Brainstorming and finding an appropiate Data Set

kaggle



Download Data Set

Import file into Python in order to make data cleaning

2nd step: Data Exploration and Cleaning

- First look at the dataset and exploration

```
In [209]: #We import the file "avocado" from our directory and we read it
          df = pd.read_csv('Data/avocado.csv')

In [210]: df
```

Out[210]:

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18244	7	2018-02-04	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	organic	2018	WestTexNewMexico
18245	8	2018-01-28	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	organic	2018	WestTexNewMexico
18246	9	2018-01-21	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	organic	2018	WestTexNewMexico
18247	10	2018-01-14	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	organic	2018	WestTexNewMexico
18248	11	2018-01-07	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	organic	2018	WestTexNewMexico

18249 rows × 14 columns

- Dropping irrelevant information and renaming columns + More data exploration

In [212]:
```
#We rename some columns in order to standarize them

dict = {'AveragePrice':'Average Price',
        'type':'Type',
        'year':'Year',
        'region':'Region',
        '4046':'Small Avocados',
        '4225':'Large Avocados',
        '4770':'XL Avocados'}

df.rename(columns=dict,
          inplace=True)
```

In [214]:
```
#We get the name of all the regions in the df
Regions = df['Region'].unique()
print (Regions)
```

```
['Albany' 'Atlanta' 'BaltimoreWashington' 'Boise' 'Boston'
 'BuffaloRochester' 'California' 'Charlotte' 'Chicago' 'CincinnatiDayton'
 'Columbus' 'DallasFtWorth' 'Denver' 'Detroit' 'GrandRapids' 'GreatLakes'
 'HarrisburgScranton' 'HartfordSpringfield' 'Houston' 'Indianapolis'
 'Jacksonville' 'LasVegas' 'LosAngeles' 'Louisville' 'MiamiFtLauderdale'
 'Midsouth' 'Nashville' 'NewOrleansMobile' 'NewYork' 'Northeast'
 'NorthernNewEngland' 'Orlando' 'Philadelphia' 'PhoenixTucson'
 'Pittsburgh' 'Plains' 'Portland' 'RaleighGreensboro' 'RichmondNorfolk'
 'Roanoke' 'Sacramento' 'SanDiego' 'SanFrancisco' 'Seattle'
 'SouthCarolina' 'SouthCentral' 'Southeast' 'Spokane' 'StLouis' 'Syracuse'
 'Tampa' 'TotalUS' 'West' 'WestTexNewMexico']
```

In [215]:
```
#We need to remove the TotalUs values since they can be misleading when it comes to analyse the dataframe, also we don'
df = df[df['Region']!="TotalUS"]
```

In [216]:
```
#We now see that the value 'Total Us' has been removed from the dataset.
Regions = df['Region'].unique()
print (Regions)
```

```
['Albany' 'Atlanta' 'BaltimoreWashington' 'Boise' 'Boston'
 'BuffaloRochester' 'California' 'Charlotte' 'Chicago' 'CincinnatiDayton'
 'Columbus' 'DallasFtWorth' 'Denver' 'Detroit' 'GrandRapids' 'GreatLakes'
 'HarrisburgScranton' 'HartfordSpringfield' 'Houston' 'Indianapolis'
 'Jacksonville' 'LasVegas' 'LosAngeles' 'Louisville' 'MiamiFtLauderdale'
 'Midsouth' 'Nashville' 'NewOrleansMobile' 'NewYork' 'Northeast'
 'NorthernNewEngland' 'Orlando' 'Philadelphia' 'PhoenixTucson'
 'Pittsburgh' 'Plains' 'Portland' 'RaleighGreensboro' 'RichmondNorfolk'
 'Roanoke' 'Sacramento' 'SanDiego' 'SanFrancisco' 'Seattle'
 'SouthCarolina' 'SouthCentral' 'Southeast' 'Spokane' 'StLouis' 'Syracuse'
 'Tampa' 'West' 'WestTexNewMexico']
```

In [217]:
```
#We can also delete de "Unnamed: 0" column
del(df['Unnamed: 0'])
df
```

Before vs After

avocado

	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2015-12-13	0.93	118220.22	794.7	109149.67	130.5	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	2015-12-06	1.08	78992.15	1132.0	71976.41	72.58	5811.16	5677.4	133.76	0.0	conventional	2015	Albany

avocadoclean

Date	Average Price	Total Volume	Small Avocados	Large Avocados	XL Avocados	Total Bags	Small Bags	Large Bags	XLarge Bags	Type	Year	Region
2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2015-12-13	0.93	118220.22	794.7	109149.67	130.5	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
2015-12-06	1.08	78992.15	1132.0	71976.41	72.58	5811.16	5677.4	133.76	0.0	conventional	2015	Albany
2015-11-29	1.28	51039.6	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany

3rd step: import into SQL in order to make data analysis

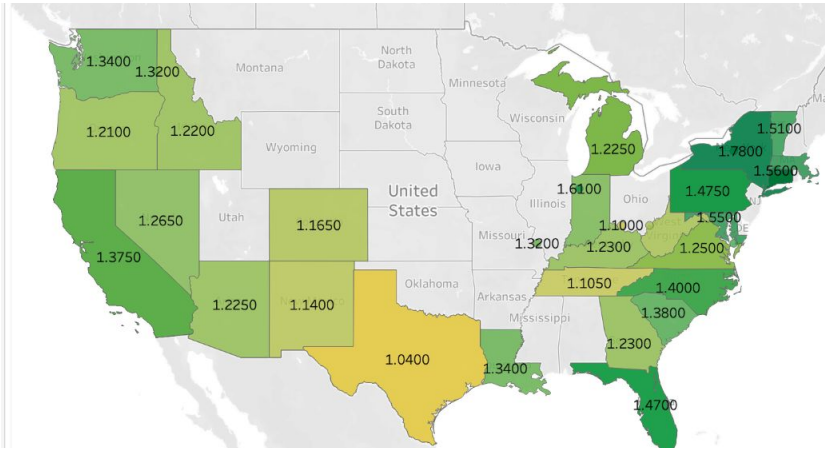
```
df.to_csv ('avocadoclean.csv', index=False)
```

```sql
#Import and check that the table works properly
SELECT * FROM AVOCADOCLEAN;

#The lowest average price ever ever for an avocado

SELECT concat(`average price`, " ",Year, " ", Region) AS 'Lowest Price ever All Regions'
FROM avocadoclean
where `Average Price` = (SELECT min(`Average Price`) from avocadoclean)
order by `Average Price` asc;

#The lowest average price ever has been in 2017, in Cincinnati Dayton, with an average price of 0.44 dollars



#Highest average price ever for an avocado
SELECT concat(`average price`, " ",Year, " ", Region) AS 'Highest Price ever All Regions'
FROM avocadoclean
where `Average Price` = (SELECT max(`Average Price`) from avocadoclean)
order by `Average Price` desc;
#The highest average price ever  has been in 2017, in San Francisco, with an average price of 3.25 dollars



#From all the regions, where can we find the one with the highest average and the lowest prices in 2015'
SELECT concat(`average price`, " ",Year, " ", Region) AS 'Region with Highest Average Price 2015'
FROM avocadoclean
where `year` = 2015
order by `Average Price` desc;

SELECT concat(`average price`, " ",Year, " ", Region) AS 'Region with Highest Average Price 2015'
FROM avocadoclean
where `year` = 2015
order by `Average Price` asc;
#The highest average price was in San Francisco and the lowest in Phoenix Tuscon
```

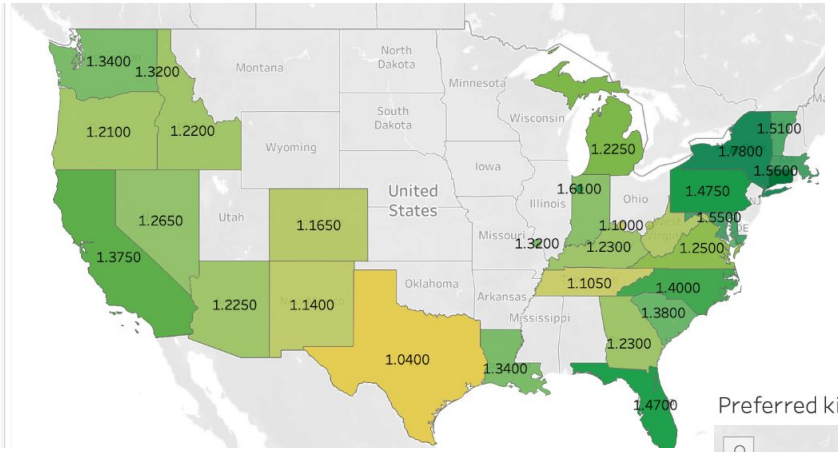
```sql
#Which is the general preferred size of avocado ever amongst the 3 types
SELECT sum(`Small Avocados`), sum(`Large Avocados`), sum(`XL Avocados`)AS 'Preferred size of avocados'
FROM avocadoclean;
#The preferred size of avocados are the Large ones


#Which is the preferred size of avocado ever amongst the 3 types in each region
SELECT round(sum(`Small Avocados`),2) as Small_total, round(sum(`Large Avocados`),2) as Large_total, round(sum(`XL Avocados`),2) as XL_total, Region AS 'Preferred size of avocados per region
CASE
  WHEN ( sum(`Small Avocados`) > sum(`Large Avocados`) ) AND ( sum(`Small Avocados`) > sum(`XL Avocados`) ) THEN "Small"
  WHEN ( sum(`Large Avocados`) > sum(`Small Avocados`) ) AND ( sum(`Large Avocados`) > sum(`XL Avocados`) ) THEN "Large"
  ELSE "XL"
END as "Best Seller"
FROM avocadoclean
GROUP BY Region;



#Which is the most sold kind of bag
SELECT sum(`Small Bags`), sum(`Large Bags`), sum(`XLarge Bags`) AS 'Preferred type of bags'
FROM avocadoclean;
#The most sold kind of bag are the Small Bags



#Which is the most sold kind of bag per region
SELECT round(sum(`Small Bags`),2) as Small_total, round(sum(`Large Bags`),2) as Large_total, round(sum(`XLarge Bags`),2) as XL_total, Region AS 'Preferred type of bags per region',
CASE
  WHEN ( sum(`Small Bags`) > sum(`Large Bags`) ) AND ( sum(`Small Bags`) > sum(`XLarge Bags`) ) THEN "Small Bags"
  WHEN ( sum(`Large Bags`) > sum(`Small Bags`) ) AND ( sum(`Large Bags`) > sum(`XLarge Bags`) ) THEN "Large Bags"
  ELSE "XLarge Bags"
END as "Best Seller Kind of Bag"
FROM avocadoclean
GROUP BY Region;
```

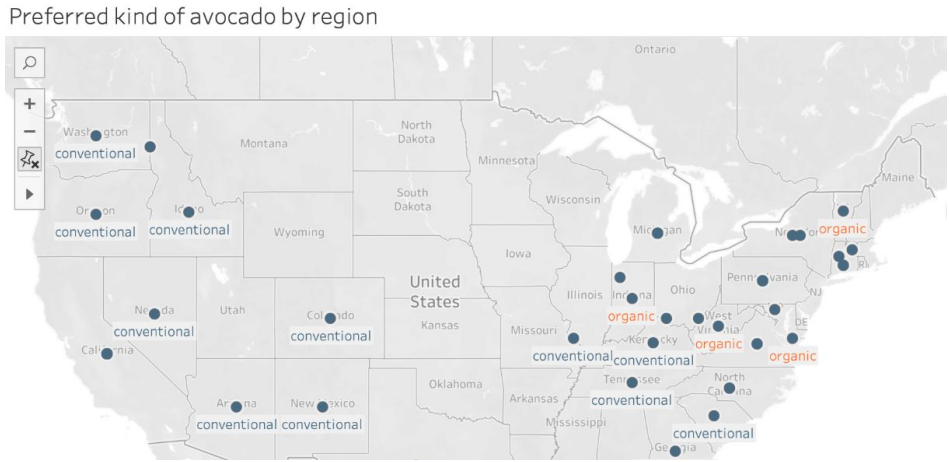
4th Step: Tableau

4th Step: Tableau

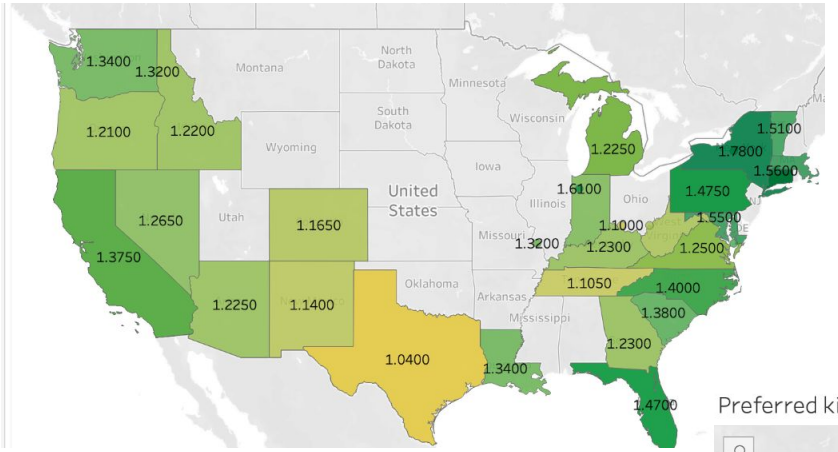


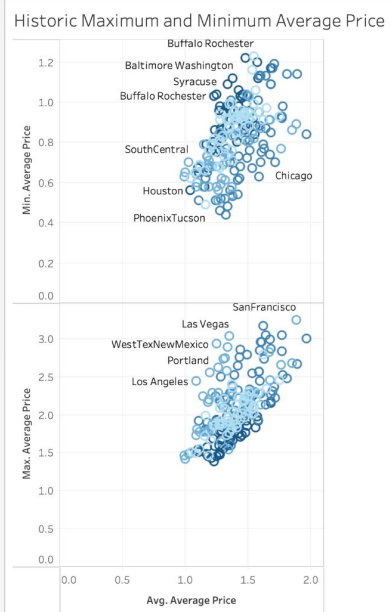
4th Step: Tableau

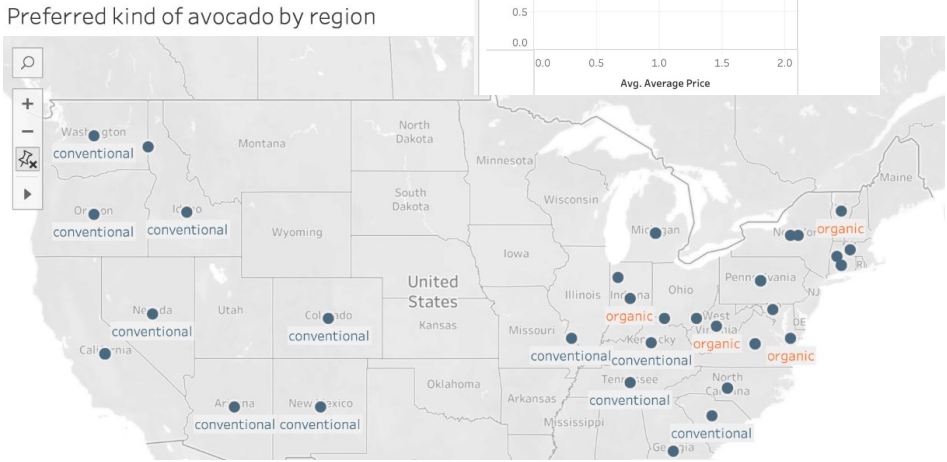




4th Step: Tableau



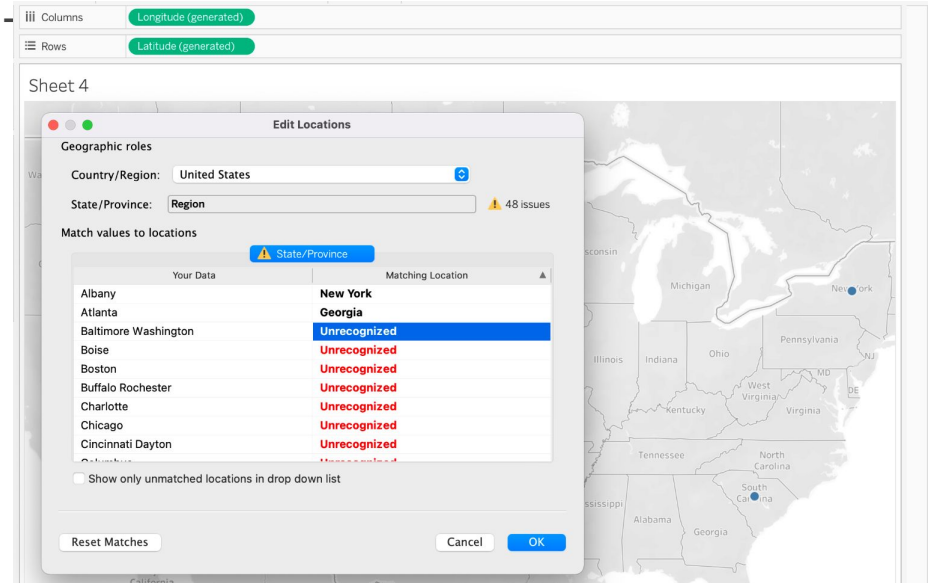


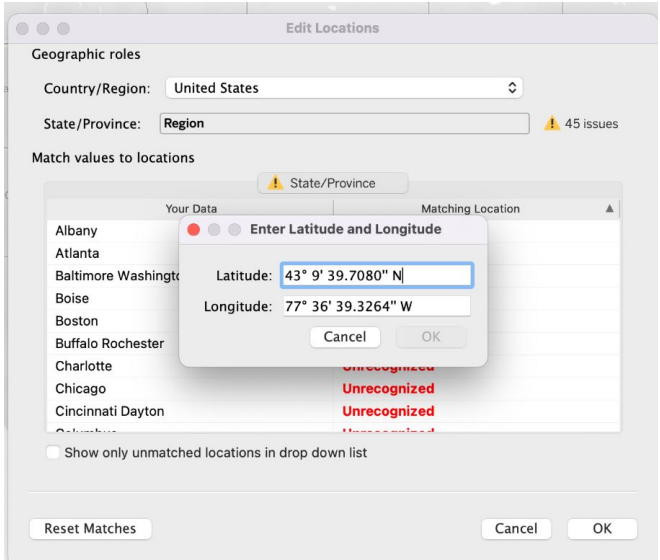


Link to Tableau profile:
https://public.tableau.com/app/profile/clara6389

Problems encountered

- Tableau not recognizing certain regions
- Having to run everything and download again the Jupyter Notebook, just because I found another irrelevant column/value
- Time management





¡Thank you all for your attention!