COFFEE QUALITY

Predicting the Quality of Coffee

27.11.2021

AGENDA

Data

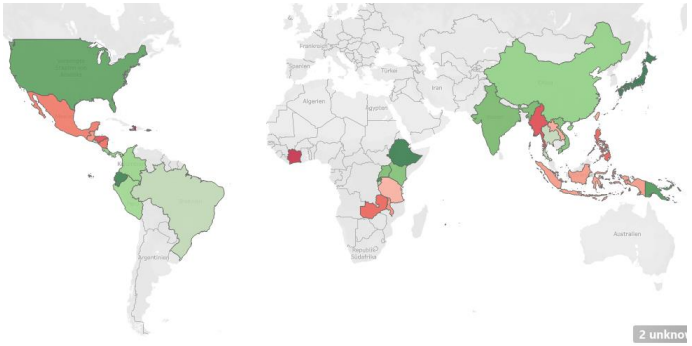
Introducing the dataset

Modelling

Exploring different models

Results

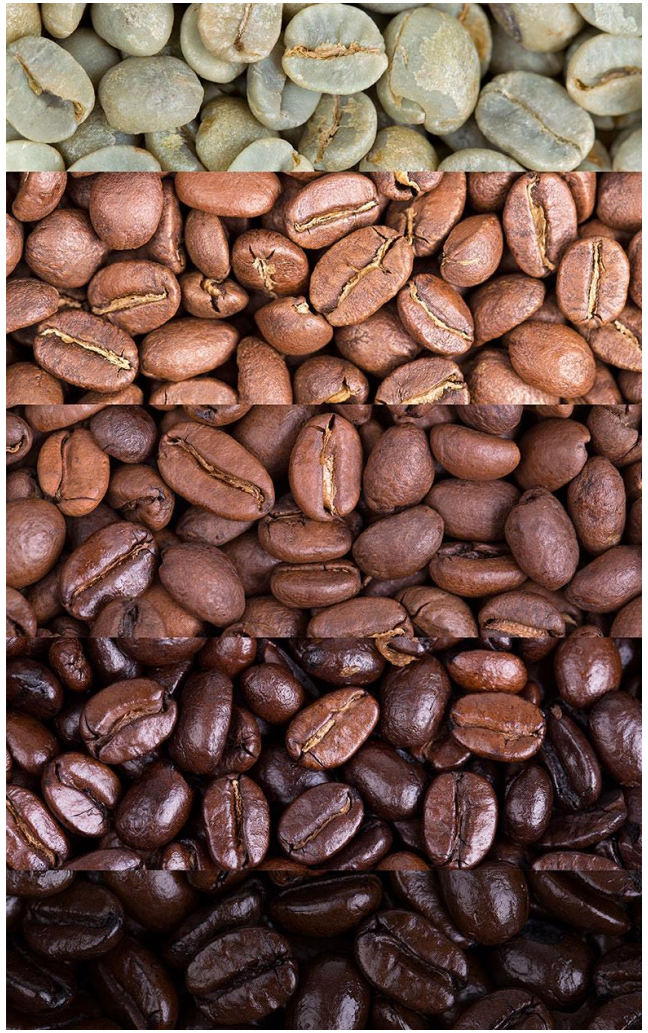
Showing results & visualizations



DATA

- Coffee Quality database from Coffee Quality Institute
- Gathered January 2018
- From kaggle
- A little bit pre cleaned
- Shape: 1339 rows x 44 columns
- Information:
    - Taste (aroma, sweetness, acidity…)
    - Bean (color, species, processing method, defects)
    - Farm (country, owner, altitude…)

➢ **Goal**: Predicting the quality of coffee based on columns of this dataset



DATA CLEANING

Missing data

```
data1.isna().sum()
```

```
species                   0
owner                     7
country.of.origin         1
harvest.year             47
variety                 226
processing.method       170
aroma                     0
flavor                    0
aftertaste                0
acidity                   0
body                      0
balance                   0
uniformity                0
clean.cup                 0
sweetness                 0
cupper.points             0
moisture                  0
category.one.defects      0
quakers                   1
color                   218
category.two.defects      0
altitude_mean_meters    230
dtype: int64
```

Random values

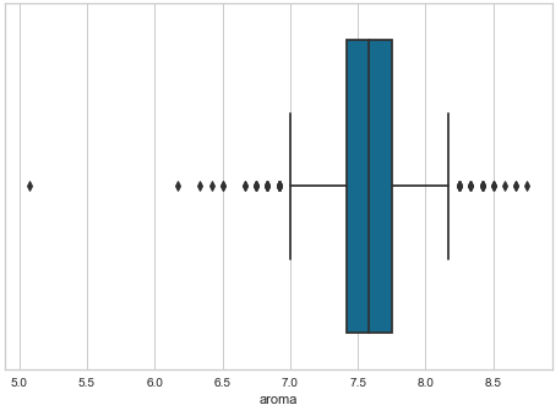
```
data1['harvest.year'].value_count
```

```
2012                          354
2014                          233
2013                          181
2015                          129
2016                          124
2017                           70
2013/2014                      29
2015/2016                      28
2011                           26
2014/2015                      19
2017 / 2018                    19
2009/2010                      12
2010                           10
2016 / 2017                     6
2010-2011                       6
4T/10                           4
March 2010                      3
2009-2010                       3
4T/2010                         3
Mayo a Julio                    3
2011/2012                       2
08/09 crop                      2
January 2011                    2
Abril - Julio                   2
TEST                            1
4t/2011                         1
May-August                      1
23 July 2010                    1
2009 - 2010                     1
4t/2010                         1
4T72010                         1
1T/2011                         1
47/2010                         1
Fall 2009                       1
August to December              1
2018                            1
1t/2011                         1
Spring 2011 in Colombia.        1
3T/2011                         1
Abril - Julio /2011             1
Sept 2009 - April 2010          1
December 2009-March 2010        1
2009 / 2010                     1
2016/2017                       1
January Through April           1
mmm                             1
```

Too many different values

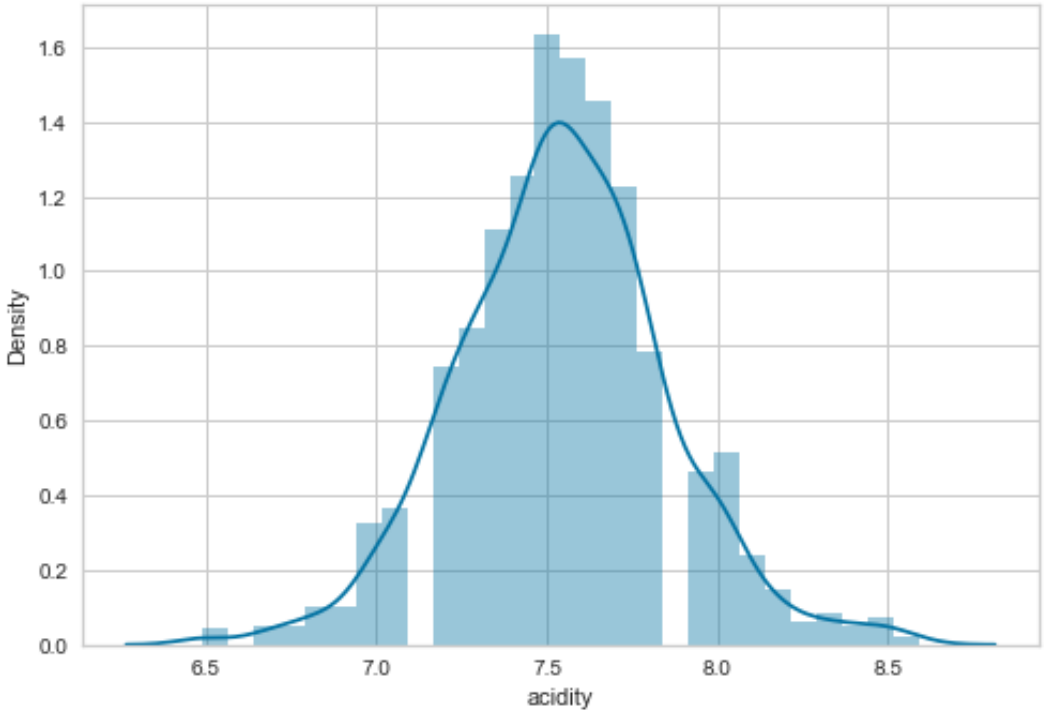
```
juan luis alvarado romero          154
racafe & cia s.c.a                  51
exportadora de cafe condor s.a      50
ipanema coffees                     50
cqi taiwan icp cqi台灣合作夥伴              47
                                   ...
jesus carlos cardenas valdivia       1
francisco hernandez lorenzo          1
immaculata john                      1
damaso martinez perez                1
wayner jimenez                       1
Name: owner, Length: 309, dtype: int64
```

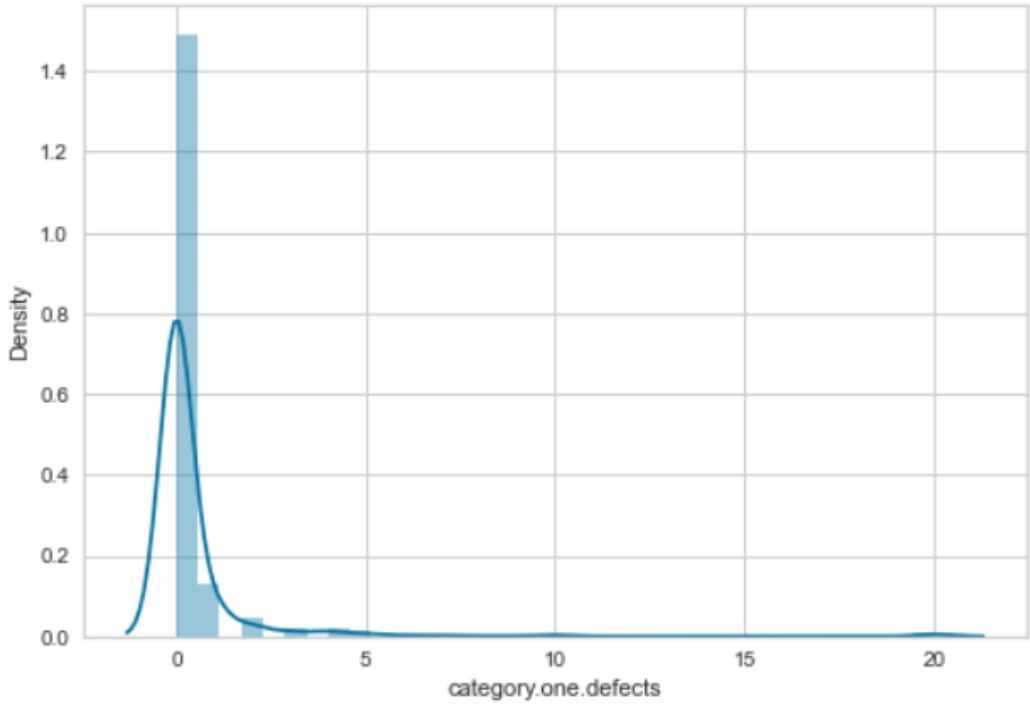
Outliers



DATA EXPLORATION

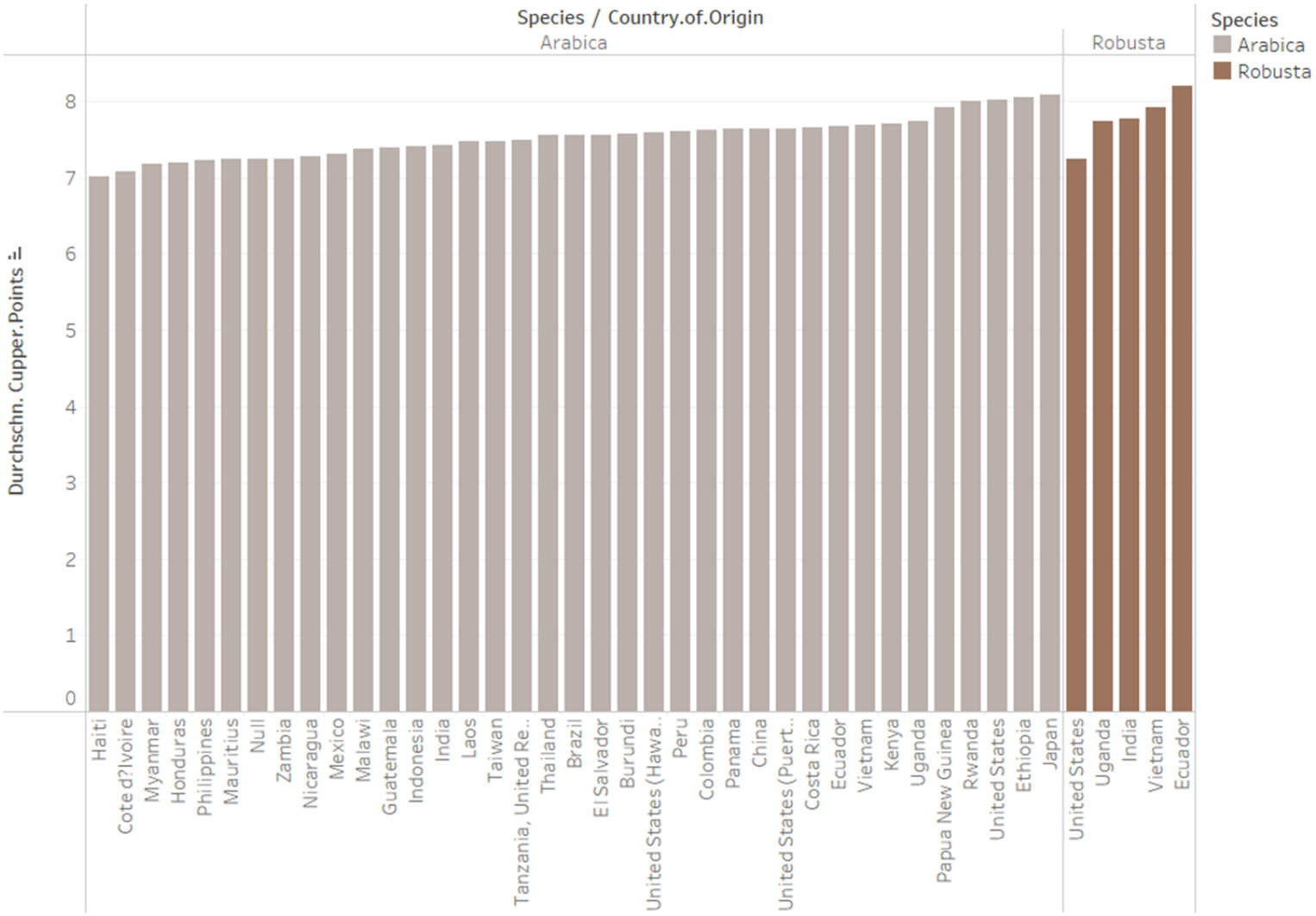
- Most columns have a (nearly) normal distribution after data cleaning
- For some columns the outliers are important





DATA EXPLORATION

- Difference of the species (arabica/ robusta)
- There are only 5 country where robusta is coming from
- ‘Best’ arabica from Japan
- ‘Worst’ arabica from Haiti
- ‘Best’ robusta from Ecuador
- ‘Worst’ robusta from US



MODELLING RESULTS

R^2 Score

Dataset		Linear Regression	Decision Tress Regressor	KNN Regressor	Random Forest Regressor
Standard Scaler	categorical & numerical	0.50582	0.36629	0.59494	0.67772
	Only numerical	0.59236	0.06086	0.53835	0.6829
Normalizer	categorical & numerical	-0.02793	-0.44514	0.19808	0.28889
	Only numerical	-0.28452	0.02685	0.06487	0.29781

RANDOM FOREST REGRESSOR

Further investigations

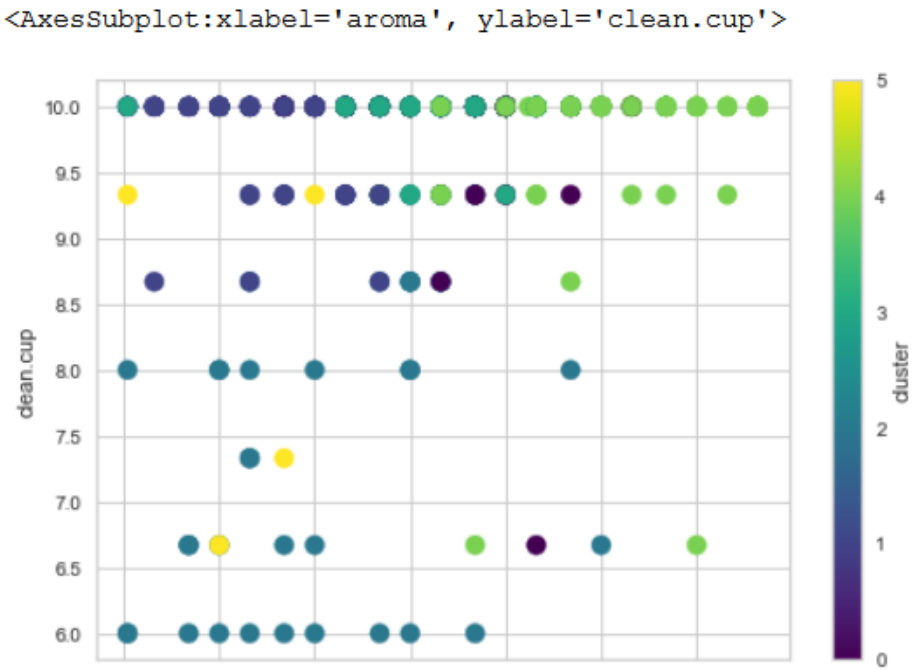
Most important Columns:

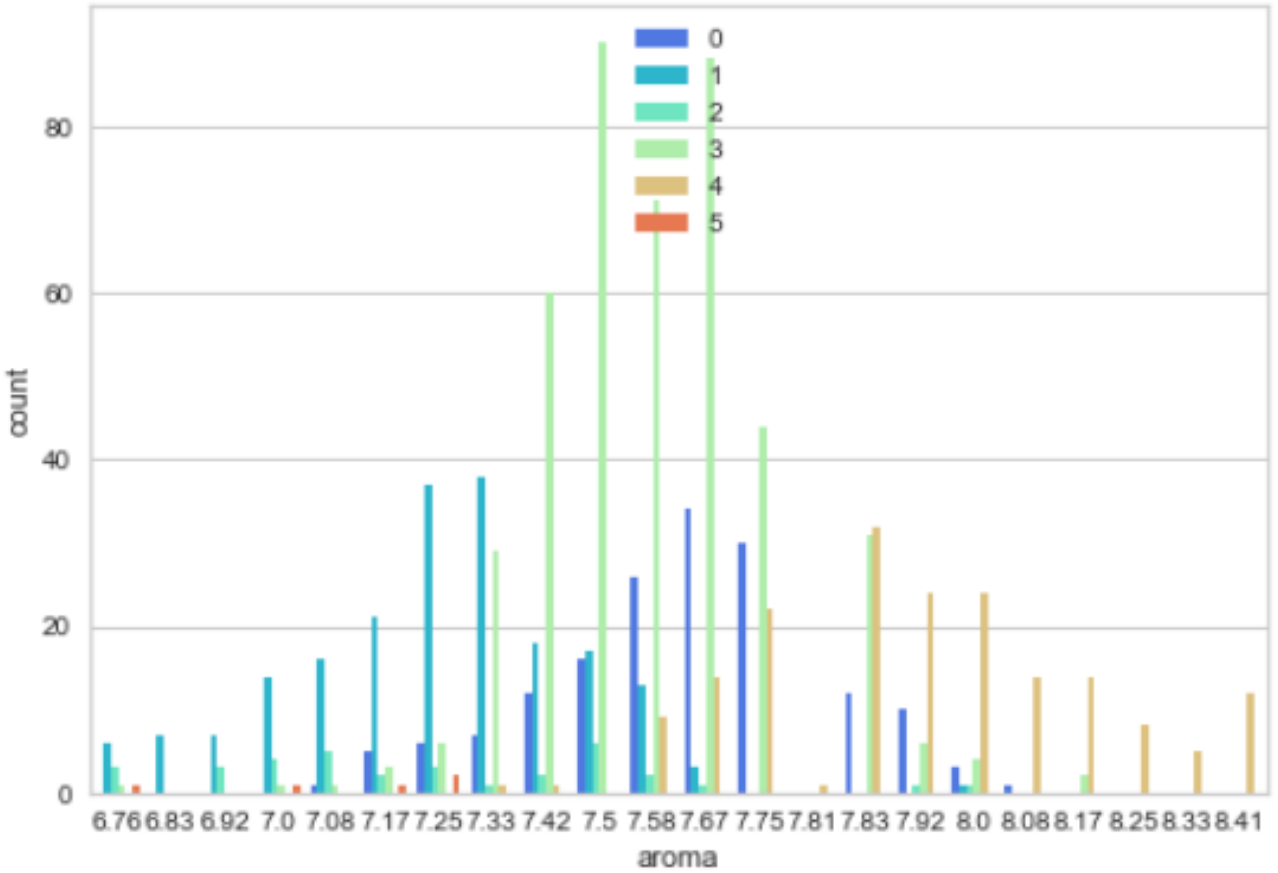
- Column 1 → Flavor
- Column 2 → Aftertaste
- Column 4 → Balance

```
Variable:                                  1 Importance: 0.37
Variable:                                  2 Importance: 0.25
Variable:                                  4 Importance: 0.19
Variable: owner_aulia arif syahri Importance: 0.04
Variable:                                  0 Importance: 0.02
Variable: harvest.year_2017    Importance: 0.02
Variable:                                  3 Importance: 0.01
Variable:                                 11 Importance: 0.01
Variable: owner_cqi taiwan icp cqi台灣合作夥伴 Importance: 0.01
Variable: owner_lydiah mwangi  Importance: 0.01
Variable: country.of.origin_Vietnam Importance: 0.01
Variable: variety_SL28         Importance: 0.01
Variable: processing.method_Other Importance: 0.01
Variable: color_Green          Importance: 0.01
Variable:                                  5 Importance: 0.0
Variable:                                  6 Importance: 0.0
Variable:                                  7 Importance: 0.0
Variable:                                  8 Importance: 0.0
Variable:                                  9 Importance: 0.0
Variable:                                 10 Importance: 0.0
Variable: species_Robusta      Importance: 0.0
Variable: owner_alejandro garcia palacios Importance: 0.0
Variable: owner_alfredo bojalil Importance: 0.0
Variable: owner_andreas kussmaul Importance: 0.0
Variable: owner_andrew hetzel  Importance: 0.0
Variable: owner_armando luis pohlenz martinez Importance: 0.0
Variable: owner_bismarck castro Importance: 0.0
Variable: owner_bourbon specialty coffees Importance: 0.0
Variable: owner_brent hall     Importance: 0.0
Variable: owner_cadexsa        Importance: 0.0
Variable: owner_cafe de don balbino s.c. de r.l. de c.v. Importance: 0.0
Variable: owner_cafe politico  Importance: 0.0
Variable: owner_cafebras       Importance: 0.0
```

CLUSTERING

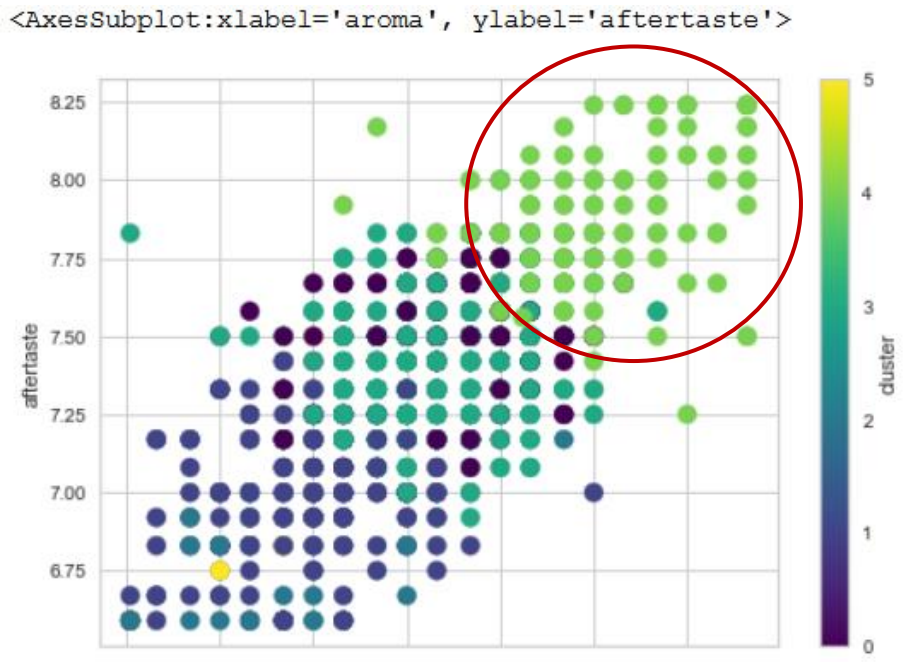
- 6 clusters
- Not all clear defined

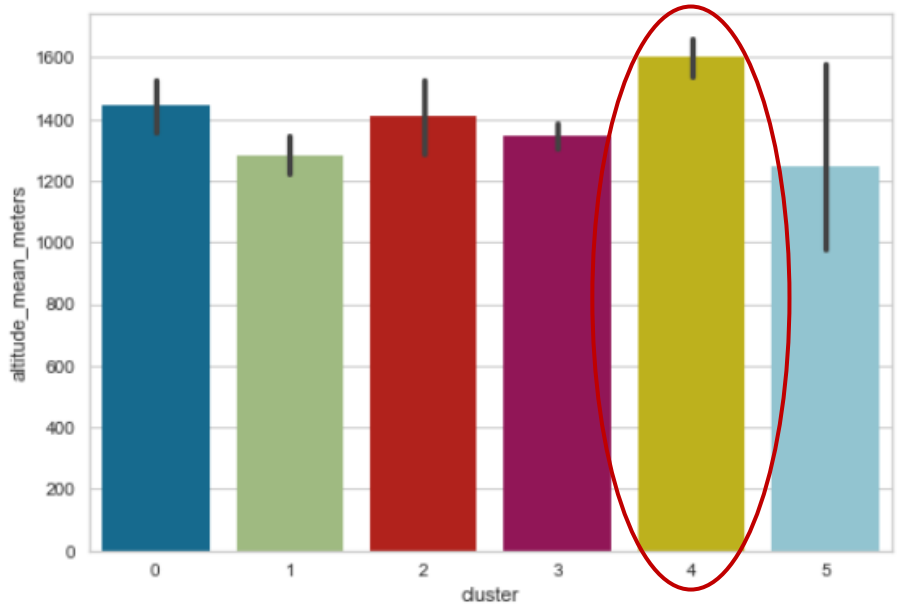




CLUSTERING

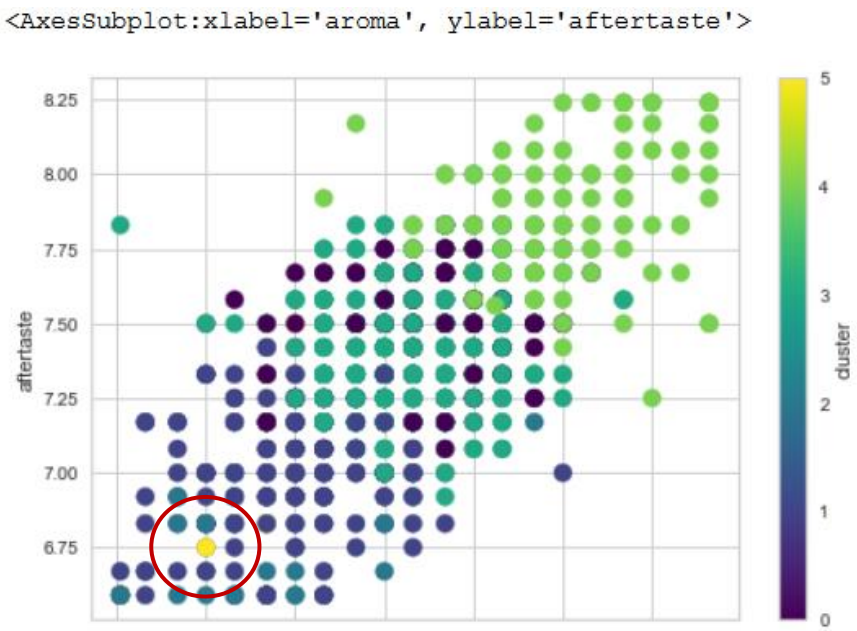
- Cluster 4 = Good coffee
- good flavor profile
- Grown in high altitude -> rich aroma

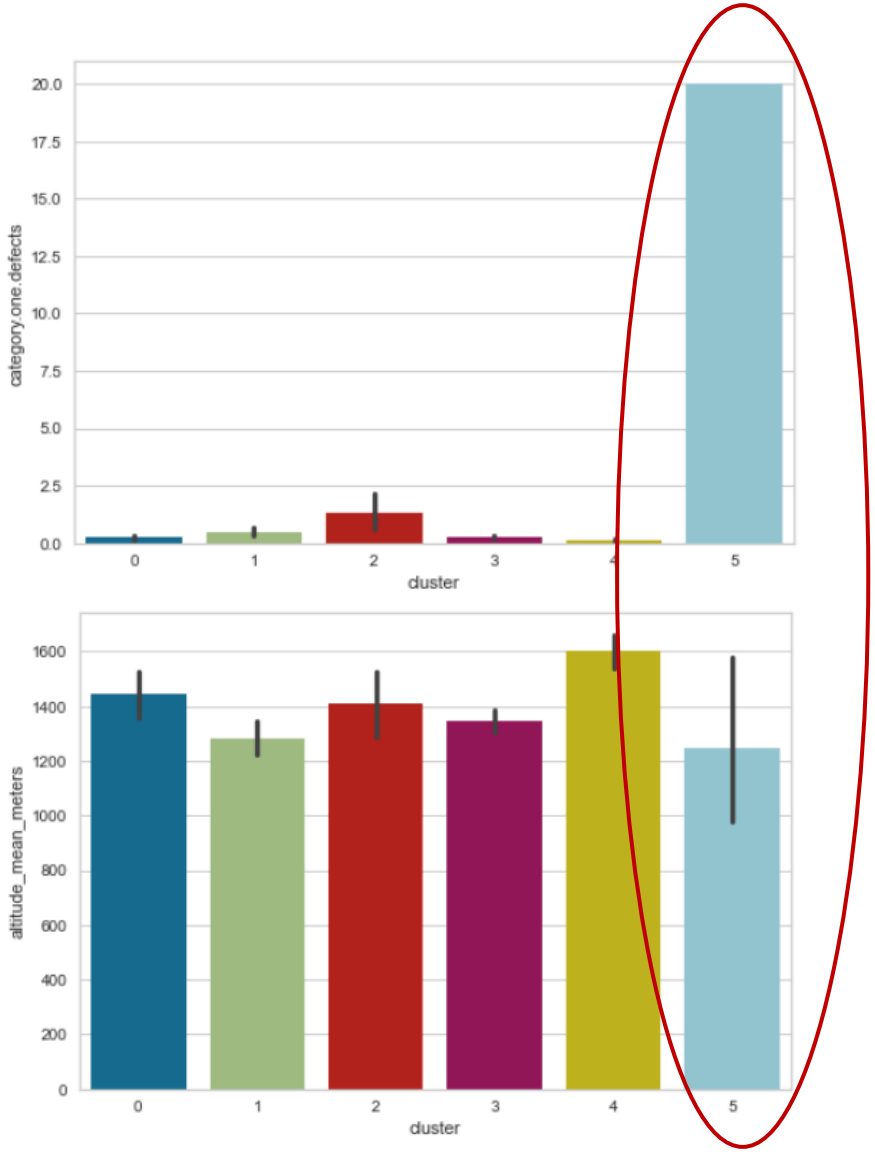




CLUSTERING

- Cluster 5 = Bad coffee
- Low aroma & aftertaste, grown in low altitude
- A lot of defects





THANK YOU!

AND ENJOY YOUR COFFEE!