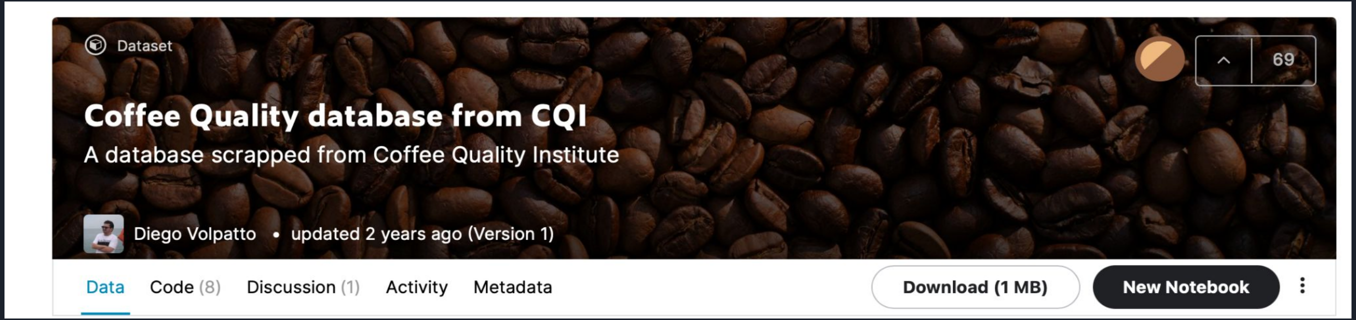
Coffee quality: data analysis of coffee quality database from CQI (institute)

What qualities make a good cup of coffee?

Dataset



Dataset

Coffee Quality database from CQI

A database scrapped from Coffee Quality Institute

Diego Volpatto • updated 2 years ago (Version 1)

- Dataset (from Kaggle) about coffee quality.
- Data gathered from Coffee Quality Institute (CQI) in January, 2018.
- 3 csv files
- Initially: 44 columns, 1339 values (including NaN and null)
- Inside the dataset: Quality Measures, Bean Metadata,Farm Metadata
- One row= one 2kg sample of green coffee
- One column = one categorical or numeric value

Limits and issues

- Unbalanced representativity in samples: more arabica and less robusta
- Difficulty in understanding the meaning of some column names (specific to coffee)
- Minimum grading : 80 to be certified, while some samples are below the target
- Biased analysis: the grade depending on individual perception and taste of each expert regarding each coffee sample.

Steps and tools

Exploratory analysis with tableau:

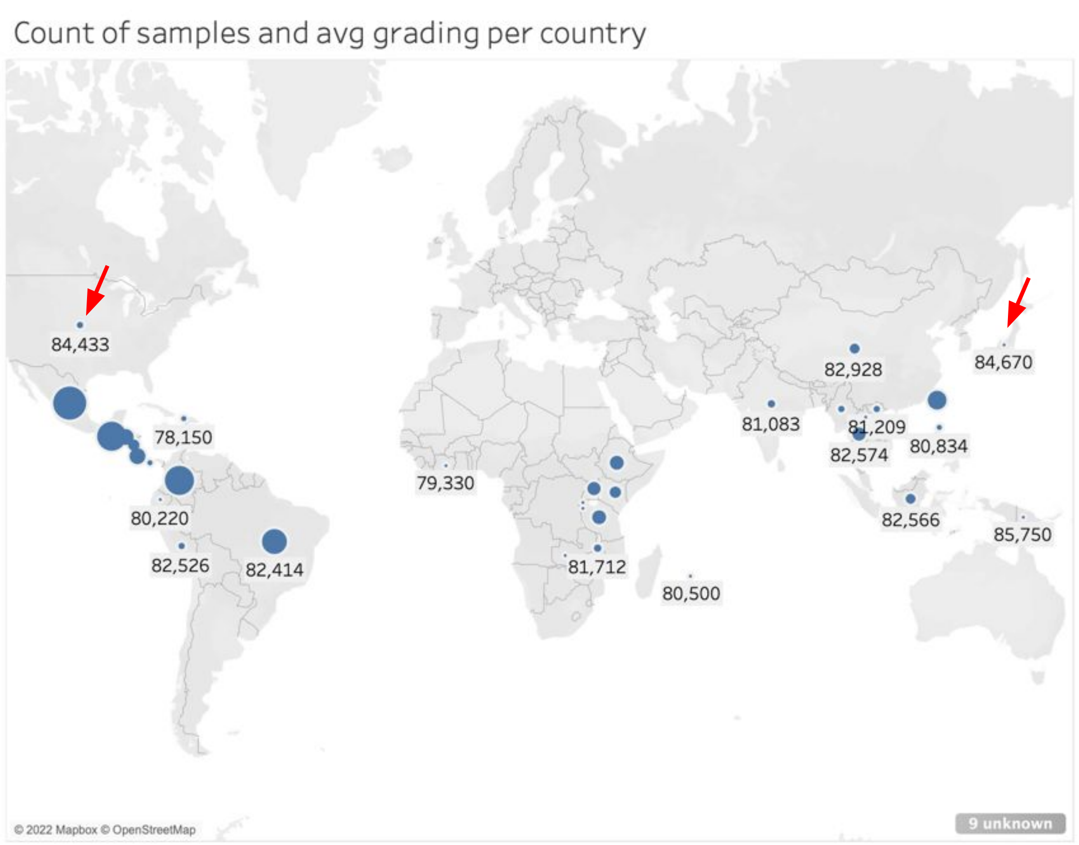
- merged csv with all of data has been used
- geographical maps with **farm metadata**

Exploratory analysis with pandas profiling in python:

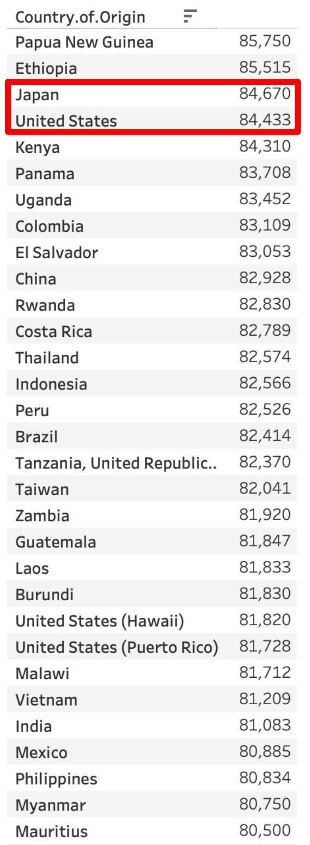
Data cleaning process on the merged csv:

- useless columns
- rows with null values
- columns with a lot of NaN values
- rows with ‘ft’ as measurement
- analysis on **quality measures data** compared with total cup points

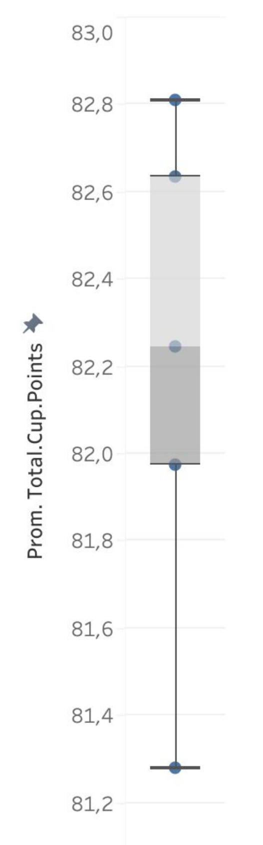
Insights (1) Analysis with tableau



Count of samples and avg grading per country



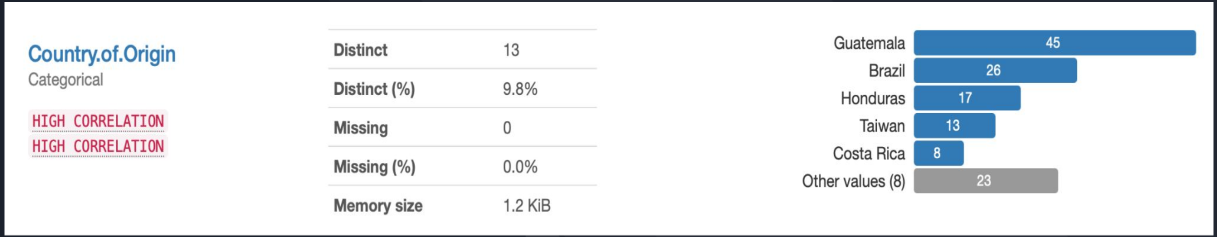
Country.of.Origin	
Papua New Guinea	85,750
Ethiopia	85,515
Japan	84,670
United States	84,433
Kenya	84,310
Panama	83,708
Uganda	83,452
Colombia	83,109
El Salvador	83,053
China	82,928
Rwanda	82,830
Costa Rica	82,789
Thailand	82,574
Indonesia	82,566
Peru	82,526
Brazil	82,414
Tanzania, United Republic..	82,370
Taiwan	82,041
Zambia	81,920
Guatemala	81,847
Laos	81,833
Burundi	81,830
United States (Hawaii)	81,820
United States (Puerto Rico)	81,728
Malawi	81,712
Vietnam	81,209
India	81,083
Mexico	80,885
Philippines	80,834
Myanmar	80,750
Mauritius	80,500



Insights (2) Univariate analysis with Python

High frequency :

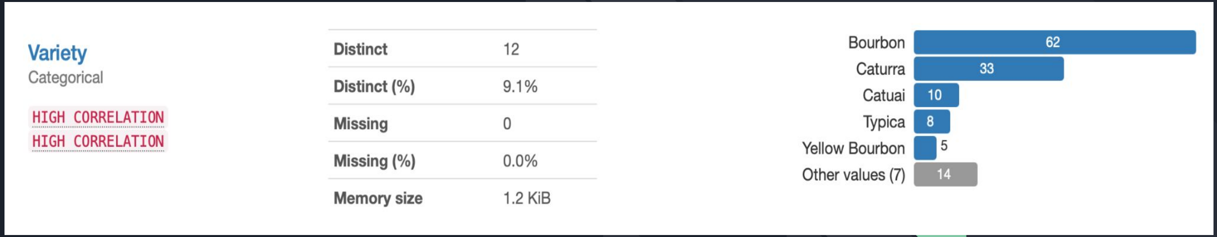
- country of Origin (Guatemala 34, 1%)
- Variety (Bourbon 47%)
- Harvest Year (2016 42,4%
- Processing Method (Washed/Wet 66,7%)



Country.of.Origin
Categorical

HIGH CORRELATION
HIGH CORRELATION

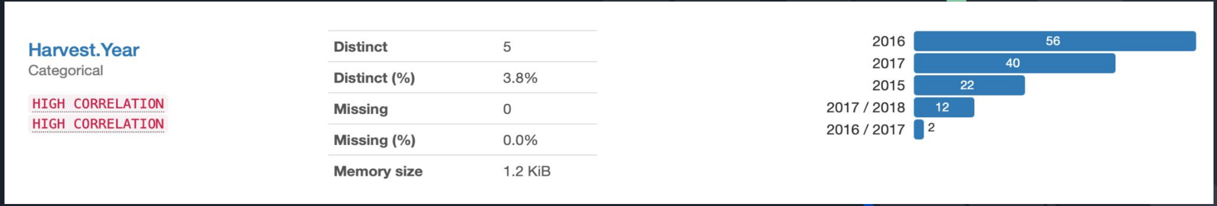
Distinct	13
Distinct (%)	9.8%
Missing	0
Missing (%)	0.0%
Memory size	1.2 KiB



Variety
Categorical

HIGH CORRELATION
HIGH CORRELATION

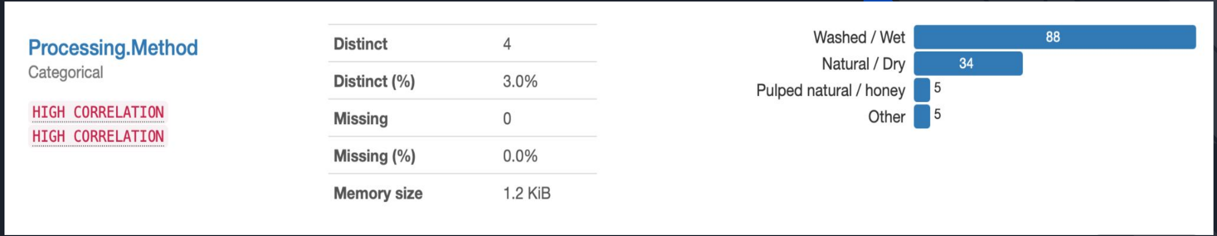
Distinct	12
Distinct (%)	9.1%
Missing	0
Missing (%)	0.0%
Memory size	1.2 KiB



Harvest.Year
Categorical

HIGH CORRELATION
HIGH CORRELATION

Distinct	5
Distinct (%)	3.8%
Missing	0
Missing (%)	0.0%
Memory size	1.2 KiB



Processing.Method
Categorical

HIGH CORRELATION
HIGH CORRELATION

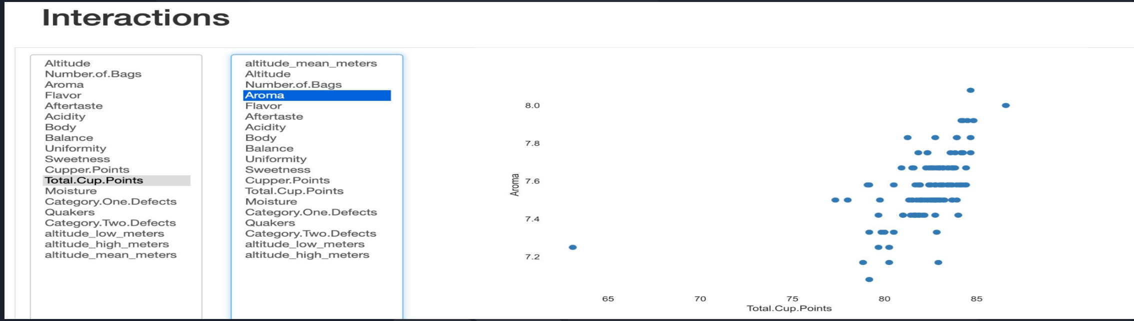
Distinct	4
Distinct (%)	3.0%
Missing	0
Missing (%)	0.0%
Memory size	1.2 KiB

Insights(3)

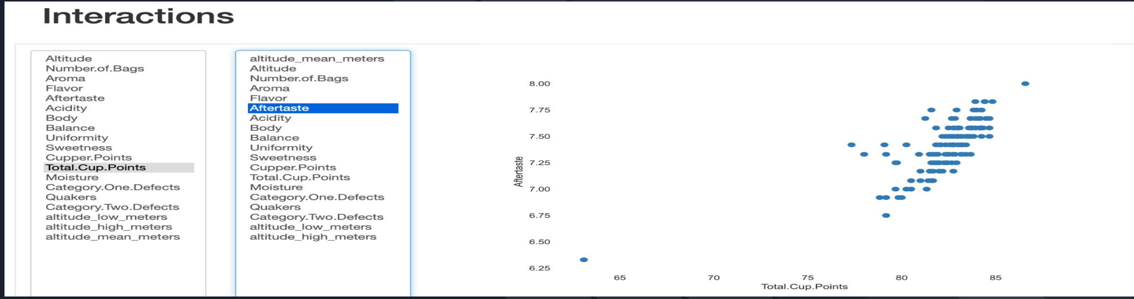
Multivariate Analysis with python

- Numeric variables: 19
- Arabica samples: 98,5%
- Robusta samples: 1,5%

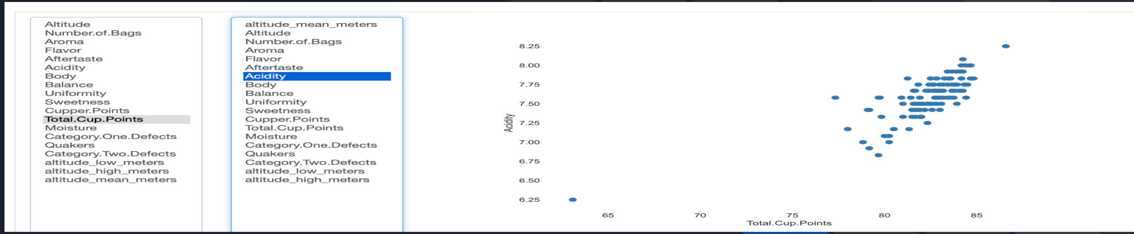
Interactions (total.cup.points) with aroma, flavor, aftertaste, acidity, body and balance



Interactions



Interactions



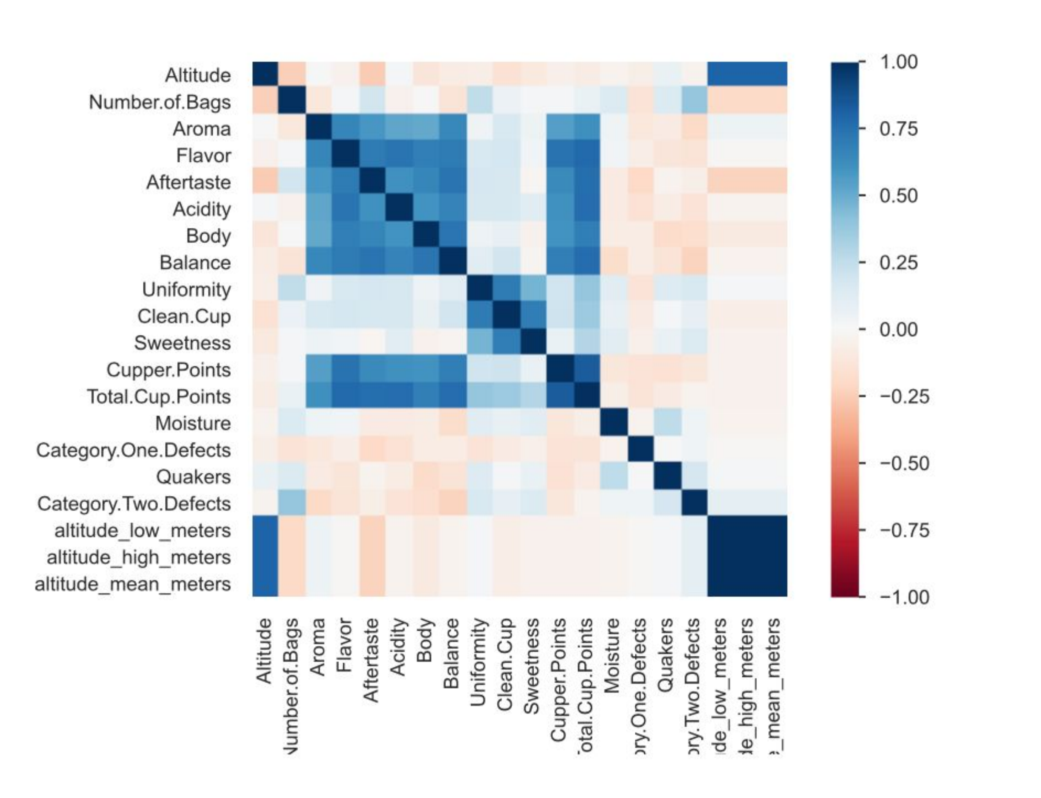
No interaction with uniformity and sweetness

Insights(4)

Multivariate Analysis with python

- Numeric variables: 19
- Arabica samples: 98,5%
- Robusta samples: 1,5%

Correlations (total.cup.points) with aroma, flavor, aftertaste, acidity, body and balance



No correlation with uniformity and sweetness

Executive summary

Question

- Identify variables that influence the quality of coffee

Data processing

- EDA with tableau public
- Cleaning data and EDA in python with pandas

Analysis Conclusion

- a few countries produce the most coffee beans in the world (South American and African countries) with score > 80
- less given samples does not mean a lower grade (case of the USA and Japan)
- The washed/wet method mainly used to process coffee beans
- Aroma, flavor, aftertaste, acidity, body and balance also influence the quality of coffee
- Uniformity and sweetness do not influence the final grade

Going further

- Why are less bags sent for analysis year after year?
- Why are sweetness and uniformity (quality measures) not considered as relevant variables for the total cup points? (see EDA)
- Why are certain years much better in terms of harvest than others?

In a larger way...

- What is the motivation behind the grading?
- Why are less bags sent for analysis year after year?
- Are there barriers for farmers to incentivize grading?

Thank you!