JADS Jheronimus Academy of Data Science

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Wine Quality Dataset

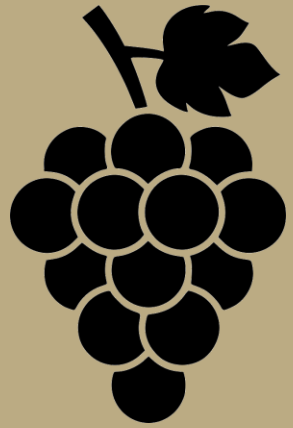
An Overview

Introduction to Data Science Elective

On behalf of the PDEng Data Science

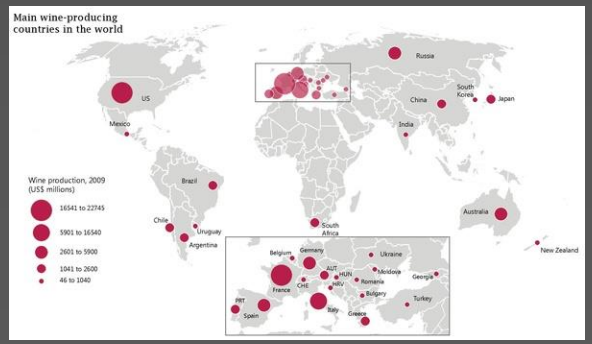
4 November 2019

Introduction



Overview

- Wine is nowadays consumed by a wide variety of customers
- Portugal is the 11th worldwide producer of wine
- Many varieties of wine exist
- "Vinho Verde" is one such variety



Source: https://www.slideshare.net/GautamSawant1/predicting-wine-quality-using-data-analytics

Overview

- Wine certification and quality control are key elements for the wine industry.
- They prevent adulteration and promote quality.
- The wines that conform with regulated standards receive the label of **Denomination of Controlled Origin**

(e.g., soil characteristics, grape varieties, vinification, and bottling)

Vinho Verde

- Vinho Verde is Portugal’s largest wine region
- Vinho Verde does not mean "green" wine. "Verde" refers to it being a young wine.
- Some facts…

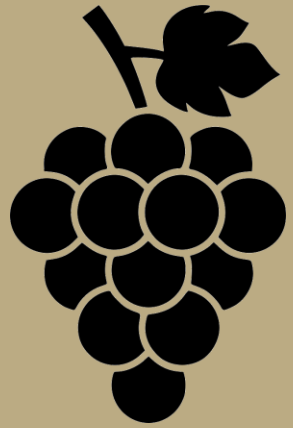
51.000
Acres Vineyards

45
Indigenous Grape Varieties

19.000
Grape Growers

2000
Year History of Wine Making

Problem Background



Grape Varieties

White Grapes

Alvarinho

Arinto

Avesso

Trajadura

Red Grapes

Espadeiro

Padeiro

Vinhão

Wine Making Process

Harvest Grapes

Pressing

Fermentation

Filtration

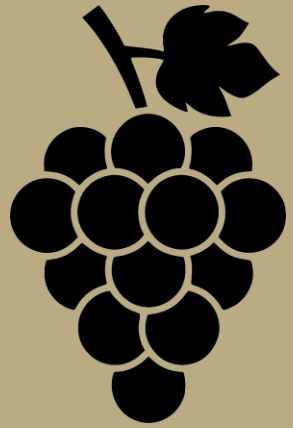
Aging

Bottling

Consumption

Grape Harvesting (“Vindima”)

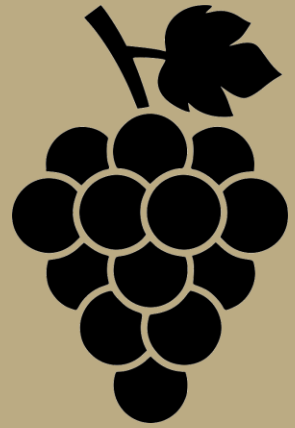
Business Understanding



Business Questions

- What are the characteristics of the wine that determine its quality? How are these related?
- Can we predict the wine quality based on results of physiochemical tests?
- What else?

Data Understanding



Wine Quality Dataset

- Data on the red and white varieties of the Portuguese **Vinho Verde** wine.
- Physiochemical tests results

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. **Modeling wine preferences by data mining from physicochemical properties**.

Wine Quality Dataset

- Number of Instances:
  - red wine – 1599
  - white wine – 4898

  (combined)
- 11 attributes
- Quality (score between 0 and 10)

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. **Modeling wine preferences by data mining from physicochemical properties**.

Wine Quality Dataset

Features:

Fixed acidity	Total sulfur dioxide
Volatile acidity	Density
Citric acid	pH
Residual sugar	Sulphates
Chlorides	Alcohol
Free sulfur dioxide	

Label: Quality Score

Attribute Description (1)

Attribute	Description	Units
Fixed acidity	Acidity is the fundamental property of wine, imparting sourness and resistance to microbial infection.	g/dm^3
Volatile acidity	Wine spoilage is legally defined by volatile acidity.	g/dm^3
Citric acid	Citric acid elicits antimicrobial activity against some molds and bacteria.	g/dm^3
Residual sugar	Sugar remaining after fermentation stops.	g/dm^3
Chlorides	Sodium chloride.	g/dm^3
Free sulfur dioxide	Free sulfites.	g/dm^3

Attribute Description (2)

Attribute	Description	Units
Total sulfur dioxide	Free sulfites + bound sulfites.	g/dm^3
Density	Wine density.	g/cm^3
pH	pH is used to measure ripeness in relation to acidity.	0 to 14
Sulphates	Potassium sulphate.	g/dm^3
Alcohol	Volume of alcohol.	%
Quality	Wine quality score.	0 to 10

Agenda

Today:

Exploratory analysis of the Wine Quality dataset

Tomorrow afternoon:

Data mining: wine quality classification

Support Material

Jupyter Notebooks:

Today: Exploratory data analysis

Tomorrow: Wine Quality classification

Available in the shared folder @ https://bit.ly/34r6YUs

References

- Dataset source (Link)
- Modelling wine preferences by data mining from physicochemical properties (Link)
- Predicting quality of wine based on chemical attributes (Link)
- Predicting wine quality using data analytics (Link)

References

- Data analysis on the wine dataset (Link)
- Wine Quality Classification (Link)
- Vinho Verde webpage (Link)

Thank you for your attention!

Enjoy!