Wine enthusiast analysis

Hruday Kinhikar

Which vineyard produces the best wine?

Best wine

Wine that is given the most points whilst simultaneously being as cheap as possible

Point to consider

One vineyard that produces the best wine vs a vineyard that consistently produces multiple highly rated wines

Vineyard that produces best wine

Unoaked at the Pam’s Cuties winery

Vineyard that consistently produces highly rated wines in the top 10 wines

Organic Grapes at the Earth’s Harvest winery

Methodology

- reorder wines based on points/price
- weight the number of times a vineyard is featured on the top 10 list with its position on the list

Three wines to recommend

Title	Pam's Cuties NV Unoaked Chardonnay (California)	ManCan NV Fizz Sparkling (California)	Earth's Harvest 2014 Organic Grapes Chardonnay (California)
Variety	Chardonnay	Sparkling Blend	Chardonnay
Description	Sweet and fruity, this canned wine feels soft and syrupy, with sugary pear as the primary flavor on the palate. It's a basic white wine in a convenient package.	This sparkling wine is the best of ManCan's three new canned offerings. It has plenty of rich fruit flavors and a buttery note. Lively bubbles settle into the smooth, creamy texture, leaving an overall impression of easygoing enjoyment.	This wine has a deep-gold color, attractive baking-spice and buttered popcorn aromas, appealing apple and pear flavors and medium body. The texture is soft and easy, and the acidity tastes low.
Points/Price	20.75	17.4	17
Taster	Jim Gordon	Jim Gordon	Jim Gordon
Wine id	59507	104412	37951
Winery	Pam's Cuties	ManCan	Earth’s Harvest

Factors important in determining the overall score of a wine

Steps to approach this problem

Step 1:

- Use correlation matrix to understand correlation between different features

Step 2:

- explore feature importance through Recursive Feature Elimination

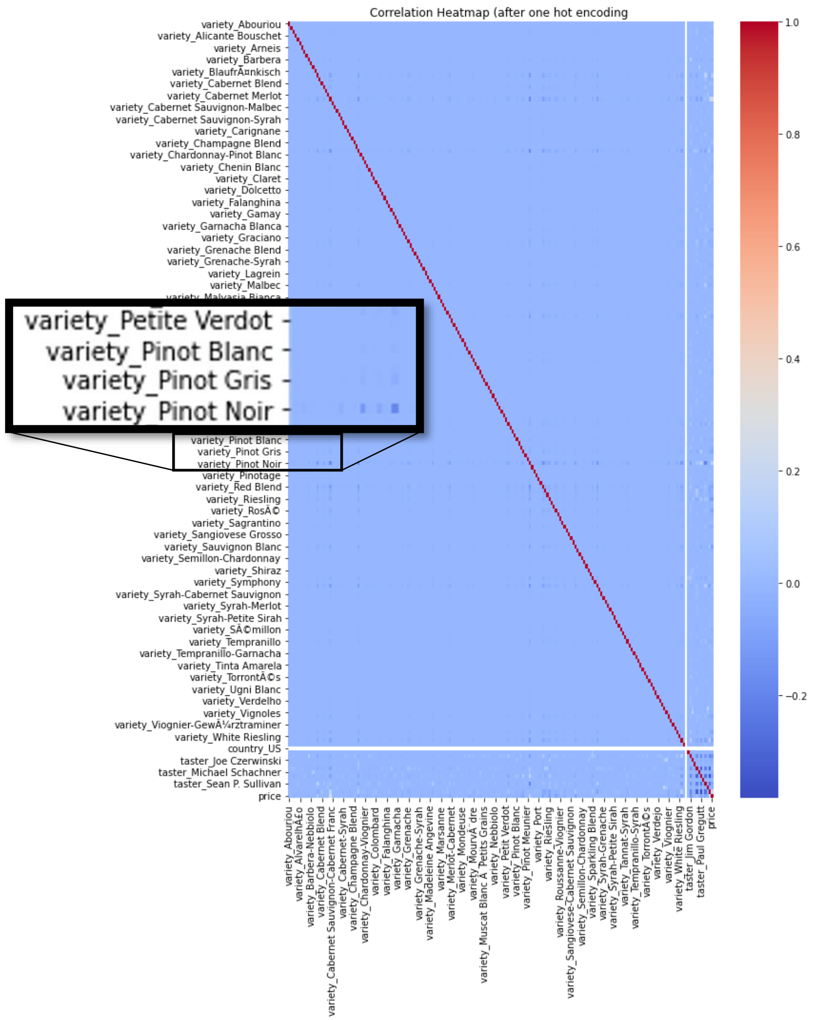
Correlation Matrix

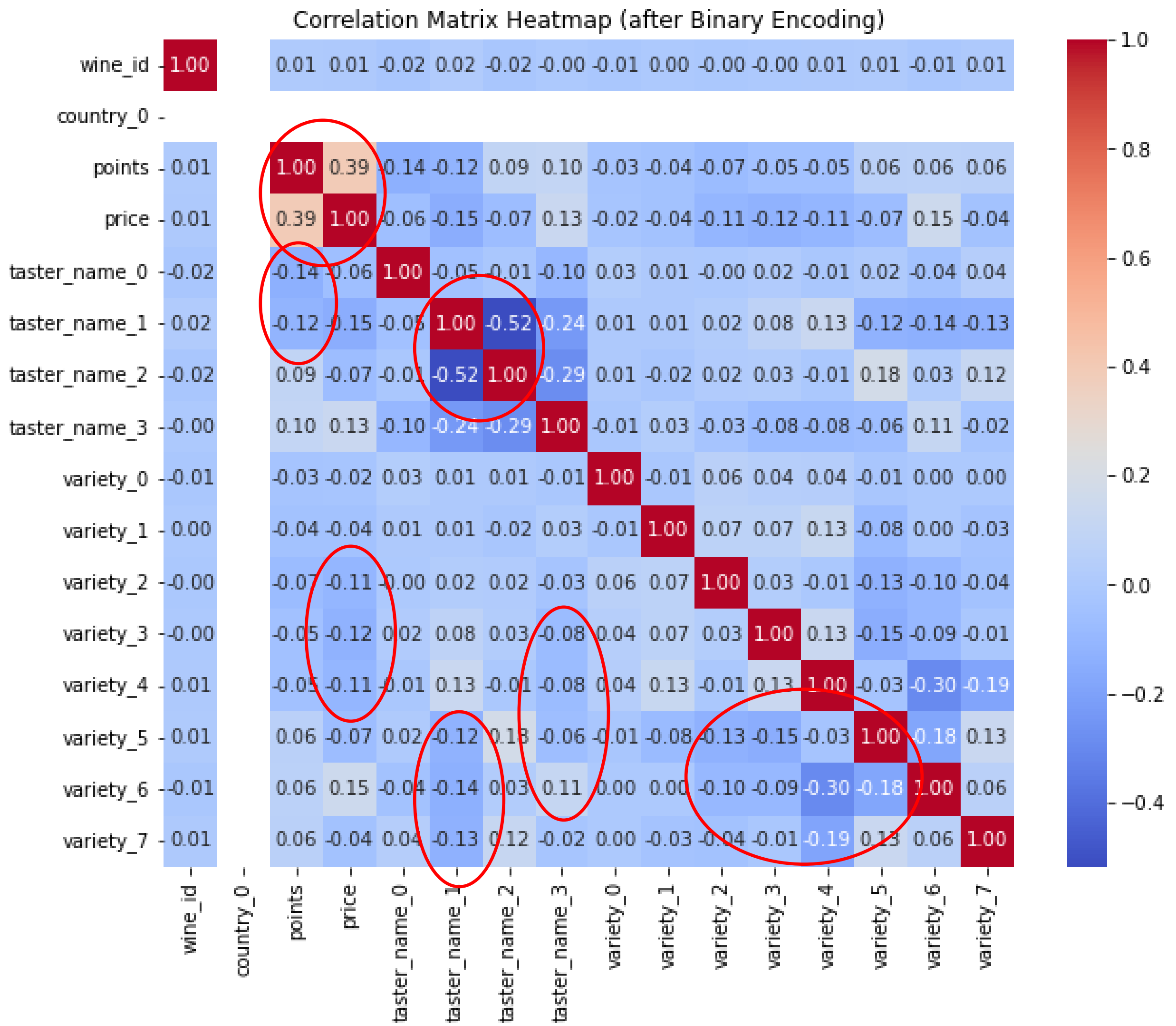
Displays the correlation coefficient between features

Implementation

One-hot encoding to convert categorical data into numerical data

Need to explore other options due to too many dimensions





Correlation Matrix Heatmap (after Binary Encoding)

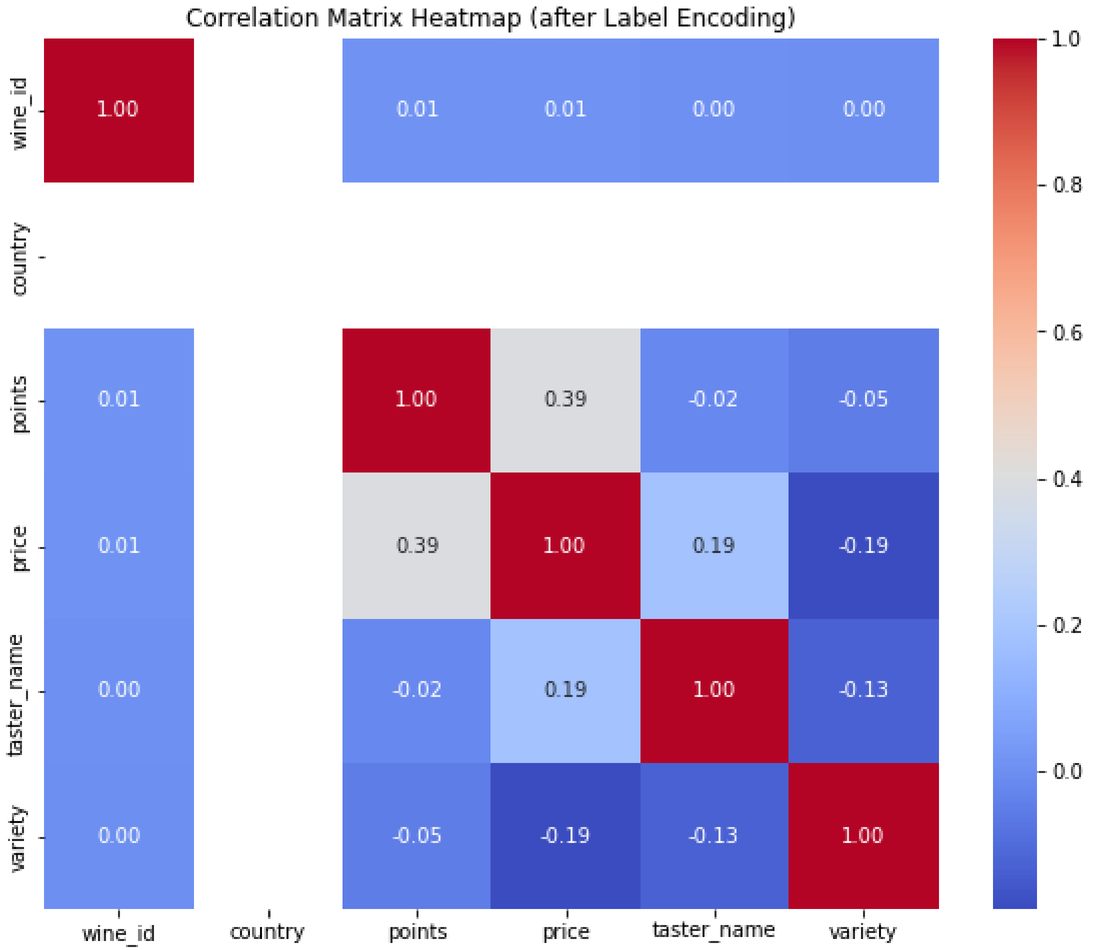
	wine_id	country_0	points	price	taster_name_0	taster_name_1	taster_name_2	taster_name_3	variety_0	variety_1	variety_2	variety_3	variety_4	variety_5	variety_6	variety_7
wine_id	1.00		0.01	0.01	-0.02	0.02	-0.02	-0.00	-0.01	0.00	-0.00	-0.00	0.01	0.01	-0.01	0.01
country_0																
points	0.01		1.00	0.39	-0.14	-0.12	0.09	0.10	-0.03	-0.04	-0.07	-0.05	-0.05	0.06	0.06	0.06
price	0.01		0.39	1.00	-0.06	-0.15	-0.07	0.13	-0.02	-0.04	-0.11	-0.12	-0.11	-0.07	0.15	-0.04
taster_name_0	-0.02		-0.14	-0.06	1.00	-0.05	-0.01	-0.10	0.03	0.01	-0.00	0.02	-0.01	0.02	-0.04	0.04
taster_name_1	0.02		-0.12	-0.15	-0.05	1.00	-0.52	-0.24	0.01	0.01	0.02	0.08	0.13	-0.12	-0.14	-0.13
taster_name_2	-0.02		0.09	-0.07	-0.01	-0.52	1.00	-0.29	0.01	-0.02	0.02	0.03	-0.01	0.18	0.03	0.12
taster_name_3	-0.00		0.10	0.13	-0.10	-0.24	-0.29	1.00	-0.01	0.03	-0.03	-0.08	-0.08	-0.06	0.11	-0.02
variety_0	-0.01		-0.03	-0.02	0.03	0.01	0.01	-0.01	1.00	-0.01	0.06	0.04	0.04	-0.01	0.00	0.00
variety_1	0.00		-0.04	-0.04	0.01	0.01	-0.02	0.03	-0.01	1.00	0.07	0.07	0.13	-0.08	0.00	-0.03
variety_2	-0.00		-0.07	-0.11	-0.00	0.02	0.02	-0.03	0.06	0.07	1.00	0.03	-0.01	-0.13	-0.10	-0.04
variety_3	-0.00		-0.05	-0.12	0.02	0.08	0.03	-0.08	0.04	0.07	0.03	1.00	0.13	-0.15	-0.09	-0.01
variety_4	0.01		-0.05	-0.11	-0.01	0.13	-0.01	-0.08	0.04	0.13	-0.01	0.13	1.00	-0.03	-0.30	-0.19
variety_5	0.01		0.06	-0.07	0.02	-0.12	0.18	-0.06	-0.01	-0.08	-0.13	-0.15	-0.03	1.00	-0.18	0.13
variety_6	-0.01		0.06	0.15	-0.04	-0.14	0.03	0.11	0.00	0.00	-0.10	-0.09	-0.30	-0.18	1.00	0.06
variety_7	0.01		0.06	-0.04	0.04	-0.13	0.12	-0.02	0.00	-0.03	-0.04	-0.01	-0.19	0.13	0.06	1.00

Binary encoding

- Rank number k – 1 in binary
- binary column does not correspond to one specific category but rather a combination of categories
- each category is represented by a combination of bits across multiple columns – difficulties mapping

Insights

- No correlation for country
- Some tasters have preference for some varieties over others
- Some varieties correlate with price
- Correlation between price and points
- Correlation within varieties as well as within tasters



Correlation Matrix Heatmap (after Label Encoding)

Label Encoding

- replaces the categorical value with a numeric value between 0 and the number of classes minus 1
- Reduces dimensionality further
- generation of priority issues during model training

Correlation between:

- price and points
- variety and price
- taster name and price
- taster name and variety

Step 2

Recursive Feature Elimination

- RFE is **wrapper** style feature selection algorithm
- Removes the weakest feature per loop until the most important feature is left

Methodology

- Pick a classifier and build a model that predicts points rated for the wine
- Measure accuracy for reference
- Implement RFE with a sample
- Increased sample size
- Extrapolate

Classifier

- Support Vector Machines(SVM) as the main classifier
- Works by implementing a hyperplane to separate data points
- Optimises hyperplane by maximizing distance between hyperplane and nearest datapoint

Validation metric

- MAE is the weighted sum of difference between predicted value and real value

Sample size = 20% of entire dataset
MAE = 2.0441889539432996
Feature hierarchy based on rankings:

Feature	Ranking
Taster_name	1
Price	2
Variety	3
Country	4

Sample size = 50% of entire dataset
MAE = 1.9586885164366323
Feature hierarchy based on rankings:

Feature	Ranking
Taster_name	1
Price	2
Variety	3
Country	4

Approximated order of importance for entire data set:

MAE of entire dataset:
1.9984124492243822

Feature	Ranking
Taster_name	1
Price	2
Variety	3
Country	4

Variety of wine for a “dry” and “citrus” flavour

Methodology

- New data frame with wine that has both ‘dry’ and ‘citrus’ in its description
- Ranking these wines according to points/price ratio
- Same concept as question 1 - weight the number of times a variety is featured on the top 10 list with its position on the list

This gives the variety of wine which fits flavour profile whilst consistently being featured among best wines

Following table displays how the best variety for mentioned condition changes based on how many of the top wines are considered

No. of top wines considered	Recommended variety
5	Pinot Grigio
10	Pinot Grigio
30	Riesling
50	Riesling