An exploration of Wine Reviews and the Big Data of wines

Giacomo Freccero, DSI General Assembly 2021

The Process:

- Some wine notions
- Goals
- Data Collection & EDA
- Findings
- Limits and potential improvements

WINE FOLLY

WINE 101

Grape Variety: Merlot

Grape Variety: Chardonnay

Wine

Wine

B: Balance (fruit vs acidity)
L: Length
I: Intensity
C: Complexity - 3 layers:
a) Aromas and Flavors of the grape and alcoh. fermentation
b) post-fermentations
c) maturation

Some wine notions

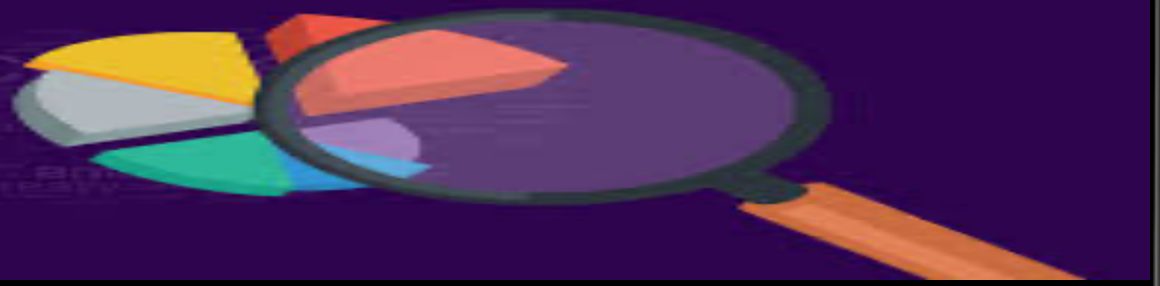
Goals

- Can we predict grape varieties from the text reviews & what words are the best predictors?
- How are the grape varieties connected in terms of common flavor characteristics? Can we create a map of flavors?
- Can we predict above or below median rating from the text reviews and what words are the best predictors?
- Can we predict ratings from text reviews?
- Can we predict value for money from the text reviews?
- Can we predict if the wine will be good,very good, superb or excellent from the text reviews?
- Can we predict Rating from the winery?
- Can we predict above/below median rating from the appellation?

Goals

- Create dictionaries of flavors that are non biased for each gv
- Get descriptors that are an indication of quality
- Understand what flavor characteristics gv share, mapping them and start a wine recommender. E.g. if you liked Californian Cab you will like Italian Aglianico.
- Which flavors are good value for money and which are expensive to buy
- Which flavors are associated with exceptional wines, which with bad wines?
- Which wineries consistently produce high rated wines (good investments)
- Which appellation is likely to predict above average wines
- Who would care? Wine Education/Wine Clubs/Wholesalers/Retail /Wine Investors/Wine consumers

Data Collection



[41,894 total results]

RELATED REVIEWS 21-40 of 41,894 | Filter by | Sort by

Ormanni 2017 Chianti Classico TUSCANY
Violet, red berry and dark spice aromas lift out of the glass. ...
SEE FULL REVIEW
91 Points
$19

Tenuta Santa Maria 2013 Riserva (Amarone della Valpolicella Classico) VENETO
This opens with aromas of violet, cooking spices and underbrush. The firm, ...
SEE FULL REVIEW
91 Points
$88

Contesa 2017 Montepulciano d'Abruzzo CENTRAL ITALY
Dense aromas of brandy-soaked red cherry and plum meet a firm blend ...
SEE FULL REVIEW
90 Points
$19

Violet, red berry and dark spice aromas lift out of the glass. On the savory palate, firm, fine-grained tannins accompany ripe black cherry, blood orange and licorice. Drink through 2022. —*KERIN O'KEEFE*

RATING	91 POINTS
PRICE	$19, BUY NOW
VARIETY	Red Blends, Red Blends
APPELLATION	Chianti Classico, Tuscany, Italy
WINERY	Ormanni
Print a Shelf Talker Label	
ALCOHOL	14%
BOTTLE SIZE	750 ml
CATEGORY	Red
IMPORTER	Virtuoso Selections
DATE PUBLISHED	3/1/2021
USER AVG RATING	Not rated yet [Add Your Review]

WINE ENTHUSIAST

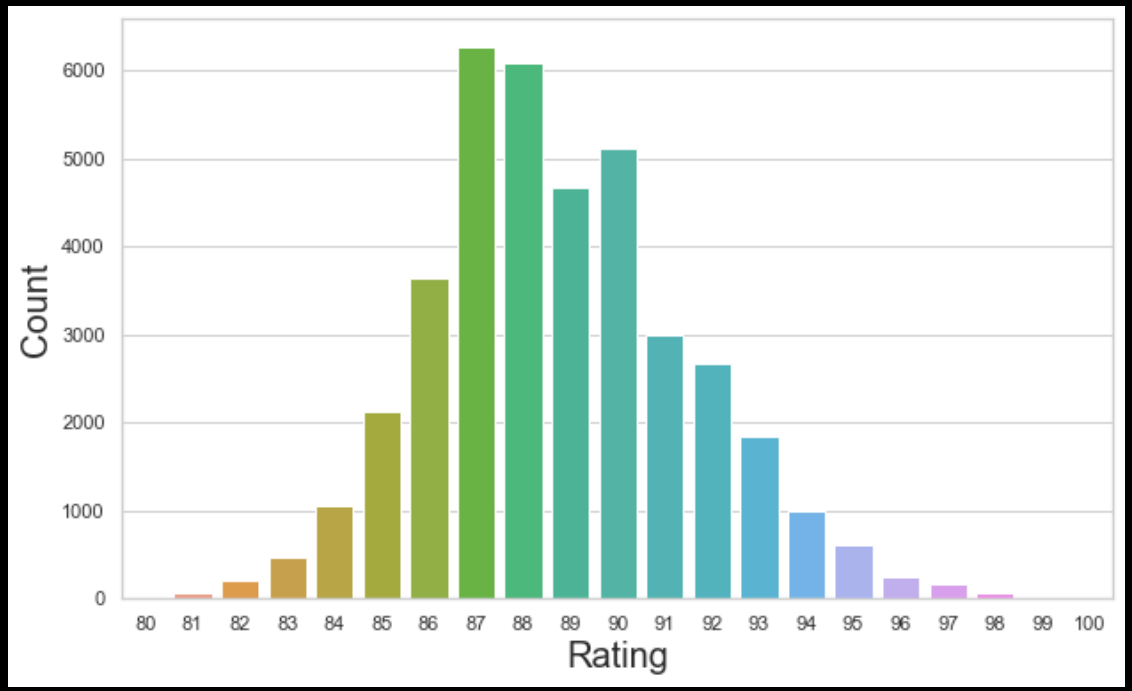
EDA

	name	review	taster	rating	price	designation	variety	appellation	winery	alc	...
0	Fattoria Giuseppe Savini 2017 Rondineto (Monte...	ripe cherry and wild berry aromas take on a bi...	ALEXANDER PEARTREE	86	12.0	Rondineto	Montepulciano, Italian Red	Montepulciano d'Abruzzo, Central Italy, Italy	Fattoria Giuseppe Savini	13.0	...
1	Feudo Antico 2017 Organic (Montepulciano d'Abr...	spiced raspberry and cherry meld with a bit of...	ALEXANDER PEARTREE	86	10.0	Organic	Montepulciano, Italian Red	Montepulciano d'Abruzzo, Central Italy, Italy	Feudo Antico	13.0	...

Columns(['name', 'review', 'taster', 'rating', 'price', 'designation', 'variety', 'appellation', 'winery', 'alc', 'bottle_size', 'category', 'importer', 'review_date', 'variety1', 'variety2', 'appellation1', 'region', 'nation', 'vintage', 'value_for_money', 'final_judgement', 'final_score', 'above_below_median_rating', 'review_test'],

'leather, tilled earth, grated clove and black cherry aromas take center stage along with a whiff of fragrant blue flower. the taut, velvety palate doles out raspberry extract, ripe cranberry, tobacco and licorice alongside firm acidity and close-grained tannins that leave a drying finish. drink

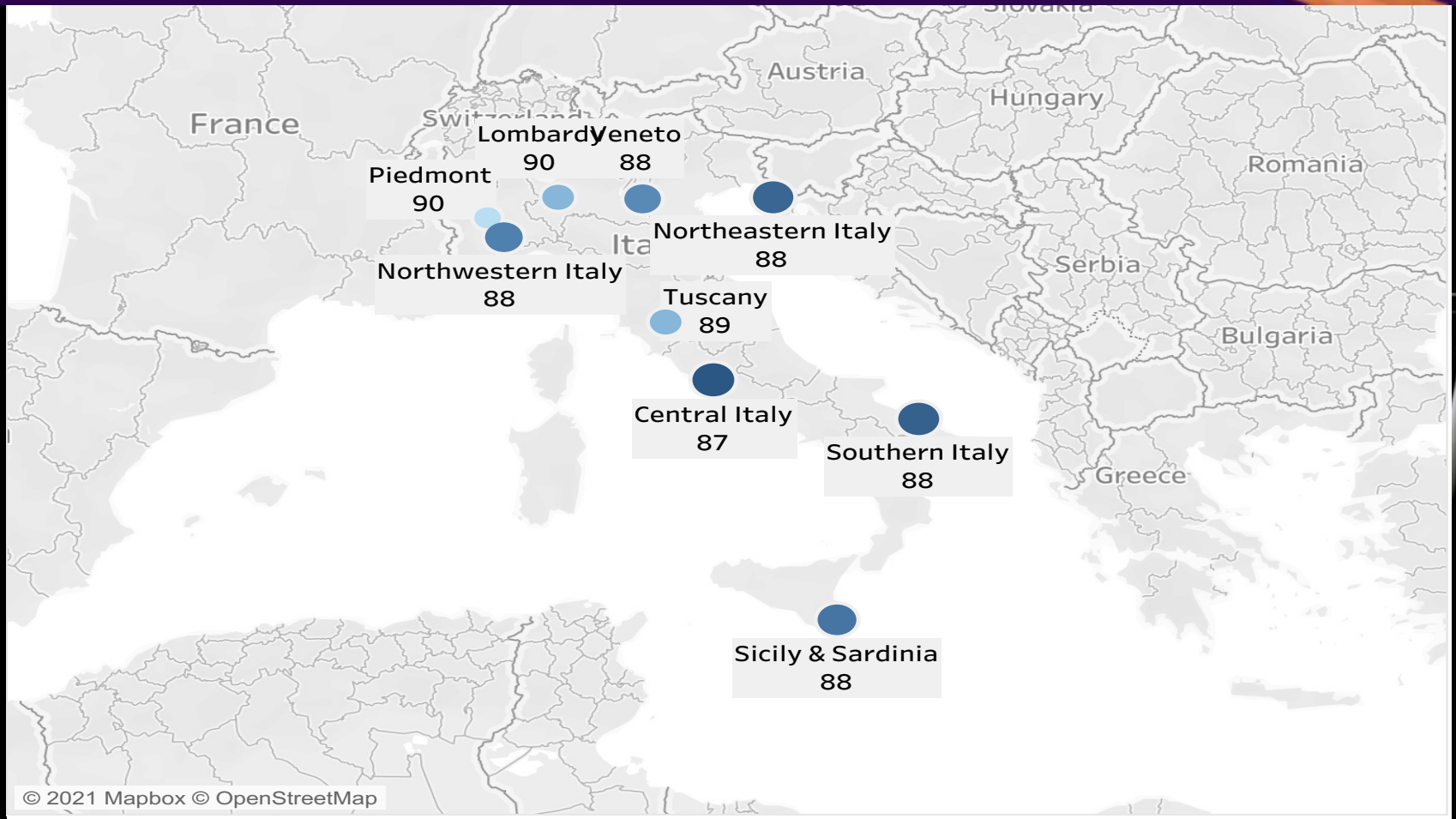
EDA



- **Wine Enthusiast's** 100 - point **wine scoring scale**:

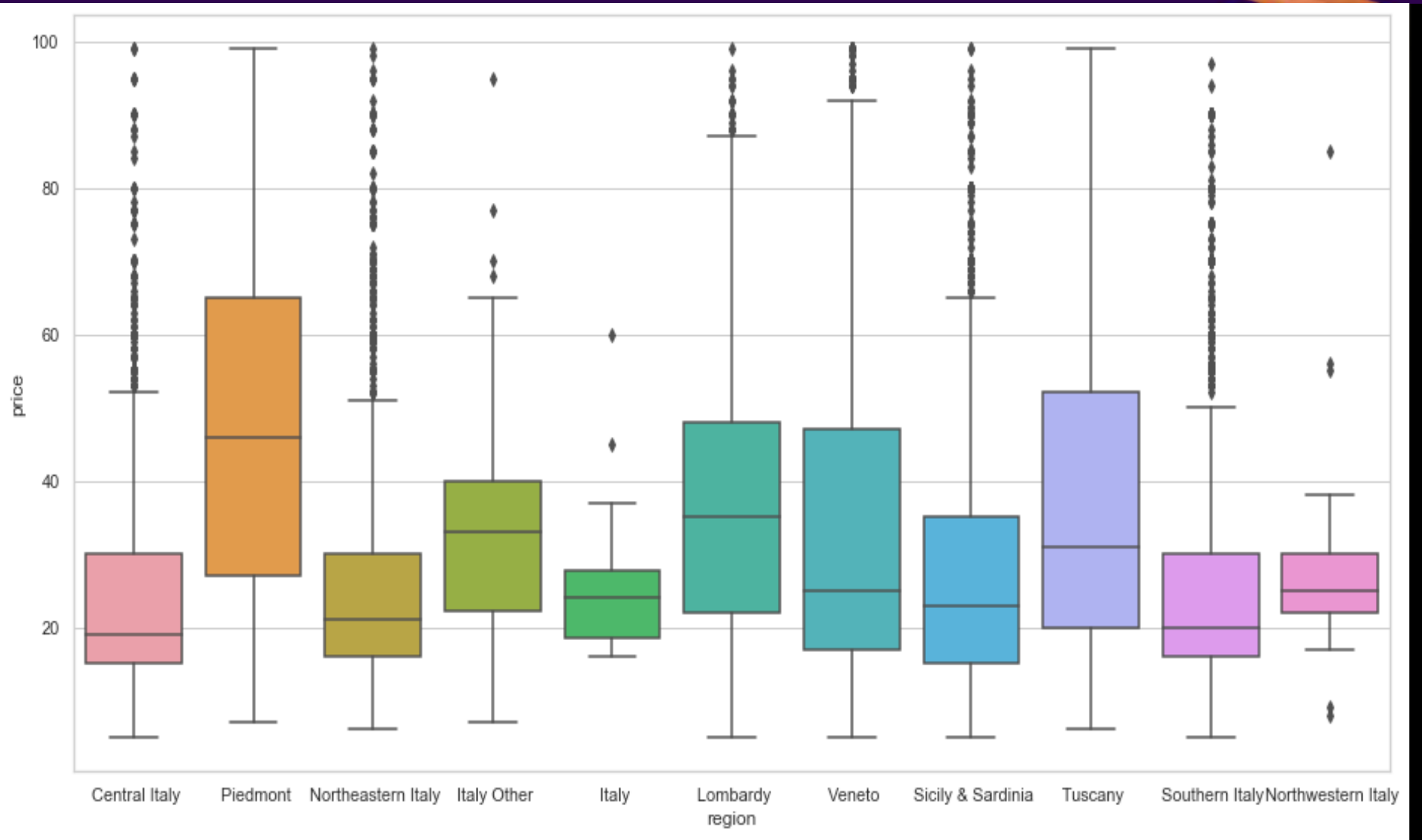
- 98–100 – Classic.
- 94–97 – Superb.
- 90–93 – Excellent.
- 87–89 – Very good.
- 83–86 – Good.
- 80–82 – Acceptable.

EDA

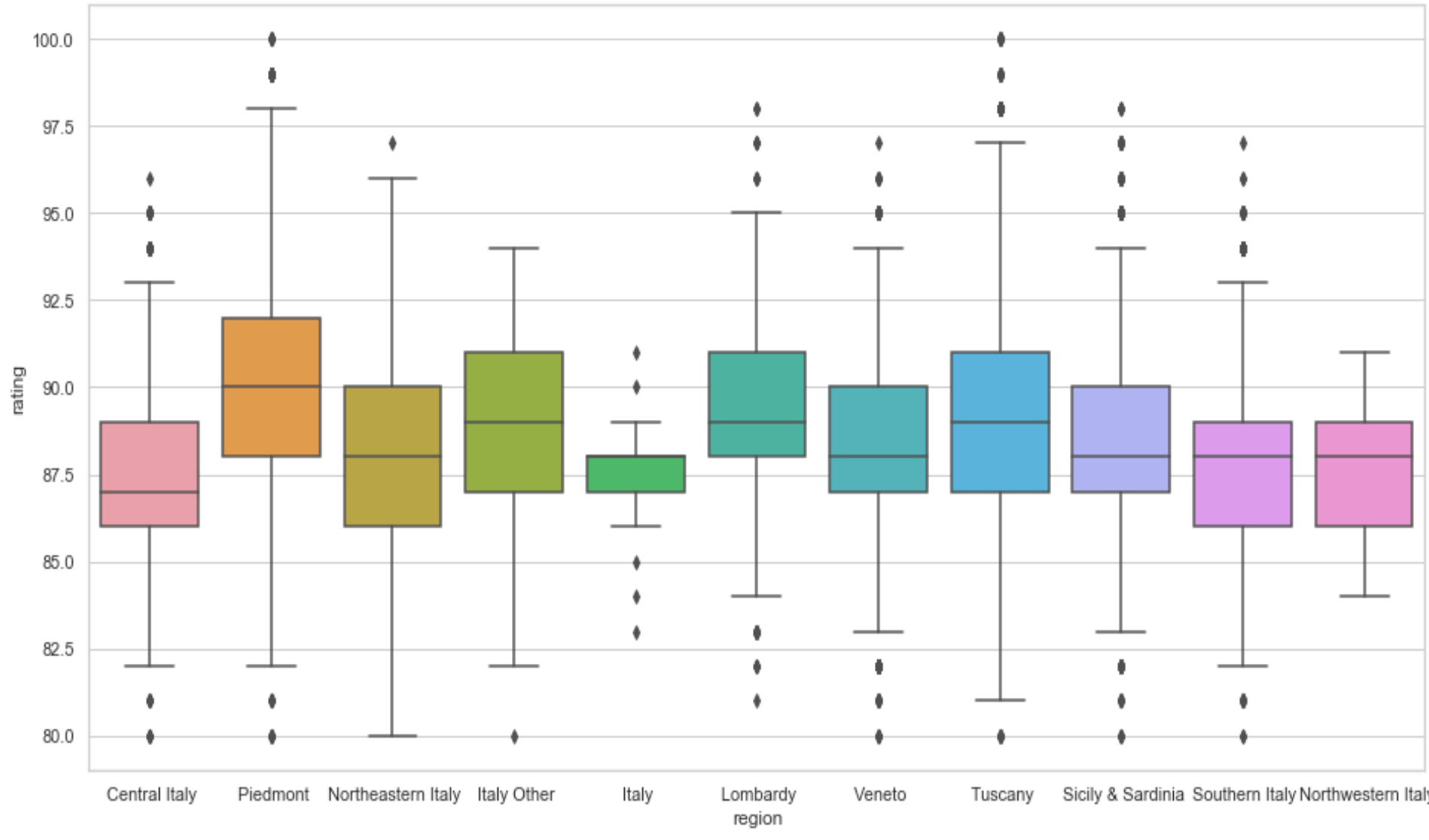


Map based on Longitude (generated) and Latitude (generated). Colour shows median of Value For Money. Size shows median of Value For Money. The marks are labelled by Region and median of Rating. Details are shown for Region. The data is filtered on Value For Money, which excludes Null.

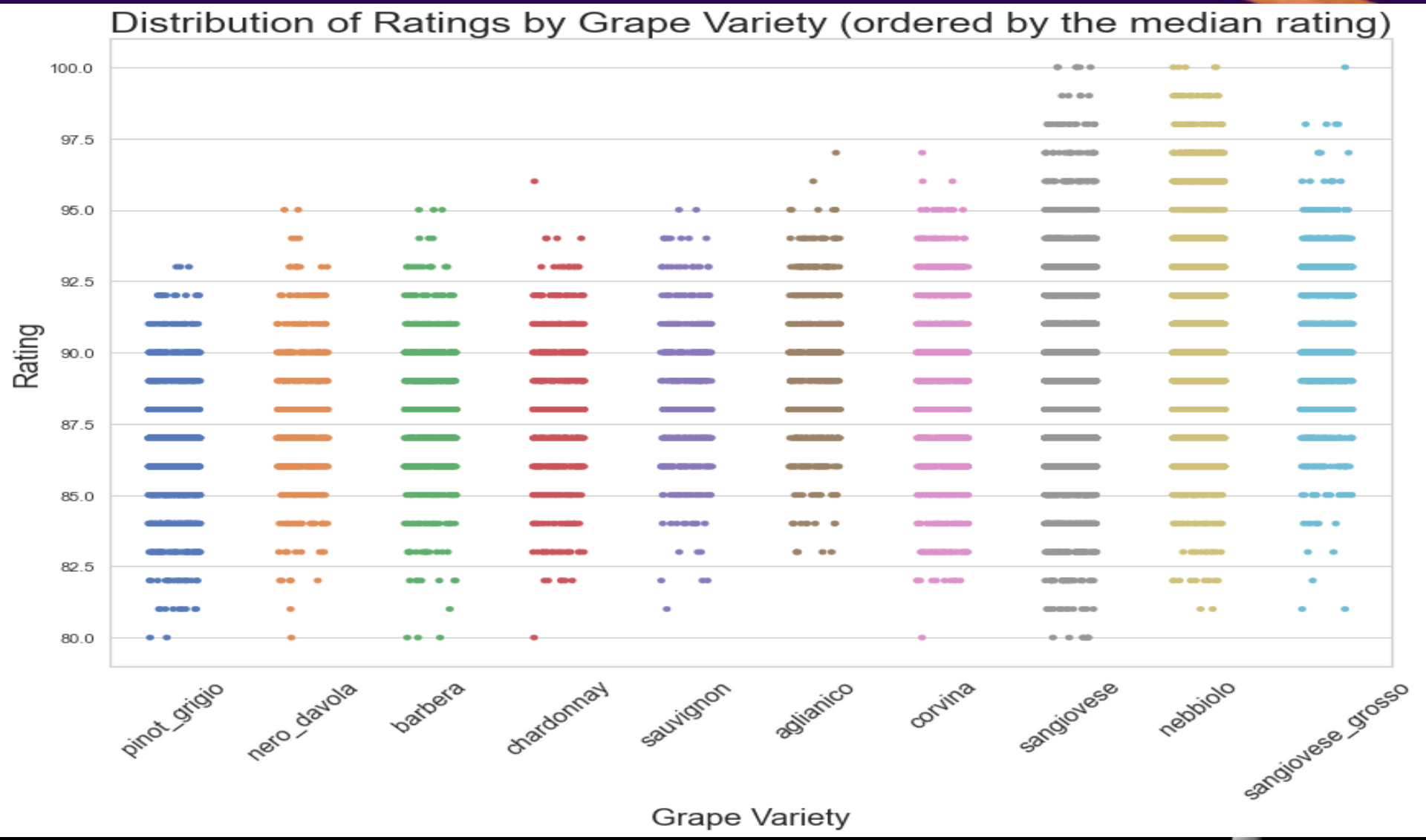
EDA


100
80
60
40
20
price
Central Italy
Piedmont
Northeastern Italy
Italy Other
Italy
Lombardy
Veneto
Sicily & Sardinia
Tuscany
Southern Italy
Northwestern Italy
region

EDA


rating
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
Central Italy
Piedmont
Northeastern Italy
Italy Other
Italy
Lombardy
Veneto
Tuscany
Sicily & Sardinia
Southern Italy
Northwestern Italy
region

EDA


Distribution of Ratings by Grape Variety (ordered by the median rating)
Rating
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
pinot_grigio
nero_davola
barbera
chardonnay
sauvignon
aglianico
corvina
sangiovese
nebbiolo
sangiovese_grosso
Grape Variety

Distribution of Price by Grape Variety (ordered by the median price)

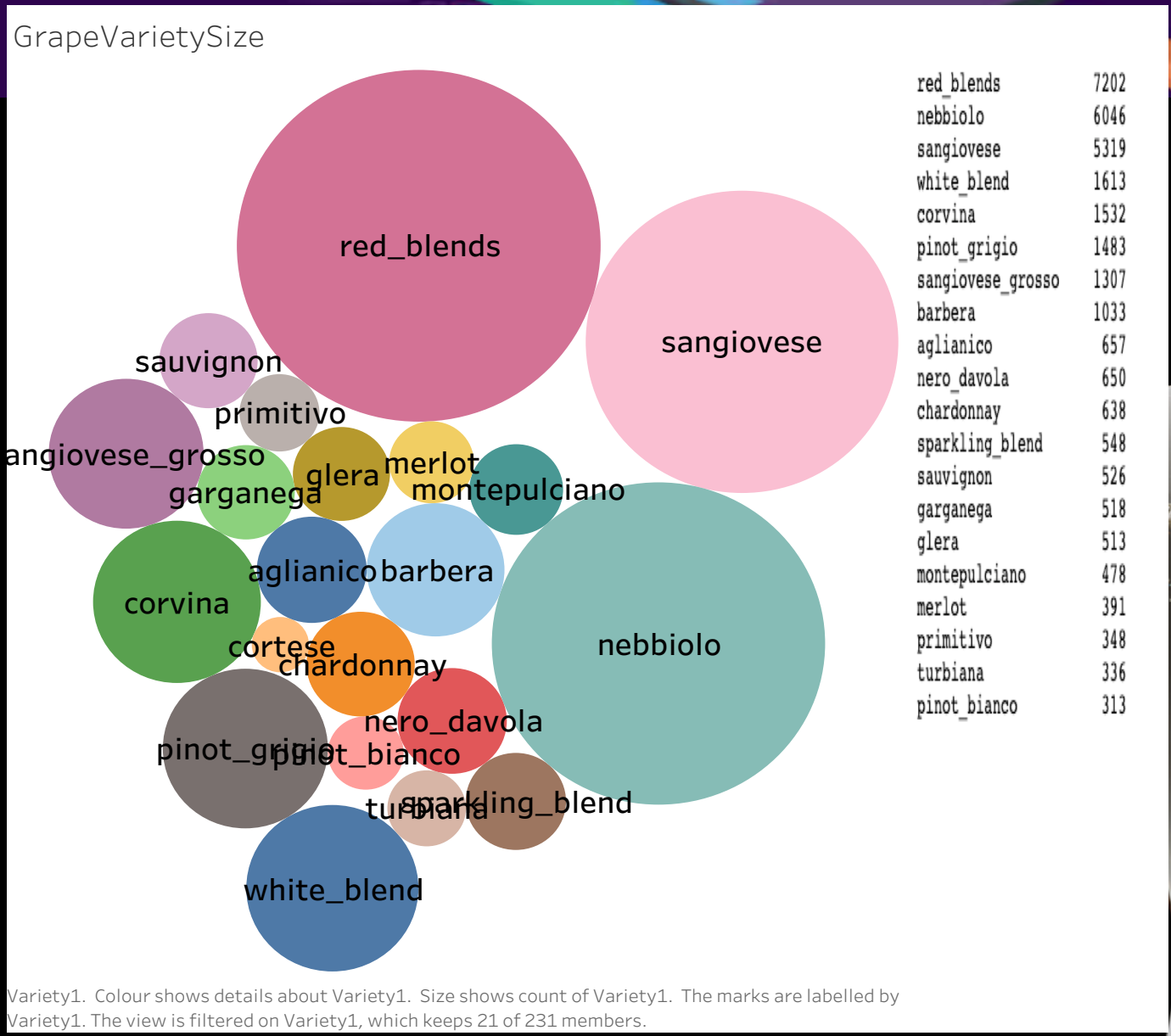
Price

300
250
200
150
100
50
0

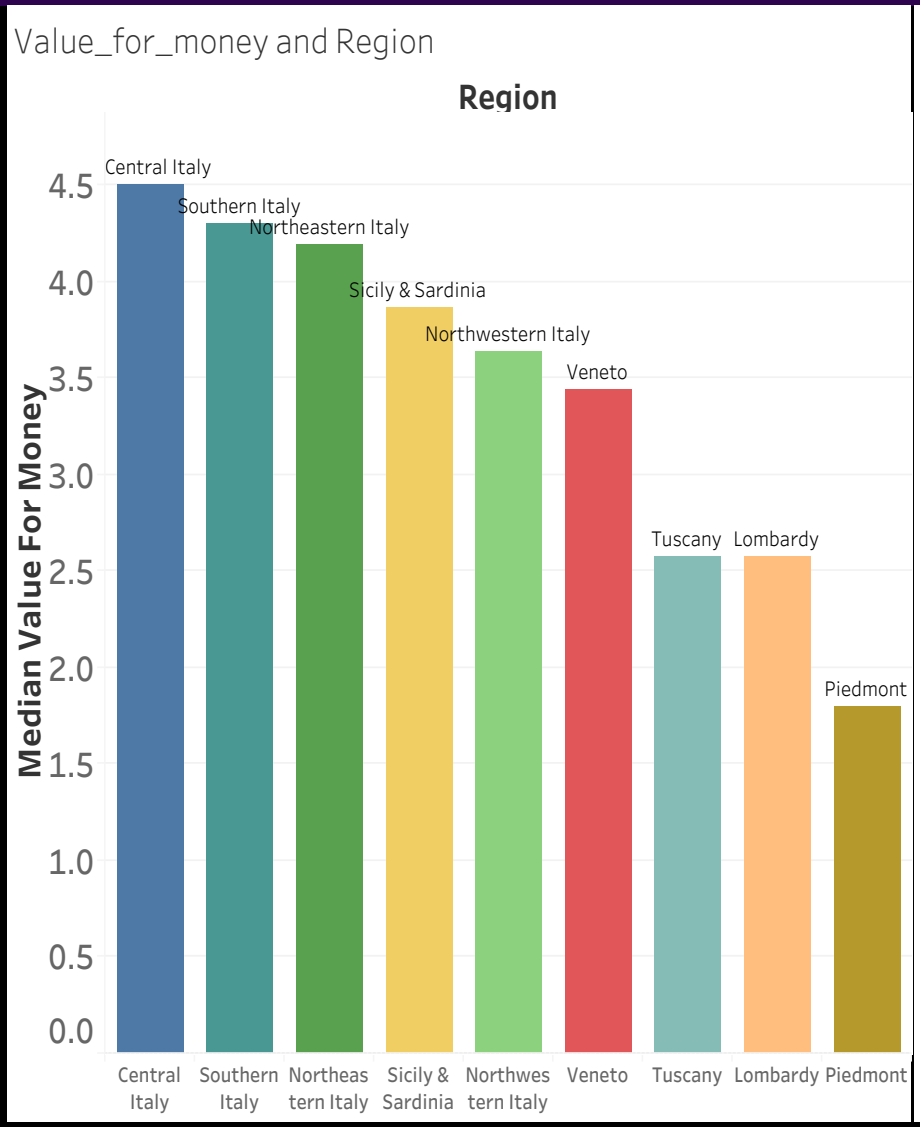
pinot_grigio
nero_davola
barbera
sauvignon
chardonnay
aglianico
sangiovese
corvina
nebbiolo
sangiovese_grosso

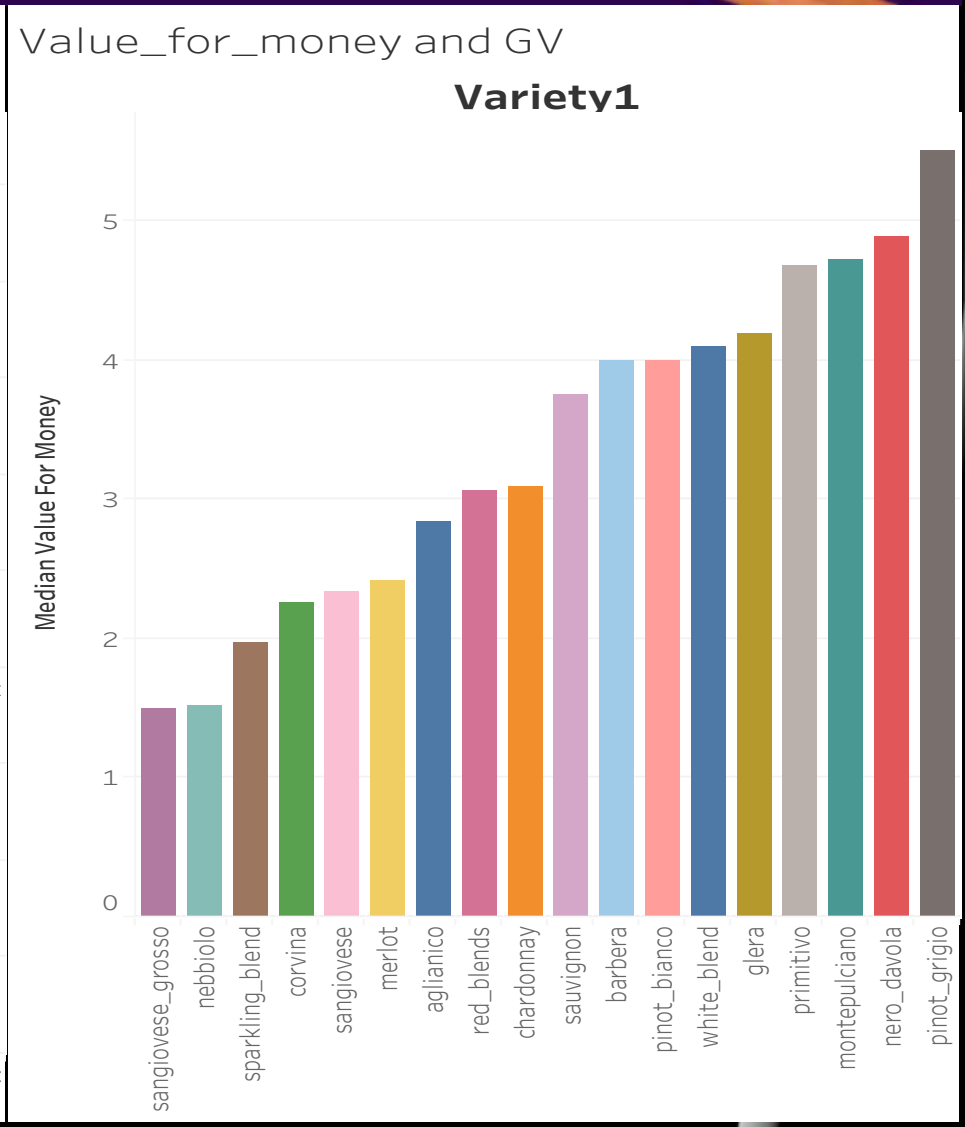
Grape Variety

EDA

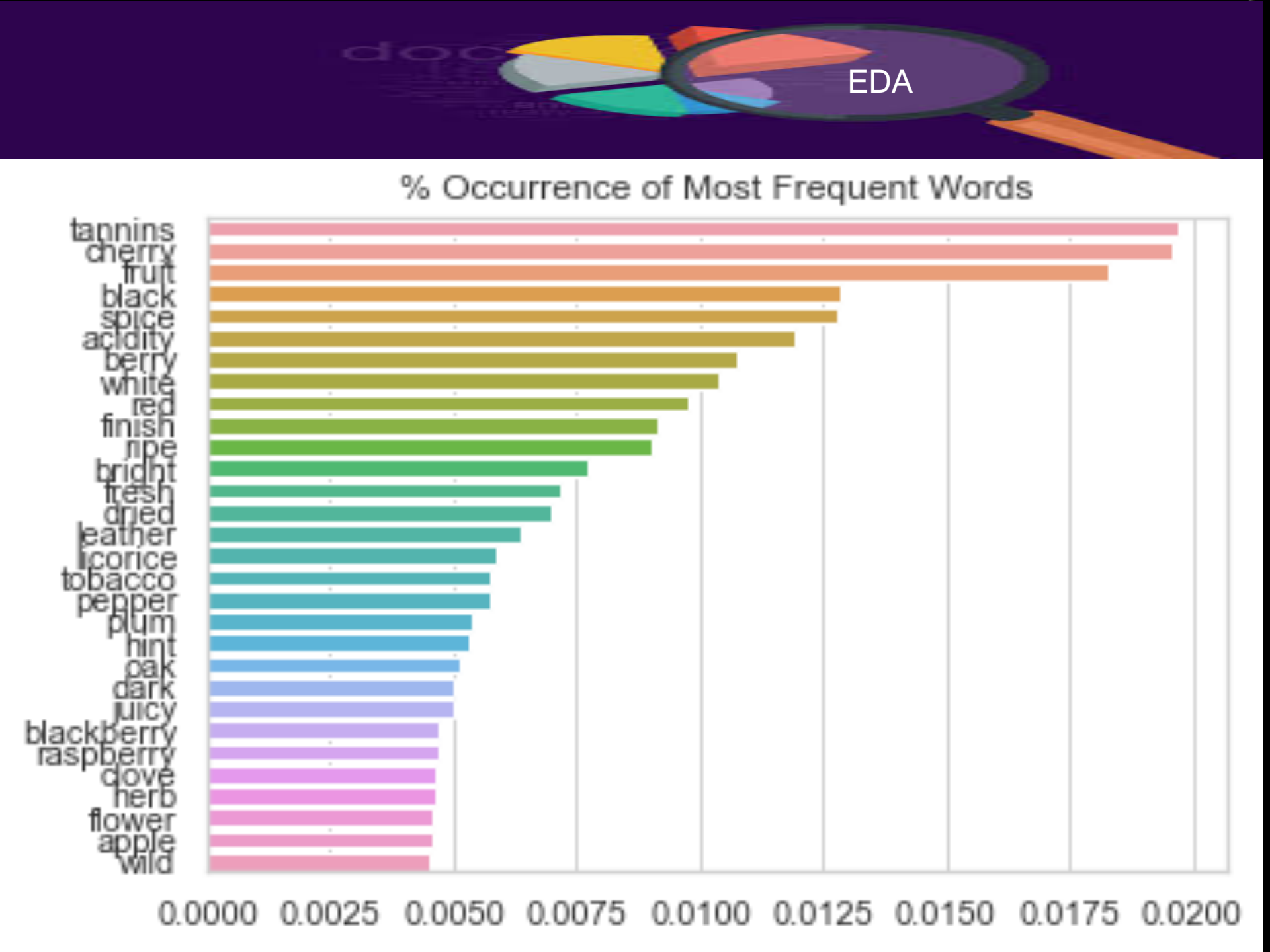

GrapeVarietySize
red_blends
sangiovese
sauvignon
primitivo
angiovese_grosso
glera
merlot
garganega
montepulciano
aglianico
barbera
corvina
nebbiolo
cortese
chardonnay
nero_davola
pinot_grigio
pinot_bianco
turbiana
sparkling_blend
white_blend
red_blends 7202
nebbiolo 6046
sangiovese 5319
white_blend 1613
corvina 1532
pinot_grigio 1483
sangiovese_grosso 1307
barbera 1033
aglianico 657
nero_davola 650
chardonnay 638
sparkling_blend 548
sauvignon 526
garganega 518
glera 513
montepulciano 478
merlot 391
primitivo 348
turbiana 336
pinot_bianco 313
Variety1. Colour shows details about Variety1. Size shows count of Variety1. The marks are labelled by Variety1. The view is filtered on Variety1, which keeps 21 of 231 members.

EDA

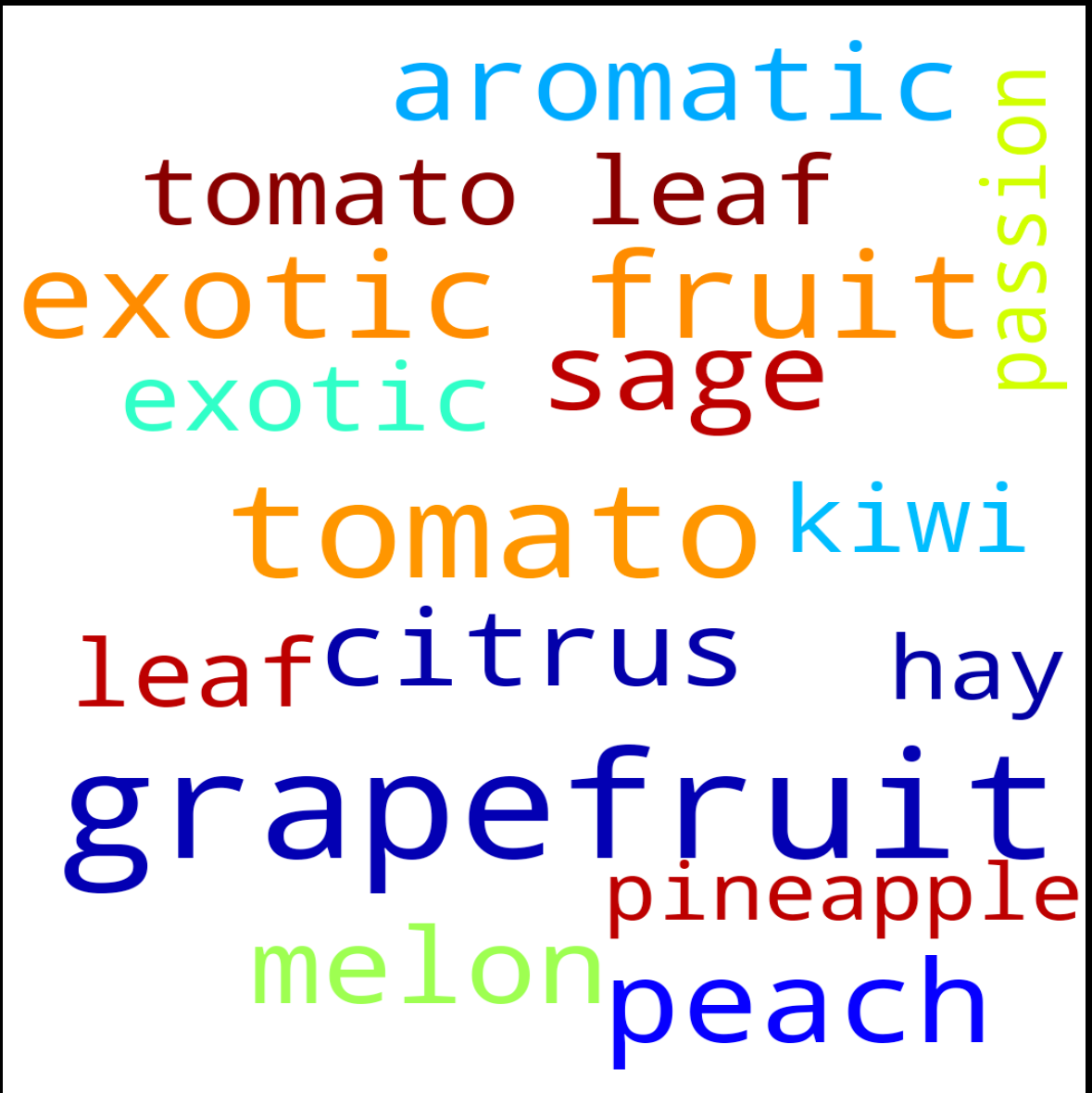

Value_for_money and Region
Region
Median Value For Money
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Central Italy
Southern Italy
Northeastern Italy
Sicily & Sardinia
Northwestern Italy
Veneto
Tuscany
Lombardy
Piedmont
Central Italy
Southern Italy
Northeastern Italy
Sicily & Sardinia
Northwestern Italy
Veneto
Tuscany
Lombardy
Piedmont


Value_for_money and GV
Variety1
Median Value For Money
5
4
3
2
1
0
sangiovese_grosso
nebbiolo
sparkling_blend
corvina
sangiovese
merlot
aglianico
red_blends
chardonnay
sauvignon
barbera
pinot_bianco
white_blend
glera
primitivo
montepulciano
nero_davola
pinot_grigio

EDA


% Occurrence of Most Frequent Words
tannins
cherry
fruit
black
spice
acidity
berry
white
red
finish
ripe
bright
fresh
dried
leather
licorice
tobacco
pepper
plum
hint
oak
dark
juicy
blackberry
raspberry
clove
herb
flower
apple
wild
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200

Predict Grape Variety from Reviews – SAUVIGNON BLANC



Predict Grape Variety from Reviews

What's inside the bottle?

Sangiovese

Nero d"Avola

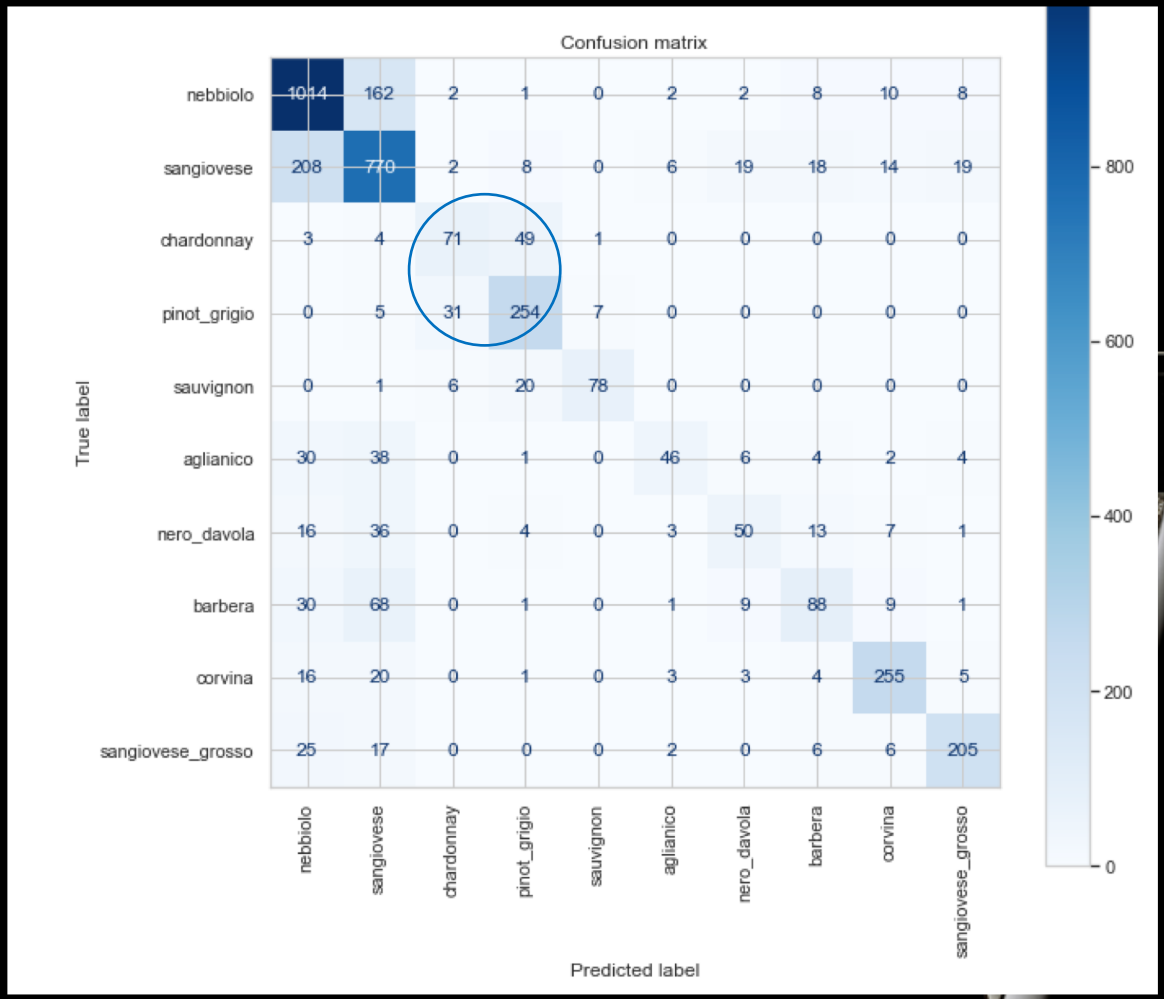
Nebbiolo

Sauvignon

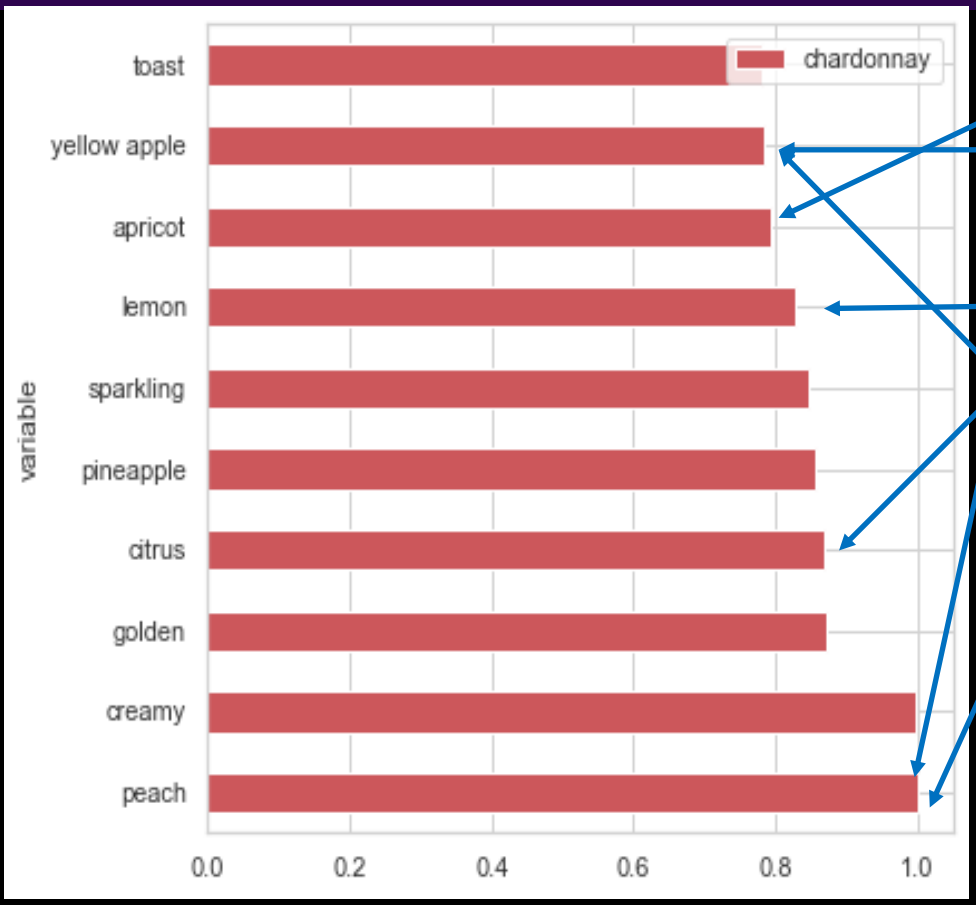
Predict Grape Variety from Reviews

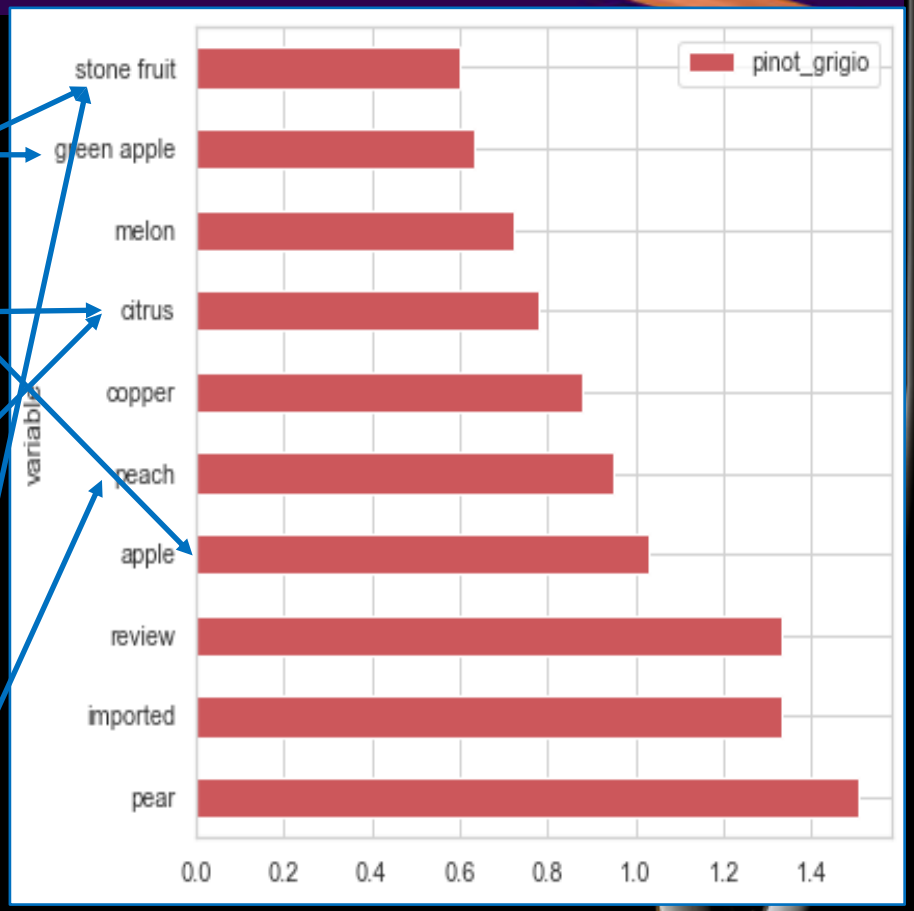
Cvec-Tvec Lr
CV 0.739
SVM 0.7

	precision	recall	f1-score
aglianico	0.73	0.35	0.47
barbera	0.62	0.43	0.51
chardonnay	0.63	0.55	0.59
corvina	0.84	0.83	0.84
nebbiolo	0.76	0.84	0.79
nero_davola	0.56	0.38	0.46
pinot_grigio	0.75	0.86	0.80
sangiovese	0.69	0.72	0.70
sangiovese_grosso	0.84	0.79	0.81
sauvignon	0.91	0.74	0.82
accuracy			0.74
macro avg	0.73	0.65	0.68
weighted avg	0.73	0.74	0.73

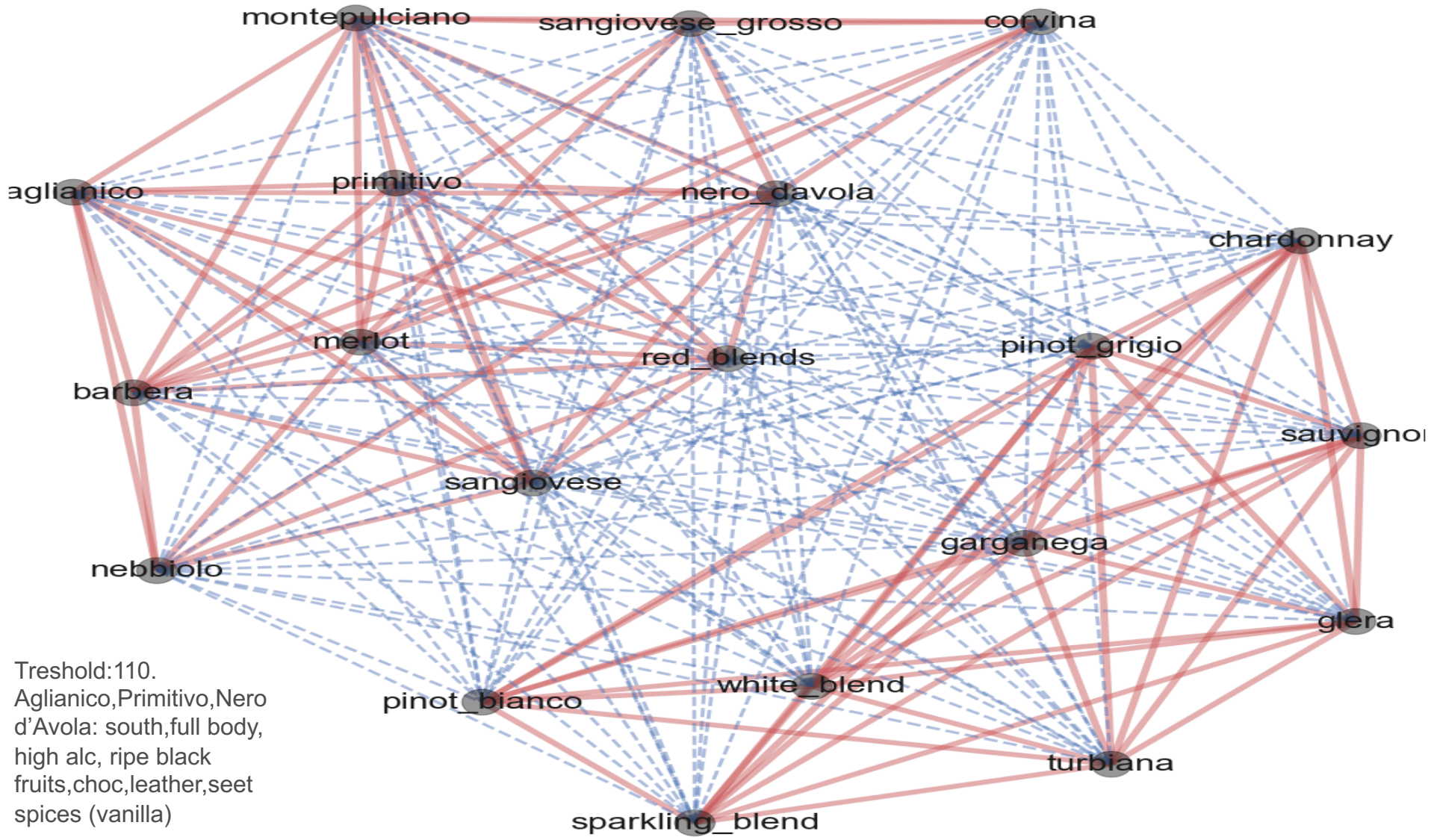


Predict Grape Variety from Reviews



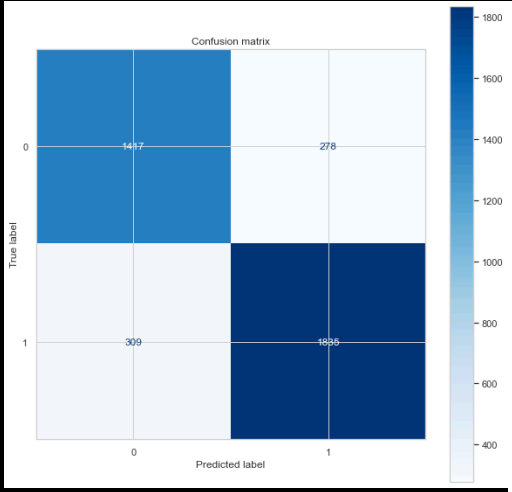


How are the grape varieties connected in terms of common flavor characteristics?

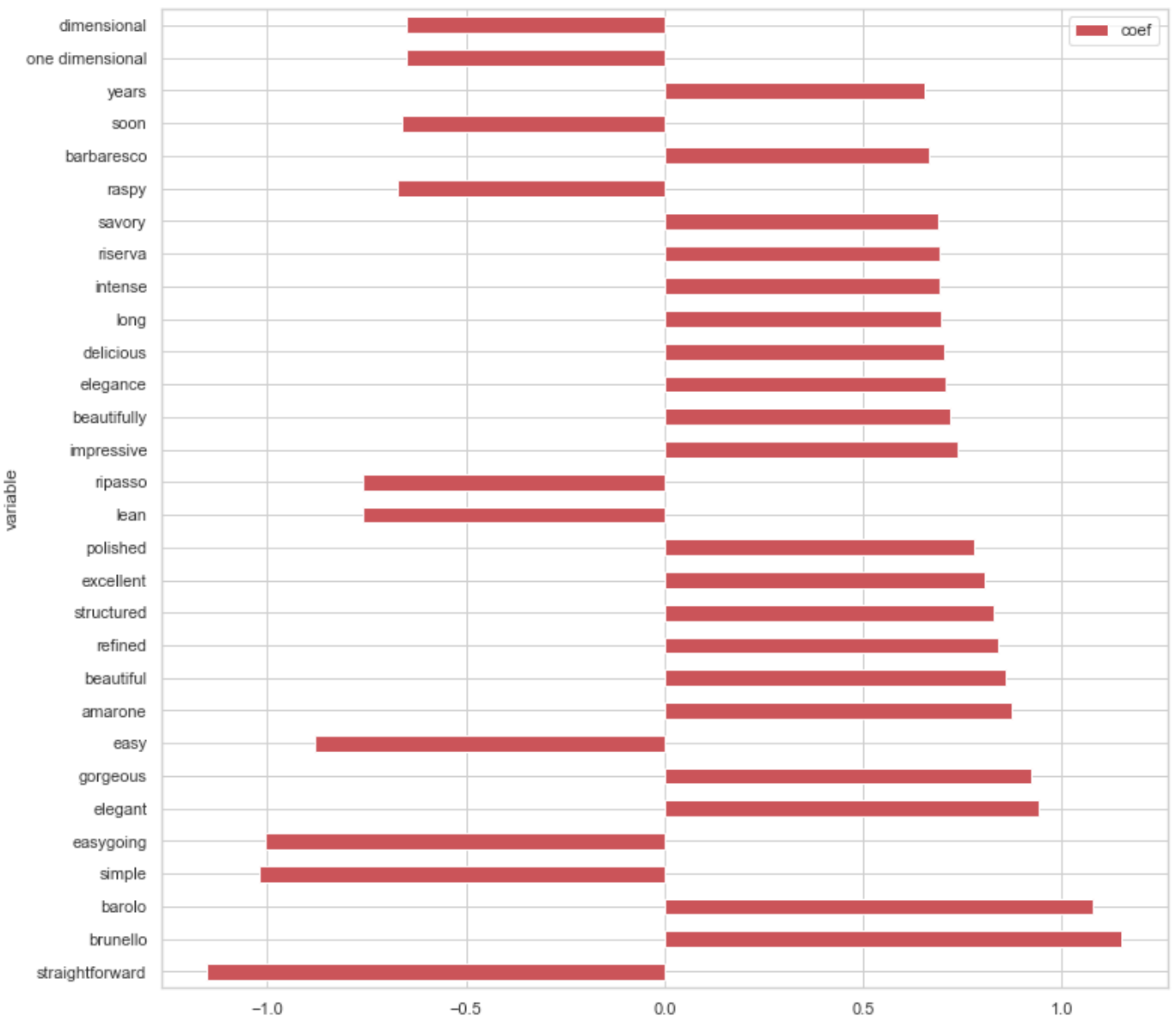


Treshold:110.
Aglianico,Primitivo,Nero d'Avola: south,full body, high alc, ripe black fruits,choc,leather,seet spices (vanilla)

Predict above or below median rating from the text reviews

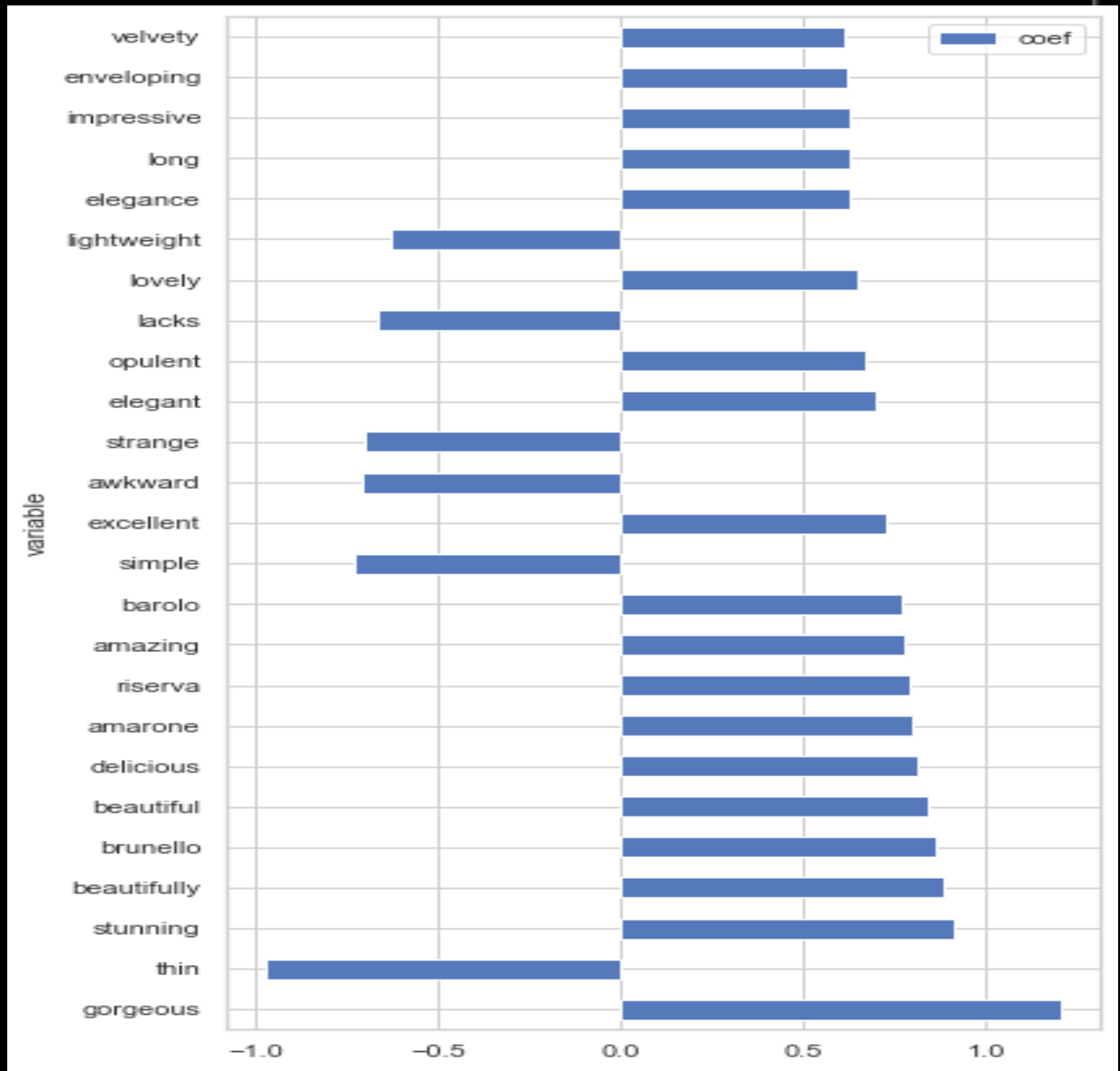


LR CV= 0.84 (train .99)



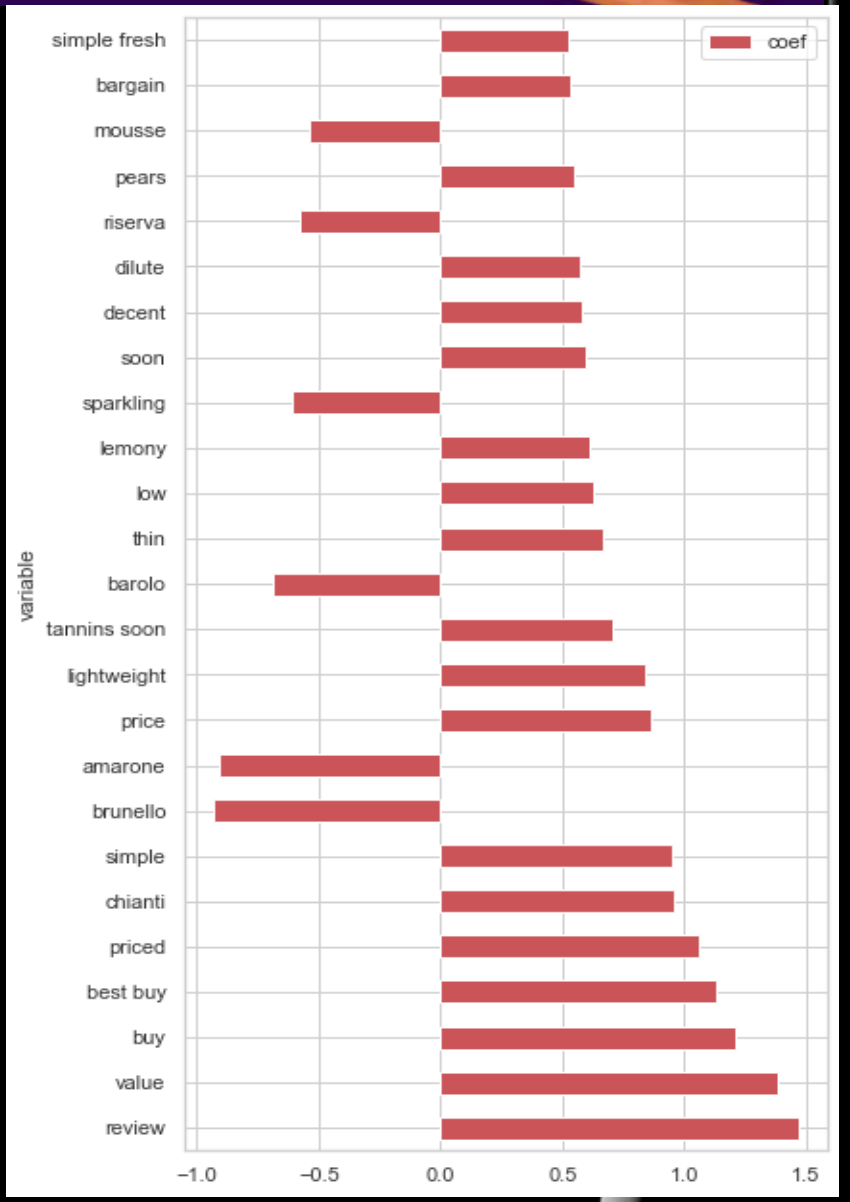
Predict ratings from the text reviews

Regression→ Ridge Cv =.75 (92 train score)
Random Forest-Dec Tree CV .25



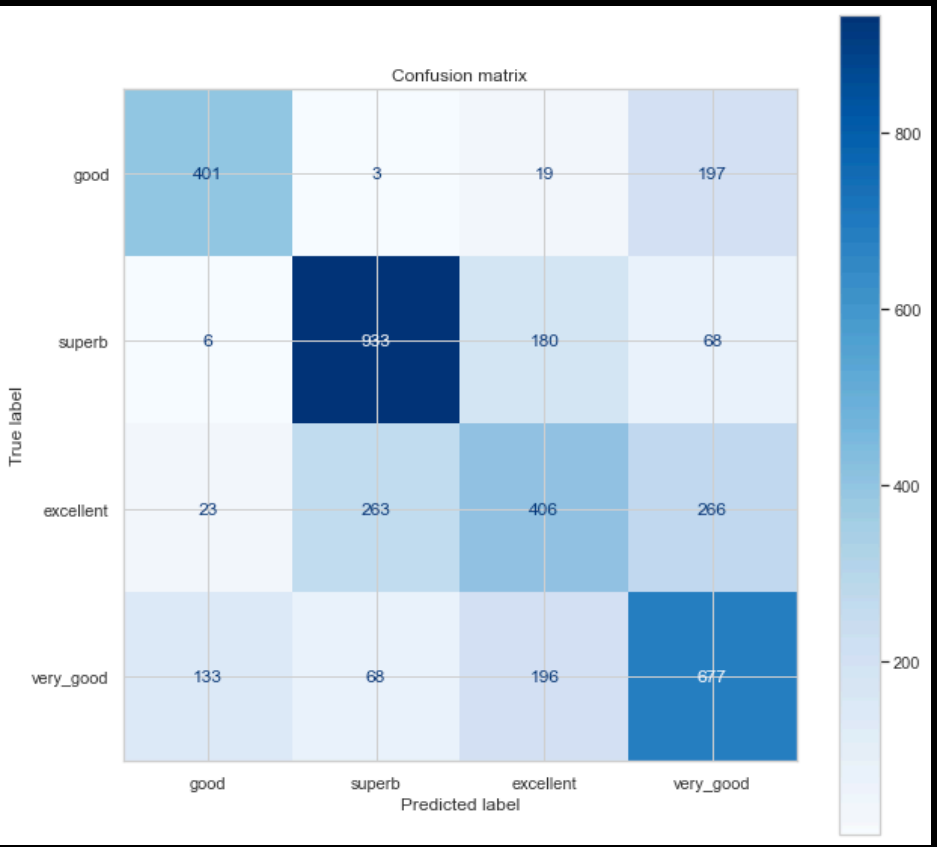
Predict Value for Money from the text reviews

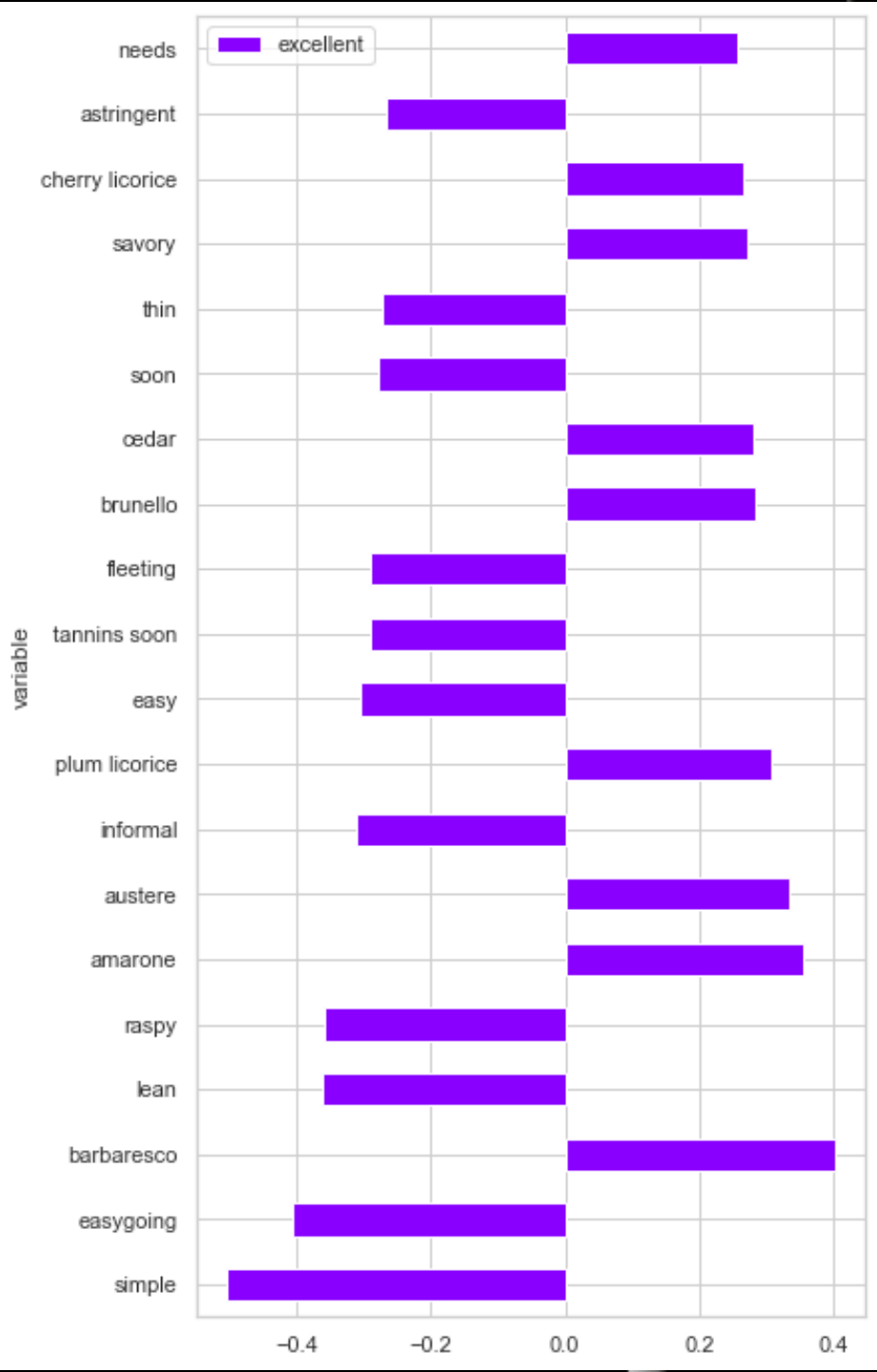
Regression→ Ridge Cv =.54
(86 train score)
Lasso/Net much lower



Predict Good,VeryGood,Excellent, Superb from the text reviews

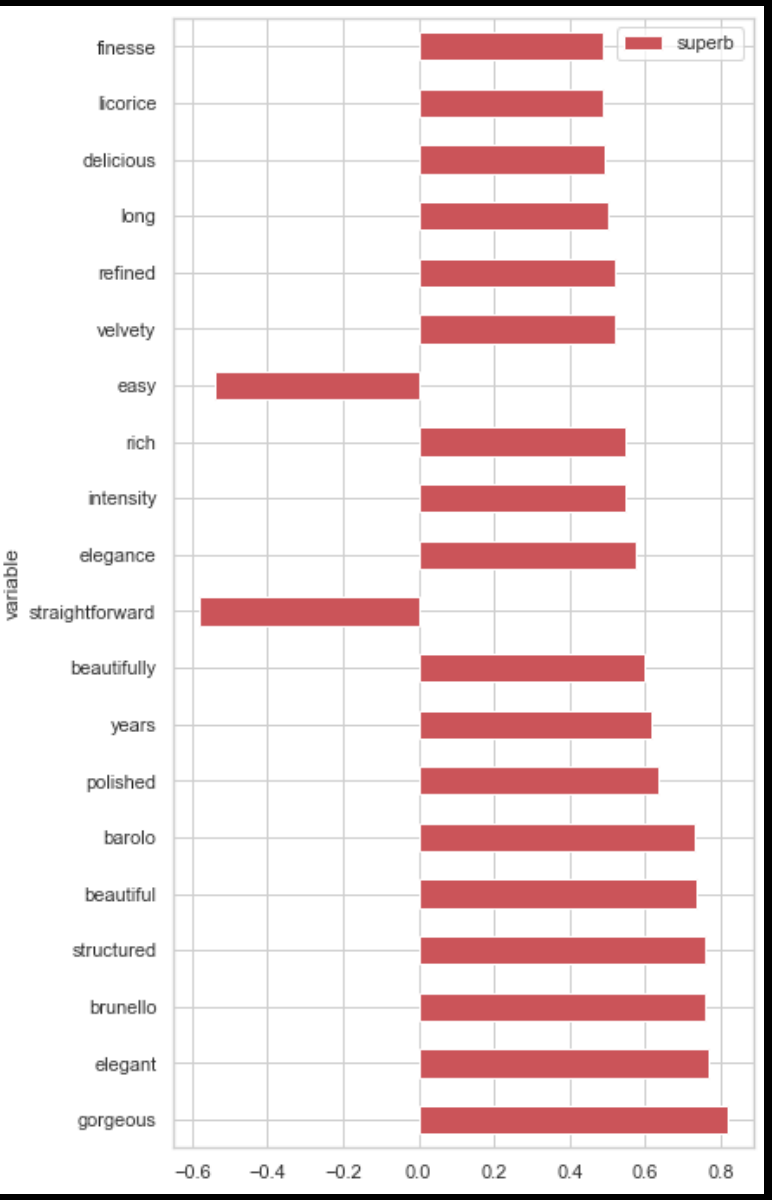
Classif.→ LR CV =.63 (89 train score)

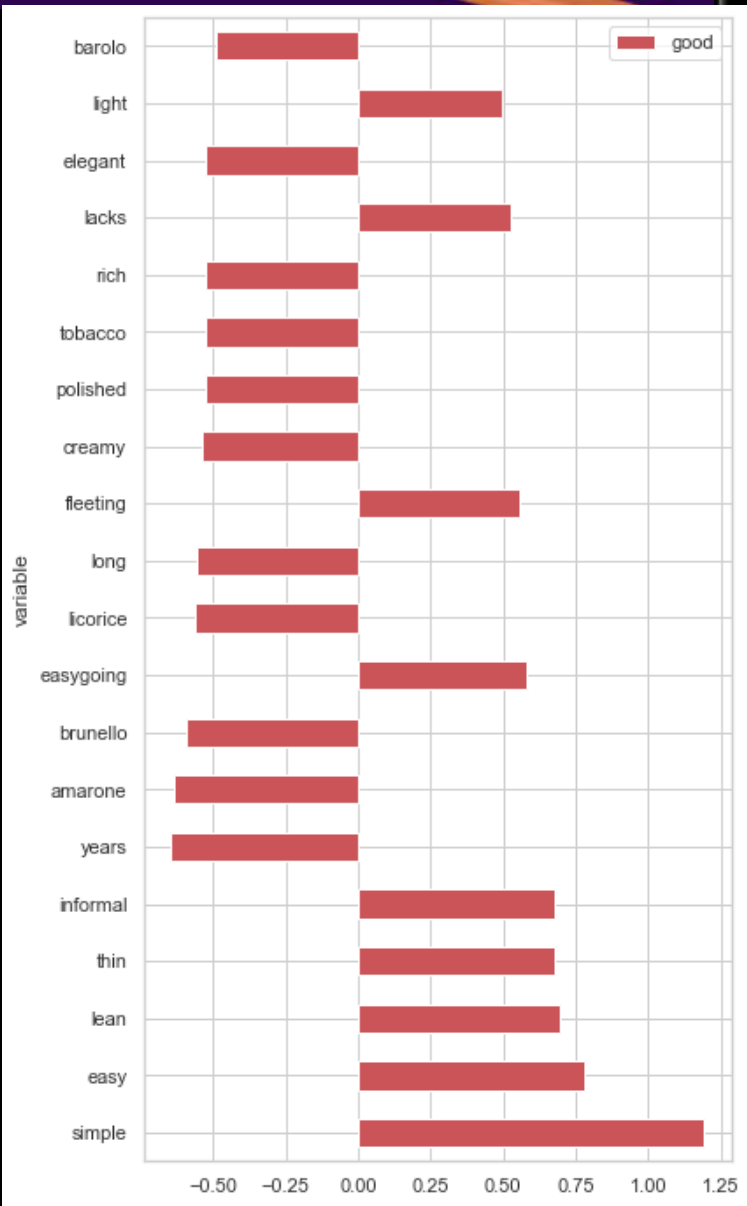




Predict Good,VeryGood,Excellent, Superb from the text reviews

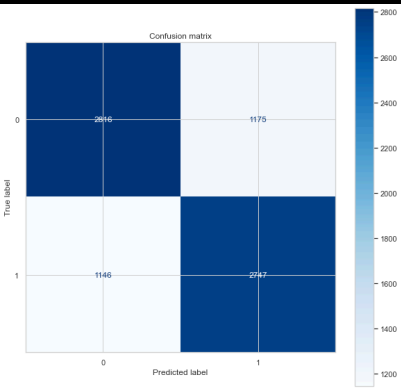
Classif.→ LR CV =.63 (89 train score)



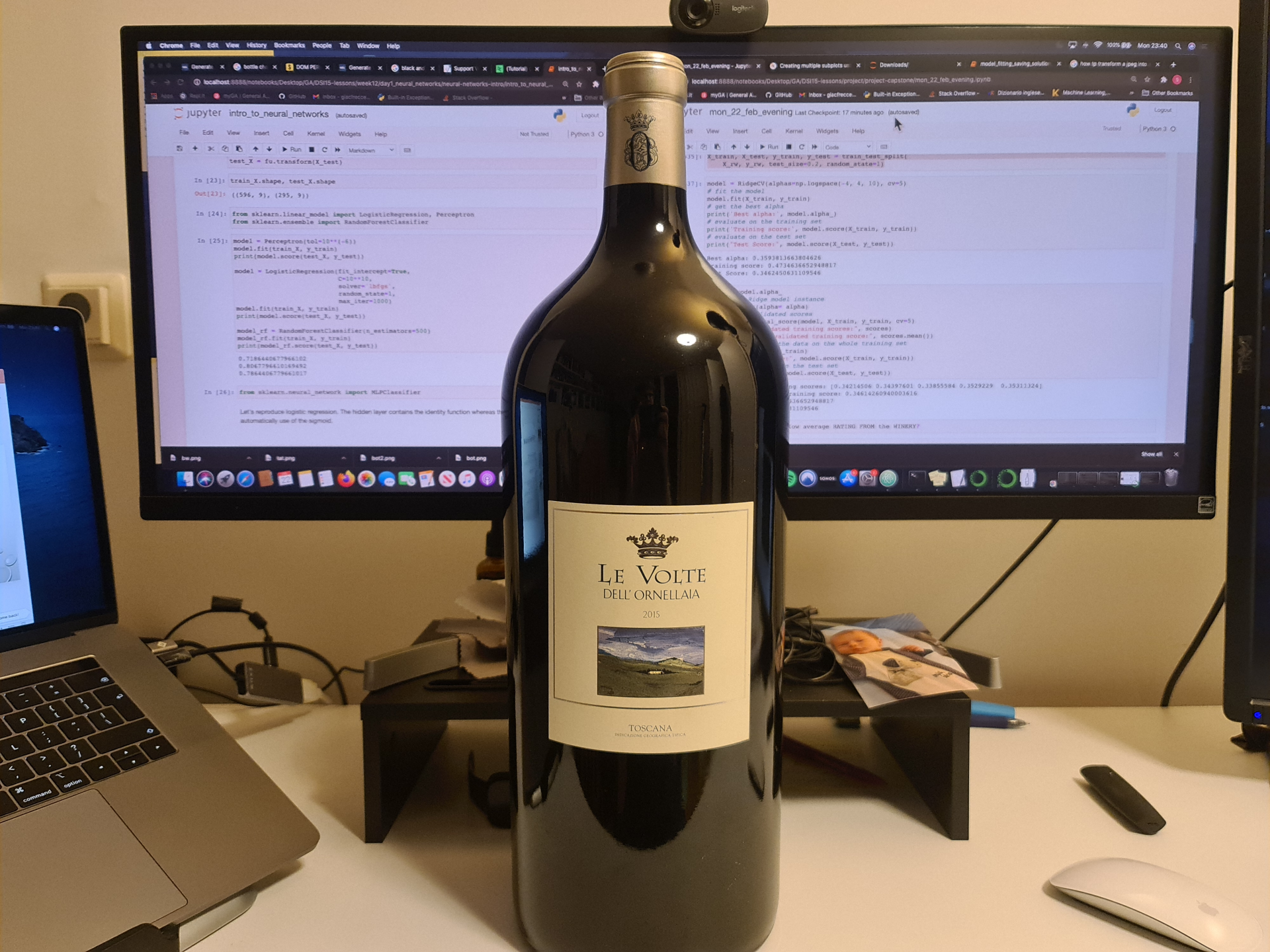


Predict above/below median Rating from Appellation

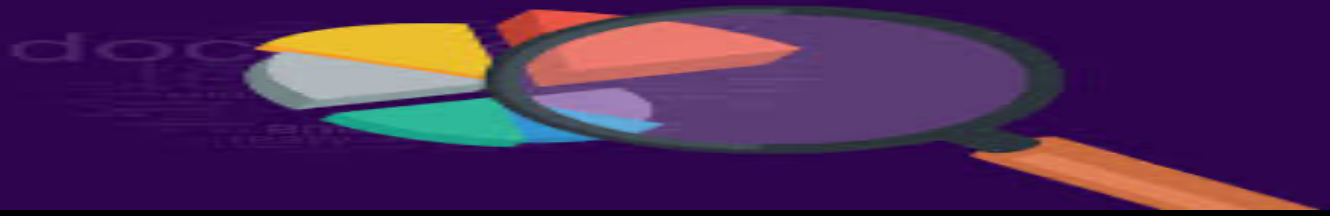
Classif.→
LR CV =.70
(70 train score)




WHAT'S NEXT?


jupyter intro_to_neural_networks (autosaved)
jupyter mon_22_feb_evening Last Checkpoint: 17 minutes ago (autosaved)
LE VOLTE
DELL' ORNELLAIA
2015
TOSCANA

LIMITS & IMPROVEMENTS



DATASET : biased towards red wines, need more data and expand beyond Italy
MODELS: more and more
CLASSIFIER: try to create one with more data
CLUSTERING: more and other techniques