A glass of ML

Wine pricing interpretation and prediction

Chechkov Eugene

Main task

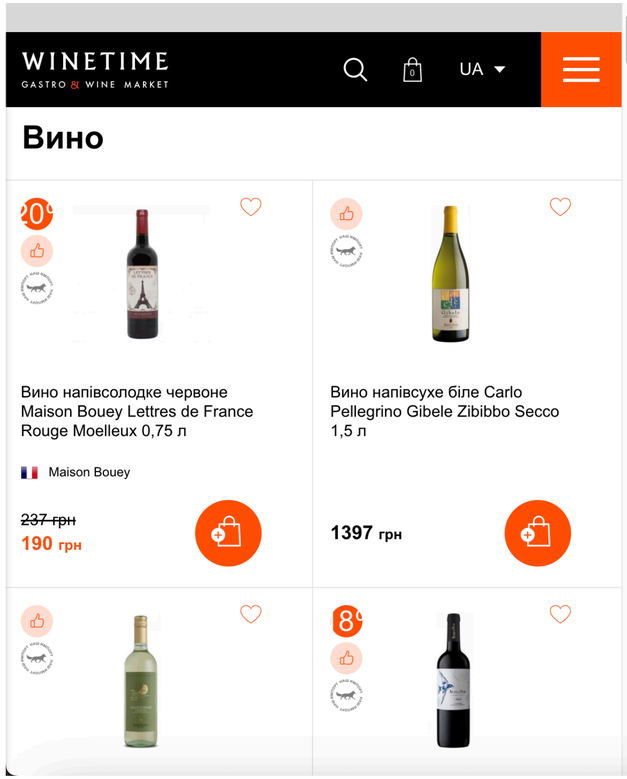
Can we understand by the data how wine prices are being formed?

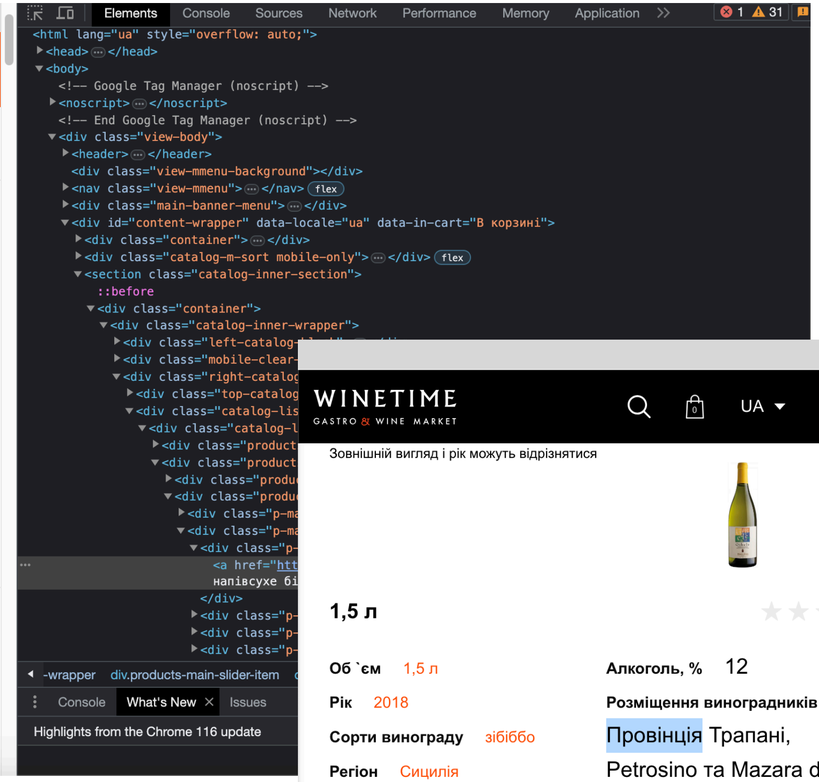
- Which features have most impact on wine price
- Which features have least impact on wine price
- Predict wine prices and whether the accuracy will be sufficient

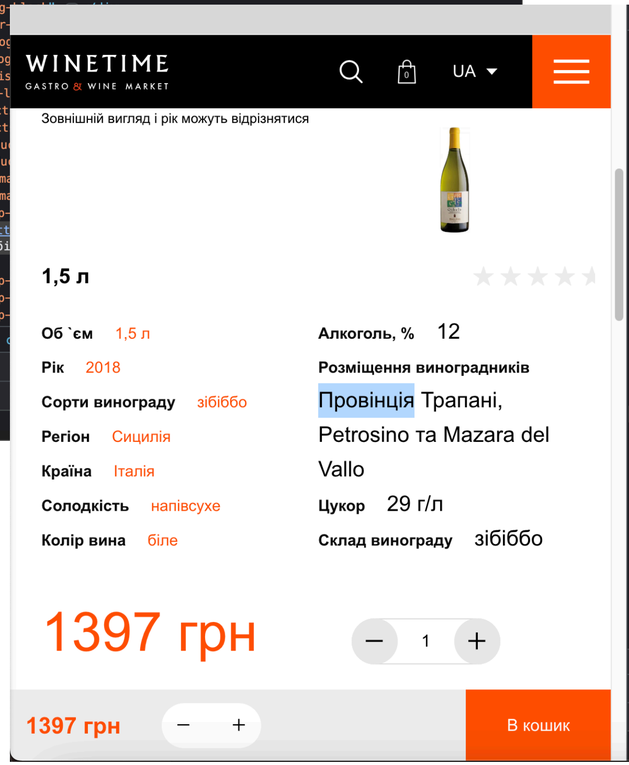
Who could be interested

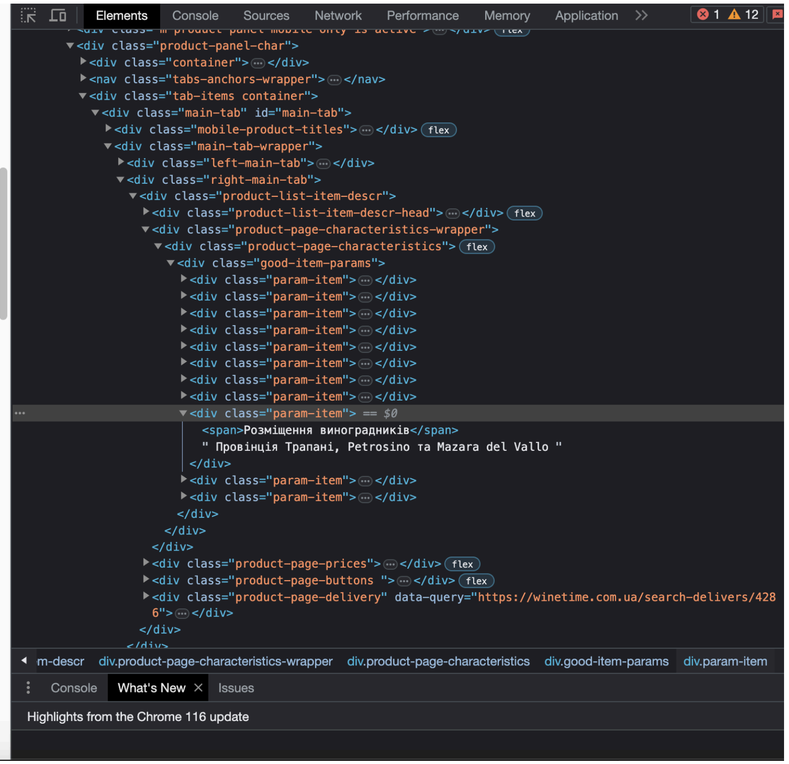
- Retailers
- Wine producers

Let’s parse a retailer









WINETIME
GASTRO & WINE MARKET

Preprocessing. Preprocessing? Preprocessing!

Data example

	Title	Producer	Vendor_code	Price	Temperature	Grape	Technology	Volume	Year	Brand	...	Sweet	Alcohol	Harvest
0	Вино сухе біле Dominio de Punctum Viento Alise...	Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.
1	Вино напівсолодке червоне Castelnuovo Vino Ros...	Castelnuovo	10369400	217 грн	14-15°C	корвіна, рондінелла	NaN	0,75 л	NaN	Castelnuovo	...	35 г/л	11.0	NaN
2	Вино сухе червоне Vintae El Picaro 0,75 л	Vintae	15426282	557 грн	16-18°C	тінта де торо	Біодинамічний спосіб ведення господарства. Фер...	0,75 л	2022.0	Vintae	...	2.4 г/л	14.5	Добірний урожай з 90-100-річних лоз, зібраний ...
3	Вино сухе біле Dominio de Punctum Viento Alise...	Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.

Producer	Vendor_code	Price	Temperature	Grape	Technology	Volume	Year	Brand	...	Sweet	Alcohol	Harvest	Additional_color
Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.	Яскравий золотистий зі відблисками лайму.
Castelnuovo	10369400	217 грн	14-15°C	корвіна, рондінелла	NaN	0,75 л	NaN	Castelnuovo	...	35 г/л	11.0	NaN	Багатий рубіновий.
Vintae	15426282	557 грн	16-18°C	тінта де торо	Біодинамічний спосіб ведення господарства. Фер...	0,75 л	2022.0	Vintae	...	2.4 г/л	14.5	Добірний урожай з 90-100-річних лоз, зібраний ...	Насичений, глибокий рубіновий.
Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.	Яскравий золотистий зі відблисками лайму.

Aroma	Taste	Interesting	Style	Potential	Degustations
Насичений із нотами запашних квітів.	Збалансований з нотами магнолії і персикового ...	NaN	NaN	NaN	NaN
Освіжаючий букет з нотками полуниці та смородини.	Вишуканий, солодкий смак з привабливим освіжаю...	NaN	NaN	NaN	NaN
Інтенсивний соковитий аромат темних ягід (чорн...	Теплий, соковитий смак, в якому домінують стиг...	На етикетках вин Matsu зображені фотографії ви...	NaN	NaN	NaN
Насичений із нотами запашних квітів.	Збалансований з нотами магнолії і персикового ...	NaN	NaN	NaN	NaN

Feature engineering

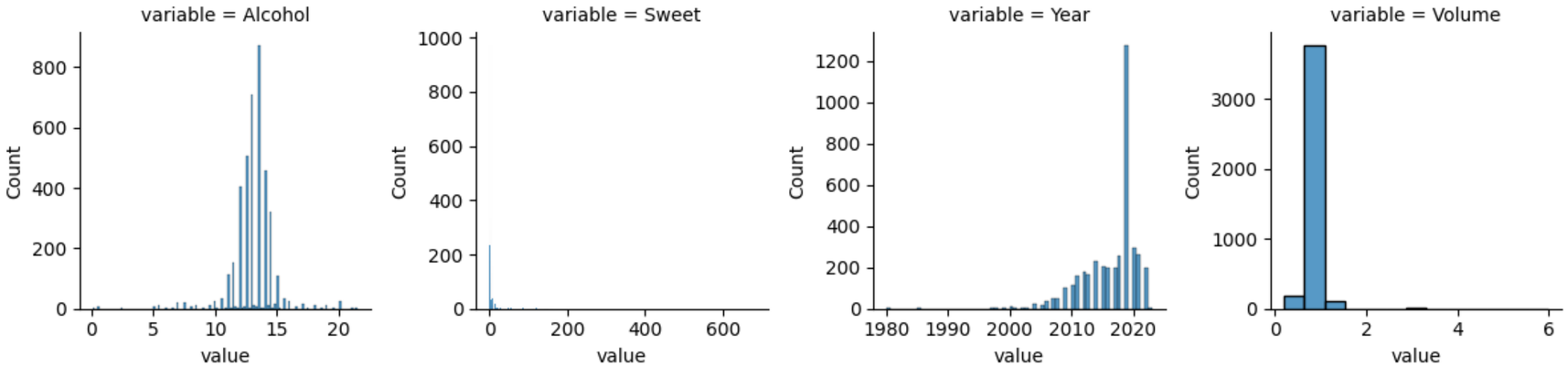
Classification

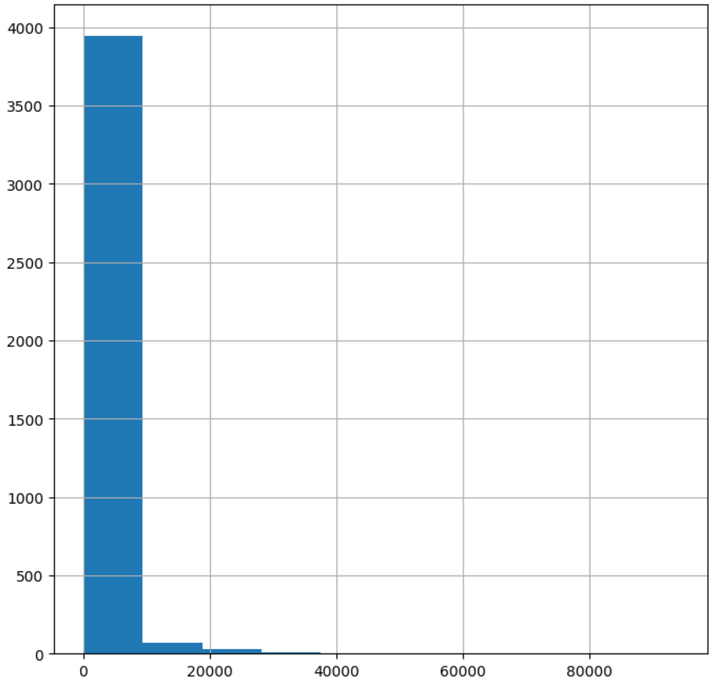
- Data for a couple of countries
- Countries classifications are different
- Spelling is different
- Result - binned into 3 categories : Regional wines, table wines, best wines

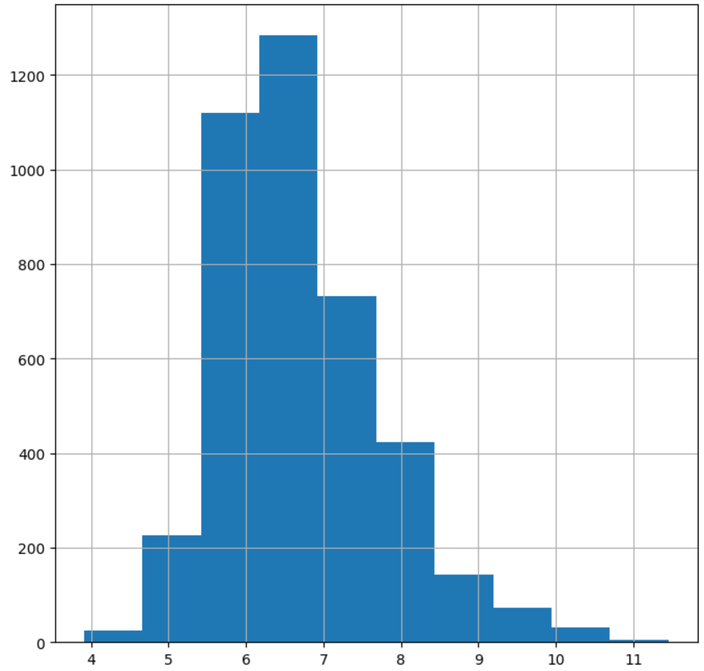
```
df['Classification'].unique()
```

```
array(['Indicazione Geografica Tipica (IGT)', nan, 'Toro DO', 'Kurant',
       'Denominazione di Origine Controllata (DOC)', 'Reserve',
       "Appellation d'Origine Controlee (AOC)", 'Vin de Table (VDT)',
       'Denominazione di Origine Controllata e Garantita (DOCG)',
       'Denominacion de Origen Calificada', 'Denominacion de Origen (DO)',
       'D.O. Carinena', 'KOP', 'Denominazione Di Origine Controllata',
       'Denomination of Origin', 'DOP', 'Appellation Bordeaux Contrelee',
       'витримане', 'Denominacion de Origen (D.O.)', 'Vin de Pays (VdP)',
       'DOCa Rioja', 'Indicazione Geografica Protette', 'DOC',
       'Appellation Alsace Controlee', 'Denominacao de Origine Protegida',
       'Qualitatswein',
       'Vino a Denominazione di Origine Controllata e Garantita (DOCG)',
       'DOC Douro', 'Denominazione Di Origine Controllata E Garantina',
       'D.O.C.G', 'IGP', 'Indication Geographique Protegee (IGP)',
       "Appellation d'Origine Protegee", 'Barolo D.O.C.G',
       'Appellation Medoc Controlee',
       'Appellation Saint-Emilion Grand Cru Controlle',
       'Indication geographique protegee (I.G.P.)',
       "Appellation d'Origine Protegee (AOP)", "Vin de Pays d'Oc",
       'Appellation Bordeaux Controlee', 'Barbaresco D.O.C.G',
       'Qualitatswein trocken', 'Appellation Cahors Controlee', 'IGT',
       'Appellation Medoc Contrelee',
       'Indication Geographique Protegee\xa0(IGP)',
       'Indicazione Geografica Tipica (IGT).', 'Eiswein',
       'Denominacao de Origem Controlada (DOC)',
       'Appellation Pomerol Contrelee', 'ординарне столове',
       'Denominazione di Origine Protetta (DOP)',
       'Vino ad Indicazione Geografica Protetta, Terre Siciliane',
```

Numerical features and target variable



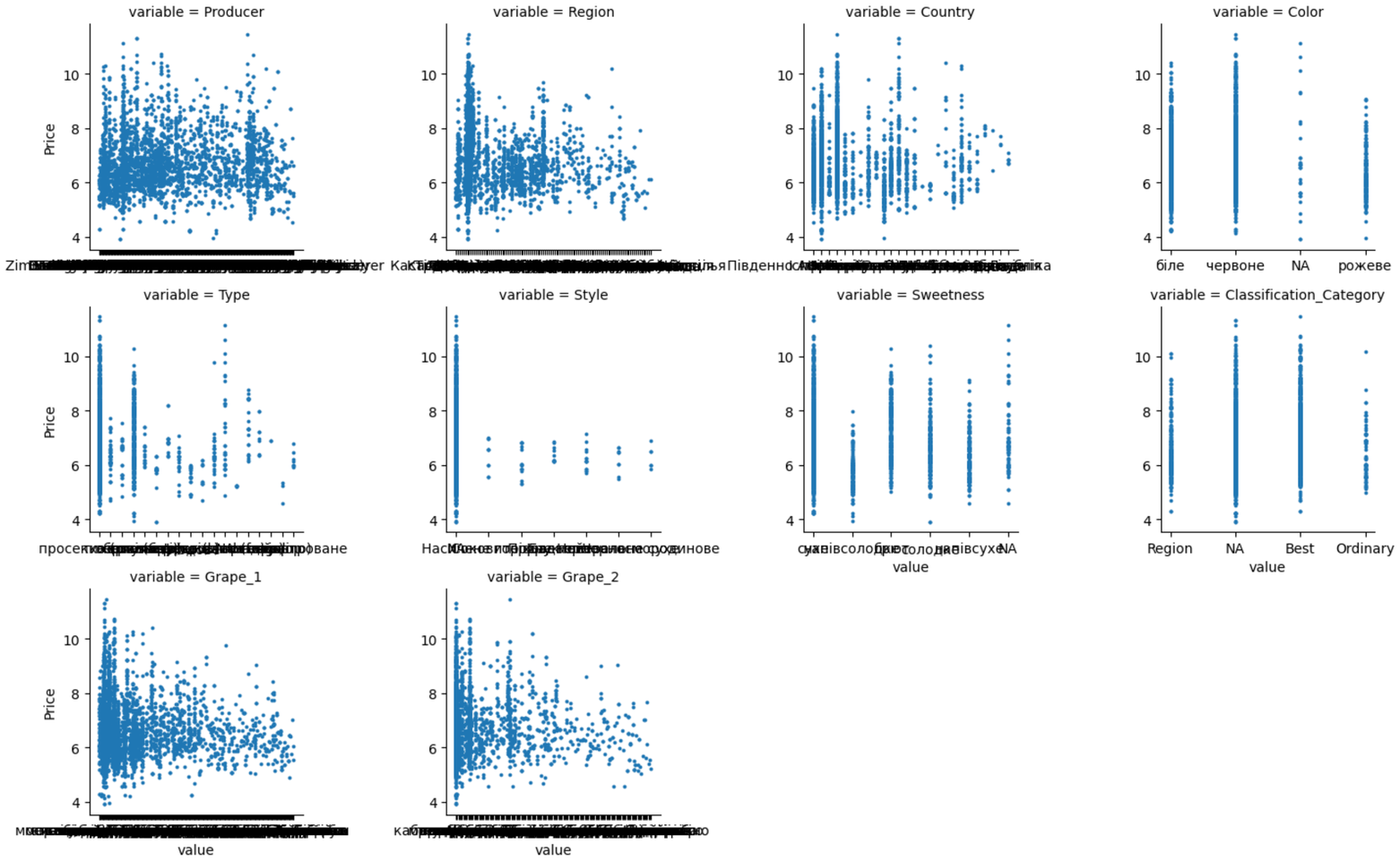




Categorical features

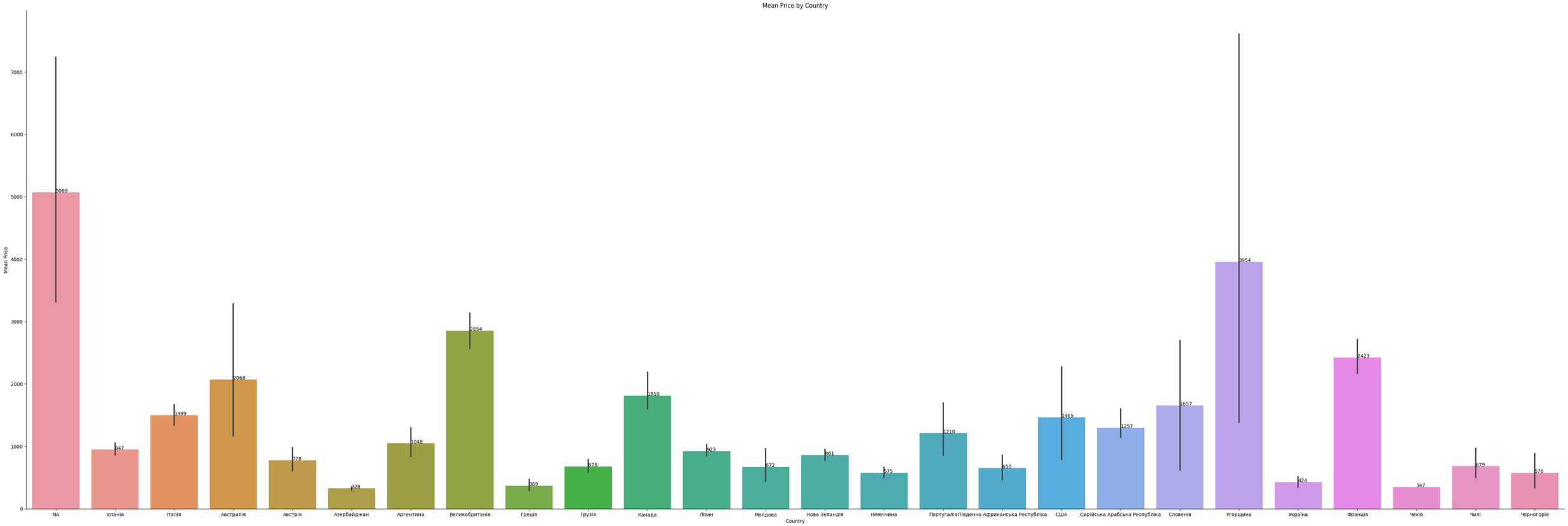
Main problem

- A lot… A LOT of unique categories. That could be a problem



Another problem

Some countries are absent



Let’s test some models

- Naive approach (just take average value
- Lasso/Elastic regression
- XGBoost

And the winner is definetely: dmlc XGBoost

Some results

Not so good (

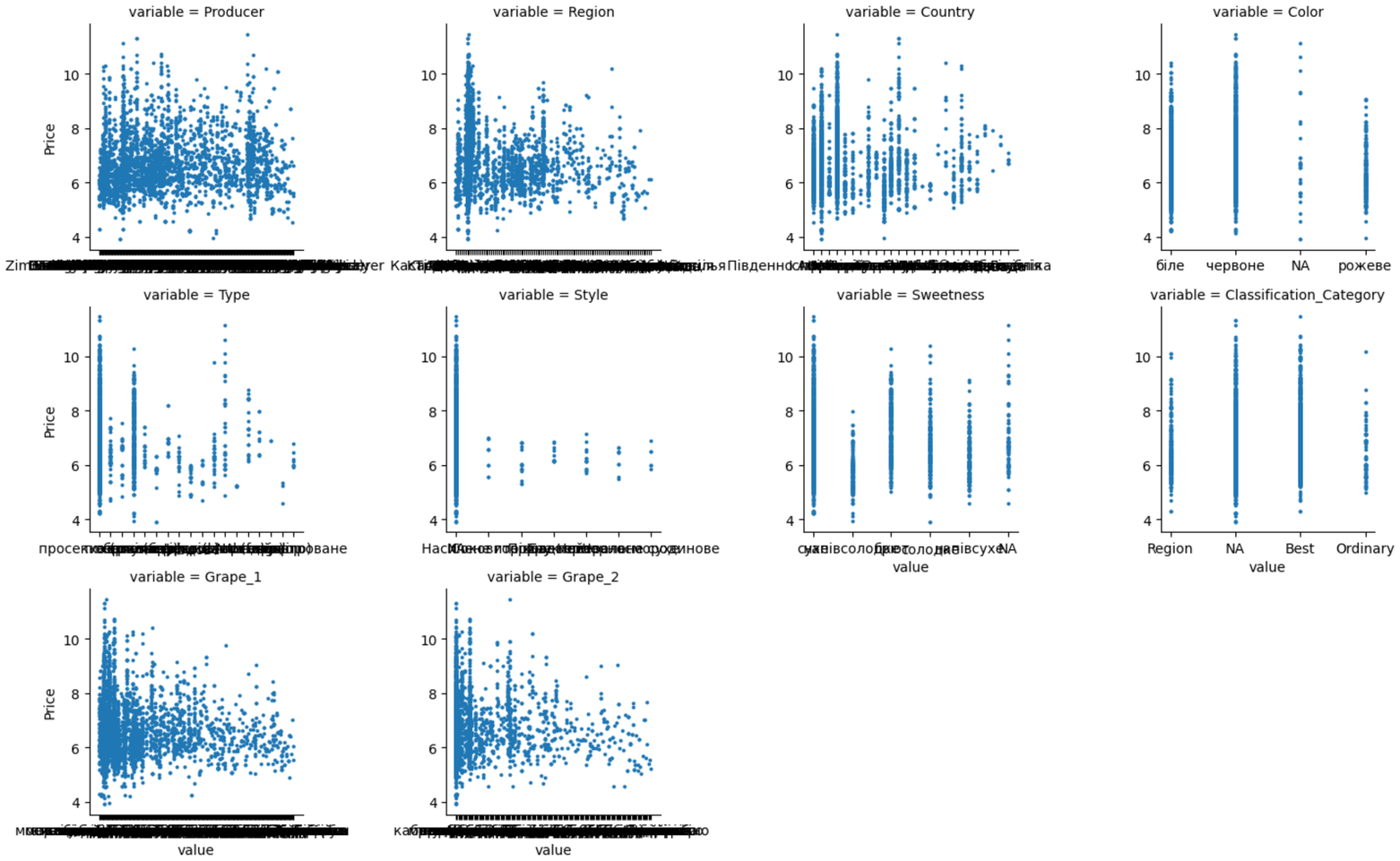
- We can achieve results almost twice as good as naive on RMSE and MAE and 2.4 times better on MAPE
- Still no perfect

	error_type	naive	lasso_cv	elastic_cv	xgb_regressor
0	MAE	1316.632166	1063.017128	1075.738863	663.959524
1	RMSE	4482.624008	4169.771960	4223.256121	2802.594270
2	R2	-0.040911	0.099314	0.076060	0.593117
3	MAPE	93.010235	56.099967	56.217935	39.029760

Categorical features

Main problem

- A lot… A LOT of unique categories. That could be a problem



Results

Main thoughts

- Champaigns and brut wines features have the biggest impact
- Most of impact features related with country/region or producer or wine type/ sweetness
- Alcohol, color, category have no impact
- A lot of features with small impact

Weight	Feature
0.0888	Region_Шампань
0.0350	Sweetness_брют
0.0275	Country_Україна
0.0184	Producer_LD Vins
0.0146	Country_Угорщина
0.0145	Producer_E.Guigal
0.0124	Grape_2_ віоньє
0.0118	Producer_Albert Bichot
0.0110	Type_ігристе
0.0108	Sweetness_солодке
0.0101	Region_Тоскана
0.0099	Producer_Cristom Vineyards
0.0098	Producer_Georg Breuer Weingut
0.0092	Producer_Les Grands Chais
0.0090	Region_Брда
0.0088	Producer_Bodega Chacra
0.0086	Country_Південно Африканська Республіка
0.0081	Producer_Maison Louis Latour
0.0079	Country_Італія
0.0078	Color_червоне
	... 1064 more ...

That’s all and thank you