INSEAD
The Business School for the World®

Wine Pricing Algorithm

Data Science Final Group Project

February 2020

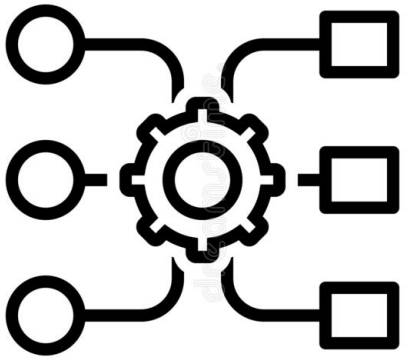
Section AA Group 8

- CUSTER Mike
- LEYVA SALAS Carlos
- LIU Bokai
- RIEHEMANN Jan
- SALVI Giuditta

Imagine you have a lovely small shop that sells wine …

… but how do you give the right price to each and every one of these bottles?

You don’t have to – our machine learning algorithm does it for you



Overview of our approach

- The data
- Cleaning process
- Factor engineering
- Analytics employed
- Add-on: Vintage analysis

We downloaded a massive dataset of wine

INSEAD
The Business School for the World®

136’000

Wine reviews in the database

13

Columns of data per wine review

1

Source website: WINE ENTHUSIAST

As many fields as on 19 thousand chess boards!

Cleaning the data included creating a new column, deleting values, combining categories and re-scaling scores

INSEAD
The Business School for the World®

1

Created new column for year

Initially, data did not have a column showing the year

Searched for text starting with “20” in the title and added next two digits

2

Deleted rows with too many missing values

We found that some rows had several missing values

After deleting rows with missing values, dataset still has 96k values

3

Combined rare categories

We combined rare categories into “others”

Because of the large number of countries, we grouped them by region (e.g. CEE, MENA)

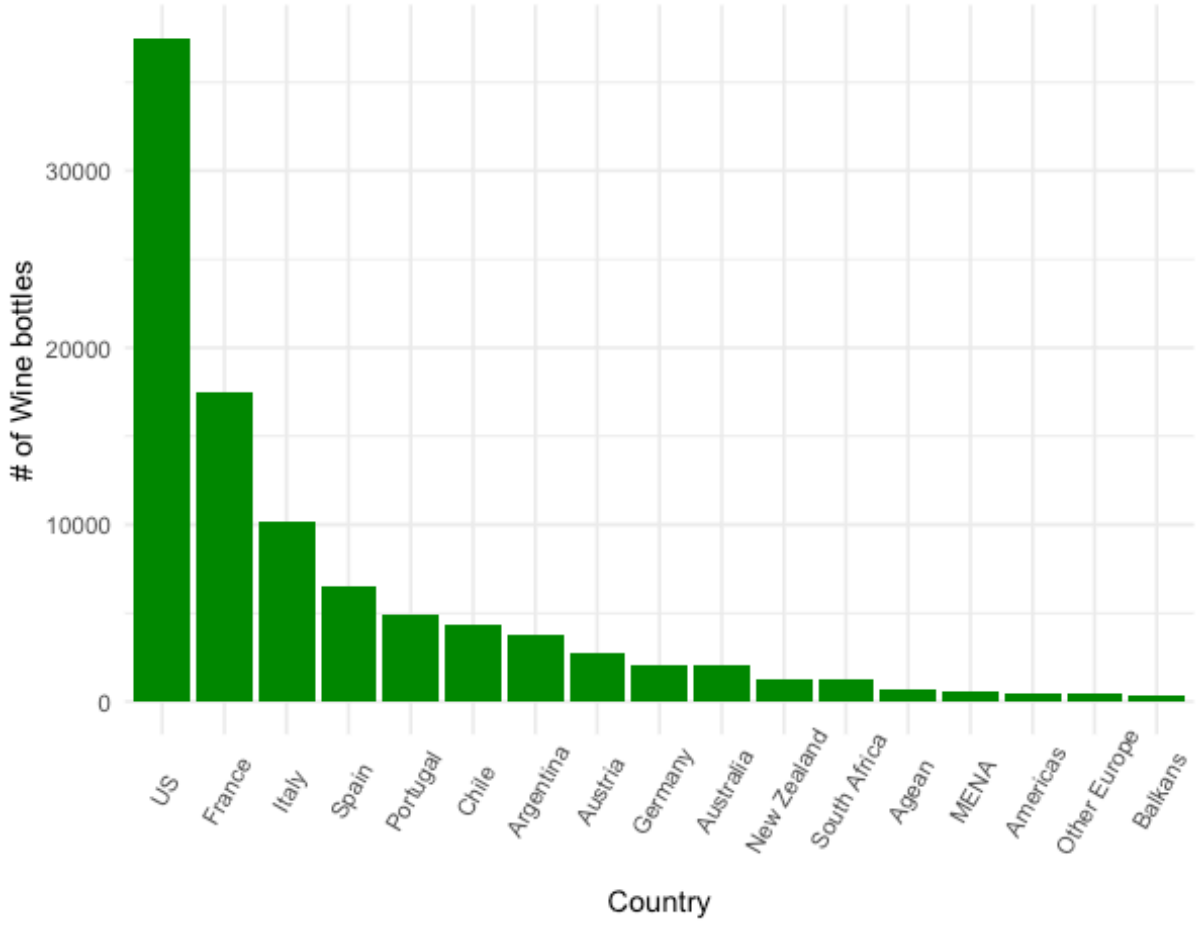
4

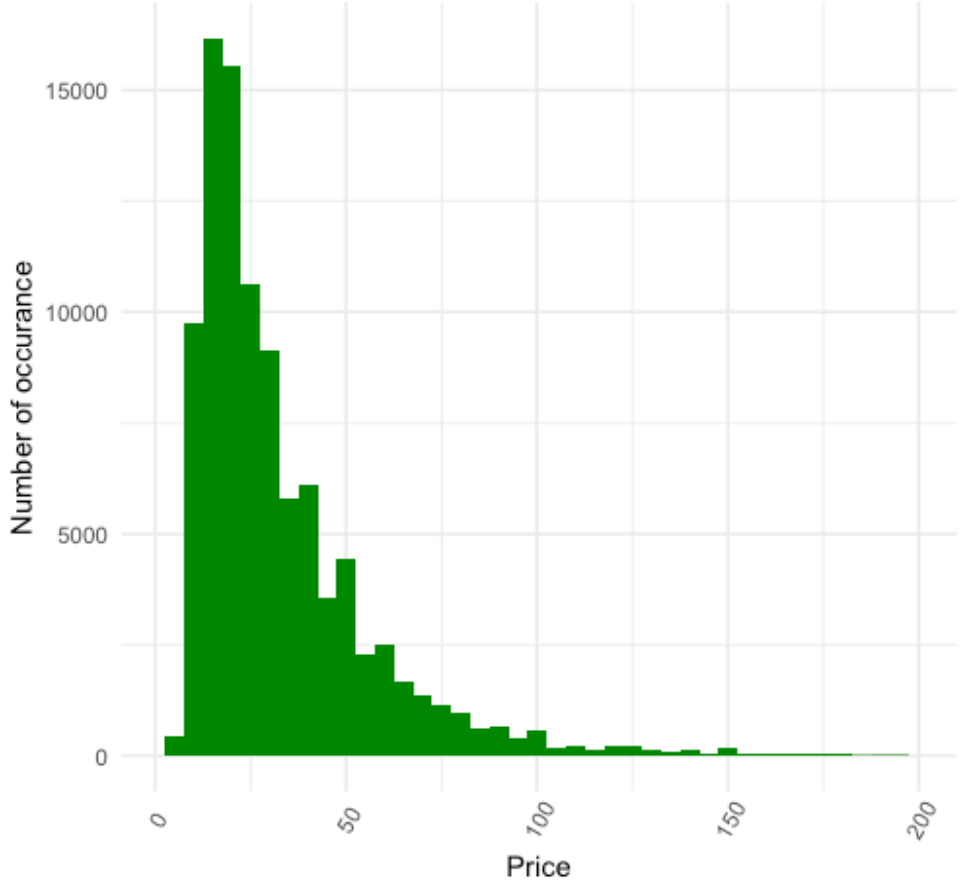
Scaled score for a normal distribution

Wine scores mostly between 80-100 and not evenly distributed

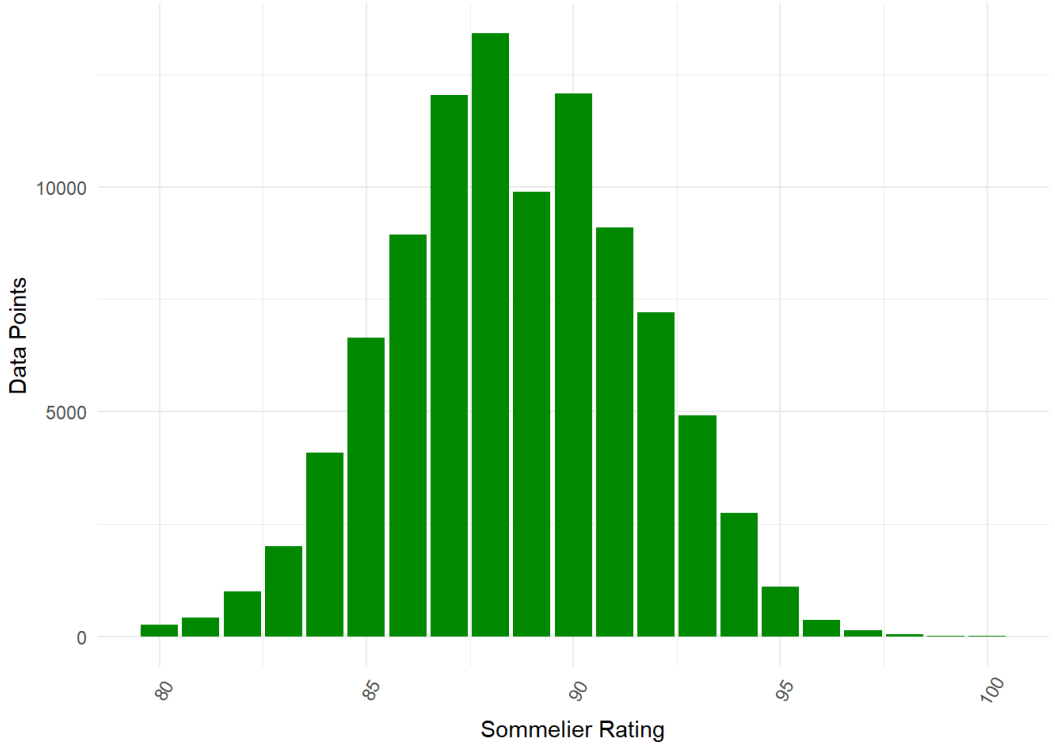
Thus we scaled ”Score” to a normal distribution with mean = 0 and std. dev. = 1

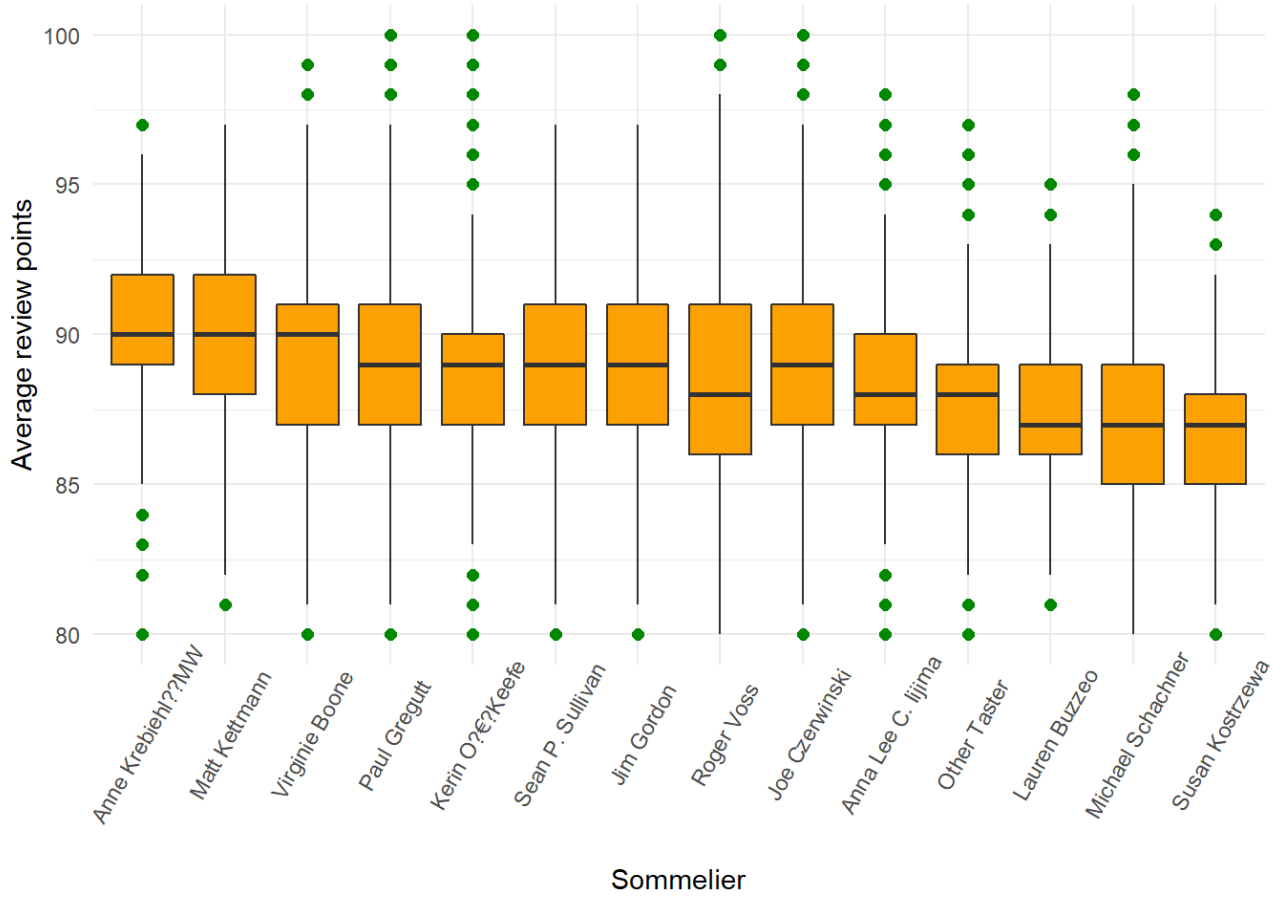
Some more interesting details on the data(1/2)





Some more interesting details on the data(2/2)





Factor engineering 1: We created several new variables through text analysis

INSEAD
The Business School for the World®

A Length of review as measure for how detailed the review was

B Average length of words used as proxy for “high-brow-ness” of reviewer

C Sentiment Analysis

Factor engineering 2: Sentiment analysis used to quantify opinion of reviewer

INSEAD
The Business School for the World®

We initially used a generic sentiment analysis library (AFINN) to evaluate reviews …

Using multiple libraries, we analyzed how positive or negative words sommeliers used to describe each wine

However, the sentiment library used very generic positive and negative words, yielding limited significance

… and eventually created our own library of words that are used to describe wine

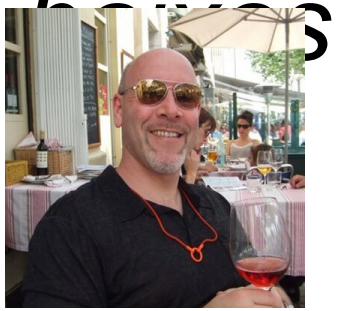
We analyzed the top and bottom 10% of reviews for the most used words

Manually sorting out meaningless words, we selected words that are specific to wine and attached them to positive or negative values between -3 (e.g. rotten) and +3 (e.g. phenomenal)

Our analysis provided wonderful Insights into the extravagances of wine reviewers - here is a “-3”

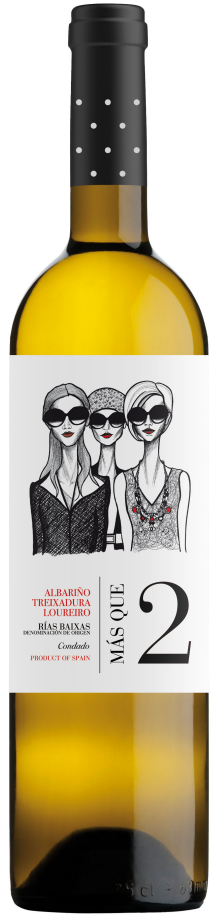
INSEAD
The Business School for the World®

“Pinched austere aromas of cat pee and oily sweat are extreme and unusual for a beiras wine…”

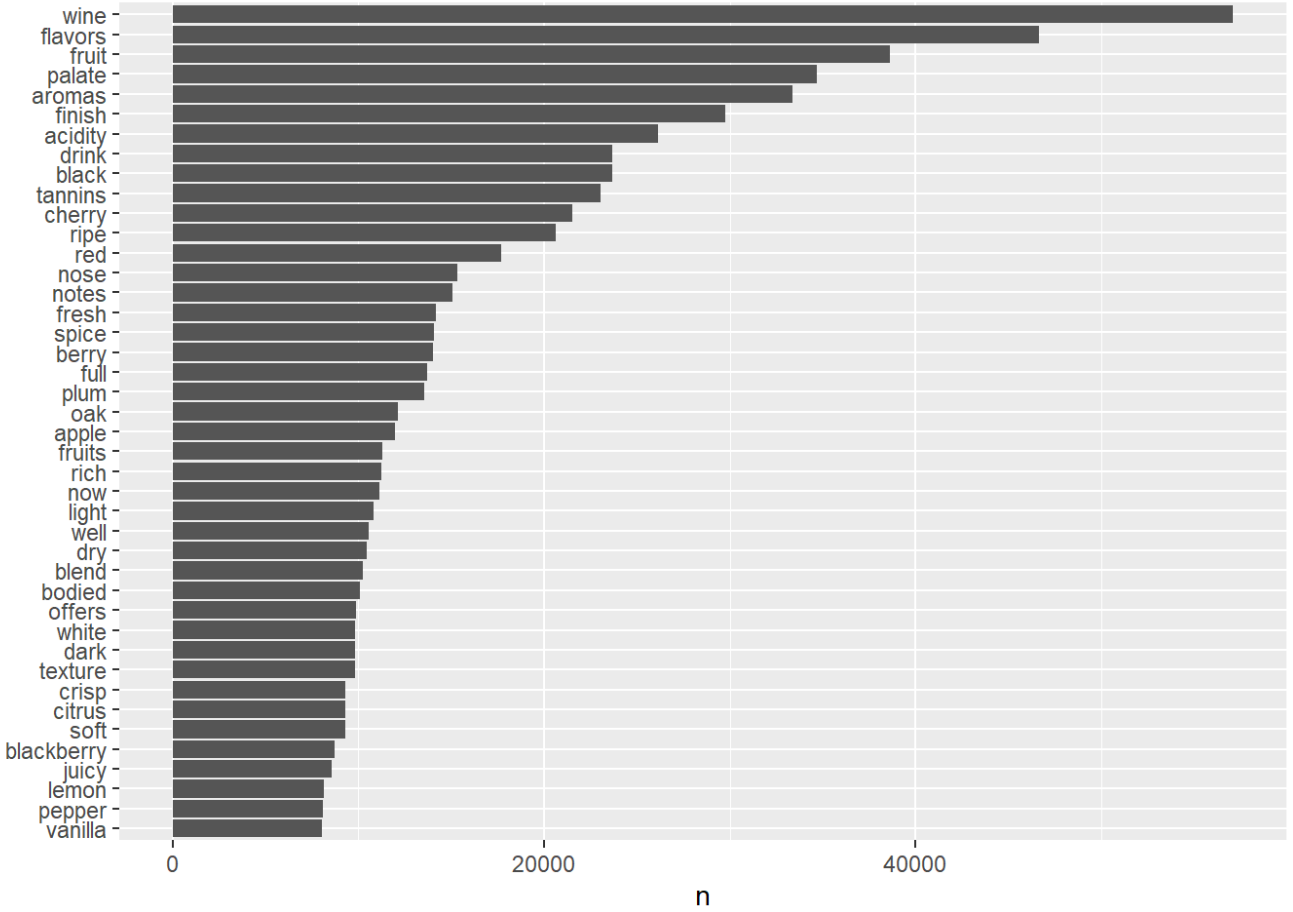


Michael Schachner

contributing editor to Wine Enthusiast Magazine



Steps to sentiment analysis – Remove stop words and analyze the most common words used in reviews



Next was to put every word into one row and use the two lexicons we selected to assign scores toe ach review

INSEAD
The Business School for the World®

1 Make every word a row

	observation	word
1	1	aromas
2	1	include
3	1	tropical
4	1	fruit
5	1	broom
6	1	brimstone
7	1	dried
8	1	herb
9	1	palate
10	1	overly
11	1	expressive
12	1	offering
13	1	unripened
14	1	apple
15	1	citrus
16	1	dried
17	1	sage
18	1	alongside
19	1	brisk
20	1	acidity
21	2	ripe
22	2	fruity
23	2	wine
24	2	smooth
25	2	still

2 Use the two lexicons to assign values to all words

Our own lexicon

21	screwcap	-1
22	full	1
23	bodied	1
24	juicy	2
25	bright	2
26	firm	2
27	chemicals	-3

AFINN

	word	value
1	abandon	-2
2	abandoned	-2
3	abandons	-2
4	abducted	-2
5	abduction	-2
6	abductions	-2
7	abhor	-3
8	abhorred	-3
9	abhorrent	-3
10	abhors	-3
11	abilities	2
12	ability	2

3 Sum up values by description

sentscore	winewordsscore
2	0
2	9
0	0
0	-2
5	0
1	2
2	3
0	3
7	2
4	2
4	2
0	1
2	5

Multiple linear regression algorithm resulting in MAPE of 34.56, which corresponds to fair prediction quality

INSEAD
The Business School for the World®

We used a linear regression algorithm , considering some important details

To train the algorithm, we split data into 80% training data and 20% testing data

Model included log(price) to correct for prices that are highly skewed to the left

Included combination of year and province in the model, making model specific to the combination of year and province

(… so that the model can think "*oh, a 2012*Bordeaux, how refined!*")

Algorithm provided fair prediction quality, especially for some countries

Model achieved an overall MAPE of 34.56%, which was significantly lower than initial >50%

Lower r-squared in the US and Austria, however higher r-squared in combined categories and Australia

As the model contains many factor variables, the final model contained 402 degrees of freedom

INSEAD
The Business School for the World®

Log(Price) = Country + Province + Variety + Review Length + High Brow + Sommelier+ AFINN Score + Own Sent Score + Province*Year

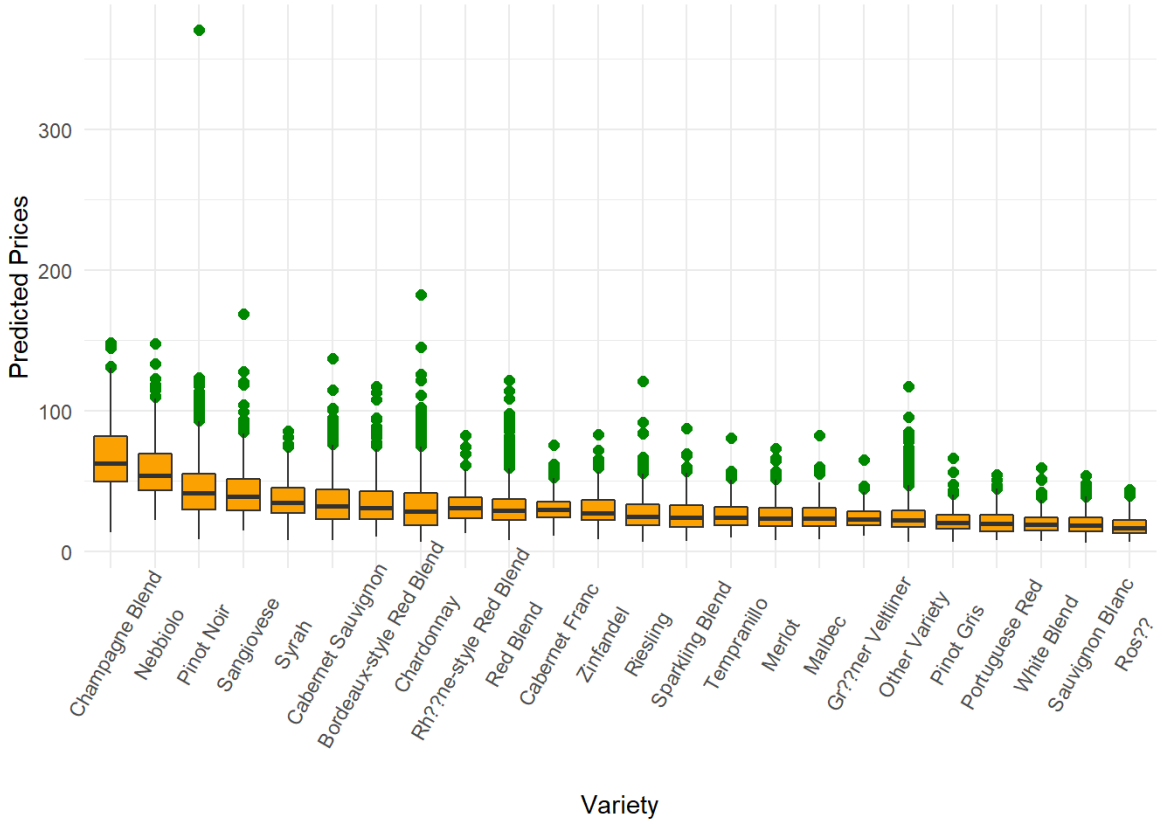
Model Insights

- Certain Country's influence on price (Germany, France and the US) was more significant than others (i.e. Italy, Australia) on price than others
- Both the AFINN score and our own Sent Score had a highly significant and inverse relationship with the price of the wine, i.e. the more positive the review
- Variety was the most reliable predictor of price of price (based on significance), it has the most significant factors

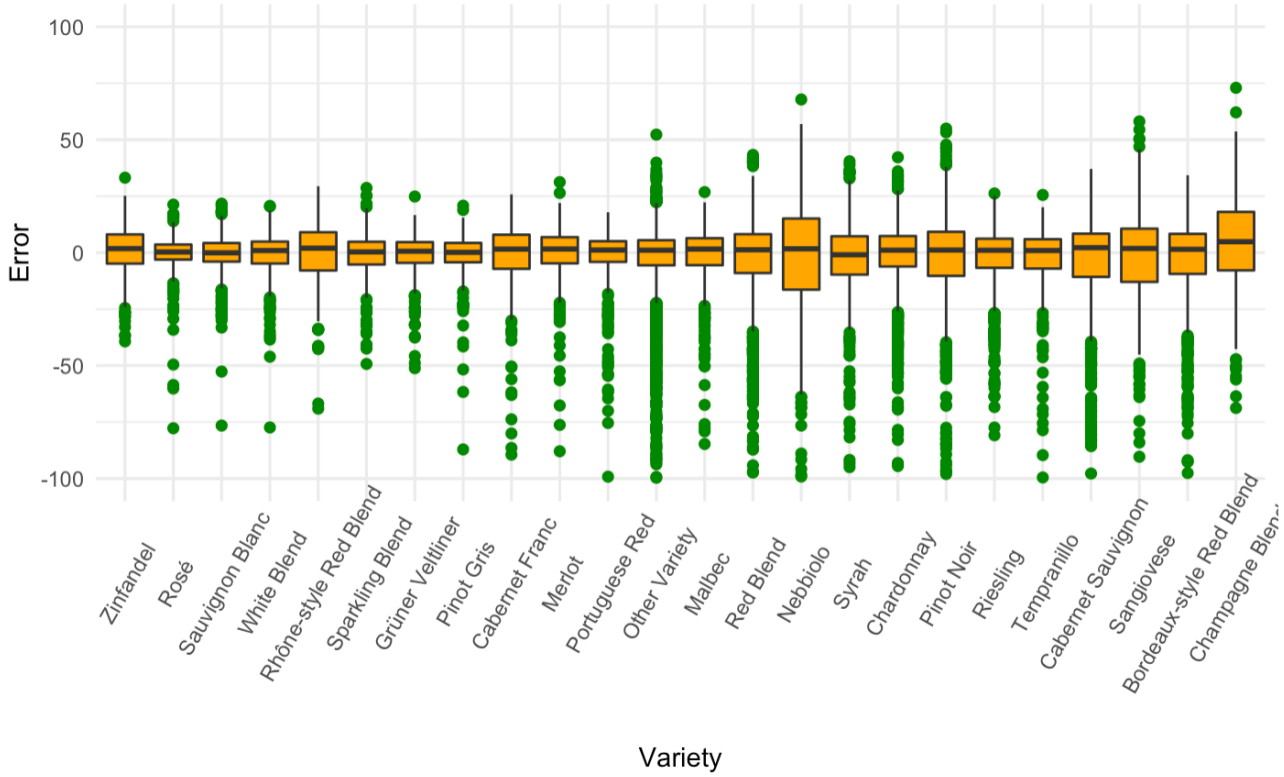
Prediction results

INSEAD
The Business School for the World®

1 Predicted Prices by variety



2 Error by variety of wine



Finally, we created an additional analysis of US wines, adding an external factor describing the quality of each vintage

INSEAD
The Business School for the World®

United States

	Wine Variety	Region	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
CALFORNIA	Cabernet Sauvignon	Napa	95	94	95	95	89	89	89	92	95	90	95	90	90	93	98	85	93	85	96	93	95	97	90	93	94	92
		Sonoma	93	93	94	93	88	87	87	90	92	87	89	87	89	88	93	84	90	84	91	90	91	92	88	89	90	89
	Chardonnay	Napa	90	92	91	91	87	87	88	87	90	86	86	85	85	90	90	88	88	87	94	90	92	90	87	90	90	91
		Russian River Valley	92	94	94	93	89	89	87	90	92	86	91	93	91	97	93	90	92	88	95	92	94	91	88	91	92	92
		Carneros	90	94	93	92	88	88	87	89	91	85	87	92	91	95	93	89	89	85	91	90	89	91	87	89	90	89
		Central Coast	91	92	92	92	90	91	88	90	92	87	90	93	92	94	90	89	89	85	94	88	90	92	87	89	90	91
		Santa Barbara	92	93	93	93	92	92	90	91	92	90	89	92	92	94	91	88	91	88	95	90	91	94	86	90	91	90
	Pinot Noir	Anderson Valley	94	93	95	96	91	93	93	92	95	88	94	90	90	87												
		Sonoma Coast	92	94	95	95	90	92	91	93	95	87	95	93	89	88	89	87	90	85	90	88	89	89				
		Russian River Valley	92	94	95	95	89	91	90	92	96	87	95	93	90	89	90	89	91	85	91	90	91	92	90	92	91	90
		Carneros	91	93	94	93	87	90	89	90	94	86	93	89	89	85	87	85	87	83	85	86	88	90	86	88	88	89
		Central Coast	93	93	92	92	88	93	89	89	94	88	92	89	93	95	91	88	89	86	93	87	88	90	86	87	89	89
		Santa Barbara	94	93	93	94	90	92	91	92	95	90	95	93	94	94	92	89	90	84	95	89	90	92	88	90	91	91
	Syrah	North Coast	92	91	91	92	89	88	88	88	93	88	85	89	89	88	92	84	89	83								
		Central Coast	91	92	90	89	88	89	87	87	90	85	84	86	92	88	91	84	86	83								
		South Coast	88	88	88	88	87	90	90	88	94	91	87	89	91	89	92	85	88	83								
	Zinfandel	Sonoma	93	93	93	92	89	89	89	88	93	86	90	90	89	88	91	86	90	86	88	90	91	92	92	90	91	90
		Napa	93	93	93	92	89	89	89	89	94	87	87	90	88	86	91	85	90	85	89	88	90	91	89	89	92	91
		Paso Robles	91	91	92	91	87	87	87	87	92	87	85	87	86	86	87	85	89	84	86	86	87	88	86	87	87	87
		Sierra Foothills	94	93	93	92	89	91	90	88	90	87	87	89	87	86	85	84	85	84	84	85	86	84	84	85	86	85
OR	Pinot Noir	Willamette Valley	94	93	92	93	91	87	91	92	86	94	89	92	86	88	89	91	94	92	84	87	86	90	91	88	87	90
	Whites	Willamette Valley	93	93	93	92	89	88	92	94	86	91	90	93	86	88	89	90	83	90	85	87	85	91	90	87	88	90
	Reds	Southern Oregon	88	90	89	90	90	89	88	90	89	87	87	90	85	87	88	89	90	89	84	85	84	89	88	87	86	88
WA	Cabernet, Merlot	Columbia Valley	91	93	91	95	89	92	92	88	91	92	95	89	92	94	92	88	96	91	87	86	89	90	87	88	86	89
	Syrah	Columbia Valley	90	90	90	95	91	89	91	90	89	91	93	90	91	94	93	88	92	89	87	88	89	89	87	87	86	89
	Whites	Columbia Valley	89	90	88	96	92	90	89	94	89	90	92	88	90	91	89	89	87	90	88	89	89	89	90	89	89	88
NY	Reds	Finger Lakes	89	89	88	92	85	91	85	89	89	85	89	86	84	89	91	86	90	88	86	84	91	88	90			
	Whites		90	89	89	90	88	90	88	90	88	89	90	88	90	91	93	89	92	89	90	88	92	90	91			
	Reds	Long Island	90	90	92	92	84	92	85	89	88	87	87	90	84	85	90	87	86	93	89	86	92	90	89	80	91	90
	Whites		89	89	88	89	86	90	86	89	90	88	87	91	88	89	90	88	89	93	90	88	92	90	89	84	91	90

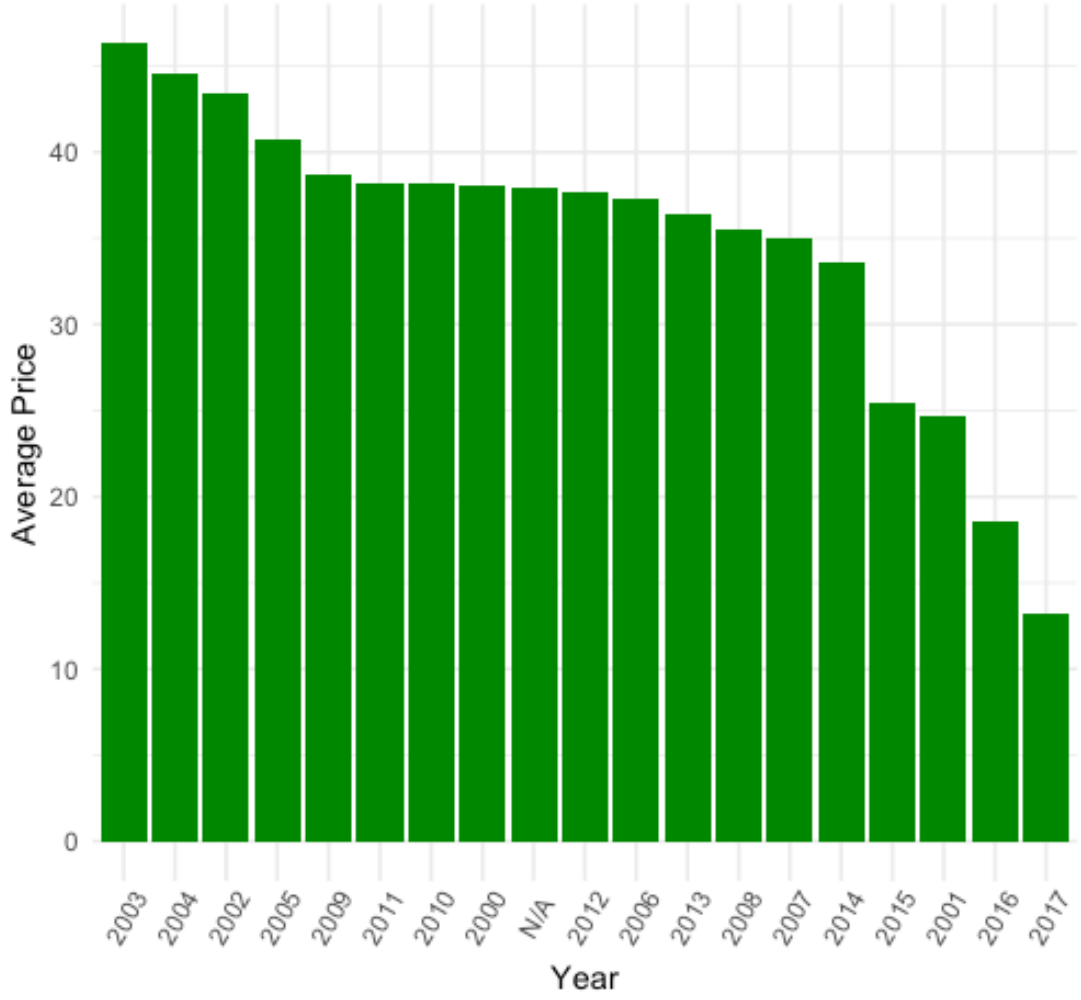
In a second attempt to build a more predictive model …

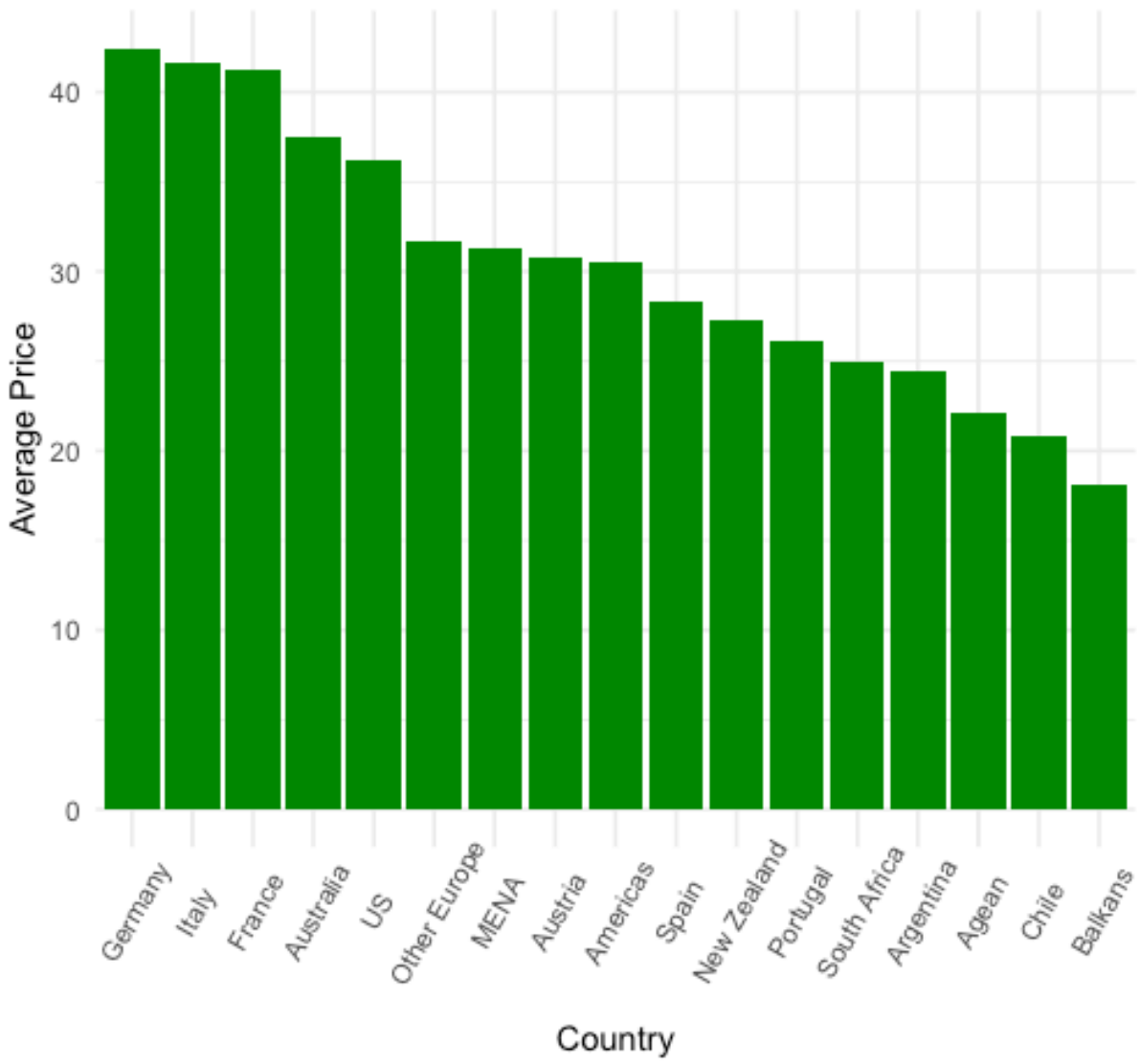
… we focused on US wines and included an external source that shows recommendation when to drink a certain wine (region & year)

Overall, predictive quality of model was high at 26.56% MAPE

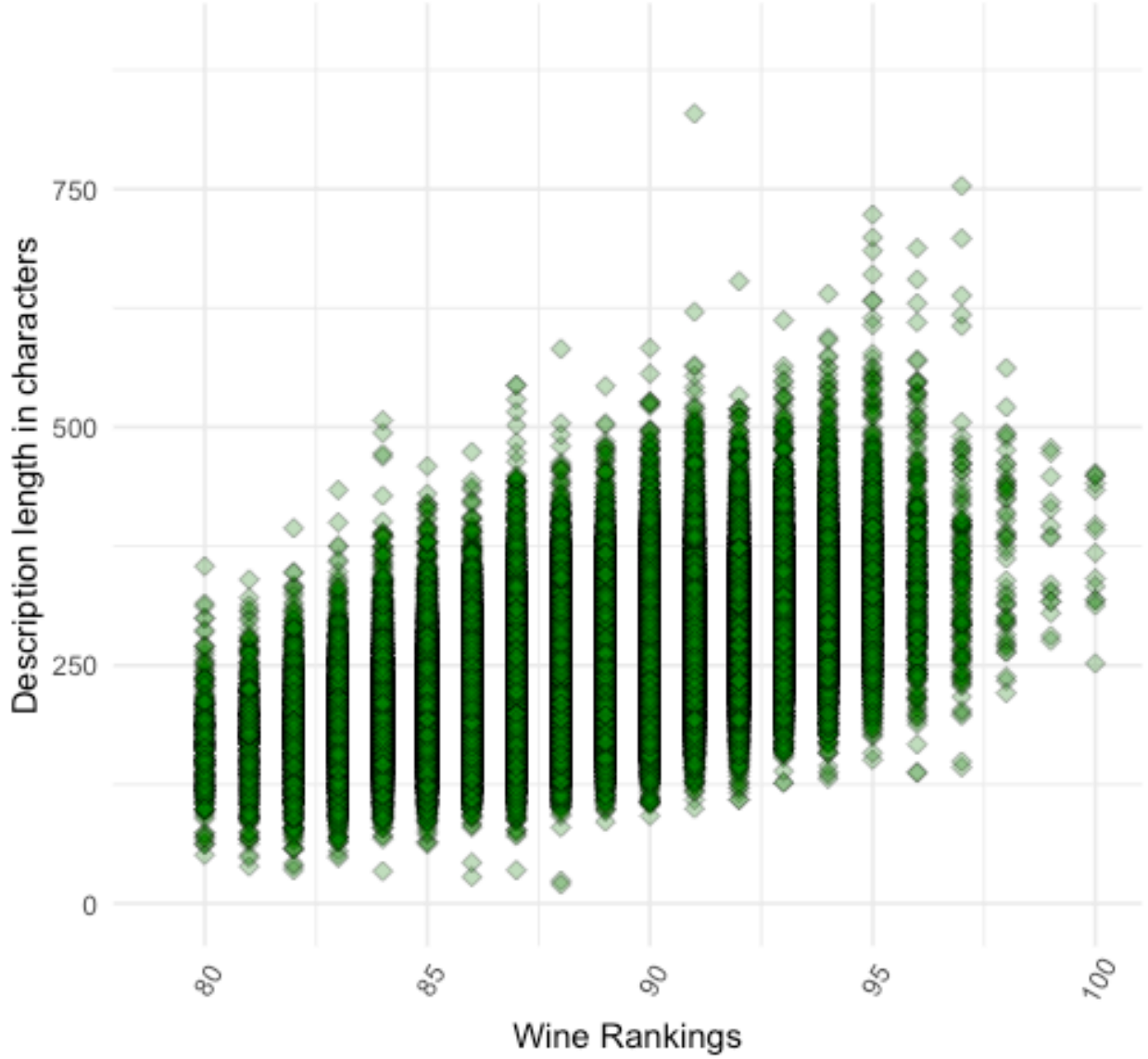
APPENDIX

Some more interesting details on the data





Indeed, we did find higher rankings with longer reviews



The Business School
for the World®

Europe | Asia | Middle East