Wines Around the World

YZ Analytics

Project Overview & Motivation

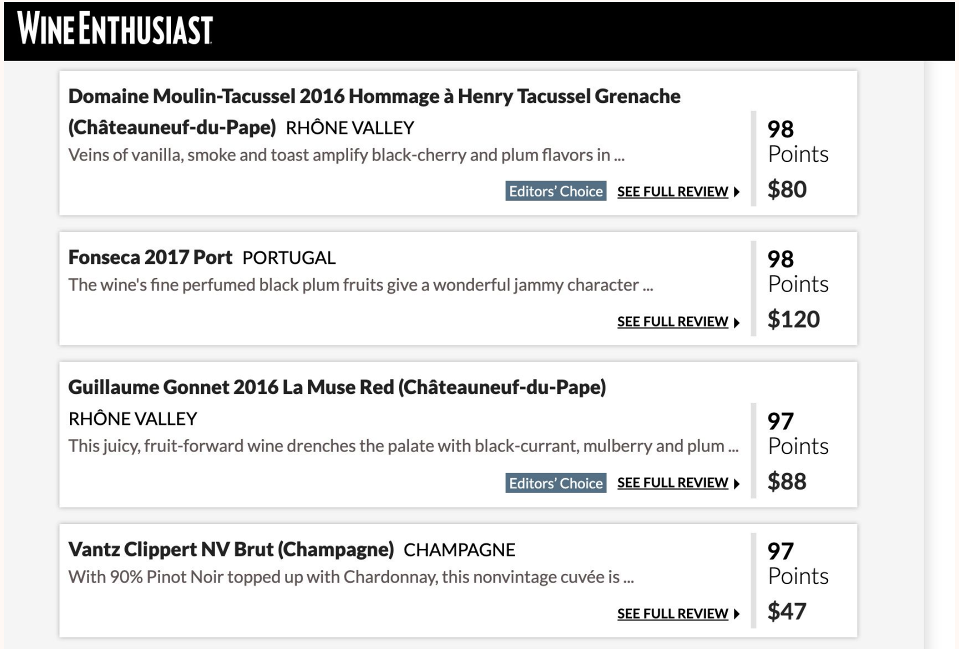
- Wine is one of the most popular alcoholic drinks, but besides sommeliers, many people do not know much about the different features of wine

What are the most important features in determining a good wine?

How can we visualize information about wine for the average layperson to explore?

Data

- Kaggle data of 130k wine reviews originally from Wine Enthusiast Magazine (Thoutt, 2017)
- 13 variables: country, description, designation, points, price, province, region_1, region_2, taster_name, taster_twitter_handle, title, variety, winery



WINE ENTHUSIAST

Domaine Moulin-Tacussel 2016 Hommage à Henry Tacussel Grenache (Châteauneuf-du-Pape) RHÔNE VALLEY
Veins of vanilla, smoke and toast amplify black-cherry and plum flavors in ...
Editors' Choice SEE FULL REVIEW ▸
98 Points **$80**

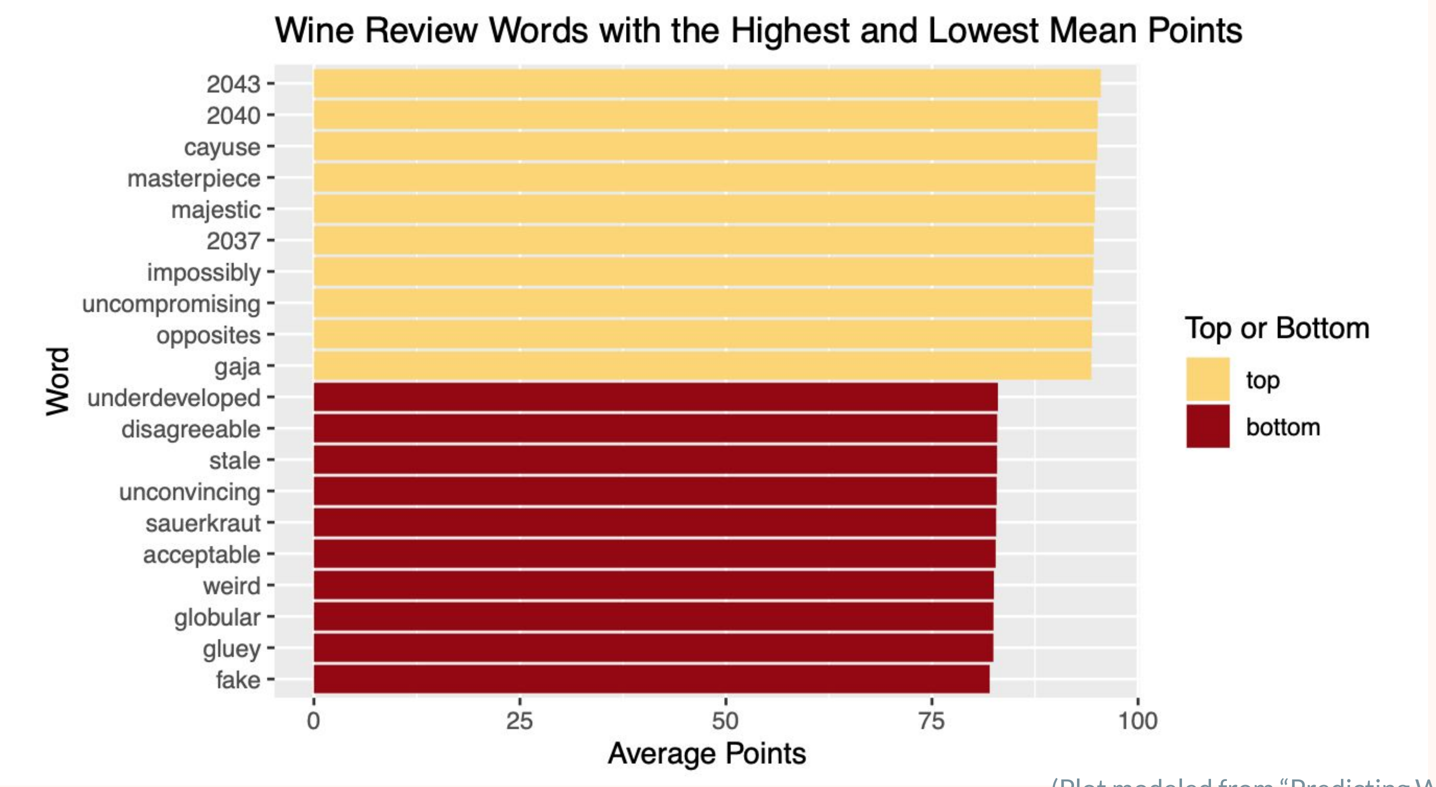
Fonseca 2017 Port PORTUGAL
The wine's fine perfumed black plum fruits give a wonderful jammy character ...
SEE FULL REVIEW ▸
98 Points **$120**

Guillaume Gonnet 2016 La Muse Red (Châteauneuf-du-Pape) RHÔNE VALLEY
This juicy, fruit-forward wine drenches the palate with black-currant, mulberry and plum ...
Editors' Choice SEE FULL REVIEW ▸
97 Points **$88**

Vantz Clippert NV Brut (Champagne) CHAMPAGNE
With 90% Pinot Noir topped up with Chardonnay, this nonvintage cuvée is ...
SEE FULL REVIEW ▸
97 Points **$47**

(Wine Enthusiast, 2019)

Text Analysis



(Plot modeled from “Predicting Wine Ratings,” 2018)

Creating features from the description variable

- Ranked words by mean Term Frequency-Inverse Document Frequency, or TF-IDF (Manning et al., 2008)
  - **Term Frequency:** how often a term appears in a document
  - **Inverse Document Frequency**: natural log of the total numbers of documents/number of documents with term
- Created a Document-Term Matrix (DTM) with the top 200 words with the highest mean TF-IDF values (Nabi, 2018)
  - DTM example: “The cow says moo.” & “The dog says woof.”

Document	**the**	**cow**	**says**	**moo**	**dog**	**woof**
Sentence 1	1	1	1	1	0	0
Sentence 2	1	0	1	0	1	1

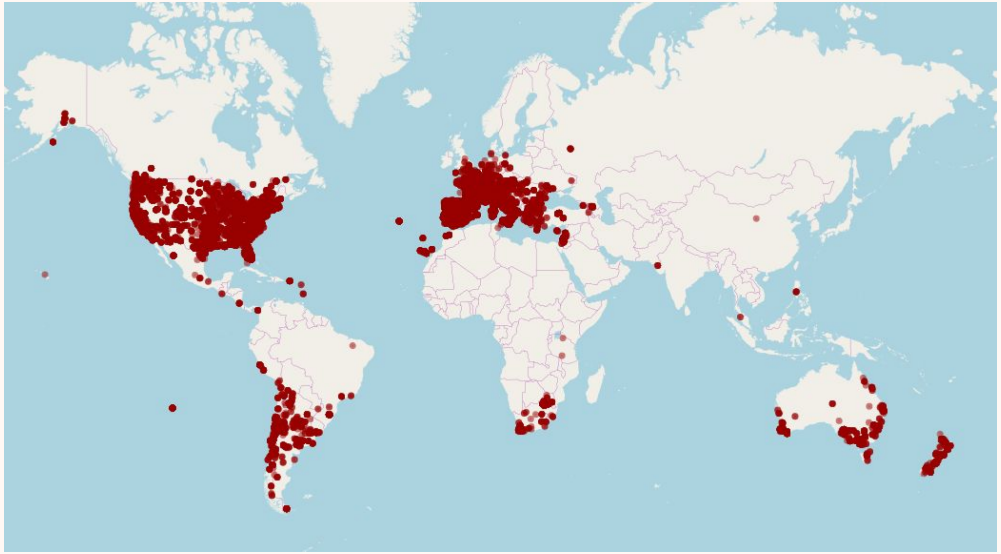
Predicting Wine Quality

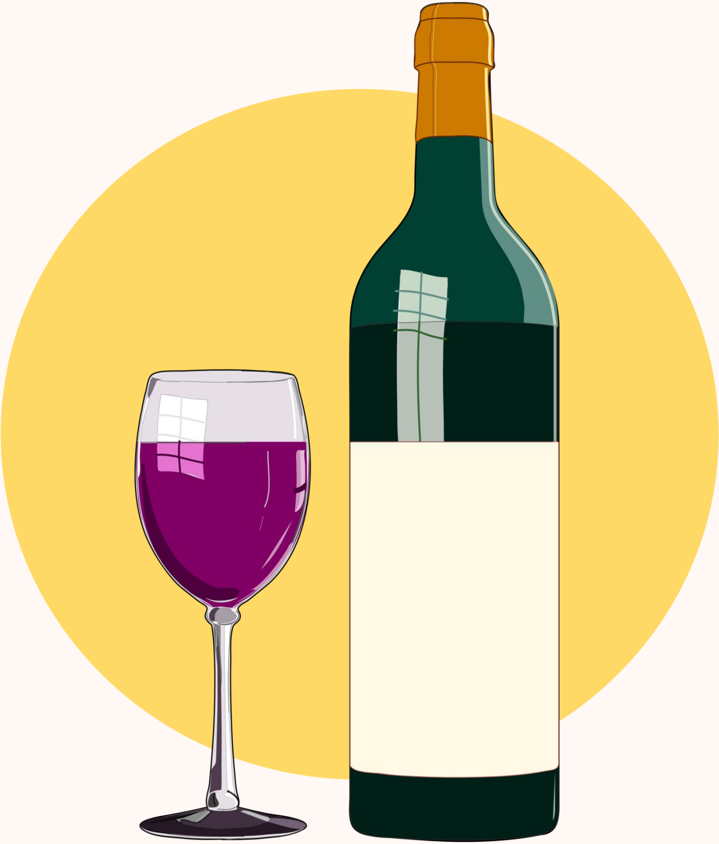
- Predicted points through gradient boosting algorithm (**gbm** package)
- Evaluated performance through square root of MSE - how many points on average the prediction is off by on the test set
    - 5 trees: 2.66
    - 500 trees: 1.83

Variable	Relative Influence
price	44.2872441
variety	12.8744278
province	12.8074735
rich	1.6195324
complex	1.5243999
simpl	1.4679213
long	1.1043904
delici	1.0599025
black	0.9946763
concentr	0.9875296

Geocoding and Visualization

- Goal: Create an interactive map and wine catalog based on the data
- Geocoded 80% of observations in original dataset





Shiny App

https://szablah.shinyapps.io/wine/

https://r.amherst.edu/apps/szablah20/wine/

Limitations & Future Directions

Scrape newer wine reviews, 2018-present

Use newer and faster gradient boosting frameworks (e.g. XGBoost, LightGBM)

References

2018. Predicting Wine Ratings Using LightGBM + Text2Vec [Blog post]. Kaggle. https://www.kaggle.com/nnnnick/predicting-wine-ratings-using-lightgbm-text2vec

2019. WineEnthusiast. https://www.winemag.com/?s=&drink_type=wine&page=0

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press.

Nabi, J. (2018). Machine Learning - Text Processing [Blog post]. *Towards Data Science.* https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958

Thoutt, Z. (2017). Wine Reviews. Kaggle. https://www.kaggle.com/zynicide/wine-reviews

Vector images from https://publicdomainvectors.org/