Predicting Wine Ratings

Aviel Stern
Data Scientist

Diploma Certificate, Data Science
BrainStation

Master of Science, Geoscience
University of Massachusetts, Amherst

Bachelor of Science, Earth Science
University of California, Santa Cruz

Project Duration
3 weeks

Tools Used

NLTK jupyter python scikit learn +ableau

Why predict wine ratings?

- Predicting how a wine will perform can benefit consumers and sellers of wine.
- A winemaker can gain insights into why its Pinot Noir is rated higher than its Chardonnay.
- Wine distributers can make better decisions about pricing and purchasing by gaining insights into whether an expensive wine will be profitable.

THE Wine SHOP

Fairmont ROYAL YORK

LET'S GET LCBO TOGETHER

Trius WINERY

MID FIELD WINE BAR & TAVERN

SIP WINE BAR

chez nous wine bar

cibo WINE BAR

Table of Contents

Data Collection

1) Data is taken from **Kaggle** website.

2) The Kaggle user scraped the wine data from a United States wine review magazine website: **www.winemag.com**

3) The data is an accumulation of information about wines across the world that were reviewed, scored, and sold in the United States.

4) This dataset specifically provides consumer insights for the US market.

Data Description

1) Data contains **129,971 rows**

2) **Text Data**:

- Country (e.g. United States)
- Province (e.g. California)
- Wine Title (e.g. Nicosia 2013 Vulkà Bianco)
- Varietal (e.g. Pinot Noir, Pinot Grigio, Merlot)
- Wine Description (e.g. This is ripe and fruity, a wine that is smooth)

3) **Numeric Data**:

- Price of wine in the United States ($ USD)
- Wine rating between 80 – 100

4) **Missing values** in columns: Country, Province, and Price

Deriving meaning from the dataset

Before producing actionable insights from the dataset, the data first needs to be processed through data cleaning, mining, and matrix transformation.

Account for missing data

Replacing missing values with the median of that column ensures that a significant amount of rows will not be deleted. This is important because removing rows reduces the sample size which represents all wines sold in the United States.

Transform text data into numeric data

Transforming text data into numeric data is a technique that provides the capability for text data to be manipulated, processed, and analyzed.

Data Cleaning and Mining

Handling Missing Values

- Column ‘Country’: Filled NaNs with most frequent country in that column.
- Column ‘Province’: Fill NaNs with most frequent province in that column.
- Column ‘Price’: Filled NaNs with Price median

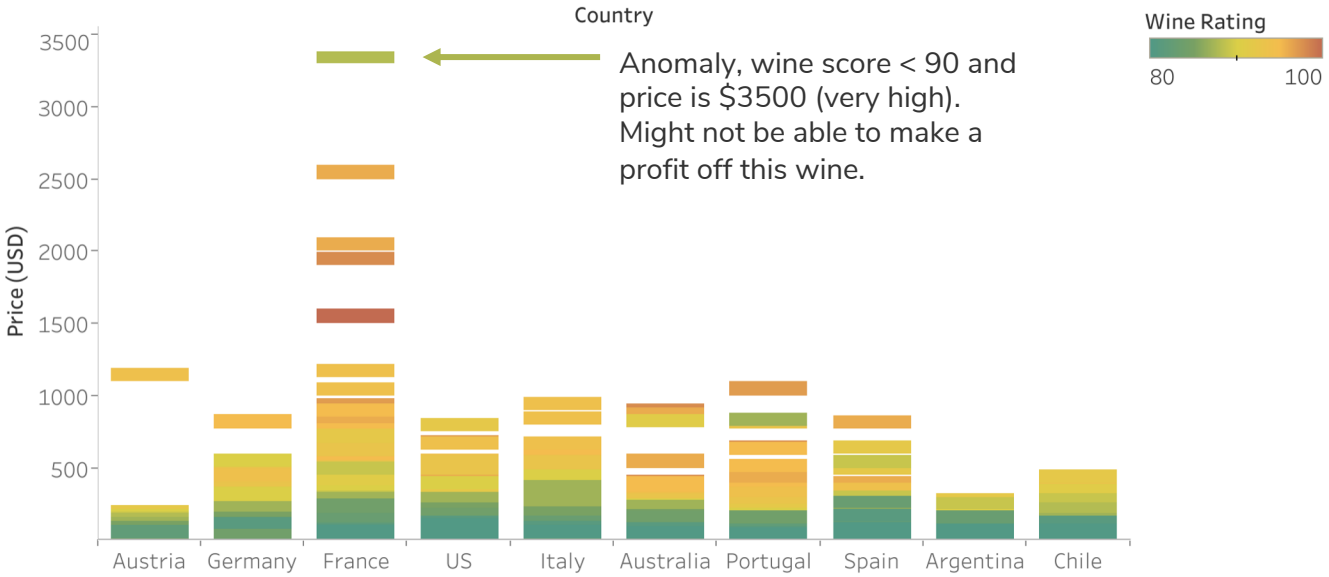
Feature Engineering

- Use of Regex to extract year from the wine title
- Add additional feature column with year of the wine.

Example of wine title: Producteurs Plaimont ***2008*** Château de Crouseilles Tannat-Cabernet Franc (Madiran)

Exploratory Data Analysis

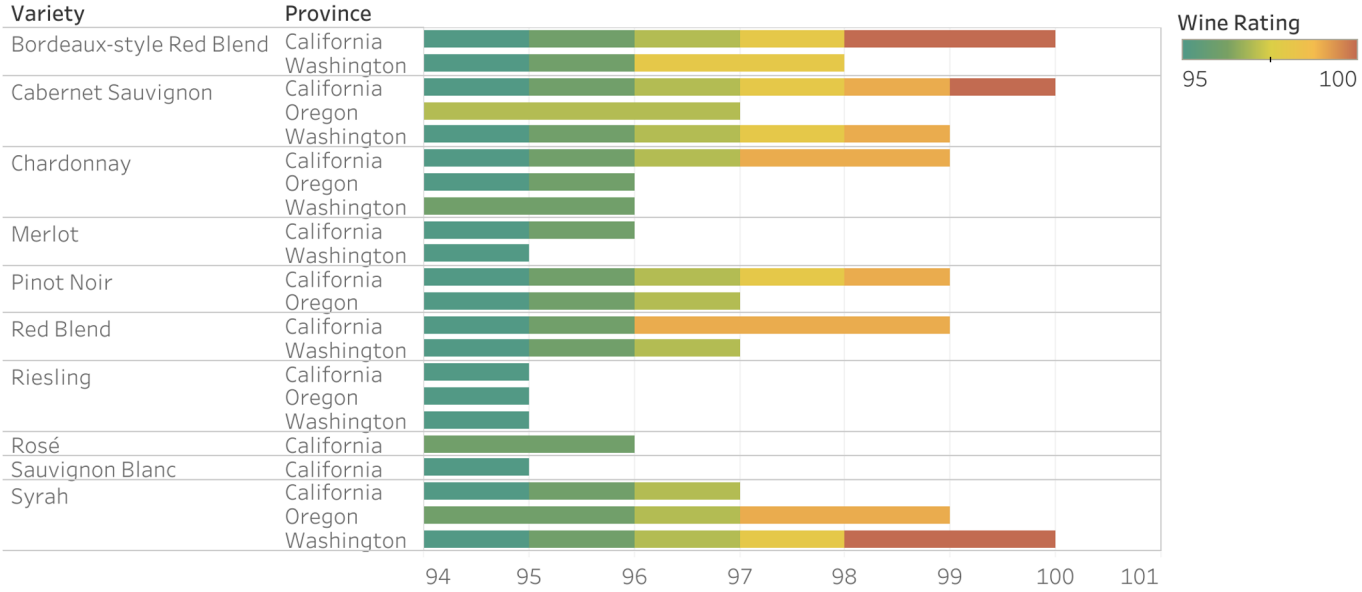
Wine Price in the United States vs. Top 10 Countries, associated with Wine Rating



- Price in the United States is positively correlated to wine rating
  - Wine rating < 80 → Lower Price Wines
  - Wine rating > 90 → Higher Price wines
- French wines are more expensive in the United States but are not rated significantly higher than wines from Italy, US, and Australia.

Exploratory Data Analysis

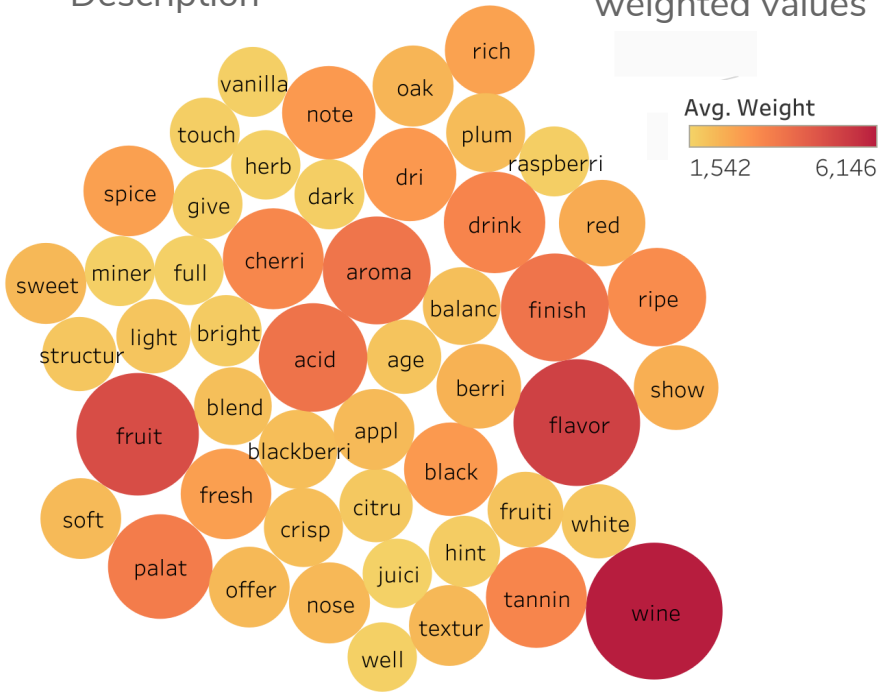
Top performing wines from California, Oregon, and Washington



- Red Wines perform very well from California, Oregon and Washington.
- White wines do not perform as well in the United States
- Suggestion for selecting outstanding wines from these regions:
  - California: Cabernet Sauvignon, Bordeaux blend/Pinot Noir, Red Blend
  - Washington: Syrah and Cabernet Sauvignon
  - Oregon: Syrah

Top weighted words from wine description

TF-IDF on Wine Description → Transforming words into weighted values



Example wine description:

"Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering ripened apple, citrus and dried sage alongside brisk acidity."

Natural Language Processing for Business Results

NLP Methods:

Transform text data into numeric matrix in order to process data and run predictive models.

Wine Descriptions:

- TF-IDF
- Stemming to obtain root word
- Apply lowercase to all text

Country, Province, and Varietal:

- One-hot-encoding

Business Result: This provides guidance for specific words in tasting notes to support sales.

Data Analyst Summary

- Pricing of wines in the United States are correlated to wine score.
- Red Wines from California, Oregon, and Washington perform the best in the United States.
- Use of TF-IDF to identify significant words used to describe a wine.

Suggestions

- In general, a higher price wine will be well received if the wine score is high and will likely make a profit.
- If purchasing wines from the US for the purpose of distribution, the analysis suggests that wines will perform well when Pinot Noir and Red blends are from California, Syrah's from Oregon, and Cabernet Sauvignon from Washington or California.
- Provides guidance for specific words in tasting notes to support sales. Tasting notes describes specific properties in a wine such as smell, taste, body, acidity, color, and tannin.

Predictive algorithms and additional insights

06

Classification Methods

Label wine ratings for classification ML models

07

Machine Learning

Accuracy scores for each predictive model

08

Key Insights

Findings from predictive models

Classification Methods

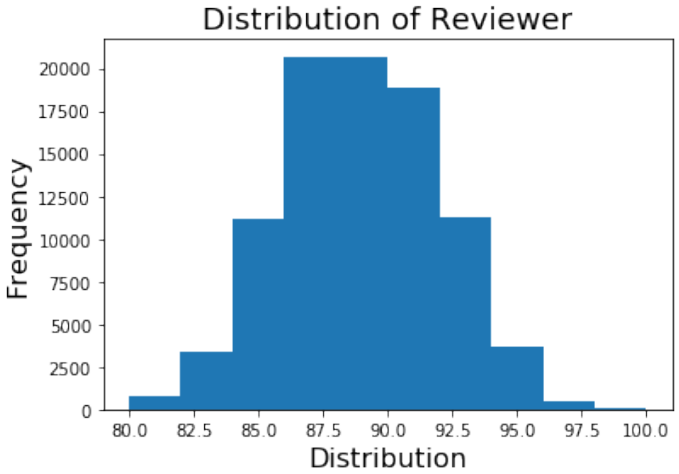
Wine ratings range between 80 – 100 from www.winemag.com

Three potential approaches for predicting wine ratings

For this project, I use Approach 3.

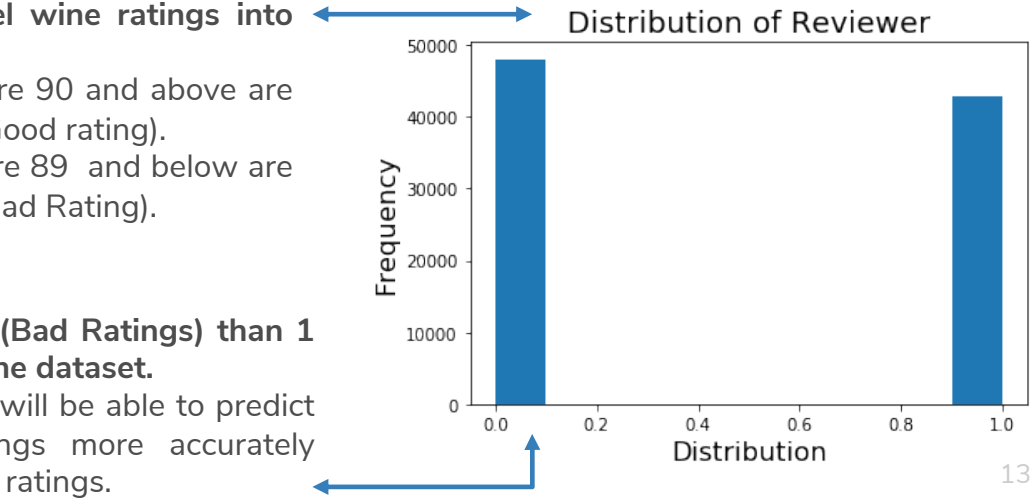
Approach 1: Use regression model and treat the wine scores as a continuous quantity.

Approach 2: Classification model to predict the wine reviews and treat the wine reviews as 20 classifiers.



Approach 3: Label wine ratings into two categories.

- Ratings that are 90 and above are labeled as 1 (Good rating).
- Ratings that are 89 and below are labeled as 0 (Bad Rating).



There are more 0 (Bad Ratings) than 1 (Good Ratings) in the dataset.

- Predictive model will be able to predict ”bad” wine ratings more accurately than “good” wine ratings.

Machine Learning

Methods for improving model accuracy:

1) Remove features. Possibly start with less features and then add more features slowly to determine best model outcome.

2) Remove bias data. There are significantly more wine reviews from the United States than the rest of the world in this dataset. Therefore predicting wine scores in the United States will be more accurate than other regions.

How well did the classification models fit?

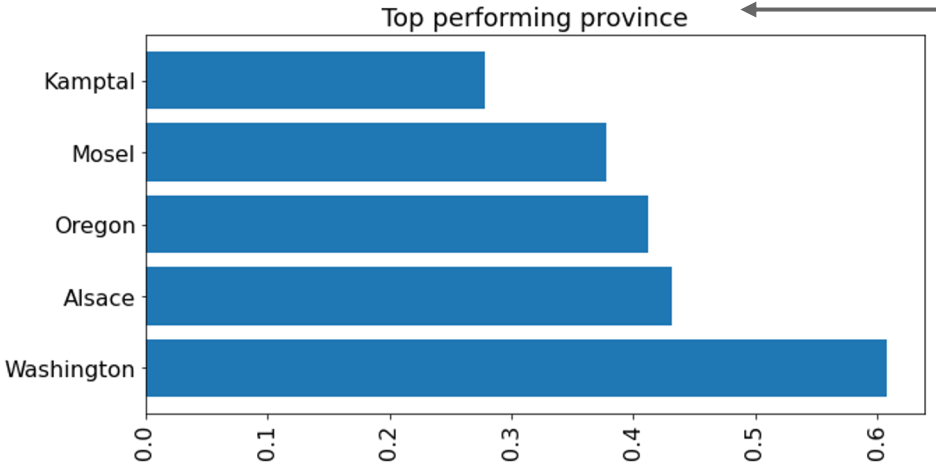
Logistic Regression performs slightly better than the rest of the three models.

	Accuracy Score	Precision	Recall
Logistic Regression	**82.3%**	**82.1%**	**79.8%**
Random Forest	**81.8%**	**81.4%**	**79.4%**
XGBoost	**79.2%**	**78.4%**	**76.9%**

The precision and recall scores are close to the accuracy score, suggesting that the accuracy for predicting a wine score as 1 is nearly the same as predicting a wine score as 0.

Key Insights

Which model features influence wine scores?

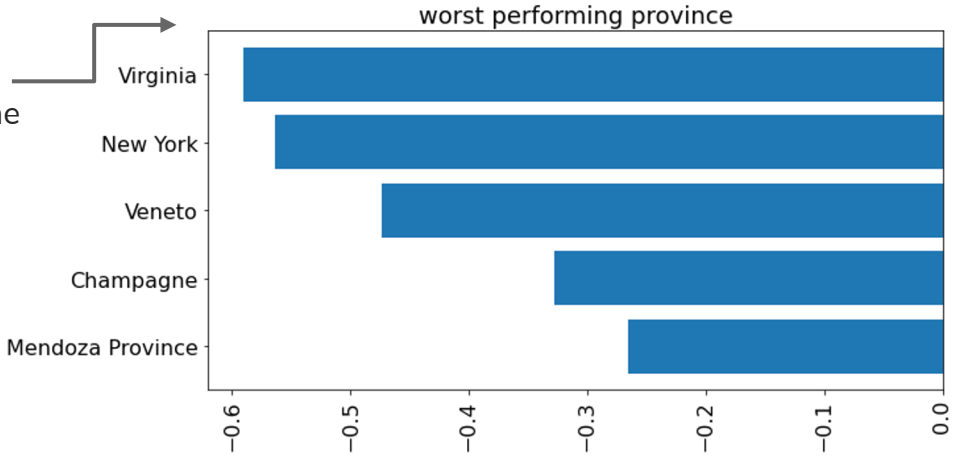


Positive coefficients positively influence the model.

Wines from Washington, Oregon, Alsace (France), Mosel (Germany), and Kamptal (Austria) will score high in the US.

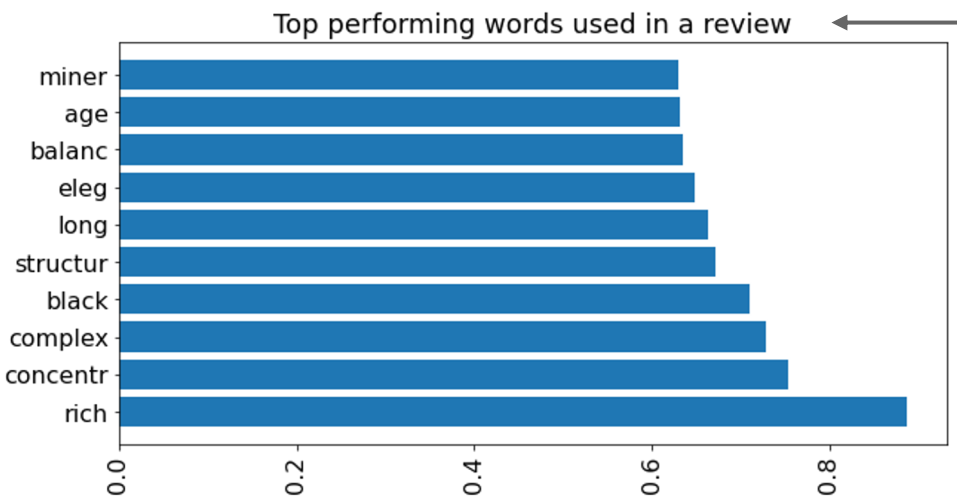
Negative coefficients negatively influence the model.

Wines from Virginia, New York, Veneto (Italy), Champagne (France), Mendoza (Argentina) will score low in the US.



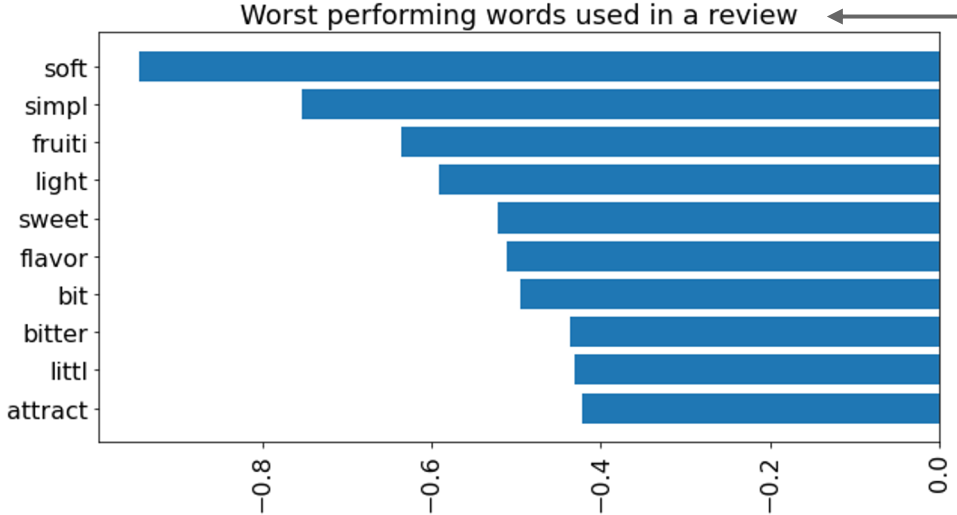
Key Insights

Which model features influence wine scores?



Positive coefficients positively influence the model.

Wines perform well in the US when described as aged, balance, rich, complex, concentrate, black, structure, long, elegant



Negative coefficients negatively influence the model.

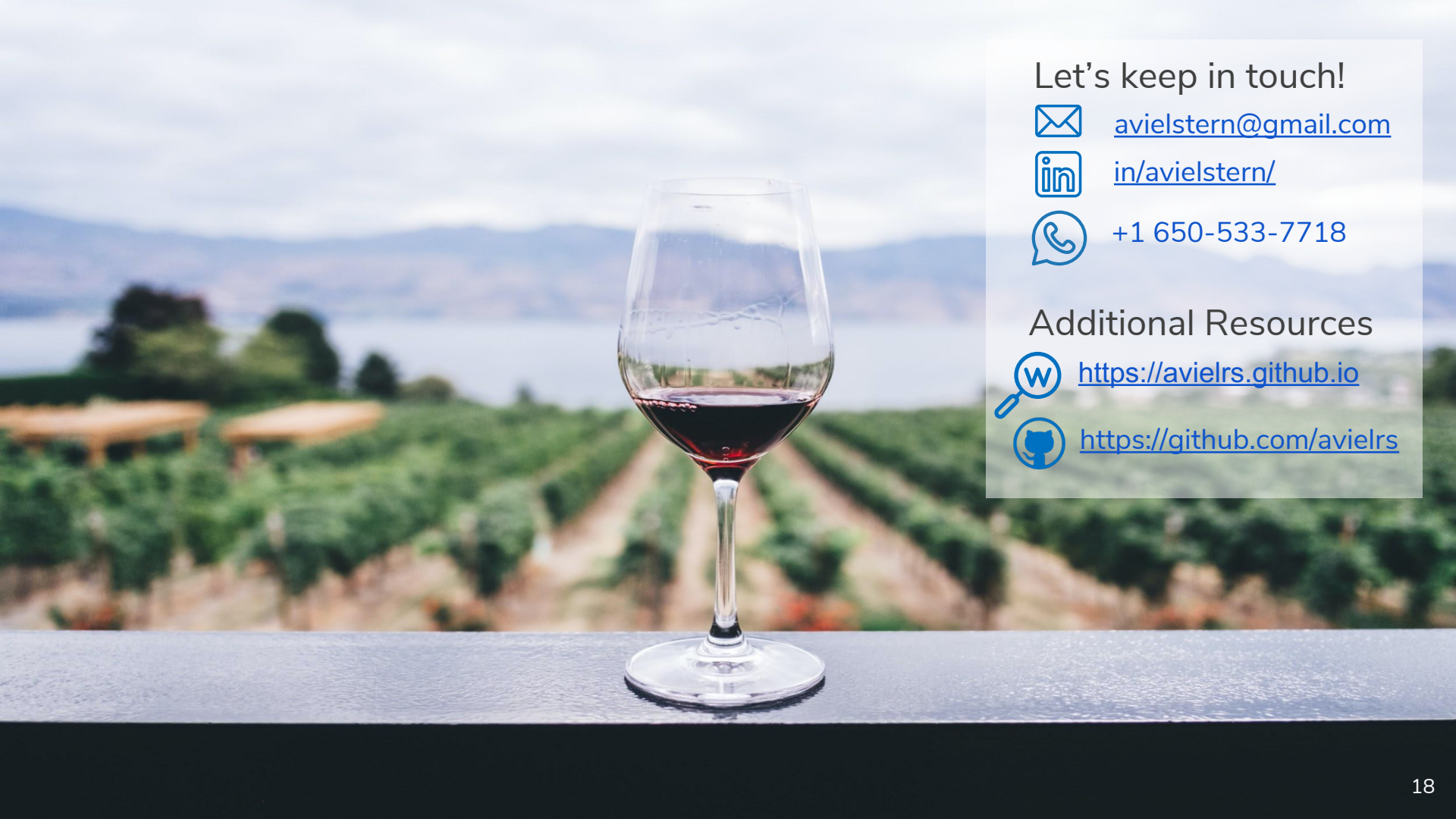
Wines do not perform well in the US when described as soft, simple, fruity, sweet, light, bitter, little

Key Insights Summary

- Across the world, wines from Washington, Oregon, Alsace (France), Mosel (Germany), and Kamptal (Austria) will score high.
- Wines perform well when described aged, balance, rich, complex, concentrate, black, structure, long, elegant
- Bias is present in data because there are significantly more data on wines from the United states than wines from other countries.

Suggestions

- Distributing wines from these locations are likely to result in good sales in the United States.
- Identifying these specific words within the wine description and wine notes will indicate a wine that is more likely perform well in the United States.
- While these suggestions demonstrate the potential impact of this project, more data needs to be collected in order to remove bias and to make confident decisions.



Let's keep in touch!

avielstern@gmail.com

in/avielstern/

+1 650-533-7718

Additional Resources

https://avielrs.github.io

https://github.com/avielrs