California versus France: The Search for Value in Wine using Data Analytics

Team Members: Zach Elson, Katherine Sullivan, Martin Wehrli

Project Goals: Determine if Californian wine represents a better value than French wine using price data and wine ratings made available by Wine Enthusiast magazine.

Specifically, compare price versus rating for both Chardonnay and Cabernet Sauvignon / Bordeaux Red Wine Blends.

Answer the following questions:

- Is there a clear value winner?
- Does rating level change the answer?
- Is it possible to predict wine prices from these reviews?
- What is the impact of region and subregion on pricing and rating?
- From a wine review perspective, what words are more likely to be used as rating increase?

Data Source: Data for this project came from Wine Enthusiast Magazine (www.winemag.com). The wine data was downloaded from Kaggle.com in the form of 16 separate CSVs. The data in Kaggle was originally sourced from the Wine Enthusiast.

- The entire dataset includes approximately 250,000 wine reviews.
- The subset used is French and Californian Cabernet and Chardonnay.

		France	US
Chardonnays		2795	10893
Cabernet / Bordeaux		3991	14065

Data Analytics Tools Used:

- Machine Learning
  - Scikit Learn, Tensorflow: Keras, DictVectorizer, OneHotEncoding, LabelEncoding, EarlyStopping, GridCV, KNN, Matplotlib, SQLalchemy
- Python - All machine learning written in Python in Jupyter NBs
- Pandas
- Postgres - All data maintained in Postgres database; migrated from CSVs.
- Tableau

Project Challenges:

Machine Learning – Predict price of French and US wines by "rating", "region" and "subregion" (type, country, region, subregion, subsubregion, varietal, alcohol %, vintage, winery)

- Discovered "alcohol" and "vintage" not significant indicators of price; "winery" difficult as data points per winery are small.
- Focused on rating and geographic variables: "rating", "type", "country", "region", "subregion", "subsubregion".
- Looked at classification (with 4 price buckets) and regression; focused on regression.
- Also experimented with predicting rating from price (another day…)

Data Integrity – Extensive data cleaning required

- Missing data (price, region, subregion, subsubregion)
- Erroneous Data (vintage)
- Inconsistent Data (region, subregion, subsubregion)
- Ratings: Wine ratings and reviews are subjective in nature and performed by different professionals not readily identifiable.
- Price: not adjusted for inflation or currency exchange fluctuations over time; may fluctuate greatly by location as well; variability due to industry structure / makeup / distribution, etc.
- Data Collection / Input: methods inconsistent; varying nomenclature and requirements for fields.
- Industry-related: Origin of French wines are typically associated with villages; US with county/region.

Python:

- Code for Wine Descriptor word search code became an iterative exercise in identifying 'Stop' words that were not relevant in describing wine.
- Running regressions takes a lot of time.

Conclusions:

Is there a clear price / value winner? No.

Does rating level change the answer? Yes.

French wines appear to be a better value for wines rated under 95 points. Over 95 points, there is a significant increase in price for rare and exclusive French Chardonnay (aka white Burgundy) and Bordeaux red wine blends.

Does region or sub-region matter? Absolutely.

In California, Napa is by far the most expensive region. Within Napa, there is also a wide disparity in subregions. Oakville is 4x more expensive for a 5 point increase in rating.

French wine prices and ratings, especially red wines from Bordeaux, vary ENORMOUSLY village to village (subregion to subregion). US Chardonnays vary much less in rating and price.

From a wine review perspective, what words are more likely to be used as rating increase?

A large majority of wine descriptors are the same within Chardonnay and Cabernet reviews. The differences by rating:

- 95-100 point wines often include descriptors such as ‘beautiful’, ‘classic’, ‘magnificent’
- 80 – 84 point wines often include “thin”, “fruity”, “weak”, “simple”

Is it possible to predict wine prices from these reviews? Kinda.

Machine learning yielded obvious relationships between rating and price as well as geographic locations. Predictions are more accurate with Chardonnays than with Cabernet/Bordeauxs. The latter having a greater range in its “scatter” price/rating values.