MY VIVINO

Prepared by Nikita Gaidamachenko and Andrew Bonham for Qwasar Technologies

CONTENTS

01. INTRO

MyVivino's CEO approves revision of wine recommendation tool

Data science and machine learning are tools of the new user search feature

Wine searches allow for increased specification tailored to the public and connoisseurs

GOALS

- Provide users with a comprehensive search tool based on:
    - Meal pairings
    - Note selection
    - Traditional factors (wine type, country, winery, etc.)
- Sort the selected wines in terms of price *and* rating

02. Methods

- HTML for Each Wine
- Building DataFrames From HTML
- Q-Rating
- Comprehensive Search Tool

HTML for Each Wine

01. Libraries Used

- Asyncio
- Aiohttp
- Chromedriver
- Selenium
- BeautifulSoup

02. Source of Data

- Wineenthusiast.com
- ~1000 wines
- Includes type, price, rating, country, etc.

03. Asynchronous approach

- Improved speed of data retrieval
- Reduced time from hours to ~25 min.

04. Selenium

- Webdriver class used to render webpage java

05. BeautifulSoup

- Content accessed via BeautifulSoup

06. Conclusions

- Content converted to string for parsing

Building DataFrames From HTML

Library used: Pandas

- HTML string was parsed to generate dictionary
- All wine data was compiled into a single dictionary
- Dictionary was converted to a DataFrame

- DataFrame was saved to CSV
- CSV was examined and edited for minor errors
- CSV was converted to SQL DB for storage
- Final DataFrame was generated from DB for analysis

- Motivation:
    - Rank wines on rating *and* price
    - Recommend low-priced and highly-rated wines first
    - Recommend high-priced and low-rated wines last
- Model used:

$$q = c\frac{r^a}{p^b}$$

- *r* is the rating, *p* is the price, and *a, b,* and *c* are positive constants.

Q-RATING

- Carry out regression on a linearized version of *q*:

$$q' \equiv \log q = c' + ax + b'y$$

$$c' = \log c, \quad x = \log r, \quad b' = -b, \quad y = \log p$$

- Model is “trained” with asserted values of q
    - $q(p_{min}, r_{max}) = 100$
    - $q(p_{min}, r_{min}) = q(p_{max}, r_{max}) = 30$
    - $q(p_{max}, r_{min}) = 10$
- Asserted values can be changed at developer’s discretion
- Search results are ranked by *q*-rating

COMPREHENSIVE SEARCH TOOL

- Wine-food pairing
  - User input is matched to a dictionary containing pairs
  - The data is then filtered based on the pairs
- General wine search
  - Pandas was used to traverse the dataset based on user input
  - The filtered wines are then passed to the notes search
- Notes search
  - Vectorized unigrams and bigrams were computed and shown
  - User selects from the top notes of available wines
  - The final filter based on notes is conducted
- Results are sorted by q-ranking and printed

03.

ANALYSIS

General Statistic, Q-Rating
and Comprehensive Search

ASSUMPTIONS

Generic Picks

General public is looking for a bottle of wine to either accompany their meal or to present it as a gift

Q-Rating

Q-rating helps to find a perfect gift in their price category or to find a bottle to balance the dish.

Budget

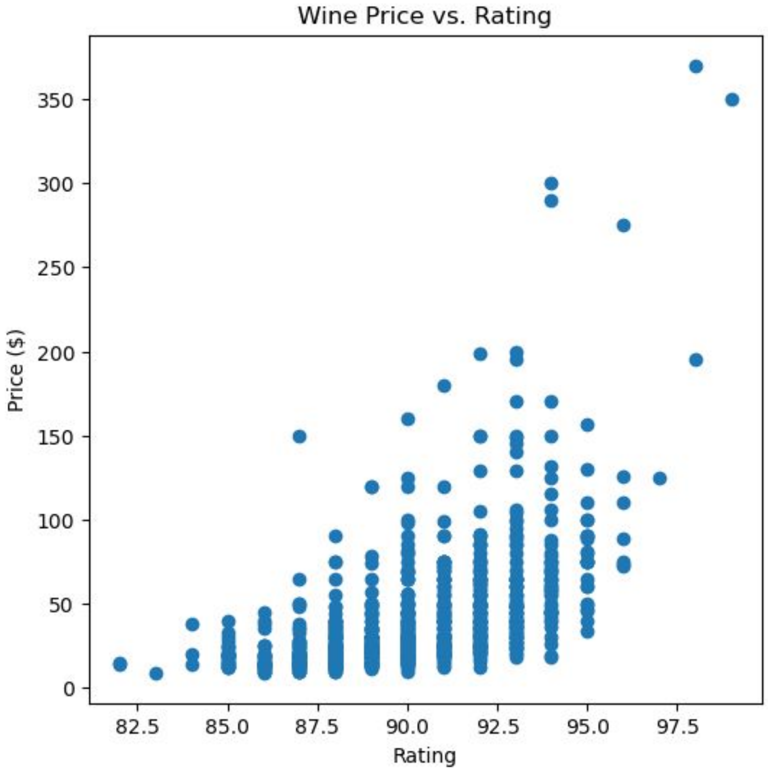
Main wine picking criteria for the general public is their budget and personal preferences

Preferences

More experienced users want to find wine with specific notes

PRICE VS RATING

- Price and rating follow a general trend, as expected
- Majority of outliers are those of high price or rating
- Average price: $38.83
- Average rating: 90.0
- Price and rating correlation coefficient: 0.56



WINES BY TYPE

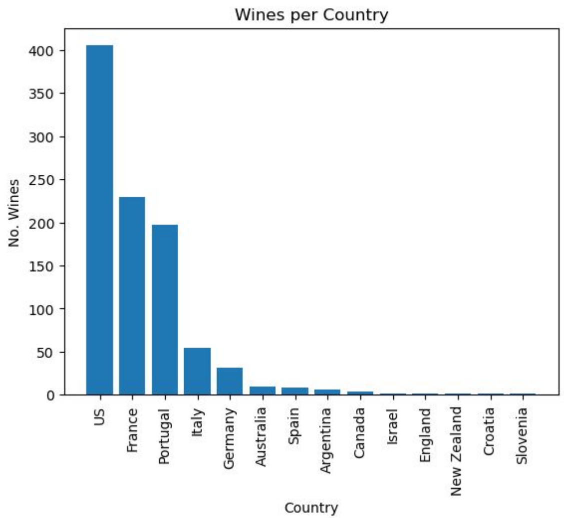
Over 80% of the database consisted of red and white wines, which were moderately priced compared to the rest of the wine types.

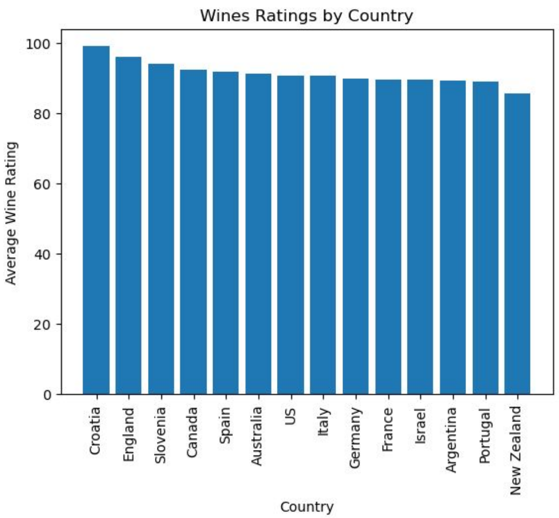
Fortified wines were the most expensive, and wine types had average ratings ranging from 87 to 91.8.

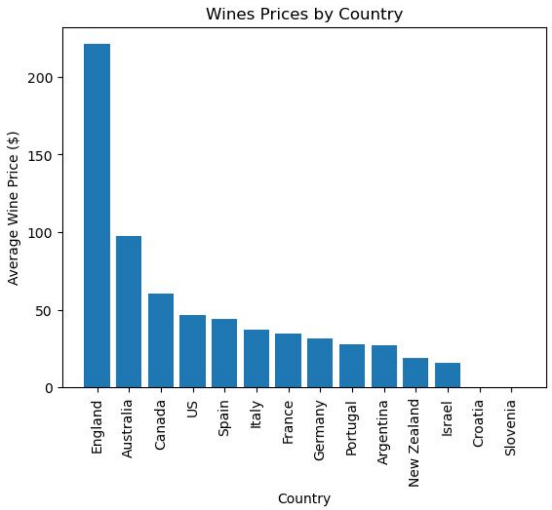
Wine Type	No. of Wines	Avg. Price ($)	Avg. Rating
Red	514	42.06	90.3
White	325	29.88	89.5
Sparkling	71	70.49	91.2
Rose	26	16.25	87
Port/Sherry	7	38.29	89.1
Dessert	5	28	90.4
Fortified	5	149.6	91.8

WINE BY COUNTRY

- The US, France, and Portugal were the top three countries for the number of wines in the dataset
- Croatia, England, and Slovenia had the 3 highest average ratings, although Croatia and Slovenia lacked price data.
- England had by far the most expensive wines on average, although these consisted of only two wines priced at $72 and $370.







COMPREHENSIVE SEARCH

Comprehensive Search Tool is a three feature wine search & recommendation model that utilizes principles of statistics, machine learning and language processing, specifics of each grape variety and type, and meal pairing guidelines to lead the user through the wine selection process and recommend suitable wines. Three features allow for users of all levels to navigate to their next favourite bottle.

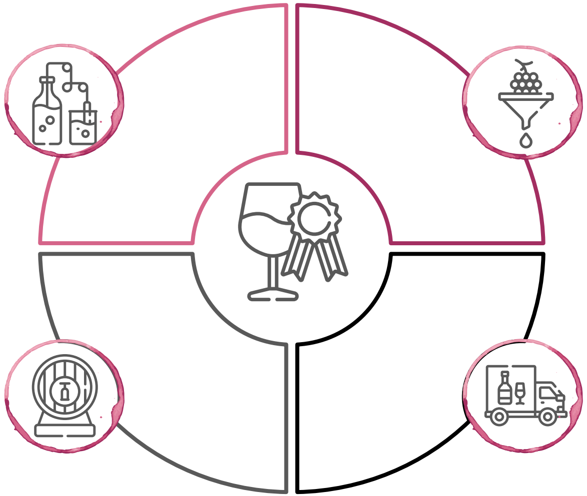
Once suitable wines are found, the Q-rating is used to prioritize highly rated wines for lower prices.

Q-RATING

Price to rating score

NOTES-BASED SEARCH

For true wine connoisseurs



FOOD PAIRING

Find the right bottle for your favourite dish

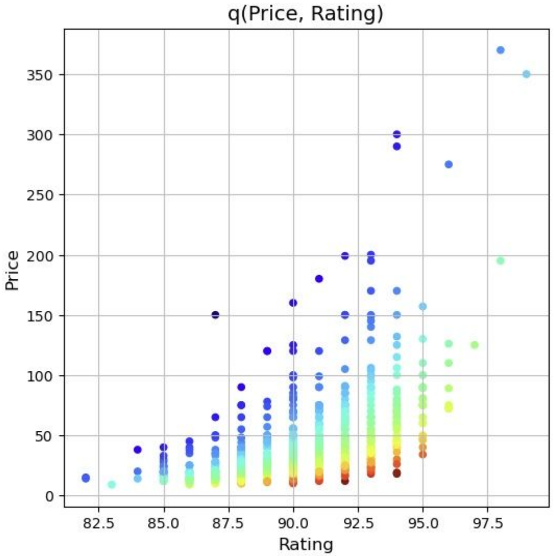
GENERAL SEARCH

Filter the results by price, rating, variety and other criteria

Q-RATING

Q-rating considers both price and rating and helps to identify niche or outlier wines that are priced significantly higher or lower than similarly-rated wines. Wines priced lower than similarly-rated wines are recommended first.

The chart below illustrates the relationship: better scoring wines are marked with warmer colours.



Q-Rating Biases

Age of Bottles

More mature bottles of certain styles may cost well over $100 and even $1000, while still being regarded as one of the best.

Size of Bottles

Some wine styles and houses offer bottles of multiple sizes, which will impact the price but not the rating.

What is perfectly normal size for a certain style, may be seen as an outlier in other groups.

Style of Wine and Variety

Cheaper wine of a different variety, but similar rating will perform better in the Q-rating analysis.
Example: Barolo that typically costs around $100.

Different wines of the same variety when one style is cheaper than another, but ratings are similar.
Example: Chianti and Brunello di Montalcino, both of which are made from Sangiovese grapes

GENERAL WINE SEARCH

It is designed to help the user narrow down the results by indicating desired parameters of the following criteria:

- Rating range
- Price range
- Producer
- Region
- Country
- Grape variety
- Wine type

This tool allows the user to set those preferences before diving into specific notes that more experienced users might be seeking.

The user can leave fields empty to broaden the search.

FOOD AND WINE PAIRING

The user starts by picking one of the 17 food categories:

- Appetizers
- Salads
- Vegetarian/Vegan
- Rich Fish
- Lean Fish
- Shellfish (including oysters)
- Poultry
- Red Meat
- Game Meat
- Pasta - Red Sauce
- Pasta - Other
- Pizza
- Spicy Foods
- Hard Cheeses
- Soft Cheeses
- Desserts - Sweet
- Desserts - Fruity

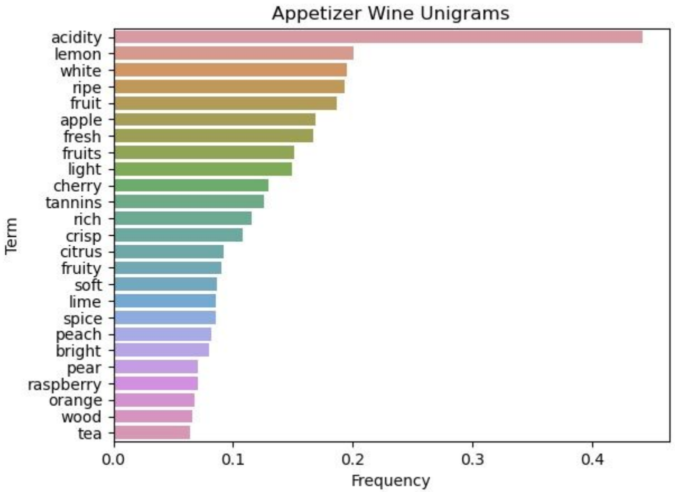
Each food category is paired with compatible grape varieties and wine types, while adhering to wine and meal pairing guidelines thus narrowing the search.

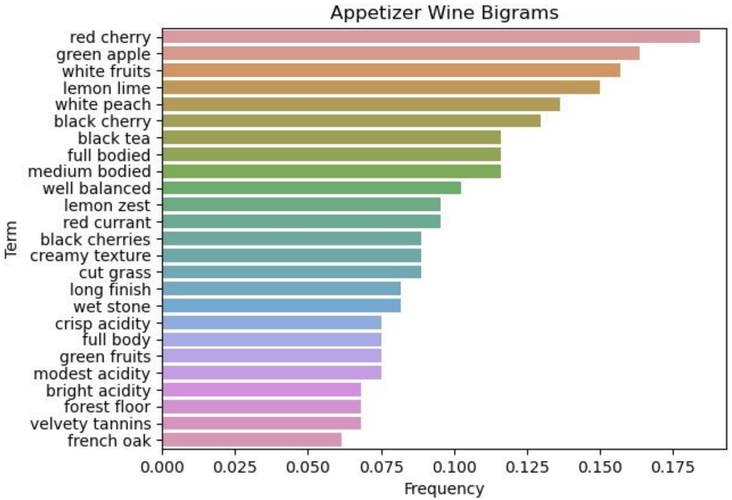
As some varieties may occur in multiple types of wine (for example, Pinot Noir is typically used in sparkling wines in addition to Red), both grape variety and type are specified there to isolate suitable bottles.

NOTES AND PAIRING BIASES

As there are many ways to describe the same note, some bigrams for common descriptors are spread over several synonym terms and therefore have less representation. For example, we see that acidity is the most mentioned unigram term, while it is underrepresented in bigrams.

Words that originate from the same stem but either represent a different part of the speech (noun/adjective - tannins vs tannic) or are multiple nouns (tannin vs tannins). Stemming and Lemming was attempted, but was abandoned as it incorrectly identified different parts of the speech as unique words rather than adjectives of the same stem.

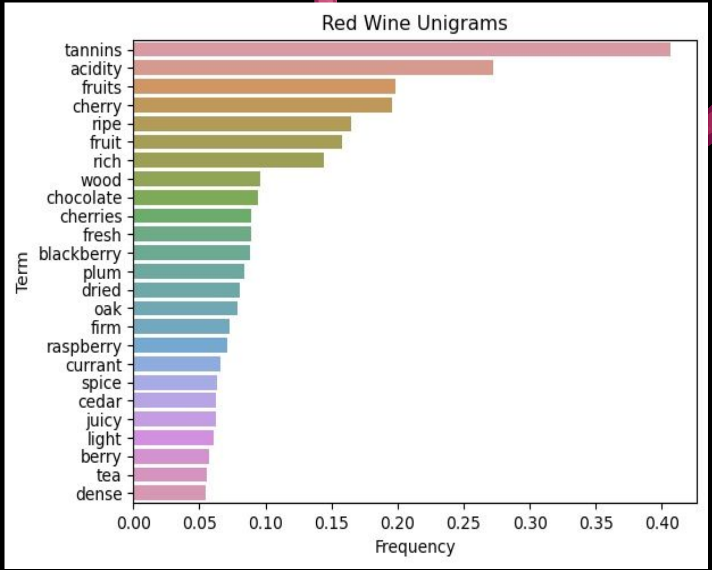




Notes Search

Users can further dive into wine notes and input desired descriptors that a bottle should have. This helps to find the specific bottle that one is seeking.

If a grape variety or a type are chosen, including those in the meal pairings section, a list of common notes for that variety or types will be shown to the user



FUTURE IMPROVEMENTS

- **Search adjustment**
  - If the user wants to change the search, he or she has to restart the search
  - Updates would require the original dataframe to be preserved
  - A UI with buttons would be the more user-friendly option
- **Unigram/bigram stemming and grouping**
  - Different forms of similar words lead to some repeat results
  - Stemming/lemming was abandoned as it produced incorrect results
  - A more complex language model could be developed
- **User interface**
  - Current search required text from user input
  - A UI could make the search more visually appealing or intuitive
  - Buttons in the UI could make the search more user friendly and easy to modify

04. CONCLUSION

- Database of ~1000 wines developed via web-scraping
- Model prioritizing both price *and* rating developed
- Comprehensive search tool allows for
  - Food-wine pairing
  - General searches for price, country, wine type, etc.
  - Note selection
- Business value added by comprehensive nature of search

THANKS

Do you have any questions?

Reach out to Nikita Gaidamachenko or Andrew Bonham in Discord

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution