Vintage Intelligence: Using Machine Learning to Recommend Your Next Favorite Wine

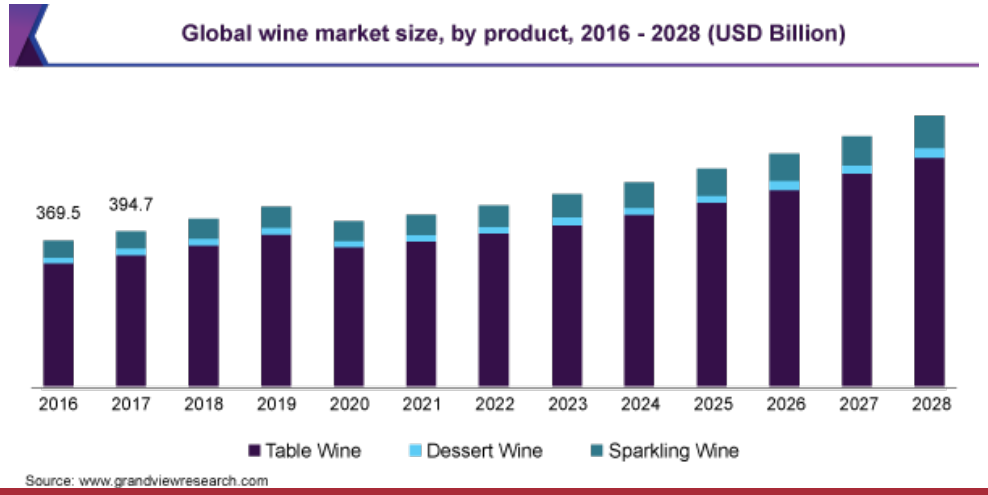
Presented by:
Nicholas Khoo

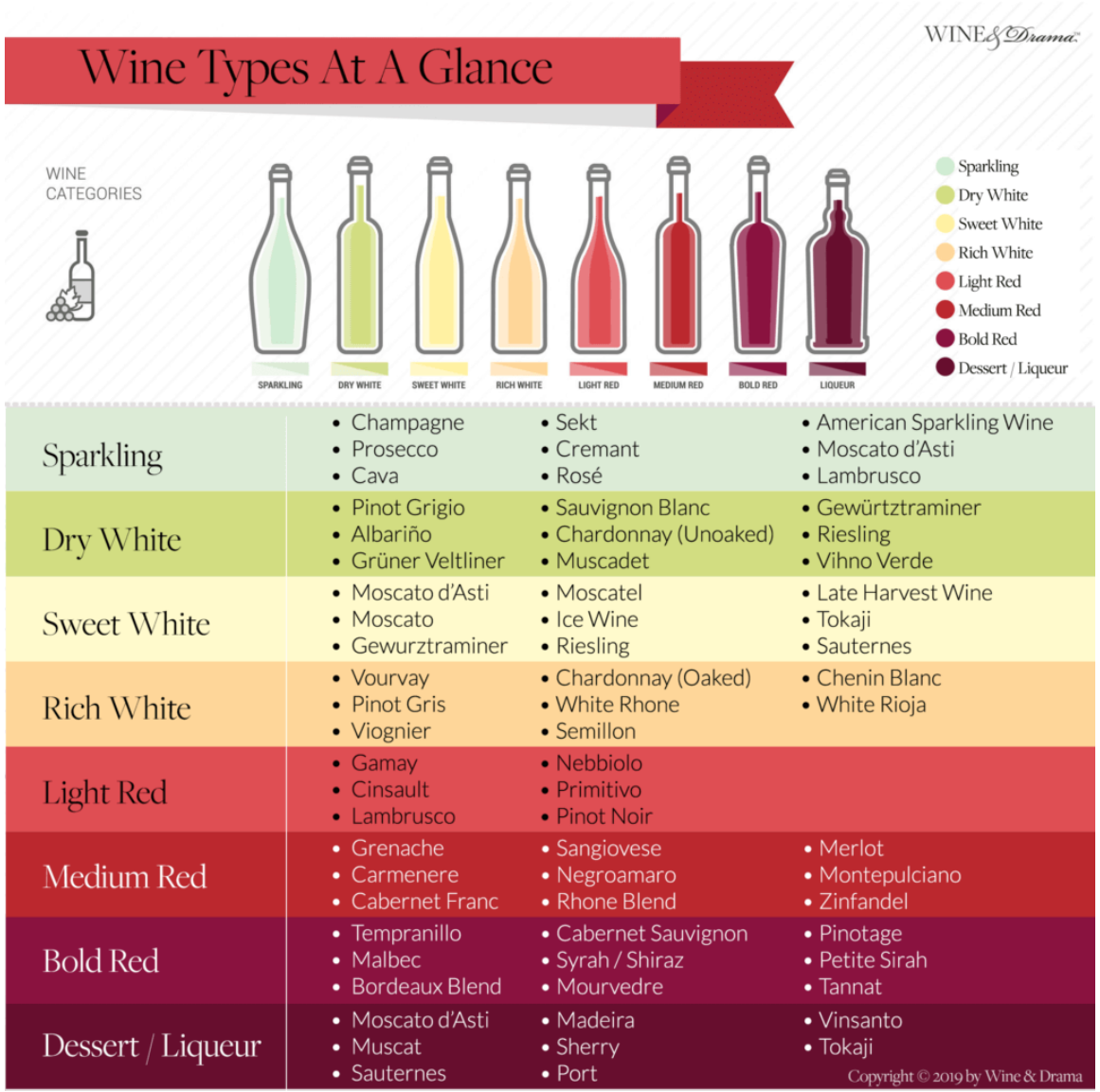
Presentation Outline

- **Background**
- Data Science Approach

Background

- The wine industry is a multi-billion dollar sector that is rapidly growing.
- Choosing the right wine can be difficult for consumers due to the overwhelming selection and subjective nature of wine tasting.
- Wine recommendation systems that use machine learning are effective in providing personalized recommendations based on various data points, increasing customer satisfaction and loyalty.





Source: Wine & Drama

Presentation Outline

- Background
- Data Science Approach

Data Science Approach

Problem Statement | Data Collection | Data Cleaning & Exploratory Data Analysis | Data Pre-processing and Machine Learning Modelling | Conclusion & Recommendation

Problem Statement

Situation

Many consumers struggle to select the right wine due to the overwhelming selection and subjective nature of wine tasting.

Complication

This creates a challenge for wineries and wine sellers to provide personalized recommendations and can lead to decreased customer satisfaction and loyalty.

Resolution

Developing a wine recommendation system using machine learning algorithms that can provide accurate and personalized wine recommendations to customers, can ultimately increase customer satisfaction and loyalty for wineries and wine sellers.

Data Science Approach

Problem Statement | **Data Collection** | Data Cleaning & Exploratory Data Analysis | **Data Pre-processing and Machine Learning Modelling** | Conclusion & Recommendation

Data Collection

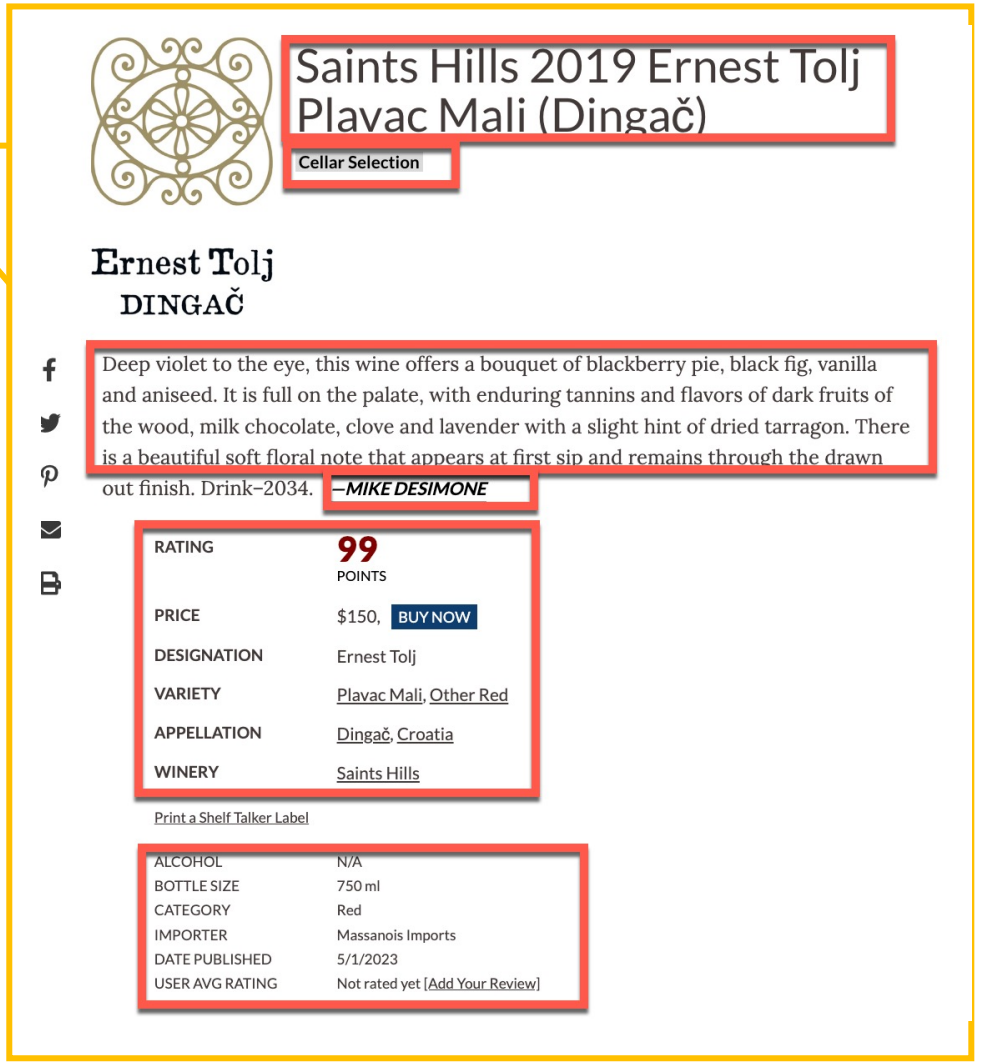
Initial Approach

Wine data was initially scrapped from Wine Enthusiast – a lifestyle magazine covering wine, food, travel, and entertaining topics.

We retain the right to use certain technologies to detect the use of specialized computer programs including, but not limited to, web robots, web wanderers, crawlers, scrapers, spiders or any other techniques designed to systematically download and/or copy our Review Content. The use of any of these programs or techniques is strictly prohibited, constitutes material violations of this agreement and is grounds for immediate termination of your subscription.

Subsequent Approach

- Wine dataset from Kaggle (https://www.kaggle.com/datasets/zynicide/wine-reviews)
- Wine Descriptors from Ronald Shuring's GitHub (https://github.com/RoaldSchuring/wine_recommender)



Data Science Approach

Problem Statement → Data Collection → **Data Cleaning & Exploratory Data Analysis** → Data Pre-processing and Machine Learning Modelling → Conclusion & Recommendation

Data Cleaning

- Irrelevant columns (e.g. taster_twitter_handle) were removed
- Duplicate entries and null values were dropped
- Scores were filtered > 88

Wine Reviews Dataset

Feature	Type	Description
country	object	The country where the wine was produced.
description	object	Description of the wine's characteristics and tasting notes
designation	object	The vineyard within the winery where the grapes for the wine were sourced from.
points	int64	The number of points assigned to the wine on a scale of 1-100 by the wine reviewer.
price	float64	The price of a bottle of the wine.
province	object	The province or state within the country where the wine was produced.
region_1	object	The first-level region within the province or state where the wine was produced (e.g. Napa Valley).
region_2	object	The second-level region within the province or state where the wine was produced (e.g. Rutherford within Napa Valley).
taster_name	object	The name of the wine reviewer.
taster_Twitter_handle	object	The Twitter handle of the wine reviewer.
title	object	The title of the wine review.
variety	object	The type of grape used to produce the wine.
winery	object	The name of the winery that produced the wine.

Wine Descriptors Dataset

Feature	Type	Description
raw descriptor	object	A descriptor for a sensory attribute of the food or drink being evaluated (e.g. "sweetness", "acidity", "herbaceousness", etc.).
level_3	object	A subcategory of the descriptor that provides additional detail about the attribute being evaluated (e.g. "fruit sweetness" for the "sweetness" descriptor).
level_2	object	A broader category that groups together related descriptors and level_3 subcategories (e.g. "flavor").
level_1	object	The highest level of categorization that groups together the level_2 categories (e.g. "aroma and flavor").

Points	Description
95-100	Wines are benchmark examples or 'classic'.
90-94	Wines are 'superior' to 'exceptional'.
85-89	Wines are 'good' to 'very good'.
80-84	Wines are 'above average' to 'good'.
70-79	Wines are flawed and taste average.
60-69	Wines are flawed and not recommended but drinkable.
50-59	Wines are flawed and undrinkable.

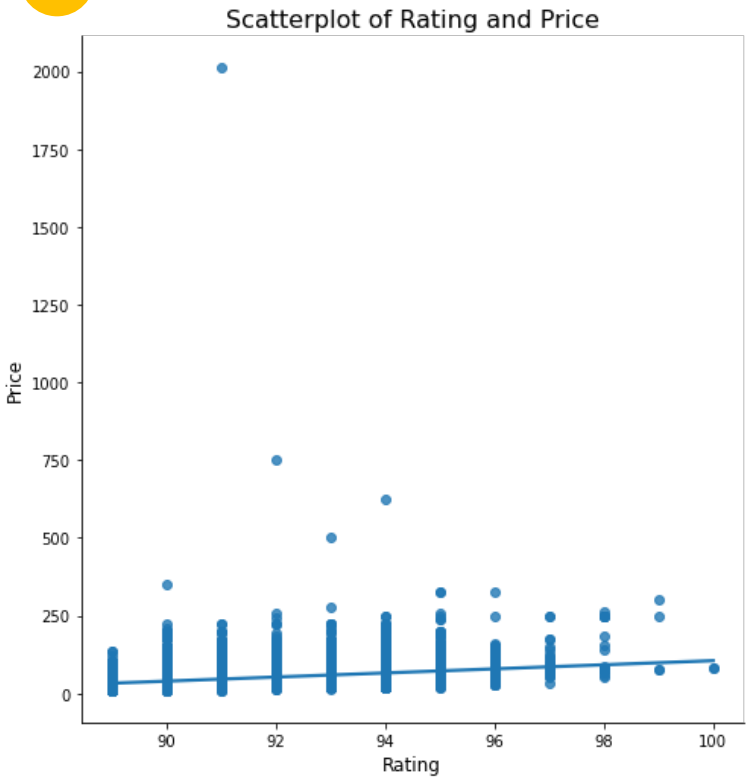
Source:
Wine Scores from Wine Searcher

Wine Prices & Ratings

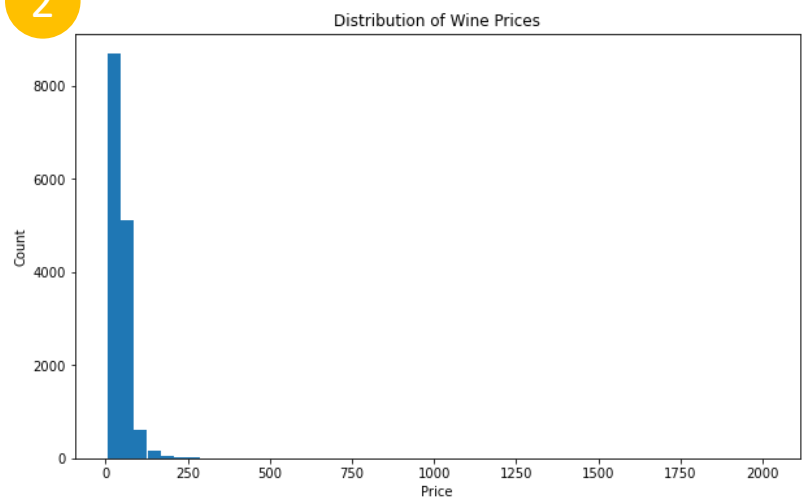
Key Findings

1. There is some relationship between "points" and "price", but it may not be a very strong one.
2. Distribution of wine prices is heavily right-skewed with majority of the wines being priced below approximately USD 100.
3. Wines with ratings of 90-92 are the most common in the dataset, while wines with ratings below 89 or above 95 are relatively rare, suggesting that wines with extreme ratings are less common and wines with ratings above 95 may be of particular interest to wine enthusiasts or collectors.

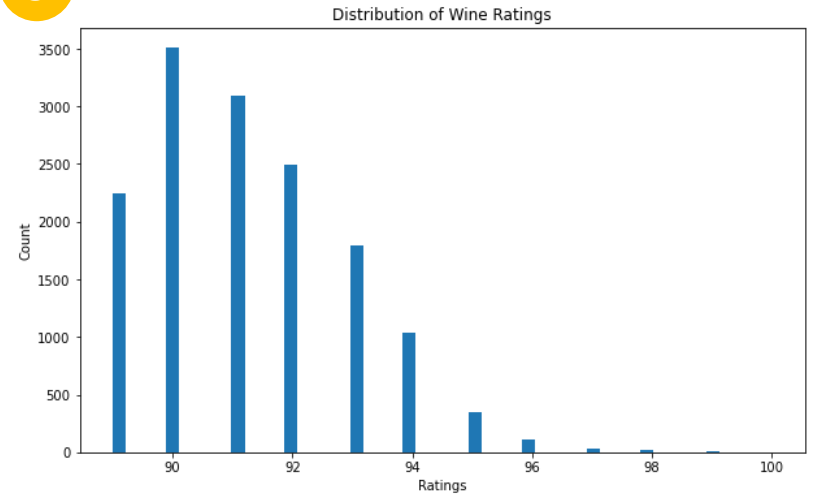
1



2



3

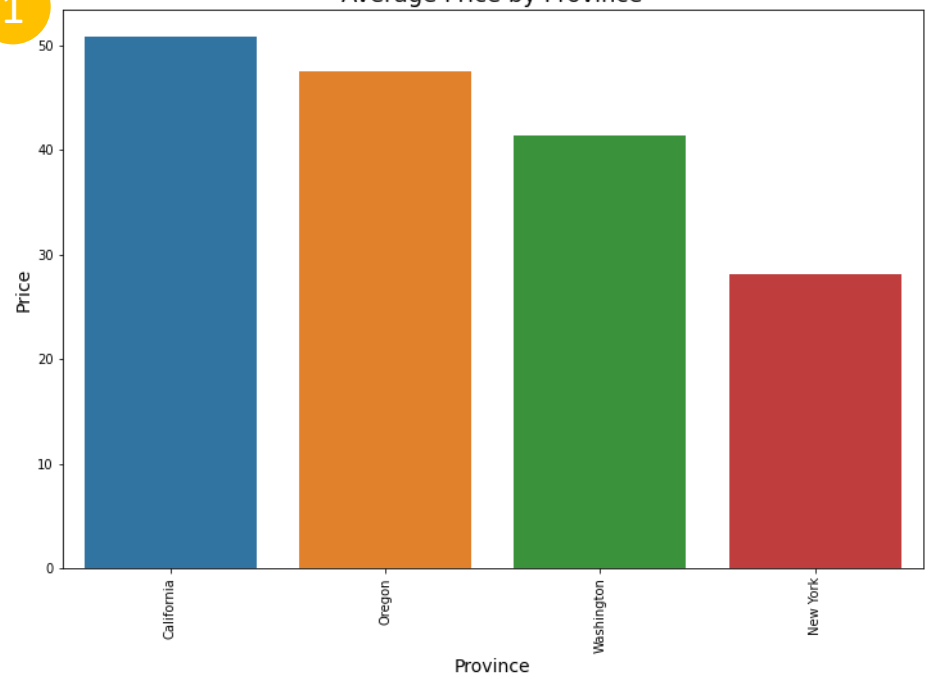


Exploration of Wine Variety and Provinces

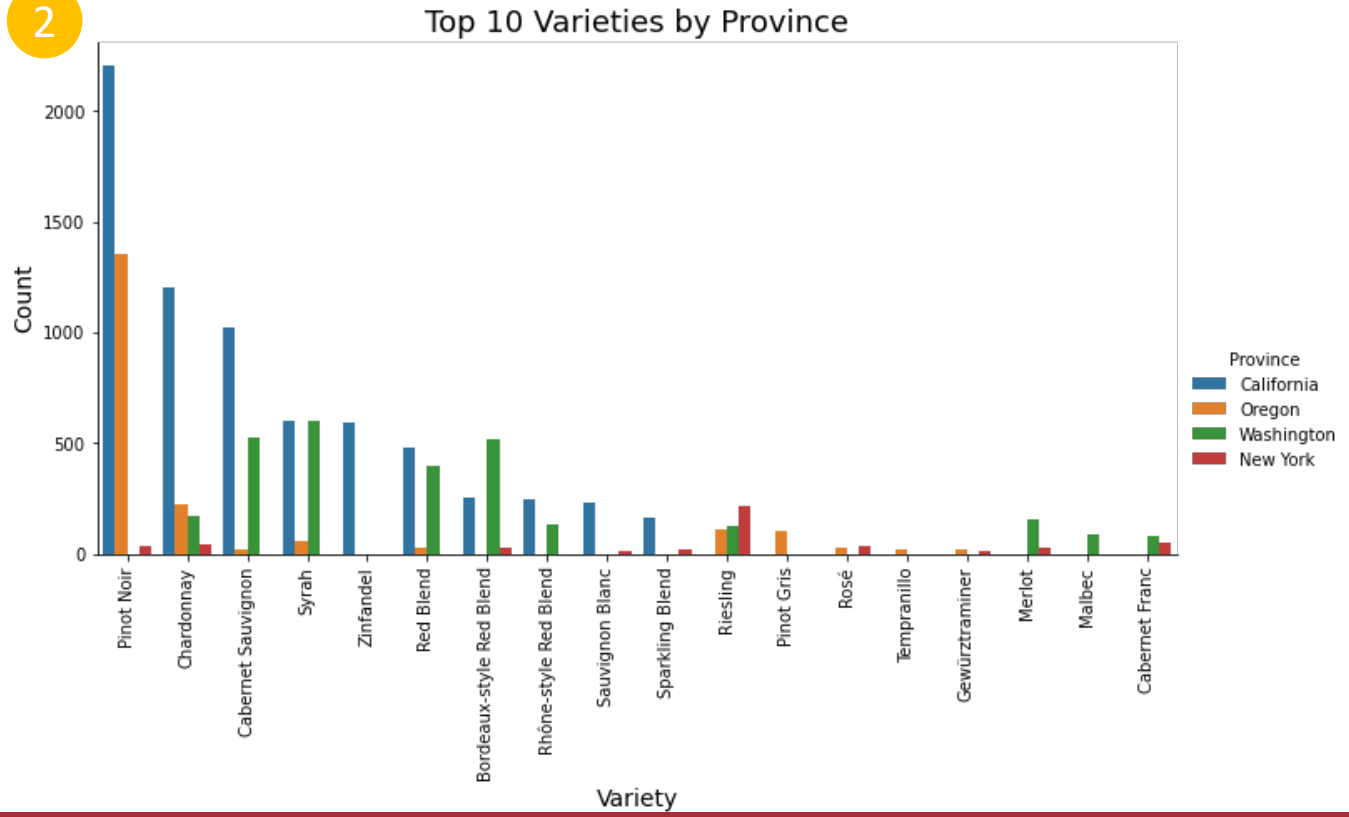
Key Findings

1. Regions with higher average prices of wine may produce higher quality wines as consumers are often willing to pay a premium for them, and production costs may be higher as well, which consequently drives up the price of the wines.
2. Pinot Noir tends to be the most popular wine variety overall, but the popularity of varieties differs province to province, with no certain pattern.

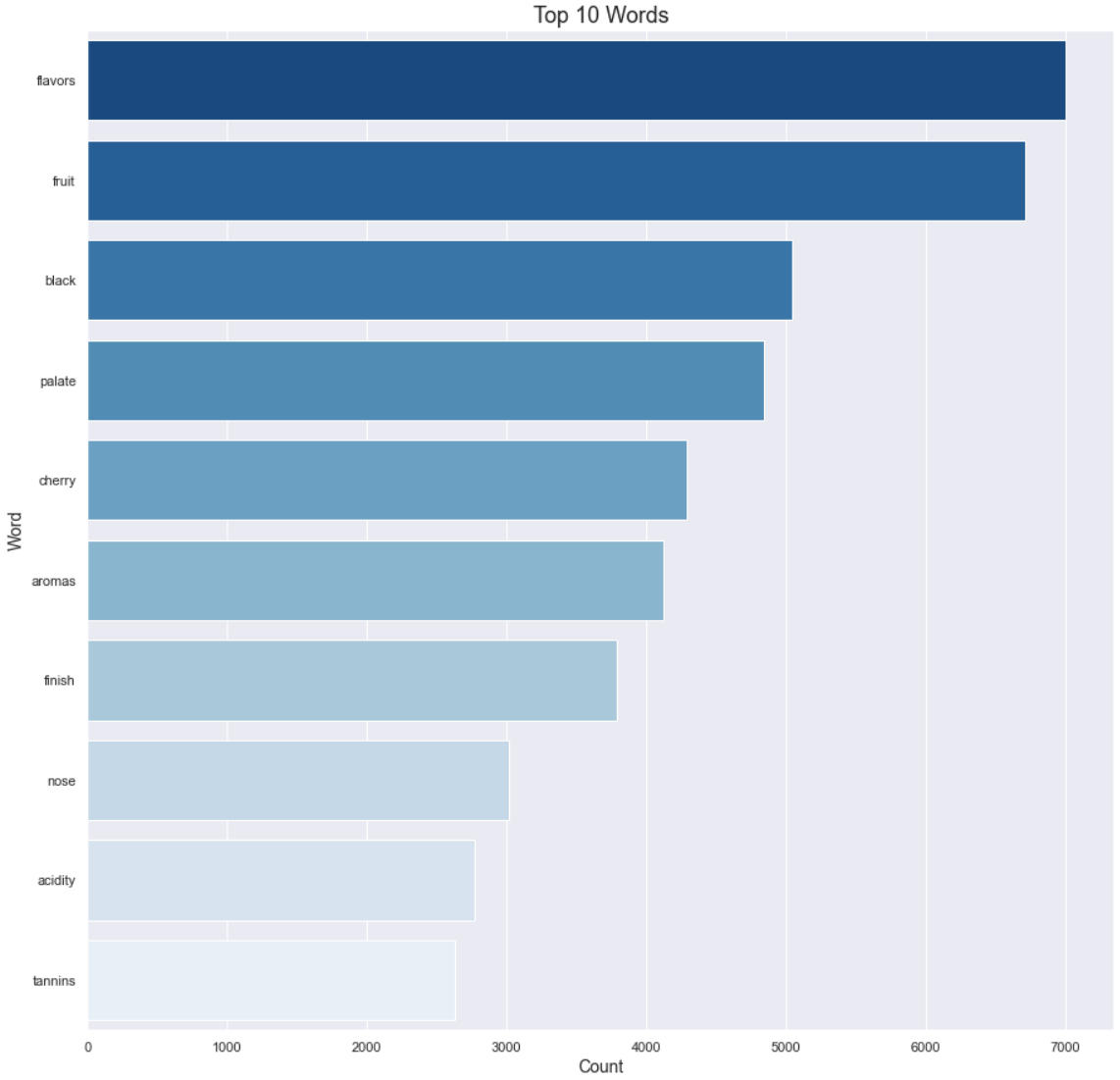
1

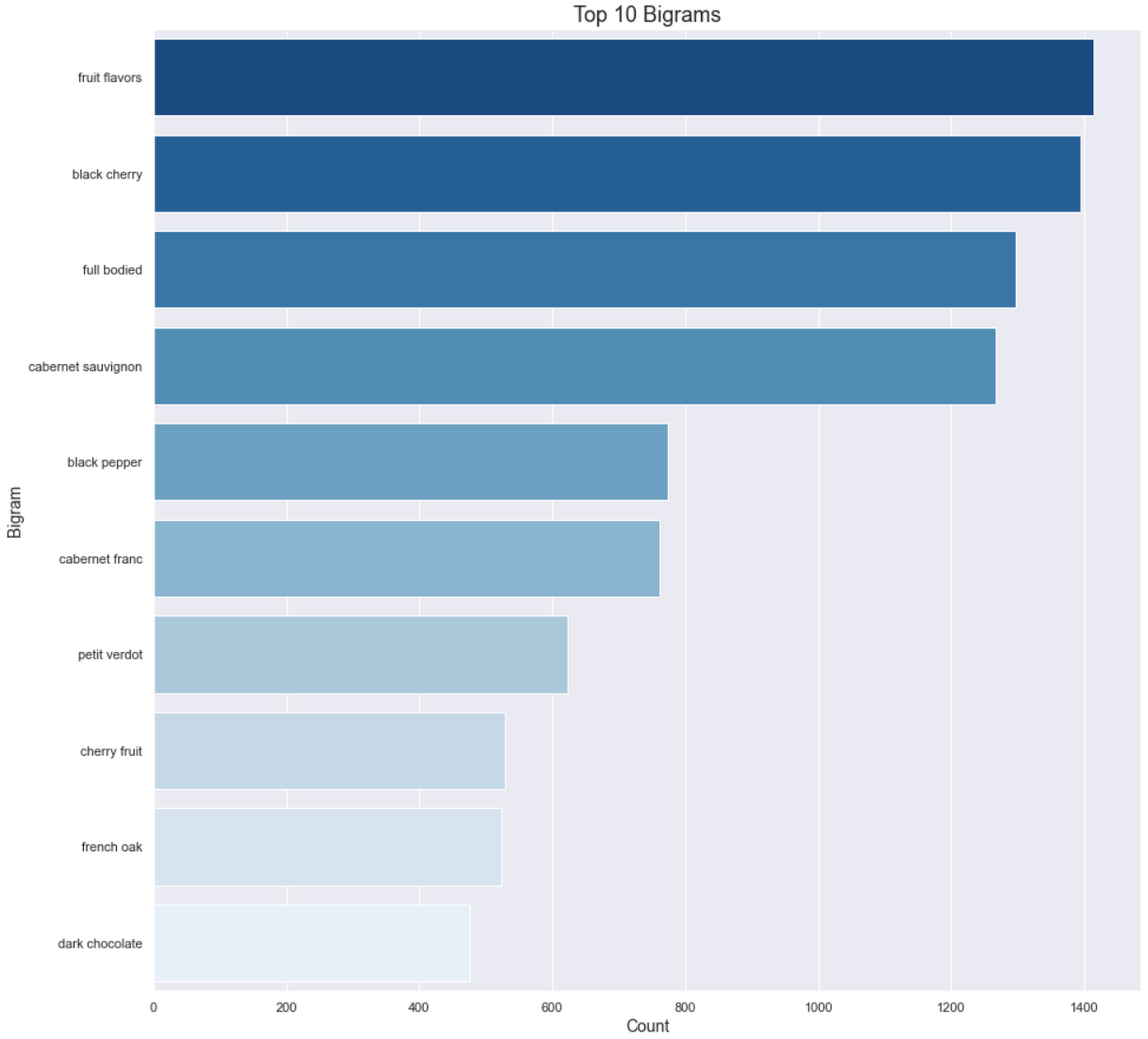


2



Top 10 Words & Bigrams

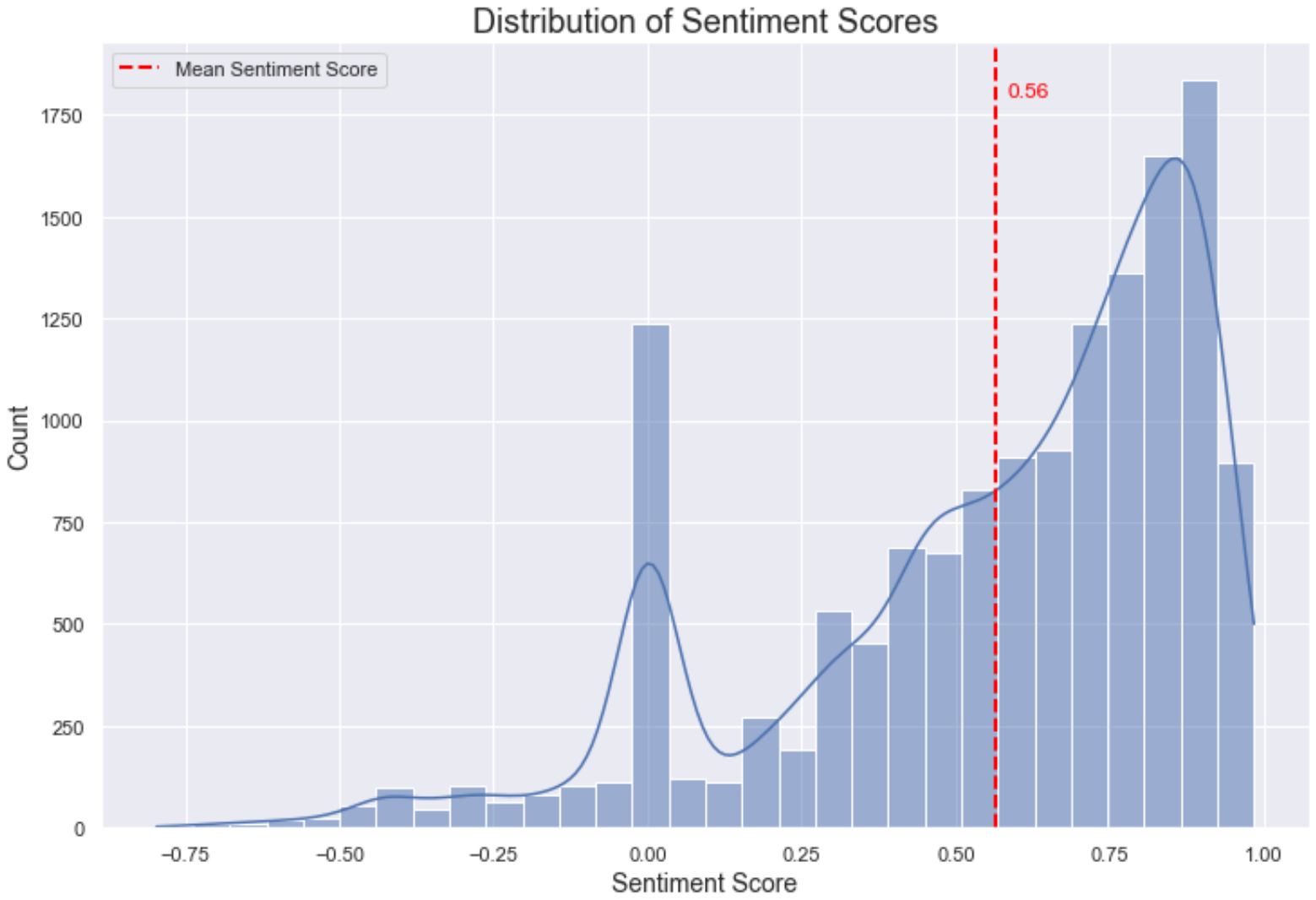




Sentiment Analysis of Wine Descriptions

Key Findings

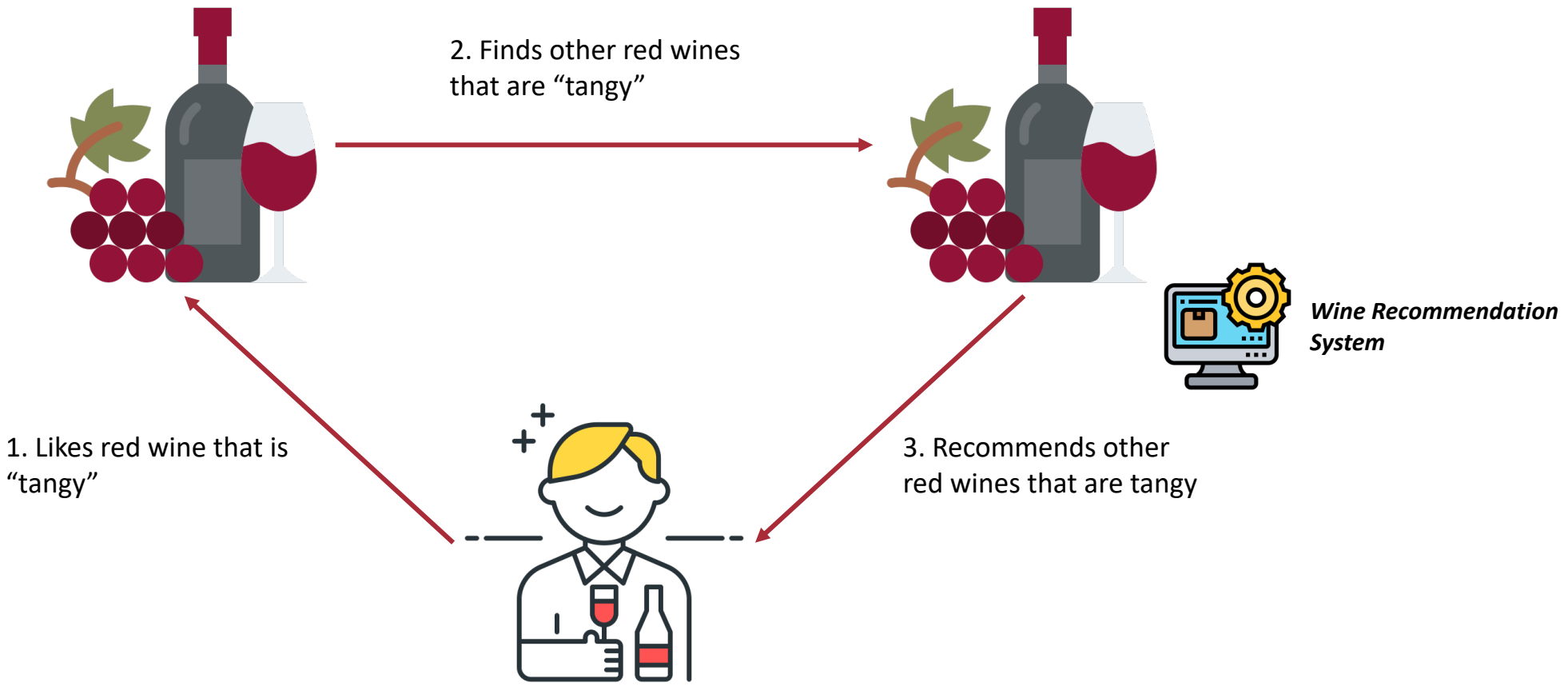
- Majority of the sentiments are positive.
- Negative sentiment scores are more spread out and less frequent, which results in a left-skewed distribution.
- The mean score being 0.56 indicates that the overall sentiment of the wine reviews is positive.



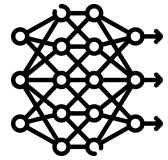
Data Science Approach

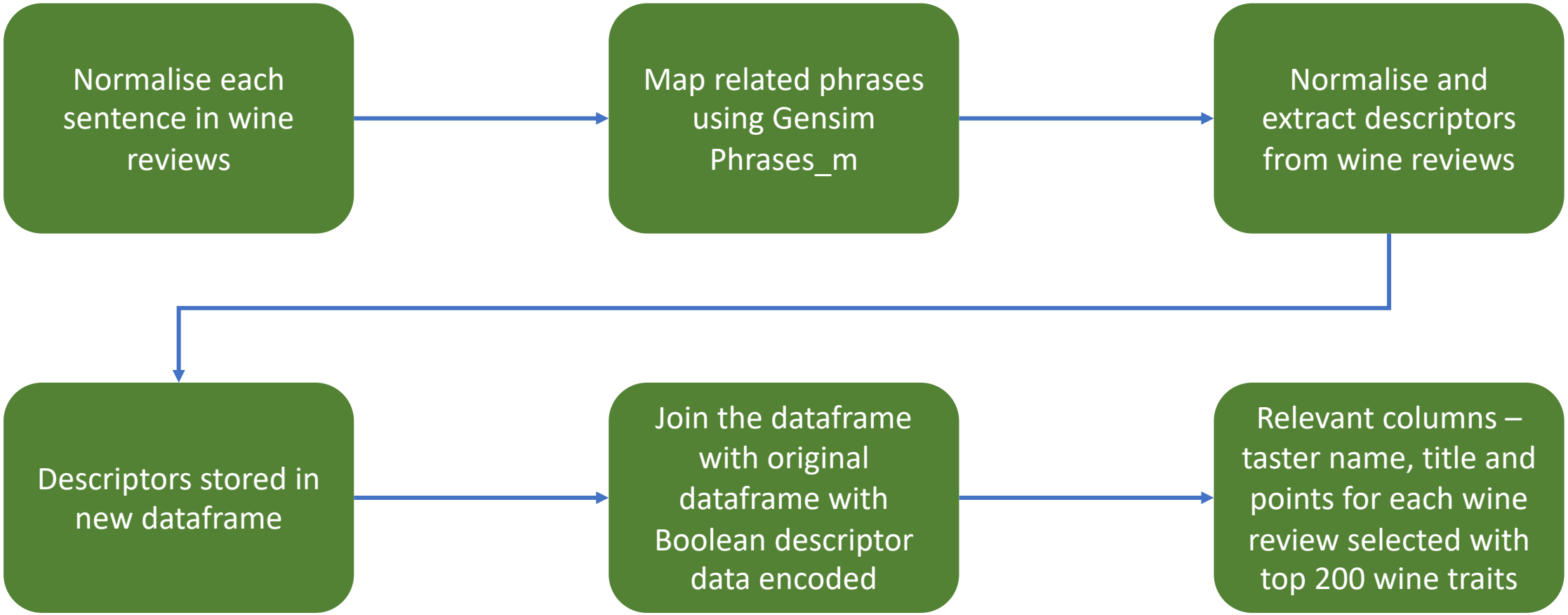
Problem Statement → Data Collection → Data Cleaning & Exploratory Data Analysis → **Data Pre-processing and Machine Learning Modelling** → Conclusion & Recommendation

Content-based Filtering Wine Recommender

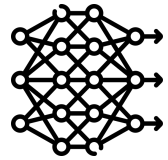


Data Preprocessing Steps





Baseline Model



- A simple recommendation system for wines by selecting 10 wines at random for each user and evaluating how many of those wines are relevant to the user based on their previous wine ratings.
- Two lists containing all the unique wine titles and taster names and set 10 wines to recommend to each user.
- Loop through each taster, selects all the relevant wines based on their previous ratings, and randomly selects 10 wines to recommend to the taster.
- Calculates the precision and recall values for these recommendations.
- Calculate the average precision and recall values across all tasters, and prints these values as the output of the code.

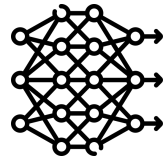
0.53

Average precision@k

0.000716

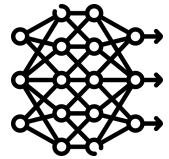
Average recall@k

Models Evaluated and Results



algorithm_name	ave_precision@k_score	ave_recall@k_score
FunkSVD	***0.867917***	***0.257780***
SVD	0.865417	0.257334
Baseline Predictor	0.822917	0.251537
KNN Baseline	0.822917	0.251537
Slope One	0.805417	0.250240
KNN Basic	0.805417	0.250240
KNN Means	0.805417	0.250240
KNN ZScore	0.805417	0.250240
Co-clustering	0.802917	0.249946
Normal Predictor	0.719583	0.205594
NonNegative Matrix Factorization	0.715417	0.246805

Post Hyper-parameter Tuning

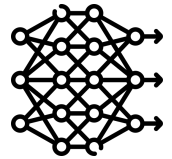


algorithm_name	ave_precision@k_score	ave_recall@k_score
Tuned FunkSVD	***0.952778***	***0.225906***
FunkSVD	0.867917	0.257780
SVD	0.865417	0.257334
Baseline Predictor	0.822917	0.251537
KNN Baseline	0.822917	0.251537
Slope One	0.805417	0.250240
KNN Basic	0.805417	0.250240
KNN Means	0.805417	0.250240
KNN ZScore	0.805417	0.250240
Co-clustering	0.802917	0.249946
Normal Predictor	0.719583	0.205594
NonNegative Matrix Factorization	0.715417	0.246805

Hyperparameters

- **'n_factors': 100**
- **'n_epochs': 40**
- **'lr_all': 0.01**
- **'reg_all': 0.02**

Model Performance Testing using Tuned FunkSVD



title	matching points	points
Wayfarer 2014 Wayfarer Vineyard Chardonnay (Fort Ross-Seaview)	91.689440	98
Quady 2006 Starboard Dessert Wine Port (Amador County)	91.613462	95
Ryan Cochrane 2015 Solomon Hills Vineyard Chardonnay (Santa Maria Valley)	91.592035	96
Williams Selyem 2014 Heintz Vineyard Chardonnay (Russian River Valley)	91.562822	96
Gary Farrell 2014 Ritchie Vineyard Chardonnay (Russian River Valley)	91.547778	96
Ryan Cochrane 2014 Solomon Hills Vineyard Chardonnay (Santa Maria Valley)	91.536482	95
Schramsberg 2008 Extra Brut Sparkling (California)	91.533279	94
Sandhi 2013 Sanford & Benedict Chardonnay (Sta. Rita Hills)	91.524916	95
Dragonette 2013 MJM Syrah (Santa Ynez Valley)	91.520572	95
Laetitia 2013 La Coupelle Single Vineyard Pinot Noir (Arroyo Grande Valley)	91.517997	96

Wine Traits

- **“Warm”**
- **“Tangy”**

Data Science Approach

Problem Statement → Data Collection → Data Cleaning & Exploratory Data Analysis → Machine Learning Modelling → Conclusion & Recommendation

Conclusion and Recommendation

- The Tuned FunkSVD model performed well in a simple wine recommender system, allowing users to input several traits they look for in a wine to generate the top 10 recommendations.
- Limitations of the project include the dataset not being representative of the entire wine industry and lacking some relevant features that could affect wine recommendations.
- In the future, a larger datasets with more comprehensive wine features can be explored to improve the recommender system and deploy the model onto a Flask app.

Thank You

Annex

- Average Points by Province

Average Points by Province