MARCH 18, 2021 | DATA MINING PRINCIPLES

# WINE REVIEWS & PREDICTION

Anna Willman, Chenchen Shentu, Fan Yang,
Olivia Yang, Wilson McDermott

# Presentation Summary

Business Case & Value

Data Source

Exploratory Data Analysis

Model Methodology & Results:
Sentiment Analysis, Price & Point Prediction, Recommendation System

Challenges

Future Extensions

Team Bio

# Business Case

## GOAL

Make business decisions easier by predicting sentiment, price, and point rating of a specific wine, then build a recommendation system to suggest similar wines

## APPROACH

Use a wine deep learning dataset from Kaggle to execute a price & point prediction, sentiment analysis, and recommendation system based on each wine's attributes

# Business Value

*Wine Distributor Use Case*

### Should we carry this wine?

Run a **sentiment analysis** to decide if distributor should purchase a new wine based on the sentiment extracted from its review

### How should we price this wine?

Execute a **price prediction analysis** to determine an appropriate price to sell the wine for based on similar wines

### When do we recommend this wine?

Create a **recommendation system** to provide similar suggestions based on client flavor or grape preferences

# Data Source

Wine Review Dataset (Kaggle):
Scraped from WineEnthusiast magazine

**Original Dataset**

Records: 280K

Attributes: 14

*cleaning* →

- Dropped attributes with majority null entries
- Dropped duplicate records
- Removed geo attributes that were too broad or specific
- Removed price outliers

**Final Dataset**

Records: 152K

Attributes: 7

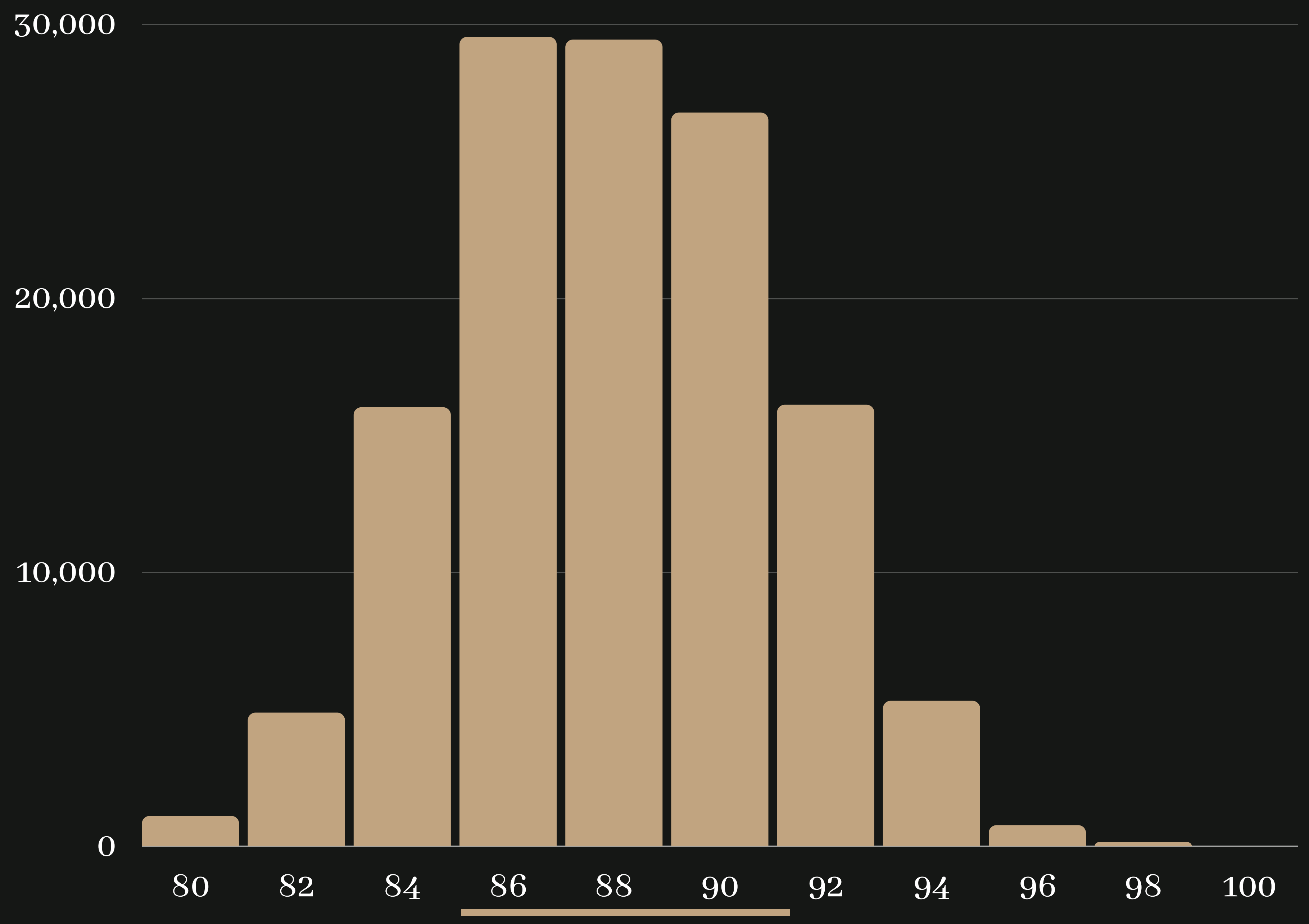
# Final Dataset

| Description | Designation | Points | Price | Province | Variety | Winery |
|---|---|---|---|---|---|---|
| Teh aromas bring notes of herb, sweet tobacco and ash. The plum flavors are tart and full in fell, with the tannins giving a (quite) chalky squeeze. | NAN | 87 | 16.0 | Idaho | Cabernet Sauvignon | Sawtooth |
| Full-bodied, rich and unctuous, this is an exotic, flamboyant white Châteauneuf-du-Pape for drinking over the next year. Grilled pineapple is drizzled with caramel and cinnamon, wrapping up long and lush. | Vieilles Vignes | 94 | 66.0 | Rhōne Valley | Rhōne-style White Blend | Tardieu-Laurent |
| You might mistake this for a young coastal Pinot. It's crisp, light-bodied and silky, with cherry, cola, herb tea and spicy, smoky flavors. If only the wine were dry. | Castelleto | 94 | 22.0 | California | Sanglovese | Mount-Palomar |
| Produced in one of the estates belonging to the Lapalu family, this wine is soft, rounded and ready to drink now. Gentle tannins give shape and structure to the black currant-fruit that finish on a fresh, crisp note. | NaN | 86 | 19.0 | Bordeaux | Bordeaux-style Rad Blend | Château Lacombe Noaillac |
| The bouquet of cassis, blackberry and controlled oak is welcoming, while the palate shows a spot of piercing acidity along with snappy black berry, cassis and light olive flavors. This is quintessential Chilean Cab with a hint of herbal character...... | Las Vascos Reserve | 89 | 20.0 | Colchagua Valley | Cabernet Sauvignony | Domaines Barons de Rothschild (Lafite) |

source: https://www.kaggle.com/zynicide/wine-reviews?select=winemag-data-130k-v2.csv

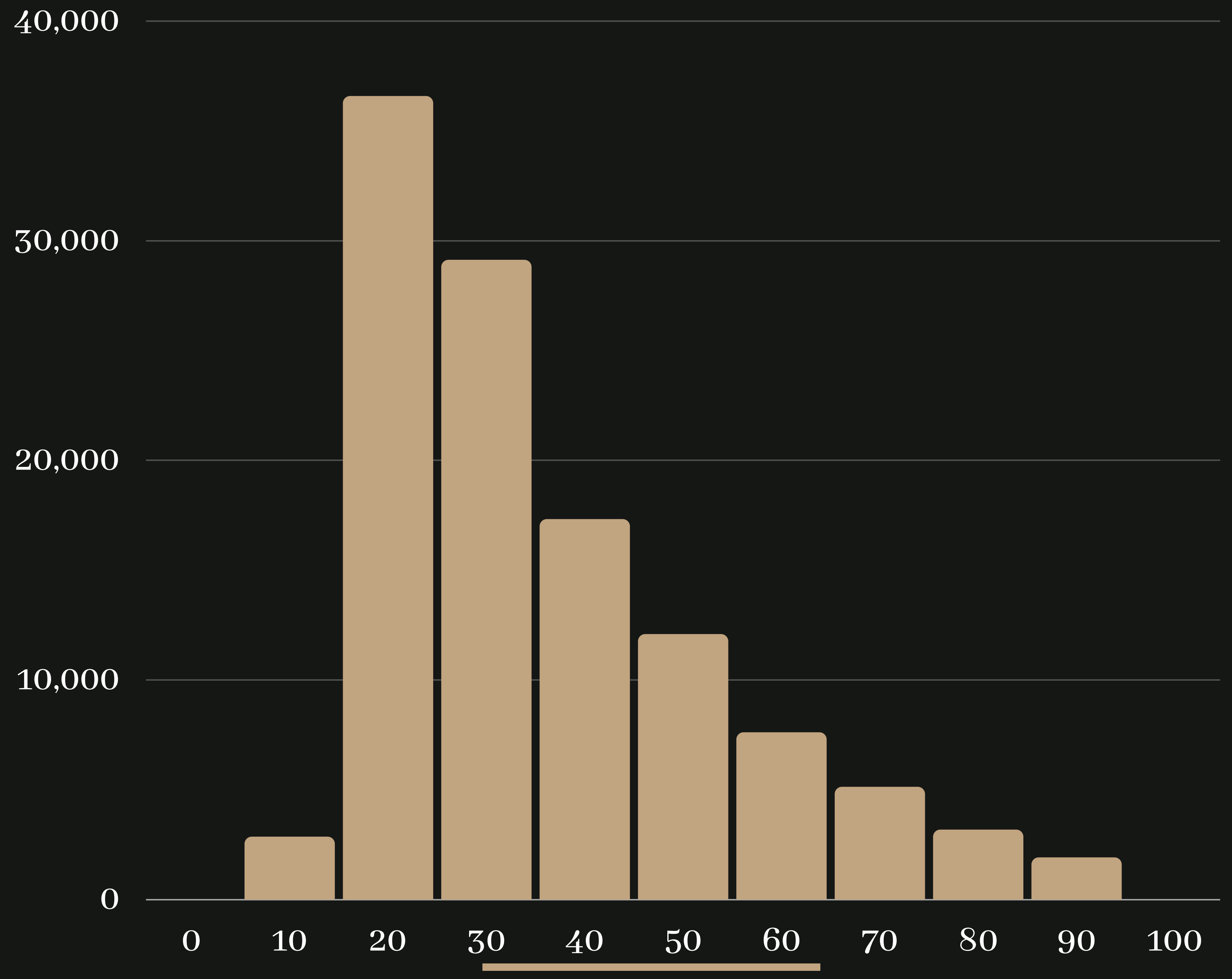
# Exploratory Data Analysis

## DISTRIBUTION OF POINTS:



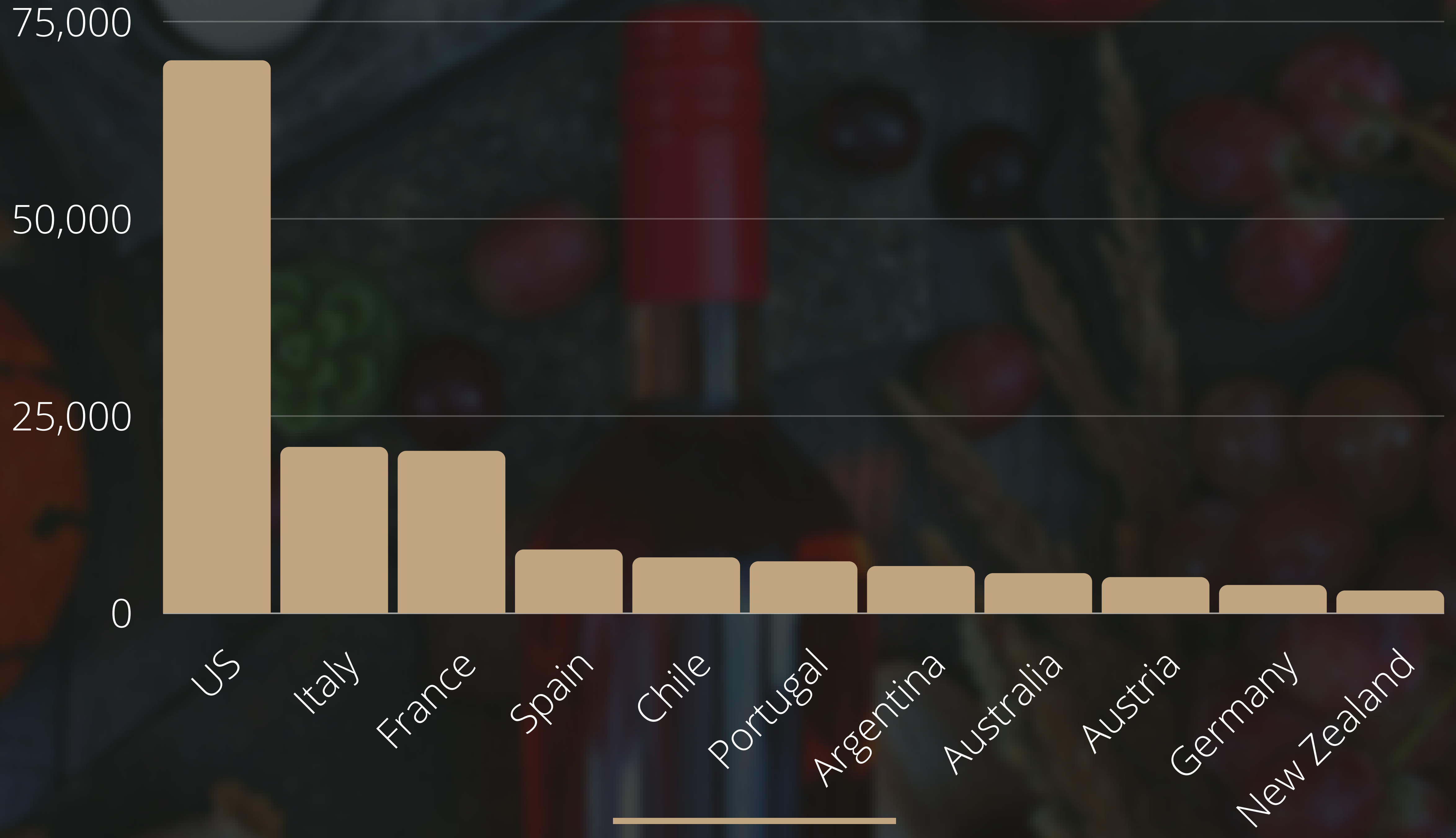

Average Points: 88
Median Points: 88

## DISTRIBUTION OF PRICES:




Average Price: $30
Median Price: $25

# GEOGRAPHIC DISTRIBUTION OF WINES:




**Top 3 :**

US, Italy, and France

## COUNTRIES WITH HIGHEST AVERAGE POINTS:

| Country | Points | Price |
| --- | --- | --- |
| England | 91.8 | 52.6 |
| Austria | 90.0 | 31.6 |
| India | 89.3 | 14.4 |
| Germany | 89.3 | 40.4 |
| Canada | 89.1 | 35.5 |

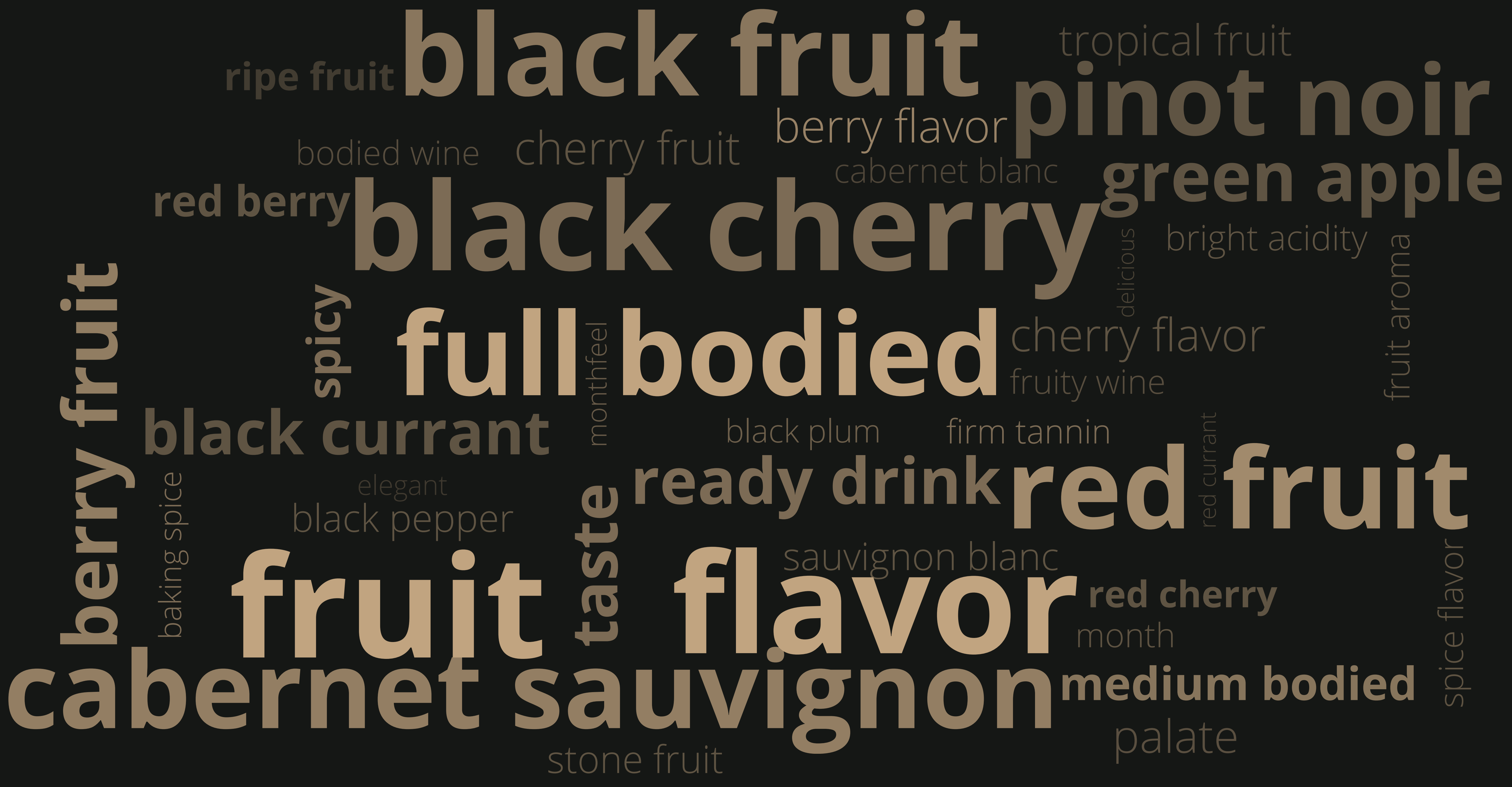
**Top 3** :
England, Austria, & India

## COUNTRIES WITH HIGHEST AVERAGE PRICES:

| Country | Points | Price |
| --- | --- | --- |
| Switzerland | 88.1 | 65.1 |
| England | 91.8 | 52.6 |
| US-France | 88.0 | 50.0 |
| Hungary | 88.4 | 43.3 |
| France | 88.7 | 42.9 |

**Top 3**:
Switzerland, England, & US-France

# Most Common Words in Description



# Model Methodology & Results

# Model Summary

### Sentiment Analysis

Use BERT to predict sentiment of review

### Price Prediction

Price prediction of wine based on sentiment, points, variety, and province

### Recommendation System

Wine & Grape recommendations based on key words in descriptions

# Sentiment Analysis

## GOAL

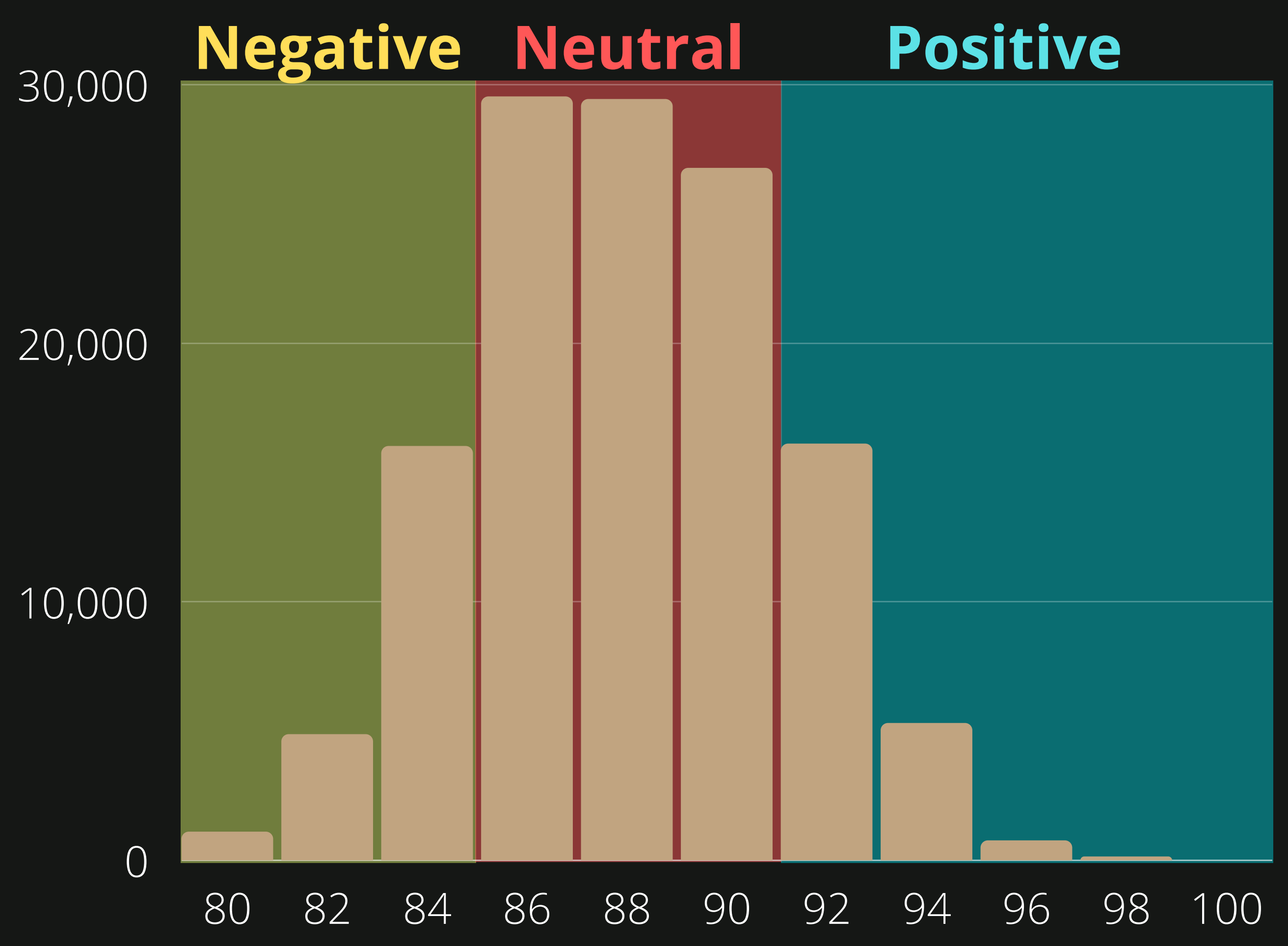
Predict sentiment towards each wine based on description

## APPROACH

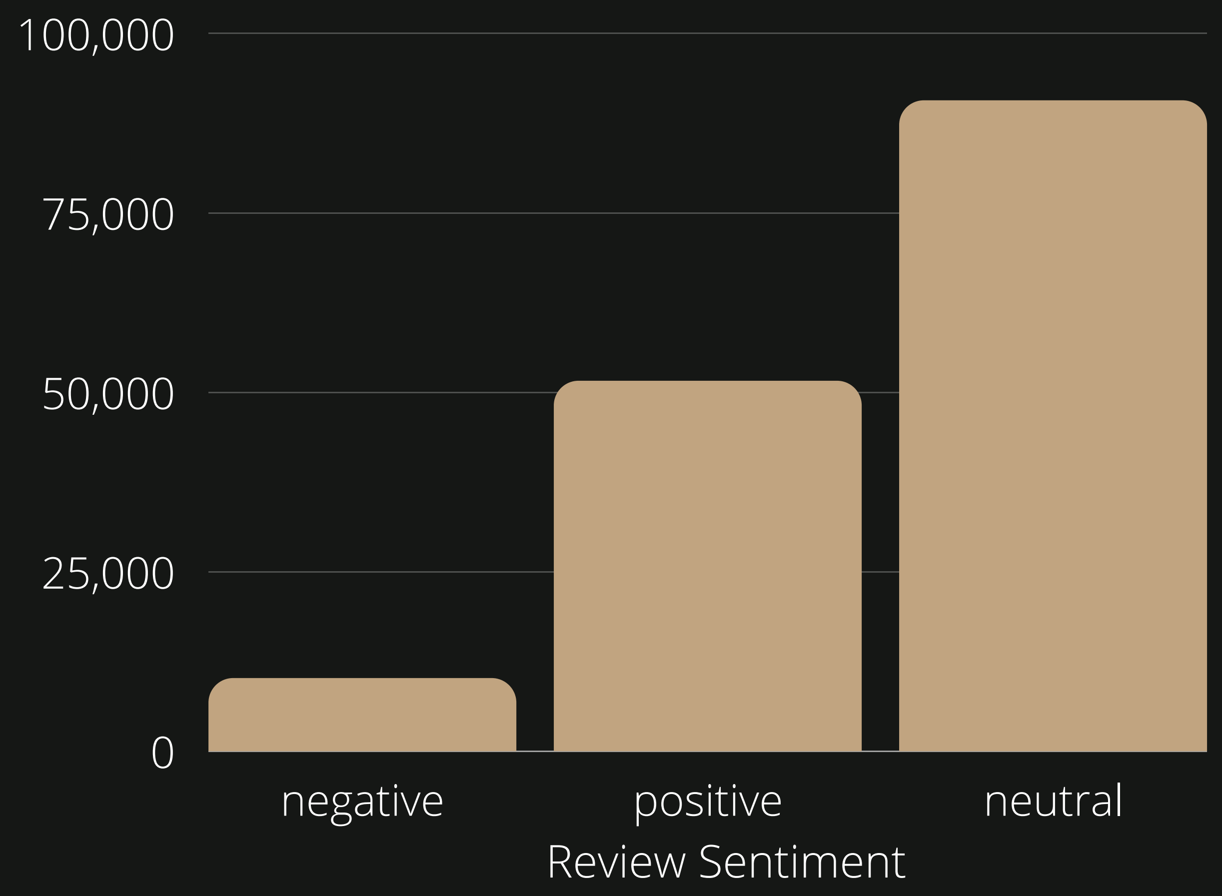
Use Transfer Learning from BERT to build Sentiment Classifier model using the Transformers library

# SENTIMENT SUMMARY

## POINTS HISTOGRAM




## SENTIMENT COUNTS




Data is **imbalanced**:

- Downsampling decreased model performance
- Upsampling & SMOTE increased process time and not enough memory to run models

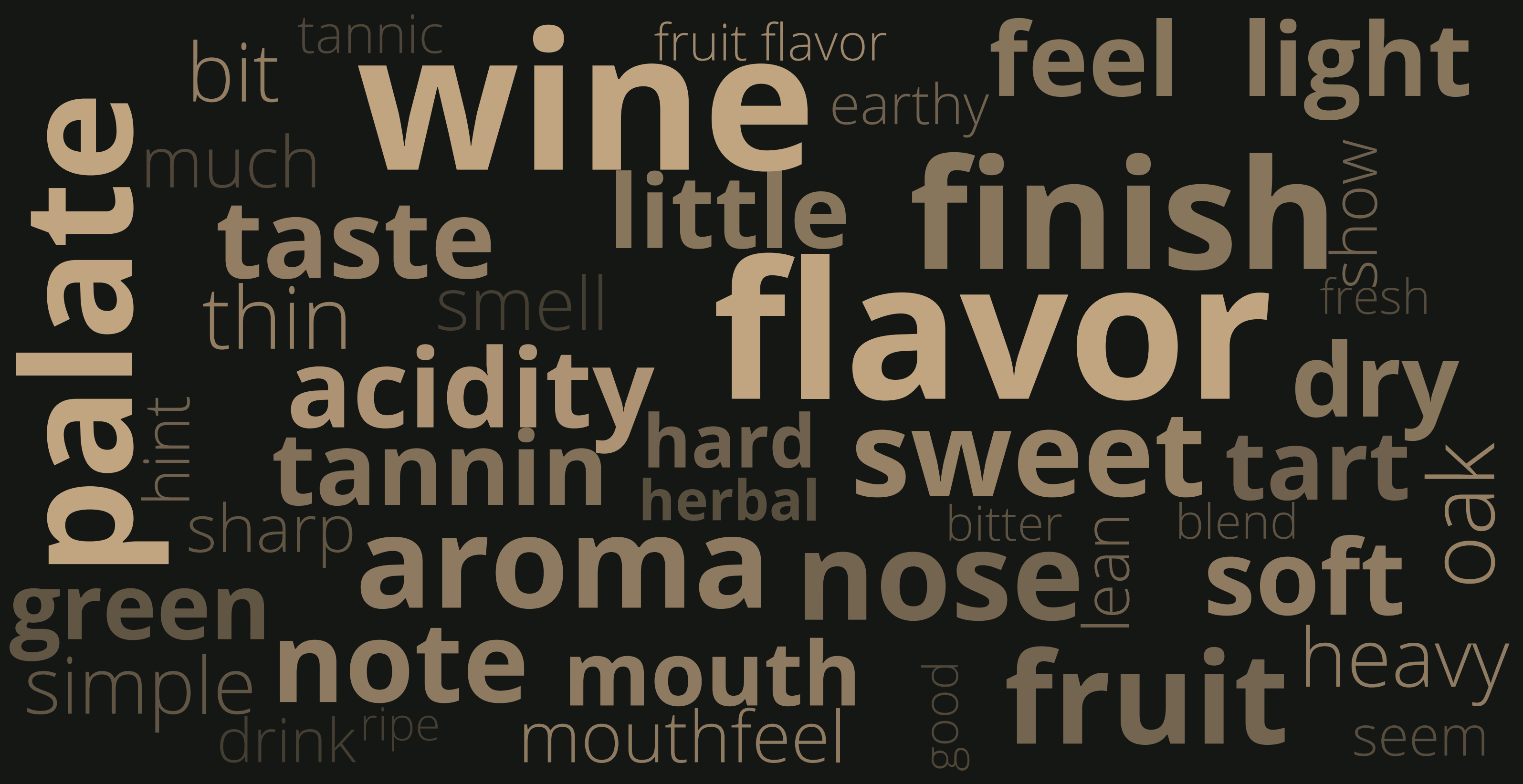
**Negative** *"Too sweet and sugary. The relatively low alcohol (13.7%) seems to have been accomplished at the cost of residual sugar, making the cherry and blackberry fruit taste like a dessert wine."*

**Neutral** *"Apple, melon, saline and buttered popcorn aromas set up a lively palate with snappy acidity. Apple and melon flavors turn a bit stalky and bitter on the finish."*

**Positive** *"From Mia Klein, this is a seriously good Cabernet Sauvignon, even better than the winery's fine 2004. It shows a great balance of ripe tannins and fine acidity, with a judicious application of smoky oak."*

# COMMON PHRASES BY SENTIMENT

## NEGATIVE




*__Negative__ reviews focused on wines' finish & flavor and used sensory language around scent & taste to describe wine*

## POSITIVE


*Full-bodied, fruit flavored (ie. black cherry) wines tend to receive more __positive__ reviews, especially cabernet sauvignons*

# DATA PREPROCESSING

## WHAT IS BERT?

**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
NLP model pre-trained by Google conditioned on both left and right context of text

## CONVERT TEXT TO NUMBER TOKENS USING PRE-TRAINED BERTTOKENIZER:

**Sentence:** A very delicious wine, rich in fruits and spices, and easy to drink for its softness.

**Tokens:** ['A', 'very', 'delicious', 'wine', ',', 'rich', 'in', 'fruits', 'and', 'spices', ',', 'and', 'easy', 'to', 'drink', 'for', 'its', 'soft', '##ness', '.']

**Token IDs:** [138, 1304, 13108, 4077, 117, 3987, 1107, 11669, 1105, 25133, 117, 1105, 3123, 1106, 3668, 1111, 1157, 2991, 1757, 119]

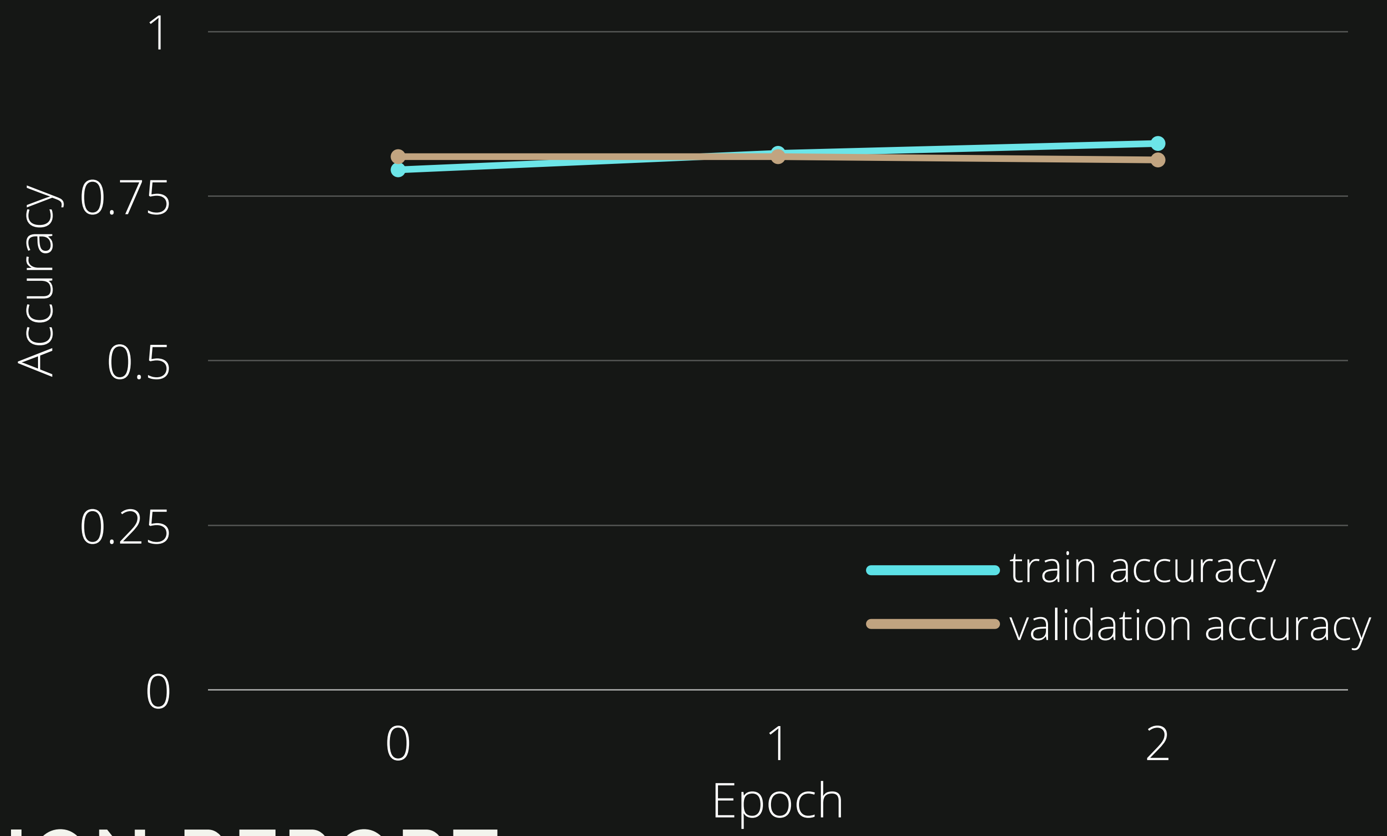
## STORE TOKENS IN TENSOR, ADD PADDING, & CREATE ATTENTION MASK:

**Input ID Tensor:**
tensor([ 101, 138, 1304, 13108, 4077, 117, 3987, 1107, 11669, 1105, 25133, 117, 1105, 3123, 1106, 3668, 1111, 1157, 2991, 1757, 119, 102, 0, 0, 0, 0, 0, 0, .... 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

**Attention Mask:**
tensor([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, .... 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

# MODEL TRAINING & EVALUATION

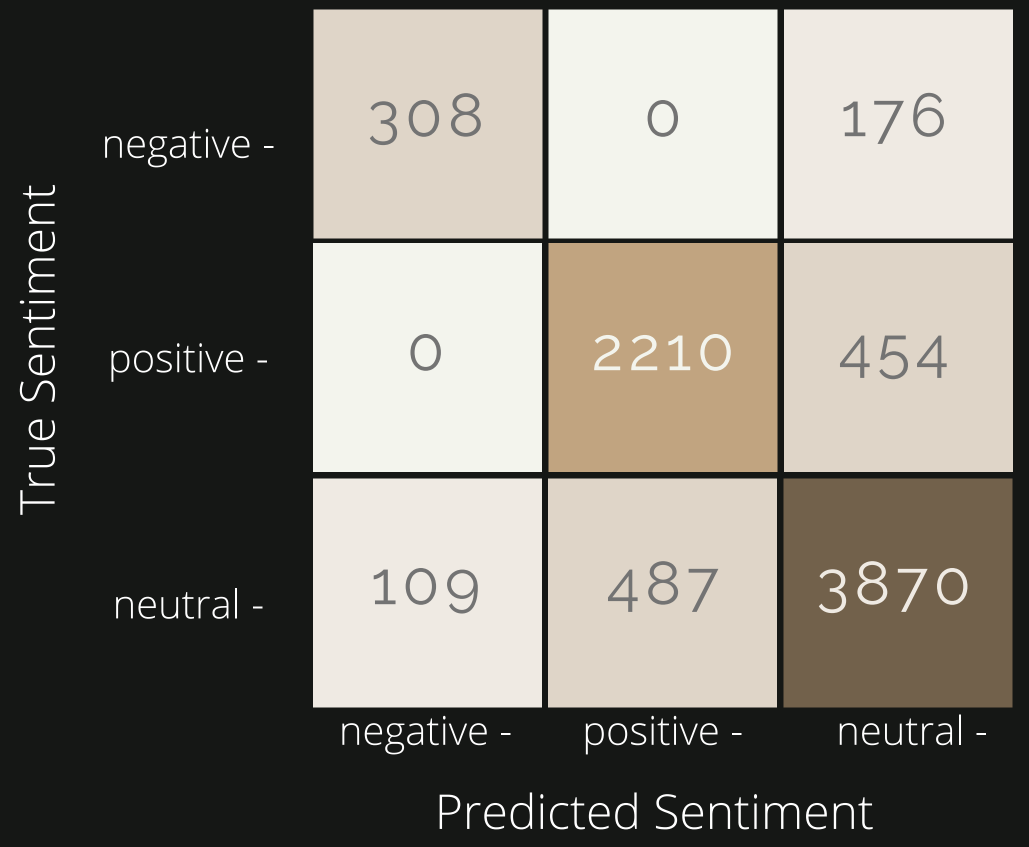
## TRAINING HISTORY



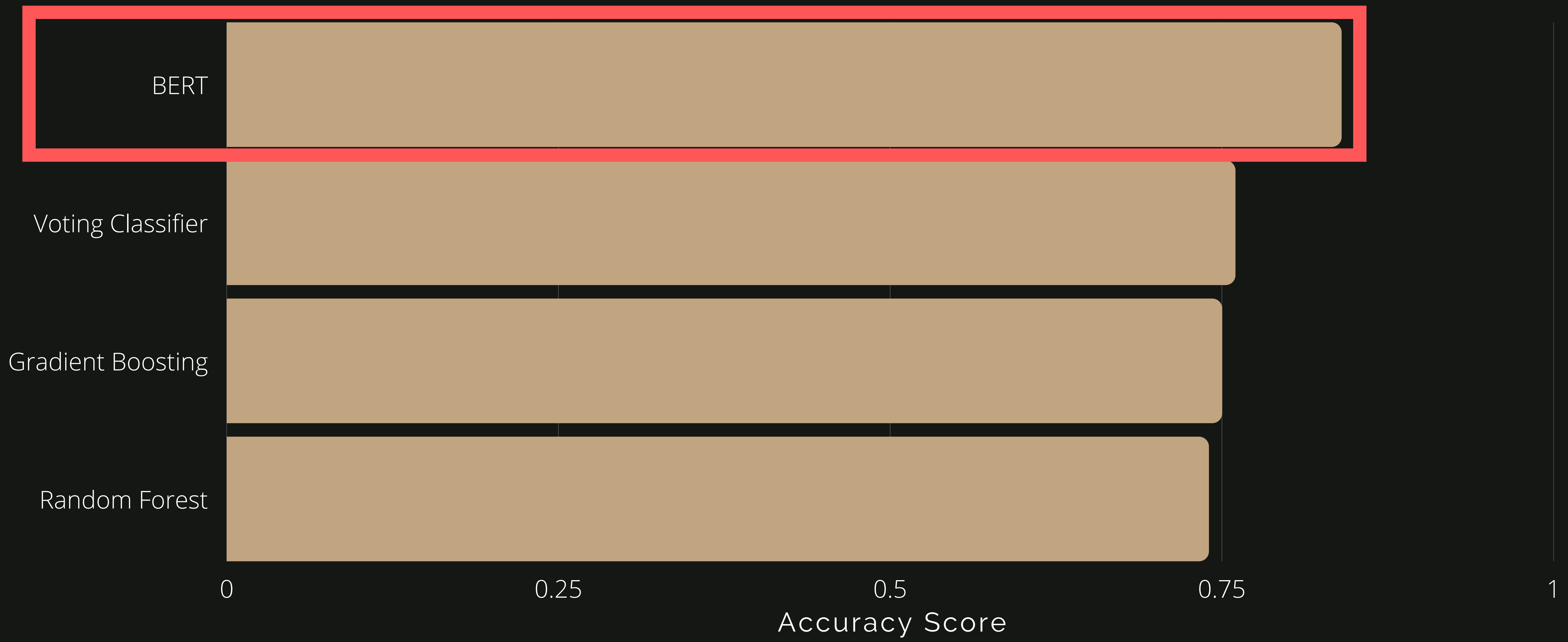

## CLASSIFICATION REPORT

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Negative | 0.74 | 0.64 | 0.68 | 484 |
| Positive | 0.82 | 0.83 | 0.82 | 2664 |
| Neutral | 0.86 | 0.87 | 0.86 | 4466 |
| Accuracy | | | 0.84 | 7614 |
| Macro Avg | 0.81 | 0.78 | 0.79 | 7614 |
| Weighted Avg | 0.84 | 0.84 | 0.84 | 7614 |

## CONFUSION MATRIX



# MODEL COMPARISON




Recommendation: **BERT model**

Compared to other classification models, BERT did the best job accurately determining which descriptions were positive, negative or neutral.

# Price & Point Prediction

## GOAL

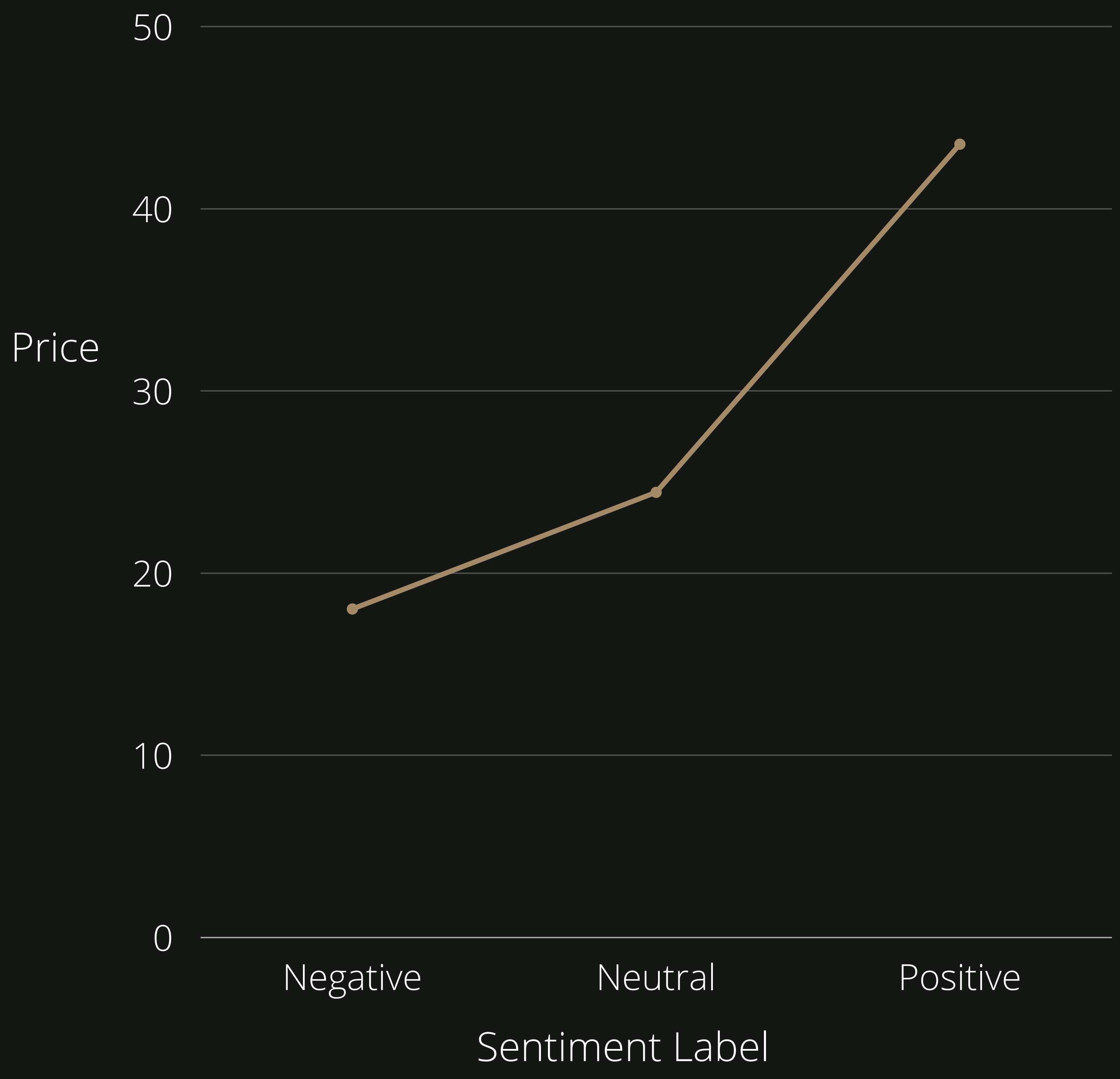
Predict price range and points of wine based on features

## APPROACH

1 - Classification of wine as "Expensive", "Mid-Range" or "Cheap"

2 - Regression problem to predict the points of each bottle of wine

# SENTIMENT & PRICE:

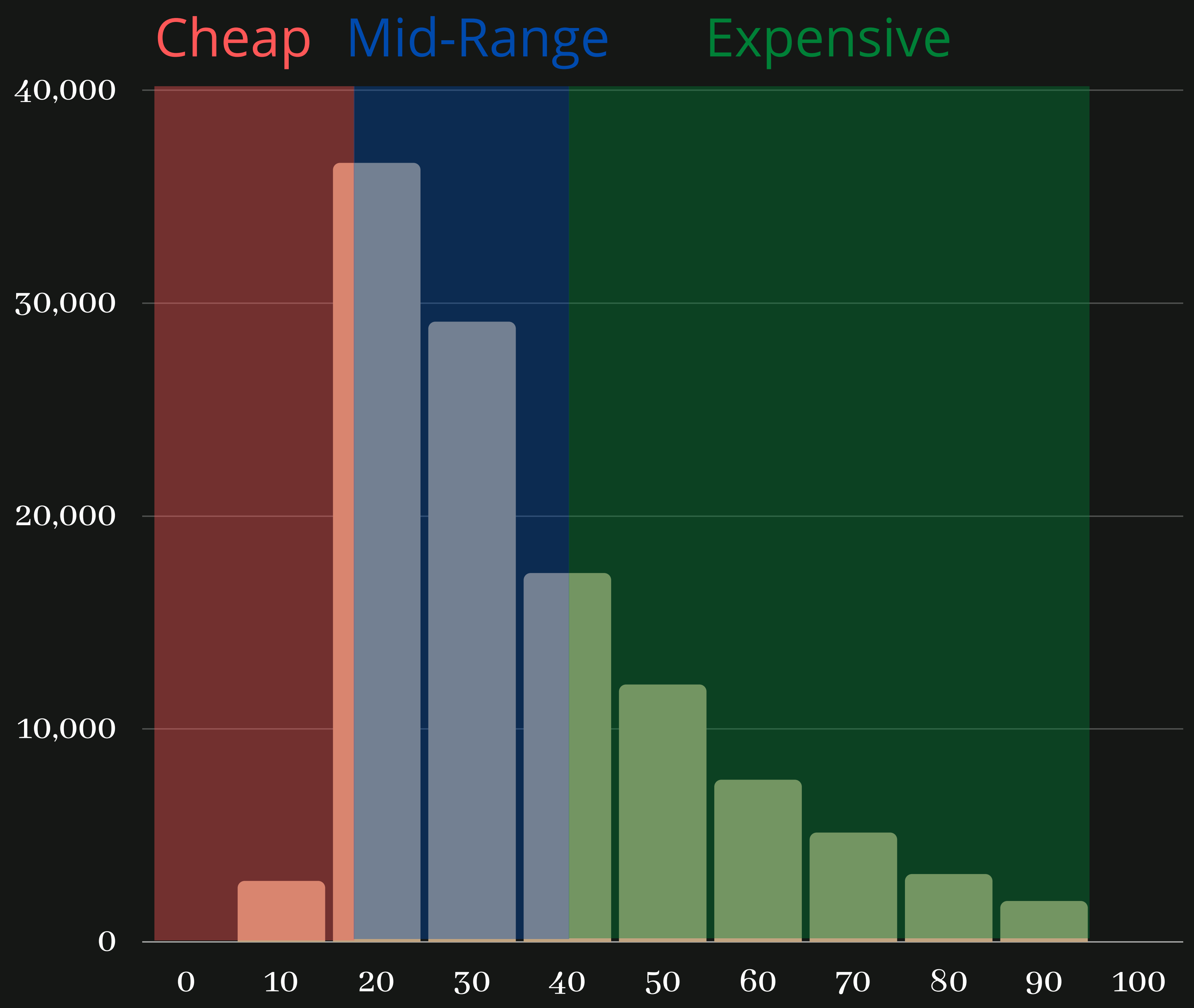



| | Avg Points | Avg Price |
|---|---|---|
| Negative: | 82 | $18 |
| Neutral: | 87 | $24 |
| Positive: | 91 | $44 |

## Other Explanatory Features:

- Province
- Variety

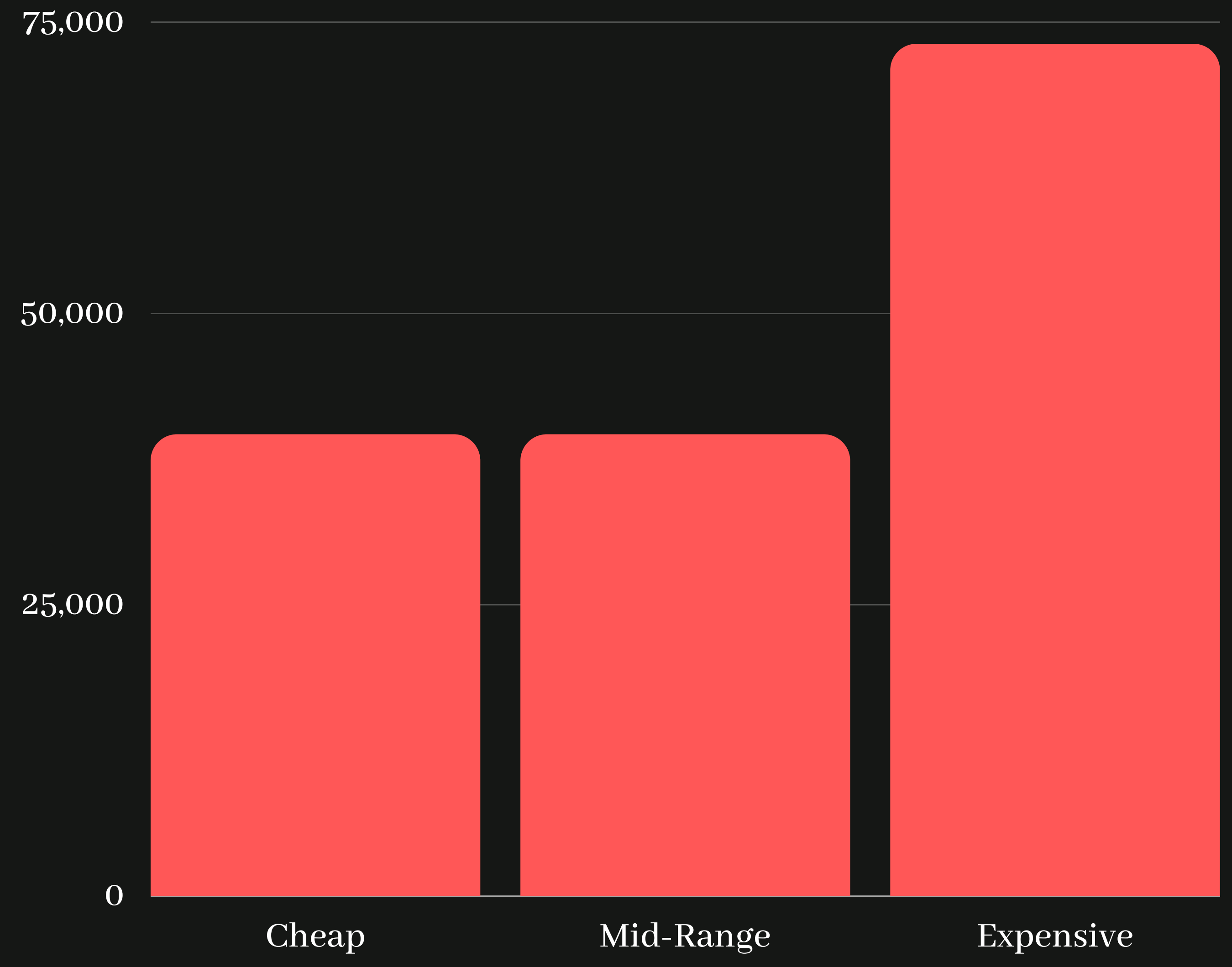
# PRICE CLASSIFICATION:




## Price

Count: 152,261
Upper Quartile: $40
Mean: $30
Median: $25
Lower Quartile: $16

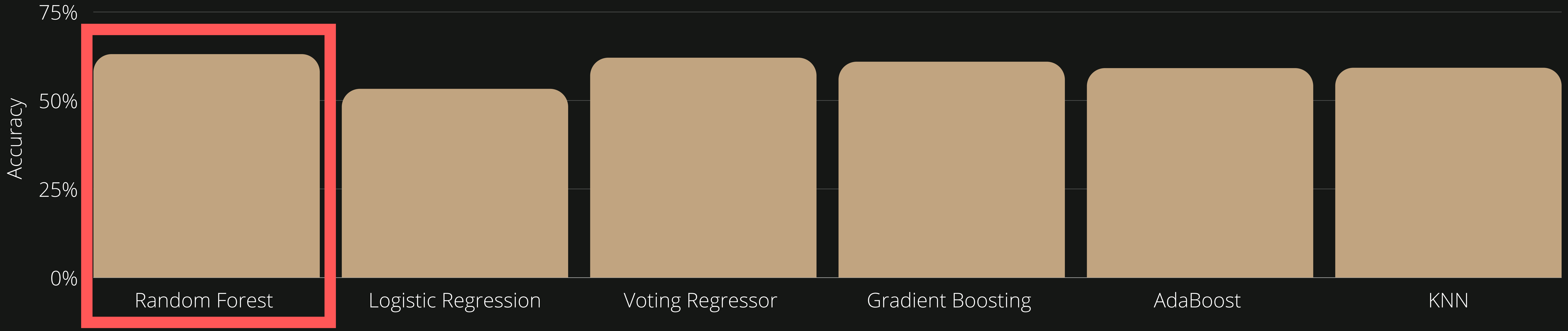
# RESAMPLING OF TRAINING SET:




*S.M.O.T.E.* →

- Split data into test and train then resampled the training data via S.M.O.T.E. to balance classes

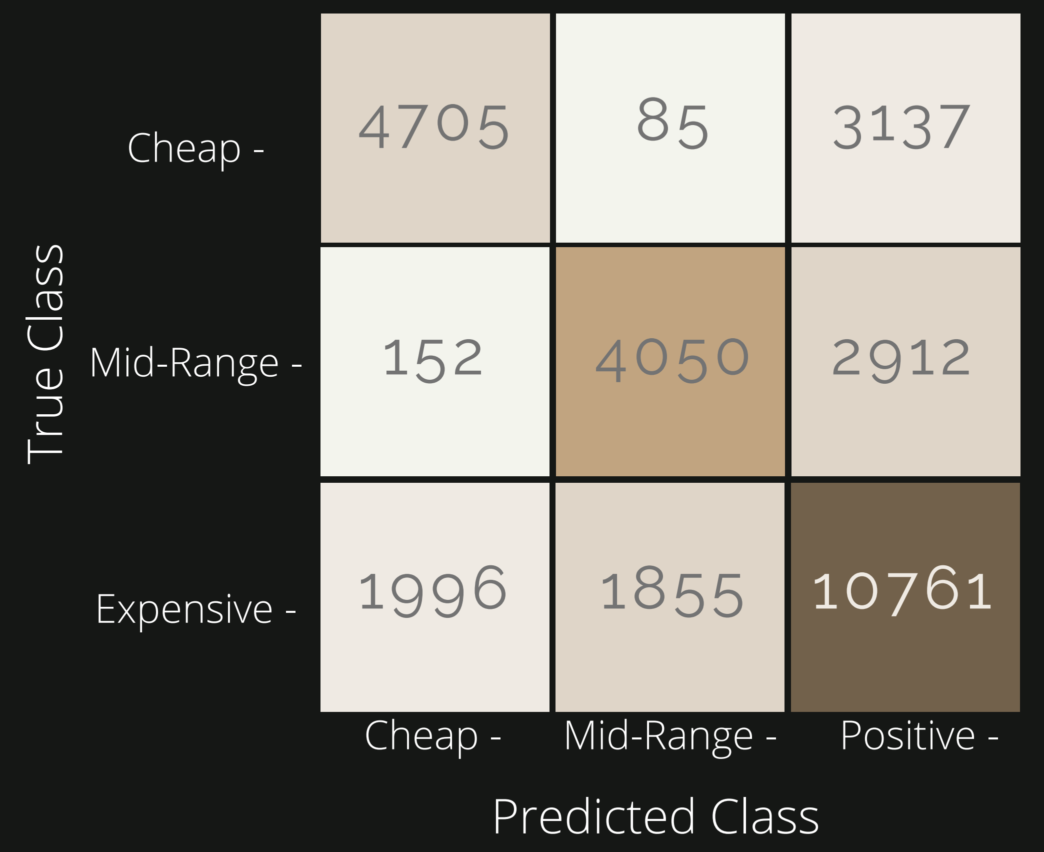
# MODELS & EVALUATION:




## RANDOM FOREST CLASSIFICATION REPORT

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Cheap | 0.59 | 0.74 | 0.66 | 7927 |
| Mid-Range | 0.61 | 0.74 | 0.67 | 7914 |
| Expensive | 0.68 | 0.51 | 0.59 | 14612 |
| | | | | |
| Accuracy | | | 0.63 | 30453 |
| Macro Avg | 0.63 | 0.66 | 0.64 | 30453 |
| Weighted Avg | 0.64 | 0.63 | 0.63 | 30453 |

## CONFUSION MATRIX



# Point Prediction

Is the relationship between province, variety, price, and sentiment strong enough to accurately predict points?

# PROVINCE / VARIETY & POINTS:

## PROVINCES WITH HIGHEST AVERAGE POINTS:

| Province | Points |
| --- | --- |
| Südburgenland | 94.0 |
| Martinborough Terrace | 93.0 |
| Mittelrhein | 92.3 |
| England | 91.8 |
| Santa Cruz | 91.5 |

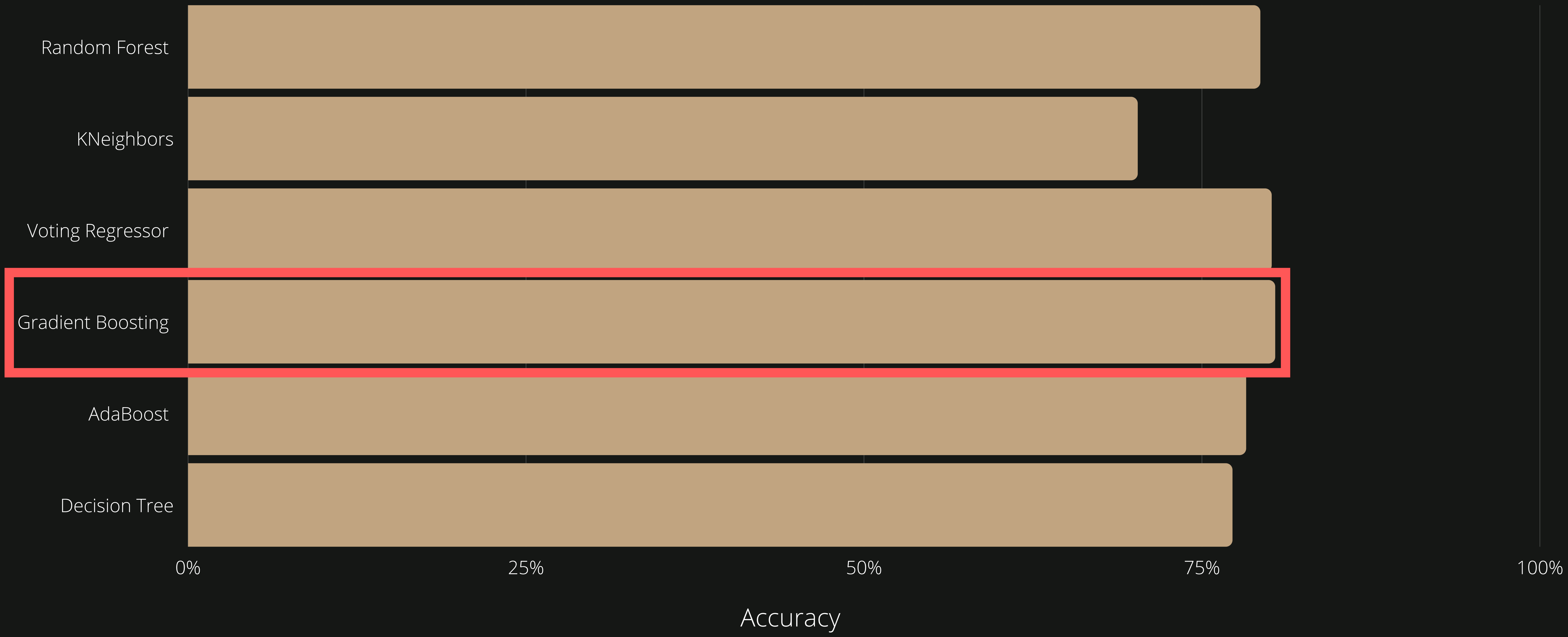
## VARIETIES WITH HIGHEST AVERAGE POINTS:

| Variety | Points |
| --- | --- |
| Gelber Traminer | 95.0 |
| Tinta del Pais | 95.0 |
| Riesling-Chardonnay | 94.0 |
| Blauburgunder (Pinot Noir) | 93.0 |
| Garnacha-Cariñena | 93.0 |

**Top 3** :
Südburgenland, Martinborough Terrace, & Mittelrhein

**Top 3**:
Gelber Traminer, Tinta del Pais, & Riesling-Chardonnay

# MODELS & EVALUATION:




## Best Model: Gradient Boosting Regressor

Accuracy (r-squared): 80.4%
Test set RMSE: 1.35
Train Set RMSE: 1.36

# Recommendation System

## GOAL

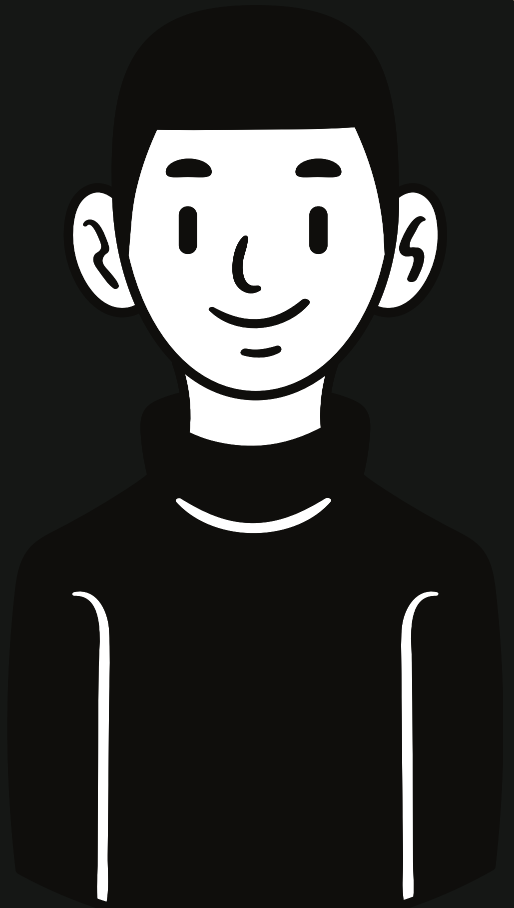
Recommend similar wines or grapes based on key words in descriptions

## APPROACH

- Wine Title (Doc2Vec)
- Variety (Content-based recommendation system)

# WINE RECOMMENDATION BASED ON DESCRIPTION

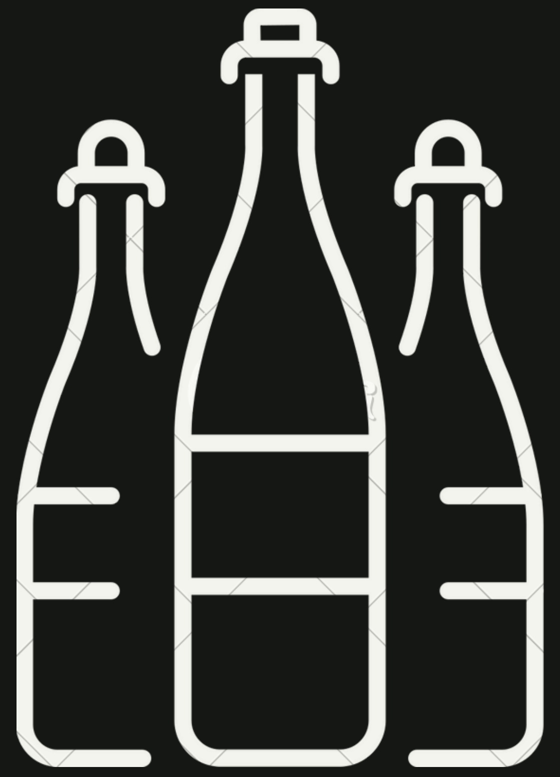
I want wine with tropical fruit, broom, brimstone and dried herb aroma without overly expressive palate. I also like unripened apple, citrus and dried sage alongside brisk acidity.



Which one?

# WINE RECOMMENDATION: MODEL PREPARATION

1. Extract key words and remove stop words: Rake
2. Tag words and keep n. & adj. words only: nltk.pos_tag
3. Use vector to represent sentence: Doc2Vec
4. Train the model using all descriptions
5. Give Recommendations based on user's description



# WINE RECOMMENDATION: RESULTS

**Input test sentence:**

Aromas include tropical fruit, broom brimstone and dried herb. The palate is not overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity

**Top 3 similar description:**

- Grass, herb and passion-fruit aromas are followed by citrus and tropical flavors. It's pleasant but the concentration seems lacking. (similarity score: 0.445)
- Light aromas of pineapple and other tropical fruit are accented by herb, floral and citrus flavors. The concentration is very light. (similarity score: 0.436)
- It resembles it in many ways, offering concentrated tropical and citrus fruit flavors, highlighted by brisk acidity and wrapped into a creamy texture. (similarity score: 0.409)

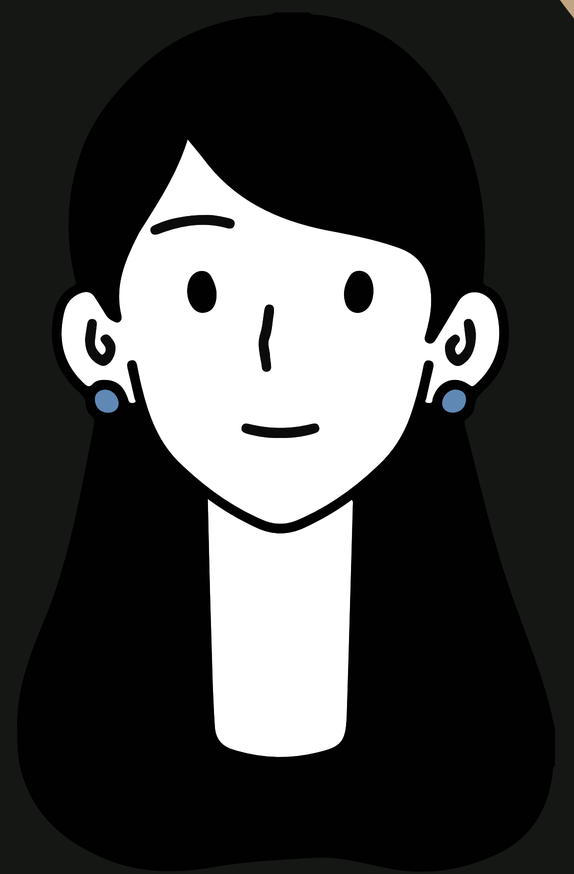
14 Hands The Reserve Sauvignon Blanc (2014)

Washington Hills Sauvignon Blanc (2015)

Talbott Logan Chardonnay (2011)

# WINE RECOMMENDATION BASED ON CONTENT

I usually drink Red Blend wines, but I would like to try something new. Could you recommend me some wines with similar taste?



# WINE RECOMMENDATION: SETTING UP

| Variety | # of Descriptions |
| --- | --- |
| Pinot Noir | 15503 |
| Chardonnay | 14439 |
| Cabernet Sauvignon | 12269 |
| Red Blend | 10317 |
| Sauvignon Blanc | 6549 |
| ...... | ...... |
| Roditis-Moschofilero | 1 |
| Centesimino | 1 |

**Group 1:**
Variety that have >1 descriptions
(589 elements)

**Group 2:**
Variety that have =1 description
(146 elements)

| Variety | Common Words in Description |
| --- | --- |
| Pinot Noir | pinot noir, black cherry, cherry fruit... |
| Chardonnay | buttered toast, tropical fruit, fruit flavors... |
| Cabernet Sauvignon | black berry, black current... |
| Red Blend | cabernet sauvignon, black berry... |
| Sauvignon Blanc | passion fruit, tropical fruit... |
| ...... | ...... |

**Common Words**

# WINE RECOMMENDATION: RESULTS

**Input (grape variety):** Red Blend

**Output (Top 5 recommended grape varieties):**

| Recommended Grape Varieties | Similarity Score | Top Common Words |
|---|---|---|
| Sangiovese | 0.833785 | black cherry, lead nose, grained tannins, blue flower... |
| Barbera | 0.804758 | black cherry, barbera alba, fruit flavors, skinned berry... |
| Aglianico | 0.778970 | black cherry, black fruit, blue flower, black pepper... |
| Cabernet Sauvignon-Merlot | 0.775810 | cabernet sauvignon, sauvignon merlot, black cherry, merlot blend... |
| Cabernet Franc | 0.757077 | cabernet franc, black cherry, fruit flavors, cherry flavors... |

Recommend **Top 5 types of wines** the customer may want to try most based on his/her current favorite grape variety

# Challenges

Processing Time (Complex Models)  / Large Data Set

Unbalanced Data (Sentiment & Price Tiers)

Limited / Redundant Features

# Future Extensions

Incorporate more features into models (such as cost of wine)

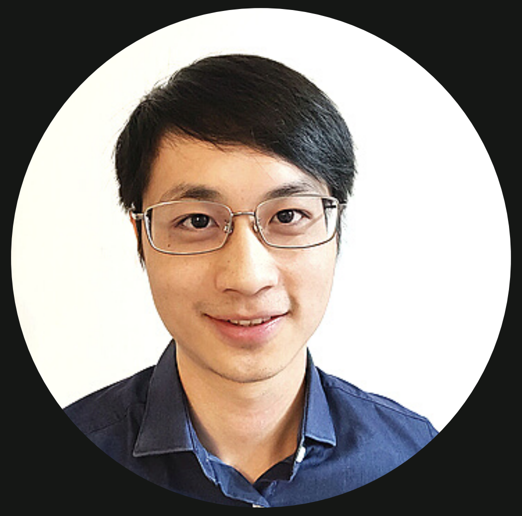
Build dashboard incorporating predictions & recommendations to easily analyze new wines

# Team Bio

## THE PEOPLE BEHIND THIS

Anna
Willman

Chenchen
Shentu



Fan
Yang

Olivia
Yang

Wilson
McDermott

## THANK YOU!

# *Questions?*