Wine Recommender

Lee Wan Xian

Agenda

01 Introduction

Background
Problem Statement

02 Exploratory Data Analysis

Overview of Datasets
Data Cleaning
EDA

03 Preprocessing

Wine Traits embedding & encoding from reviews

04 Modelling

Apply algorithms from Scikit-Surprise
Model Evaluation

05 Conclusion

Recommendation
Future Work
Application Experience

01

Introduction

Background | Problem Statement

Background

- More than 10,000 varieties of wine globally
- As a consumer, draining task to pick the right wine
- Retailers worried that the negative experience would be bad for business

Problem Statement

- Client: A wine retailer

- Develop a wine recommender system that gives suggestions cater to consumer's needs
  - Mitigate decision fatigue burden

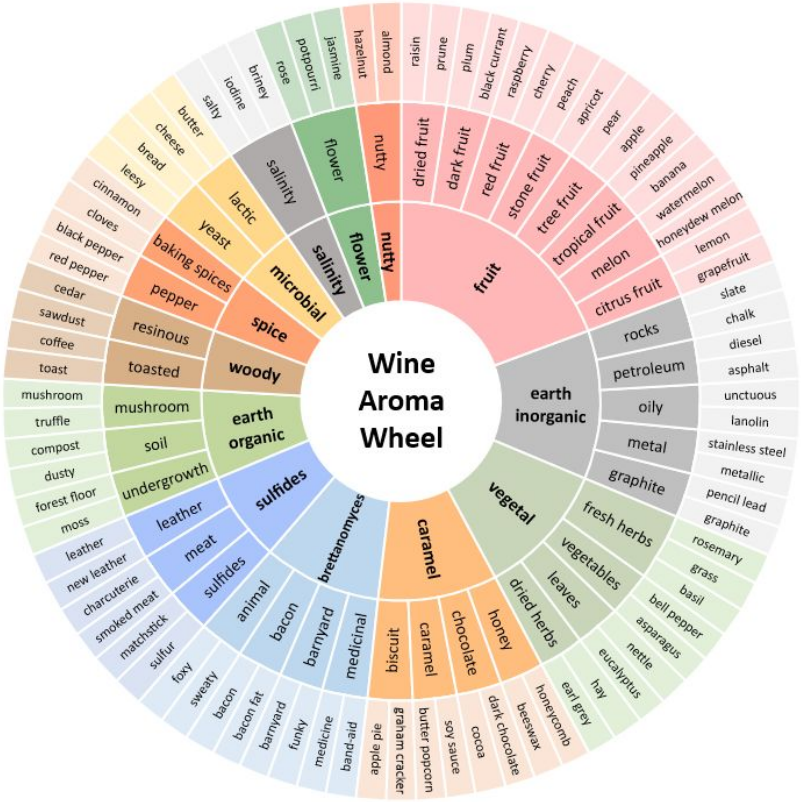
02

Exploratory Data Analysis

Datasets Overview | Data Cleaning | EDA

Datasets

- Wine reviews that were scraped from *WineEnthusiast* webpage
- List of standardized wine traits descriptors
  - Derived from *RoboSomm wine wheels*



Datasets

Wine ratings, given by wine tasters, follow the 100-points scale

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

Data Cleaning

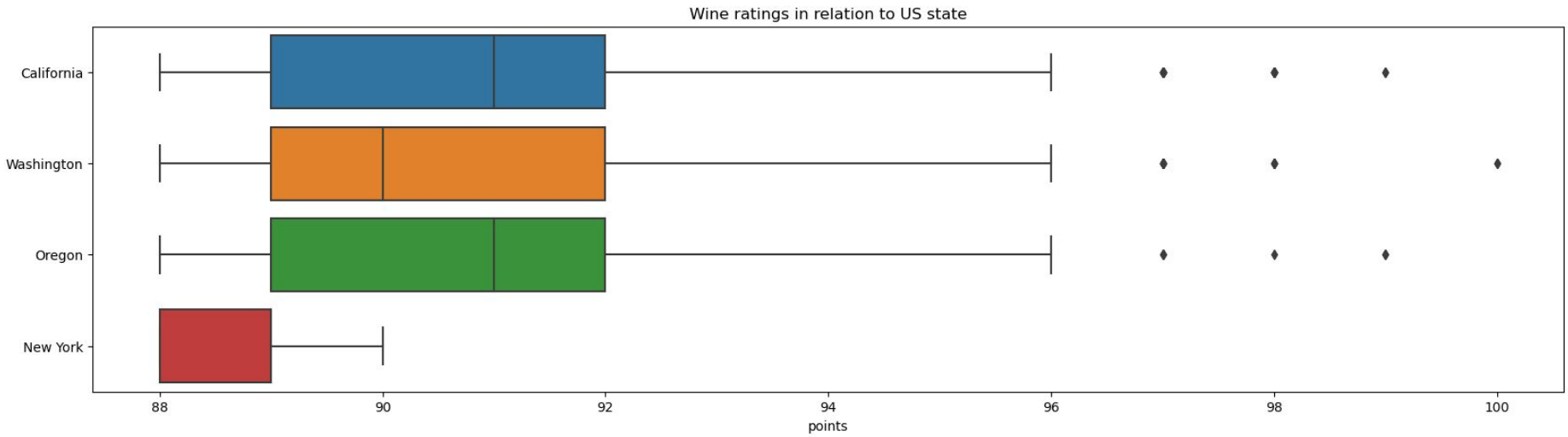
Missing values

- Wine Reviews = Removed
- Wine Traits Descriptor = No issue

Duplicate records

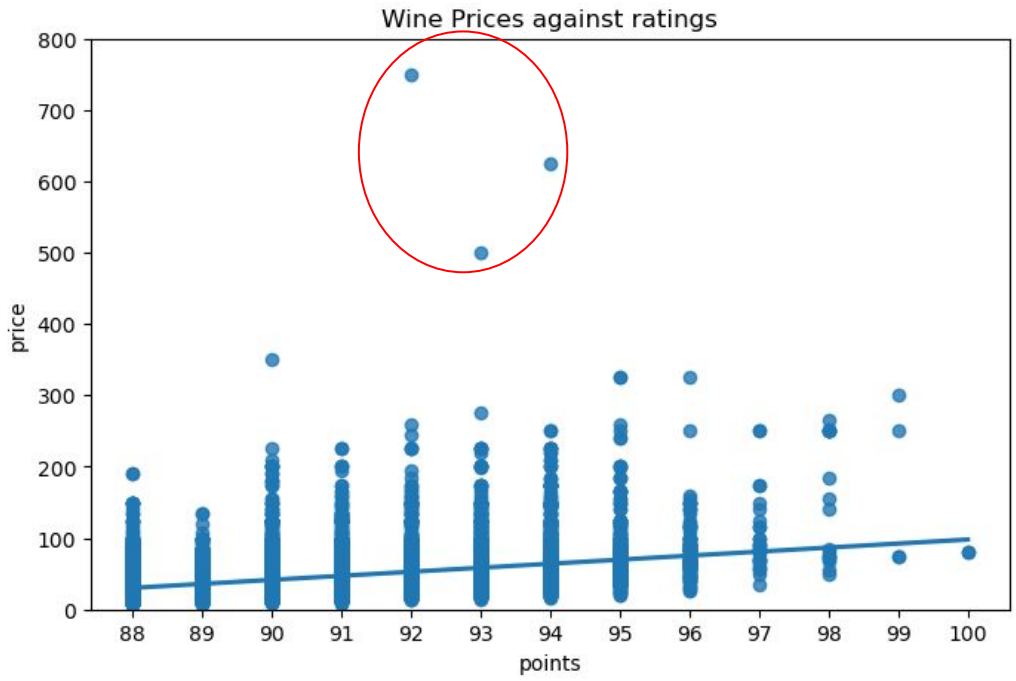
- Wine Reviews = Left untouched
  - Small volume (8% of total dataset)
- Wine Traits Descriptor = No issue

EDA



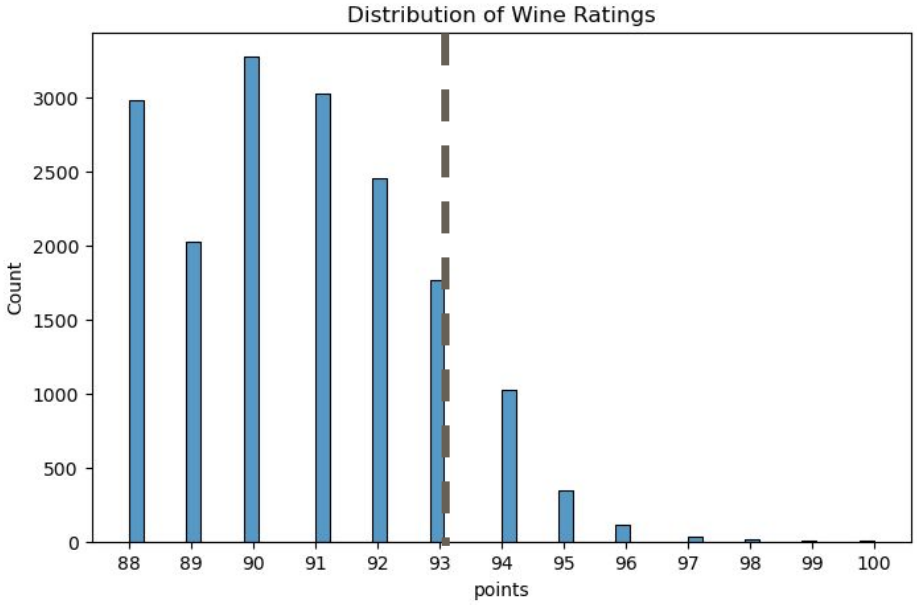
- New York Wines fared poorly in comparison to wines from other states
- Other states wines fared similarly to each other

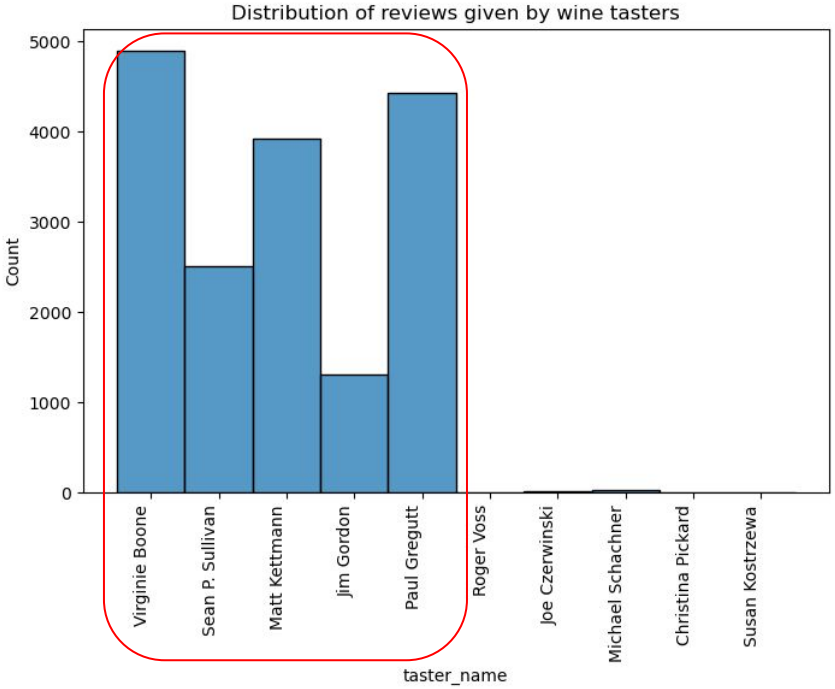
EDA



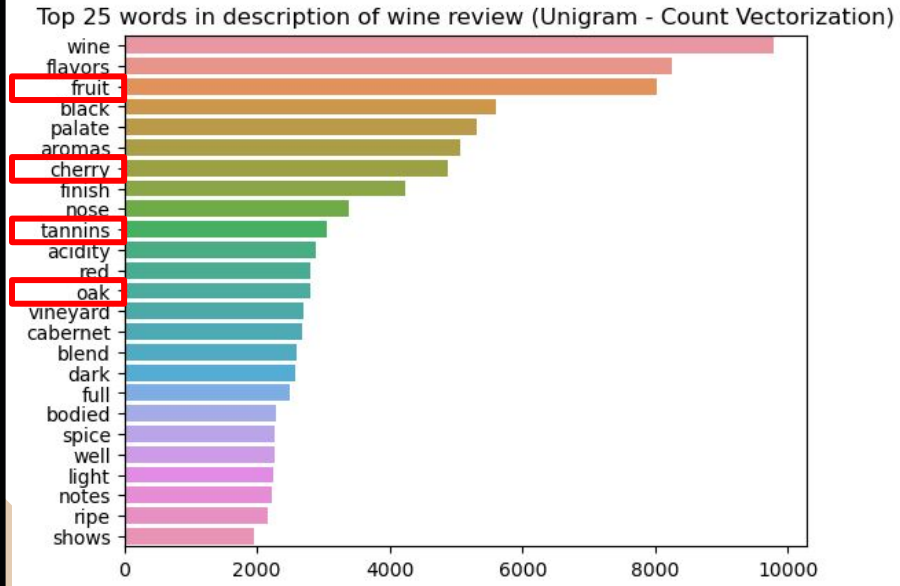
- Weak Positive correlation (0.354)
- Outliers where expensive wines are rated below cheaper ones

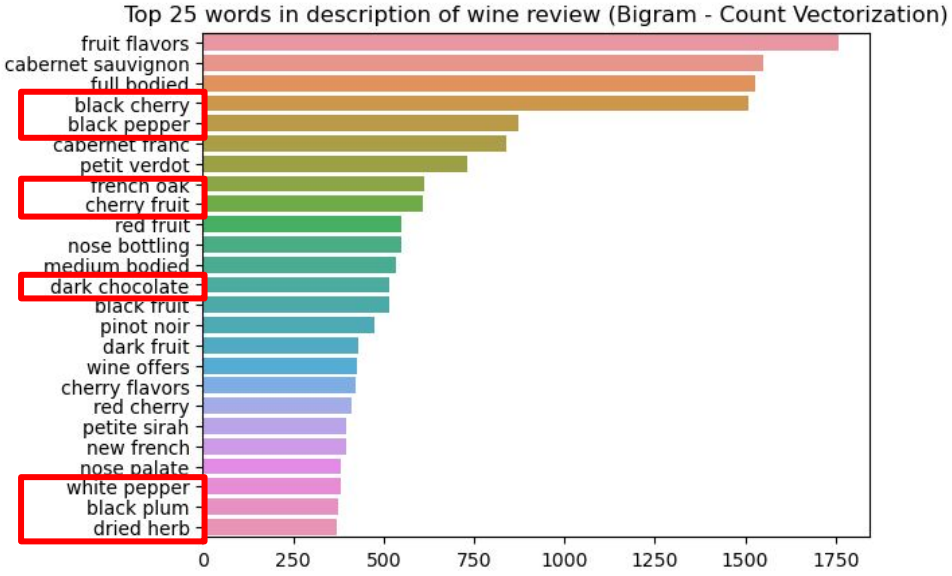
EDA


Distribution of Wine Ratings
Count
points
0
500
1000
1500
2000
2500
3000
88
89
90
91
92
93
94
95
96
97
98
99
100


Distribution of reviews given by wine tasters
Count
taster_name
0
1000
2000
3000
4000
5000
Virginie Boone
Sean P. Sullivan
Matt Kettmann
Jim Gordon
Paul Gregutt
Roger Voss
Joe Czerwinski
Michael Schachner
Christina Pickard
Susan Kostrzewa

EDA

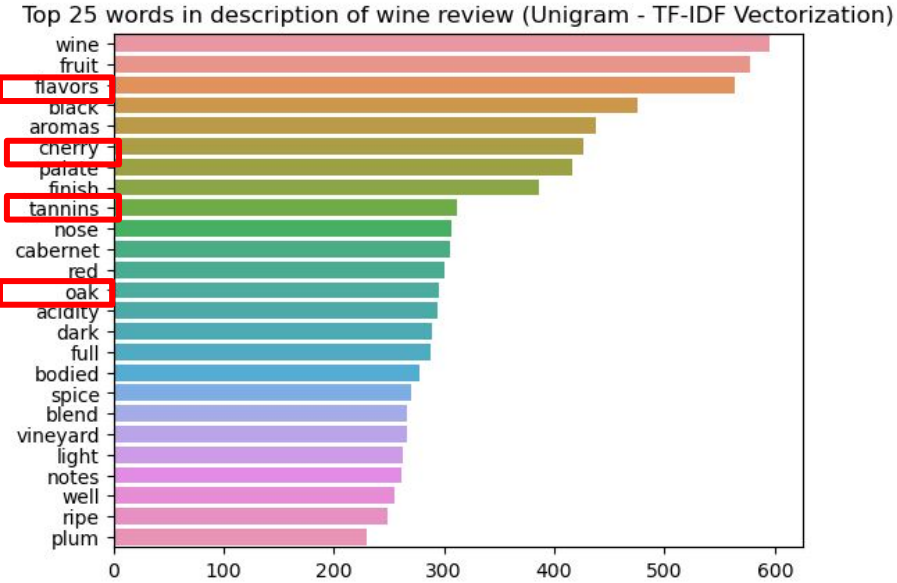


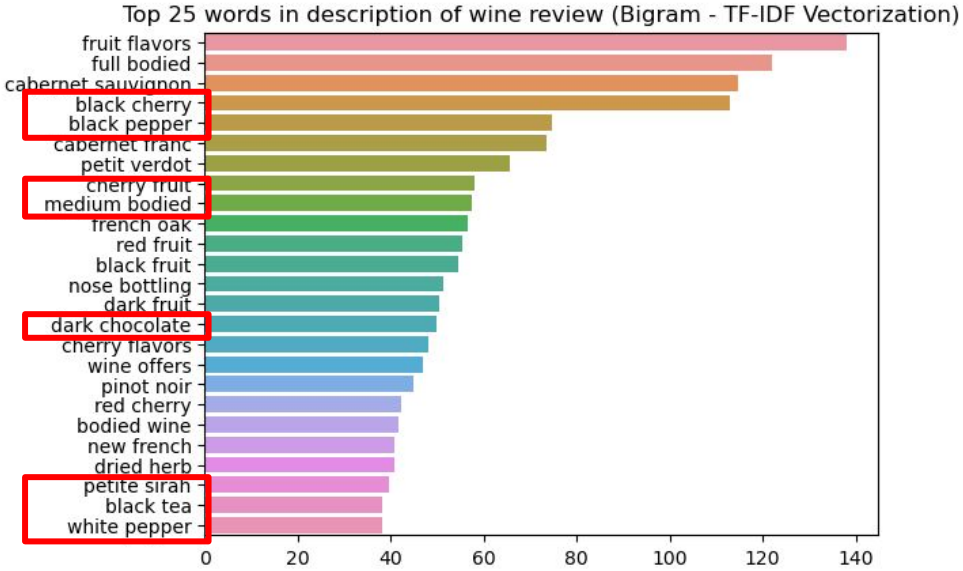


Some key wine traits are present in most of the review descriptions

- Top 25 words in terms of Count

EDA





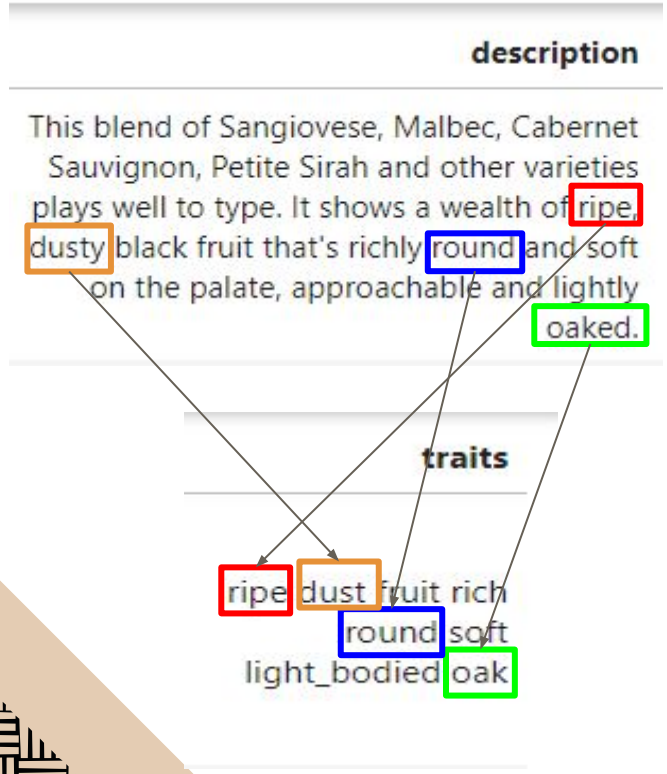
- Similar wine traits are in the top 25 words when accounting for Term Frequency-Inverse Document Frequency

03

Preprocessing

Wine Traits embedding & encoding from reviews

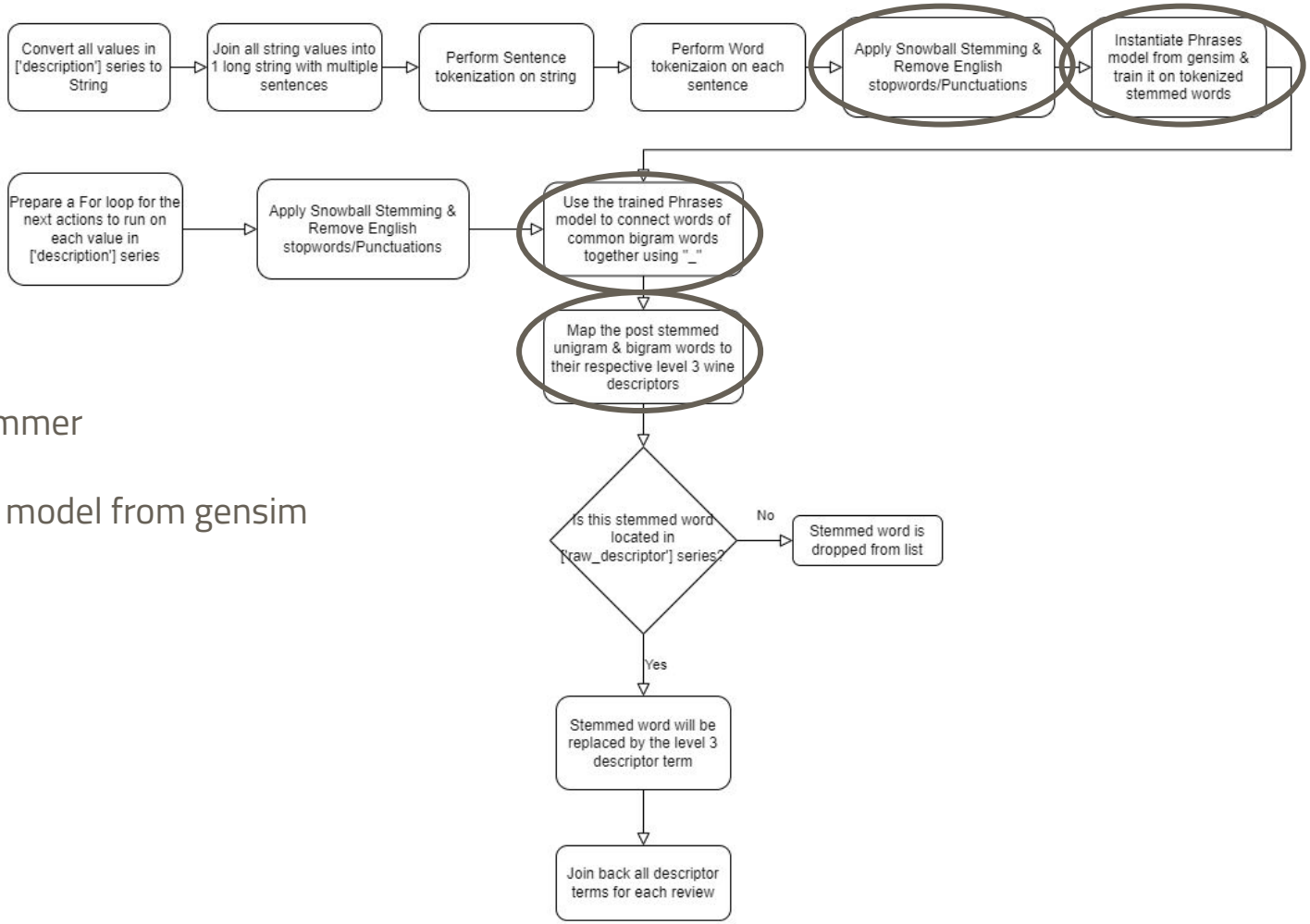
Preprocessing



Wine Traits embedding

- Extract out key words in review descriptions
- Map raw words into level 3 wine trait descriptors (RoboSomm Wine Wheels)
- Append the wine traits terms into a string under a new column

Preprocessing



- Snowball Stemmer
- Train Phrases model from gensim

Preprocessing

description	desc_wd_count	traits
This blend of Sangiovese, Malbec, Cabernet Sauvignon, Petite Sirah and other varieties plays well to type. It shows a wealth of ripe, dusty black fruit that's richly round and soft on the palate, approachable and lightly oaked.	37	ripe dust fruit rich round soft light_bodied oak

- Scope down the wine traits terms to the top 150
  - Reduce from 653 trait terms
- Perform binary encoding to set the wine traits as a filter criteria

Preprocessing

[27]:

	taster_name	title	points	variety	designation	winery	country	province	region_1	region_2	...	vanilla	velvety	vibrant	violet	warm	weight	wet_rocks	white	white_pepper	wood
0	Virginie Boone	Ferrari-Carano 2014 Siena Red (Sonoma County)	88	Red Blend	Siena	Ferrari-Carano	US	California	Sonoma County	Sonoma	...	0	0	0	0	0	0	0	0	0	0

- If specific wine traits are relevant to the wine, the integer under the wine trait term will show as 1
- Otherwise, it will show as 0

04

Modelling

Apply various algorithms from Scikit-Surprise
Model Evaluation

Modelling

surpr!se

A Python scikit for recommender systems.

- A Python scikit library
  - Build recommender systems that deal with explicit rating data.
- Recommender Algorithms available
  - Basic (Normal, Baseline)
  - k-NN based (KNN Basic, KNN with Means, KNN with ZScore, KNN Baseline)
  - Matrix Factorization based (SVD, NMF)
  - Slope One
  - CoClustering

Modelling

$$\text{RMSE} = \sqrt{\frac{1}{|\hat{R}|} \sum_{\hat{r}_{ui} \in \hat{R}} (r_{ui} - \hat{r}_{ui})^2}.$$

- Root Mean Squared Error (RMSE) is calculated based on comparing the estimated rating to the actual rating for each taster-wine pair
- The lower the RMSE, the better the model performs

Modelling

$$\text{Precision@k} = \frac{|\{\text{Recommended items that are relevant}\}|}{|\{\text{Recommended items}\}|}$$

- A wine is considered relevant if actual rating > threshold.
- A wine is considered recommended if
  a. Estimated rating > Threshold
  b. Among the *k* highest estimated ratings (i.e. Top 10 recommendations if *k*=10)
- The rating threshold for relevant wines is set at 90
  - Closest point to the median score

Modelling

Model	Root Mean Squared Error (RMSE)	Precision@k
Pure Randomized Recommender (Baseline)	n/a	0.36
KNN Baseline (Tuned)	**1.828586**	**0.751706**
Baseline Predictor	1.921808	0.751706
KNN Basic	1.853119	0.751706
KNN Means	1.853373	0.751706
KNN ZScore	1.853372	0.751706
KNN Baseline	1.828589	0.751706
Slope One	1.853719	0.751706
Co-clustering	1.858452	0.749706
SVD	1.830108	0.659683
Normal Predictor	2.67958	0.605464
NonNegative Matrix Factorization	2.123324	0.582563

Modelling

Recall the rating threshold for relevant wines is set at 90 (Closest point to median)

Threshold for Relevant wines	89	90	91
Precision@k	0.892143	0.751706	0.633214

- When tuning the rating threshold for relevant wines
  - Threshold = 89, model becomes too lenient
  - Threshold = 91, model becomes too strict

05

Conclusion

Recommendation | Future Work | Streamlit App Experience

Conclusion & Recommendation

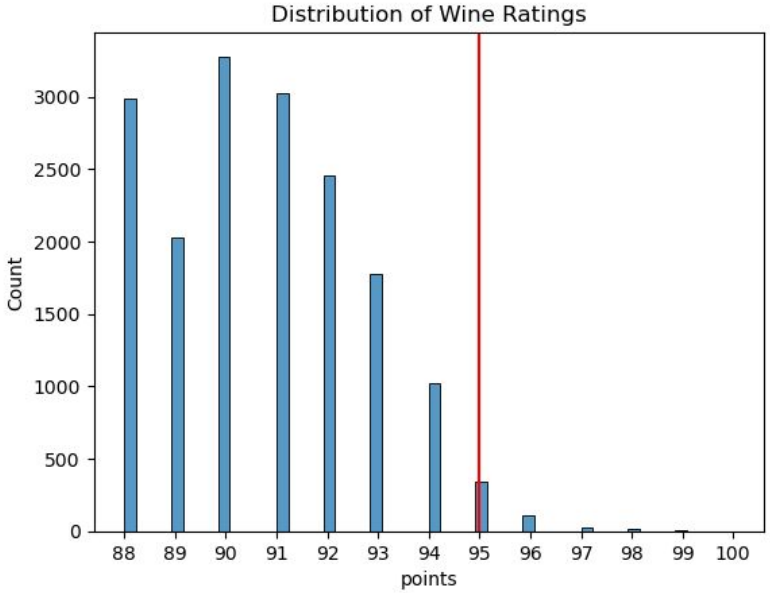
Best Recommender System Model

- KNN Baseline (Tuned)
  - Maximum no. of neighbors to account for aggregation = 35
  - Cosine Similarity Measure

- Precision@k = 0.751706 (Highest)
- RMSE = 1.828589 (Lowest)

Conclusion & Recommendation

Name	Estimated Rating	Actual Rating
Charles Smith 2006 Royal City Syrah (Columbia Valley (WA))	91.557454	100
Cayuse 2008 Bionic Frog Syrah (Walla Walla Valley (WA))	91.557454	100
K Vintners 2013 The Hidden Northridge Vineyard Syrah (Wahluke Slope)	91.523150	95
Cayuse 2011 En Chamberlin Vineyard Syrah (Walla Walla Valley (OR))	91.466545	99
Cayuse 2009 En Chamberlin Vineyard Syrah (Walla Walla Valley (OR))	91.466545	99
Alpha Omega 2012 Stagecoach Vineyard Cabernet Sauvignon (Atlas Peak)	91.448946	99
Alpha Omega 2012 ERA Red (Napa Valley)	91.448946	99
Doyenne 2008 Grand Ciel Vineyard Syrah (Red Mountain)	91.429456	95
Williams Selyem 2012 Eastside Road Neighbors Pinot Noir (Russian River Valley)	91.397192	95
Dutton-Goldfield 2013 Dutton Ranch Cherry Ridge Vineyard Syrah (Russian River Valley)	91.397192	95

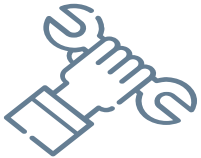


60% of recommendations are in the highest tier of quality

Future Work

- Collect more data & update database regularly
  - New wines (Items)
  - Ratings from Tasters (Users)

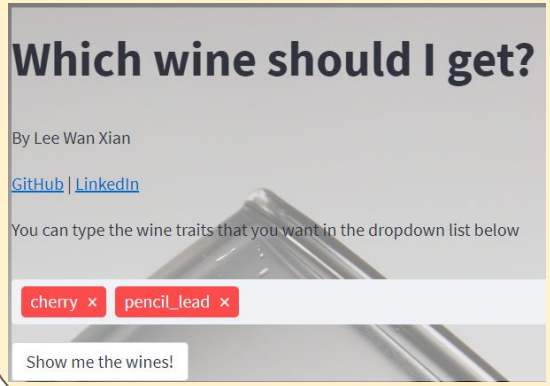
- To re-train the wine recommender regularly with new data



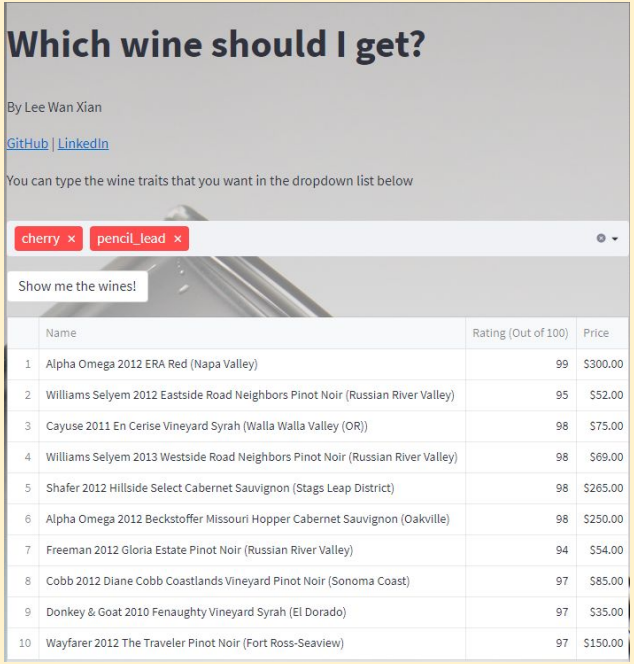
App Workflow

User Inputs
Select wine traits from the dropdown list or leave it blank if indifferent

Process
Click “Show me the wines!”



Output



	Name	Rating (Out of 100)	Price
1	Alpha Omega 2012 ERA Red (Napa Valley)	99	$300.00
2	Williams Selyem 2012 Eastside Road Neighbors Pinot Noir (Russian River Valley)	95	$52.00
3	Cayuse 2011 En Cerise Vineyard Syrah (Walla Walla Valley (OR))	98	$75.00
4	Williams Selyem 2013 Westside Road Neighbors Pinot Noir (Russian River Valley)	98	$69.00
5	Shafer 2012 Hillside Select Cabernet Sauvignon (Stags Leap District)	98	$265.00
6	Alpha Omega 2012 Beckstoffer Missouri Hopper Cabernet Sauvignon (Oakville)	98	$250.00
7	Freeman 2012 Gloria Estate Pinot Noir (Russian River Valley)	94	$54.00
8	Cobb 2012 Diane Cobb Coastlands Vineyard Pinot Noir (Sonoma Coast)	97	$85.00
9	Donkey & Goat 2010 Fenaughty Vineyard Syrah (El Dorado)	97	$35.00
10	Wayfarer 2012 The Traveler Pinot Noir (Fort Ross-Seaview)	97	$150.00

Streamlit App

https://leewanxian-wine-recommender-app-44l3c7.streamlit.app/

Thanks!

Any Questions?

https://www.linkedin.com/in/wanxianlee/

https://github.com/leewanxian