Wine Review Classification Analysis

By: Meichan, Nicholas, Eryka

IST 644: Natural Language Processing

Project Summary

Introduction

Data source

Data preprocessing

Data descriptive analysis

3 classification models based on the accuracy scores

Data Source

150k reviews from Wine Enthusiast magazine from Kaggle

10 attributes

	country	description	designation	points	price	province	region_1	region_2	variety	winery
0	Portugal	This is a deliciously creamy wine with light w...	Assobio Branco	87	14.0	Douro	NaN	NaN	Portuguese White	Quinta dos Murças
1	US	Black plum juice, black pepper, caramel and sm...	NaN	87	25.0	California	Paso Robles	Central Coast	Cabernet Sauvignon	Western Slope
2	Georgia	Aromas of green apple and white flowers prepar...	NaN	87	14.0	Lechkhumi	NaN	NaN	Tsolikouri	Teliani Valley
3	Kosovo	This wine has aromas of black berry, dried red...	NaN	87	13.0	Rahoveci Valley	NaN	NaN	Shiraz	Stone Castle
4	Italy	A blend of organically cultivated Groppello, M...	San'Emiliano Chiaretto	87	13.0	Lombardy	Valtènesi	NaN	Rosato	Pratello

Project Research Goals :

(1) How accurate can we use wine reviews to predict their quality (points)?

(2) How accurate can we use wine reviews to predict their price?

(3) How accurate can we use wine reviews to predict their varieties?

Data Preprocessing

1| Import data as pandas data frame

2| Examine data types

3| Handle missing values (remove NAs and impute NAs with mode):

- country
- province
- prices

4| Preprocess the textual data of "Review":

- POS tagging to extract nouns, adjectives, noun phrases, and adjective phrases
- Tokenize, clean, and extract bag of words and bigrams

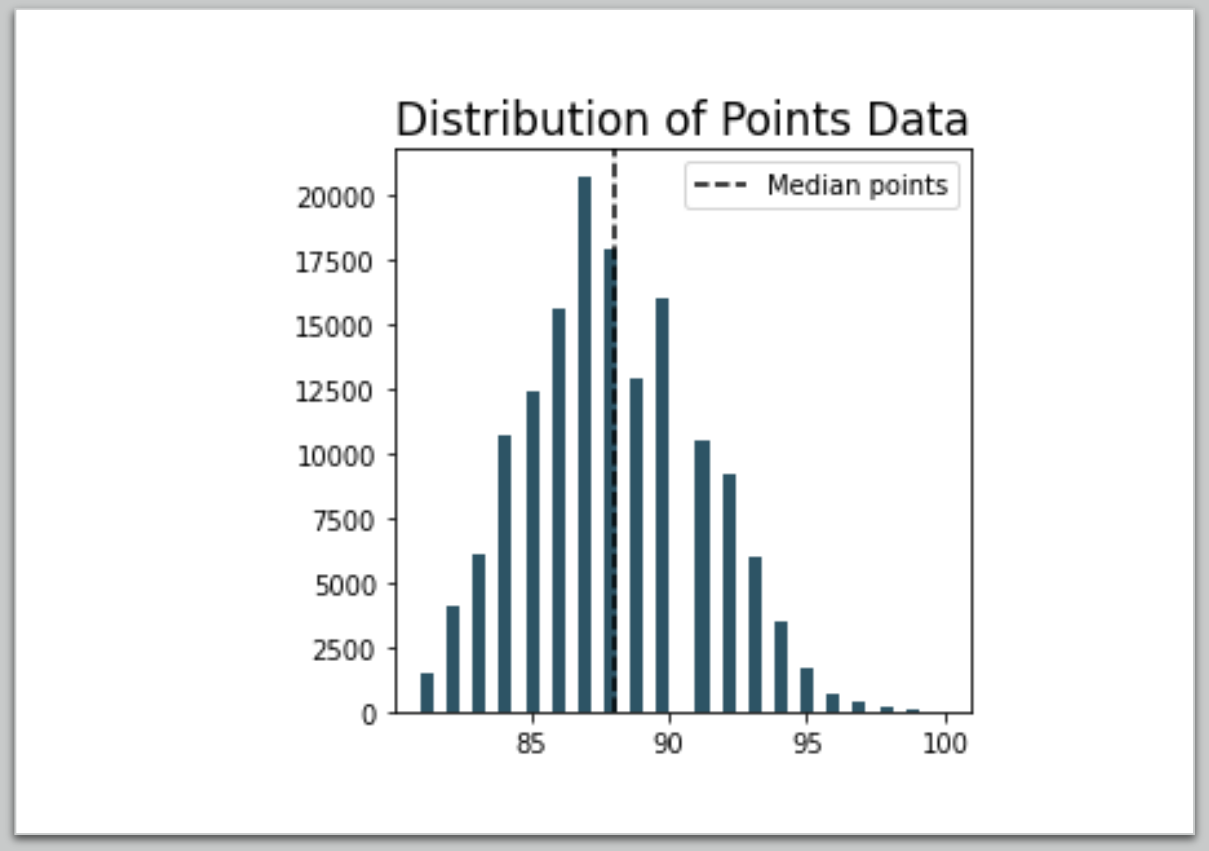
5| Categorize the price and points into "Price_val" and "Quality"

Final Dataset for Classification Tasks

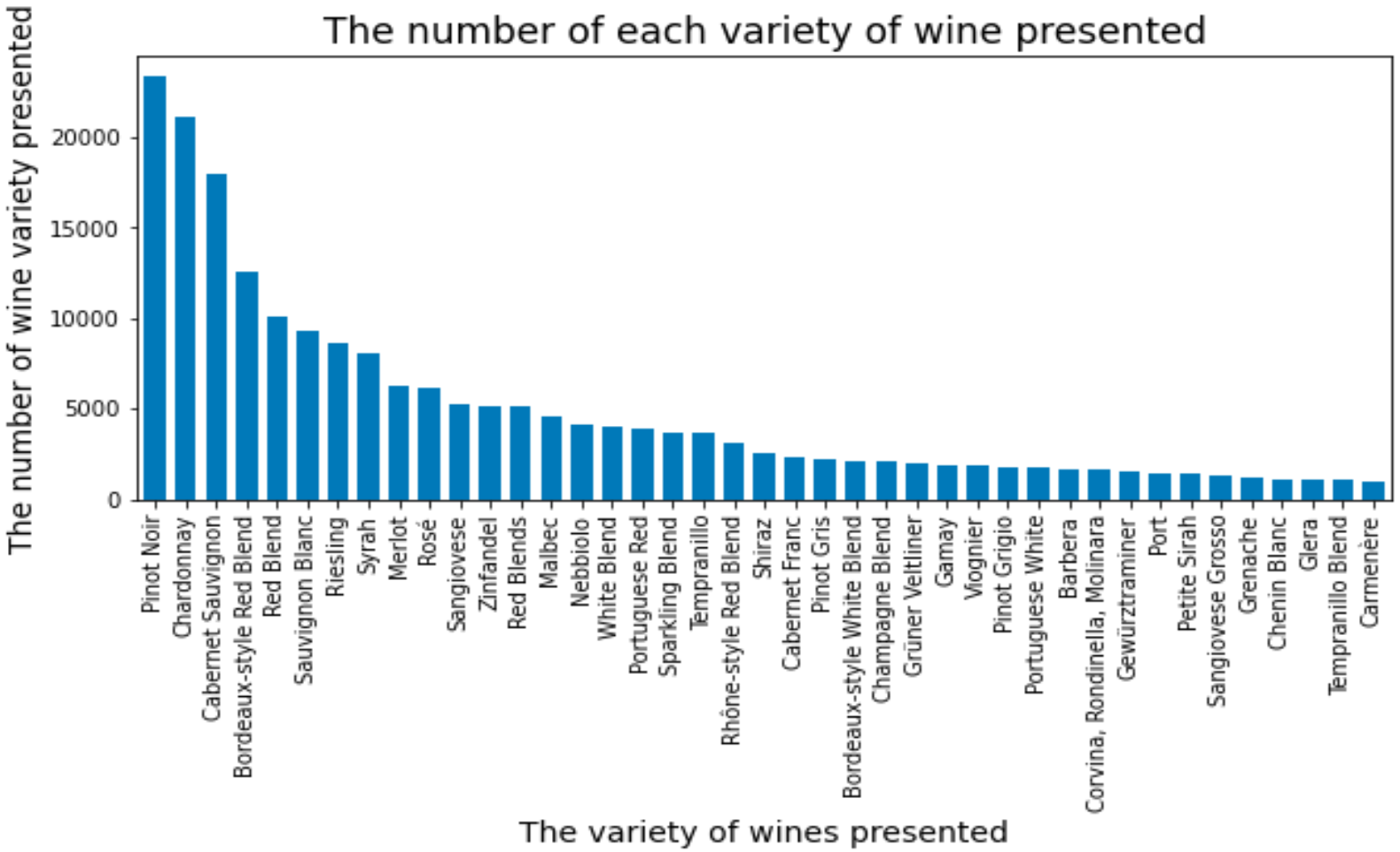
	Country	Review	Points	Price	Variety	TaggedReview	nouns	adjectives	word_token_cleaned	adjphrases	nounphrases	Quality	Price_val
0	Spain	ripe aromas of fig, blackberry and cassis are ...	96	110.0	Tinta de Toro	[('ripe', 'JJ'), ('aromas', 'NN'), ('of', 'IN'...	[aromas, fig, blackberry, cassis, slathering, ...	[ripe, oaky, full, intense, rich, black, heady]	[ripe, aromas, fig, blackberry, cassis, soften...	['tremendously delicious ']	['mac ', 'watson ', 'the memory ', 'a wine ', ...	Excellent	Above 100
1	US	mac watson honors the memory of a wine once ma...	96	90.0	Sauvignon Blanc	[('mac', 'NN'), ('watson', 'NN'), ('honors', '...	[mac, watson, memory, wine, mother, gold, colo...	[delicious, balanced, complex, white, dark, pe...	[mac, watson, honors, memory, made, mother, tr...	[]	['% ', 'new french oak ', 'fruit ', 'ponzi ', ...	Excellent	50-100
2	US	this spent 20 months in 30% new french oak, an...	96	65.0	Pinot Noir	[('this', 'DT'), ('spent', 'VBD'), ('20', 'CD'...	[months, %, oak, fruit, ponzi, aurora, abetina...	[new, french, aromatic, dense, toasty, aromas,...	[spent, months, new, french, oak, incorporates...	[]	['the top wine ', 'la bégude ', 'point ', 'the...	Excellent	50-100
3	France	this is the top wine from la bégude, named aft...	95	66.0	Provence red blend	[('this', 'DT'), ('is', 'VBZ'), ('the', 'DT'),...	[wine, bégude, point, vineyard, feet, structur...	[top, la, highest, considerable, extra, tari]	[top, la, bégude, named, highest, point, viney...	[]	['dense ', 'pure ', 'the opening ', 'bell ', '...	Great	50-100

Descriptive Results

Price and Points



Descriptive Results: Varieties



Descriptive Results

Token and Bigram Frequencies

20 Most Frequent Tokens

There were a total of 3,441,539 unigram tokens after removing all punctuations, numbers, and a list of stop words.

```
[('fruit', 56505),
 ('finish', 37724),
 ('aromas', 35820),
 ('acidity', 32602),
 ('tannins', 32181),
 ('cherry', 30659),
 ('palate', 29008),
 ('ripe', 26720),
 ('black', 24590),
 ('dry', 22978),
 ('spice', 22643),
 ('sweet', 21286),
 ('rich', 21172),
 ('oak', 19675),
 ('notes', 19606),
 ('red', 19187),
 ('soft', 17745),
 ('fresh', 17666),
 ('good', 17291),
 ('berry', 17098)]
```

20 Most Frequent Bigrams

```
(('black', 'cherry'), 0.00180762153211107)
(('cabernet', 'sauvignon'), 0.0013726417163949037)
(('pinot', 'noir'), 0.0012305541212812058)
(('black', 'fruit'), 0.0009431826865829502)
(('cherry', 'fruit'), 0.0007865667075107967)
(('sauvignon', 'blanc'), 0.0007836610307191056)
(('black', 'currant'), 0.0007615778871022528)
(('crisp', 'acidity'), 0.0007429815556354294)
(('ripe', 'fruit'), 0.0007206078443394075)
(('berry', 'fruit'), 0.0006947473208933562)
(('tropical', 'fruit'), 0.0006656905529764445)
(('red', 'fruit'), 0.0006642377145805989)
(('firm', 'tannins'), 0.000636924352738702)
(('green', 'apple'), 0.0006357620820220256)
(('black', 'pepper'), 0.000611644964650989)
(('barrel', 'sample'), 0.0005761957077923569)
(('stone', 'fruit'), 0.0005517880227421511)
(('red', 'berry'), 0.0005294143114461292)
(('blackberry', 'cherry'), 0.0005143047921293351)
(('long', 'finish'), 0.0005122708183751514)
```

Descriptive Results

Frequencies Based on POS

```
[('wine', 88260),
 ('flavors', 77815),
 ('fruit', 56053),
 ('finish', 33140),
 ('acidity', 32557),
 ('aromas', 32195),
 ('tannins', 31111),
 ('palate', 28440),
 ('cherry', 26067),
 ('spice', 22030),
 ('notes', 19508),
 ('oak', 18153),
 ('berry', 16106),
 ('%', 15768),
 ('blackberry', 14217),
 ('blend', 13285),
 ('vanilla', 13209),
 ('plum', 13032),
 ('years', 12759),
 ('fruits', 12147)]
```

There were a total of 1,980,620 tokens of nouns

20 Most Frequent Nouns

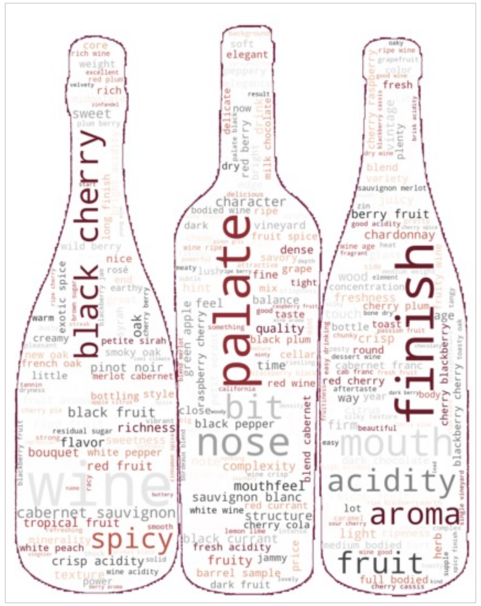
```
[('black', 24590),
 ('dry', 21426),
 ('rich', 20957),
 ('red', 18328),
 ('soft', 17710),
 ('fresh', 17511),
 ('good', 17258),
 ('ripe', 17069),
 ('sweet', 16854),
 ('white', 11852),
 ('green', 10642),
 ('full', 9017),
 ('bright', 8884),
 ('dark', 7756),
 ('clean', 7709),
 ('light', 7656),
 ('tannic', 7508),
 ('fine', 7370),
 ('more', 7210),
 ('great', 6879)]
```

There were a total of 981,885 tokens of adjectives

20 Most Frequent Adjectives

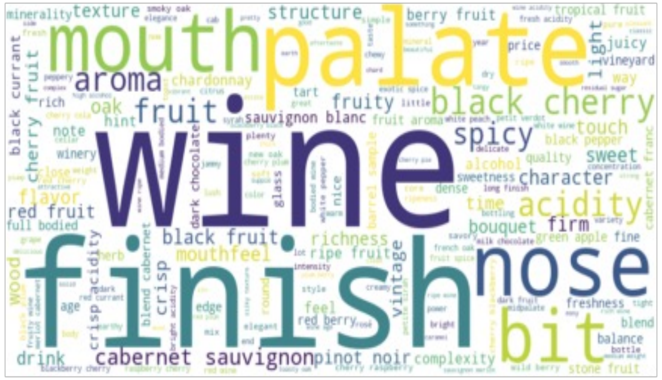
Word Clouds

Wine Reviews



Noun Phrases

Adjective Phrases



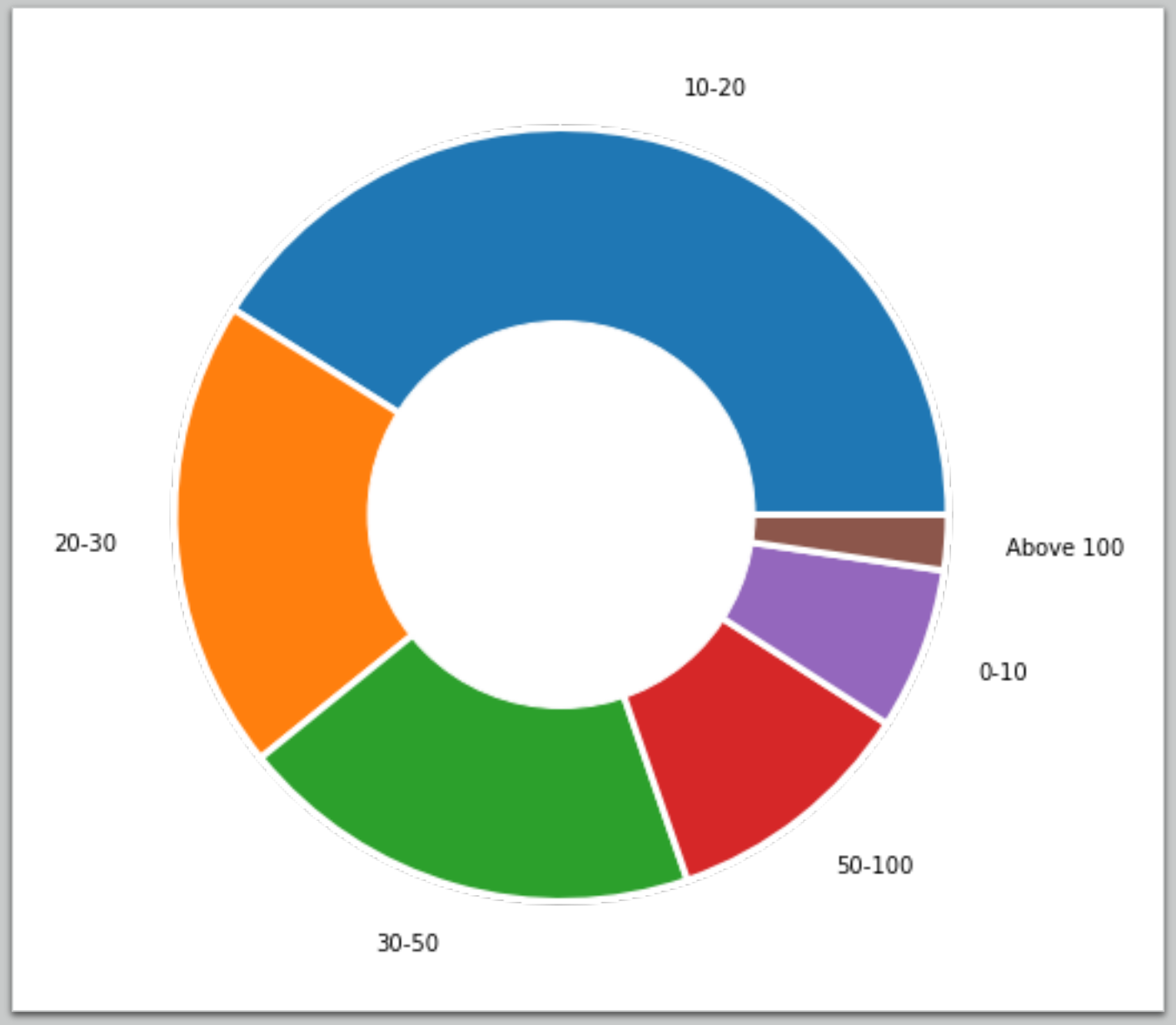
Categorized price and points

	COUNTRY	VARIETY	ADJECTIVES	QUALITY	PRICE_VAL
0	Spain	Tinta de Toro	[ripe, oaky, full, intense, rich, black, heady]	Excellent	Above 100
1	US	Sauvignon Blanc	[delicious, balanced, complex, white, dark, pe...	Excellent	50-100
2	US	Pinot Noir	[new, french, aromatic, dense, toasty, aromas,...	Excellent	50-100
3	France	Provence red blend	[top, la, highest, considerable, extra, tari]	Great	50-100
4	Spain	Tinta de Toro	[dark, ripe, black, cool, feels, toasty, heady...	Great	50-100

Price categories of wines

10-20	61651
20-30	29610
30-50	29194
50-100	15940
0-10	10167
Above 100	3464

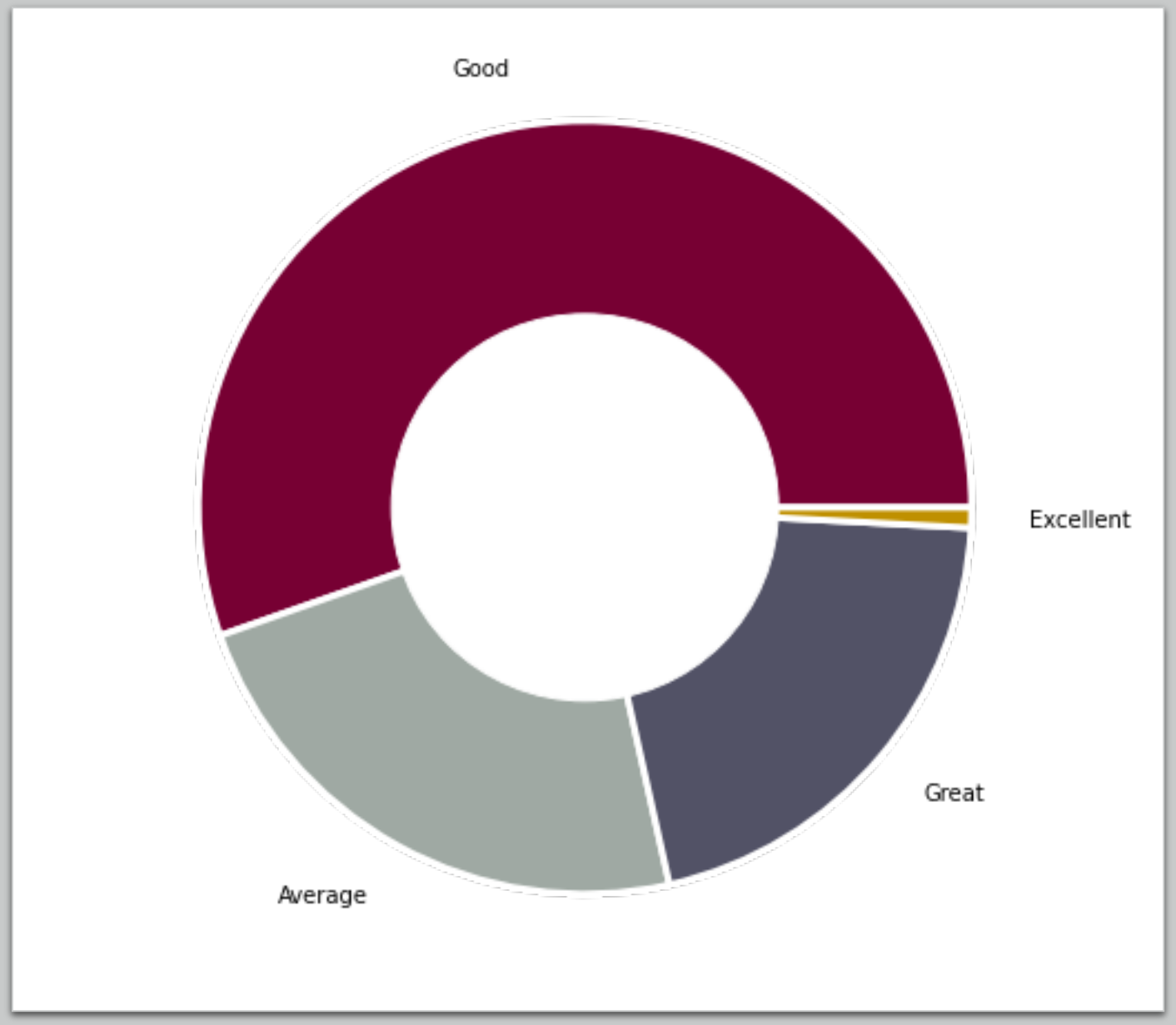
Name: Price, dtype: int64



Quality categories of wines

Good	83083
Average	34707
Great	30972
Excellent	1264

Name: Quality, **dtype:** int64



Naïve Bayes Classification

1| Created multiple training/test sets to run Naïve Bayes

2| Using tokenized data columns (tokenized words, adjectives, nouns)

3| Using TfidfVectorizer() for the features extraction

4| Model with different variables (price, points, varieties) to see what can the text variables best predict

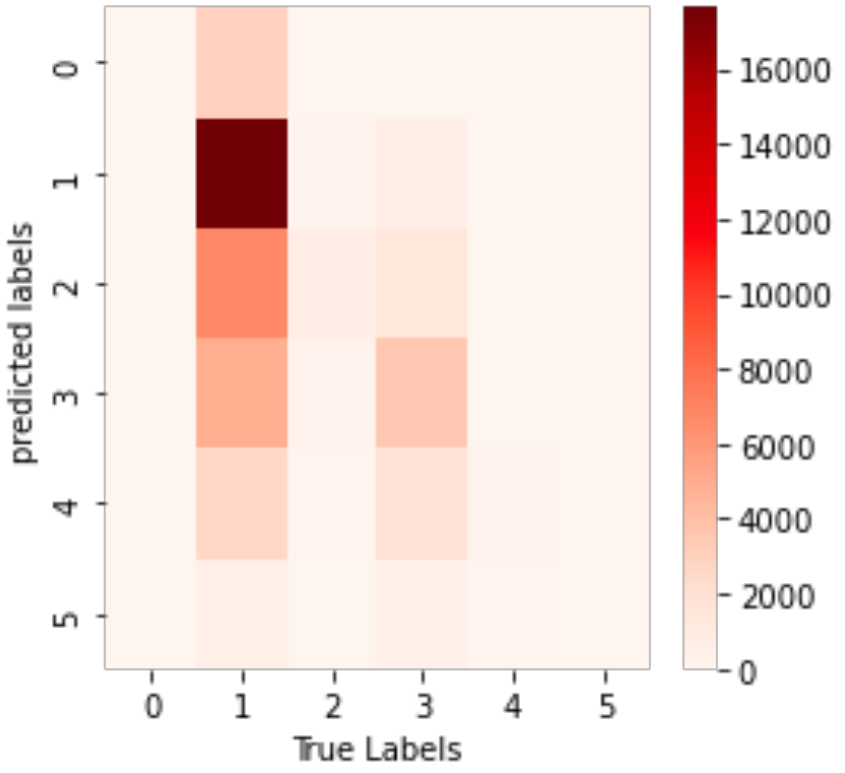
Wine quality predicted by reviews – WORST :(

- Model wine price variable with tokenized words.
- Tokenized words had the highest accuracy score of 49.33% , compared to adjectives and nouns.

```
The score of Mulitnomial Naive Bayes for predicting the price of wine with tokenized words only is 49.32900817632421 %
Model accuracy score for this second model is: 0.4933
/Users/meichanhuang/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-scor
e are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

	precision	recall	f1-score	support
0-10	0.83	0.00	0.01	3050
10-20	0.50	0.95	0.65	18496
20-30	0.56	0.08	0.14	8883
30-50	0.45	0.41	0.43	8758
50-100	0.61	0.06	0.12	4782
Above 100	0.00	0.00	0.00	1039
accuracy			0.49	45008
macro avg	0.49	0.25	0.22	45008
weighted avg	0.52	0.49	0.39	45008

```
array([[   10,  2979,    21,    40,     0,     0],
       [    0, 17636,   182,   639,    39,     0],
       [    0,  6738,   691,  1438,    16,     0],
       [    1,  4922,   234,  3555,    46,     0],
       [    0,  2566,    92,  1814,   310,     0],
       [    1,   579,    18,   345,    96,     0]])
```



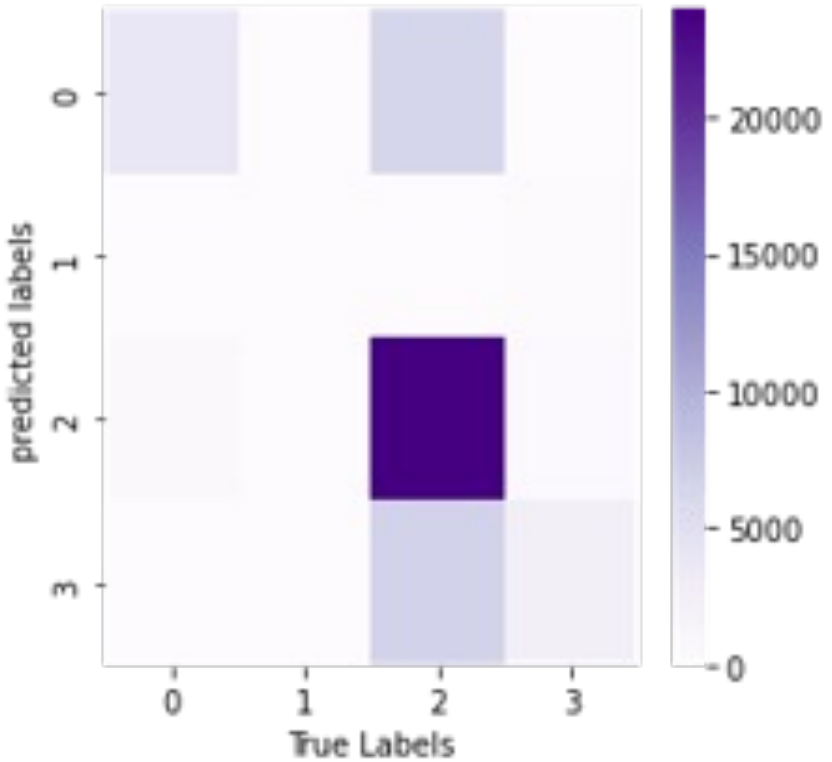
Wine quality predicted by reviews - MEDIOCRE :|

- Model wine quality variable with tokenized words
- Tokenized words performed again the best out of all three types of texts, with an accuracy score of 67.55%.

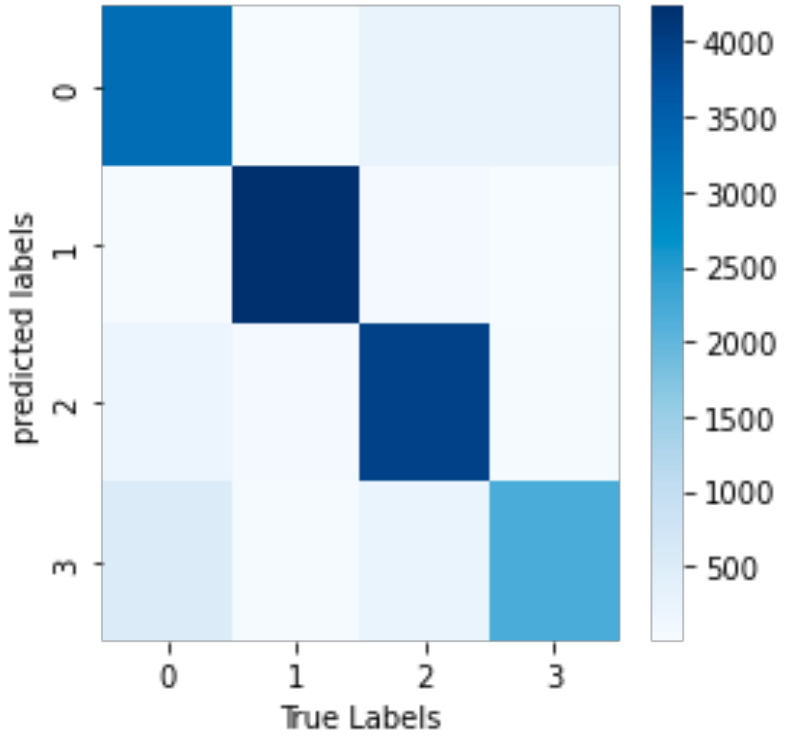
```
The accuracy score of Mulitnomial Naive Bayes model for the quality of wine with word tokens is 67.75017774617845 %
Model accuracy score with tokens to predict the quality of wine is: 0.6775
/Users/meichanhuang/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classification.py:1318: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

	precision	recall	f1-score	support
Average	0.87	0.39	0.54	10412
Excellent	0.00	0.00	0.00	379
Good	0.64	0.96	0.77	24925
Great	0.78	0.27	0.41	9292
accuracy			0.68	45008
macro avg	0.58	0.40	0.43	45008
weighted avg	0.72	0.68	0.63	45008

```
array([[ 4023,     0,  6374,    15],
       [    0,     0,   107,   272],
       [  584,     0, 23930,   411],
       [    3,     0,  6749,  2540]])
```



Wine variety predicted by reviews – BEST :)



- Model wine varieties (top four varieties only) with tokenized words
- Surprisingly (or not), Tokenized words performed again the best out of all three types of texts, with an accuracy score of 88.75%.

```
array([[3261,   17,  271,  264],
       [  22, 4235,   61,    5],
       [ 199,   58, 3981,   38],
       [ 537,   22,  240, 2197]])
```

Model accuracy score for this second model is: 0.8875

	precision	recall	f1-score	support
Cabernet Sauvignon	0.81	0.86	0.83	3813
Chardonnay	0.98	0.98	0.98	4323
Pinot Noir	0.87	0.93	0.90	4276
Red Blend	0.88	0.73	0.80	2996
accuracy			0.89	15408
macro avg	0.89	0.87	0.88	15408
weighted avg	0.89	0.89	0.89	15408

Closing Remarks

Fine-tuning the modeling to increase accuracy:

- Adjustments to datasets/sample size
- Try other classification algorithms, e.g. DecisionTree or SVM, which was not working well in our experiments

Questions?