Wine Review Classification Analysis

By: Meichan, Nicholas, Eryka

IST 644: NLP

Project Summary

Introduction

Data source

Data preprocessing

Data descriptive analysis

Prediction modeling with classification

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

Data preprocessing

1| Import data as pandas data frame

2| Examine data types

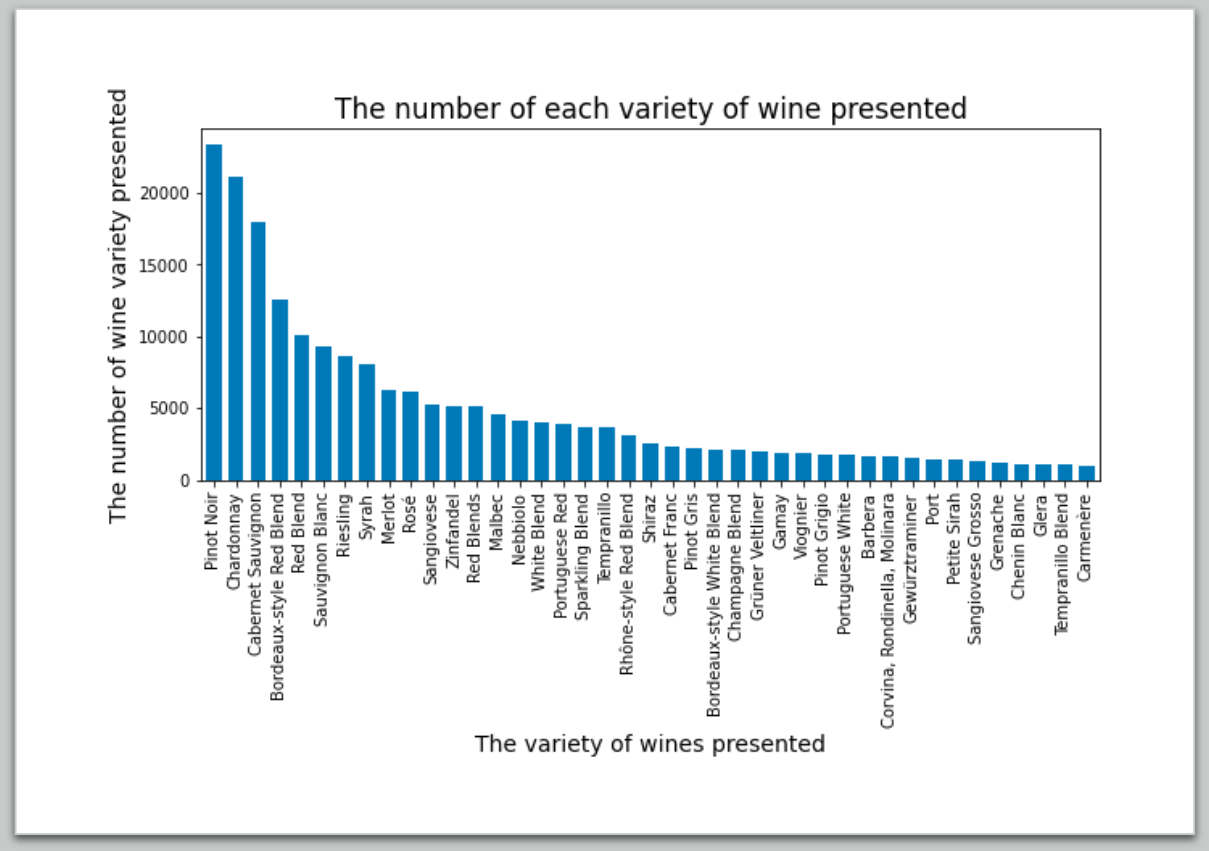
3| Handle missing values (remove NAs and impute NAs with mode):

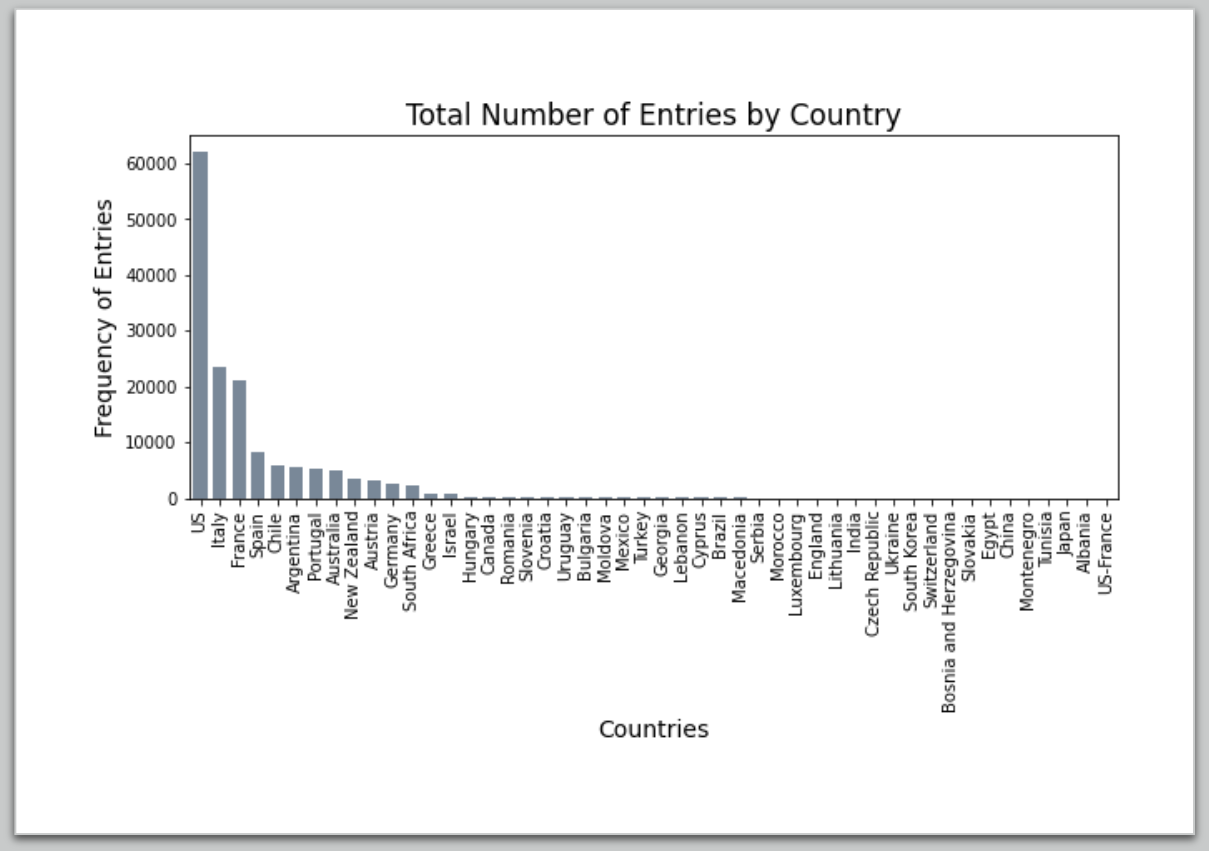
- country
- province
- prices

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

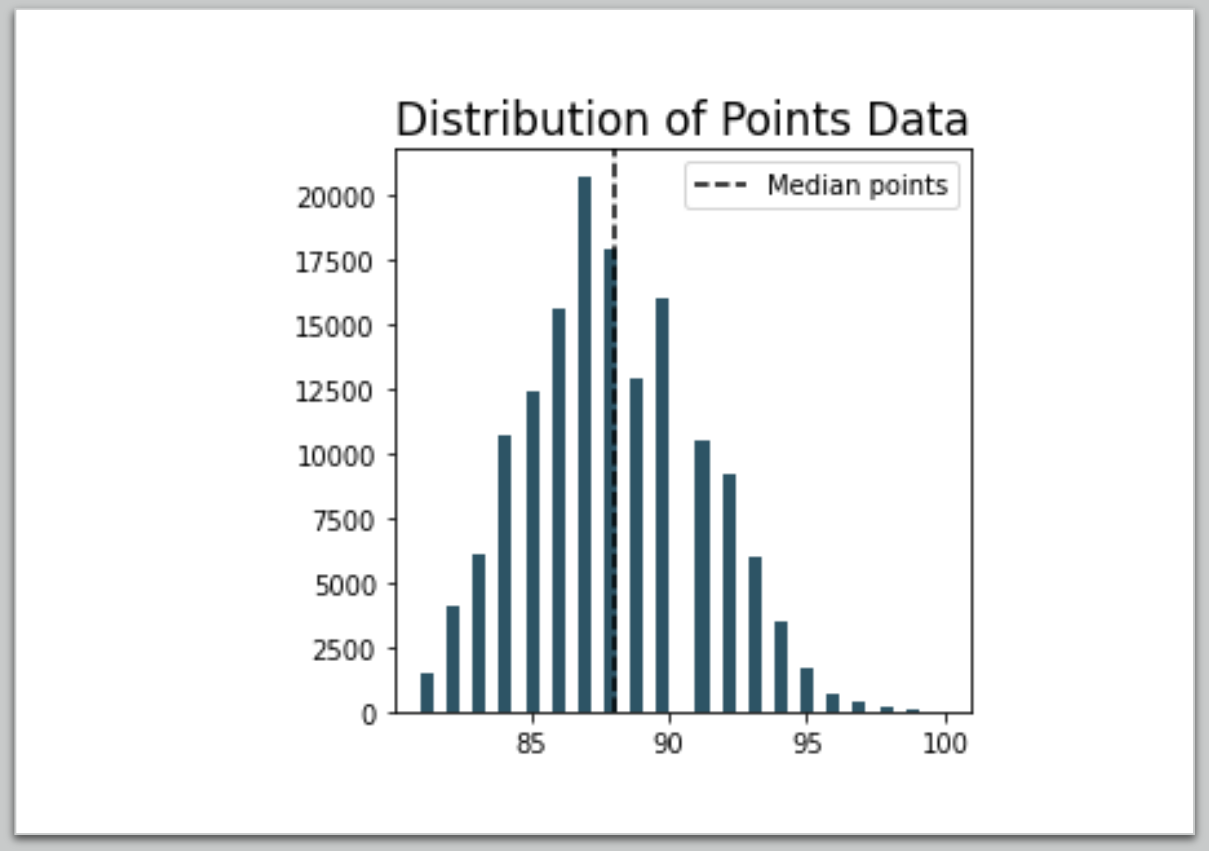
Descriptive Results





Descriptive Results

Price and Points



Descriptive Results

Token and Bigram Frequencies

20 Most Frequent Tokens

```
[('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)]
```

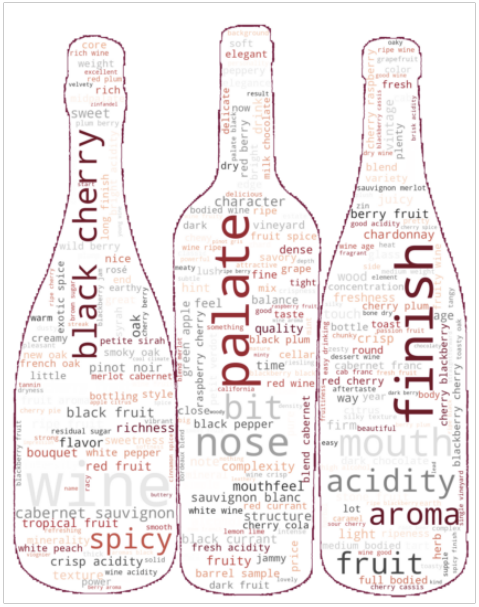
20 Most Frequent Nouns

```
[('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)]
```

20 Most Frequent Adjectives

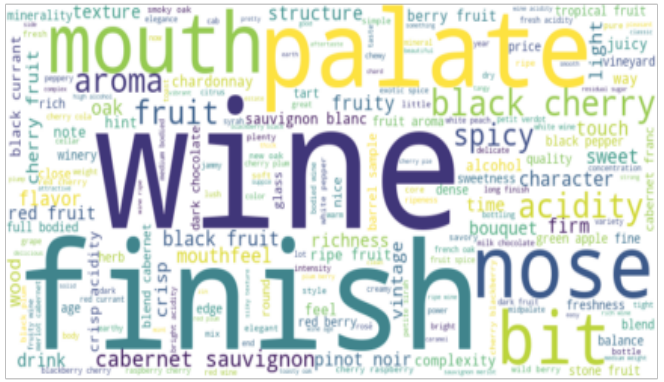
Word Clouds

Wine Reviews



Noun Phrases

Adjective Phrases



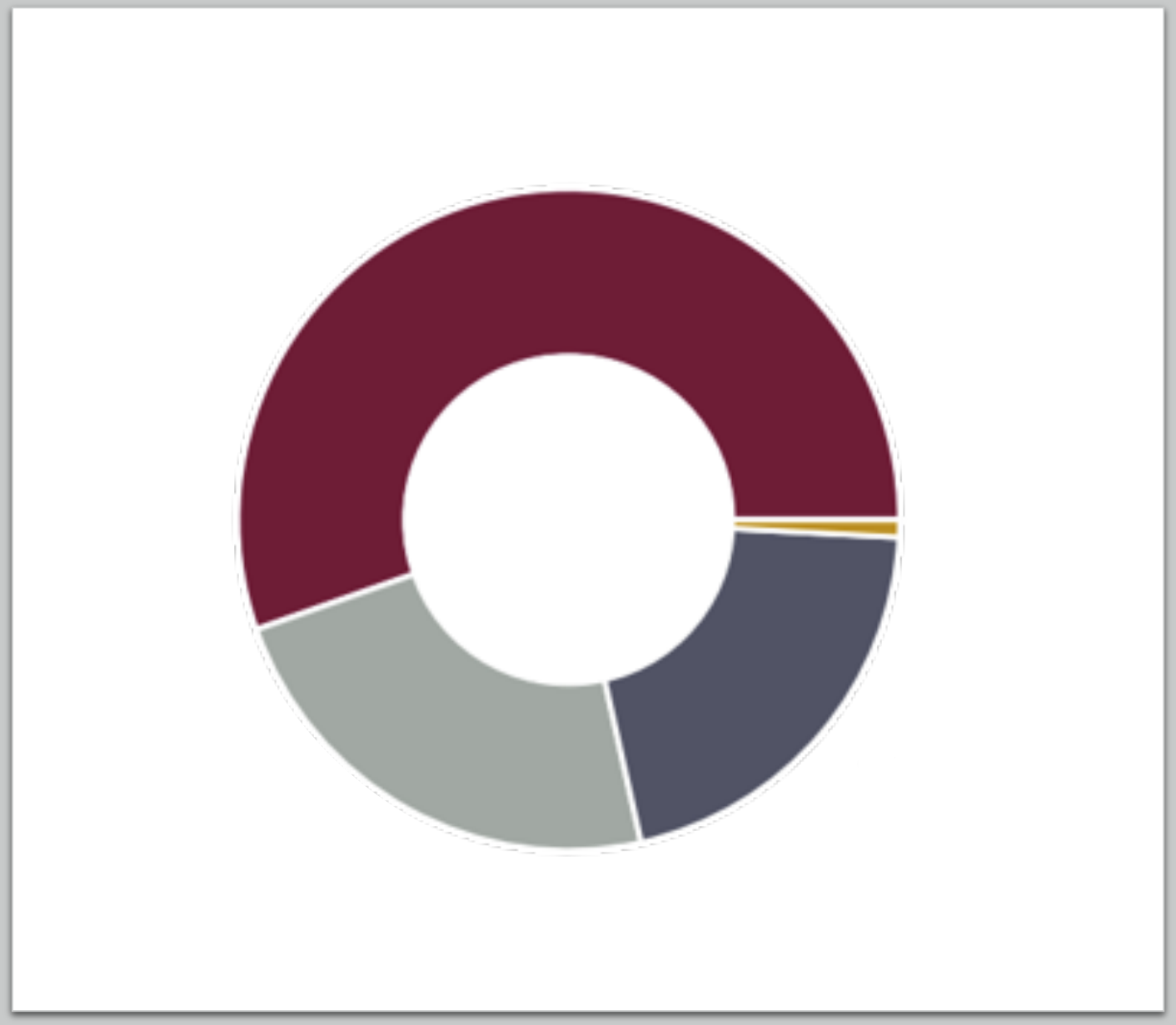
Categorize Points into Quality

	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

Review Quality

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 winereview and reviewdf (tokenized)

3| Different sample size sets

4| Combination of different columns to have best accuracy outcome

56.67%

Highest Accuracy

Variety & Quality

8%

Lowest Accuracy

Price & word_token_cleaned

Closing Remarks

Fine-tuning the modeling to increase accuracy:

- Using adjectives for modeling
- Using nouns for modeling