Wine Review Analysis

By: Meichan Huang

IST 652 FALL 2022: Scripting for Data Analysis

Project Summary

Introduction

Data source

Data preprocessing

Data analysis

Findings

	country	description	points	price	province	variety	tokenized_words
0	Italy	Aromas include tropical fruit, broom, brimston...	87	20.0	Sicily & Sardinia	White Blend	[aromas, include, tropical, fruit, broom, brim...
1	Portugal	This is ripe and fruity, a wine that is smooth...	87	15.0	Douro	Portuguese Red	[ripe, fruity, smooth, structured, firm, tanni...
2	US	Tart and snappy, the flavors of lime flesh and...	87	14.0	Oregon	Pinot Gris	[tart, snappy, lime, flesh, rind, dominate, gr...
3	US	Pineapple rind, lemon pith and orange blossom ...	87	13.0	Michigan	Riesling	[pineapple, rind, lemon, pith, orange, blossom...
4	US	Much like the regular bottling from 2012, this...	87	65.0	Oregon	Pinot Noir	[much, like, regular, bottling, comes, across,...

	geometry	Unnamed: 0	country	description	country_code	latitude	longitude
	MULTIPOLYGON (((-122.84000 49.00000, -122.9742...	14	Canada	399	CA	56.130366	-106.346771
	MULTIPOLYGON (((-122.84000 49.00000, -120.0000...	0	United States	86678	US	37.090240	-95.712891
	MULTIPOLYGON (((-68.63401 -52.63637, -68.25000...	7	Argentina	5587	AR	-38.416097	-63.616672
	MULTIPOLYGON (((-68.63401 -52.63637, -68.63335...	5	Chile	6068	CL	-35.675147	-71.542969

	Datetime	User	Location	Followers	TotalTweets	Likes
0	1669334354000	markreinoso	Senoia, GA	101	2595	3
1	1669334280000	jacy1273		65	317	0
2	1669334273000	Dedric84P		12	4553	0
3	1669333828000	InfoAuMax1		80	10736	0
4	1669333657000	bonesondisplay	Pennsylvania, USA	1830	17118	0

Data source

- 130k wine reviews from 2017 from Kaggle
- 80K wine reviews from 2017 – 2020 from Kaggle
- World Country Longitude and Latitude data from Kaggle
- 10k wine twitters with #wine, #winereview from 11/14– 11/24/2022 scrapped using snscrape

Attributes: Country, Province, Points, Price, Variety, Review

Research questions

Question 1: Which countries and provinces had the MOST and LEAST reviews (demographic distributions of wine reviews)?

Question 2: What is the MEAN, MIN and MAX points received for wines from the most and least reviewed countries and provinces?

Question 3: What was the average price, min, and max price of wine by country?

Question 4: How many wines were reviewed by category? What are the top 10 wine varieties reviewed?

Question 5: What are the max scores, min scores, particular for top 10 most reviewed varieties?

Question 6: What was the average price, min, and max price of wine by top 10 reviewed varieties?

Question 7: For each country, what type of wine is most reviewed?

Question 8: What is the correlation between the price and score of a bottle of wine?

Question 9: What types of descriptors were frequently associated categories of wine in the reviews? e.g. what kind of fruit flavor were typically used in the wine reviews?

Question 10: Who were the top 10 accounts posted during the time period of 11/14/2022 to 11/24/2022 and what types?

Question 11: How many users tweeted during the week prior to Thanksgiving based on the Date?

Question 12: How many users tweeted for each hour of the day during the time period? In other words, at what time did users tweet more frequently compared to other hours?

Data preprocessing

1| Import data as pandas data frame

2| Examine data types

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

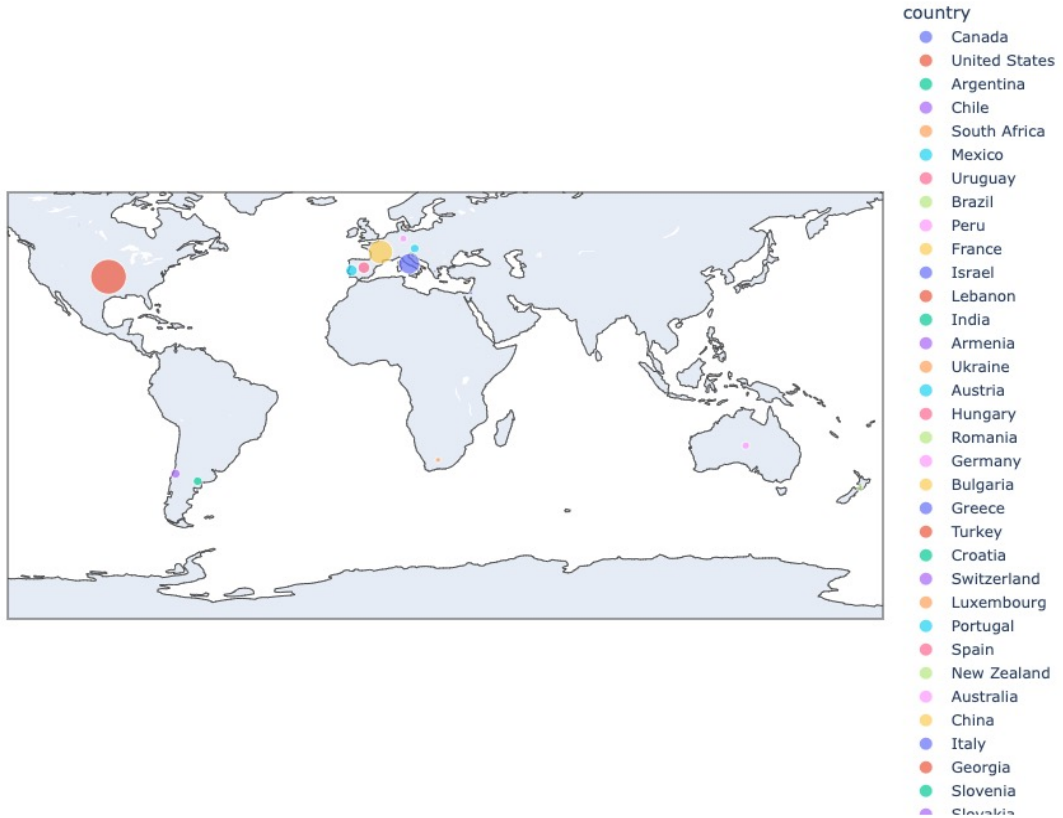
- country
- province
- prices

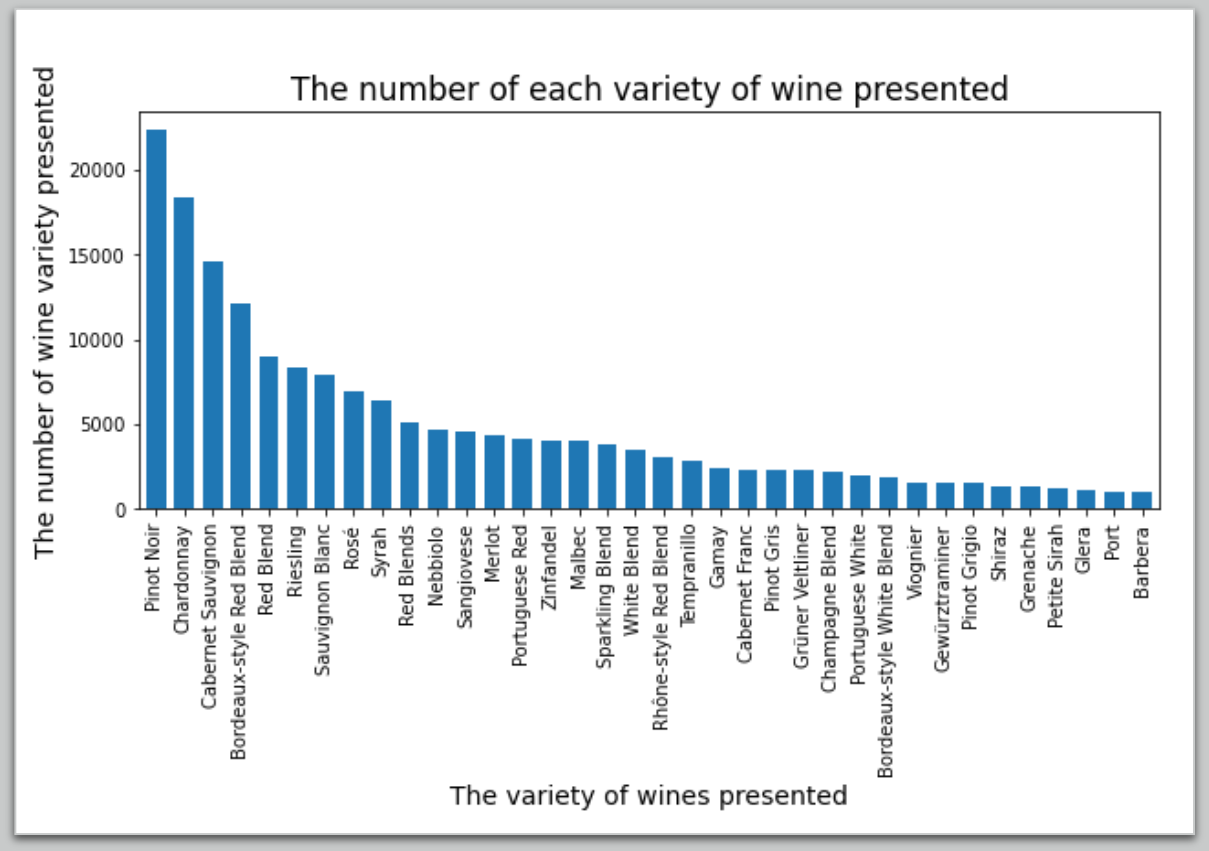
4| Join data frames of longitude and latitude dataset and reviews by countries)

5| Preprocess the textual data of ”description":

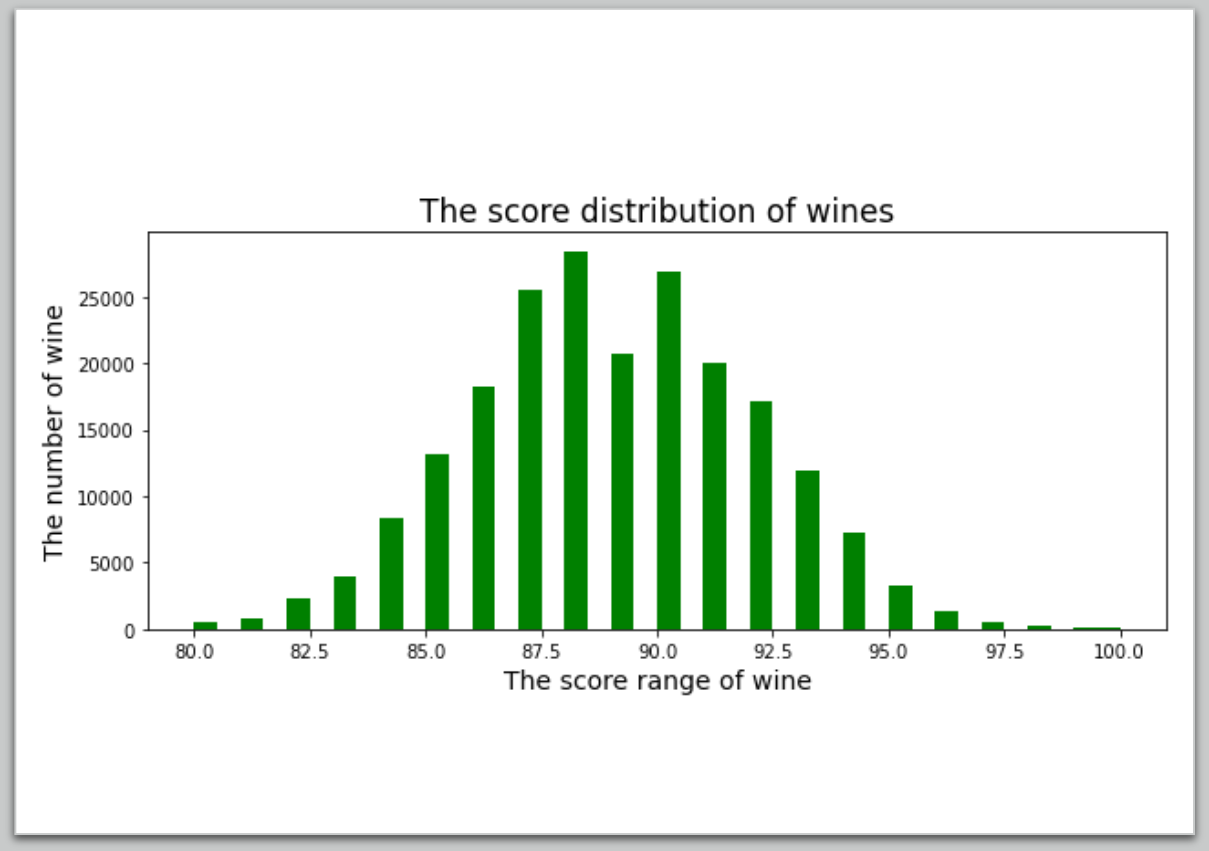
- Tokenize
- Remove stopwords and punctuations
- Extract bag of words and bigrams

Descriptive results: wine countries and wine varieties



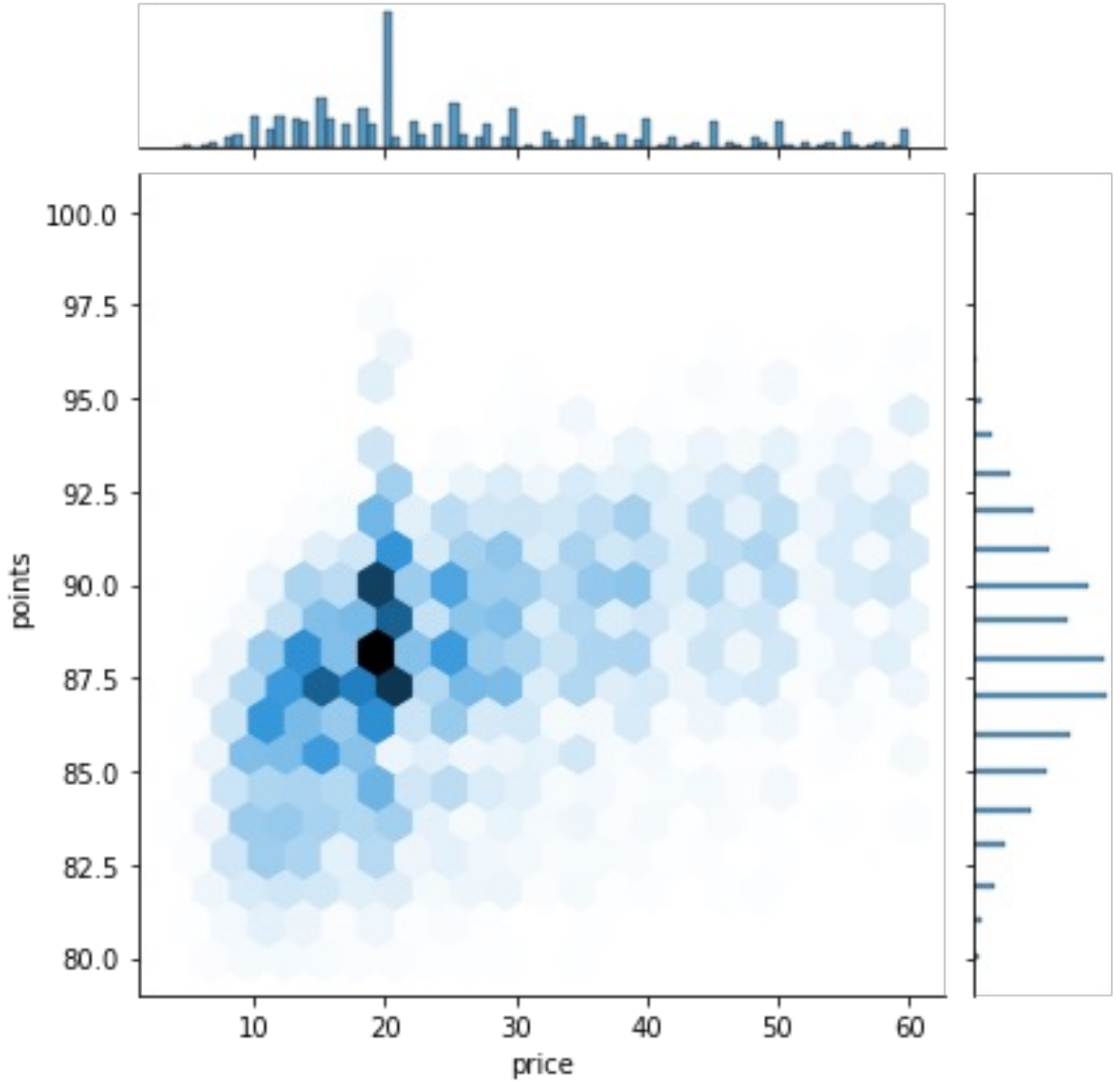


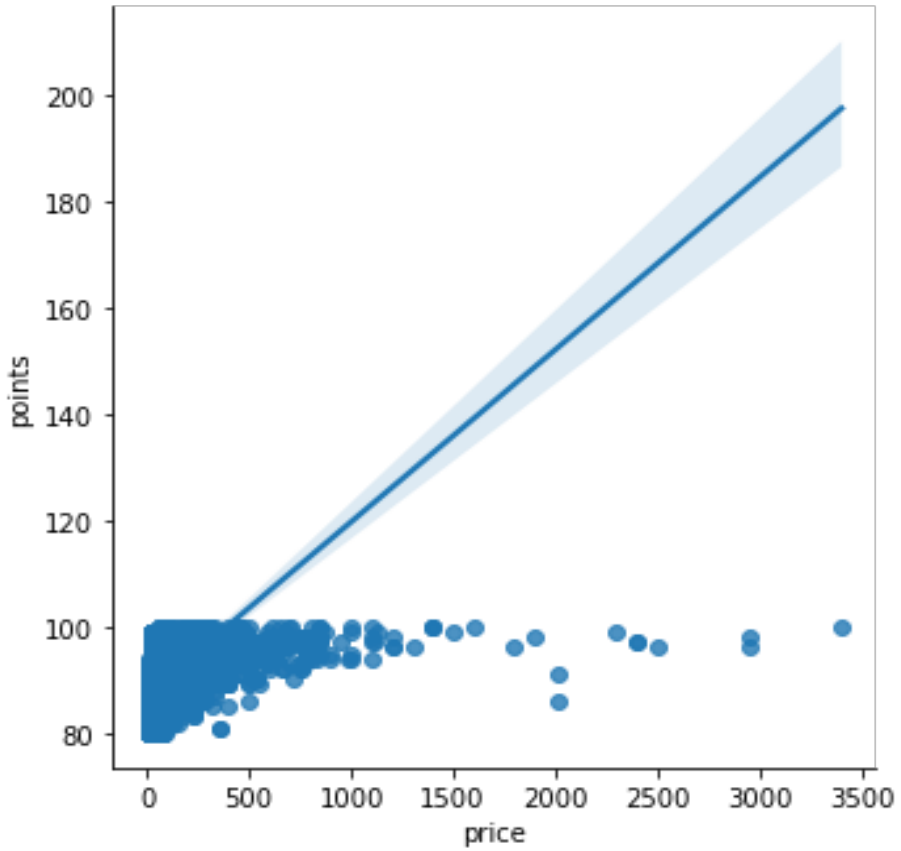
Descriptive Results: Price and Points

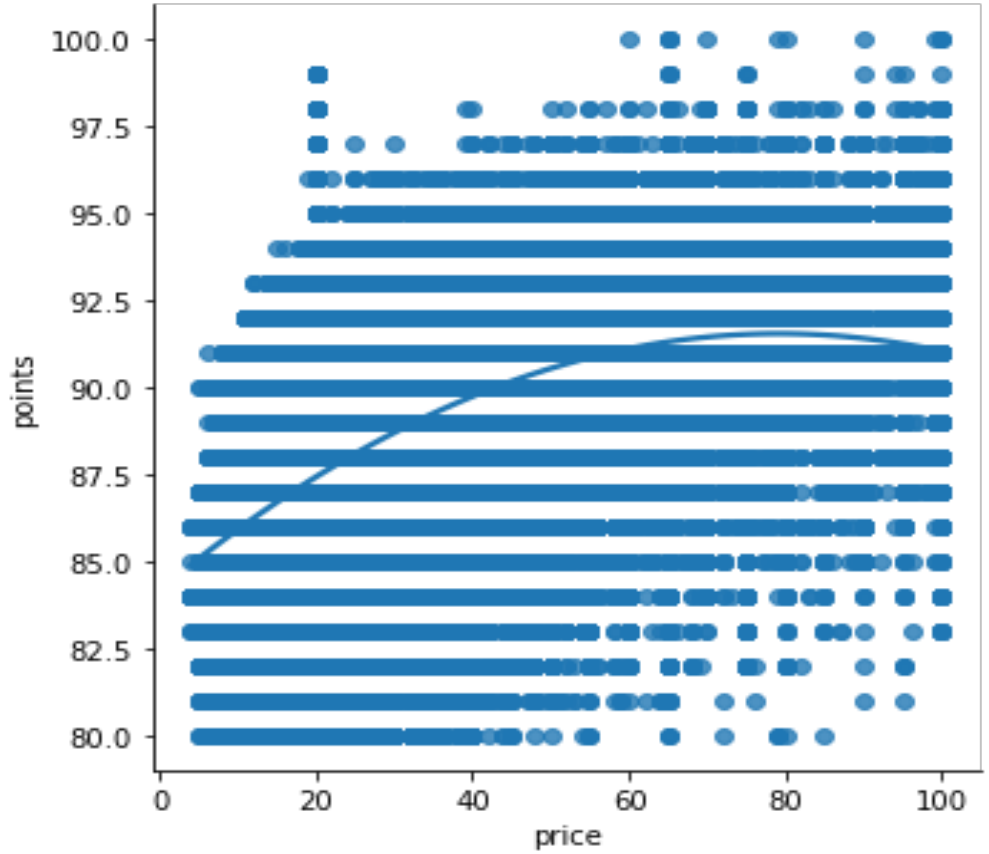


What is the correlation between the price and score of a bottle of wine?

- Wines between 12- 18 dollars have the highest point value and price density
- You can get the best value of your dollars with wines in this range







- There is not a linear correlation between price and points
- When computing all price points in the model, only 14% variance in points explained
- When only including wine from $100 or below, 27% of variance in points explained

Can we get excellent wine with cheaper price?

	country	description	points	price	province	variety	tokenized_words
345	Australia	This wine contains some material over 100 year...	100	350.0	Victoria	Muscat	['contains', 'material', 'years', 'old', 'show...
7332	Italy	Thick as molasses and dark as caramelized brow...	100	210.0	Tuscany	Prugnolo Gentile	['thick', 'molasses', 'dark', 'caramelized', '...
36512	France	This is a fabulous wine from the greatest Cham...	100	259.0	Champagne	Champagne Blend	['fabulous', 'greatest', 'champagne', 'vintage...
39268	Italy	A perfect wine from a classic vintage, the 200...	100	460.0	Tuscany	Merlot	['perfect', 'classic', 'vintage', 'masseto', '...

- 53 wines with perfect scores (100)
- Price ranging from $ 60 – $3400
- Oregon has a pinot noir for $60 dollars
- France has a Bordeaux-style red blend for $3400 dollars

	country	description	points	price	province	variety	tokenized_words
48871	France	A beautiful, pure wine that combines freshnes...	98	20.0	Burgundy	Pinot Noir	['beautiful', 'pure', 'combines', 'freshness',...
109353	Austria	Opulent honey and lemon aromas waft from the g...	98	20.0	Burgenland	Welschriesling	['opulent', 'honey', 'lemon', 'aromas', 'waft'...
111704	France	A big, bold wine with unbelievable power and c...	99	20.0	Bordeaux	Bordeaux-style Red Blend	['big', 'bold', 'unbelievable', 'power', 'conc...
111705	France	Stern, almost severe initially, this great win...	99	20.0	Bordeaux	Bordeaux-style Red Blend	['stern', 'almost', 'severe', 'initially', 'gr...
111706	France	A great wine that is just starting out. The hi...	98	20.0	Bordeaux	Bordeaux-style Red Blend	['great', 'starting', 'high', 'proportion', 'c...

- 49 wines with 98 points or above and below $20 dollars
- Mostly Bordeaux-style red blend from France
- But if you do not mind, there are 600+ wine with 95 points or above for that price range

Implications

- There is not a linear correlation between price and commercial wine
- You can get excellent wines with cheap price (e.g. 95- 98 points for 20 dollars with a wide range of options)

Future research

- Build better models examine what factors (country of origins, variety, regions, years, etc.)
- Wine recommendation apps
- Sentiment analysis with wine descriptions