OPTIMIZING WINE PURCHASE

JAMES LEE
OLIVE GARDEN INC.
FOR CHIEF MARKETING OFFICER
3/15/2019

AGENDA

- Objectives of research
- Exploratory Data Analysis (EDA)
- Unsupervised Learning
- Supervised Learning:
  - Regression
  - Classification

OBJECTIVES

- Primary Questions:
  - Price and Quality: Is more expensive wine necessarily better?
  - Review and Quality: What keywords should we look for in identifying good wines?
  - Best predictors of quality?

- Lesser questions:
  - What keywords in review tend to appear together in wine reviews?
  - What keywords predict/are associated with the variety?
  - What keywords predict/are associated with the origin of sale? (US, EU, Asia)

EXPLORATORY DATA ANALYSIS

- Data profile
- Data cleaning
- Distribution of key variables
- Preliminary analysis

DATA PROFILE

- ***Points***: the number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
- ***Variety***: the type of grapes used to make the wine (ie Pinot Noir)
- ***Description***: a few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- ***Country:*** the country that the wine is from
- *Province*: the province or state that the wine is from
- *Region 1*: the wine growing area in a province or state (ie Napa)
- *Region 2*: sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
- *Winery*: the winery that made the wine
- *Designation*: the vineyard within the winery where the grapes that made the wine are from
- ***Price***: the cost for a bottle of the wine

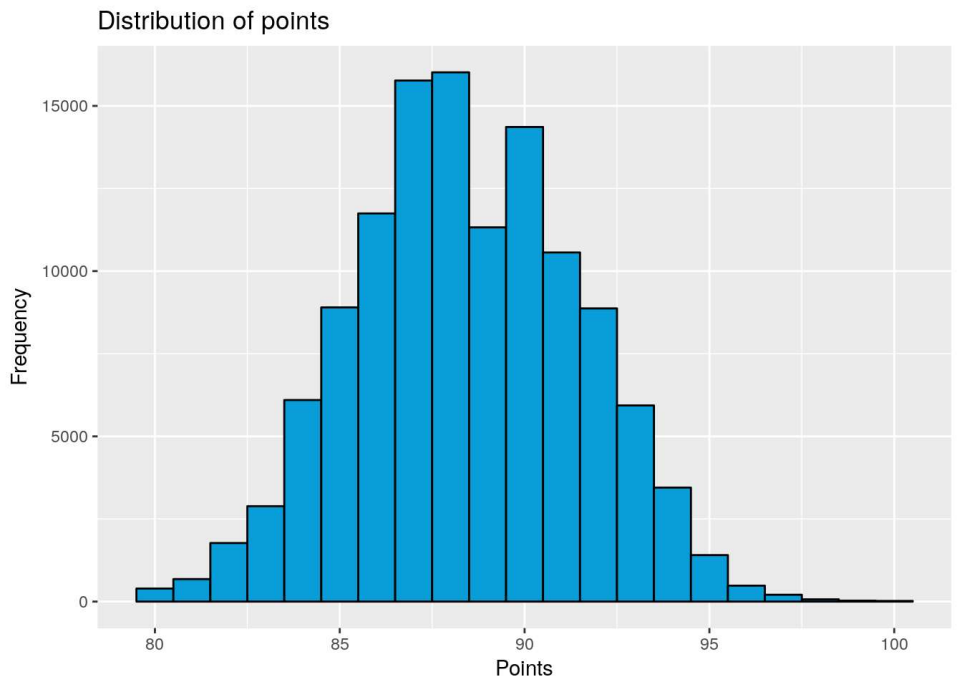
DATA CLEANING

- Remove unnamed 0 column and other non-relevant ones
- Remove rows with NaN for price/description/points
- Remove duplicates

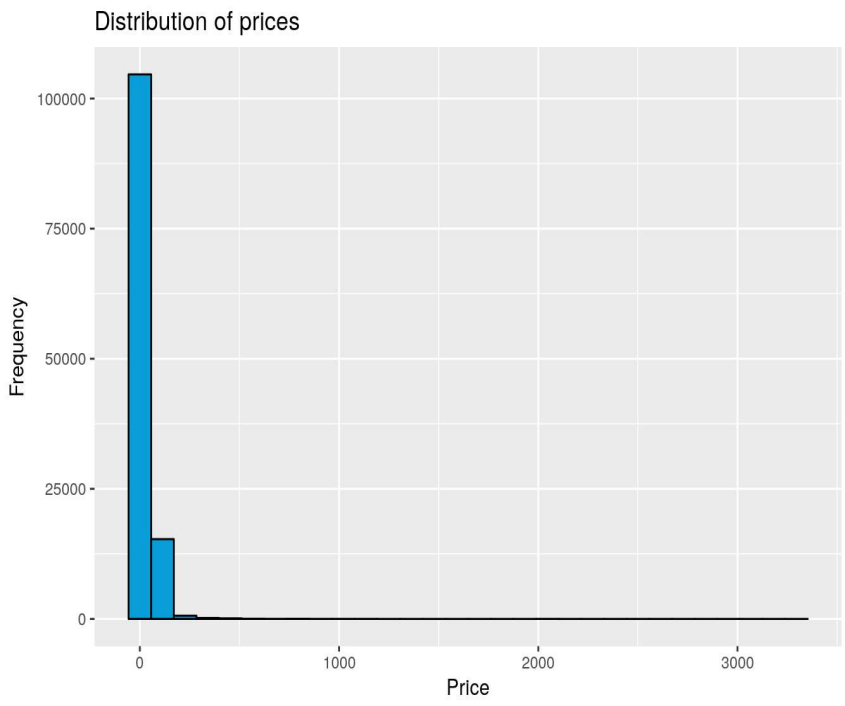
	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN

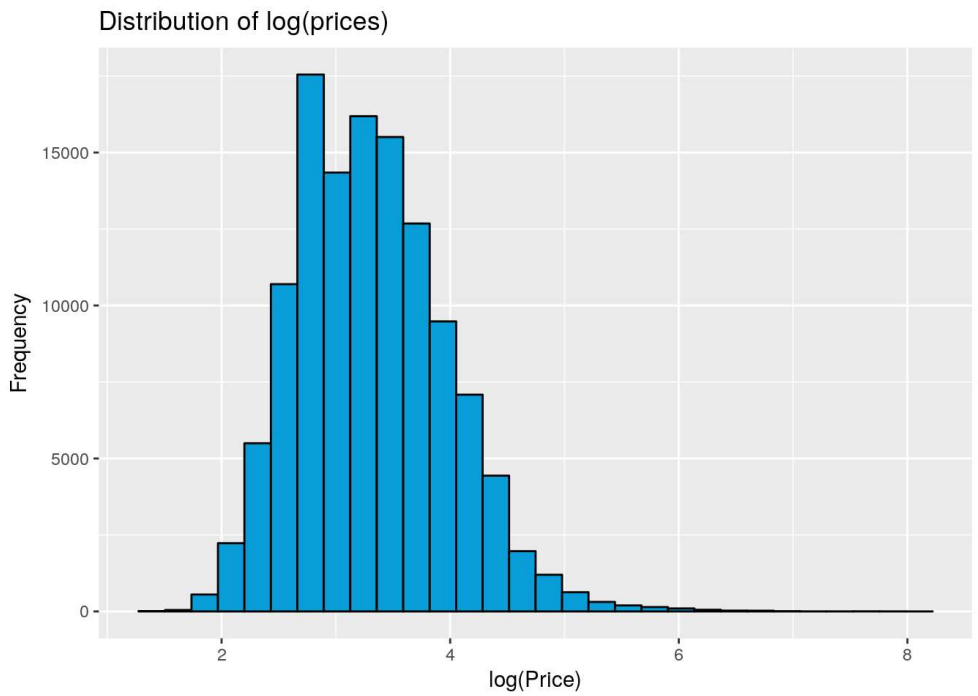
DISTRIBUTION OF VARIABLES

- Points (quality) are roughly bimodal but centered at the mean and approximately symmetric
- Mode at 88



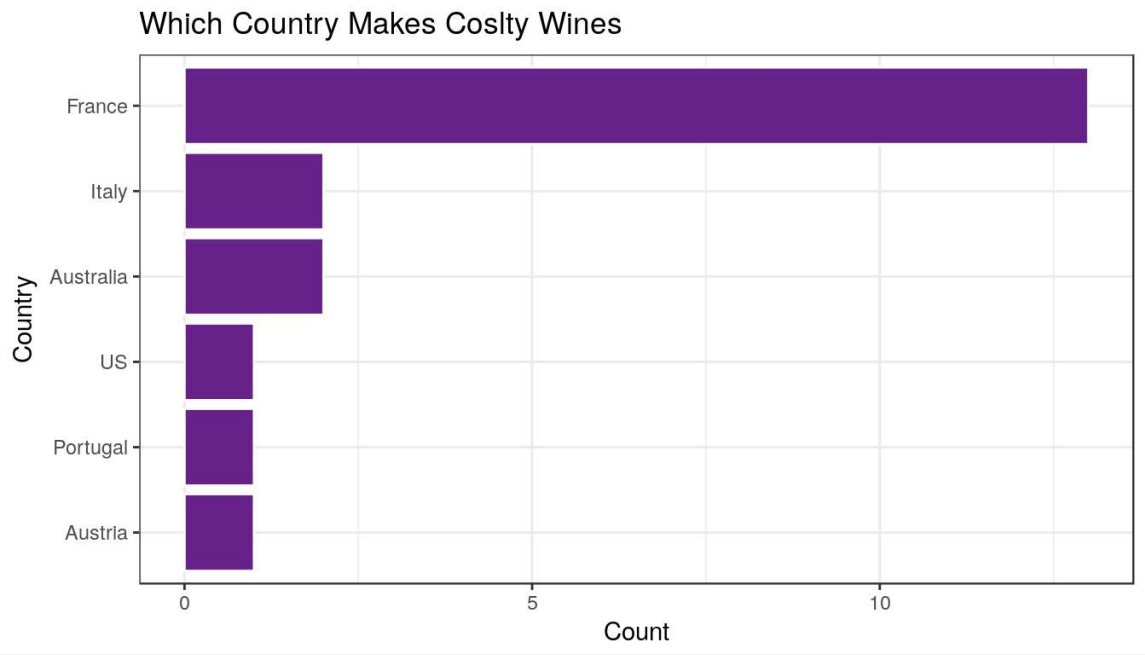
DISTRIBUTION OF VARIABLES II





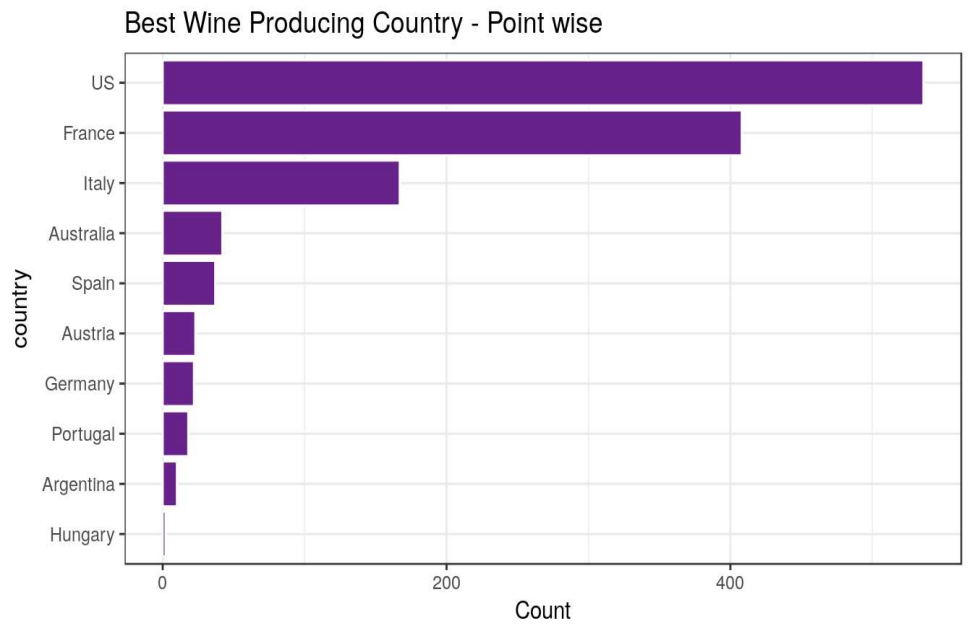
COUNTRIES AND POINTS

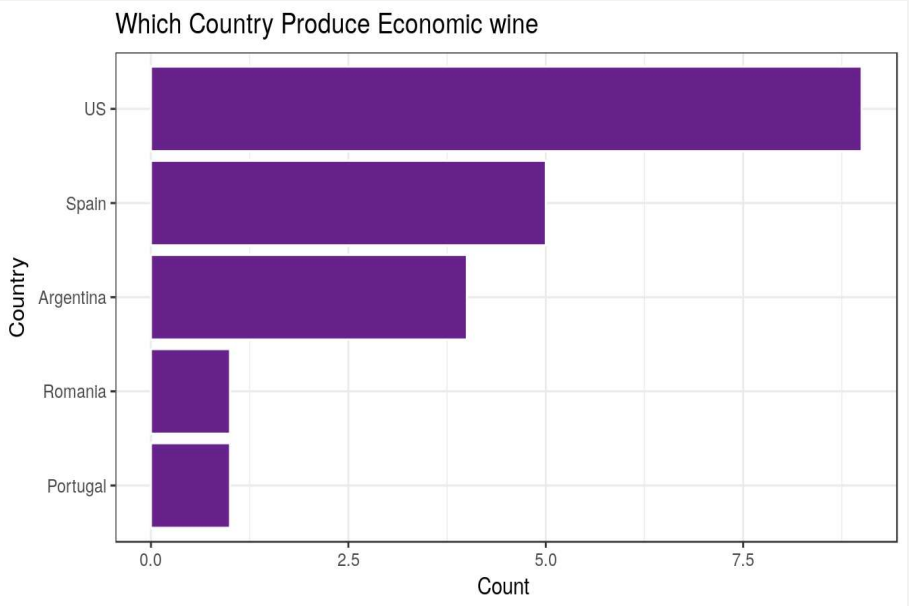
- France is on top producing costly wines.



COUNTRIES AND POINTS II

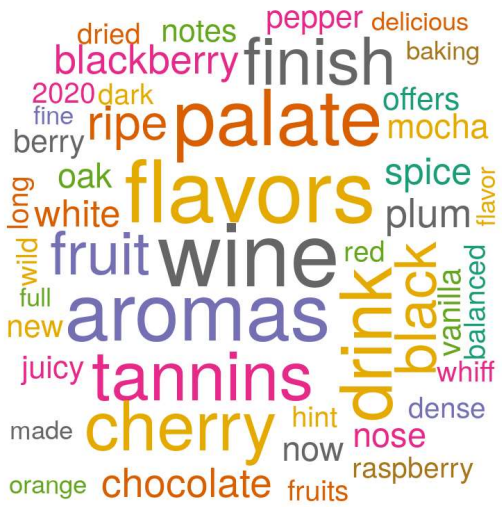
- US is where there are best winery point vise followed by France, Italy.
- US is where least cost wine produced followed by Spain, Argentina(highlighted).



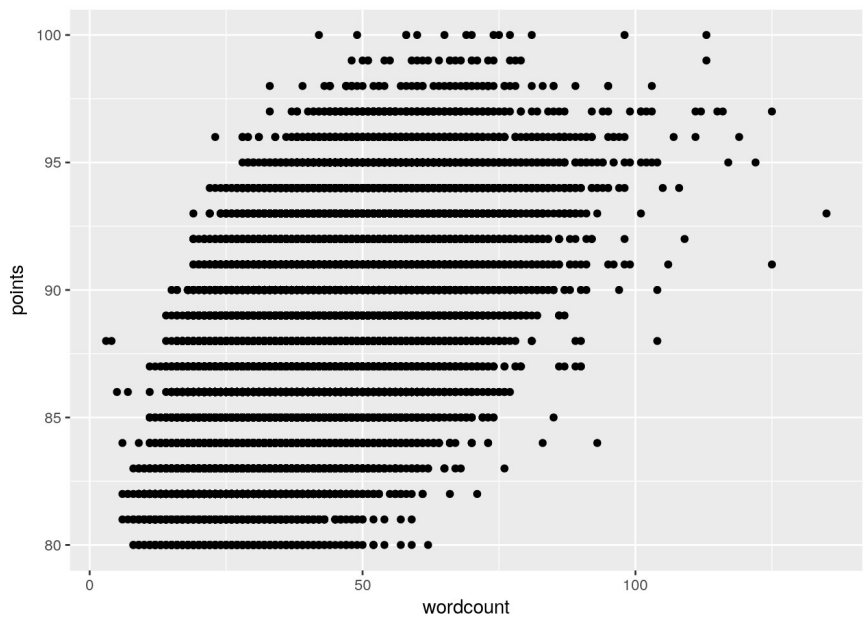


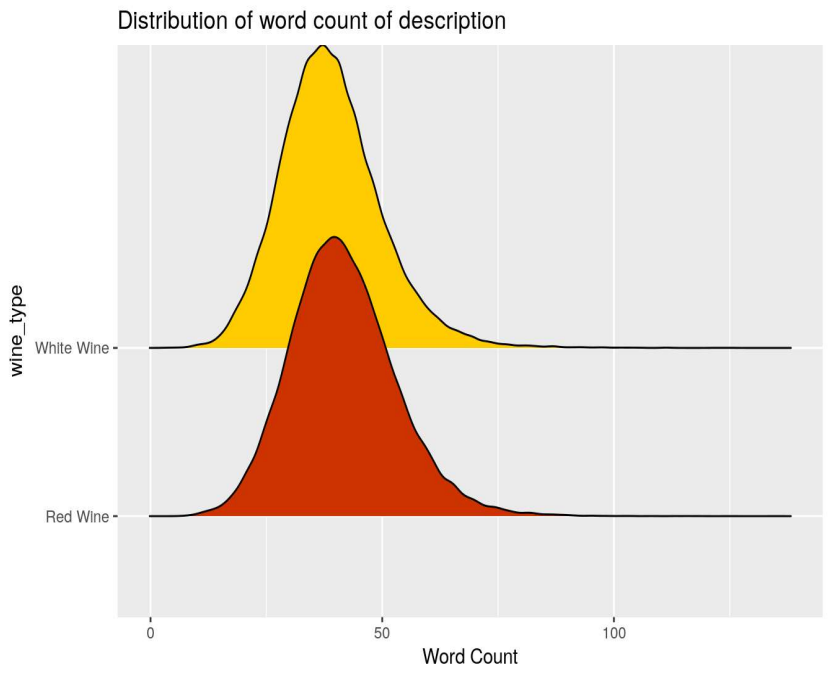
REVIEWS

- Vineyard , reserva, riserva, estate, grapes are most used words to designate a wine.
- Flavors, cherry, fruit, drink, chocolate, blackberry, plum, tannins, aromas, palate, black are most common words used to describe a wine.

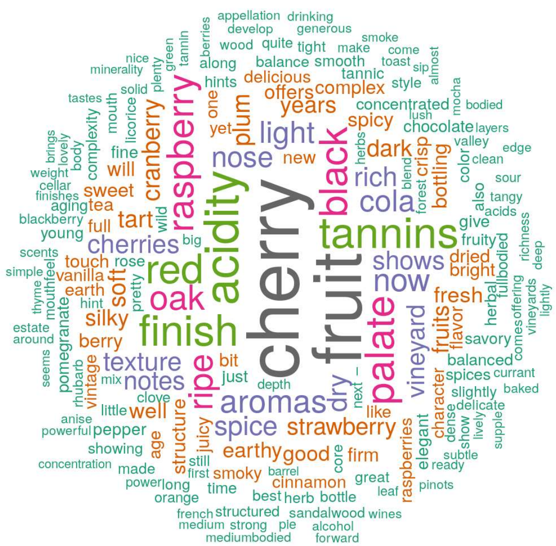


REVIEWS II


100
95
90
85
80
points
0
50
100
wordcount


Distribution of word count of description
wine_type
White Wine
Red Wine
0
50
100
Word Count

REVIEWS III


cherry
fruit
acidity
tannins
finish
palate
black
raspberry
red
oak
aromas
light
nose
rich
cola
cherries
shows
now
spice
ripe
strawberry
texture
notes
plum
years
dark
cranberry
sweet
tart
silky
earthy
good
firm
fresh
dried
bright
vineyard


acidity
fruit
apple
finish
palate
lemon
citrus
oak
aromas
pear
rich
peach
crisp
ripe
vanilla
toast
pineapple
tropical
creamy
minerality
texture
white
fresh
shows
green
notes
spice
sweet
soft
good
nose
fruits
buttered
yellow
melon
honey
orange

UNSUPERVISED LEARNING

TEXT PREPROCESSING FOR K-MEANS

- Limit to top 3000 words for efficient feature space
- Python is very useful for natural language processing:
  - Rid of stop words
  - Term frequency-inverse document frequency (TF-IDF) vectorizer
  - Stem words

K-MEANS

- 15 clusters optimal
- Words associated with each cluster
- ***What keywords in review tend to appear together in wine reviews?***
  - ***These are the keywords to be used toward recommending wines to customers***

```
0 : berri, aroma, finish, plum, palat, flavor, feel, herbal, red, nose
1 : blackberri, currant, dri, tannin, flavor, cherri, rich, oak, drink, wine
2 : citrus, peach, finish, white, lemon, flavor, lime, palat, acid, wine
3 : sampl, barrel, wine, tannin, fruit, ripe, wood, structur, juici, veri
4 : appl, green, flavor, pear, finish, palat, citrus, wine, acid, aroma
5 : wine, acid, fruiti, fresh, fruit, attract, drink, ripe, soft, red
6 : light, fruit, wine, flavor, fresh, red, finish, cherri, acid, aroma
7 : wine, fruit, cherri, flavor, finish, tannin, red, spice, berri, note
8 : cabernet, sauvignon, blend, merlot, franc, petit, verdot, wine, syrah, cherri
9 : pinot, noir, cherri, cola, silki, flavor, raspberri, dri, acid, drink
10 : sweet, cherri, flavor, soft, simpl, tast, raspberri, like, candi, wine
11 : black, cherri, palat, aroma, tannin, fruit, plum, pepper, dark, spice
12 : chardonnay, pineappl, butter, oak, toast, flavor, vanilla, acid, rich, pear
13 : wine, age, fruit, wood, structur, tannin, year, rich, firm, ripe
14 : blanc, sauvignon, flavor, citrus, crisp, acid, green, lime, grapefruit, wine
```

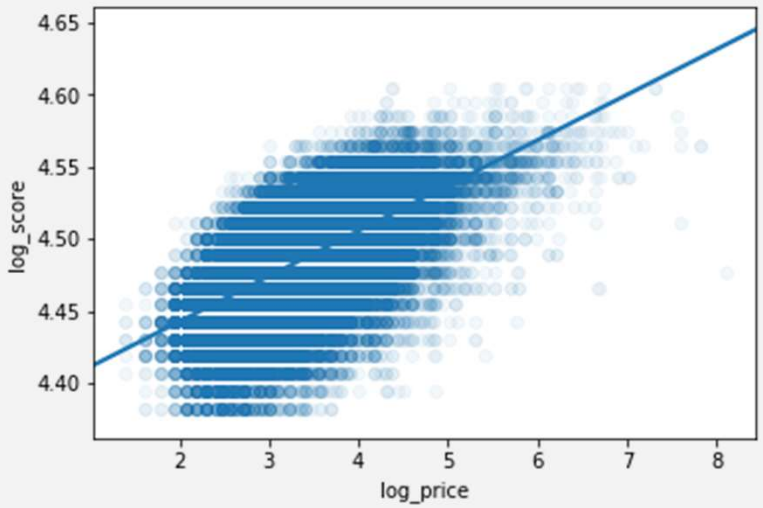
SUPERVISED LEARNING

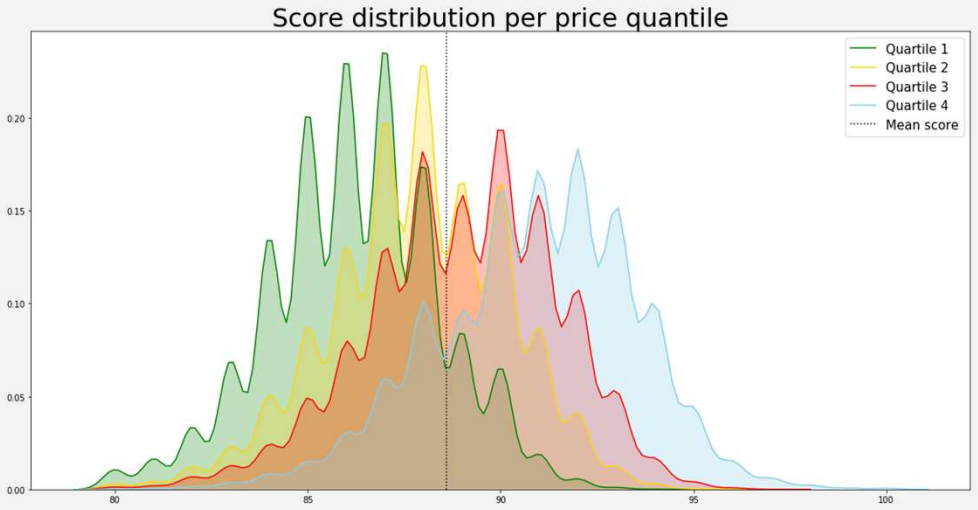
REGRESSION

- Tool:
  - Cluster Regression
  - Categorical Boosting
- Questions:
  - Price and Quality: Is more expensive wine necessarily better?
  - Best predictors of quality

REGRESSION: QUALITY VS PRICE

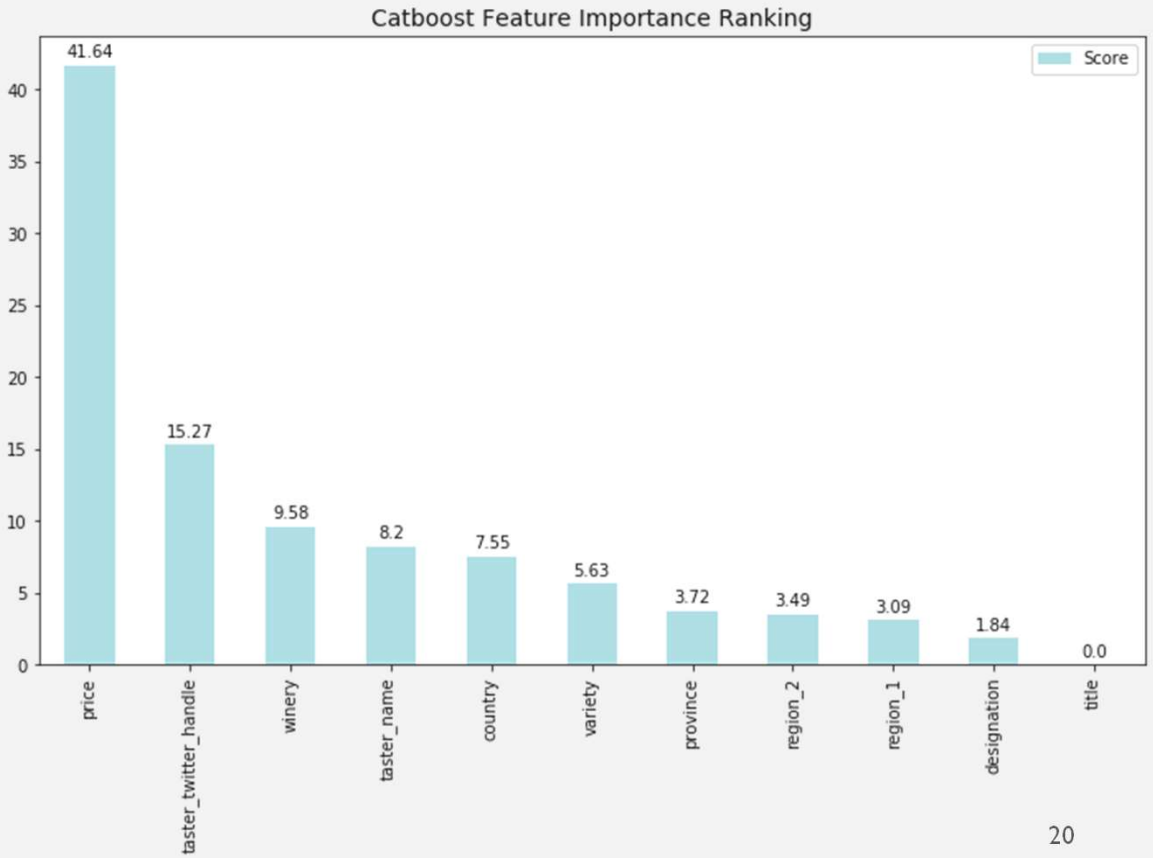
- Regress quality on price
- Despite the high correlation between price and score, there are plenty of good wine options across all price ranges
- Cluster regression on different quantiles
- **<u>Does good wine *have* to be expensive?</u>**





BOOSTING: PREDICTORS OF QUALITY

- As we can see the most important feature is price. Tester has also big impact for the points score.
- **Best predictors of quality?**
  - **Best wines tend to be expensive, tester's judgment, and makers (more important than country or grape)**

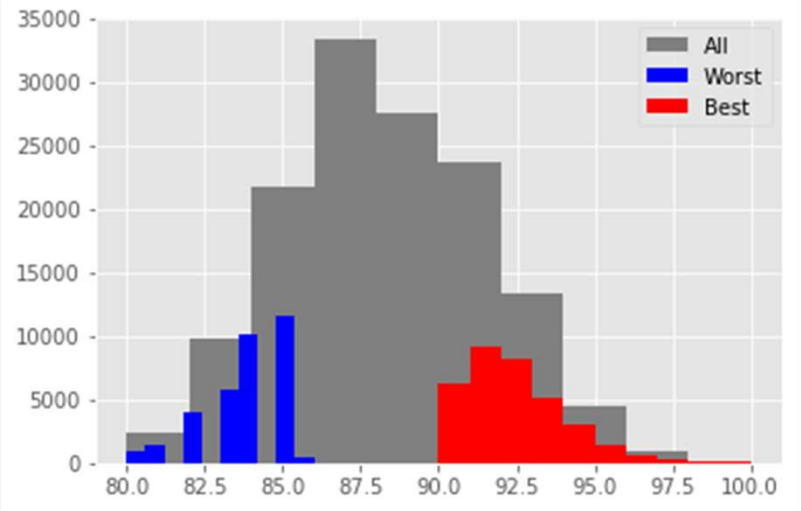


CLASSIFICATION

- Tools
  - Naïve Bayes Classifier
  - SVM
  - Multinomial Regression
- Questions
  - What keywords should we look for in identifying good wines?
  - What keywords predict/are associated with the origin of sale? (US, EU, Asia)
  - What keywords predict/are associated with the variety?

NAÏVE BAYES CLASSIFIER

- Good wine vs bad wine
- Can words predict the quality of wine? A resounding yes
- **What keywords should we look for in identifying bad/good wines**?




```
Most Informative Features
                  dilute = True           worst : best    =    163.7 : 1.0
                   weedy = True           worst : best    =    119.0 : 1.0
                   bland = True           worst : best    =     82.4 : 1.0
```

SVM

- **What keywords predict/are associated with the origin of sale? (US, EU, Asia)?**

	country	country_pred	desc	desc2
0	France	0	The wine is candied and sherbet-flavored. Acid...	wine candied sherbet-flavored acidity give fre...
1	US	1	This dense, full-bodied wine exhibits characte...	dense full-bodied wine exhibit characteristic ...
2	Italy	1	In 1998 Capezzana eliminated its "reserve" win...	1998 capezzana eliminated " reserve " wine the...
3	US	1	Brassfield's Sauvignon Blanc is proving to be ...	brassfield sauvignon blanc proving good invest...
4	France	0	In the house style of this producer this wine ...	house style producer wine soft fruity strawber...
5	Spain	0	Raw, fresh, almost scratchy aromas of cranberr...	raw fresh almost scratchy aroma cranberry plum...
6	Portugal	0	Old vines in this case means an average age of...	old_vine case mean average age 70 year vineyar...
7	Portugal	0	Sandeman's single quinta vintage is based arou...	sandeman single quinta vintage based around qu...
8	Italy	0	Fresh and bright, this has bold aromas of ston...	fresh bright bold aroma stone fruit tangerine ...
9	US	1	This is a highly unusual wine made from a trad...	highly unusual wine made traditional champagne...

	feature	tfidf
0	petite_sirah	0.042025
1	creek	0.032845
2	carneros	0.032720
3	paso_roble	0.032119
4	ava	0.031866
5	lodi	0.031360

MULTINOMIAL REGRESSION

- Prune down the varieties to 8 with observations greater than 200; small sample size can bias the coefficients for small observations
- 53% accuracy – barely better than random guess
- **Can words and price predict the variety used?**

```
from sklearn.metrics import accuracy_score
print('Accuracy Score:',accuracy_score(comparison.actual, comparison.predicted)*100,"%")
comparison.head(5)
```

	actual	predicted
0	Chardonnay	Chardonnay
1	Sauvignon Blanc	Sauvignon Blanc
2	Cabernet Sauvignon	Cabernet Sauvignon
3	Syrah	Zinfandel
4	Sangiovese	Red Blend

SUMMARY

Questions	Answers	Business Relevance
Is more expensive wine necessarily better?	No, but expensive wines strongly tend to be better	Identify best wines at the affordable price level using database
What keywords should we look for in identifying good wines?	NBC result	Use the set of key words to identify new good, affordable wines
Best predictors of quality?	Price, taster's subjectivity, and producer	Identify producers who produce the best wines at the cheapest cost
What keywords in review tend to appear together in wine reviews?	K-means result	We can expand keyword database to identify good wines
What keywords predict/are associated with the variety?	Not clear	No result
What keywords predict/are associated with the origin of sale? (US, EU, Asia)	Mostly regional and specific grape name	Uninteresting result

FUTURE WORK

- Reviews from average consumers, not sommelier
- Drill down on the questions:
  - How much does variety matter to wine rating?
  - How much does production origin matter to wine rating?