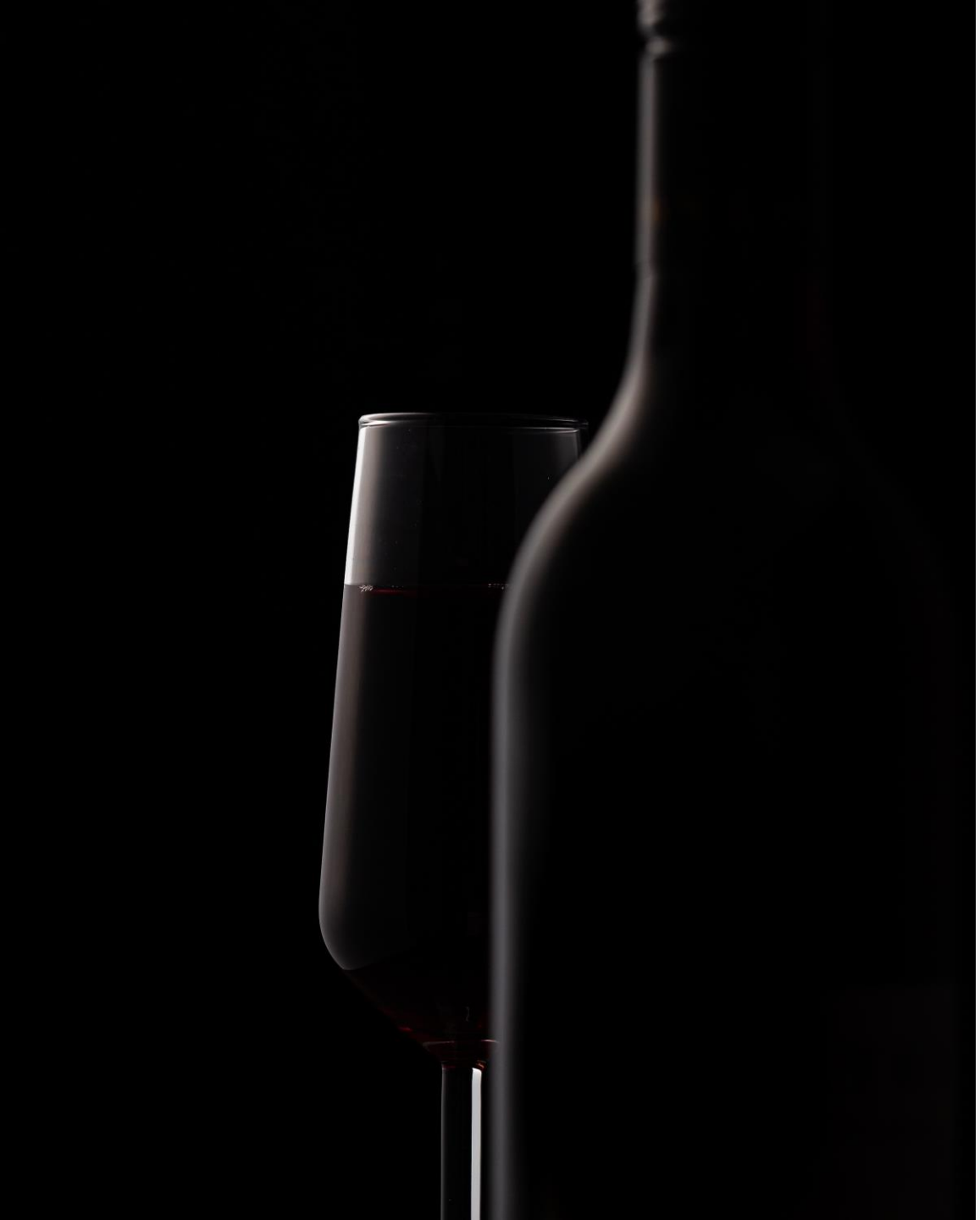
ALL YOU NEED IS…GOOD WINE

3251 - 017.

Group 6: Ali, Neda, Nisarg, Olga, Vaishali, Vinay



Agenda

1. Introduction
2. Hypothesis Testing
3. Logistic Regression
4. Linear Regression and SVM
5. Conclusions

HOW MUCH DO WE DRINK?

What does an average month of drinking look like?

10 Days of consumption per month	13 Unique occasions	25 Total drinks consumed

3.6 liters more than the world average

Across multiple categories (beer, wine, spirits, cider, RTD)

MARKET VALUE OF THE WINE INDUSTRY IN CANADA

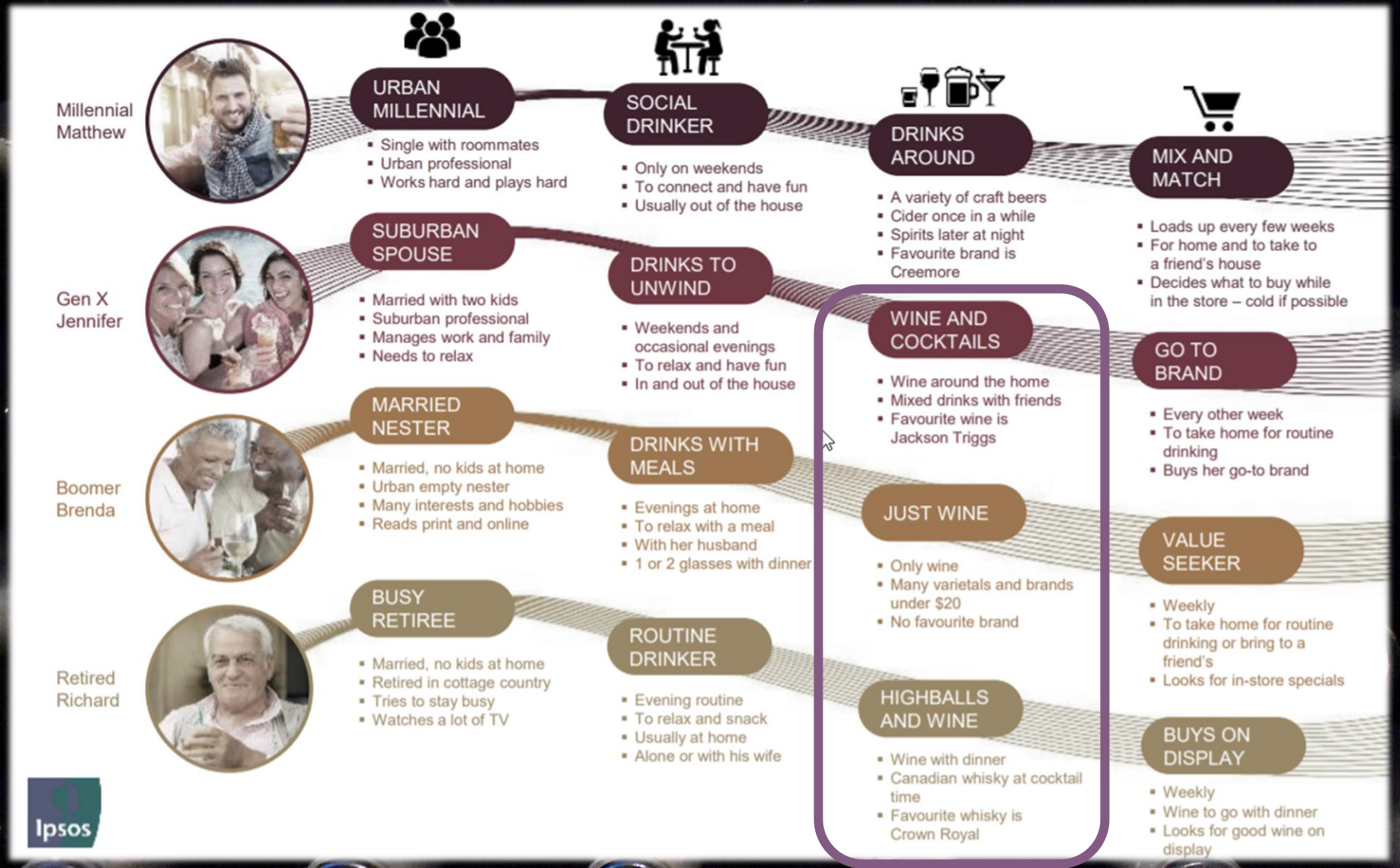
8.59bn CAD

FORECASTED MARKET VALUE OF THE WINE INDUSTRY IN CANADA

10.85bn CAD

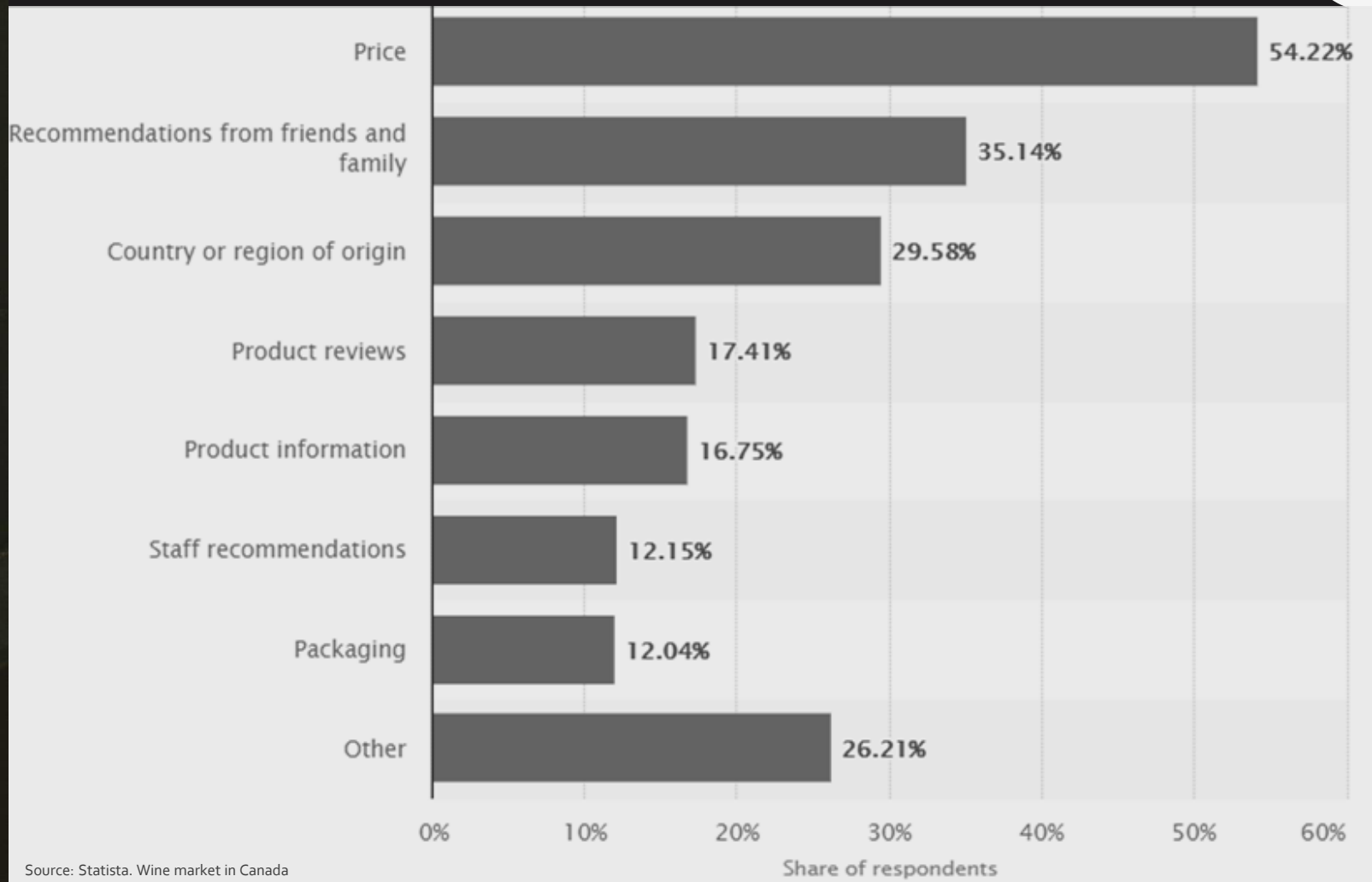
Sources: Statista. Wine market in Canada. World Health Statistics 2016

HOW WE DRINK?


Millennial Matthew
URBAN MILLENNIAL
Single with roommates
Urban professional
Works hard and plays hard
SOCIAL DRINKER
Only on weekends
To connect and have fun
Usually out of the house
DRINKS AROUND
A variety of craft beers
Cider once in a while
Spirits later at night
Favourite brand is Creemore
MIX AND MATCH
Loads up every few weeks
For home and to take to a friend's house
Decides what to buy while in the store – cold if possible
Gen X Jennifer
SUBURBAN SPOUSE
Married with two kids
Suburban professional
Manages work and family
Needs to relax
DRINKS TO UNWIND
Weekends and occasional evenings
To relax and have fun
In and out of the house
WINE AND COCKTAILS
Wine around the home
Mixed drinks with friends
Favourite wine is Jackson Triggs
GO TO BRAND
Every other week
To take home for routine drinking
Buys her go-to brand
Boomer Brenda
MARRIED NESTER
Married, no kids at home
Urban empty nester
Many interests and hobbies
Reads print and online
DRINKS WITH MEALS
Evenings at home
To relax with a meal
With her husband
1 or 2 glasses with dinner
JUST WINE
Only wine
Many varietals and brands under $20
No favourite brand
VALUE SEEKER
Weekly
To take home for routine drinking or bring to a friend's
Looks for in-store specials
Retired Richard
BUSY RETIREE
Married, no kids at home
Retired in cottage country
Tries to stay busy
Watches a lot of TV
ROUTINE DRINKER
Evening routine
To relax and snack
Usually at home
Alone or with his wife
HIGHBALLS AND WINE
Wine with dinner
Canadian whisky at cocktail time
Favourite whisky is Crown Royal
BUYS ON DISPLAY
Weekly
Wine to go with dinner
Looks for good wine on display
Ipsos

HOW WE DECIDE WHAT TO DRINK?

Canadian purchase criteria when buying wine 2019



Source: Statista. Wine market in Canada

BUT WHAT ABOUT QUALITY?

Let's consider hypotheses that can help us understand the quality of wine

DATA DESCRIPTION

Data description

Records

- White: 4,898
- Red: 1,599

Data types

- Real, integer & string

12 attributes, incl. quality ranking

No missing values

- **Acids**

Fixed, Volatile - high levels can lead to a vinegar taste

Citric acid - in small quantities adds freshness and flavor

- **Sulfur**

Free SO2, Sulphates & Total SO2 – antioxidant and antimicrobial

Alcohol

Ethanol % by volume

Residual sugar: amount of sugar remaining after fermentation
Chlorides: sodium chloride
Density: will differ from water (1) depending on the alcohol and sugar content
pH: 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
Quality: score between 0 and 10 (rated by sommeliers)

Source of data: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

HYPOTHESIS TESTING

HYPOTHESIS TESTING

	HYPOTHESIS 1	HYPOTHESIS 2	HYPOTHESIS 3
NULL HYPOTHESIS	Mean quality scores for red and white wines are the same	Red and white wines have the same average alcohol levels	Quality is the same for wines with higher and lower than average abv%
ALTERNATIVE HYPOTHESIS	Mean quality scores for red and white wines are different	There is difference in alcohol levels between red and white wines	Wines with higher ABV% are of significantly better quality than wines with lower ABV%
OUTCOME	MEAN QUALITY WINE IS HIGHER FOR WHITE WINE THAN IT IS FOR RED WINE	MEAN ALCOHOL LEVEL IS HIGHER FOR WHITE WINE THAN IT IS FOR RED WINE	WINES WITH HIGHER ALCOHOL CONTENT ARE PREFERRED
EXPLANATION	The region is known for its white wines (86%) and is the main source of a unique style of white wine. Portugal has the advantage of numerous indigenous grape varieties that are able to keep their acidity in hot climates and including these in a blend provides balancing freshness for rich white wines.	Given the climate in the North-West corner of Portugal, the grape varieties that grow in that region make wines with high ABV%.	Alcohol content of wine has spiked considerably in recent years. Demand is high for intense flavors and higher alcohol content.
TOTALS			

LOGISTIC REGRESSION

BINARY LOGISTIC REGRESSION MODEL
WHITE WINE

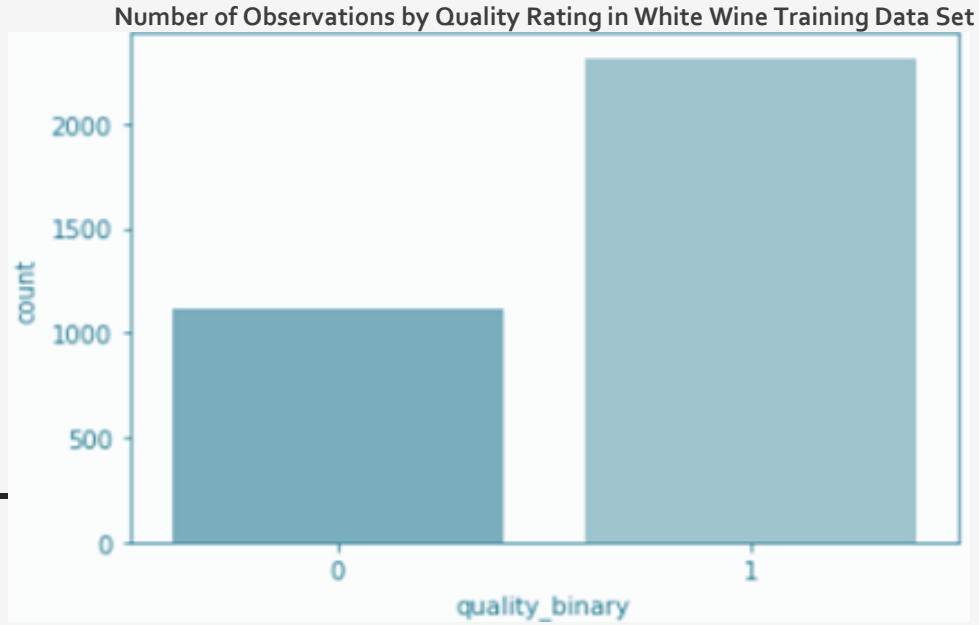
1. Reclassified wine quality based on their quality score into a binary class:

Quality score 1-10

Premium Wine >=6	1
Economy Wine <=5	0

Results: more balance data.

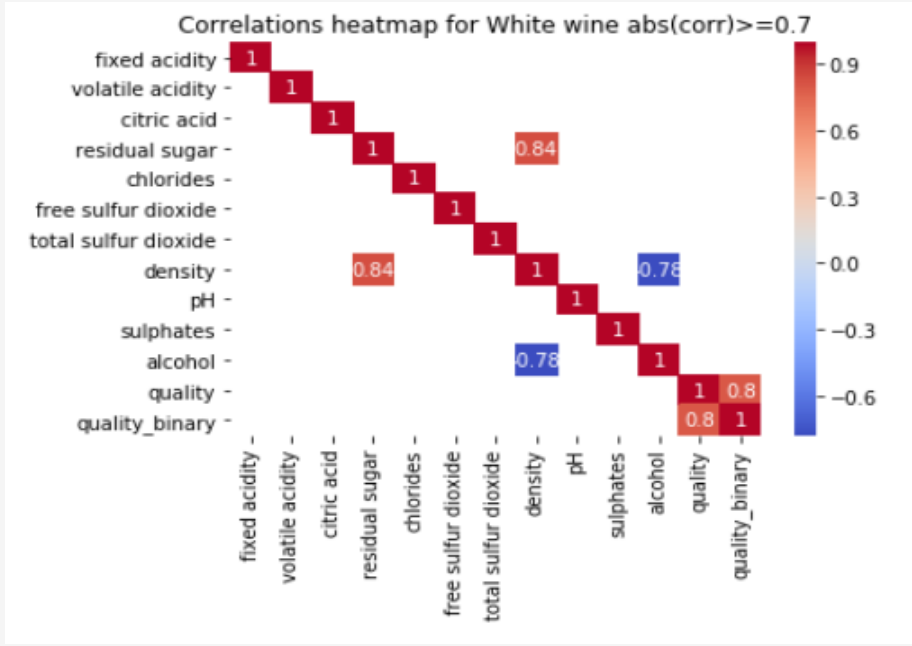
Number of positive and negative responses are comparable



2. Checking Logistic Regression Conditions:

- Observations are independent of each other (given number of samples, assumed to be reasonable)
- There should be little or no multicollinearity among predictors, removed "Density"

3. Log transformed chlorides, free sulfur dioxide, and residual sugar and split data between Train and Test (ratio=0.3).



BINARY LOGISTIC REGRESSION MODEL
WHITE WINE

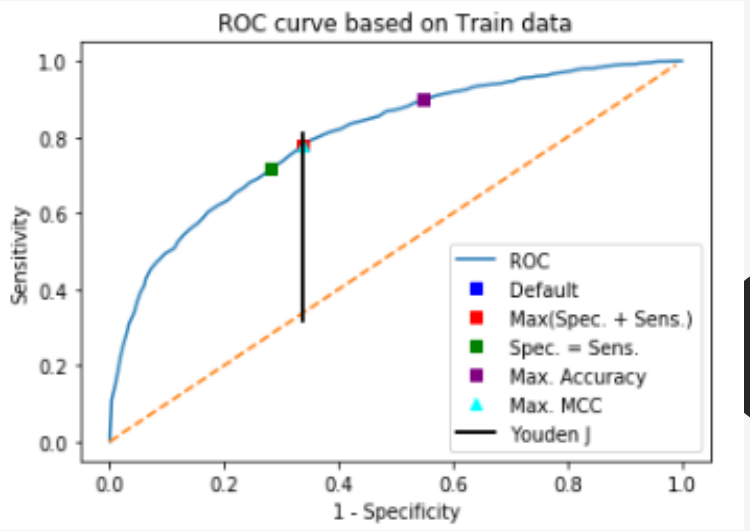
Removed statistically insignificant predictors (citric acid and chlorides) after first iteration w/o much change to overall performance.

- The wine maker must balance risk of :
  - FP (predicted as premium but is economy) to avoid disappointing customers and loss of brand equity (higher precision)
  - FN (predicted as economy but is premium) to minimize loss of revenue/profit, may lead to higher volume/customers in the segment (higher recall).
  - Tune threshold value.

		Predicted Value (y_hat)	
		1	0
True Value	1	731 TP	212 FN
	0	**187 FP**	340 TN

- Based on ROC curve, point of Max MCC was deemed to be optimum, with a threshold value of 0.62 replacing default value of 0.5 in the final model.
- Model performance with tune threshold value and Test data was "acceptable"

Data	Threshold	Precision	Recall	ACC	MCC
Train	0.50	0.774	0.898	0.754	0.400
Train	0.62	0.829	0.781	0.743	0.433
Test	0.62	0.796	0.775	0.729	0.416



LINEAR REGRESSION, SVM

LINEAR REGRESSION

Independent variables:

1) fixed acidity
2) volatile acidity
3) citric acid
4) residual sugar
5) chlorides
6) free sulfur dioxide
7) total sulfur dioxide
8) density
9) pH
10) sulphates
11) alcohol

Dependant Variable:

Wine quality grouped as continuous scalar variable.

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.295
Model:	OLS	Adj. R-squared:	0.293
Method:	Least Squares	F-statistic:	203.3
Date:	Wed, 20 Nov 2019	Prob (F-statistic):	0.00
Time:	07:30:52	Log-Likelihood:	-6464.3
No. Observations:	5847	AIC:	1.295e+04
Df Residuals:	5834	BIC:	1.304e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	109.7634	16.374	6.704	0.000	77.665	141.862
fixed_acidity	0.0866	0.018	4.944	0.000	0.052	0.121
volatile_acidity	-1.4867	0.085	-17.431	0.000	-1.654	-1.319
citric_acid	-0.0278	0.084	-0.333	0.740	-0.192	0.136
residual_sugar	0.0630	0.007	9.535	0.000	0.050	0.076
chlorides	-0.7116	0.353	-2.013	0.044	-1.405	-0.019
free_sulfur_dioxide	0.0048	0.001	5.932	0.000	0.003	0.006
total_sulfur_dioxide	-0.0014	0.000	-3.987	0.000	-0.002	-0.001
density	-108.8708	16.595	-6.561	0.000	-141.403	-76.339
pH	0.5034	0.098	5.139	0.000	0.311	0.695
sulphates	0.6928	0.081	8.586	0.000	0.535	0.851
alcohol	0.2131	0.021	10.274	0.000	0.172	0.254
white	-0.3689	0.062	-5.986	0.000	-0.490	-0.248

Omnibus:	130.269	Durbin-Watson:	1.965
Prob(Omnibus):	0.000	Jarque-Bera (JB):	292.596
Skew:	0.022	Prob(JB):	2.91e-64
Kurtosis:	4.095	Cond. No.	3.26e+05

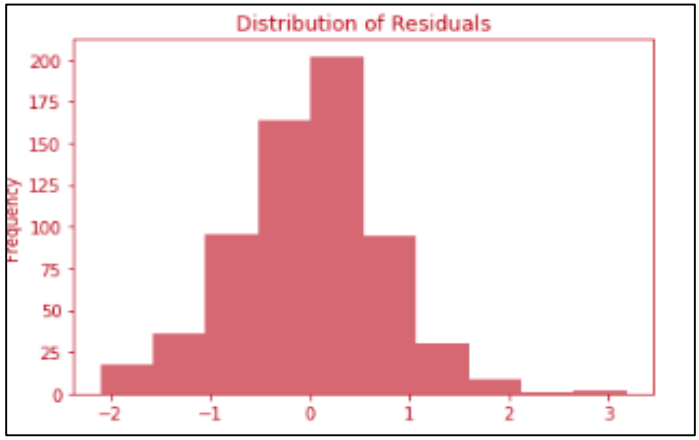
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.26e+05. This might indicate that there are strong multicollinearity or other numerical problems.

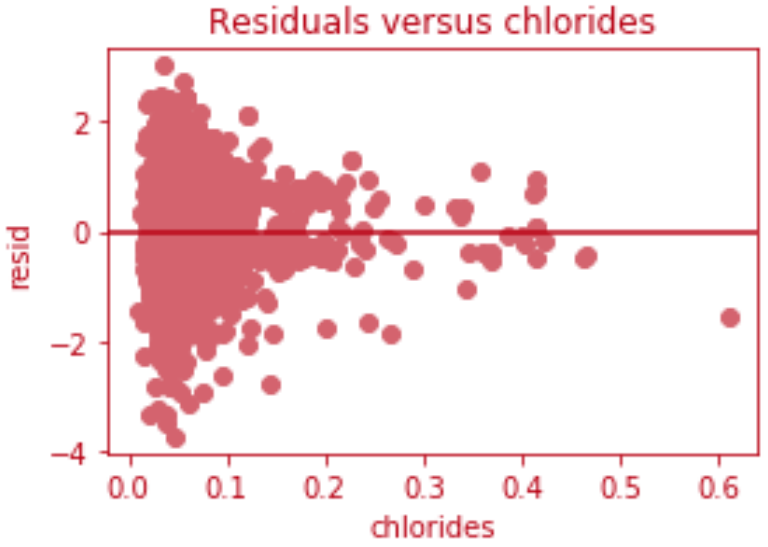
Highlights:

- 90 / 10 – test train split
- Overall score 0.3 R^2 on test

Observations:

1. weak Linearity amongst response and independent variables.
2. Independence of samples assumed as dataset is small.





3. Residuals is normal
4. No constant variability in residuals plot.

MODEL IMPROVEMENTS

Fine Tuning by backward elimination & p-value

Models have R^2 ranging from 0.26 to 0.3 which is not better than the first model.

Results of backward elimination

model with removed feature	rsquared on test	aic	bic	fvalue
fixed_acidity	0.3	12977.12	13057.2	218.65
volatile_acidity	0.27	13249.51	13329.6	184.55
citric_acid	0.31	12952.78	13032.87	221.77
residual_sugar	0.29	13043.08	13123.17	210.25
chlorides	0.31	12956.73	13036.81	221.27
free_sulfur_dioxide	0.3	12987.83	13067.91	217.28
total_sulfur_dioxide	0.31	12968.58	13048.66	219.74
density	0.3	12995.65	13075.73	216.28
pH	0.31	12979.08	13059.16	218.4
sulphates	0.29	13026.09	13106.17	212.4
alcohol	0.26	13057.51	13137.59	208.42
white	0.3	12988.47	13068.56	217.2

model with removed features	rsquared	aic	bic	fvalue
['citric_acid', 'chlorides']	0.31	12955.21	13028.62	243.37

METHODS	RIDGE REGRESSION	NORMALIZATION BY LOG TRANSFORMATION	SVR WITH NORMALIZED DATA
	Used Regularization that constrains the magnitude of coefficients to reduce model complexity	▪ Free_sulfur_dioxide variable was transformed using np.sqrt. ▪ Chlorides, volatile_acidity and sulphates variables are log-transformed	▪ Grid Search Cross Validation to find optimal SVM Params: Kernel, C, Gamma ▪ Data Normalization using Standard-Scaler
Outcome	Model coefficients are shrunk towards 0	Statistically significant coefficients as almost all of the coefficients have p-value<0.05 p-values of citric_acid and chlorides are much better.	-
R^2 on Test	**0.31**	**0.33**	**0.45**

CONCLUSIONS

- ALCOHOL CONTENT ∝ QUALITY
- RESIDUAL SUGAR ∝ 1 / (ALCOHOL CONTENT)
  - White wine has more sugar
- INGREDIENTS THAT CAN SPOIL YOUR EXPERIENCE:
  - Volatile Acid - turns to Vinegar
  - SO_2 - rotten egg like smell

RED OR WHITE?

AND THE WINNER IS…

(based on dataset used)

- SWEET FLAVOR
- FLORAL AROMA
- HIGHER ALCOHOL CONTENT

Best choice to make:

PORTUGUESE MADEIRA

SPANISH SHERRY

THANK YOU AND CHEERS!