CLASSIFY RESPONSIBLY:

NAIVE BAYES CLASSIFICATION FOR WINE QUALITY

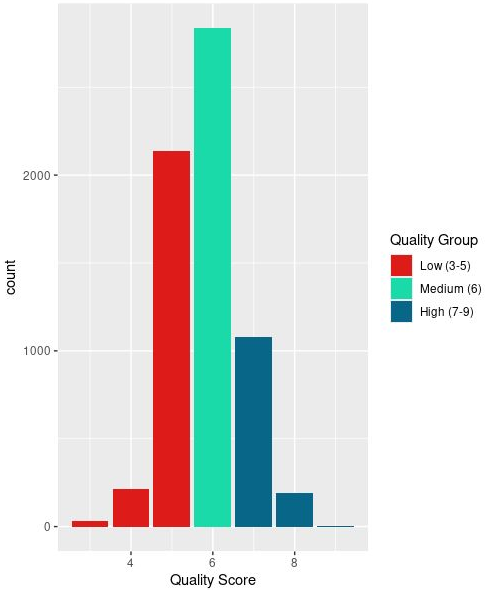
ADAM NORDQUIST

IMPORTANT NOTES

- ABV: Alcohol By Volume
- Volatile Acidity: measure of a wine's gaseous acids
    - Informs smell, which informs taste heavily
- Quality group arbitrarily selected from discrete quality score

7-9

Wines denoted as high quality if quality score falls between 7 and 9 (n = 1,277).



6

Wines denoted as medium quality if quality score falls between 5 and 6 (n = 2,836).

3-5

Wines denoted as low quality if quality score falls between 3 and 5 (n = 2,384).

EDA: 1 Glance

01. Relationships

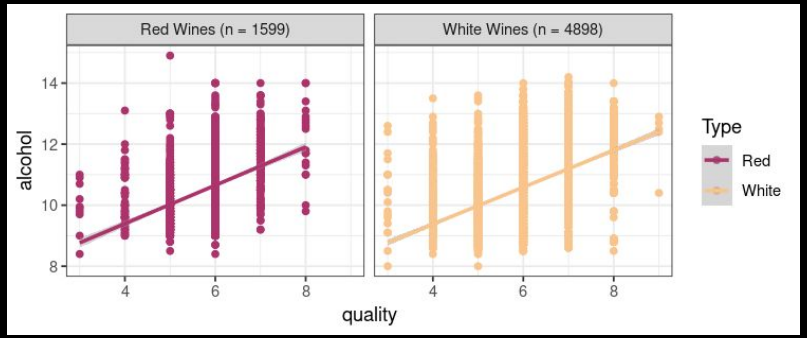
ABV + quality score
Positive, linear

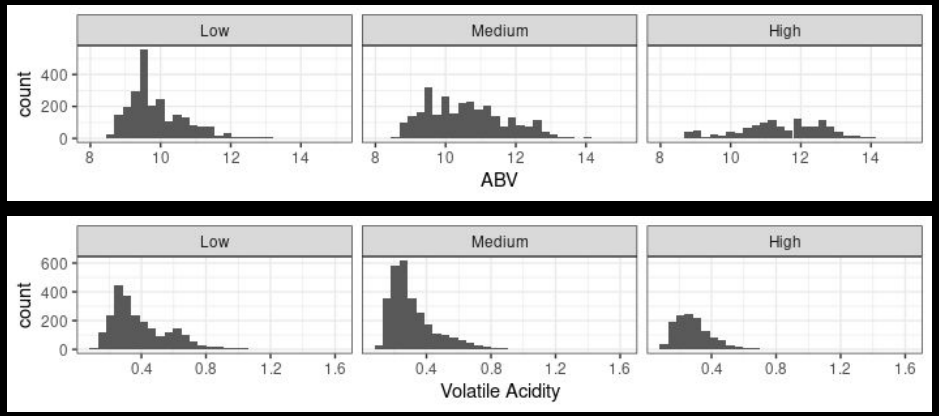
02. Correlation values

Highest 2 w/ Q:
ABV, VA

03. Distributions

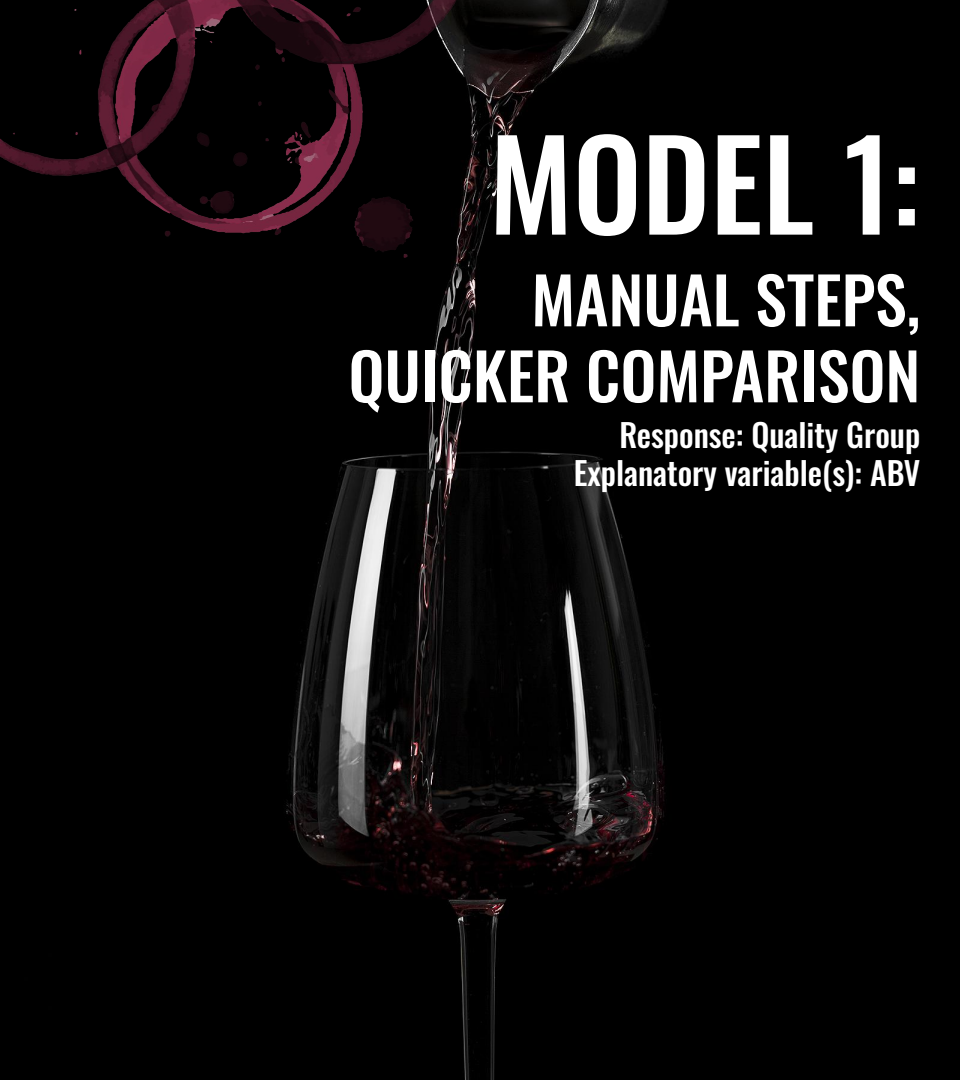
ABV, VA





MODELING

- Manual Modeling
- Quick Modeling
- Models 1-5



MODEL 1:
MANUAL STEPS, QUICKER COMPARISON

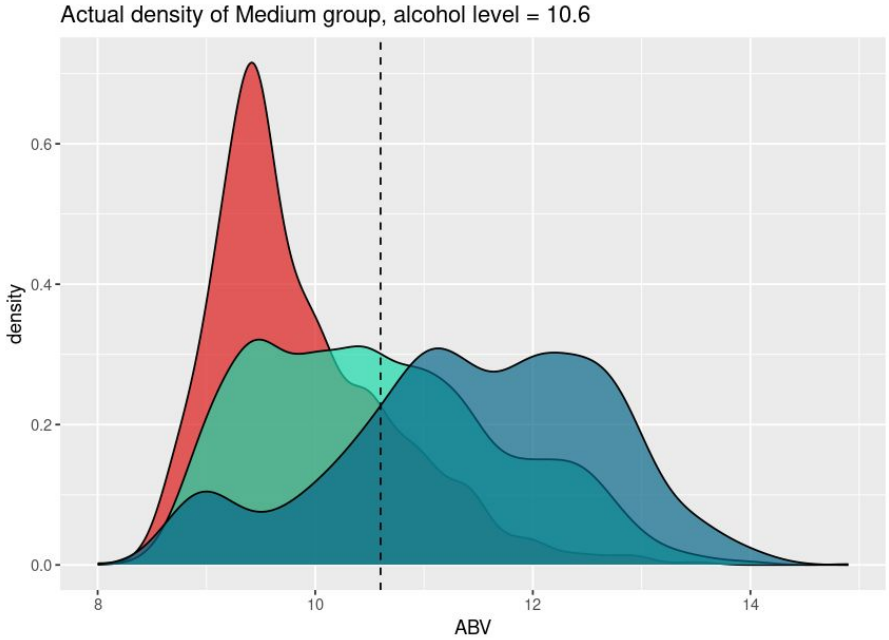
Response: Quality Group
Explanatory variable(s): ABV

Real Data — Observe real data and its density

Normal Priors — Construct normal, naive priors based on the mean and standard deviation of the data

Calculations — Calculate probabilities using Bayes' Rule

THE CHOSEN WINE / DATA



10.6 ABV (X_A)

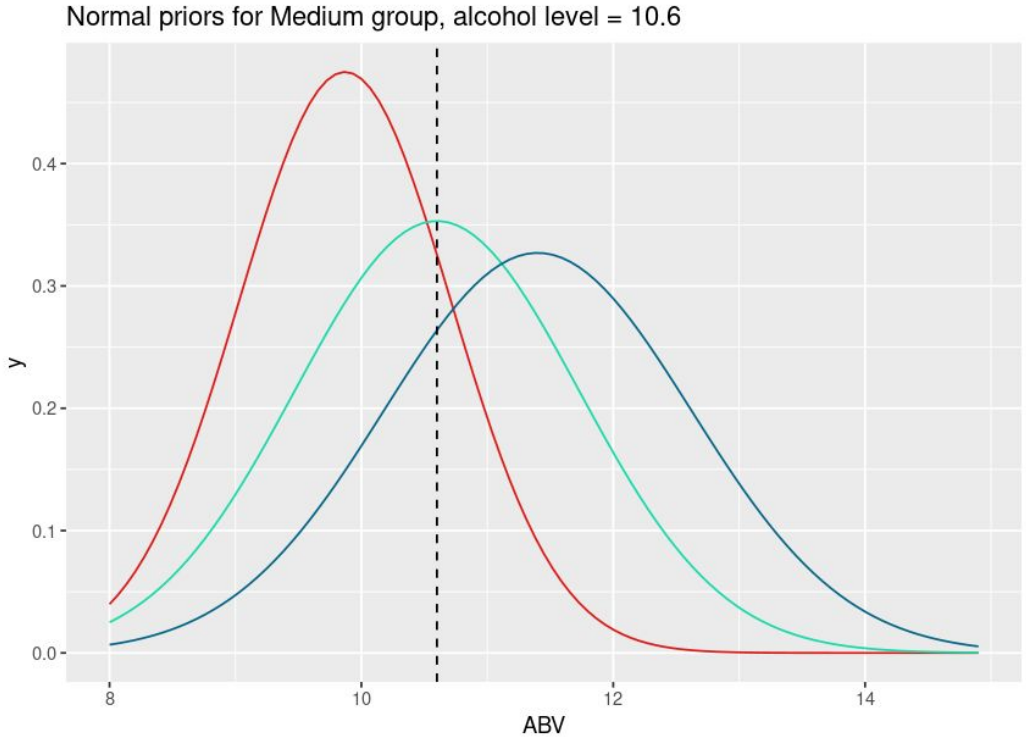
0.685 Volatile acidity (X_V)

6 | Medium Quality Score + Group (Y)

Red Wine type (X_T)

Use the sample means and standard deviations of this data to create the naive priors.
I randomly selected one wine to use as a proof of concept for these models, called the Chosen Wine.

NAIVE PRIORS



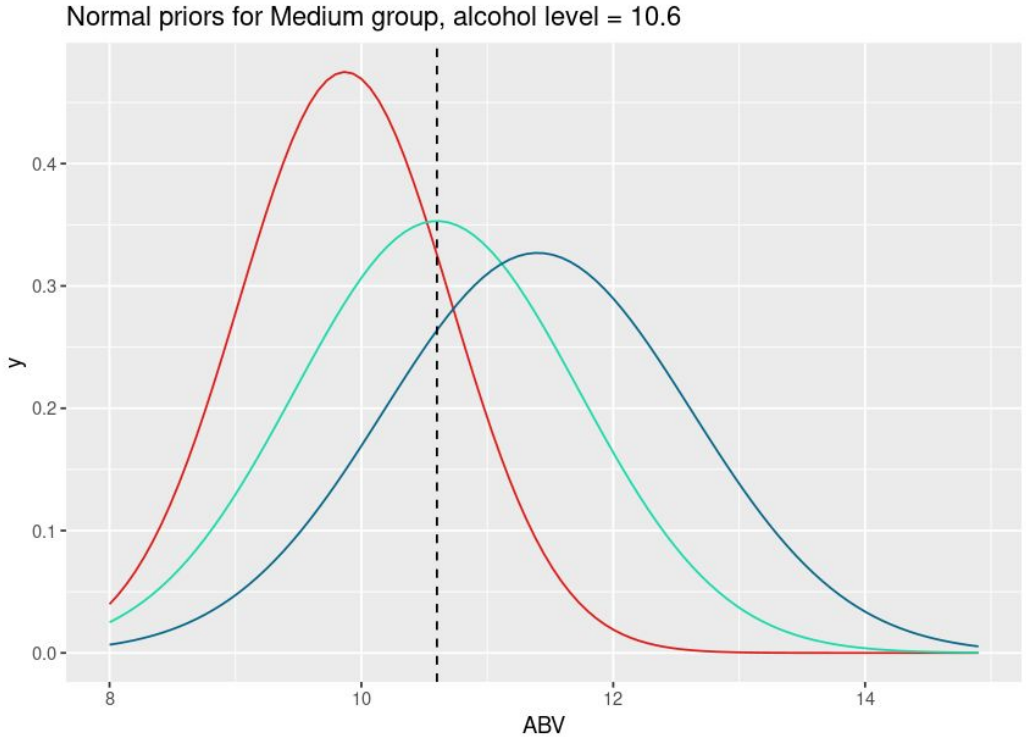
Quality Group	Mean	Median	Std. Dev.
Low	9.87	9.60	0.84
Medium	10.59	10.50	1.13
High	11.43	11.50	1.22

For this example, Y = quality group, X_A = ABV.

- $X_A \mid Y = \text{Low} \sim N(9.87, 0.84)$,
- $X_A \mid Y = \text{Med} \sim N(10.59, 1.13)$,
- $X_A \mid Y = \text{High} \sim N(11.43, 1.22)$

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

NAIVE PRIORS



Quality Group	Mean	Median	Std. Dev.
Low	9.87	9.60	0.84
Medium	10.59	10.50	1.13
High	11.43	11.50	1.22

- Calculate likelihoods ($L(X')$)
  - dnorm(ABV, μ_i, σ_i)
- Calculate normalizing constant (N.C.)
  - $\Sigma\ [\ F(X_A') * L(X_A')\ ]$
- Using Bayes' Rule, calculate probability
  - $(F(X_A') * L(X_A'))/N.C.$

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

0.159

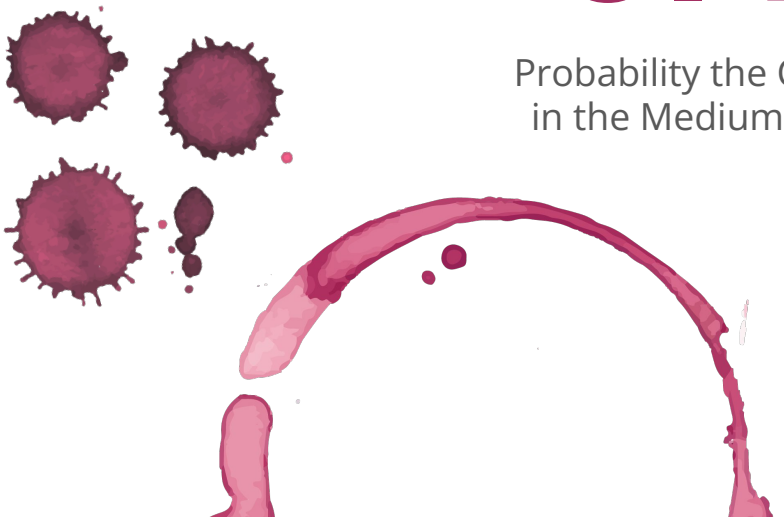
Probability the Chosen Wine is in the High quality group

0.473

Probability the Chosen Wine is in the Medium quality group

0.366

Probability the Chosen Wine is in the Low quality group



MANUAL VS QUICK

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

0.366

Probability the Chosen Wine is in the Low quality group

↓

0.368

0.473

Probability the Chosen Wine is in the Medium quality group

↓

0.475

0.159

Probability the Chosen Wine is in the High quality group

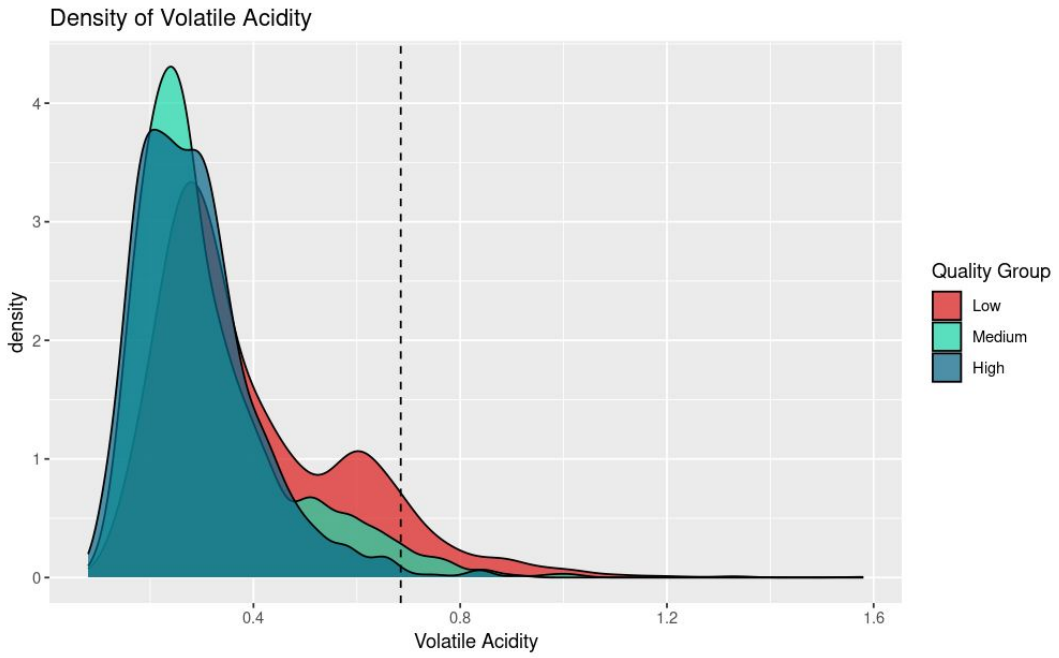
↓

0.156

This “quick” version, the naiveBayes fn from the e1071 package, is what we’ll use for each subsequent model here. It does the same work we just did.

MODEL 2

Response: Quality Group
Explanatory variable(s): Volatile Acidity (X_V)



- Generally difficult to predict
- Most likely outcome for *this specific wine* at a glance: Low (outlier)

0.825 Probability CW in Low group

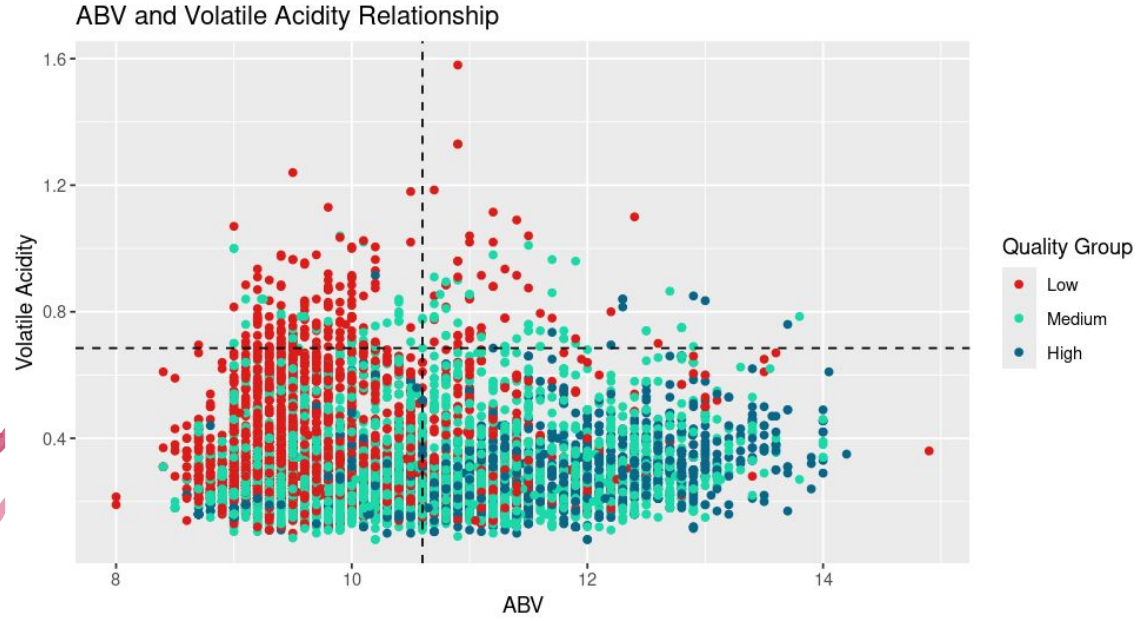
0.168 Probability CW in Medium group

0.007 Probability CW in High group

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

MODEL 3

Response: Quality Group
Explanatory variable(s): ABV (X_A), Volatile Acidity (X_V)



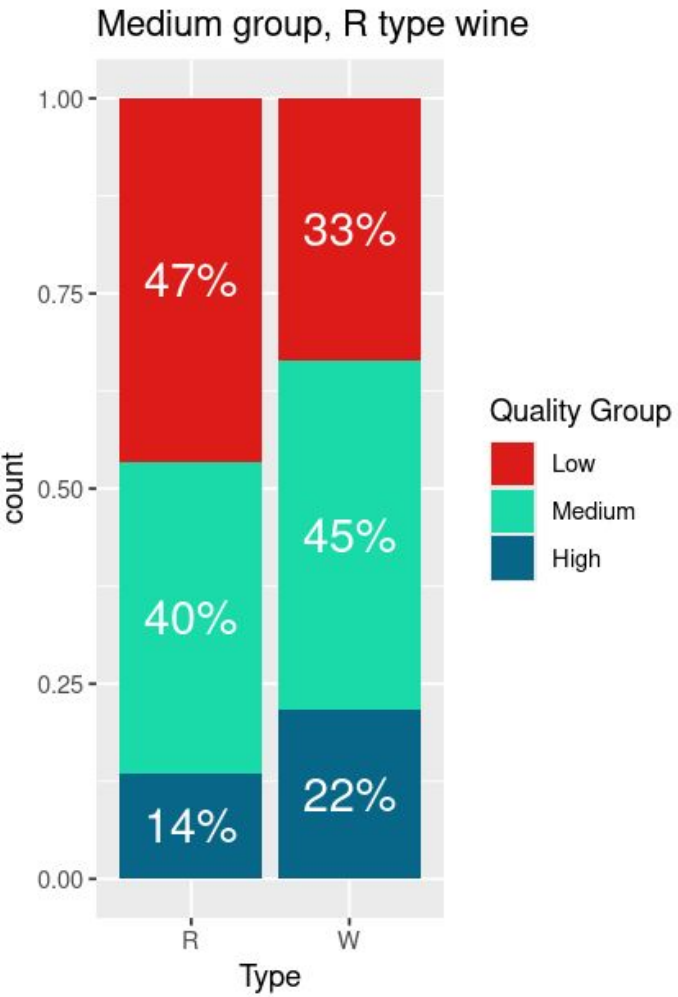
- Some clear separation of groups
- The Chosen Wine is on the precipice of Low vs Medium quality

0.815 Probability CW in Low group

0.179 Probability CW in Medium group

0.006 Probability CW in High group

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R



MODEL 4.1

Response: Quality Group
Explanatory variable(s): Wine Type (X_T)

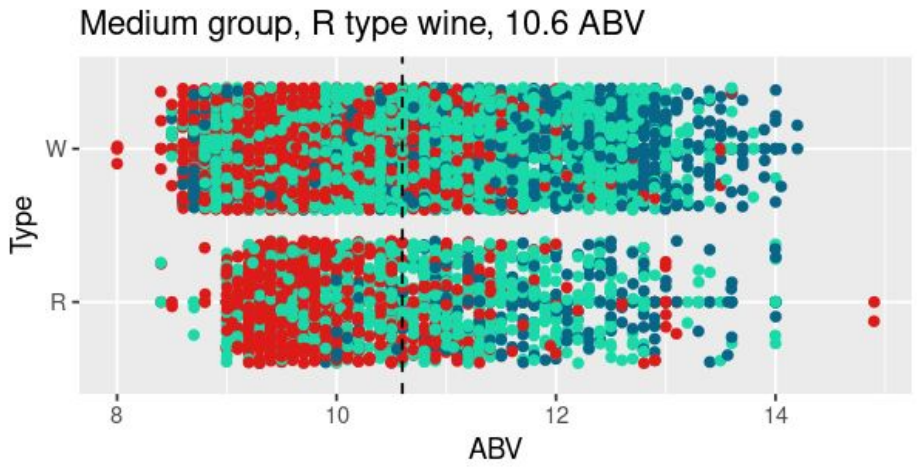
- Similar-ish proportions in W vs R
  - There are about 3 times as many Ws than Rs
- Closely models P(Y | X = R)

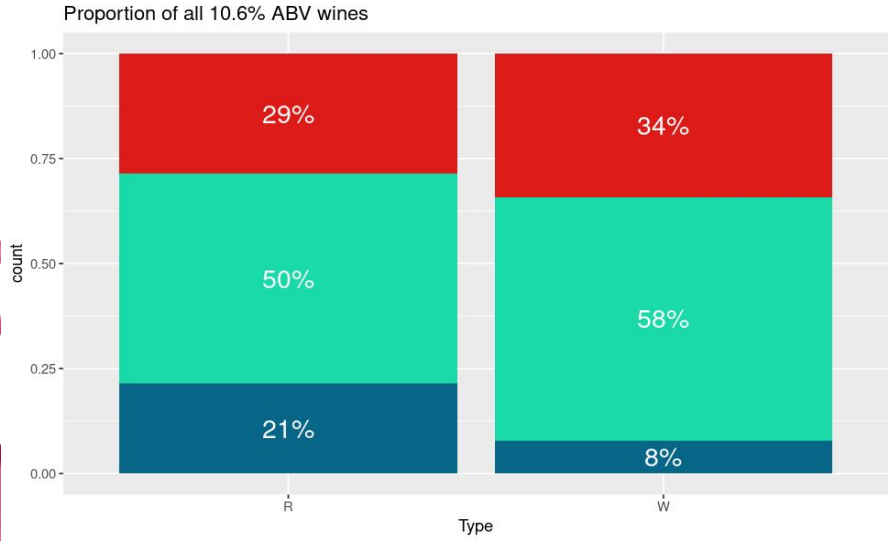
0.465 Probability CW in Low group

0.399 Probability CW in Medium group

0.136 Probability CW in High group

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R





MODEL 4.2

Response: Quality Group
Explanatory variable(s): ABV (X_A), Wine Type (X_T)

- Again, on the precipice
- Out of all 10.6 ABV wines, most likely outcome is Medium

0.463 Probability CW in Low group

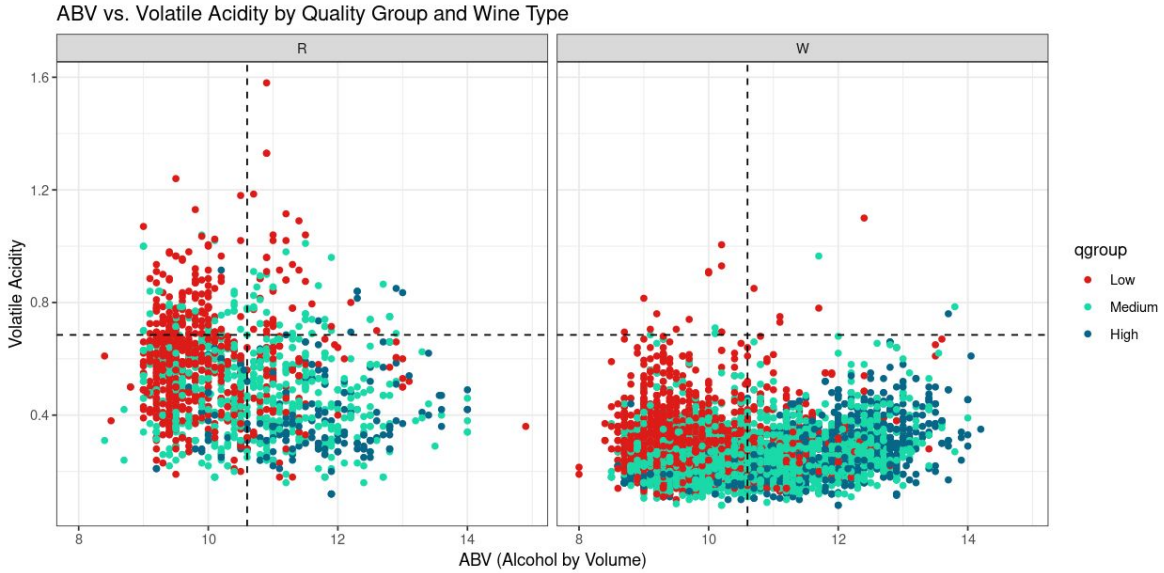
0.430 Probability CW in Medium group

0.107 Probability CW in High group

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

MODEL 5

Response: Quality Group
Explanatory variable(s): ABV (X_A), Volatile Acidity (X_V), Wine Type (X_T)



- Some differences, some clusters
- Clear where volatile acidity helps classification.. in wine type, not quality group

0.860 Probability CW in Low group

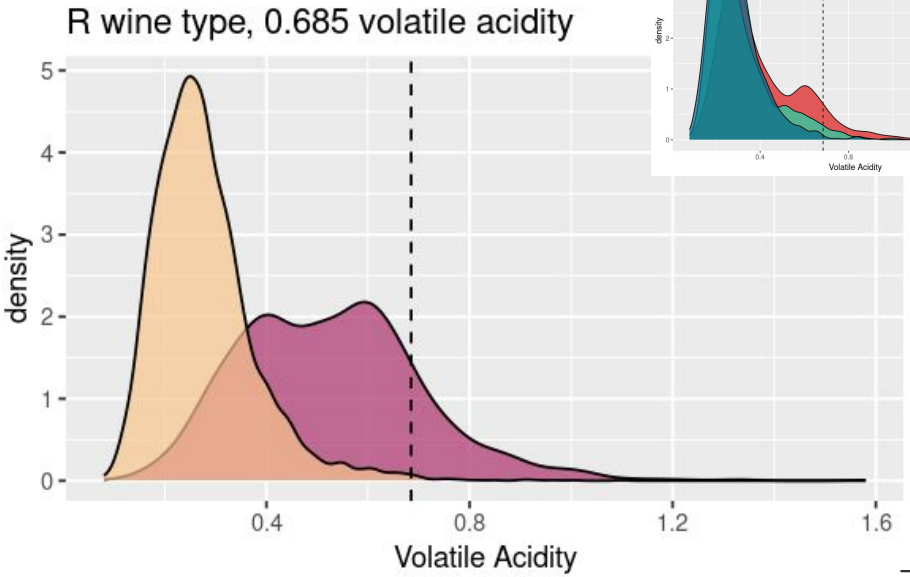
0.137 Probability CW in Medium group

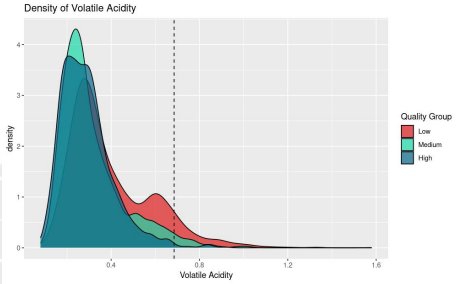
0.003 Probability CW in High group

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

MODEL DEVIATIONS

Response: Wine Type
Explanatory variable(s): Volatile Acidity





- Easier to distinguish between wine type and volatile acidity than wine quality and the same

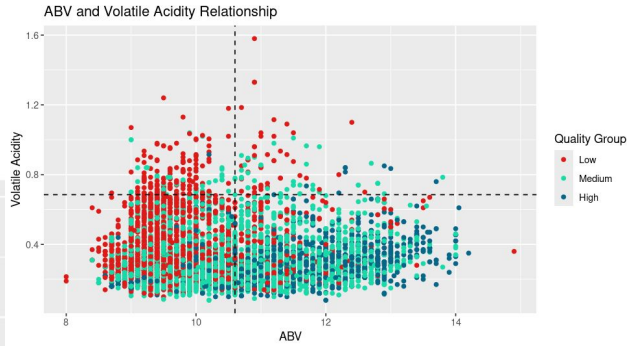
0.997 Probability CW is red wine

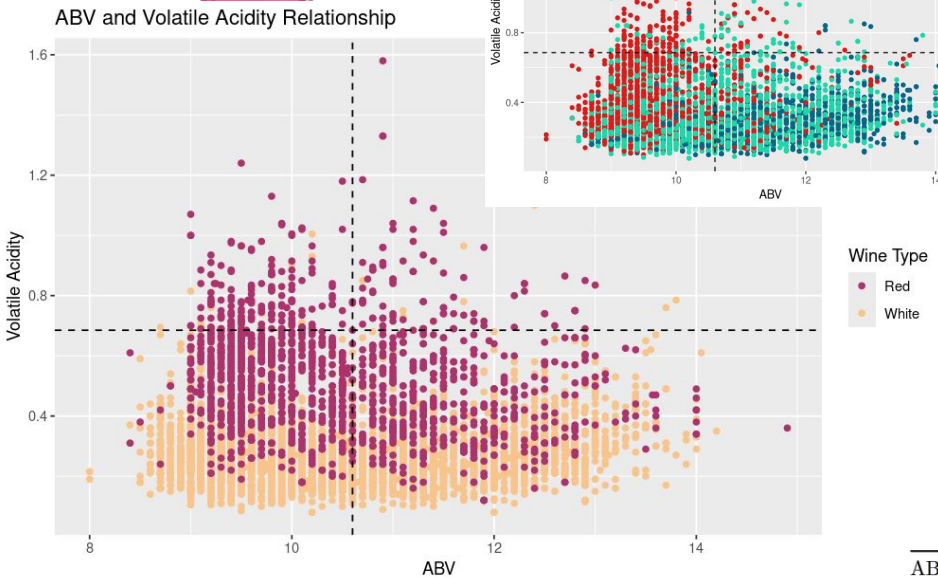
0.003 Probability CW is white wine

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

MODEL DEVIATIONS

Response: Wine Type
Explanatory variable(s): ABV, Volatile Acidity





- Both together: no noticeable improvement from just VA

0.998 Probability CW is red wine

0.002 Probability CW is white wine

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

MODEL DEVIATIONS

Response: Wine Type
Explanatory variable(s): ABV, Quality Group*

- Quality group absolutely tanks probabilities

0.249 Probability CW is red wine

0.751 Probability CW is white wine

ABV	Volatile Acidity	Quality Score	Quality Group	Wine Type
10.6	0.685	6	Medium	R

MODELING COMPARISON

CV ACCURACY: WINE TYPE MODELS

ABV + Quality Group + Volatile Acidity

Max Acc: 0.701
Avg Acc: 0.613

ABV + Volatile Acidity

Max Acc: 0.680
Avg Acc: 0.606

ABV + Quality Group

Max Acc: 0
Avg Acc: 0

CV ACCURACY: QUALITY MODELS

ABV

Max Acc: 0.563

Avg Acc: 0.525

Volatile Acidity

Max Acc: 0.526

Avg Acc: 0.481

ABV + Volatile Acidity*

Max Acc: 0.575

Avg Acc: 0.543

Wine Type

Max Acc: 0.476

Avg Acc: 0.452

ABV + Wine Type

Max Acc: 0.574

Avg Acc: 0.528

ABV + Wine Type + Volatile Acidity*

Max Acc: 0.583

Avg Acc: 0.541

- Highest average accuracy: .543 and .541, 2/3-predictor models
- 3-predictor highest single accuracy score (.583)

CONFUSION MATRIX: BEST* MODEL(S)

Model 3 (two-predictor model, ABV + Volatile Acidity)

	Low	Medium	High
Low	66.57% (1587)	32.05% (764)	1.38% (33)
Medium	33.60% (953)	54.41% (1543)	11.99% (340)
High	11.75% (150)	56.77% (725)	31.48% (402)

Model 5 (three-predictor model, ABV + Volatile Acidity + Wine Type)

	Low	Medium	High
Low	53.61% (1278)	44.84% (1069)	1.55% (37)
Medium	23.27% (660)	64.03% (1816)	12.69% (360)
High	5.32% (68)	61.47% (785)	33.20% (424)

- Quality group is not easy to predict well
- Two-predictor model good at Low accuracy but not Medium, vice versa for three-predictor
- Both models not great at predicting High quality wines
  - There are not many High quality wines (n=1277)
  - Because of that, the characteristics are less separated than they are at the L and M levels

THANKS

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**