Predicting Wine Quality for California Wine Institute

Zacharia Schmitz, Junior Data Scientist
Joshua Click, Junior Data Scientist
September 21, 2023

- Identify features
- Develop clusters
- Build a model
- Tune the model

- Identify features
- Develop clusters
- Build a model
- Tune the model

Objectives

Big Idea

Findings

Proposal

Model:
RandomForestClassifier

3 Clusters

StandardScaler

Features
Volatile Acidity
Chlorides
Density
Alcohol
is_red

- Identify features
- Develop clusters
- Build a model
- Tune the model

Objectives

Big Idea

Model:
RandomForestClassifier

3 Clusters

StandardScaler

Features
Volatile Acidity
Chlorides
Density
Alcohol
is_red

Findings

- RandomForestClassifier was best with defaults hyperparameters
- Prone to being overfit
- 5 Features & 3 Clusters

56%

Proposal

Objectives

- Identify features
- Develop clusters
- Build a model
- Tune the model

Big Idea

Model:
RandomForestClassifier

3 Clusters

StandardScaler

Features
Volatile Acidity
Chlorides
Density
Alcohol
is_red

Findings

- RandomForestClassifier was best with defaults hyperparameters
- Prone to being overfit
- 5 Features & 3 Clusters

56%

Proposal

Recommend keeping red and white wine data frames separate

Continue to feature engineer to be able to build better models in future

PLAN

ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

1. Identify the need for the model

2. Collect the data

3. Narrow down and test features

4. Develop clusters

4. Test modeling with features and clusters

PLAN

ACQUIRE

ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

The data is acquired from
data.world/food/wine-quality
as 2 datasets and combined into one

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

Combined:
6,497 rows
12 columns

Target Variable
quality

PLAN
ACQUIRE

PREPARE

EXPLORE
MODEL
DELIVER

The Features

is_red
(Red: 1 White: 0)

After preparing:

6,497 rows

6 columns

DF Name	variable type	dtype
quality	**ordinal**	**int**
volatile_acidity	continuous	float
chlorides	continuous	float
density	continuous	float
alcohol	continuous	float
is_red	nominal	int

PLAN
ACQUIRE
PREPARE

EXPLORE

MODEL
DELIVER

Train (60%)
Validate (20%)
Test (20%)

Non Parametric
&
Not Normal

α = 0.05

(Confidence level of 95%)

For continuous vs. categorical variables we binned quality into two groups and used

Mann-Whitney U

PLAN
ACQUIRE
PREPARE

EXPLORE

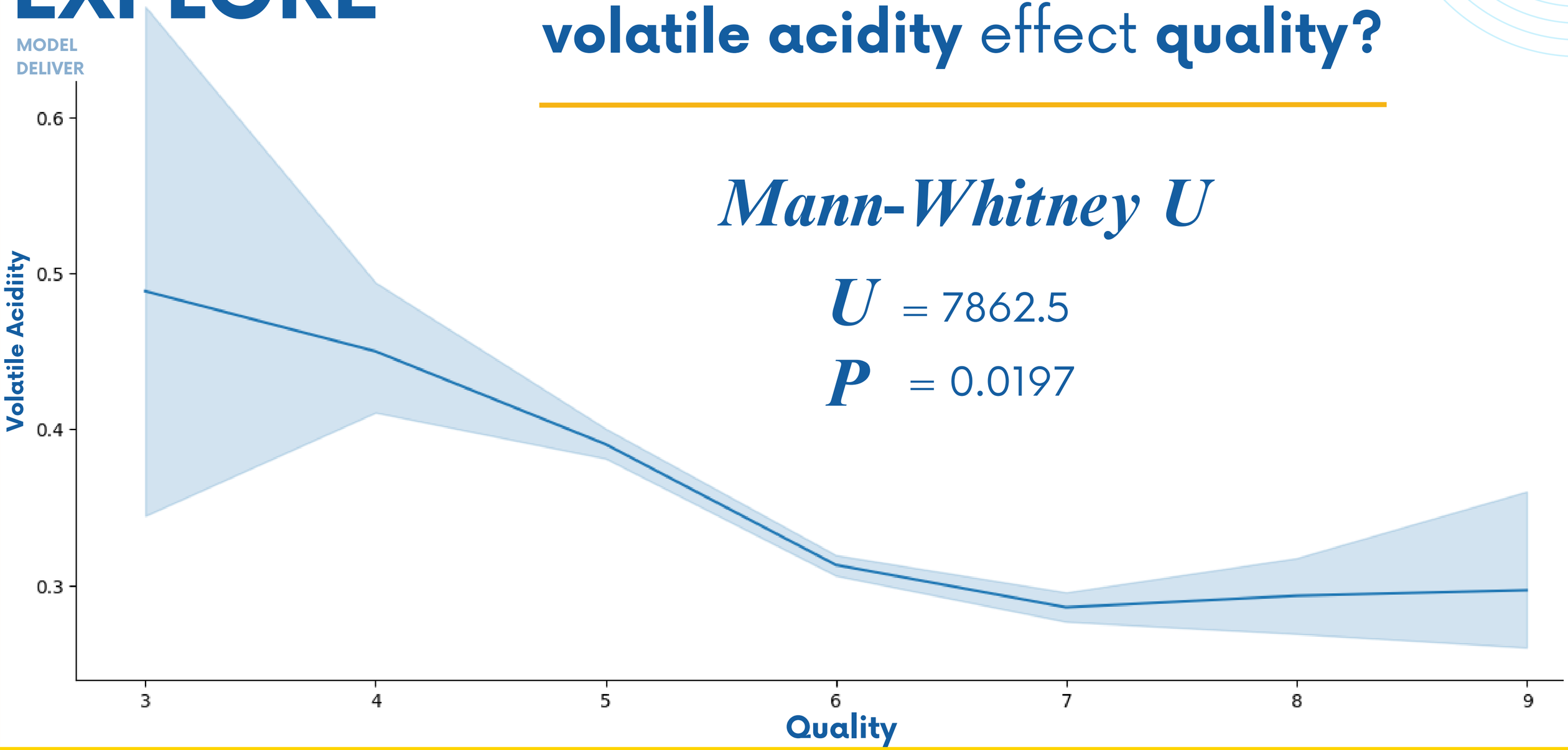
MODEL
DELIVER

How does **volatile acidity** effect **quality?**

Mann-Whitney U

$U = 7862.5$

$P = 0.0197$

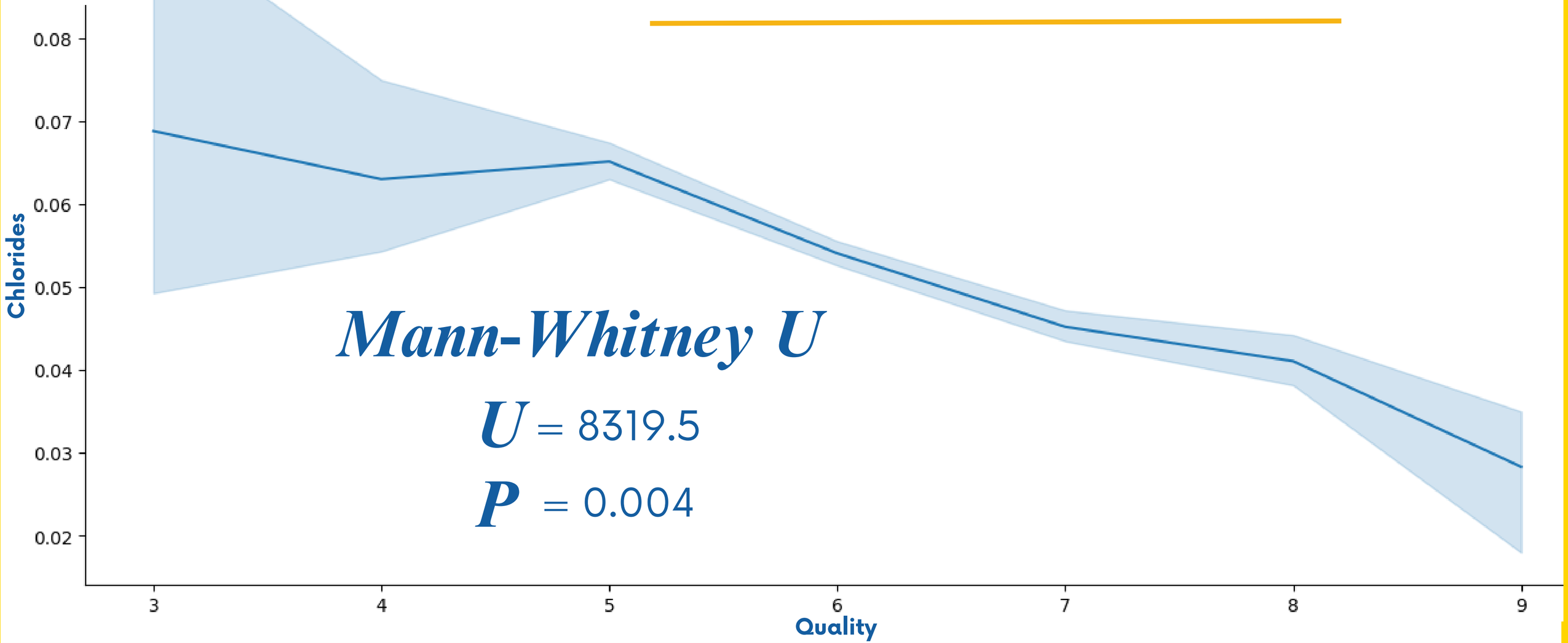


PLAN
ACQUIRE
PREPARE

EXPLORE

MODEL
DELIVER

How do **chlorides** effect **quality?**



Mann-Whitney U

$U = 8319.5$

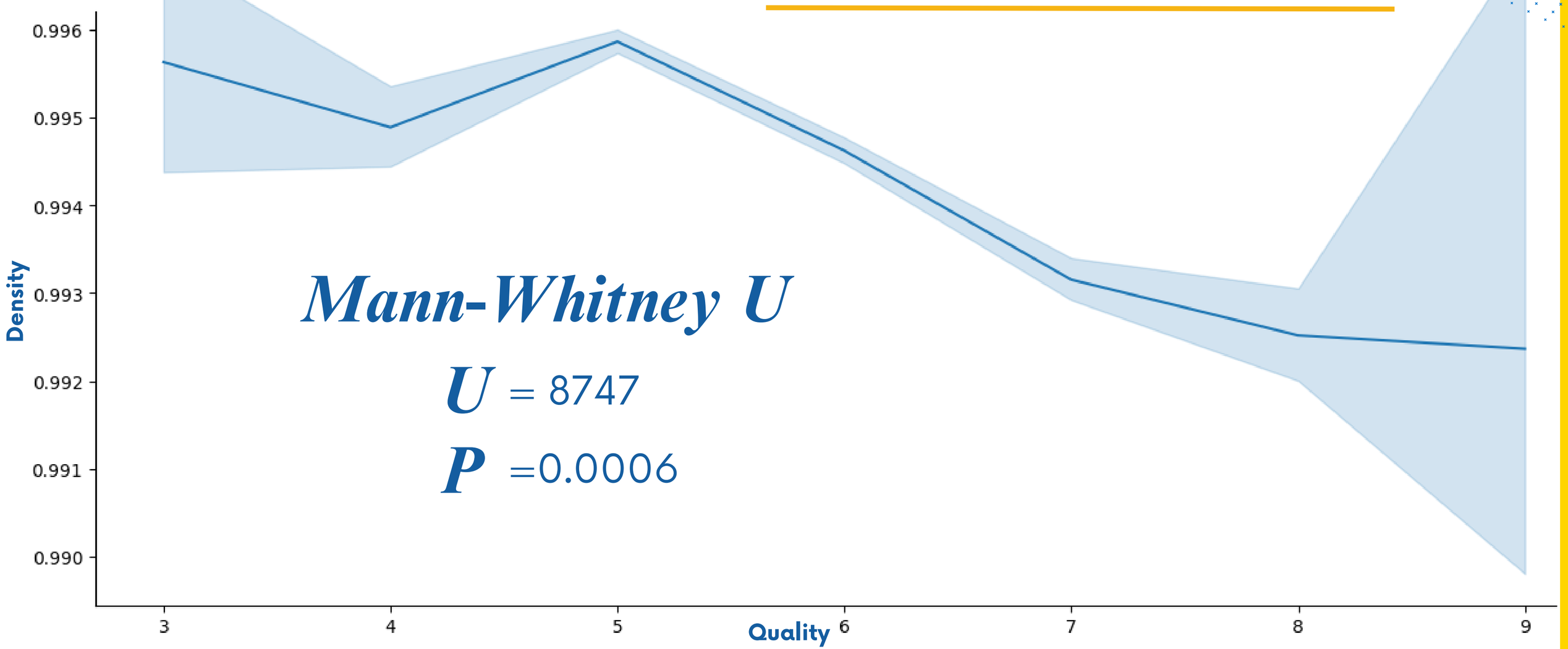
$P = 0.004$

PLAN
ACQUIRE
PREPARE

EXPLORE

MODEL
DELIVER

How does **density** effect **quality?**



Mann-Whitney U

$U = 8747$

$P = 0.0006$

PLAN
ACQUIRE
PREPARE

EXPLORE

MODEL
DELIVER

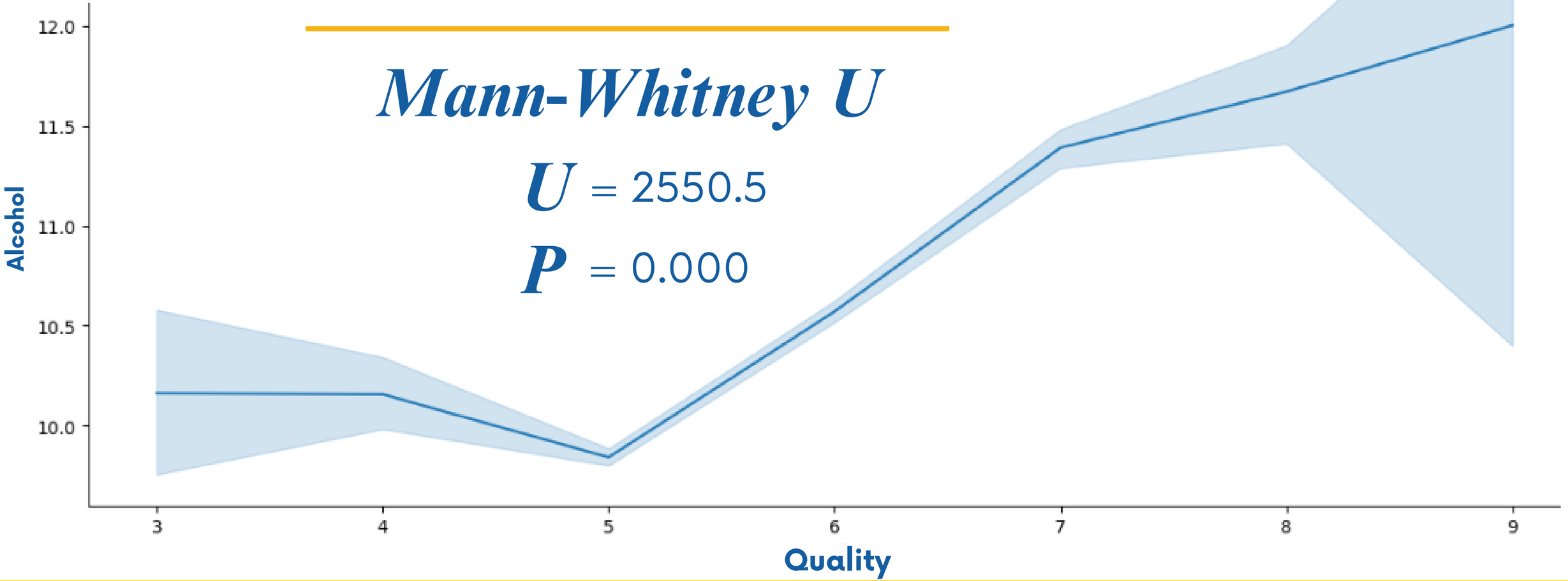
How does

alcohol effect **quality?**

Mann-Whitney U

$U = 2550.5$

$P = 0.000$



PLAN
ACQUIRE
PREPARE

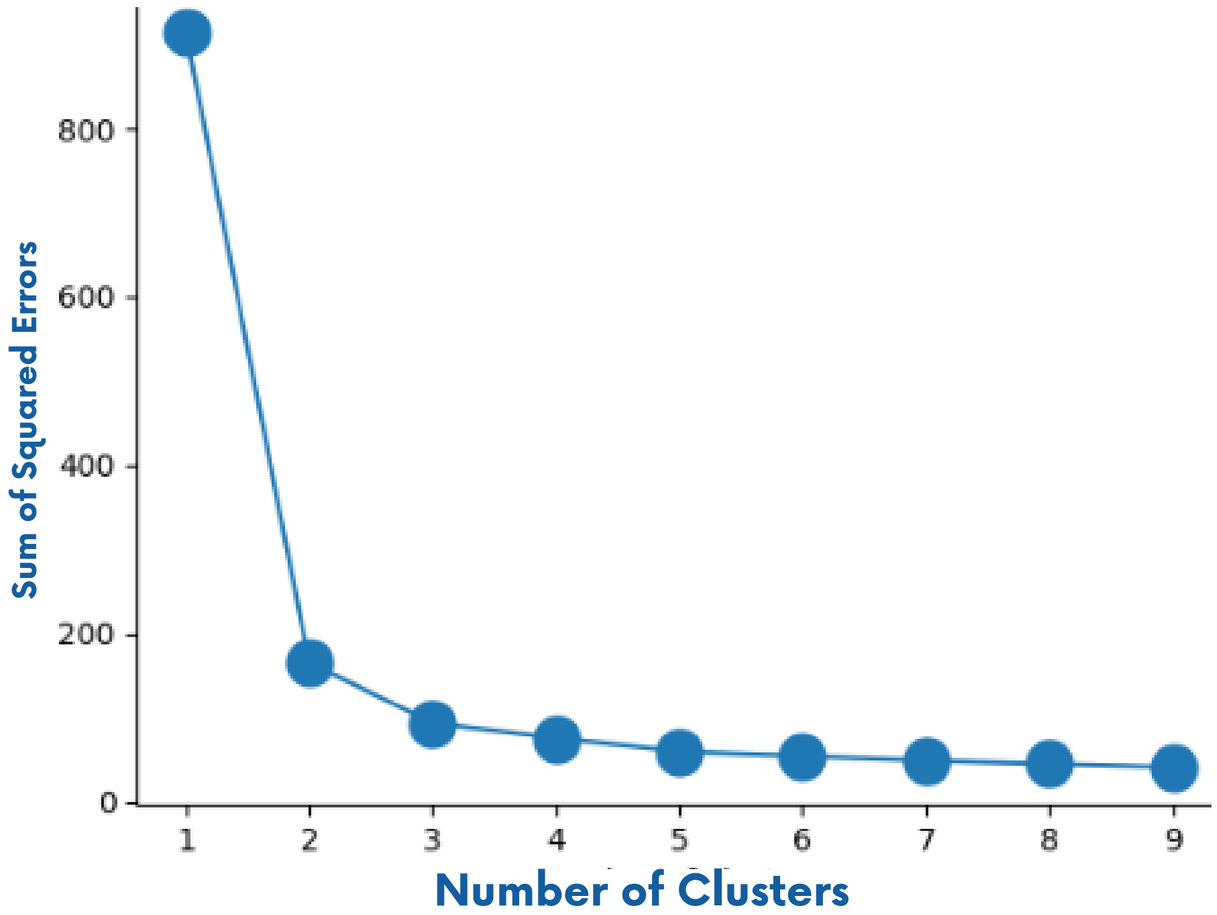
EXPLORE

MODEL
DELIVER

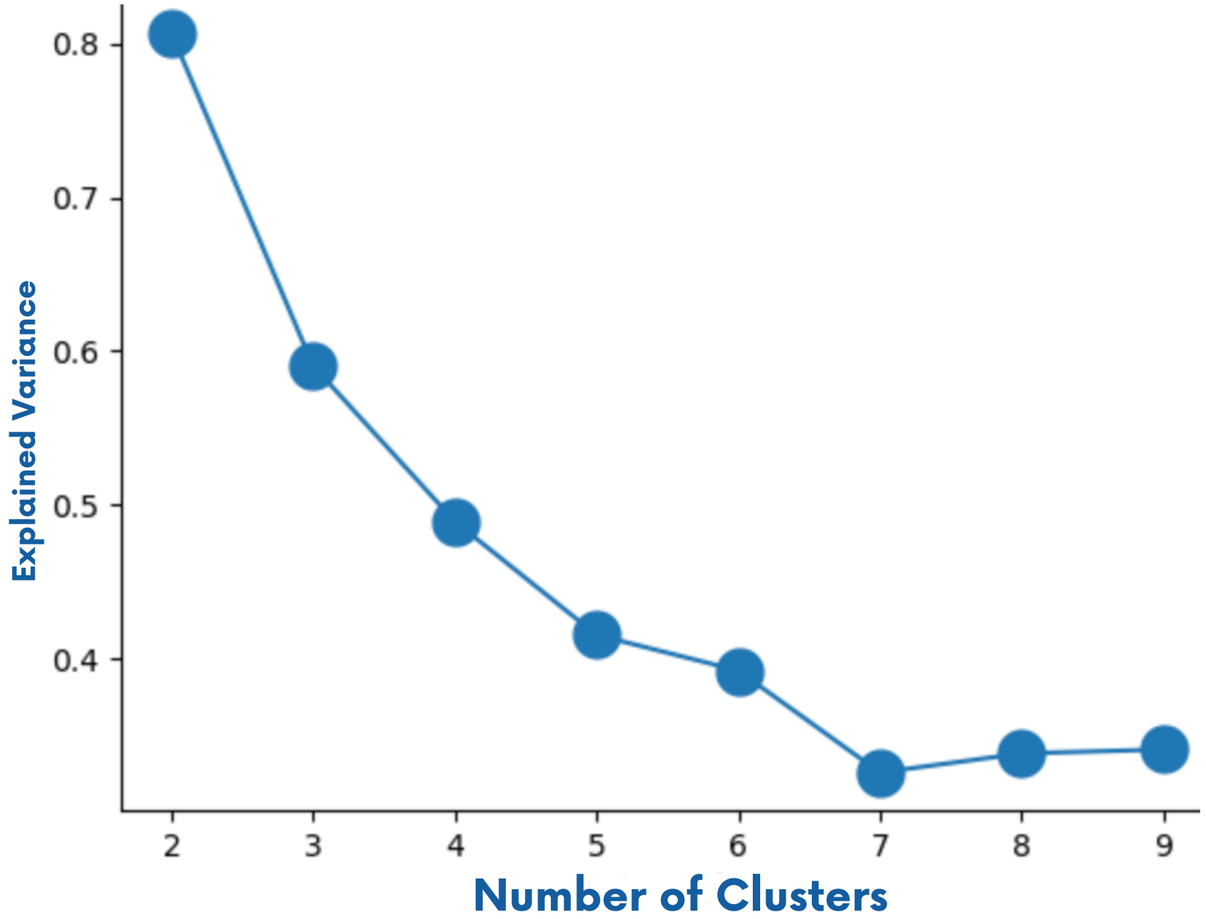
How many

Clusters?

Elbow Method



Silhouette Score



PLAN
ACQUIRE
PREPARE

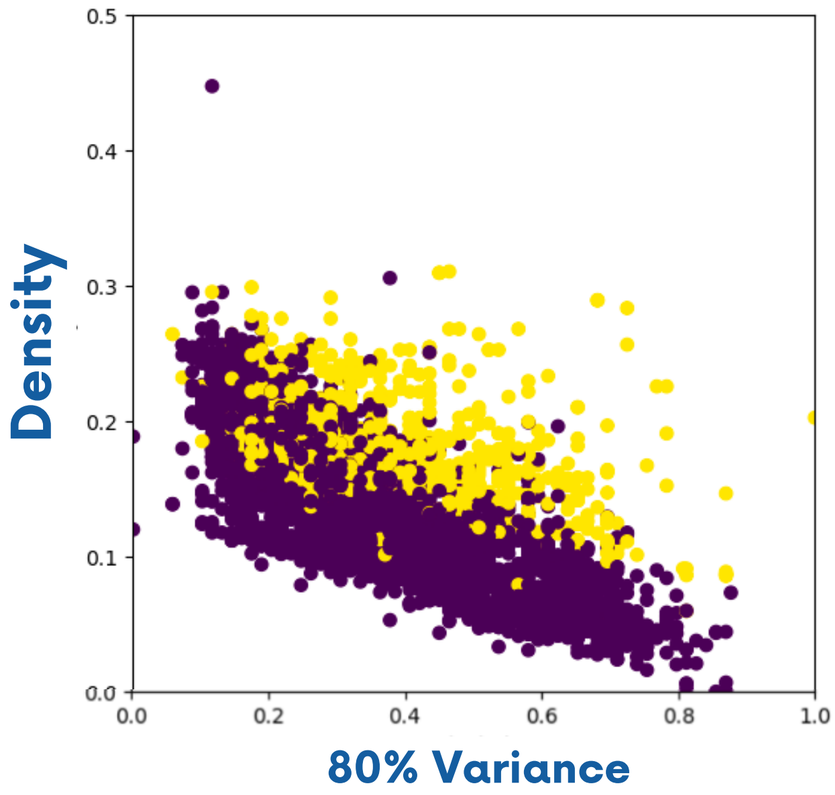
EXPLORE

MODEL
DELIVER

How many

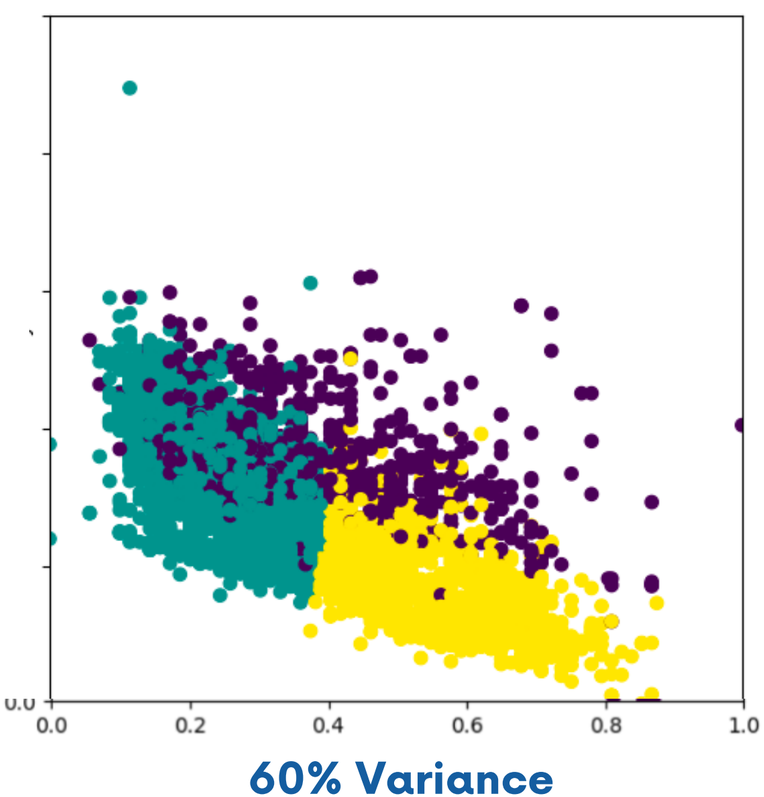
Clusters?

2 Clusters



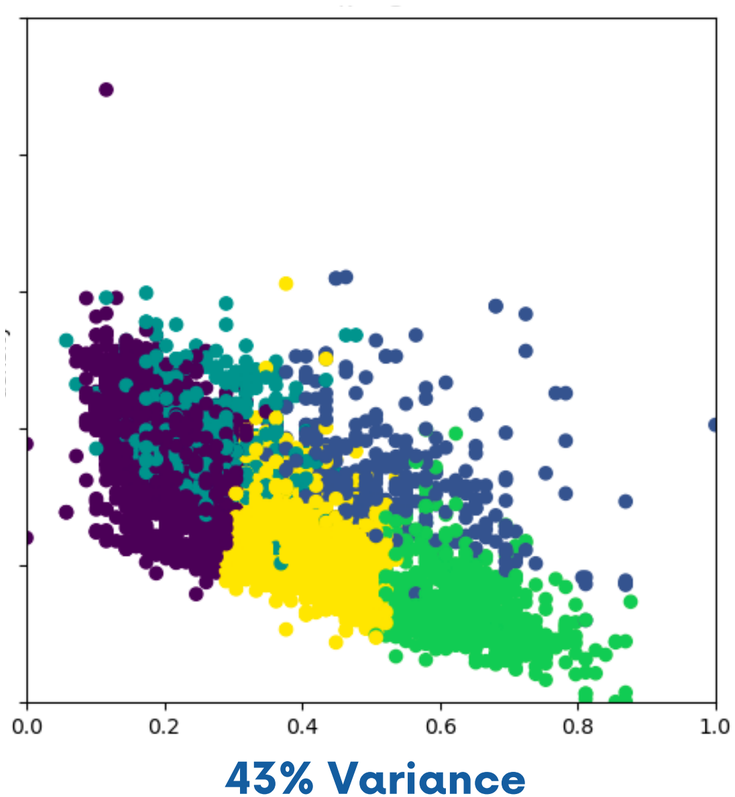
80% Variance

3 Clusters



60% Variance

5 Clusters



43% Variance

Alcohol

PLAN
ACQUIRE
PREPARE
EXPLORE

MODEL

DELIVER

594 RFC Models

These are the top 4

RandomForestClassifier					
Scaler	Parameters	Clusters	Accuracy		
			Train	Val	Test
Baseline Using Mean			0.436		
Standard	**n_estimators=300 max_depth=6 min_samples_split=10 min_samples_leaf=4**	**3**	**0.601**	**0.556**	**0.562**
MinMax	n_estimators=200 max_depth=6 min_samples_split=5 min_samples_leaf=1	2	0.604	0.555	NA
Standard	n_estimators=300 max_depth=6 min_samples_split=2 min_samples_leaf=1	5	0.609	0.555	NA
Standard	n_estimators=300 max_depth=6 min_samples_split=5 min_samples_leaf=1	3	0.604	0.555	NA

PLAN
ACQUIRE
PREPARE
EXPLORE

MODEL

DELIVER

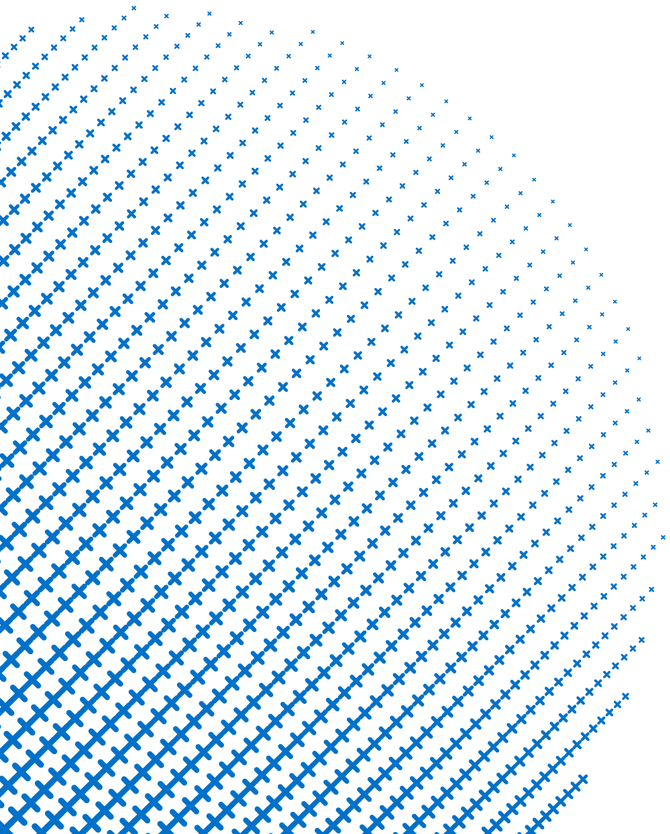
594 RFC Models

These are the top 4

RandomForestClassifier					
Scaler	Parameters	Clusters	Accuracy		
			Train	Val	Test
Baseline Using Mean			**0.436**		
Standard	**n_estimators=300 max_depth=6 min_samples_split=10 min_samples_leaf=4**	**3**	**0.601**	**0.556**	**0.562**
MinMax	n_estimators=200 max_depth=6 min_samples_split=5 min_samples_leaf=1	2	0.604	0.555	NA
Standard	n_estimators=300 max_depth=6 min_samples_split=2 min_samples_leaf=1	5	0.609	0.555	NA
Standard	n_estimators=300 max_depth=6 min_samples_split=5 min_samples_leaf=1	3	0.604	0.555	NA

PLAN
ACQUIRE
PREPARE
EXPLORE
MODEL

DELIVER



Takeaways and Key Findings

We expect the model to perform well on future data

The features we identified were valuable in the model.

With only 5 of the original features, and 3 clusters, it performed well.

PLAN
ACQUIRE
PREPARE
EXPLORE
MODEL

DELIVER

Recommendations

Model based on wine type

More ratings for qualities.

Next Steps

More features and other models with hyperparameters for increased accuracy

PLAN
ACQUIRE
PREPARE
EXPLORE
MODEL

DELIVER

Recommendations

Model based on wine type

More ratings for qualities.

Next Steps

More features and other models with hyperparameters for increased accuracy

CALIFORNIA WINES

Zacharia Schmitz, Junior Data Scientist
Joshua Click, Junior Data Scientist
September 20, 2023

Post Presentation Feedback

Title Slide:
Wordy Title

Exec Slide:
Too wordy, bullets of 3-5 words (generally)

Prepare Slide:
Explain how we binned target variable for testing

Explore Slides:
Graphs should be of binned variable

Model Slides:
Only show the test model compared to baseline, briefly talk about other models

General Feedback:
Fill in white space, utilize real estate
Tables had too much
Less script reading