Visualising
Machine Learning Models

Amit Kapoor
@amitkaps

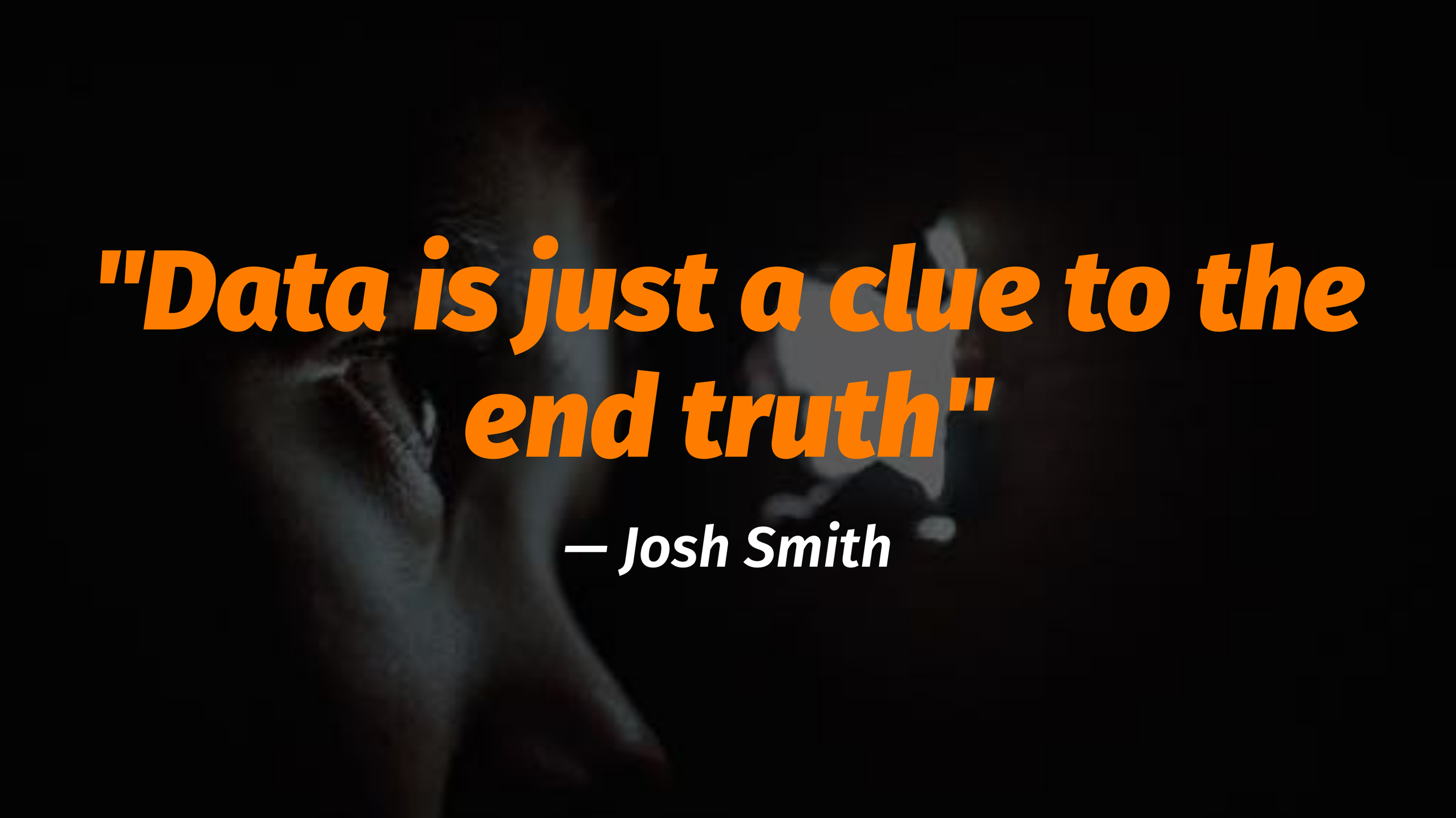

We don’t see things as they are, we see them as we are.
— Anais Nin

衆聲

摸象之圖

"And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong."

— John Godfrey Saxe


"Data is just a clue to the end truth"
— Josh Smith

Making sense of the world through a data lens

$$\theta(t) = 0.25\pi\, e^{-0.2t} cos(2.55t)$$

Layers of Abstraction

Data Abstraction
Visual Abstraction
Model Abstraction

"Geometry without algebra is dumb! Algebra without geometry is blind!"

— David Hestenes

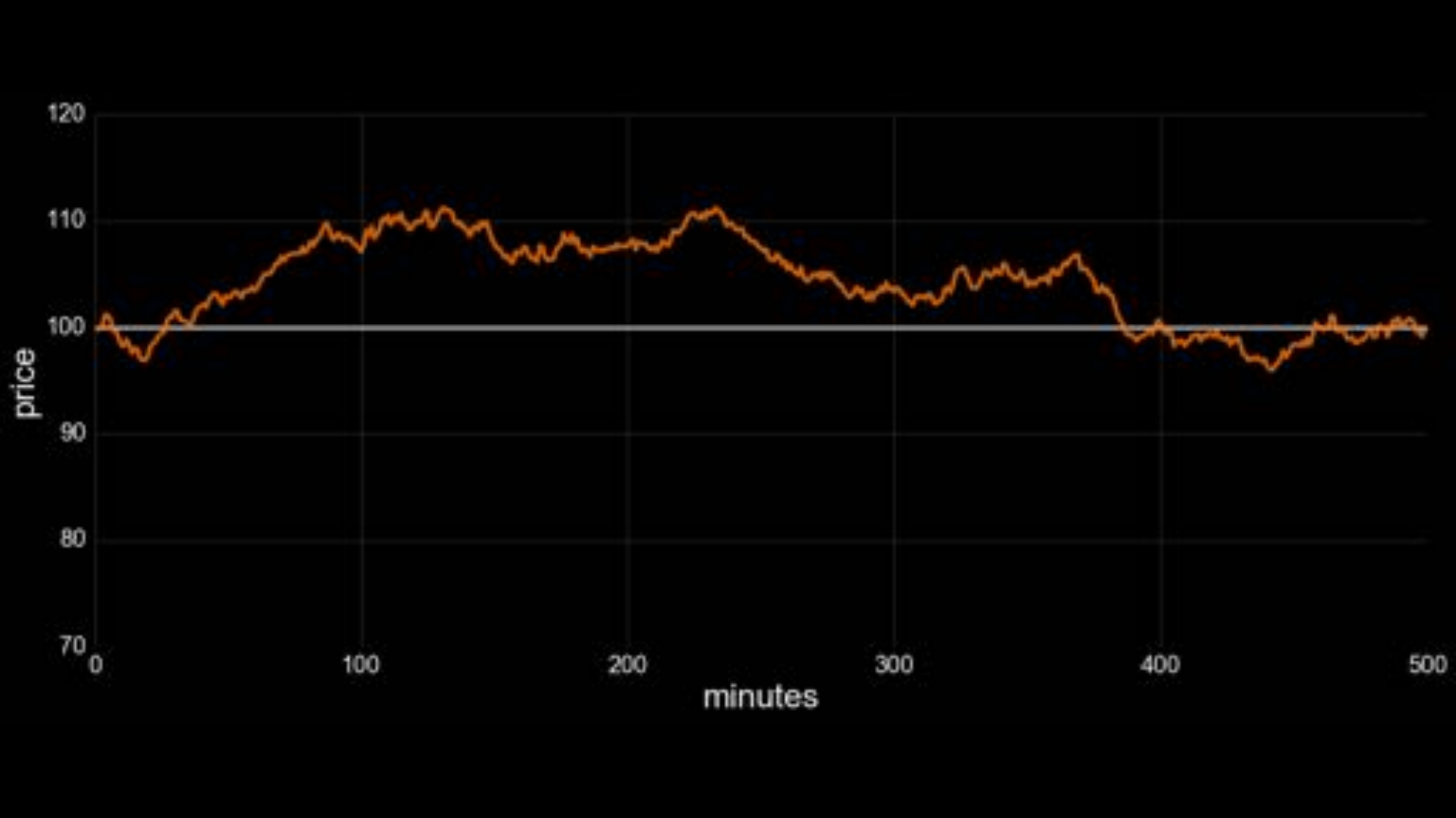

Trading Game
When to Buy, Hold or Sell the stock?


Stock of Shares
Start at 1000 shares


Price of Shares
Starting at Price SGD 100


Time in a day
500 minutes of trading

Visualise the data


120
110
100
90
80
70
price
0
100
200
300
400
500
minutes

Simple Model

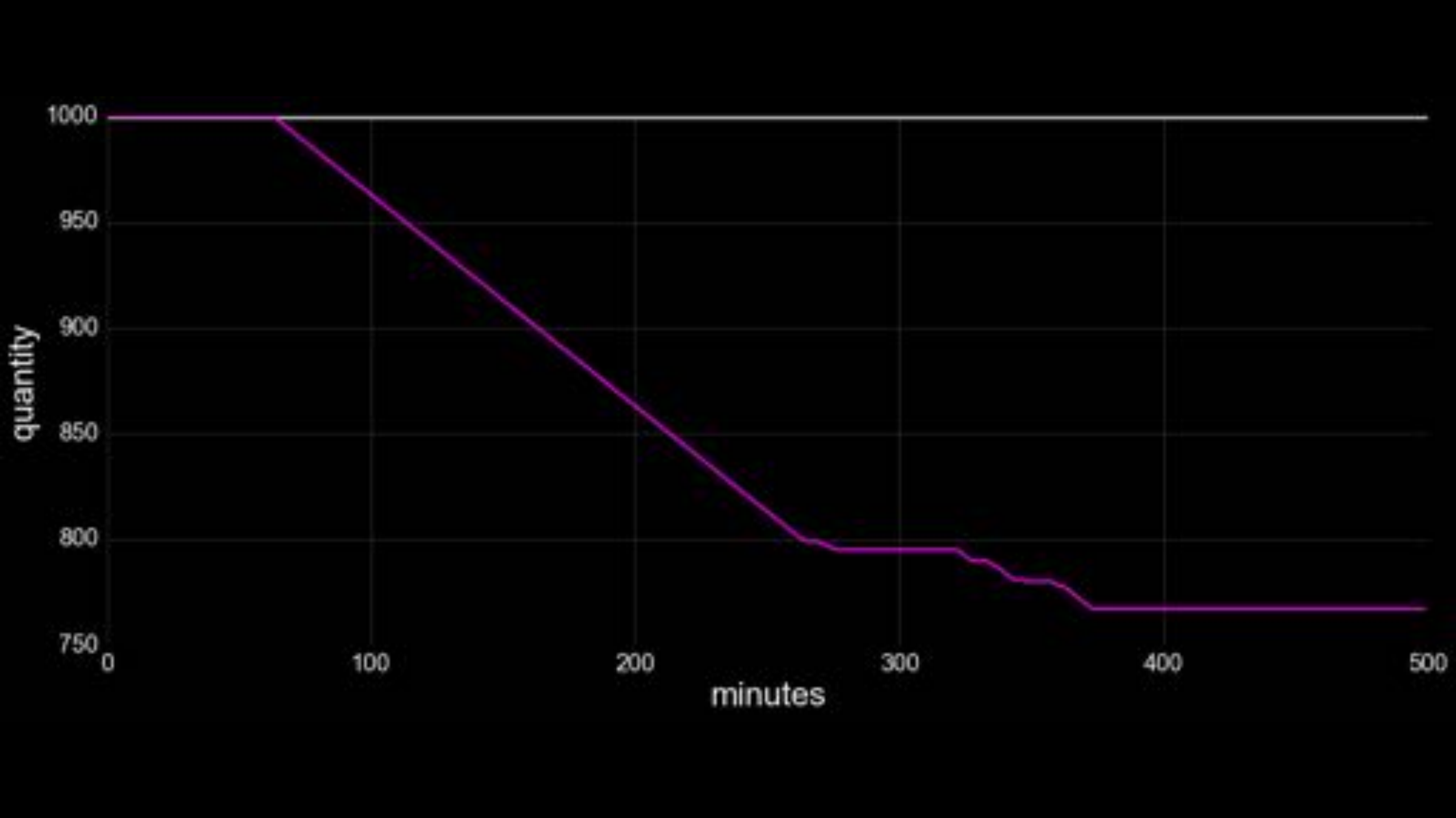
One Minute Strategy

If price > 5%, then Sell 1 share

If price < 5%, then Buy 1 share

Visualise the model

(1) within the data space

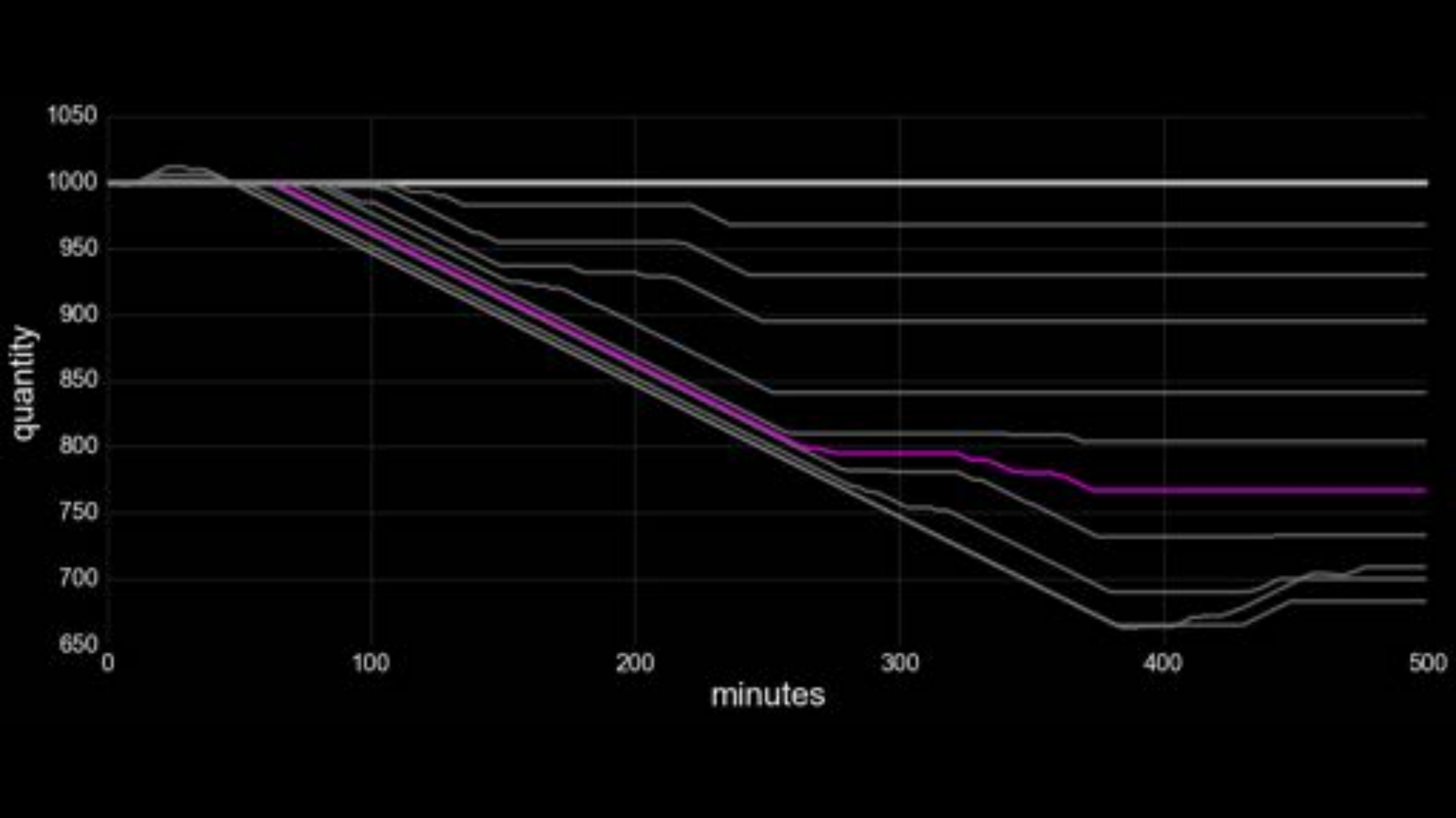

1000
950
900
850
800
750
quantity
0
100
200
300
400
500
minutes

Visualise the model

(2) with varying model parameters

(3) for the process of model fitting

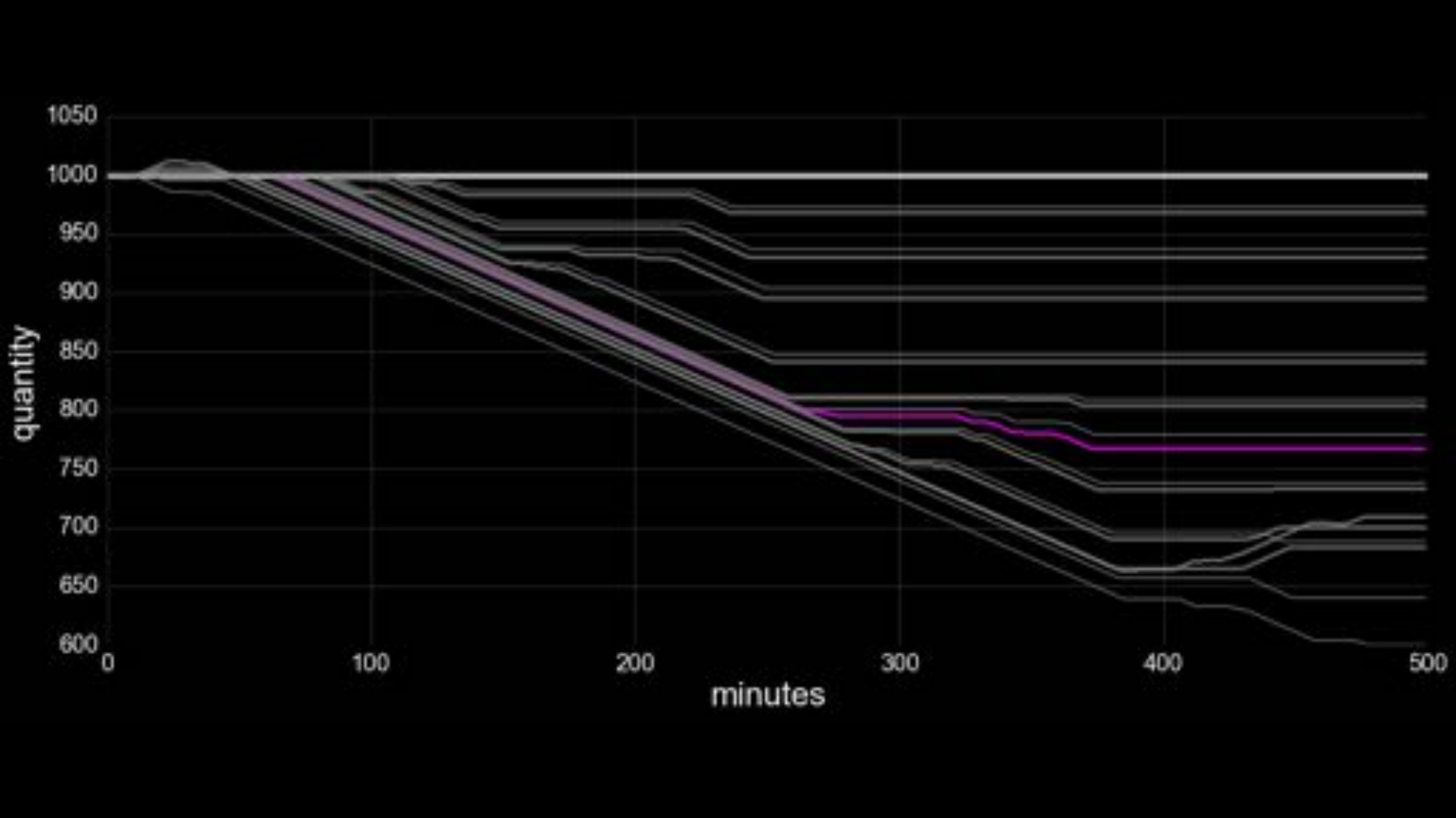
Model Bounds - From 1% to 10%


1050
1000
950
900
850
800
750
700
650
quantity
0
100
200
300
400
500
minutes

Visualise the model

(4) with entire model space

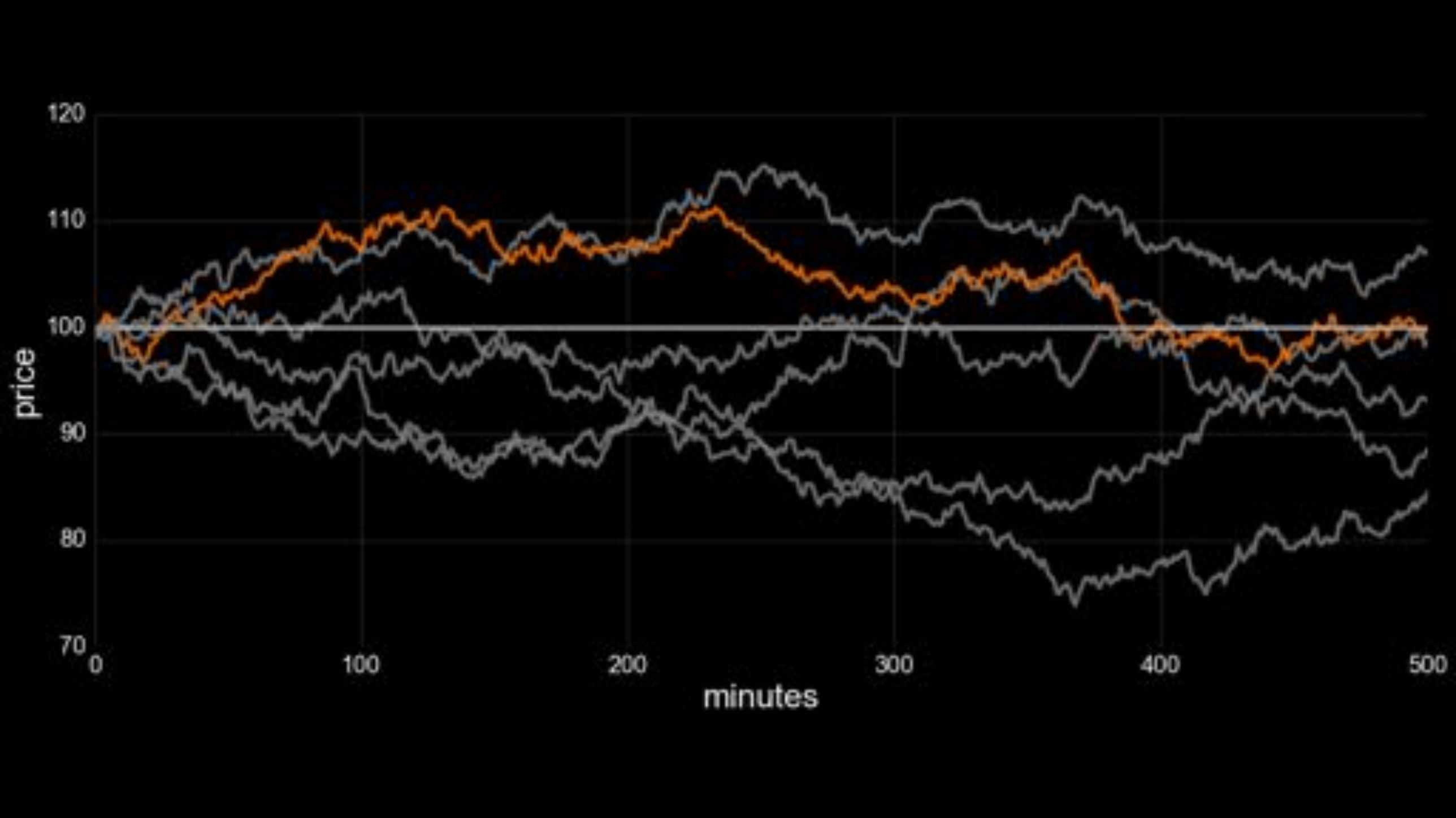
Add one more model - 2 minute strategy

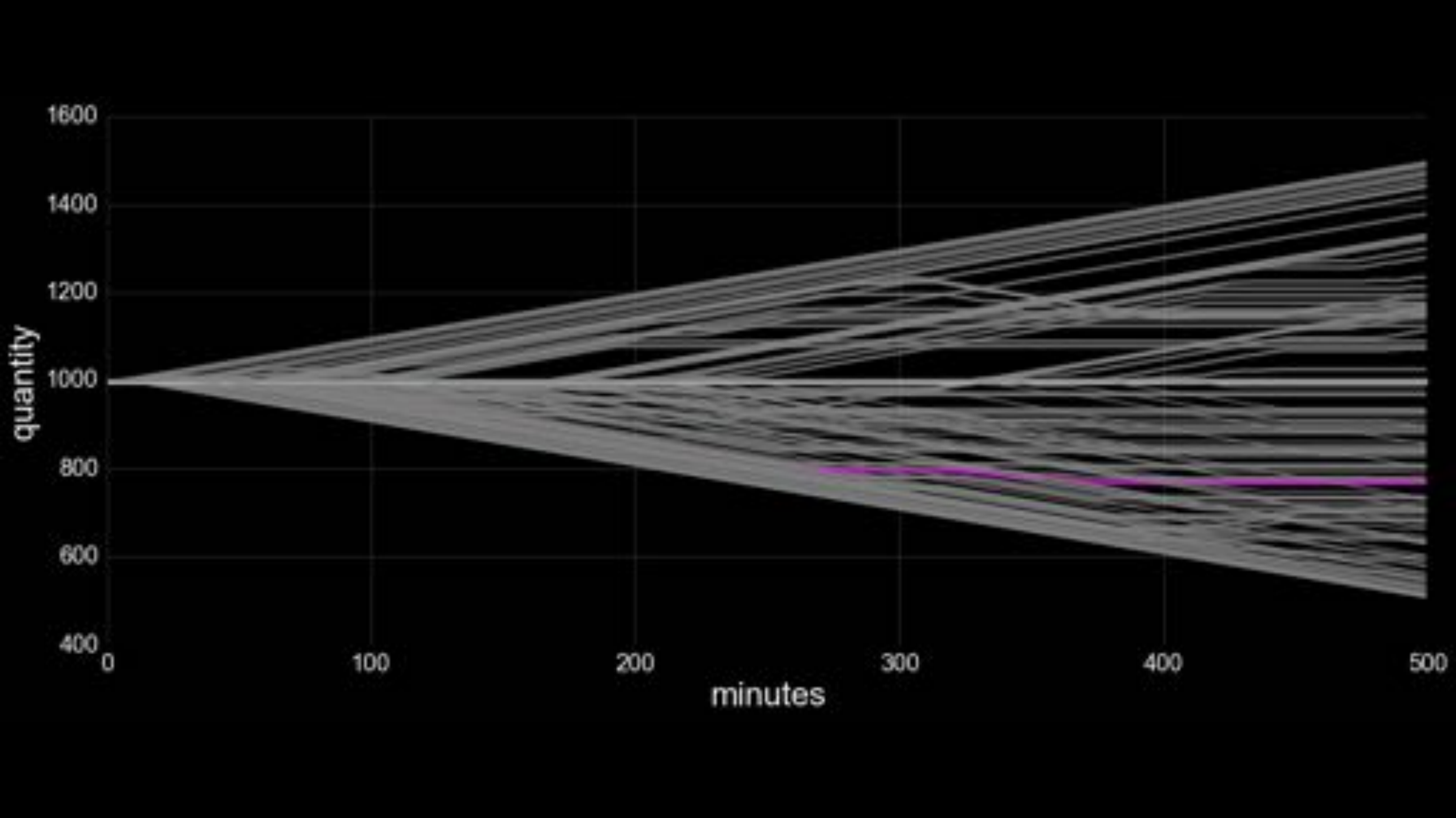

1050
1000
950
900
850
800
750
700
650
600
quantity
0
100
200
300
400
500
minutes

Visualise the model

(5) with different input datasets

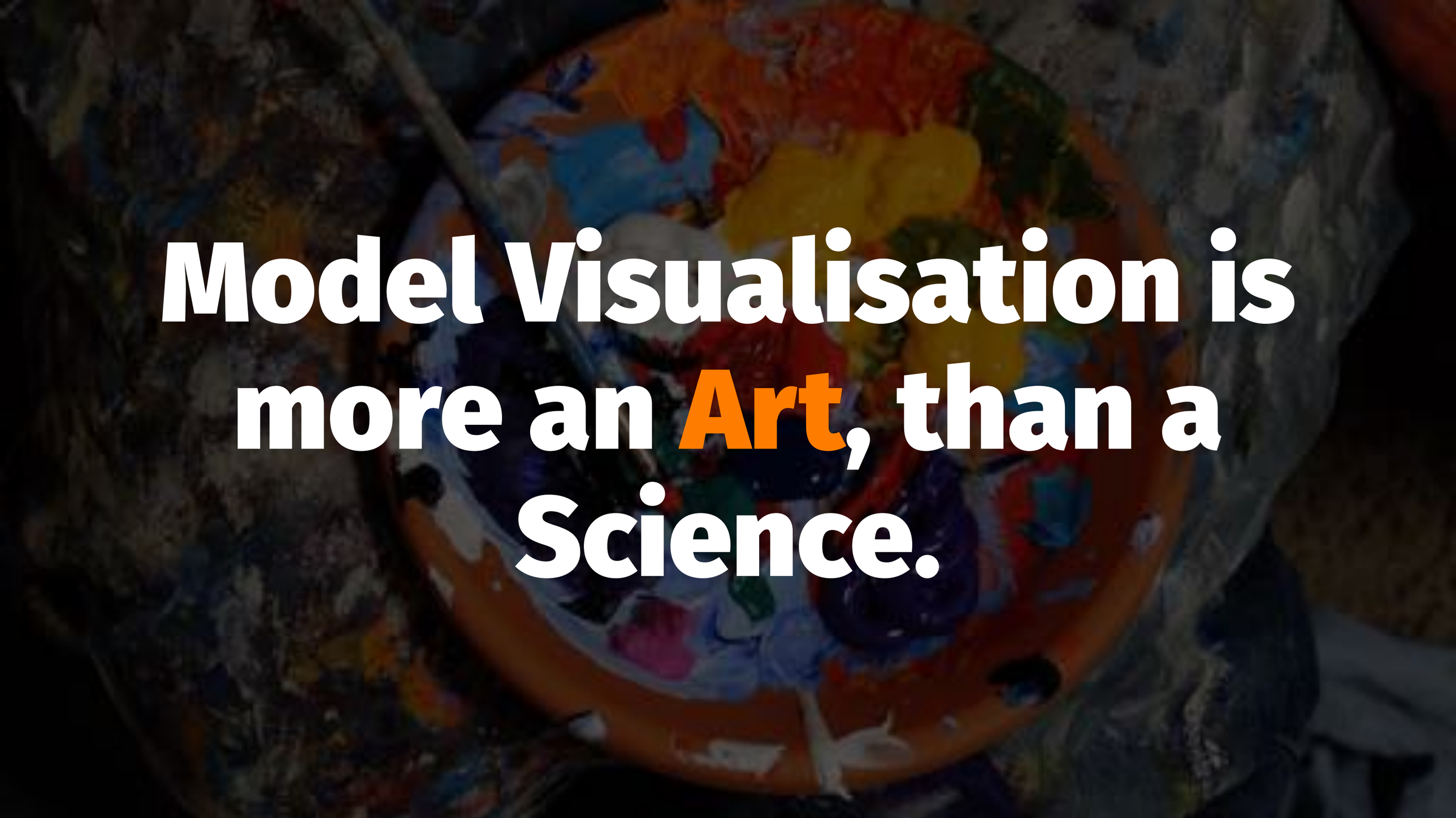
Add five input datasets

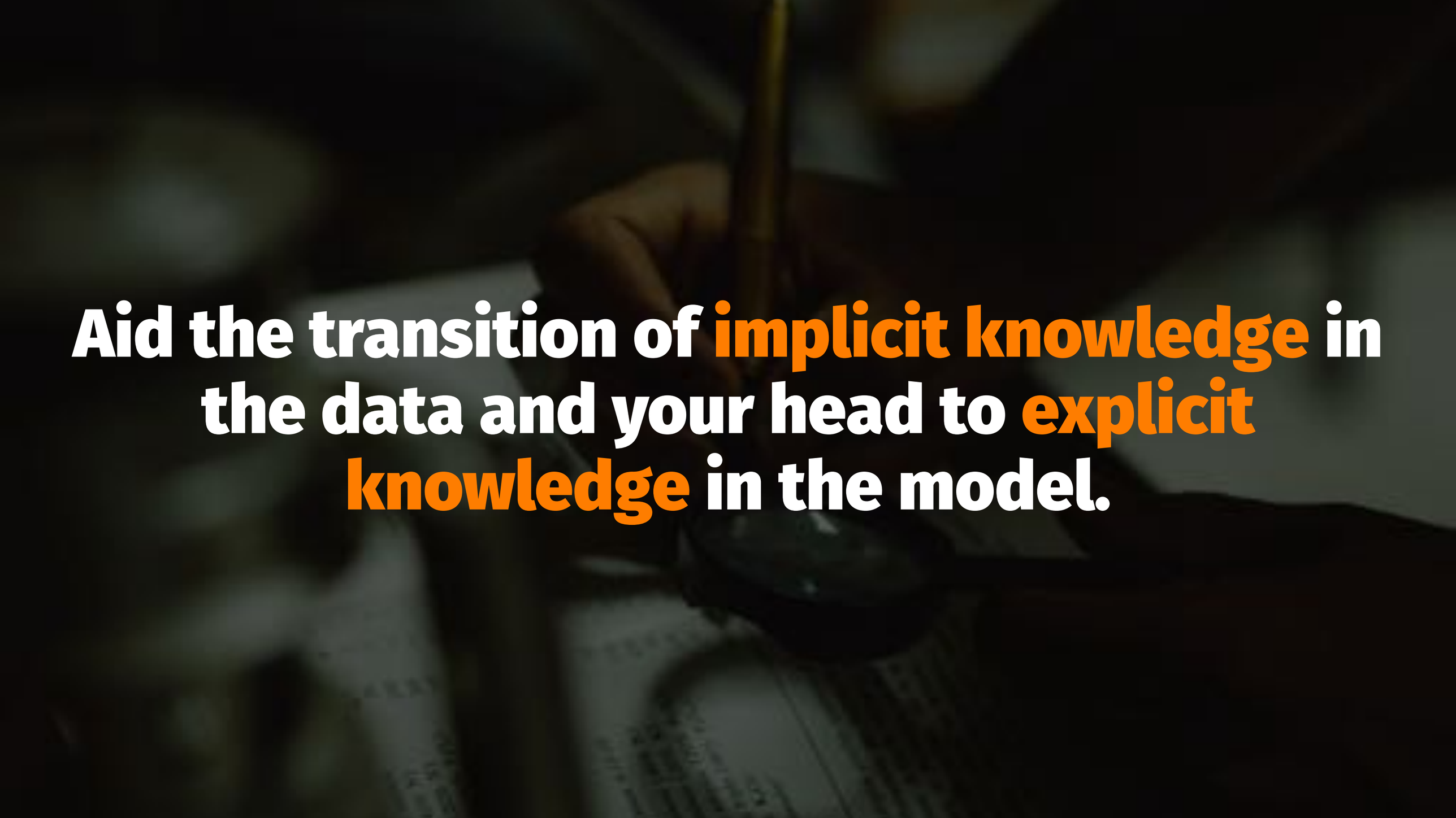

120
110
100
90
80
70
price
0
100
200
300
400
500
minutes


1600
1400
1200
1000
800
600
400
quantity
0
100
200
300
400
500
minutes

Model Vis Approach

within the **data space**
with **varying model** parameters
for the process of **model fitting**
with **entire model space**
with **different input datasets**


Model Visualisation is more an Art, than a Science.


Aid the transition of implicit knowledge in the data and your head to explicit knowledge in the model.

Frame

"An approximate answer to the right problem is worth a good deal"

Acquire

"80% perspiration, 10% great idea, 10% great output"


Refine
"All data is messy."

Transform

"The world does not work on the scales it happens to be measured on."

Explore

"I don't know, what I don't know."

Model

"All models are wrong, but some are useful"

Insight

"If you can’t explain it simply, you don’t understand it well enough."

Transform
Explore
Model

ML Approach

Focus on improving the predictive ability of the model

Being careful to fairly assess it (train vs. test)

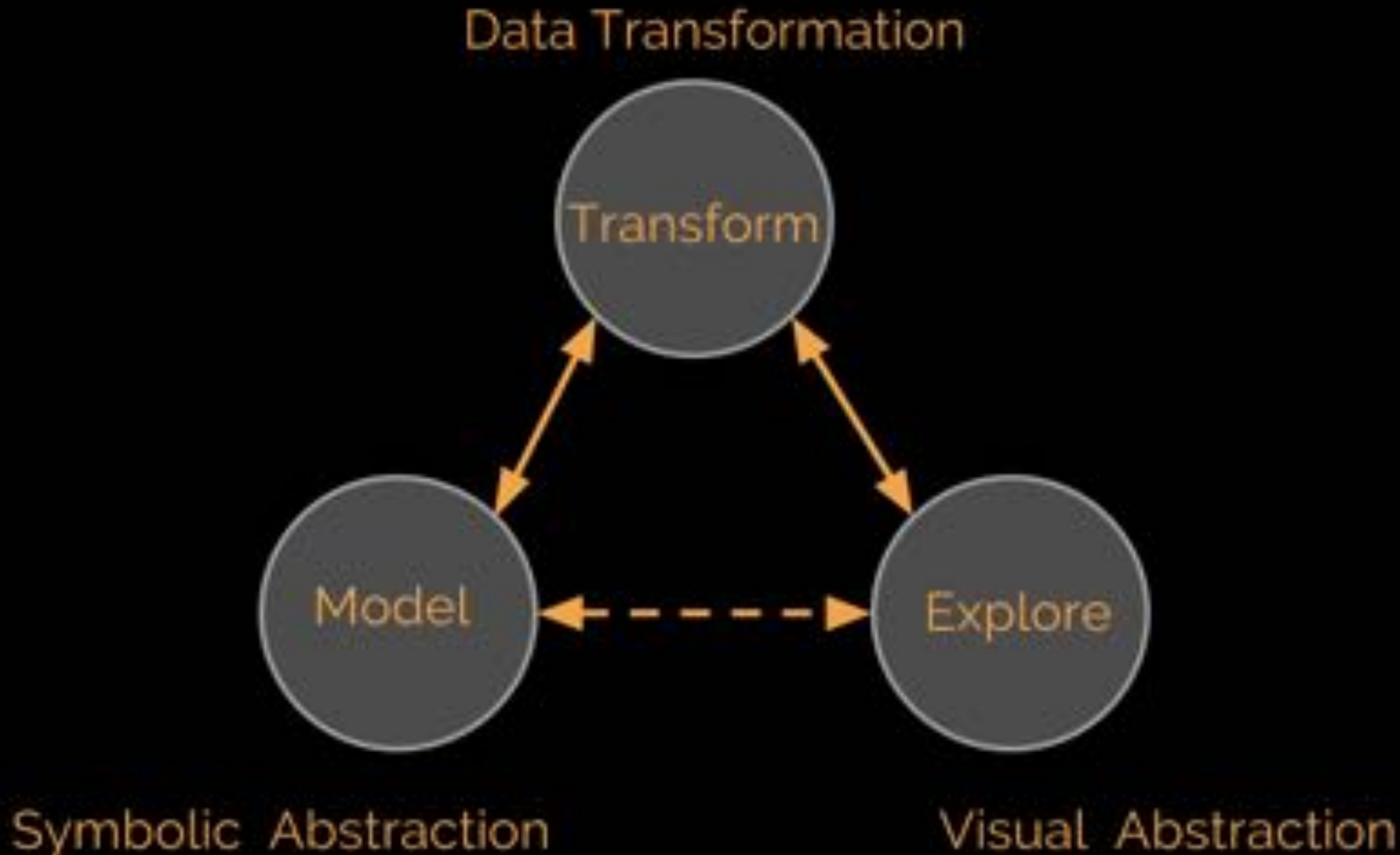
Black boxes

The model does a really good job, but you don’t know why.

Challenges

How do you apply real world knowledge to the model?

Will it work in the long-term, as fundamentals change?


Data Transformation
Transform
Model
Explore
Symbolic Abstraction
Visual Abstraction

ML Approach : Model Vis

DIMENSIONALITY REDUCTION : within the data space

FEATURE SELECTION : with varying model parameters

CROSS-VALIDATION : for the process of model fitting

ENSEMBLE : with entire model space

BOOTSTRAP : with different input datasets

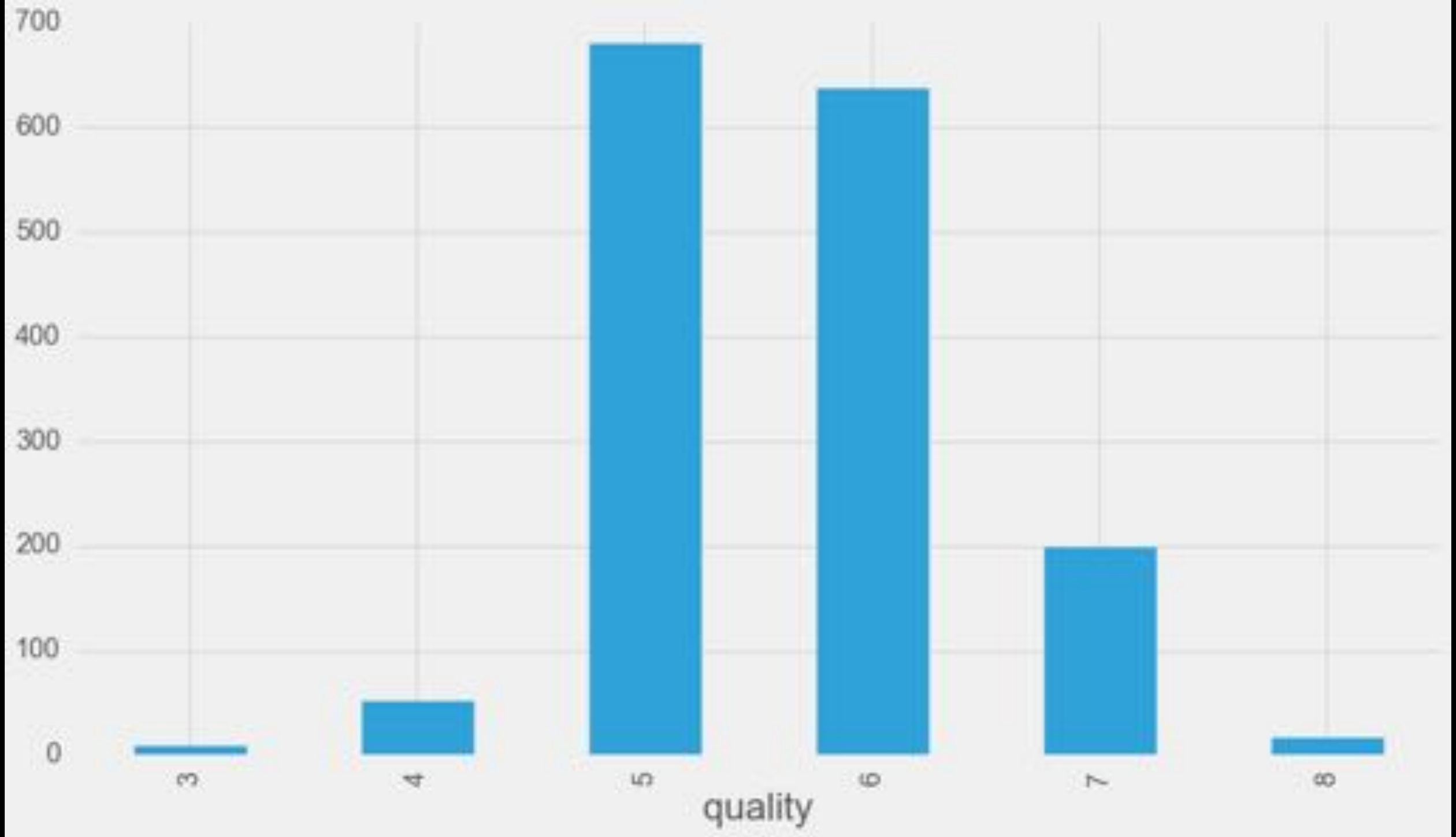
Ideas to develop on Model Visualisation

Predicting the Quality of Wine

1599 Observation with 12 dimensions

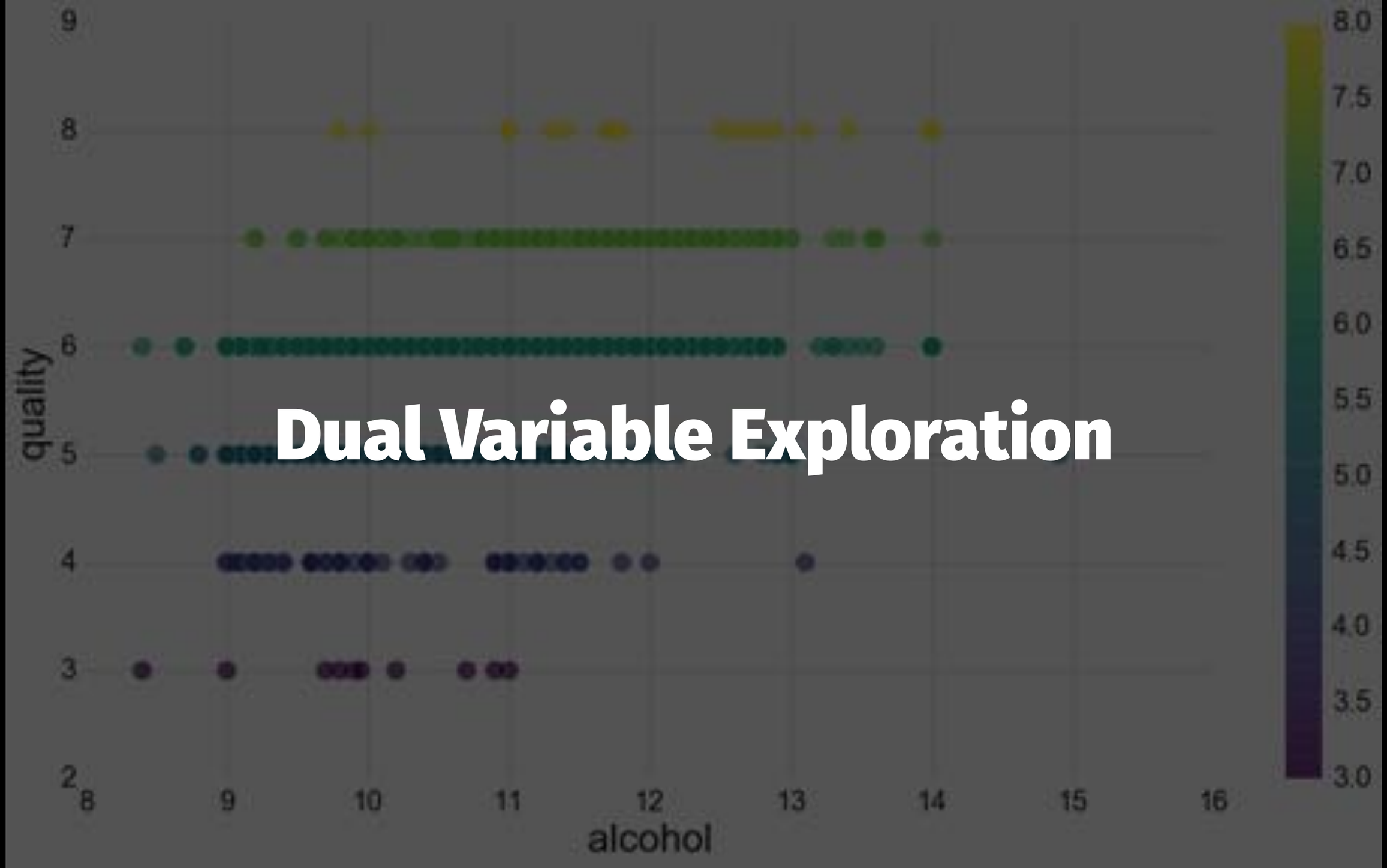
1 target
based on sensory data

Wine Quality

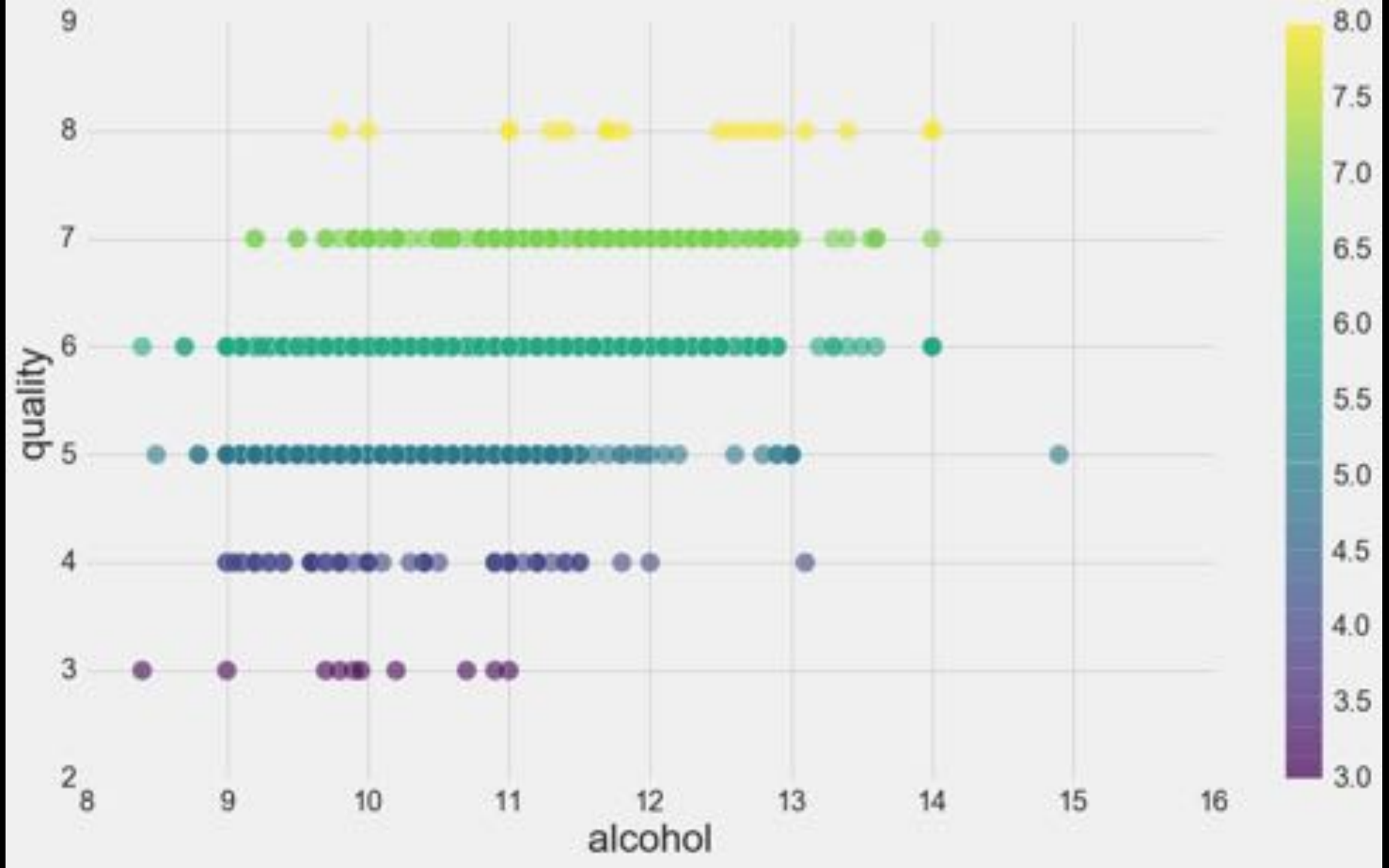

700
600
500
400
300
200
100
0
3
4
5
6
7
8
quality

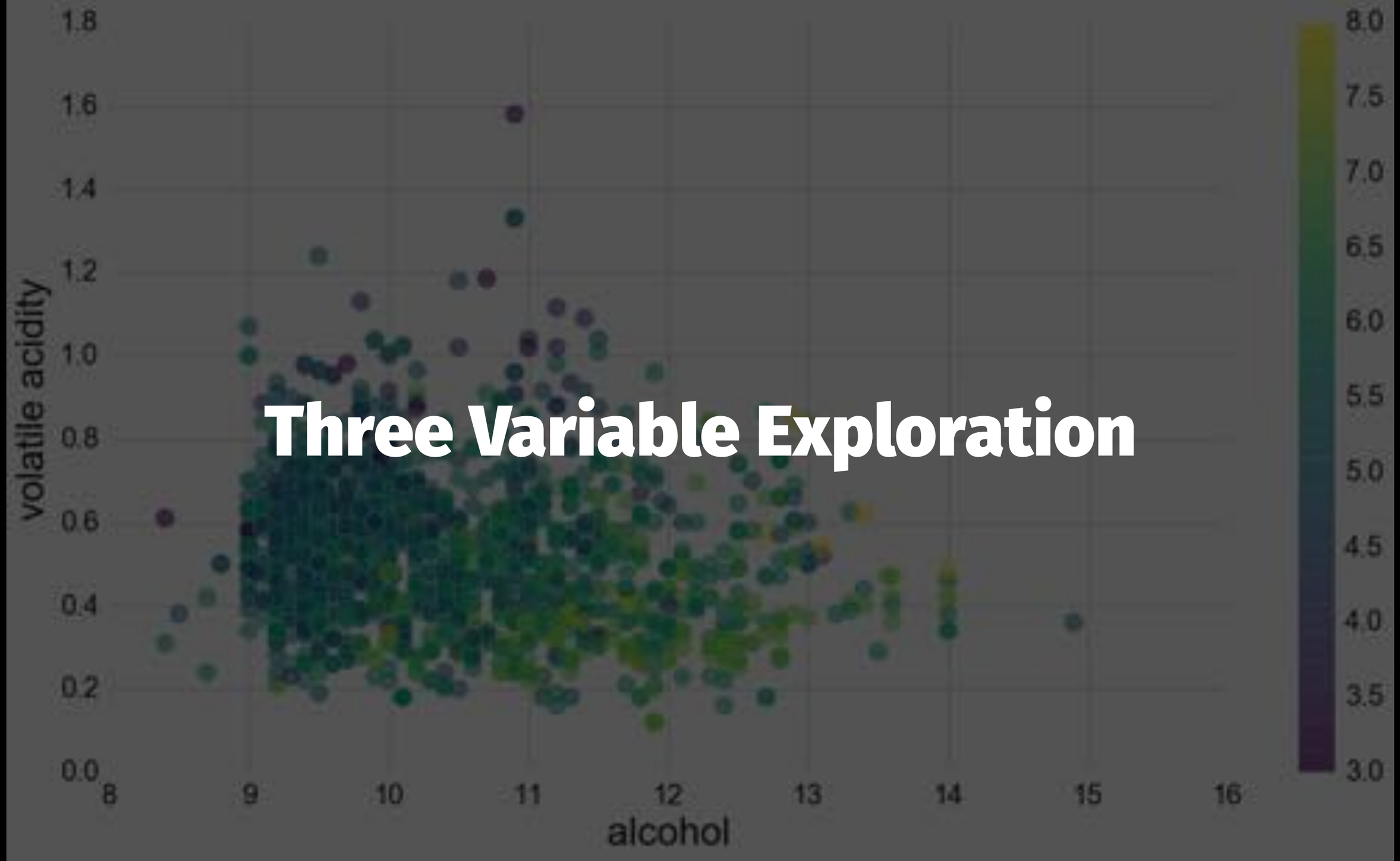
11 features
based on physicochemical tests

alcohol
density, pH
residual sugar
fixed acidity, volatile acidity, citric acid
chlorides, free sulfur dioxide, total sulfur dioxide, sulphates

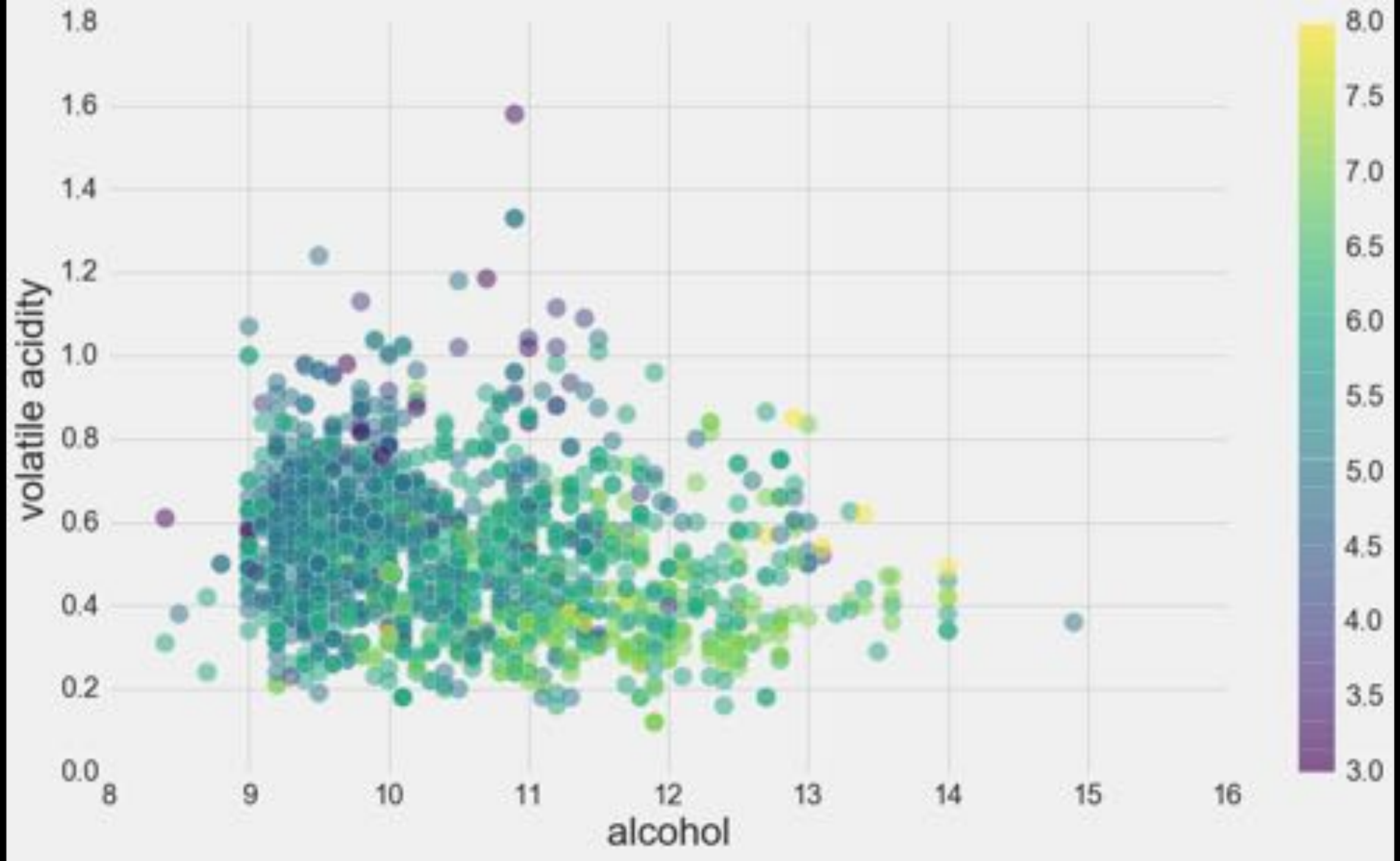


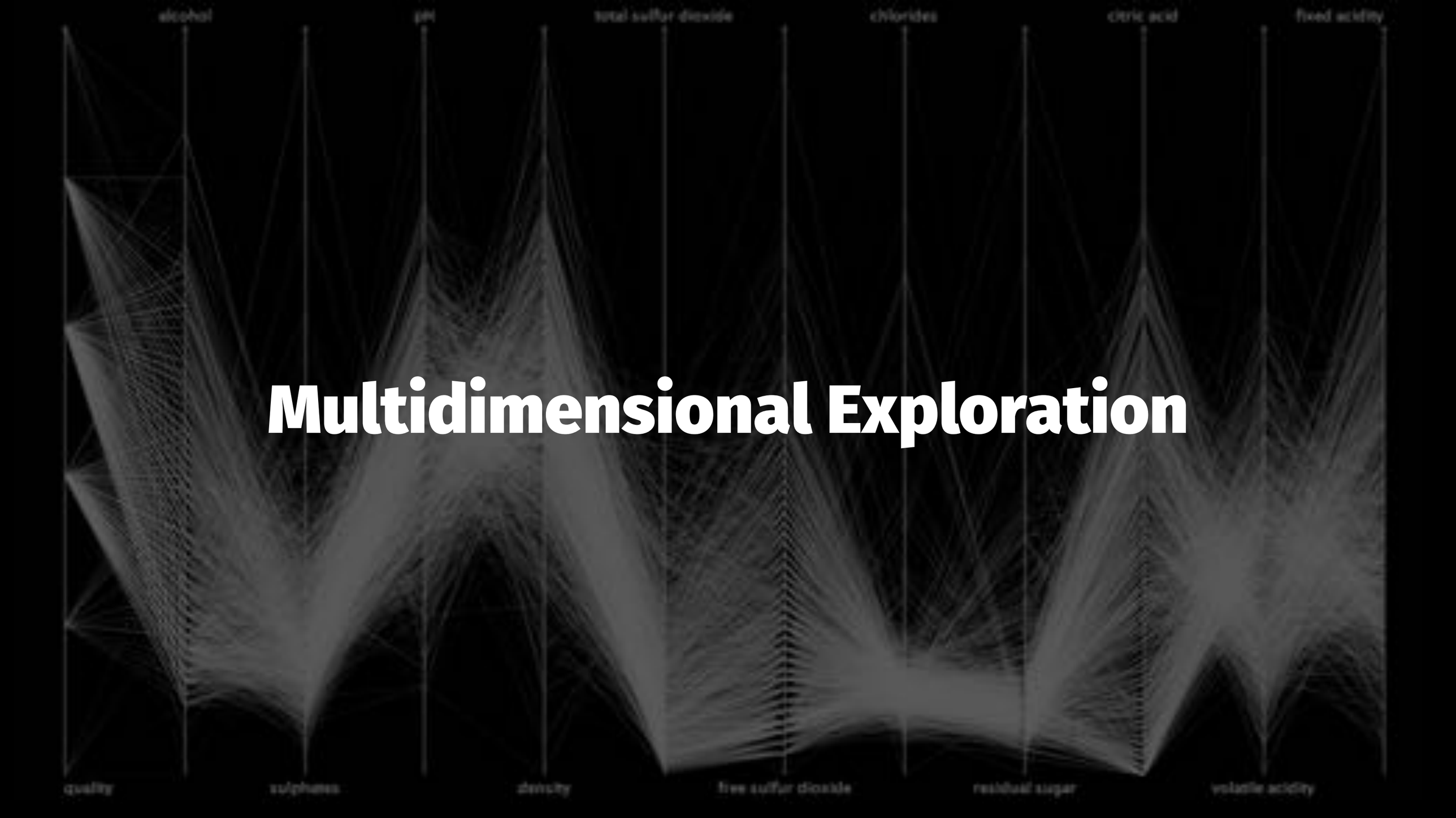
Dual Variable Exploration


quality
alcohol
9
8
7
6
5
4
3
2
8
9
10
11
12
13
14
15
16
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0

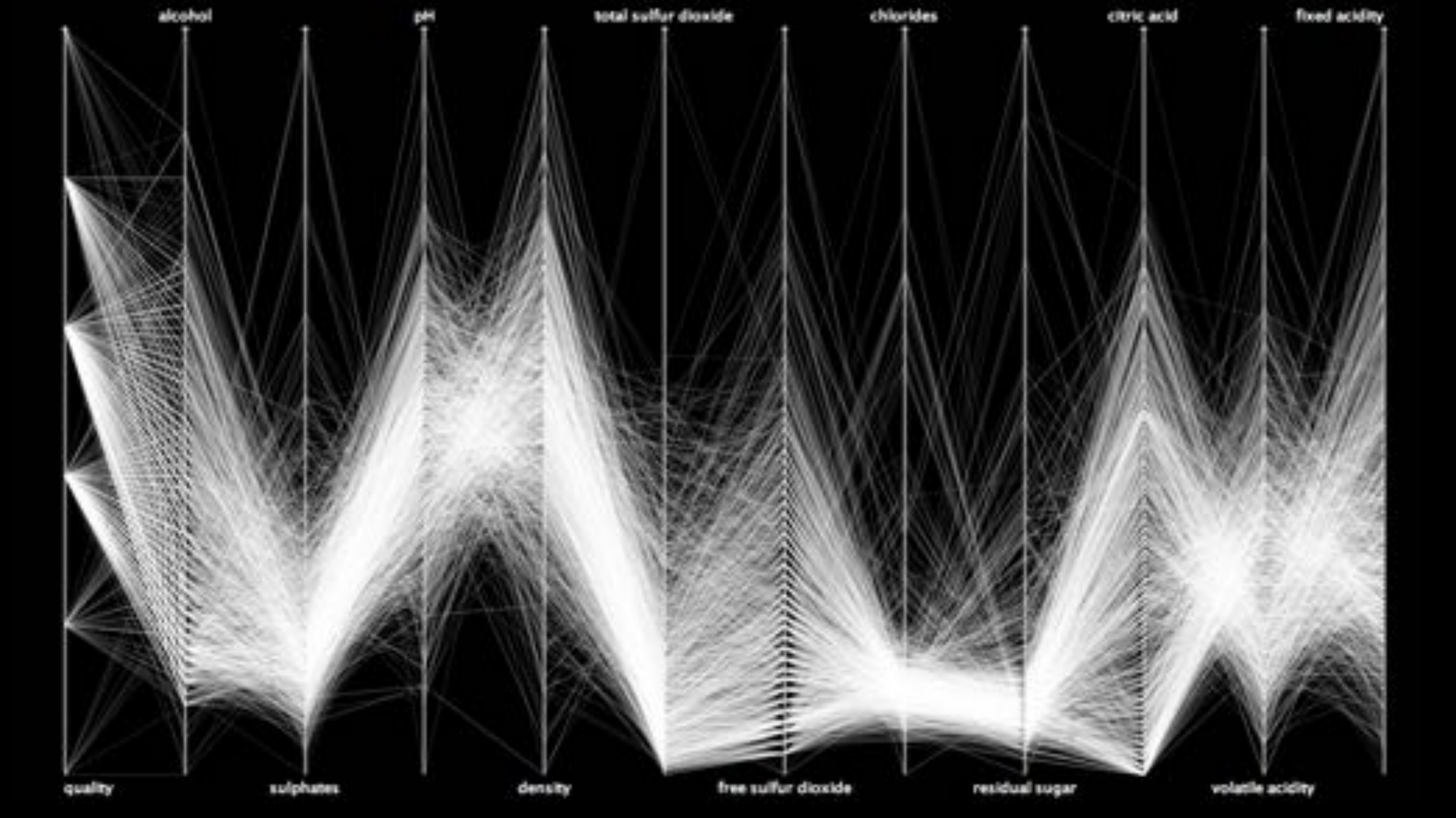


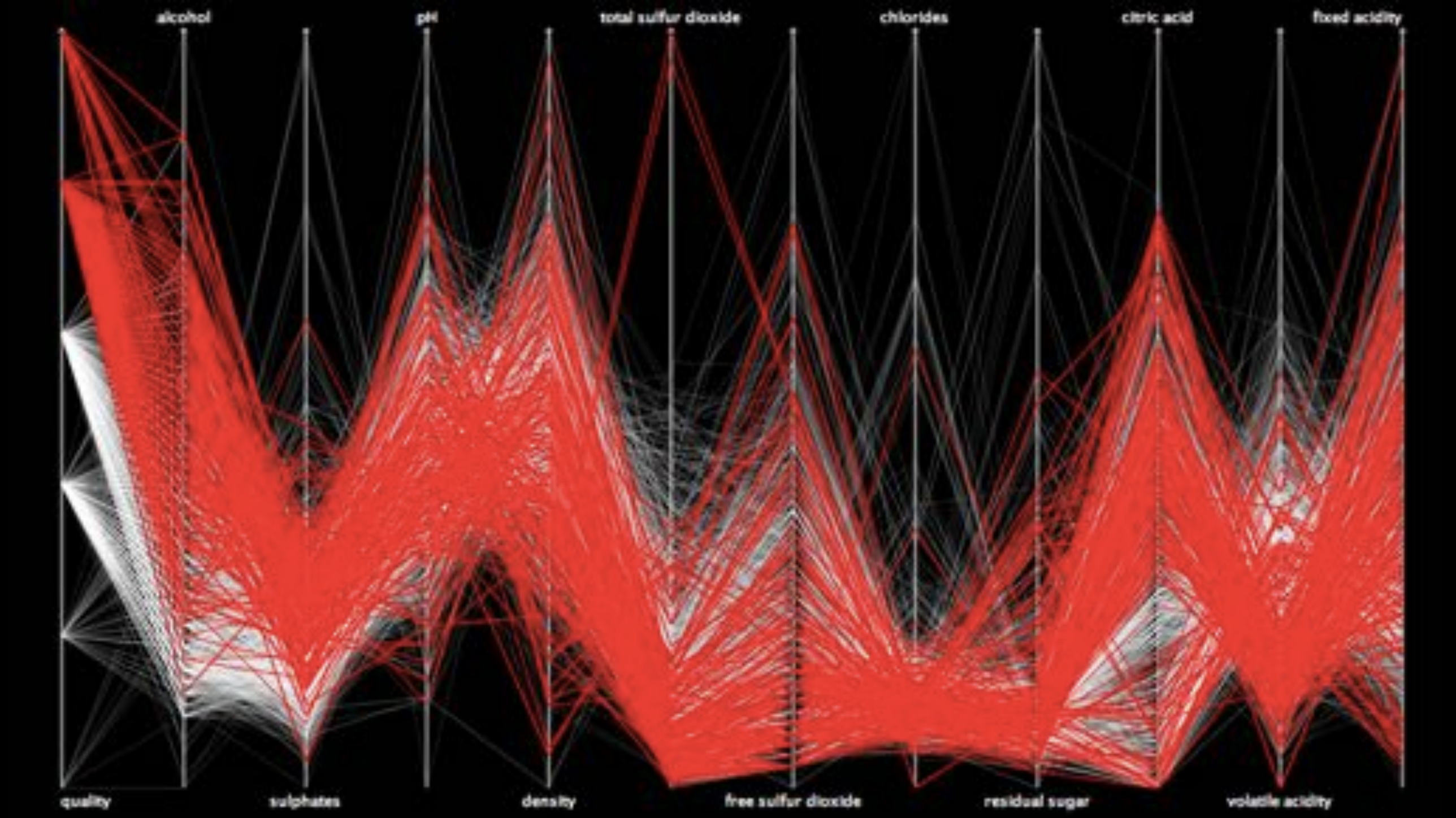
Three Variable Exploration


volatile acidity
alcohol
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
8
9
10
11
12
13
14
15
16
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0



Multidimensional Exploration


alcohol
pH
total sulfur dioxide
chlorides
citric acid
fixed acidity
quality
sulphates
density
free sulfur dioxide
residual sugar
volatile acidity


alcohol
pH
total sulfur dioxide
chlorides
citric acid
fixed acidity
quality
sulphates
density
free sulfur dioxide
residual sugar
volatile acidity

Linear Regression

quality ~ f(alcohol, pH,... sulphates)

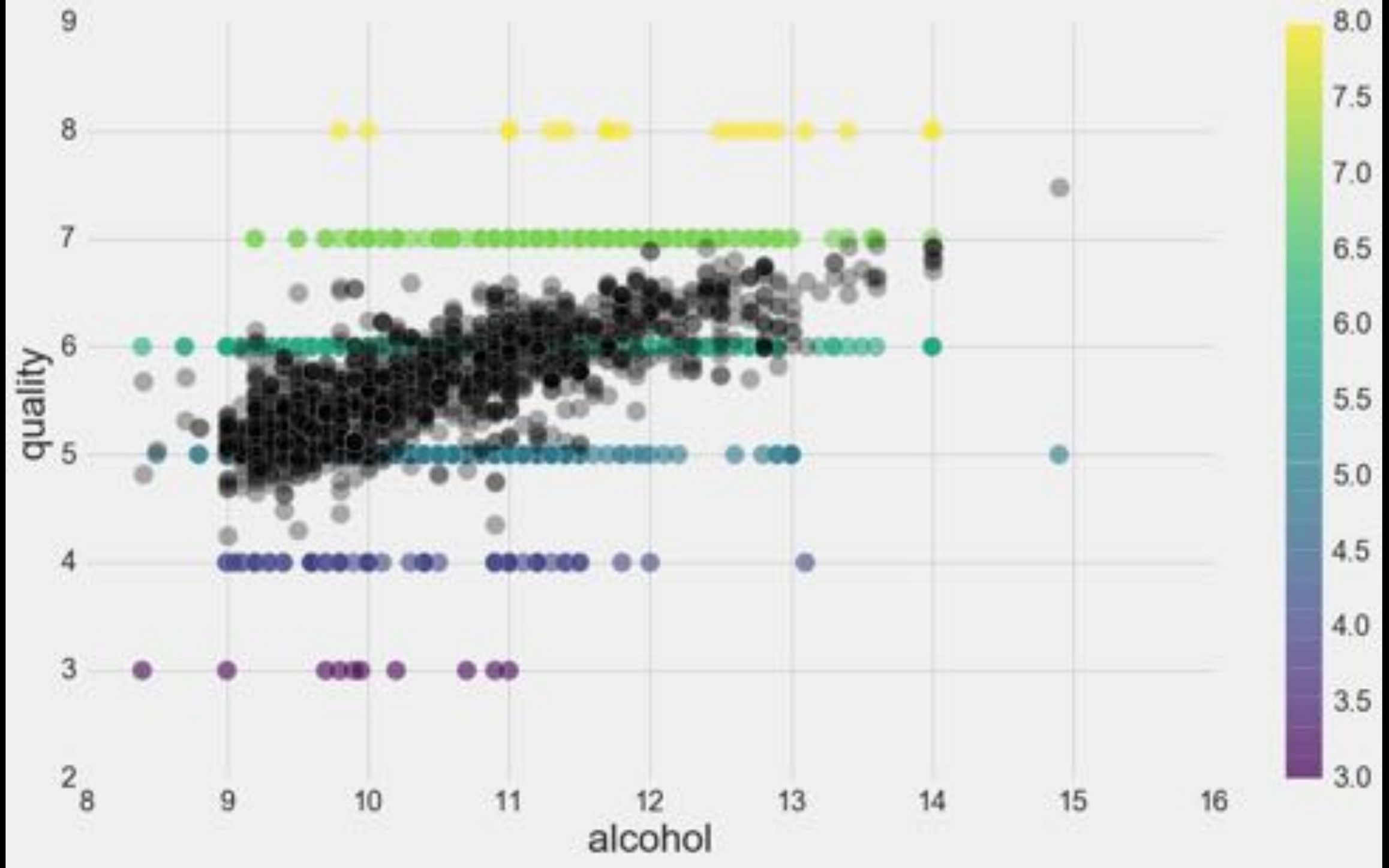
model score = 0.36

Visualise the model

(1) within the data space

Add predicted data

Use data generated from model as regular data - manipulate it and visualise it in many different ways.

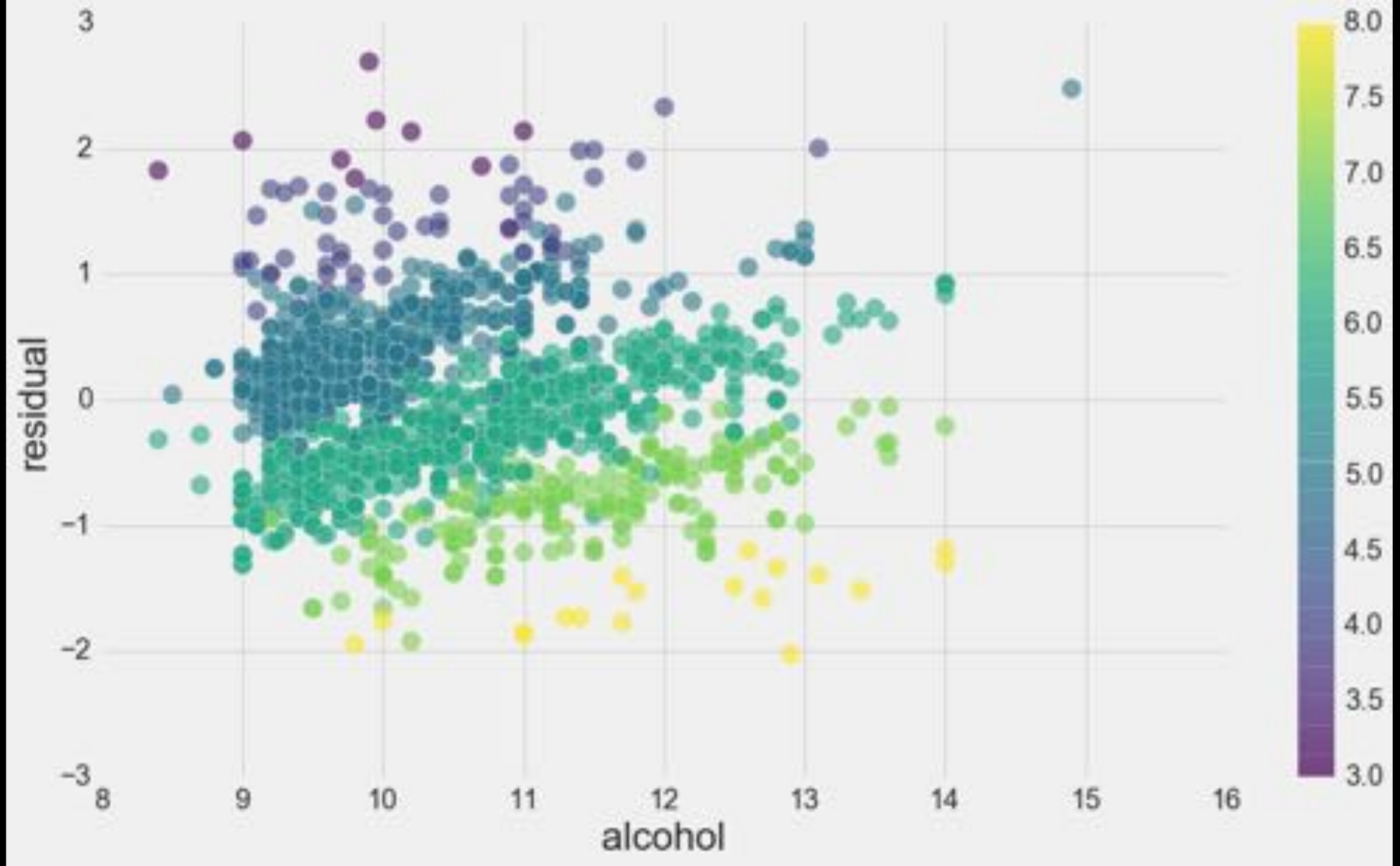

quality
alcohol
9
8
7
6
5
4
3
2
8
9
10
11
12
13
14
15
16
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0

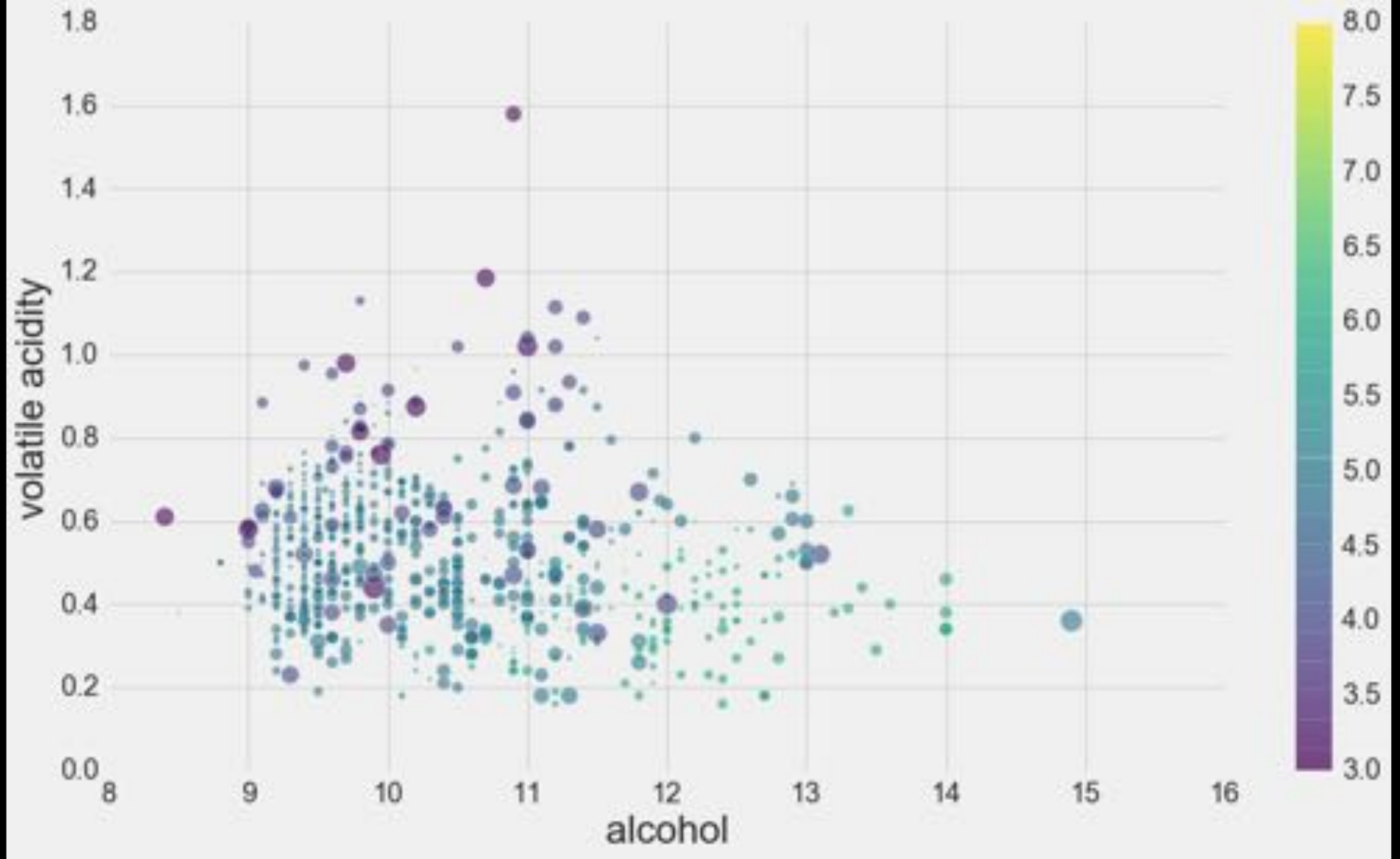
Visualise the model

(3) with the process of model fitting

Use residuals

To subtract patterns from the data, while adding them to the model

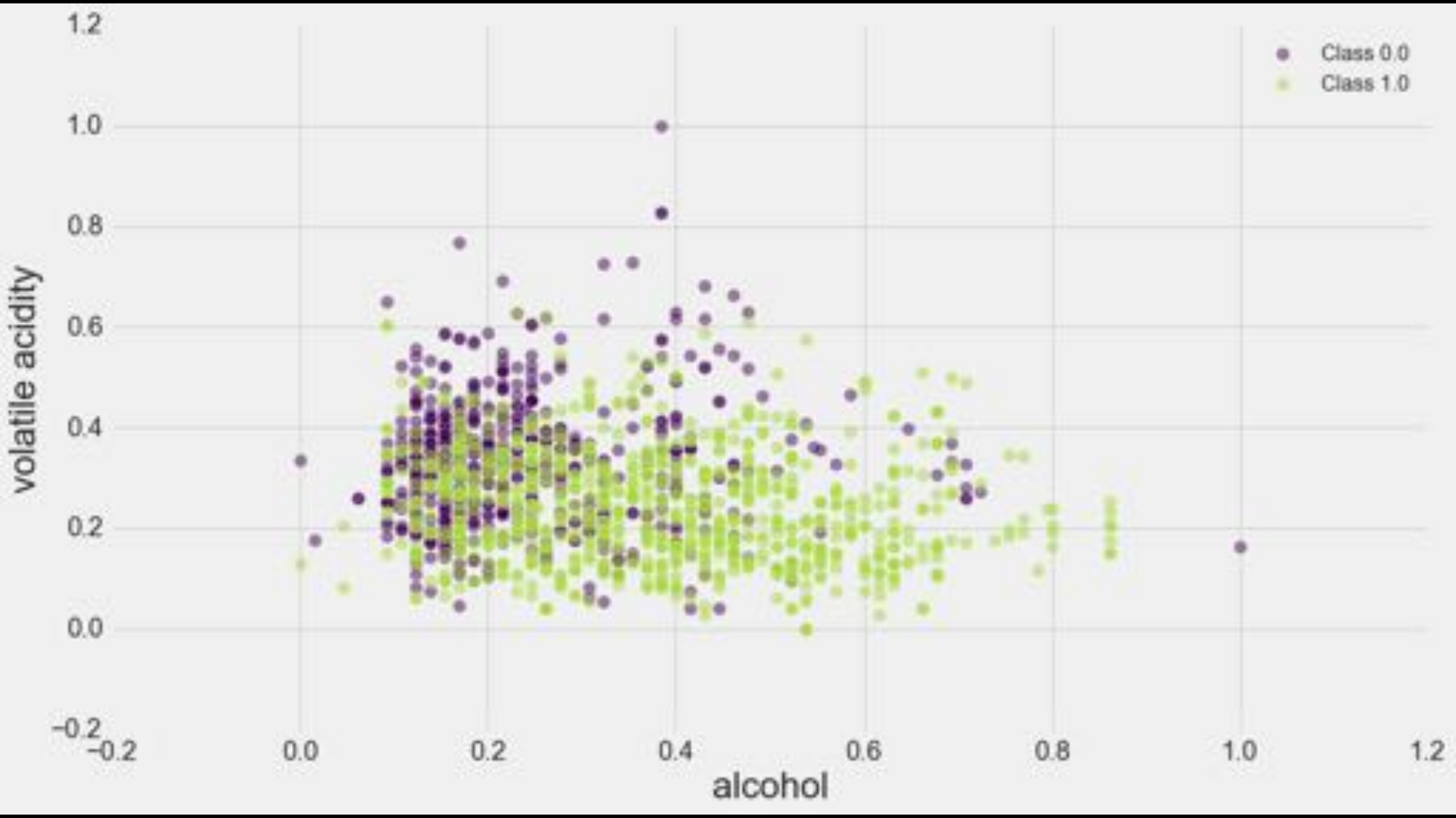

residual
alcohol
3
2
1
0
−1
−2
−3
8
9
10
11
12
13
14
15
16
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0


volatile acidity
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
8
9
10
11
12
13
14
15
16
alcohol
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0

Binary Classification

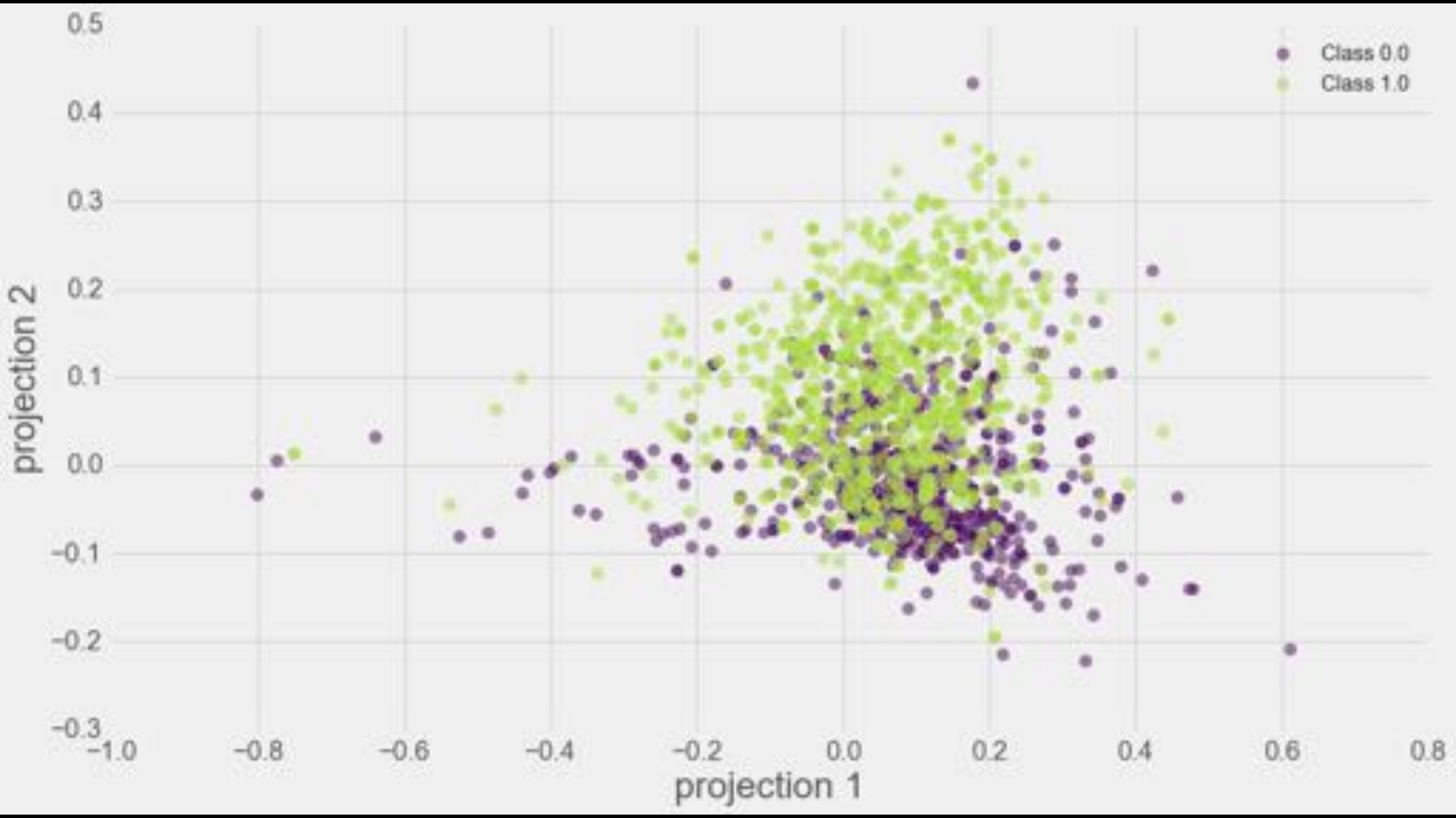
Class 0 (Low): Quality = [3,4,5]

Class 1 (High): Quality = [6,7,8]


Class 0.0
Class 1.0
volatile acidity
alcohol
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

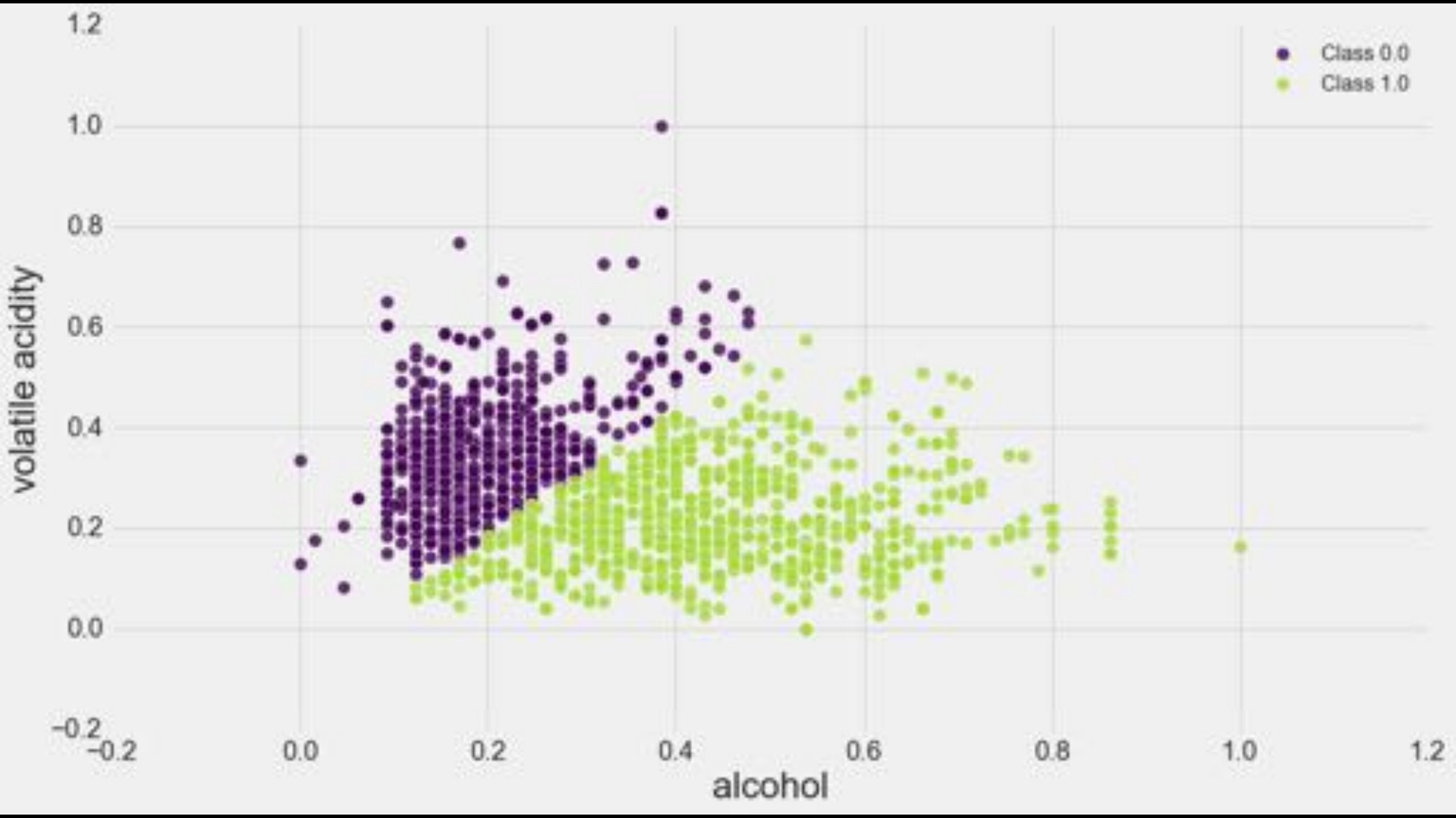
Play with Projections

Be able to view the data in multiple projections - guided tours, projection pursuit


Class 0.0
Class 1.0
projection 2
projection 1
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

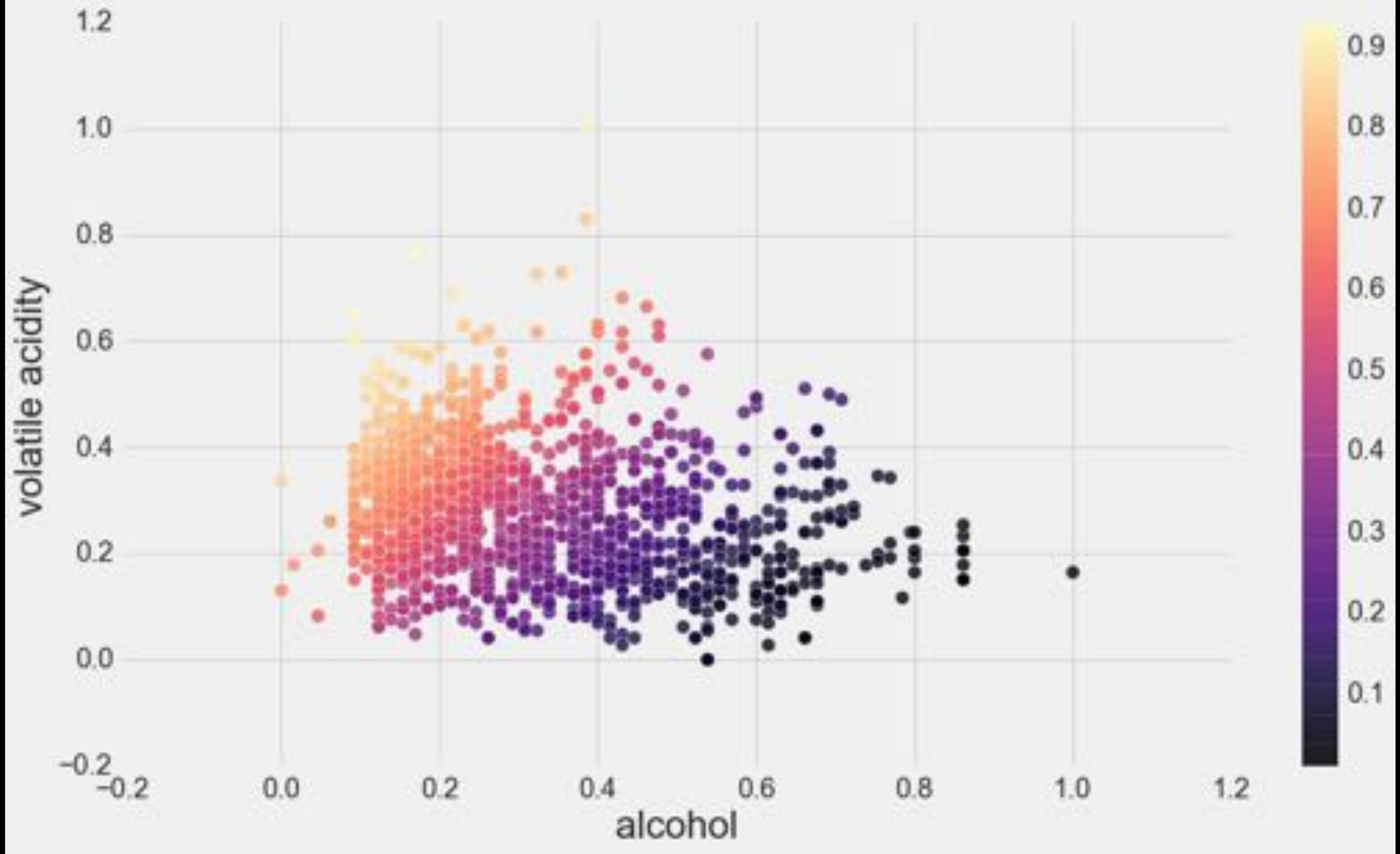
Logistic Regression

model score - 0.74


Class 0.0
Class 1.0
volatile acidity
alcohol
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

Evaluate Boundaries

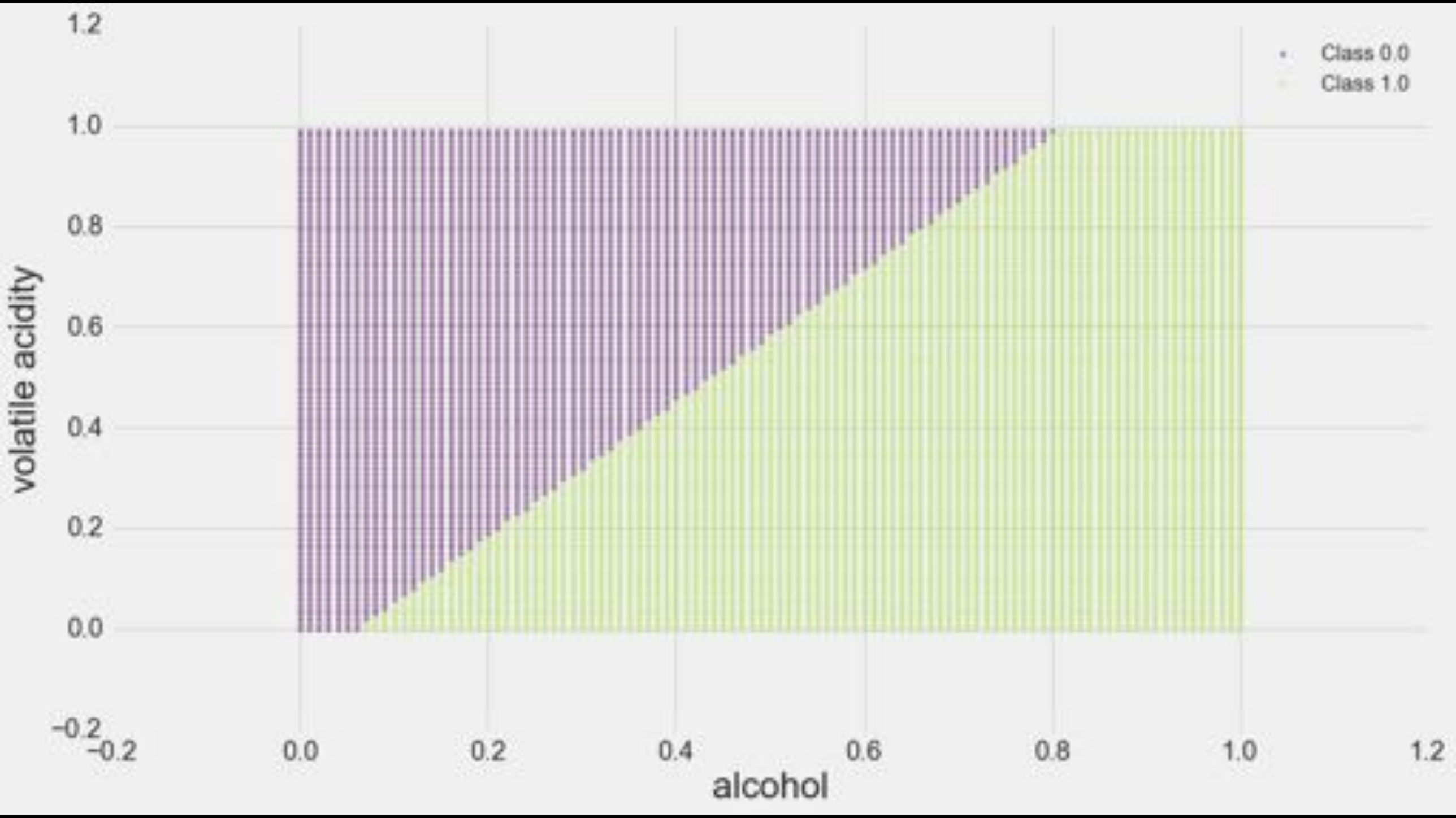
Sample data with border probabilities to create boundaries


volatile acidity
alcohol
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

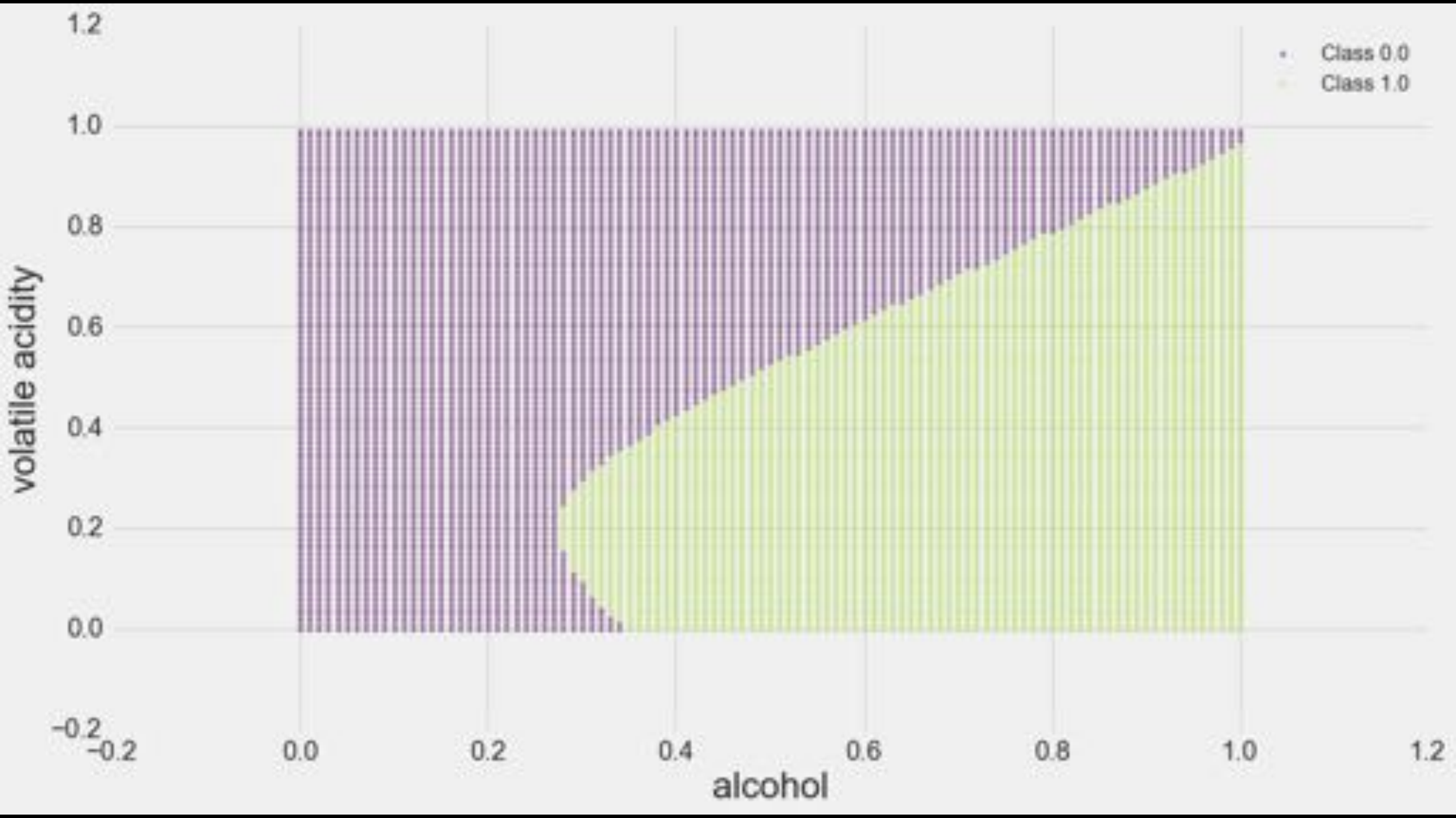
Evaluate Boundaries

Create a mesh of entire range of data

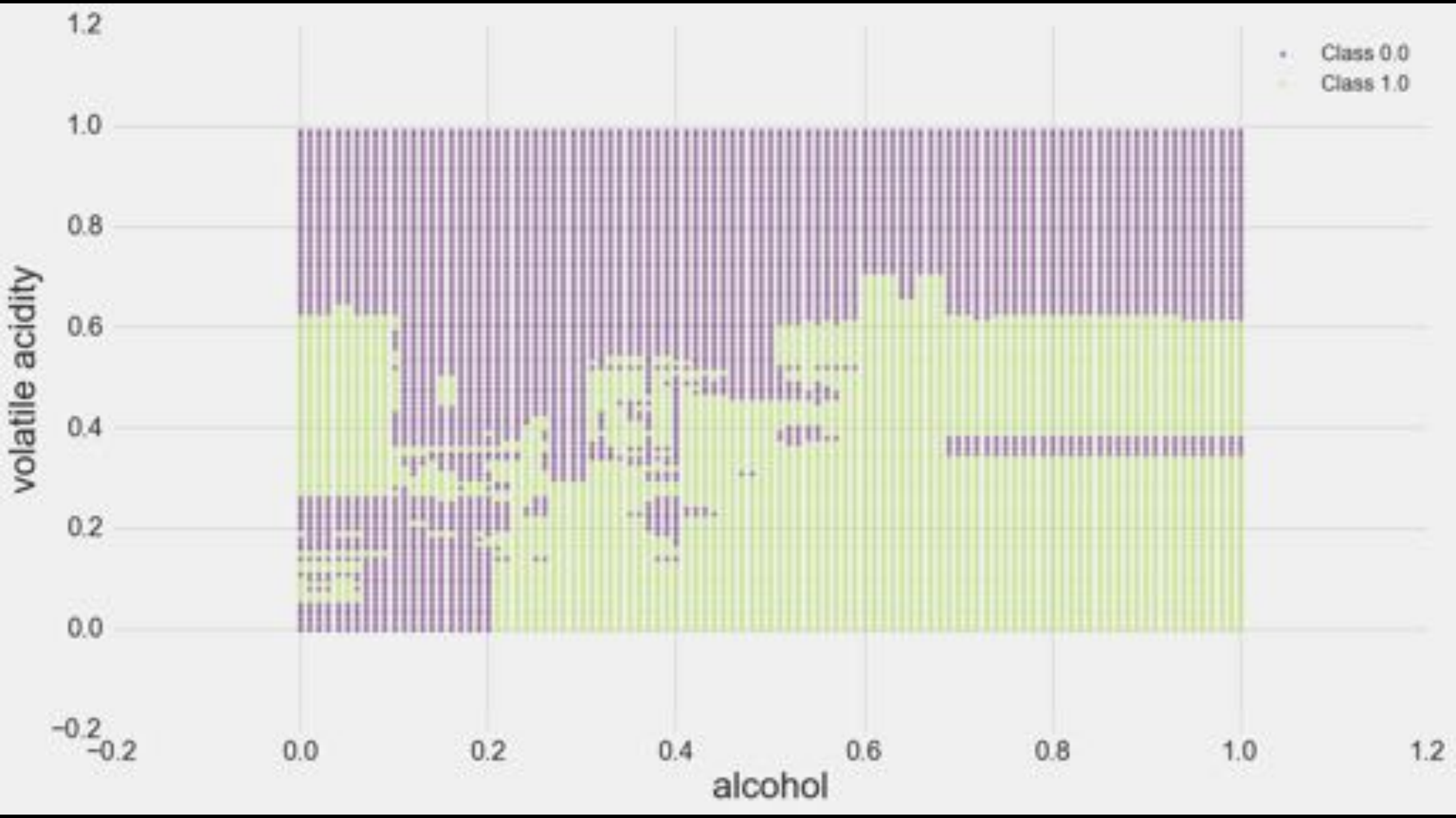
Mesh - Logistic Regression


Class 0.0
Class 1.0
volatile acidity
alcohol
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

Mesh - Quadratic Discriminant Analysis


Class 0.0
Class 1.0
volatile acidity
alcohol
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

Mesh - Random Forest


Class 0.0
Class 1.0
volatile acidity
alcohol
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

Not Easy!

Selective approaches articulated

Curse of Dimensionality

Poor interactive tools

Limited package development in Python

Work in Progress

Projection module

Bootstrap module

Cross Validation module

Model Visualisation

https://github.com/amitkaps/modelvis

Amit Kapoor

Crafting Visual Stories with Data

@amitkaps

amitkaps.com