K NEAREST NEIGHBORS

Projects: 1) Identifying Wine Color and 2) Optical Character Recognition

BEFORE WE BEGIN LET US DO A THOUGHT EXPERIMENT

I want to find somebody to spend a Saturday afternoon with and I am looking for somebody most similar to me (nearest neighbor) in terms of:

- Sex (coded as 0 for female, and 1 for male)
- Age (coded in years)
- Outdoor sports interest (coded from 0 (no interest) to 10 (enthusiast))»

(all categories matter the same to me)

LET US DO THE CALCULATION FOR A SIMILARITY SCORE

(AVERAGE ABSOLUTE DIFFERENCES)

Sake of argument: I am male (**1**), **50** years, outdoor sports score **8**:

- first candidate a student
  .
- second candidate an athletic outdoor (score=**9**) women (**0**) **51** years old
- third candidate an athletic outdoor (score=**9**), man (**1**) **53** years

LET US DO THE CALCULATION FOR A SIMILARITY SCORE

(AVERAGE ABSOLUTE DIFFERENCES – NORMALIZED TO 0 – 10)

Sake of argument: I am male (**1**), **50** years, outdoor sports score **8**:

- first candidate a student
  .
- second candidate an athletic outdoor (score=**9**) women (**0**) **51** years old
- third candidate is an athletic outdoor (score=**9**) man (**1**) **53** years»

OVERWIEW

In this session you will learn:

1. What is the underlying **idea of k-Nearest Neighbors**
2. How similarity can be measured with **Euclidean distance**
3. Why **scaling predictor variables** is important for some machine learning models
4. Why the **tidymodels package** makes it easy to work with machine learning models
5. How you can define a **recipe** to pre-process data with the `tidymodels` package
6. How you can define a **model-design** with the `tidymodels` package
7. How you can create a machine learning **workflow** with the `tidymodels` package
8. How **metrics** derived from a **confusion matrix** can be used to asses prediction quality
9. Why you have to be careful when interpreting *accuracy*, when you work with **unbalanced observations**
0. How a machine learning model can **process images** and how OCR (Optical Character Recognition) works»

ABOUT THE WINE DATASET

We will work with a publicly available wine dataset[1] containing 3,198 observations about different wines and their chemical properties.

Our goal is to develop a k-Nearest Neighbors model that can predict if a wine is red or white based on the wine's chemical properties.»

1. Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." Decision Support Systems 47 (4): 547–53. https://doi.org/10.1016/j.dss.2009.05.016

RAW OBSERVATIONS FROM WINE DATASET

```
 1  library(rio)
 2  DataWine=import("https://lange-analytics.com/AIBook/Data/WineData.rds")
 3  print(DataWine)
```

```
# A tibble: 3,198 × 13
   wineColor acidity volatileAcidity citricAcid residualSugar Chlorides
   <chr>       <dbl>           <dbl>      <dbl>         <dbl>     <dbl>
 1 red          10.8           0.32        0.44           1.6     0.063
 2 white         6.4           0.31        0.39           7.5     0.04
 3 white         9.4           0.28        0.3            1.6     0.045
 4 white         8.2           0.22        0.36           6.8     0.034
 5 white         6.4           0.29        0.44           3.6     0.197
 6 red           6.7           0.855       0.02           1.9     0.064
 7 red          11.8           0.38        0.55           2.1     0.071
 8 white         6.7           0.25        0.23           7.2     0.038
 9 red           7.5           0.38        0.57           2.3     0.106
10 red           7.1           0.27        0.6            2.1     0.074
# … with 3,188 more rows, and 7 more variables: freeSulfurDioxide <dbl>,
#   totalSulfurDioxide <dbl>, Density <dbl>, pH <dbl>, sulphates <dbl>,
#   alcohol <dbl>, quality <dbl>
```

»

OBSERVATIONS FROM WINE DATASET FOR SELECTED VARIABLES

Note we use `clean_names("upper_camel")` from th `janitor` package to change all column (variable) names to UpperCamel.

```
library(tidyverse); library(rio);library(janitor)
DataWine=import("https://lange-analytics.com/AIBook/Data/WineData.rds") %>%
  clean_names("upper_camel") %>%
  select(WineColor,Sulfur=TotalSulfurDioxide,Acidity) %>%
  mutate(WineColor=as.factor(WineColor))
print(DataWine)
```

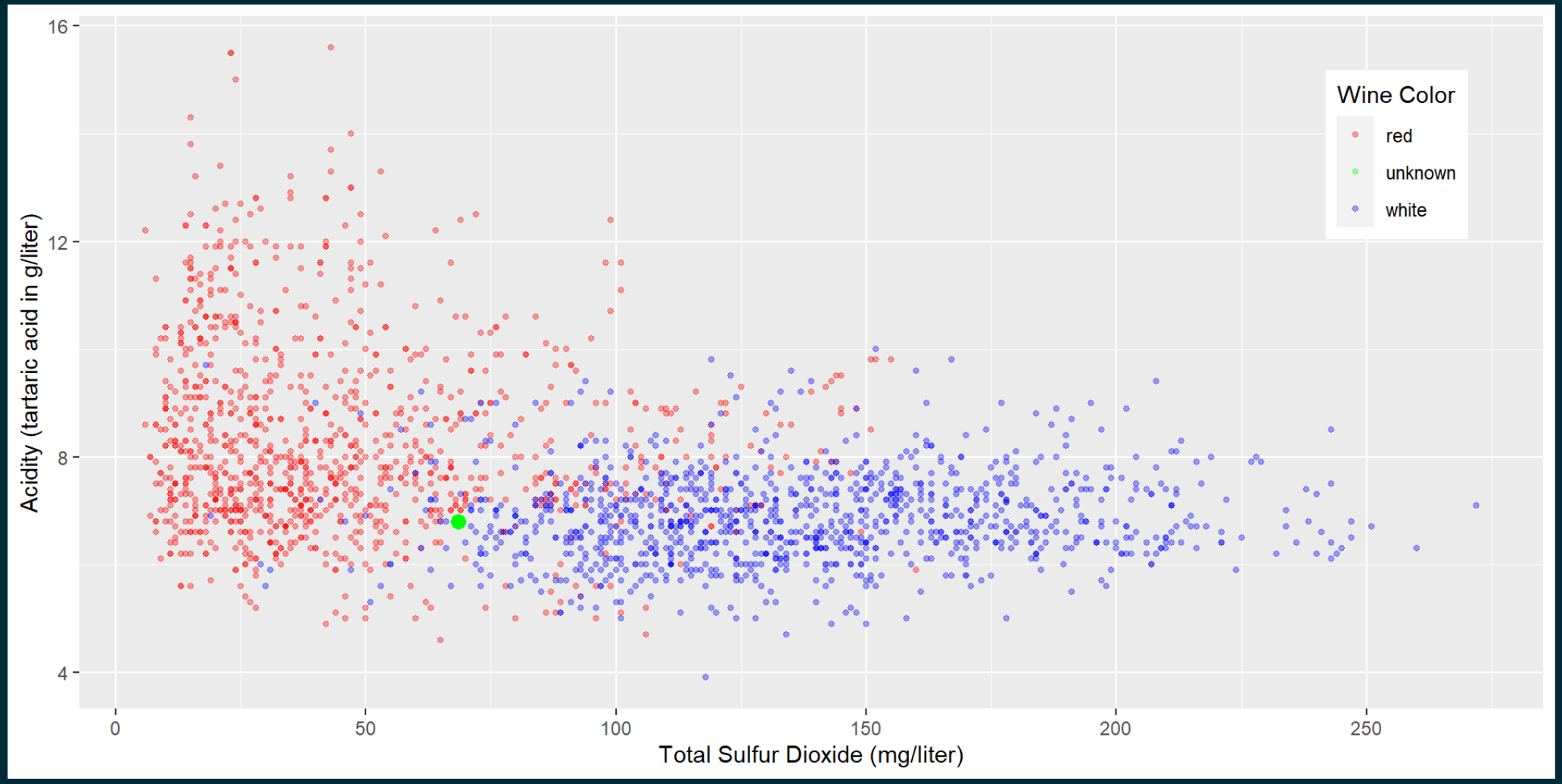
```
# A tibble: 3,198 × 3
   WineColor Sulfur Acidity
   <fct>       <dbl>   <dbl>
 1 red            37    10.8
 2 white         213     6.4
 3 white         139     9.4
 4 white          90     8.2
 5 white         183     6.4
 6 red            38     6.7
 7 red            19    11.8
 8 white         220     6.7
 9 red            12     7.5
10 red            25     7.1
# … with 3,188 more rows
```

BEFORE STARTING WITH K NEAREST NEIGHBORS

LET US FIND SOME EYEBALLING TECHNIQUES THAT ARE RELATED TO VARIOUS MACHINE LEARNING MODELS»

EYE BALLING TECHNIQUES TO IDENTIFY RED AND WHITE WINES

TRY EYEBALLING THE DATA

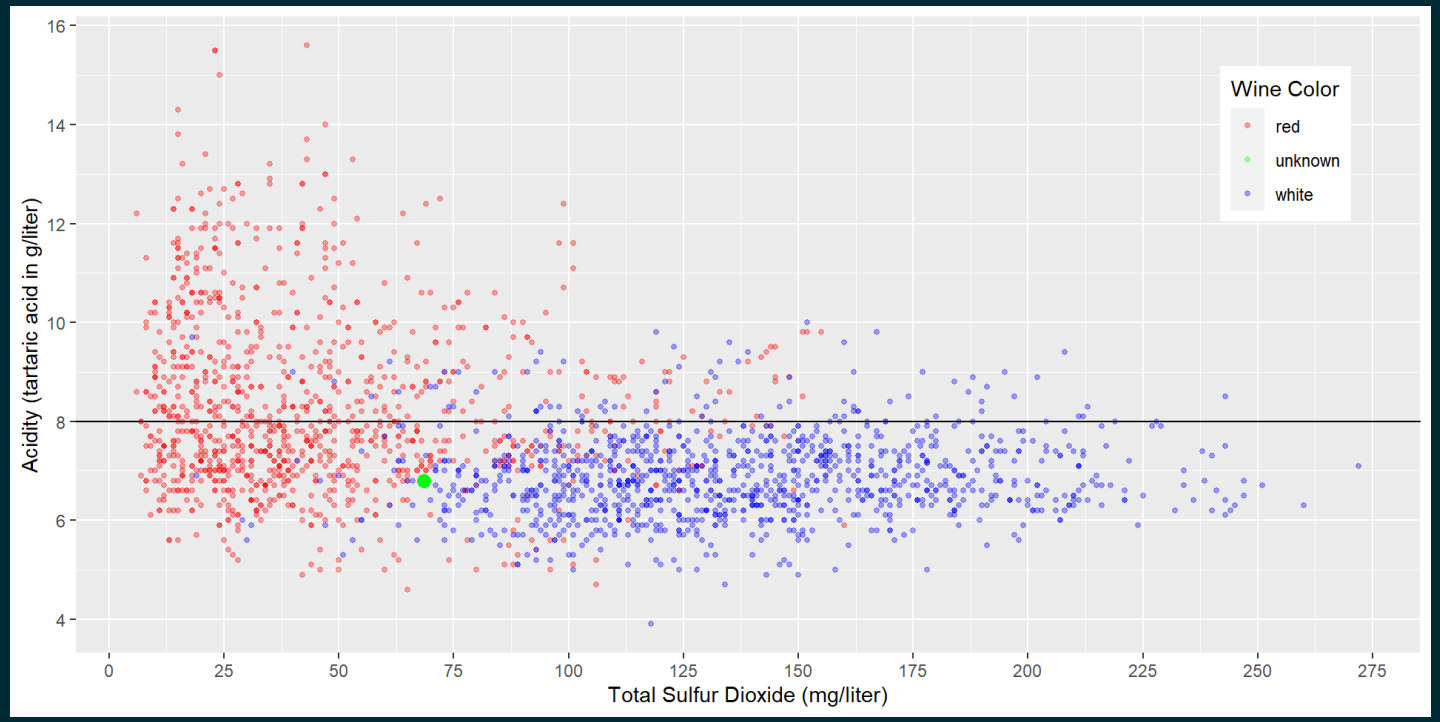


Acidity and Total Sulfur Dioxide Related to Wine Color

https://econ.lange-analytics.com/aibook/

EYE BALLING TECHNIQUES TO IDENTIFY RED AND WHITE WINES

HORIZONTAL BOUNDARY



Horizontal Decision Boundary for Acidity and Total Sulfur Dioxide Related to Wine Color

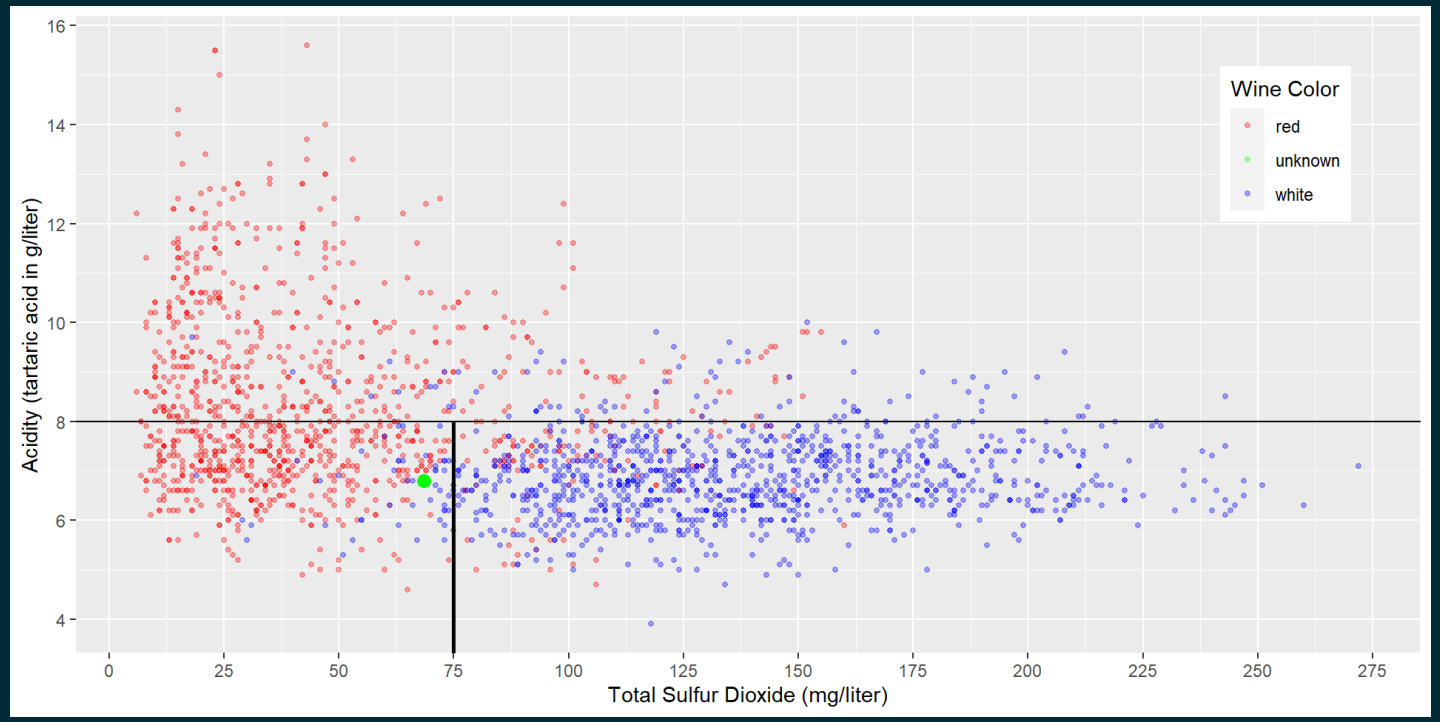
CONFUSION MATRIX

	Truth	
Prediction	Red Wine	White Wine
Red Wine	TP: 'half'	FP: 'few'
White Wine	FN: 'half'	TN: 'most'

https://econ.lange-analytics.com/aibook/

EYEBALLING TECHNIQUES TO IDENTIFY RED AND WHITE WINES

CREATING SUBSPACES LIKE SIMILAR TO A DECISION TREE



Sub-Space Boundaries for Acidity and Total Sulfur Dioxide Related to Wine Color

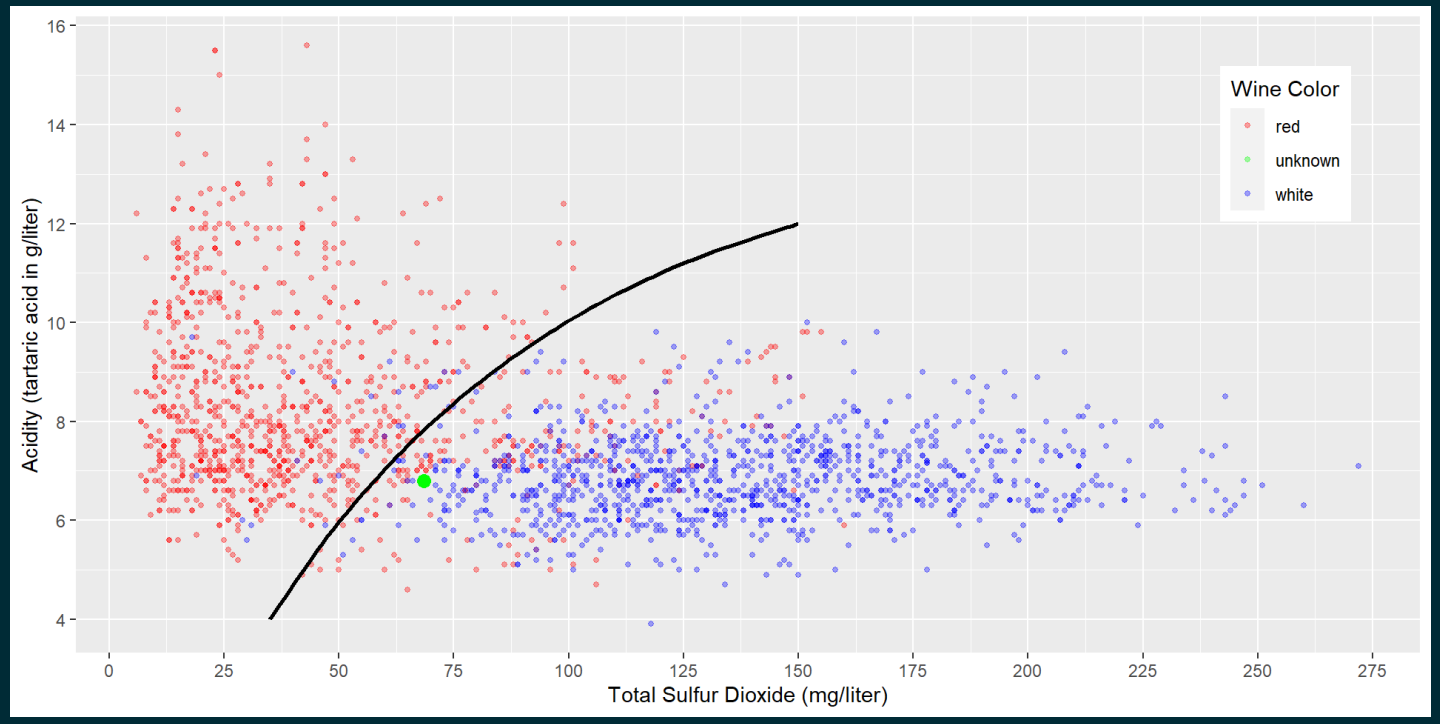
CONFUSION MATRIX

	Truth	
Prediction	Red Wine	White Wine
Red Wine	TP: 'most'	FP: 'few'
White Wine	FN: 'few'	TN: 'most'

https://econ.lange-analytics.com/aibook/

EYEBALLING TECHNIQUES TO IDENTIFY RED AND WHITE WINES

USING A NON-LINEAR DECISION BOUNDARY LIKE A NEURAL NETWORK



Curved Decision Boundary for Acidity and Total Sulfur Dioxide Related to Wine Color

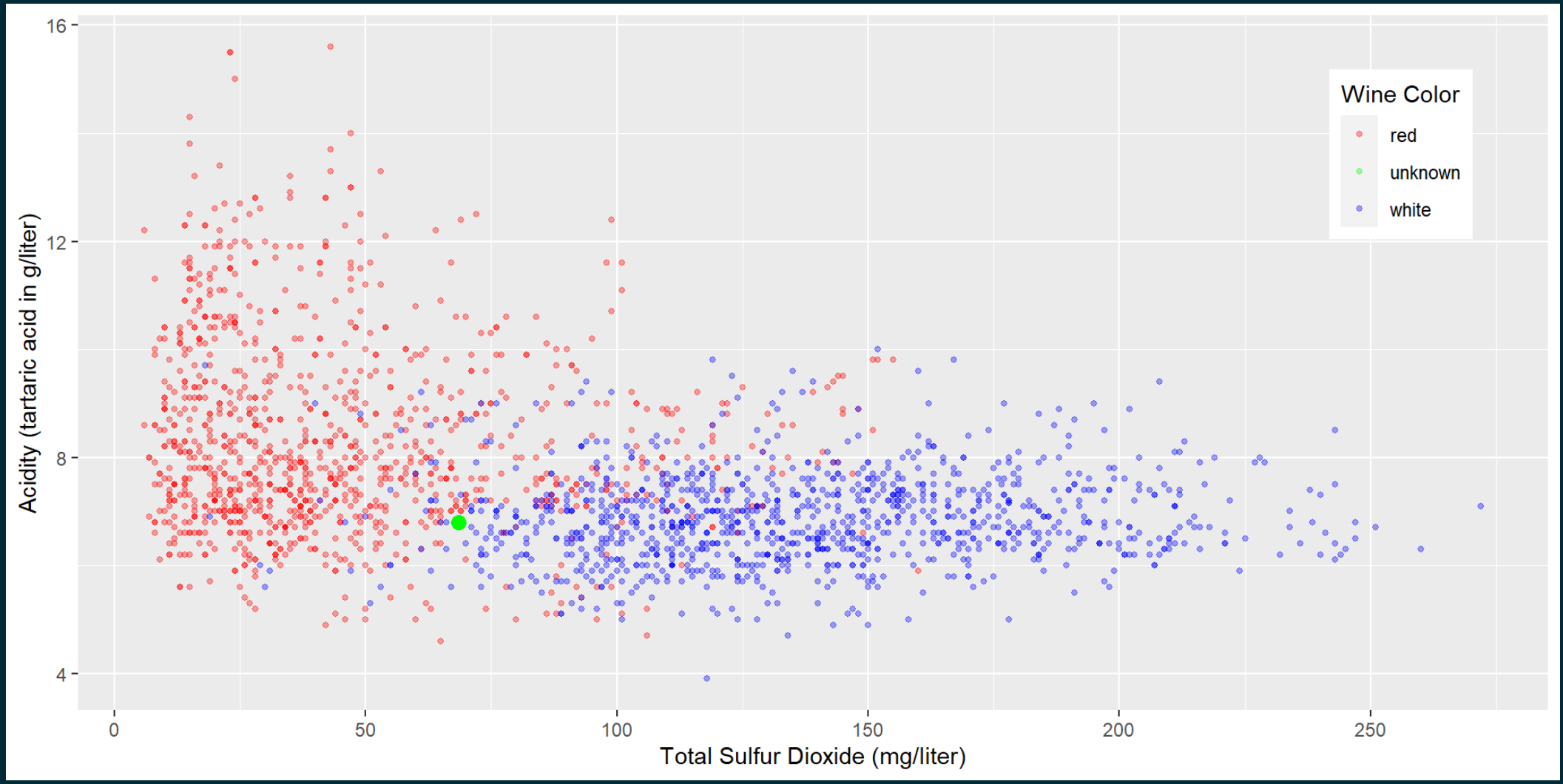
CONFUSION MATRIX

	Truth	
Prediction	Red Wine	White Wine
Red Wine	TP: 'most'	FP: 'few'
White Wine	FN: 'few'	TN: 'most'

https://econ.lange-analytics.com/aibook/

SO, HOW DOES K NEAREST NEIGHBORS WORK?

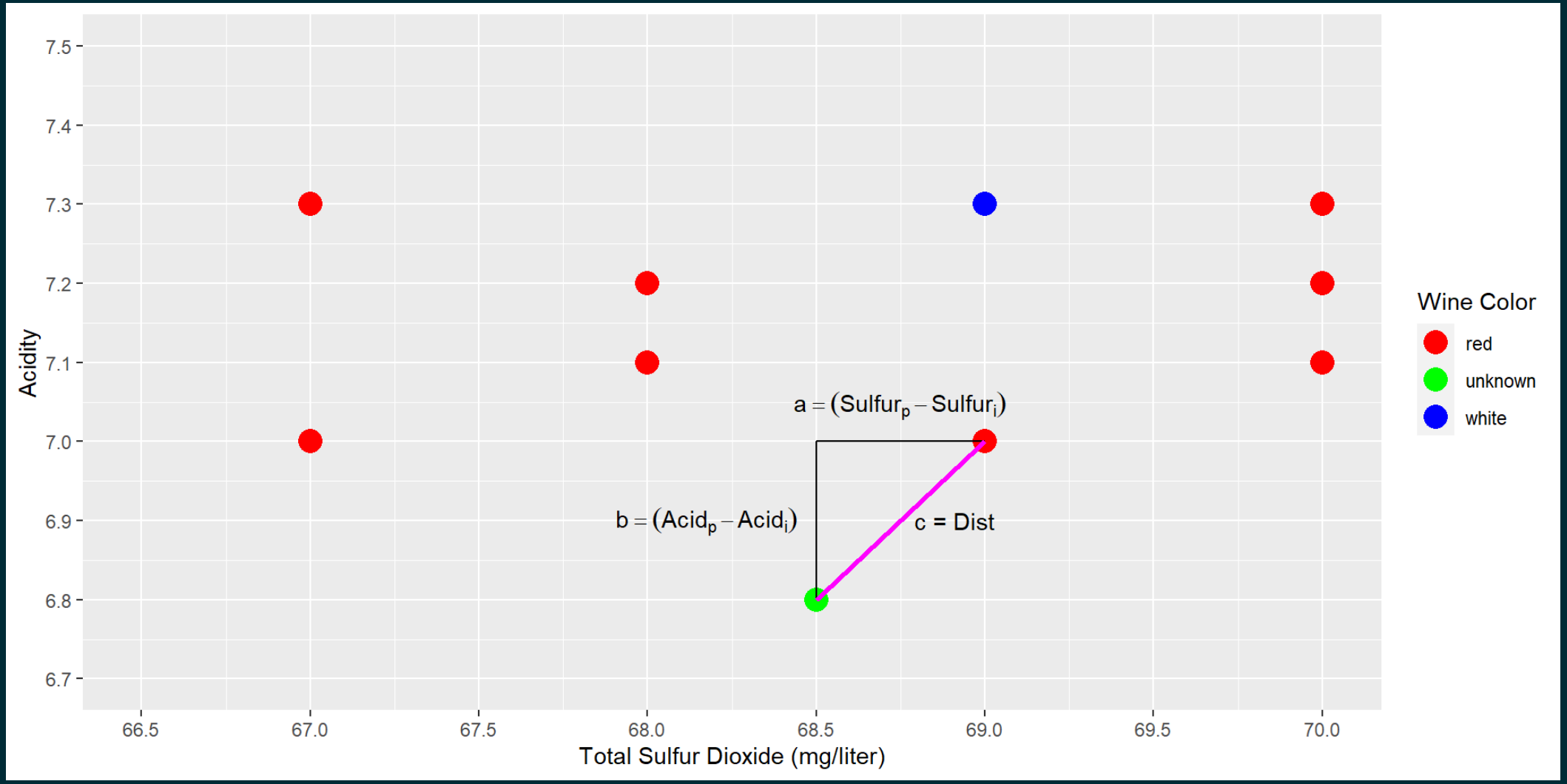
K NEAREST NEIGHBORS K=1



Acidity and Total Sulfur Dioxide Related to Wine Color

https://econ.lange-analytics.com/aibook/

K NEAREST NEIGHBORS K=1



Predicting Wine Color with k-Nearest Neighbors (k=1)

https://econ.lange-analytics.com/aibook/

HOW TO CALCULATE EUCLIDEAN DISTANCE FOR TWO VARIABLES

Assume our observations have **two predictor variables x and y**. We compare the unknown point p to one of the points from the training data (e,g., point i):

$$Dist_i = \sqrt{(x_p - x_i)^2 + (y_p - y_i)^2}$$

»

HOW TO CALCULATE EUCLIDEAN DISTANCE FOR THREE VARIABLES

Assume our observations have **three predictor variables x, y, and z.** We compare the unknown point p to one of the points from the training data (e,g., point i):

$$Dist_i = \sqrt{(x_p - x_i)^2 + (y_p - y_i)^2 + (z_p - z_i)^2}$$

»

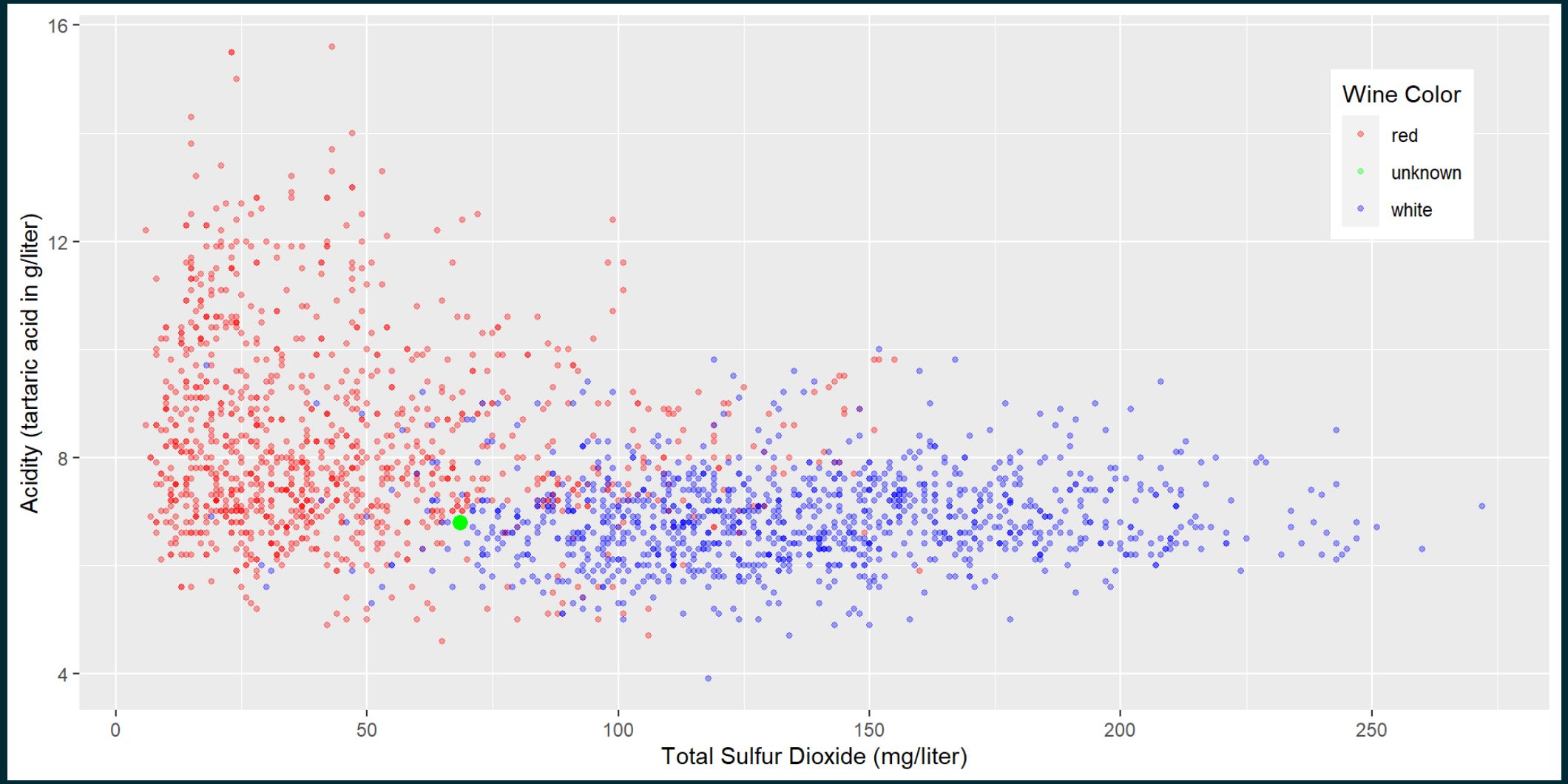
HOW TO CALCULATE EUCLIDEAN DISTANCE FOR N VARIABLES

Assume our observations have N **predictor variables** v_j **with** $j = 1...N$. We compare the unknown point p to one of the points from the training data (e,g., point i):

$$Dist_i = \sqrt{\sum_{j=1}^{N}(v_{p,j} - v_{i,j})^2}$$

»

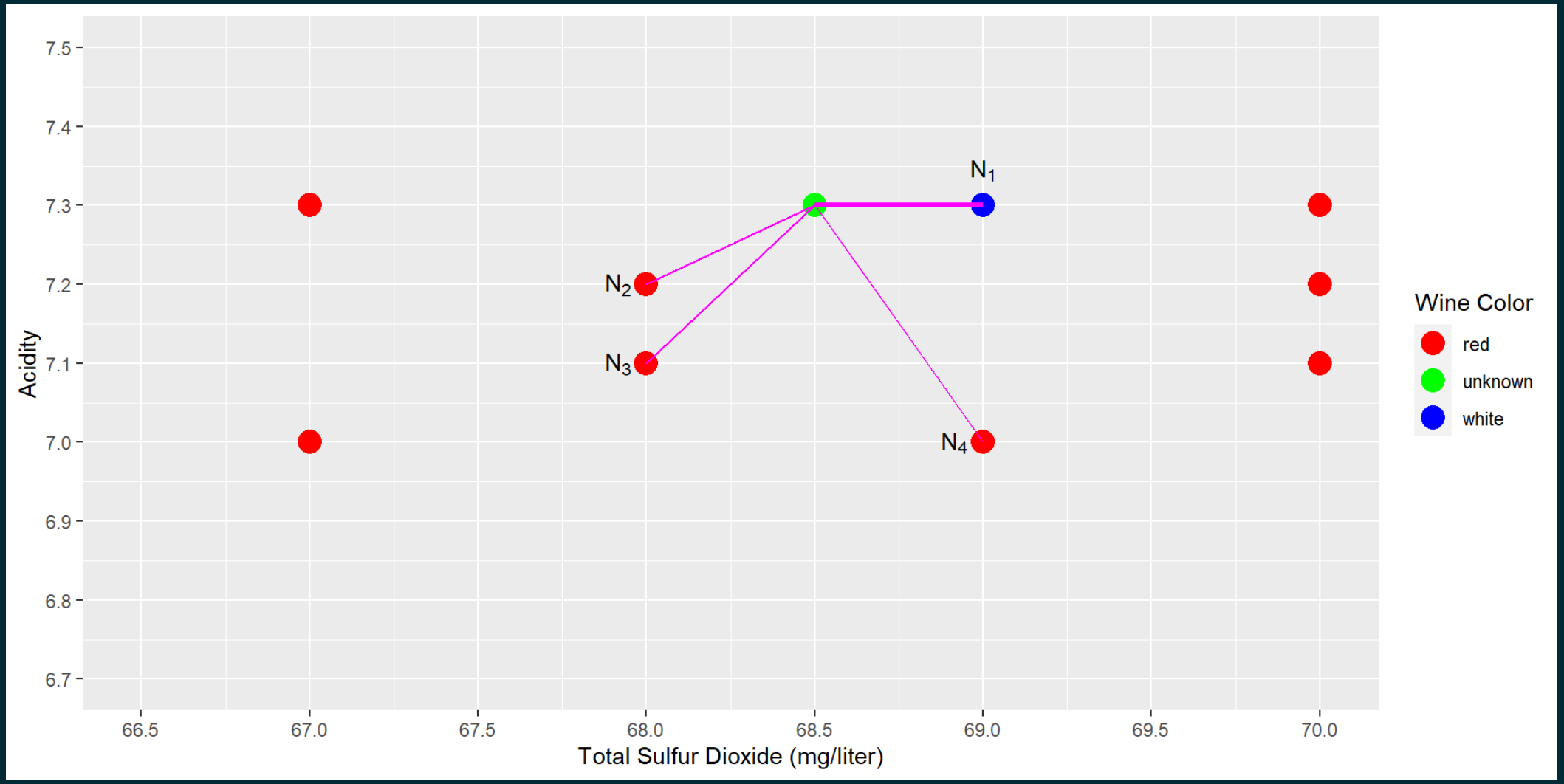
K NEAREST NEIGHBORS K=4



Acidity and Total Sulfur Dioxide Related to Wine Color

https://econ.lange-analytics.com/aibook/

K NEAREST NEIGHBORS K=4



Predicting Wine Color with k-Nearest Neighbors (k=4)

https://econ.lange-analytics.com/aibook/

TO BE CONTINUED

https://econ.lange-analytics.com/aibook/