K-means Clustering report

Get the Data

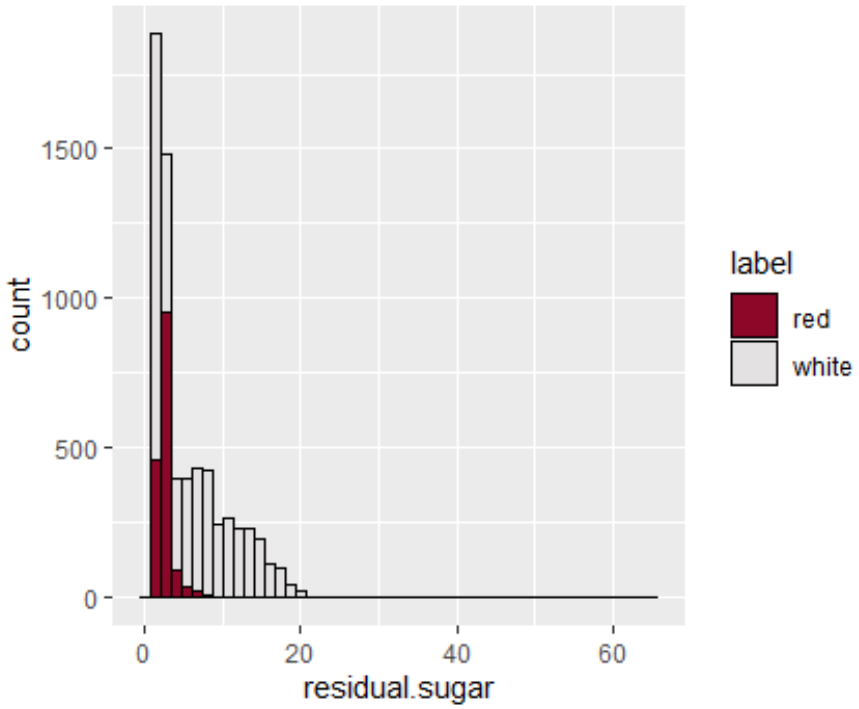
```
df1 <- read.csv("winequality-red.csv", sep = ";")
df2 <- read.csv("winequality-white.csv", sep = ";")
head(df1)
head(df2)

df1$label <- "red"
df2$label <- "white"
wine <- rbind(df1,df2)
head(wine)
str(wine)
```

EDA

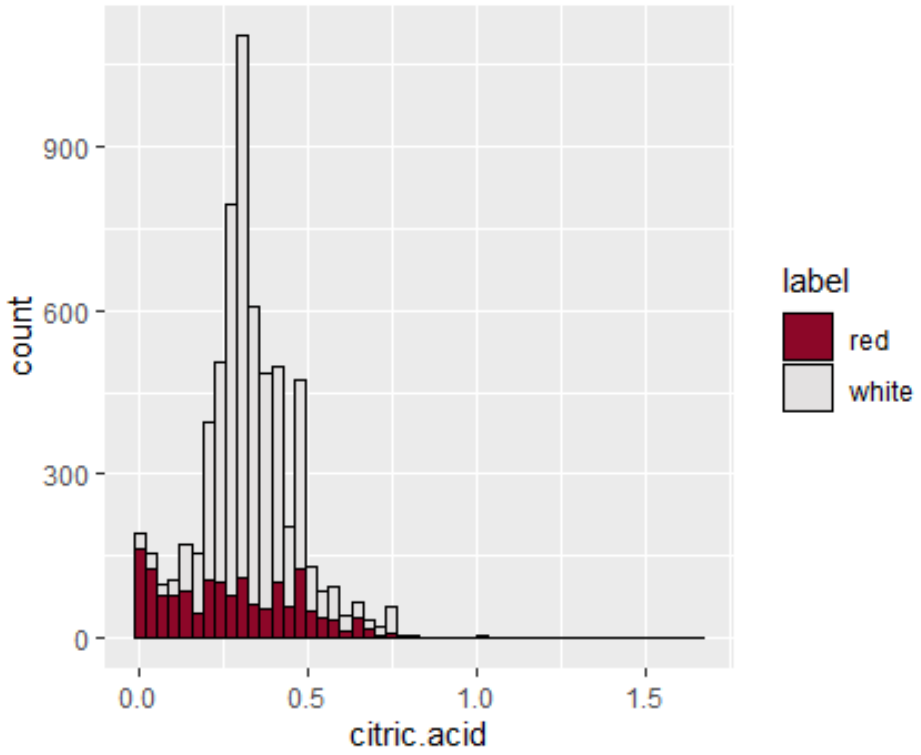
```
library(ggplot2)

# Create Histogram of residual.sugar Color by label
pl <- ggplot(wine, aes(residual.sugar))
pl + geom_histogram(bins = 50,aes(fill = label), color = "black", position = position_stack(reverse = TRUE)) + scale_fill_manual(values=c("
```



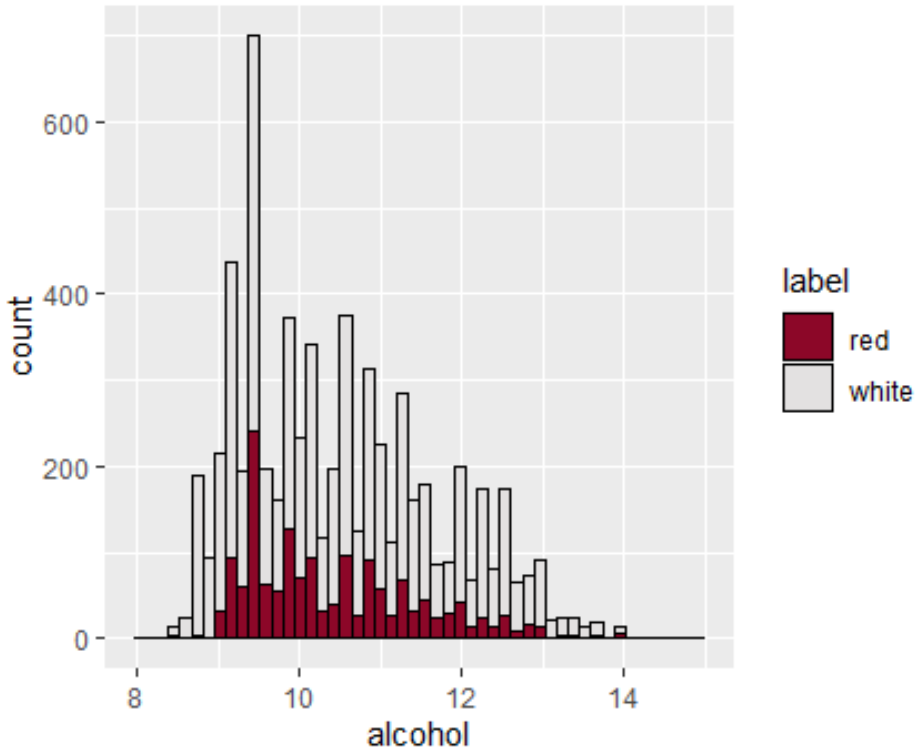
- เราพบว่าไวน์แดงมีแนวโน้มว่าจะมีจำนวนน้ำตาลที่ลดลง
- ข้อผิดพลาดที่เกิดขึ้นคือ จำนวนไวน์แดงนั้นมีค่าน้อยกว่าจำนวนไวน์ขาวมากเกินไป อาจทำให้ข้อมูลมีการเบี่ยงเบนไปเล็กน้อย
- สิ่งที่จะตามมาก็คือการตรวจจับไวน์แดงจะทำได้ค่อนข้างยาก หากไม่มีคุณสมบัติเฉพาะเจาะจงใดๆ

```
# Create Histogram citric.acid Color by label
pl2 <- ggplot(wine, aes(citric.acid))
pl2 + geom_histogram(bins = 50,aes(fill = label), color = "black", position = position_stack(reverse = TRUE)) + scale_fill_manual(values=c(
```



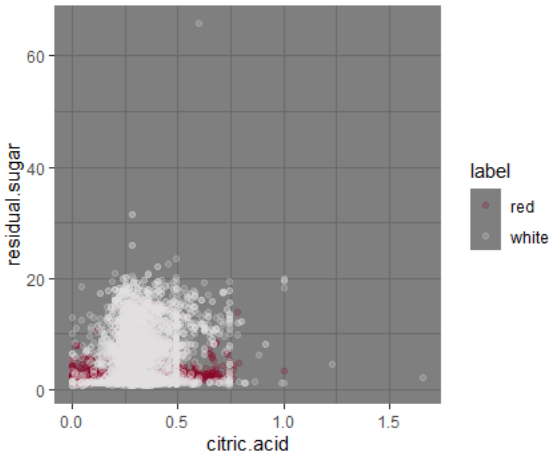
- กรณีของ citric acid ก็เช่นกันการที่เราเห็นจากกราฟว่าไวน์ขาวนั้นมากกว่า เนื่องจากปริมาณที่ต่างกันของข้อมูล
- ซึ่งอาจจะเป็นคุณสมบัติที่ไม่สามารถเอามาจัดกลุ่มได้ชัดเจน เนื่องจากอาจจะเป็นคุณสมบัติเด่นทั้งไวน์ขาวและไวน์แดง

```
# Create Histogram alcohol Color by label
pl3 <- ggplot(wine, aes(alcohol))
pl3 + geom_histogram(bins = 50,aes(fill = label), color = "black", position = position_stack(reverse = TRUE)) + scale_fill_manual(values=c(
```



- จำนวน alcohol ก็เช่นกัน
- ข้อสังเกต จะเห็นได้ว่ากราฟมีลักษณะเป็นคลื่นขึ้นลง เนื่องจากอาจจะเกิดจากการปัดเศษที่ติดในฉลากขวด จึงทำให้เกิดการจัดกลุ่มที่ค่อนข้าวจะสวิ่ง

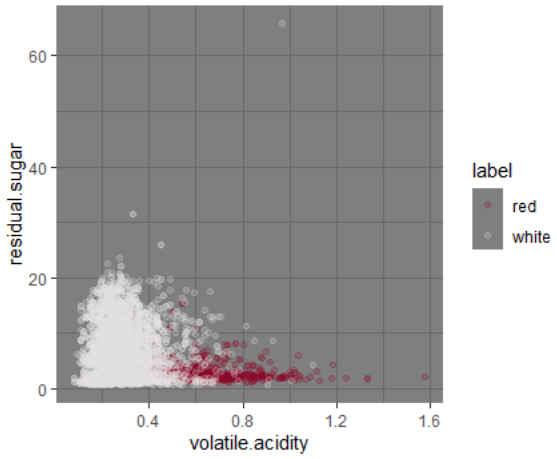
```
# Create scatterplot residual.sugar versus citric.acid, color label
pl4 <- ggplot(wine, aes(y = residual.sugar, x = citric.acid))
pl4 + geom_point(aes(color = label), alpha = 0.2) + scale_color_manual(values=c("#8c0728","#e3e1e1")) + theme_dark()
```



- มีข้อสังเกตที่ระดับน้ำตาล 60g ที่ดูไม่เกาะกลุ่ม อาจจะคัดออกจากการวิเคราะห์หรือทำการแทนที่ด้วยค่าเฉลี่ยของกลุ่ม
- อย่างเช่นเคย acid แยกแยะความแตกต่างและจำแนกกลุ่มค่อนข้างยาก

```
# Create scatterplot volatile.acidity versus residual.sugar, color by label
pl5 <- ggplot(wine, aes(x = volatile.acidity, y = residual.sugar))
```

```
pl5 + geom_point(aes(color = label), alpha = 0.2) + scale_color_manual(values=c("#8c0728","#e3e1e1")) + theme_dark()
```



- volatile acidity อาจจะดูแหละที่จะเลือกเป็นตัวแปรที่ใช้พิจารณาการแยกแยะระหว่างไวน์แดงและไวน์ขาว
- แต่ก็อาจจะเป็นเรื่องยากที่จะแยกความแตกต่างได้เนื่องจากมี Rose wine ซึ่งจัดอยู่ในประเภทไวน์ขาว ซึ่งยังคงรักษาคุณสมบัติความเป็นไวน์แดงอยู่
- ถ้าหากมีความรู้ด้านการจัดการโดเมนอาจจะมองว่าตอนนี้มีไวน์อยู่ 3 ประเภท คือ ขาว แดง ชมพู ก็ได้เช่นกัน

→ Grab the wine data without the label and call it clus.data

```
clus.data <- wine[,-13]
head(clus.data)
```

Building the Clusters

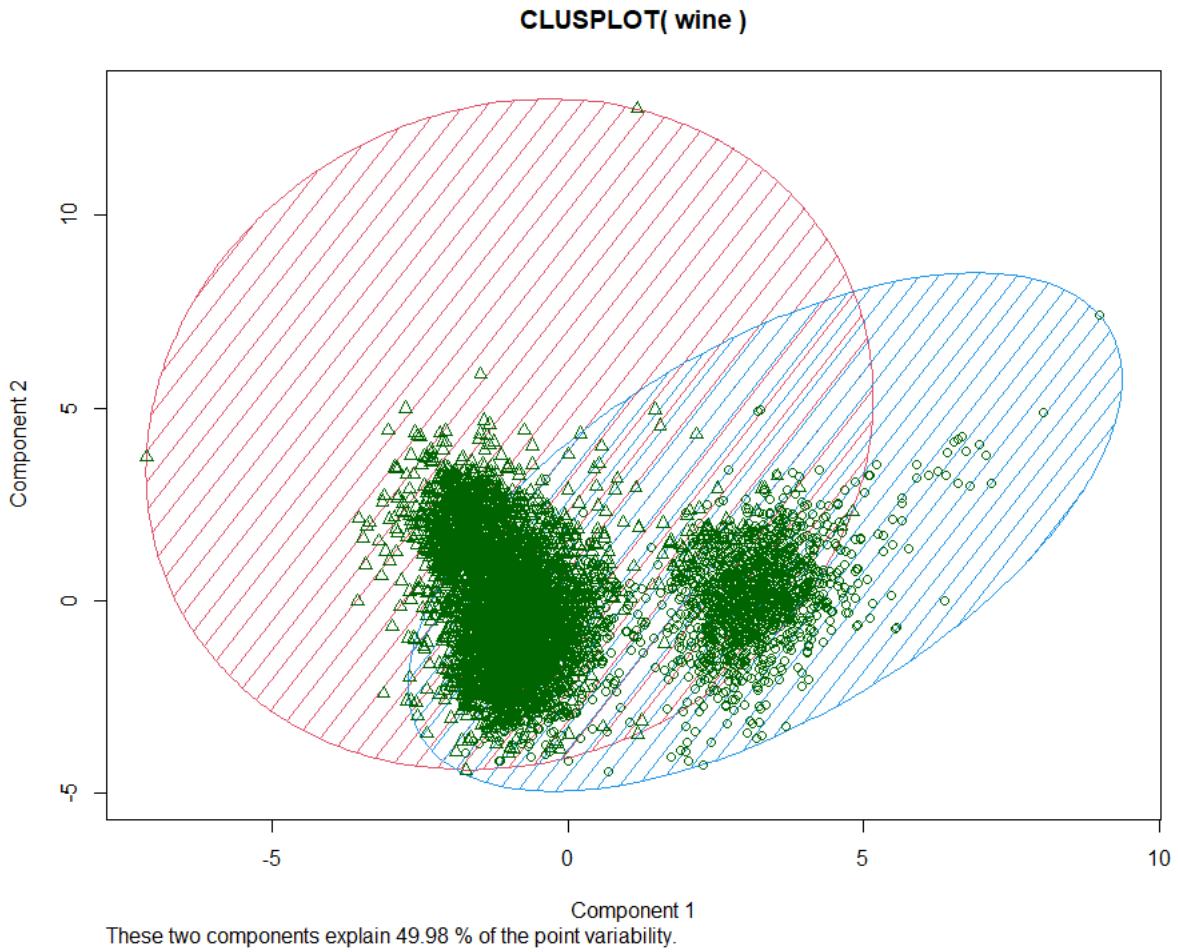
→ Call the kmeans function on clus.data and assign the results to wine.cluster.

```
set.seed(24)
dfCluster <- kmeans(clus.data, 2, nstart = 20)
```

Evaluating the Clusters

```
# evaluating the clusters
table(wine$label, dfCluster$cluster)

# Cluster Visualizations
clusplot(wine, dfCluster$cluster, color=TRUE, shade=TRUE, labels=0,lines=0, )
```



เราจะเห็นว่าสีแดงรวมเข้าด้วยกันได้ง่ายกว่า ซึ่งสมเหตุสมผลเมื่อพิจารณาจากการสร้างภาพข้อมูลก่อนหน้านี้ ไวน์ขาวดูเหมือนจะไม่ค่อยมี noise รบกวน อาจเป็นเพราะ "Rose wines" จัดอยู่ในประเภทไวน์ขาว ในขณะที่ยังคงรักษาคุณภาพของไวน์แดงเอาไว้ โดยรวมแล้วถือว่าสมเหตุสมผลเนื่องจากข้อมูลการวัดทางเคมีที่ได้รับอาจไม่สัมพันธ์กันดีนักสำหรับไวน์ที่เป็นสีแดงหรือสีขาว!

สิ่งสำคัญที่ควรทราบคือ K-Means สามารถให้เฉพาะคลัสเตอร์เท่านั้น ไม่สามารถบอกได้โดยตรงว่าควรอยู่ในกลุ่มไหน หรือแม้แต่จำนวนคลัสเตอร์ควรเป็นเท่าไร