K Means Clustering Project ¶

Usually when dealing with an unsupervised learning problem, its difficult to get a good measure of how well the model performed. For this project, we will use data from the UCI archive based off of red and white wines (this is a very commonly used data set in ML).

We will then add a label to the a combined data set, we'll bring this label back later to see how well we can cluster the wine into groups.

Get the Data ¶

[Download the two data csv files from the UCI repository](#) **(or just use the downloaded csv files).**

Use read.csv to open both data sets and set them as df1 and df2. Pay attention to what the separator (sep) is.

In [10]:

```
setwd("C:/Users/ADMIN/Desktop/DataScience/R")
df1<-read.csv("winequality-red.csv",sep = ";",header = TRUE)
df1
setwd("C:/Users/ADMIN/Desktop/DataScience/R")
df2<-read.csv("winequality-white.csv",sep = ";",header = TRUE)
df2
```

Now add a label column to both df1 and df2 indicating a label 'red' or 'white'.

In [11]:

```
df1$label<-c('red')
df1
df2$label<-c('white')
df2
```

Check the head of df1 and df2.

In [12]:

```
head(df1)
```

Out[12]:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	label
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	red
2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5	red
3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5	red
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	red
5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	red
6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	red

In [13]:

```
head(df2)
```

Out[13]:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	label
1	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6	white

2	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6	white
3	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	white
4	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	white
5	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	white
6	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	white

Combine df1 and df2 into a single data frame called wine.

In [14]:

```
install.packages('dplyr')
library(dplyr)
wine<-full_join(df1,df2,by=NULL,type='left',match='all')
wine
str(wine)
```

In [15]:

```
str(wine)
```

```
'data.frame':6497 obs. of  13 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
 $ label               : chr  "red" "red" "red" "red" ...
```

EDA¶

Let's explore the data a bit and practice our ggplot2 skills!

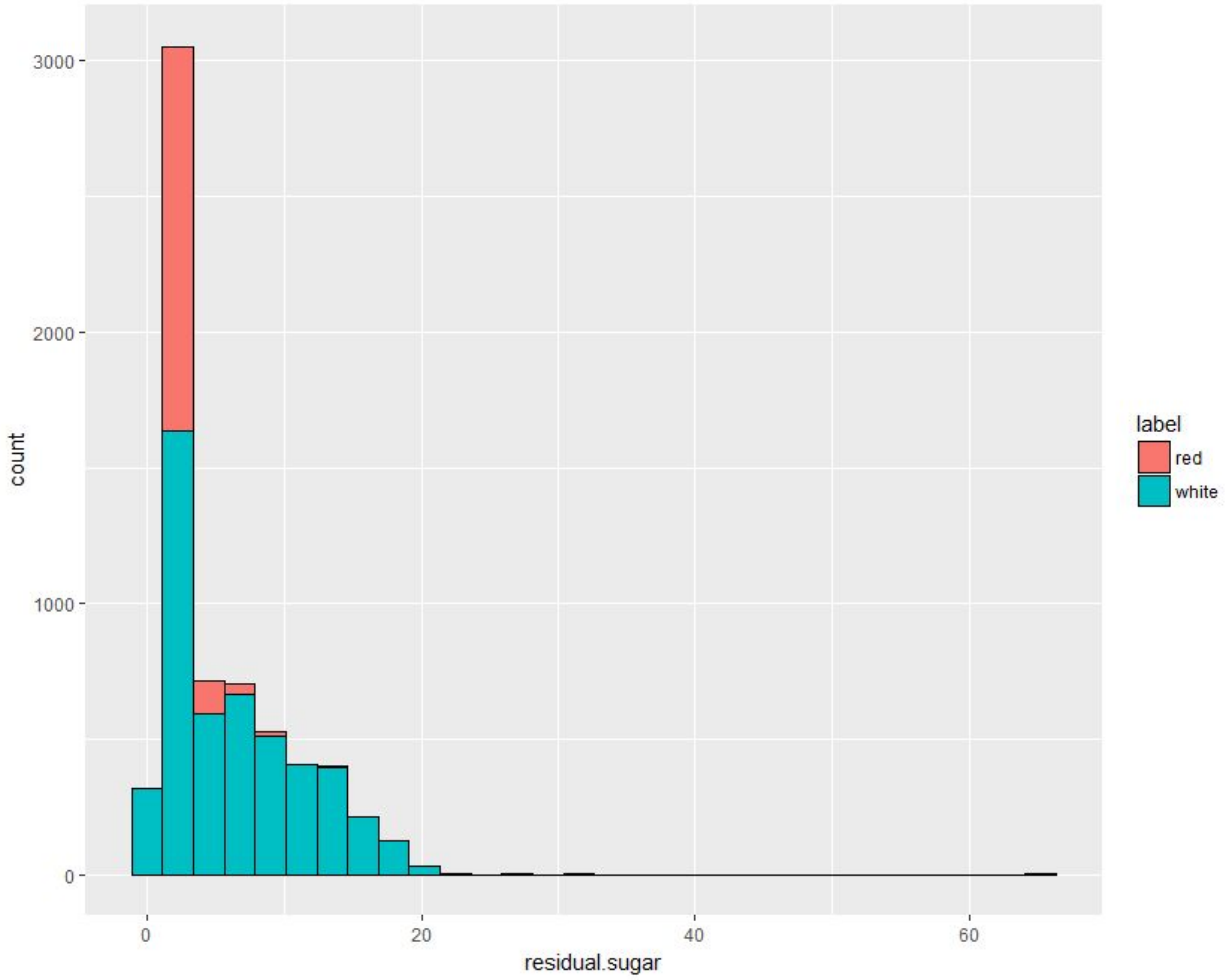
Create a Histogram of residual sugar from the wine data. Color by red and white wines.

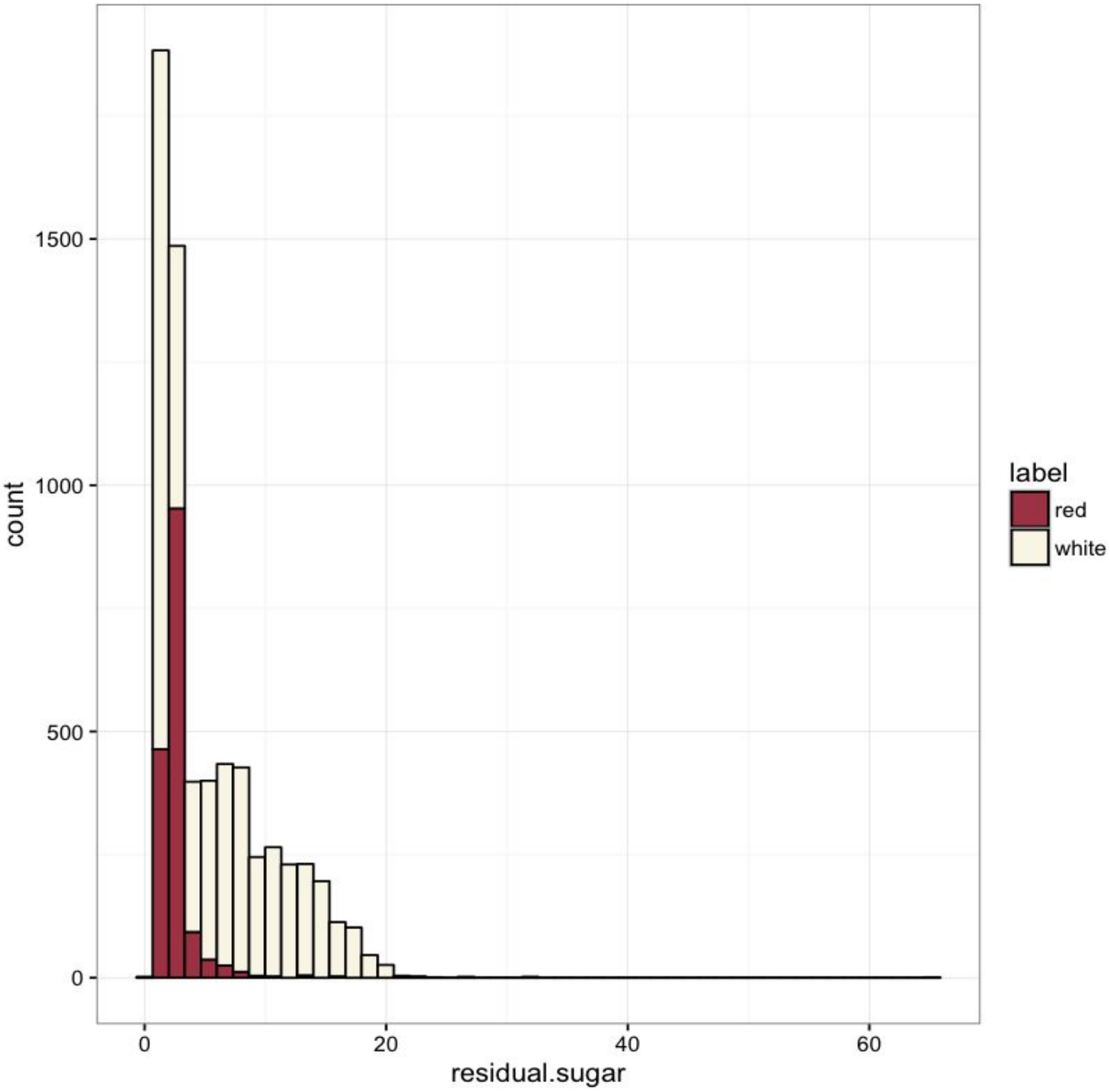
In [16]:

```
ggplot(data = wine,aes(x=residual.sugar,fill=label,colour =c("#FF0000","#ffffff")))+geom_histogram(col='black')
```

Note: Even after mentioning the colors as white and red ,graph wont changing its colors.

In [37]:

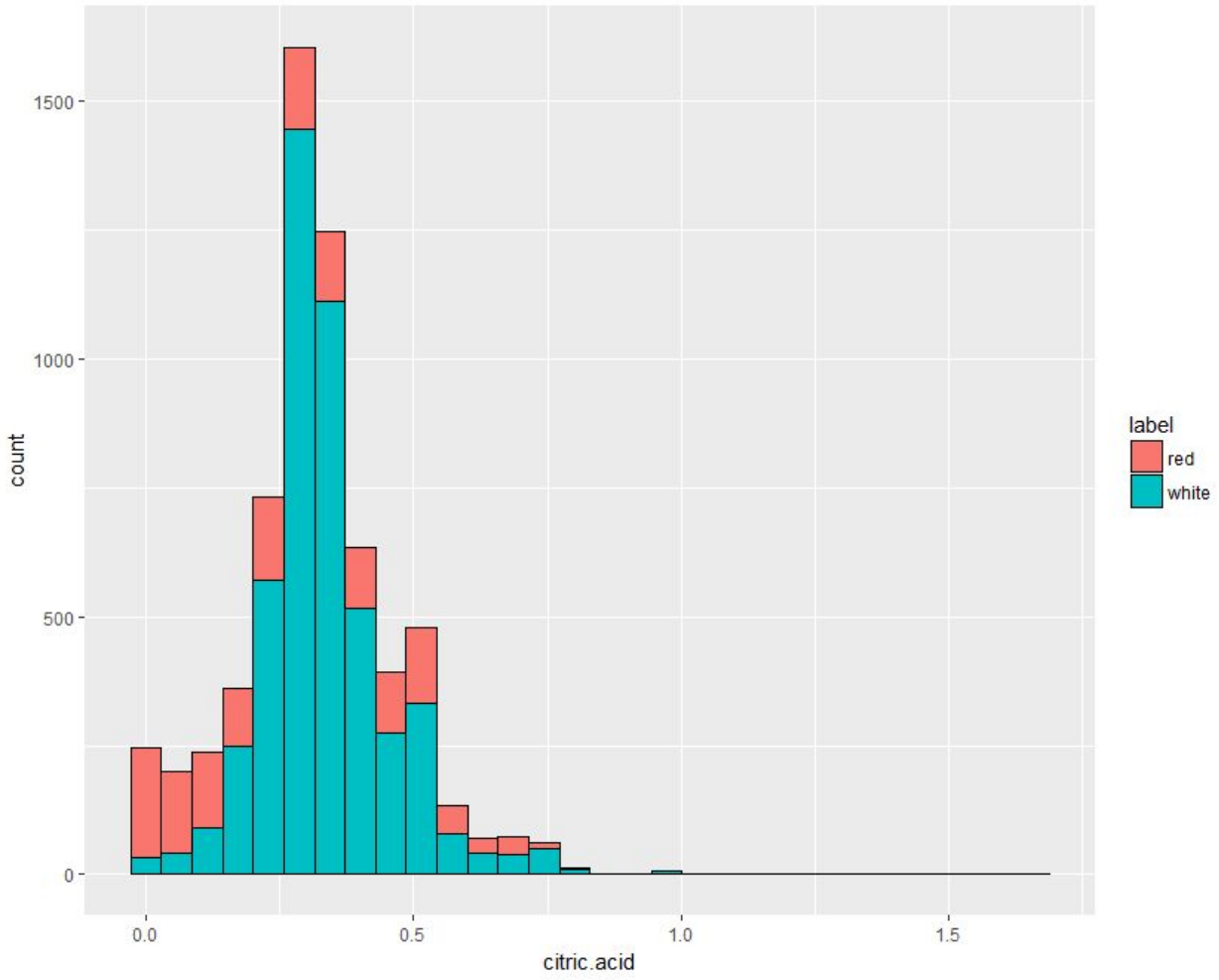


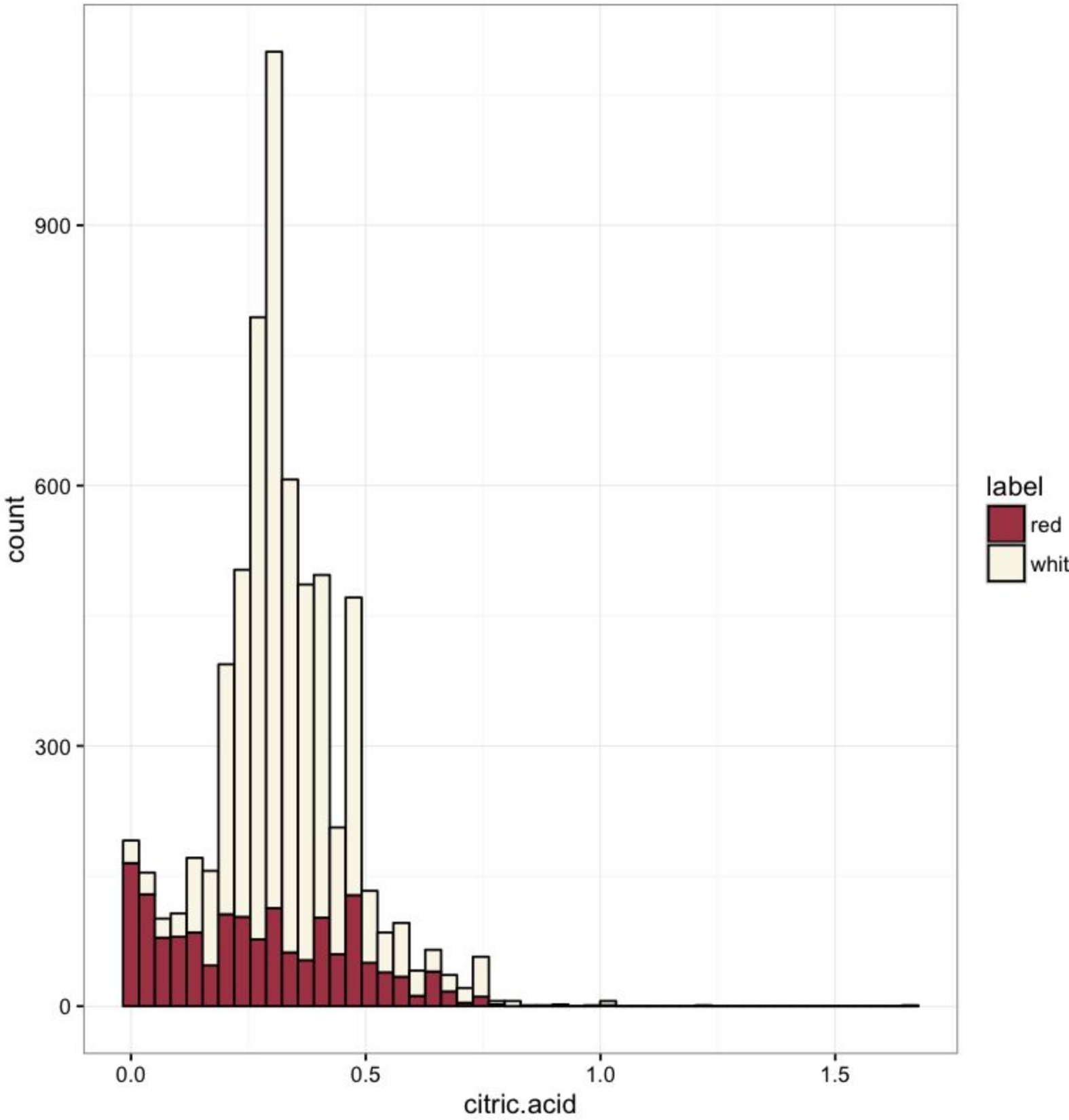


Create a Histogram of citric.acid from the wine data. Color by red and white wines.

In [39]:

```
ggplot(data = wine,aes(x=citric.acid,fill=label,colour =c("#FF0000","#ffffff")))+geom_histogram(col='black')
```

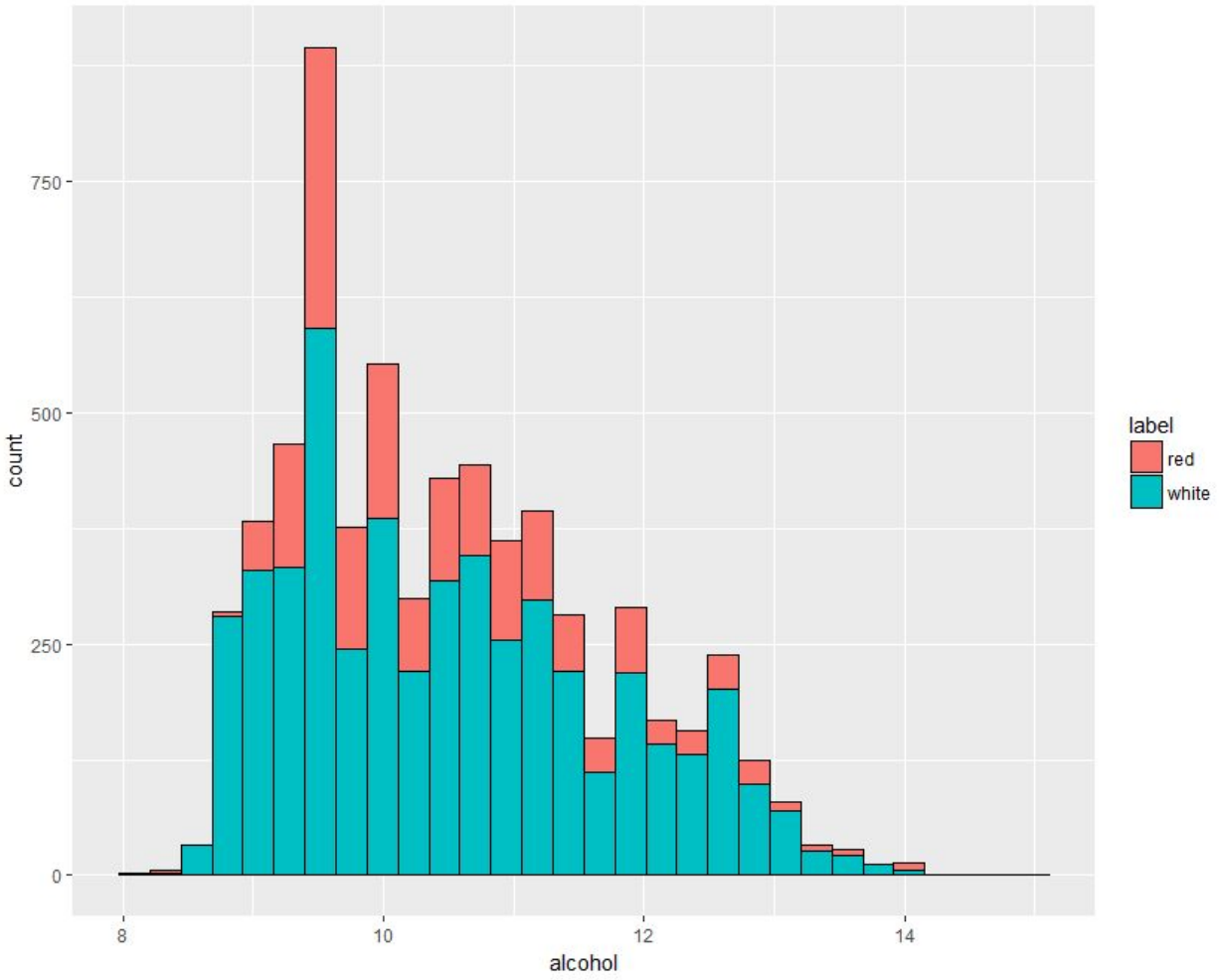

1500
1000
500
0
count
0.0
0.5
1.0
1.5
citric.acid
label
red
white

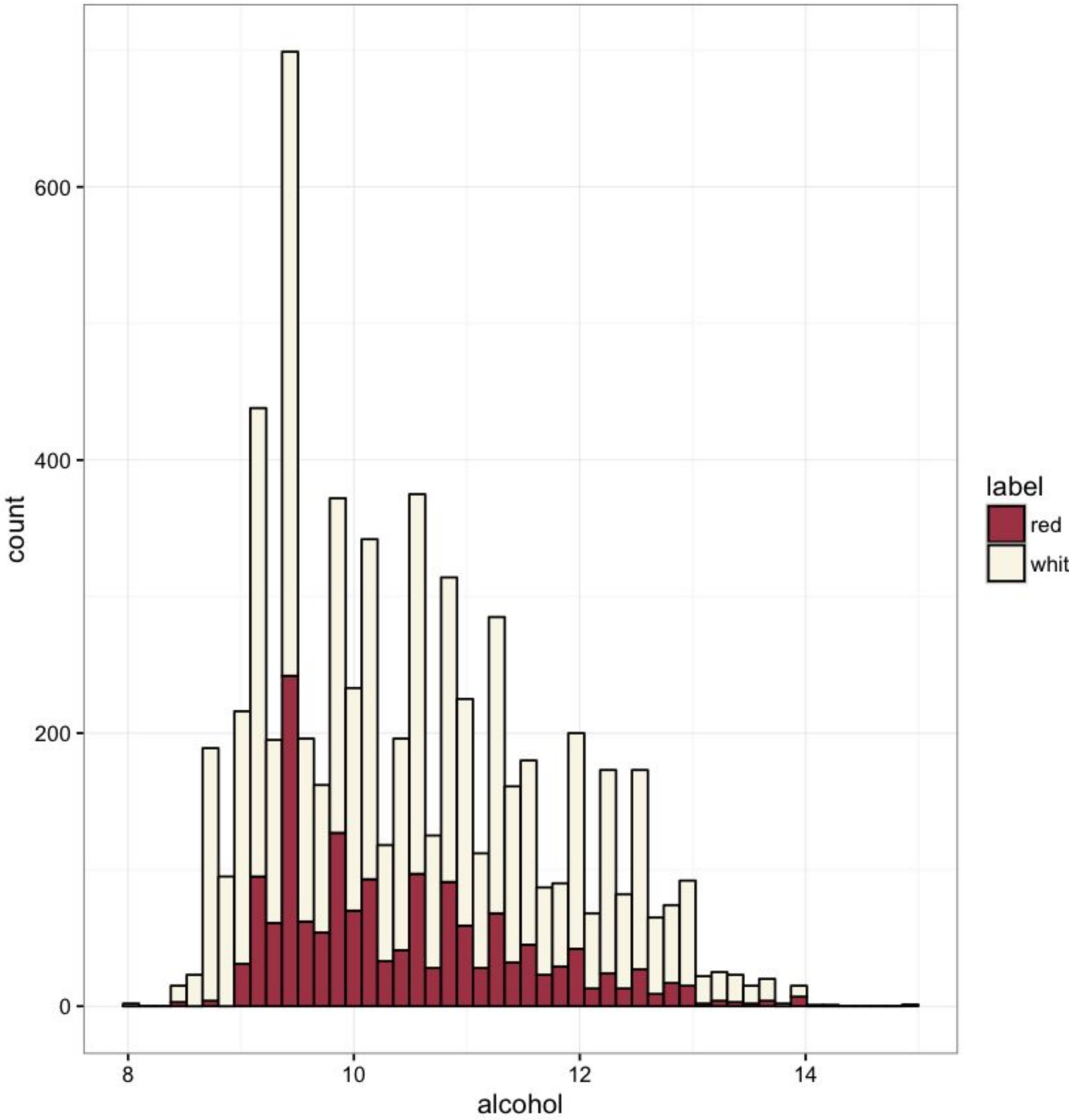


Create a Histogram of alcohol from the wine data. Color by red and white wines.

In [40]:

```
ggplot(data = wine,aes(x=alcohol,fill=label,colour =c("#FF0000","#ffffff")))+geom_histogram(col='black')
```

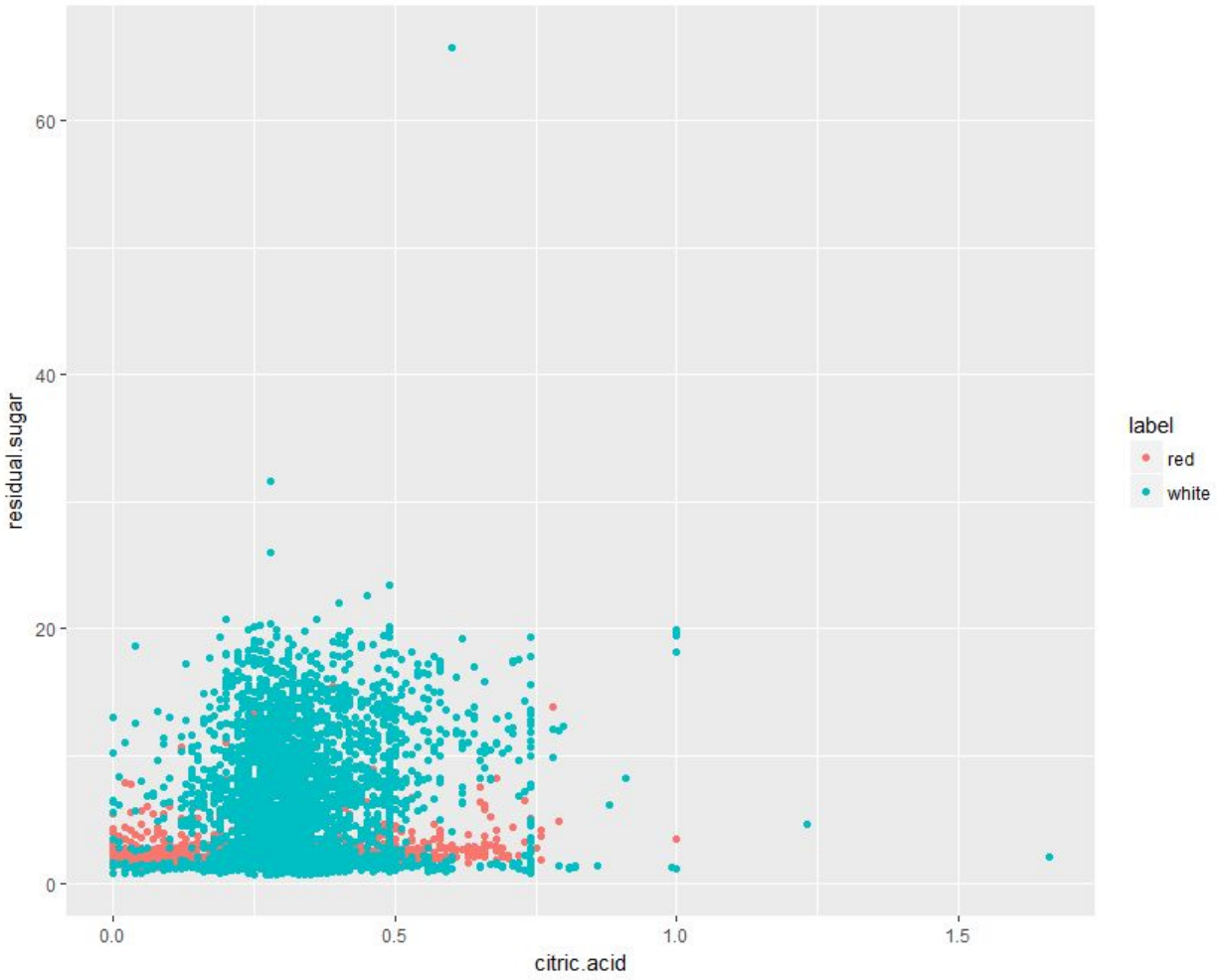


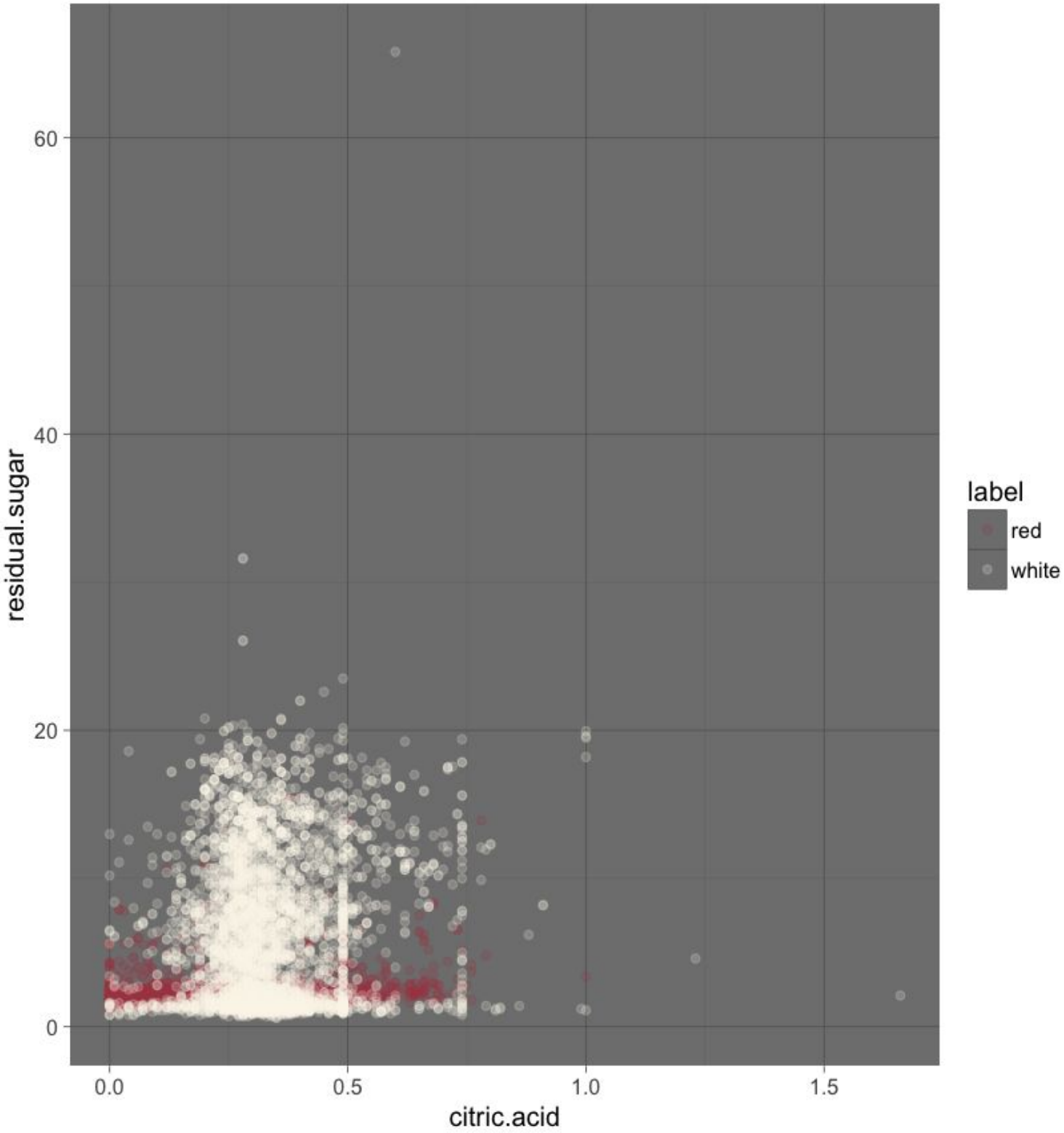


Create a scatterplot of residual.sugar versus citric.acid, color by red and white wine.

In [49]:

```
ggplot(data = wine,aes(x=citric.acid,y=residual.sugar,col=label))+geom_point()
```

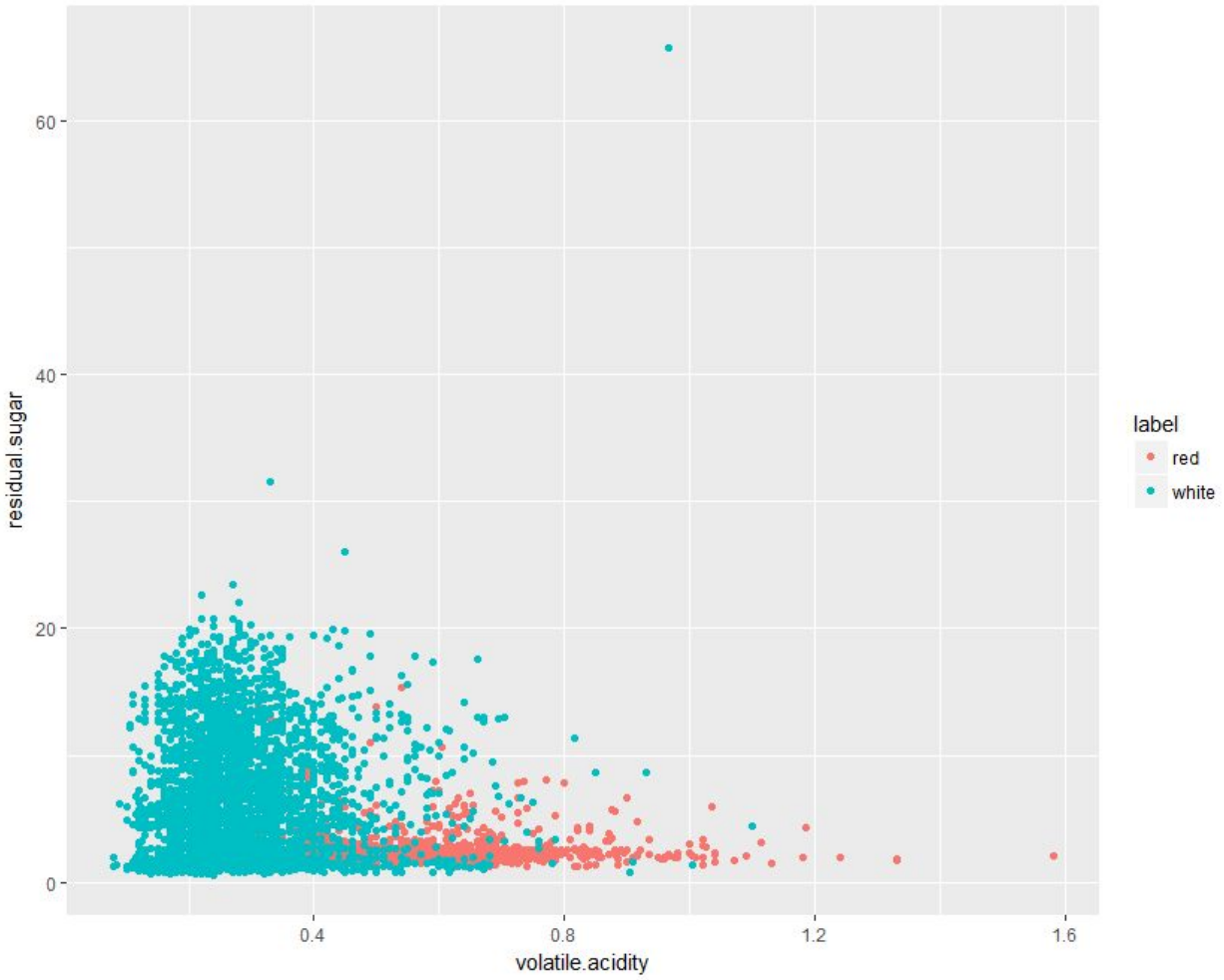


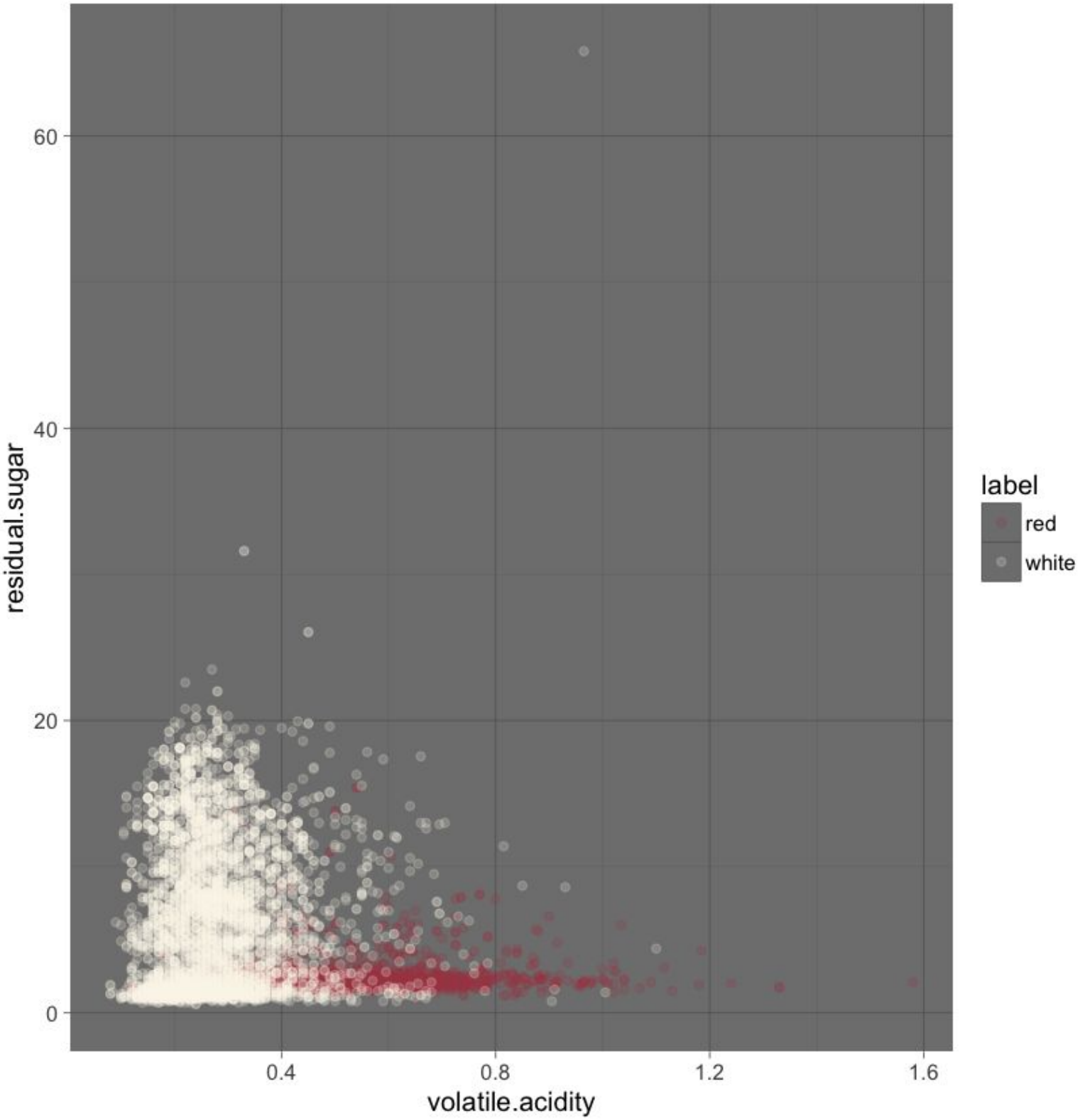


Create a scatterplot of volatile.acidity versus residual.sugar, color by red and white wine.

In [52]:

```
ggplot(data = wine,aes(x=volatile.acidity ,y=residual.sugar,col=label))+geom_point()
```





Feel free to explore the data as you see fit, we'll go ahead and move on!

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

In [65]:

```
clus.data<-wine[,-13]
```

Check the head of clus.data

In [63]:

```
head(clus.data)
```

Out[63]:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	label
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	red
2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5	red
3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5	red
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6	red

5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5	red
6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5	red

Building the Clusters¶

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

In [74]:

```
install.packages('cluster')
library(cluster)
wine.cluster<-kmeans(clus.data,centers = 2)
```

Print out the wine.cluster Cluster Means and explore the information.

In [76]:

```
wine.cluster
```

o/p:

```
Cluster means:
  fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides free.sulfur.dioxide
1      6.904812        0.2871659   0.3397642       7.244809 0.04859257            39.75590
2      7.623219        0.4086378   0.2908725       3.076425 0.06580983            18.39868
  total.sulfur.dioxide   density       pH sulphates  alcohol  quality
1            155.69246 0.9947903 3.190808 0.4999485 10.25932 5.824343
2             63.26318 0.9945736 3.254882 0.5724145 10.79722 5.810541
```

```
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
1      7.619044        0.4079451   0.2911080       3.082690 0.0656846
2      6.904698        0.2871364   0.3398094       7.259286 0.0486092
  free.sulfur.dioxide total.sulfur.dioxide    density       pH sulphates
1            18.43735             63.54832 0.9945680 3.255147 0.5718655
2            39.82503            155.90101 0.9947956 3.190308 0.5000354
   alcohol  quality
1 10.79529 5.809204
2 10.25832 5.825436
```

Evaluating the Clusters¶

You usually won't have the luxury of labeled data with KMeans, but let's go ahead and see how we did!

Use the table() function to compare your cluster results to the real results. Which is easier to correctly group, red or white wines?

In [85]:

```
table(wine.cluster$cluster,wine$label)
```

Out[85]:

```
       1    2
red   1515   84
white 1310 3588
```

Great Job!