Scribe Notes - 8/10/2015 Module 13

Bob Cook, Dan Zalewski, Nalini Agrawal

August 10, 2015

Exploratory Analysis and PCA in R

The first half of today's class was a further investigation into how to effectively conduct an exploratory analysis and run PCA in R.

Exploratory Analysis:

The most important insight in this exploratory analysis was stratifying the voting equipment variable by the poor (binary) variable. This allows us to see beyond the initial boxplot, which indicates there is nothing interesting going on with the voting equipment.

In class a question came up about boxplots versus bar charts. The boxplot contains more information (inner-quartile range, outliers, etc) that the bar chart does not. So for Exploratory Analysis, the boxplot makes more sense here.

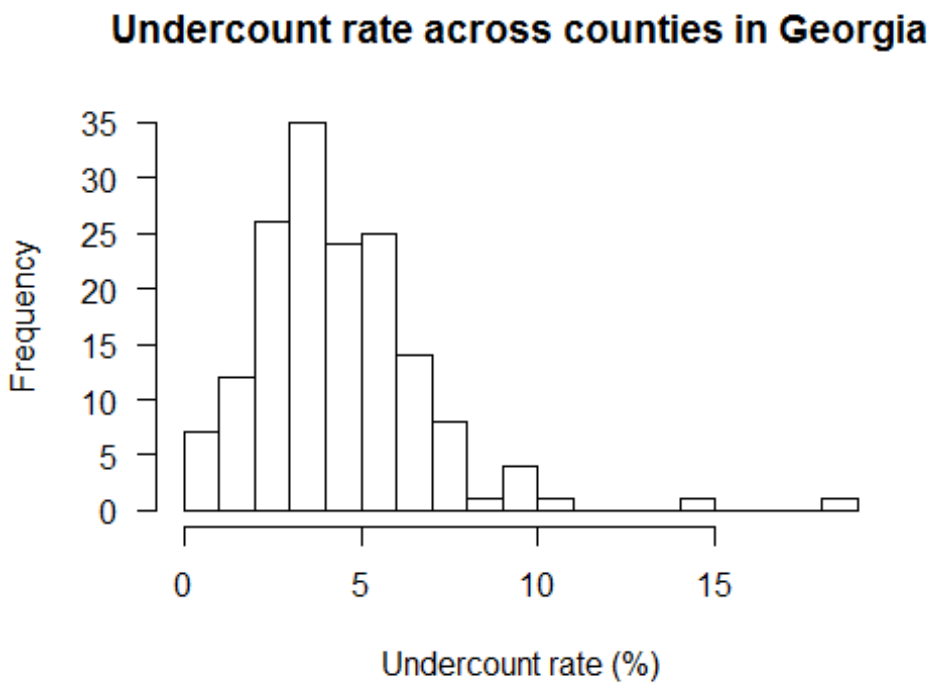
We began by discussing the need to create a variable that reflects the "rate" of undercount within each county, because we want to be able to compare counties regardless of their total number of votes:

```
georgia2000 = read.csv('C:/Users/Bob/Downloads/georgia2000.csv')
attach(georgia2000)

georgia2000$ucount_rate = (georgia2000$ballots -
georgia2000$votes)/georgia2000$ballots * 100
```

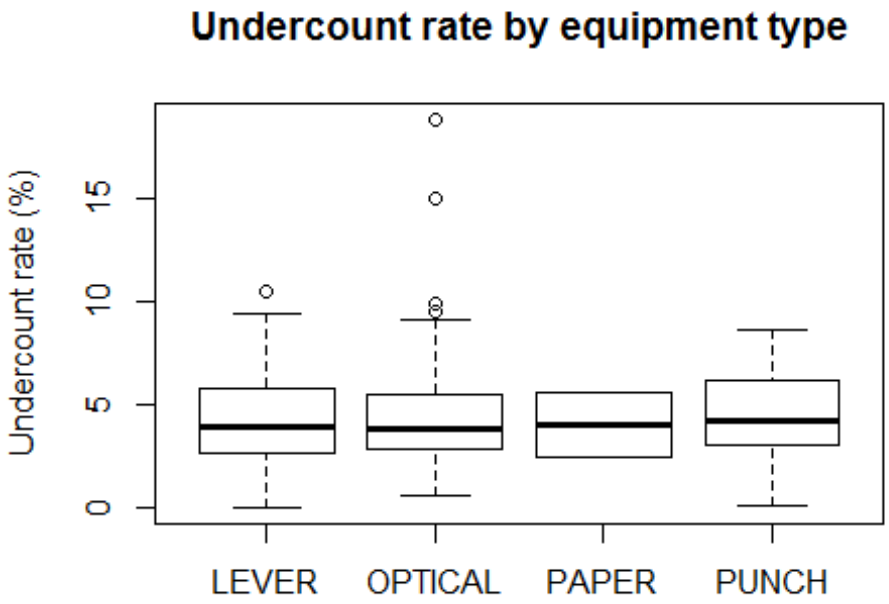
There is apparently wide variability in undercount rates across the 159 counties:

```
hist(georgia2000$ucount_rate, 20, las=1, xlab='Undercount rate (%)',
main='Undercount rate across counties in Georgia')
```



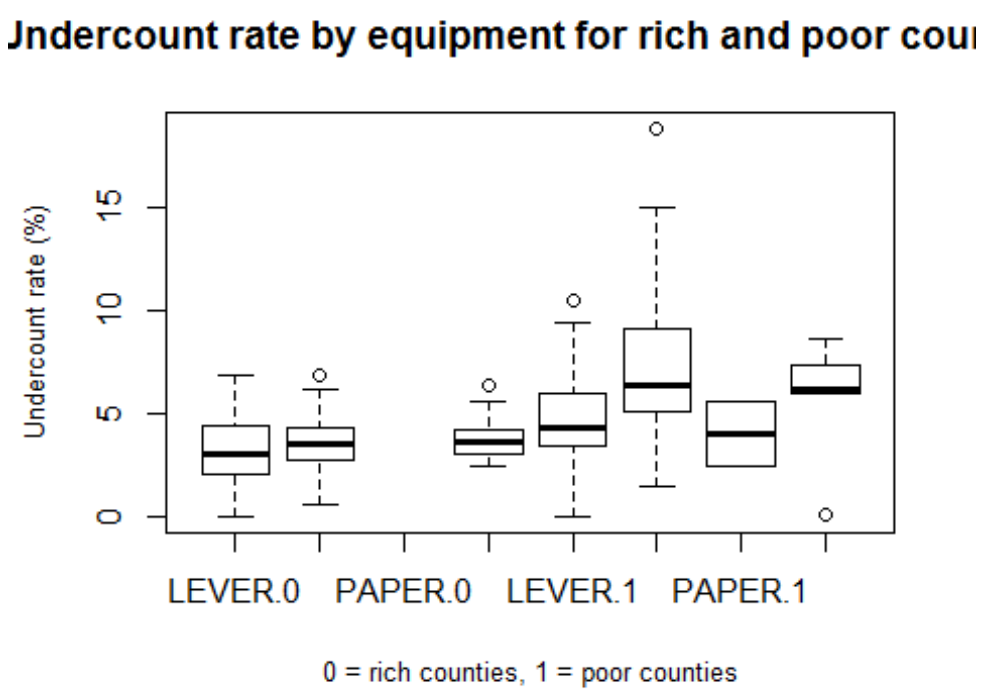
Stratifying the counties by voting-equipment type does not immediately suggest much of a relationship between equipment and undercount rate:

```
boxplot(ucount_rate ~ equip, data=georgia2000, ylab='Undercount rate (%)',
main='Undercount rate by equipment type')
```



If we look at the rich and poor counties separately, however, it appears as though optical machines have a higher undercount rate within poor counties, but not within rich counties. This is a key step in our analysis because we now see something that is potentially important within the equipment variable, which initially appeared to be uninteresting. The figure below shows this:

```
boxplot(ucount_rate ~ equip:poor, data=georgia2000, ylab='Undercount rate
(%)', main='Undercount rate by equipment for rich and poor counties', xlab="0
= rich counties, 1 = poor counties", cex.lab=0.8)
```



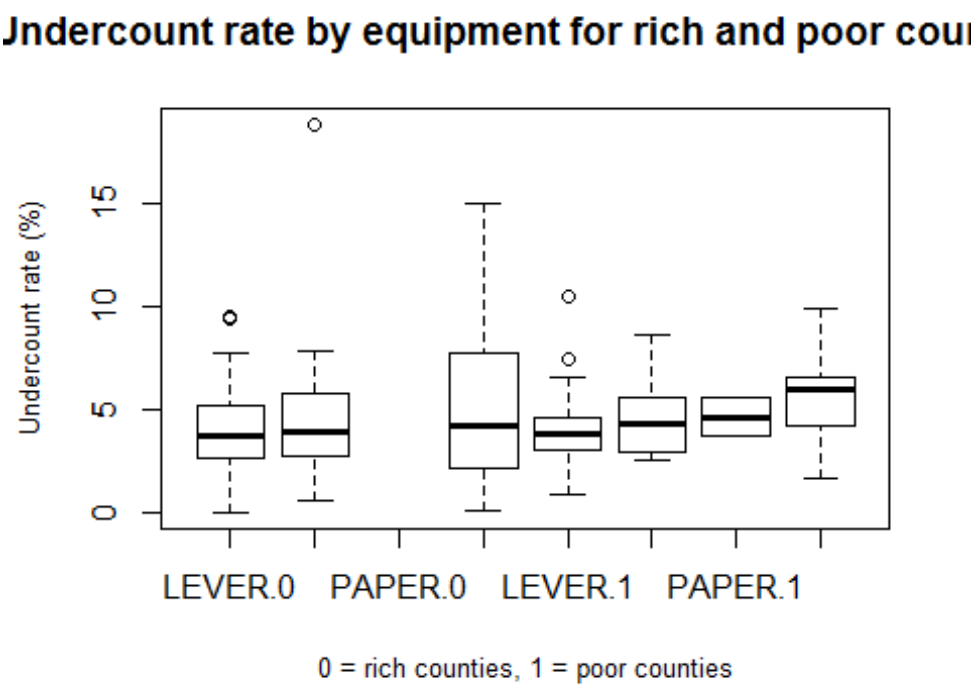
If we are curious about a particular piece of information here, such as why the 1st quartile for PUNCH.1 is so high, we can run a permutation test to see if this is significant or likely due to chance, using the shuffle function on the undercount variable:

```
library(mosaic)
```

```
## Loading required package: car
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
##
## Loading required package: lattice
## Loading required package: ggplot2
## Loading required package: mosaicData
##
## Attaching package: 'mosaic'
##
## The following objects are masked from 'package:dplyr':
##
```

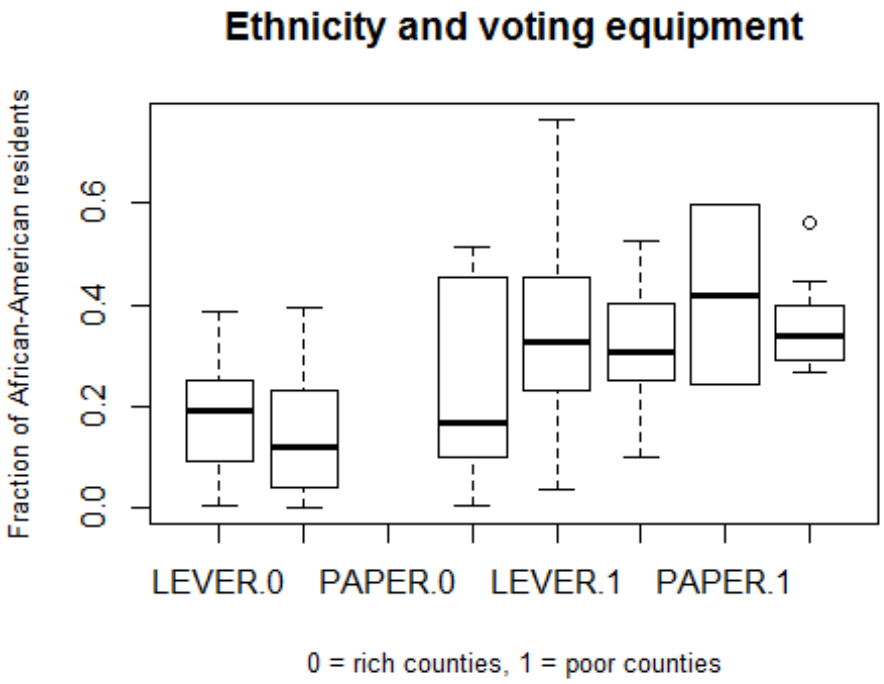
```
##     count, do, tally
##
## The following object is masked from 'package:car':
##
##     logit
##
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cov, D, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
##
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
boxplot(shuffle(ucount_rate) ~ equip:poor, data=georgia2000, ylab='Undercount
rate (%)', main='Undercount rate by equipment for rich and poor counties',
xlab="0 = rich counties, 1 = poor counties", cex.lab=0.8)
```



If we run this permutation test many times, we can see that the high 1st quartile for PUNCH.1 is not consistent, and therefore appears to be due to noise (maybe small sample size).

However, it does not seem as if African-Americans are systematically more likely than whites to live in counties with optical equipment, regardless of whether those counties are poor:



Conclusions from Exploratory Analysis: 1. There is a greater variation of voter undercount among equipment types 2. The voter undercount is not present in African American communities

Here are 2 links that might aid your understanding of PCA:

(1) Simple explanation of role of eigenvectors and eigenvalues in PCA:

https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/

(2) In-depth explanation of PCA, skip ahead to chapter 3:

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

PCA:

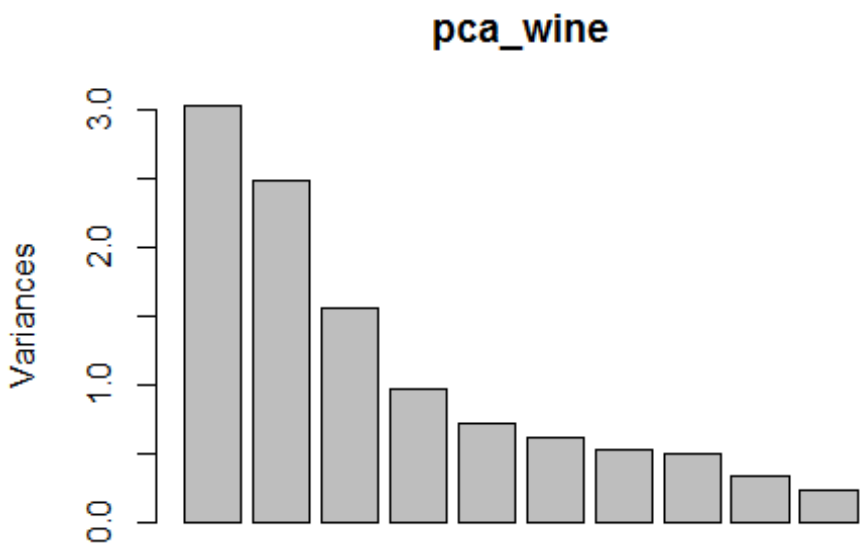
We run PCA on the wine dataset. There is less of the data variability captured by the first few Principal Components here than we saw with the iris data set, as this graph shows:

```
wine = read.csv('C:/Users/Bob/Downloads/wine.csv', header=TRUE)
X = scale(wine[,1:11])
mu = attr(X, 'scaled:center')
sigma = attr(X, 'scaled:scale')
```

```
pca_wine = prcomp(X)
pca_wine$rotation[,1]
```

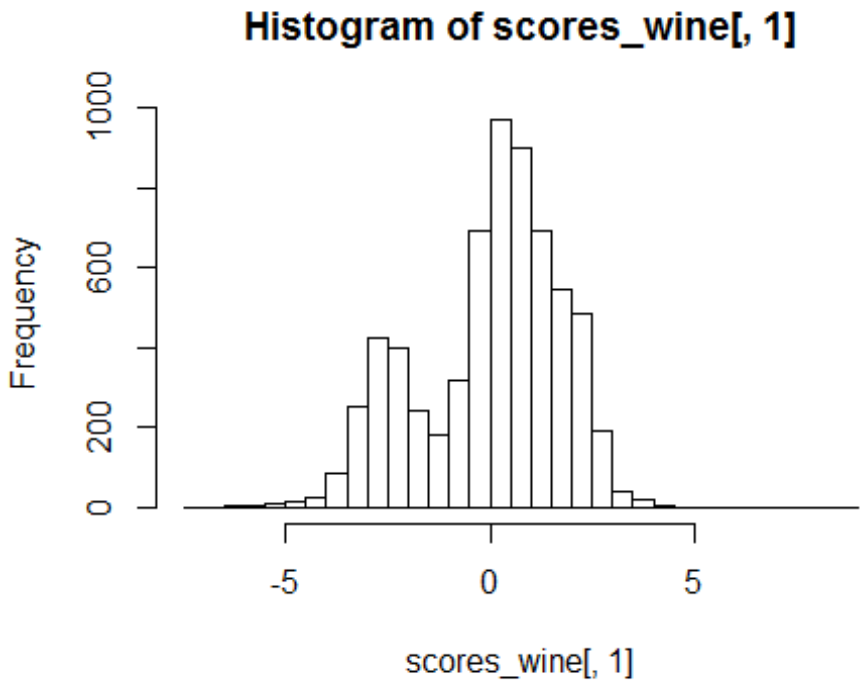
```
##        fixed.acidity     volatile.acidity          citric.acid
##           -0.23879890           -0.38075750            0.15238844
##       residual.sugar            chlorides  free.sulfur.dioxide
##            0.34591993           -0.29011259            0.43091401
## total.sulfur.dioxide              density                    pH
##            0.48741806           -0.04493664           -0.21868644
##            sulphates              alcohol
##           -0.29413517           -0.10643712
```

```
plot(pca_wine)
```



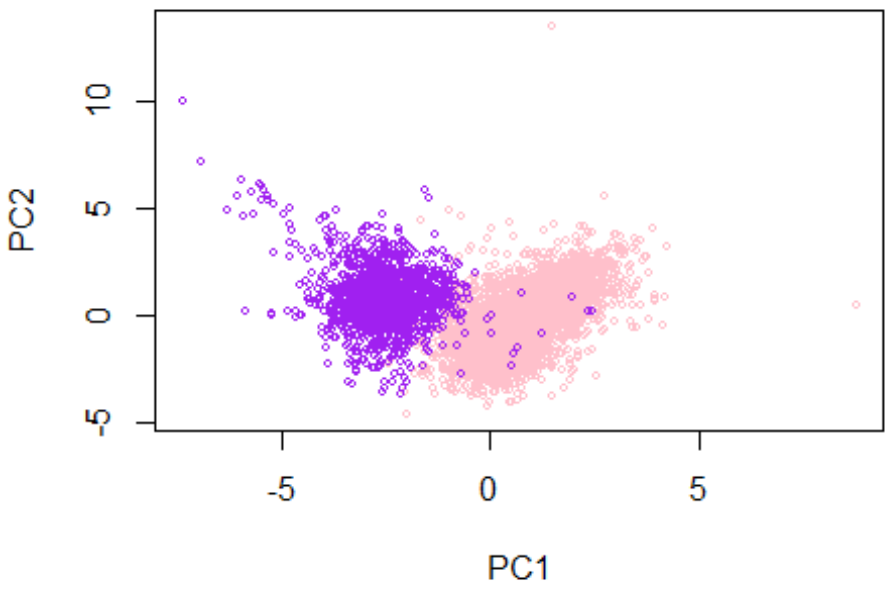
A histogram of the scores on the first PC, notice the bimodal distribution:

```
scores_wine = pca_wine$x
hist(scores_wine[,1], 25)
```



Now, we plot the data points in 2 dimensions, with one axis as our first principal component and the other axis as our second. We can clearly see the separation by red and white wines:

```
plot(scores_wine[,1:2], type='n')
points(scores_wine[wine$color=='white',1:2], col='pink', cex=0.5)
points(scores_wine[wine$color=='red',1:2], col='purple', cex=0.5)
```



K-Means:

Now, we run k-means. We have included the scaling and centering code here again to emphasize that this must be done each time clustering is run because we are relying on a consistent notion of 'distance' in the algorithm.

```
X = scale(wine[,1:11])
mu = attr(X, 'scaled:center')
sigma = attr(X, 'scaled:scale')
cluster_wine = kmeans(X, 2)
```

Now we can tabulate cluster membership by the color of the wine:

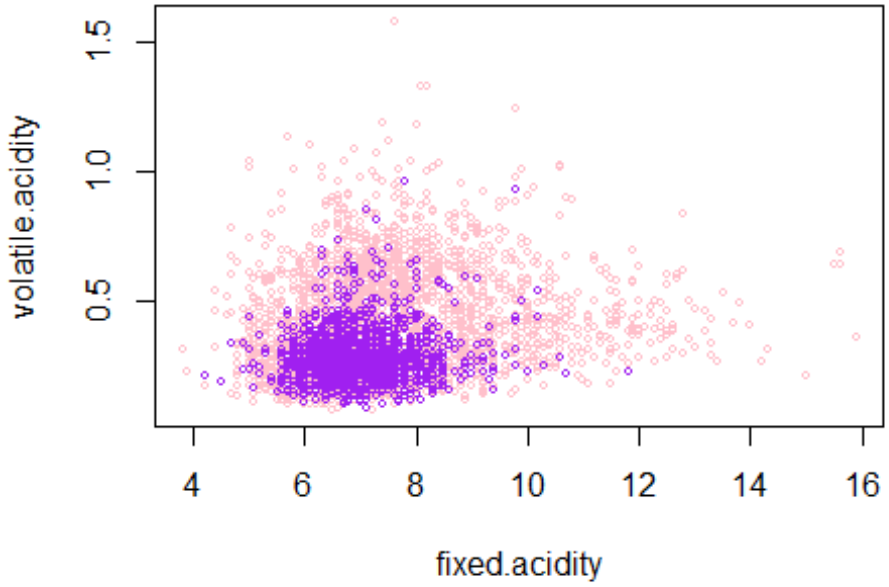
```
xtabs(~wine$color + cluster_wine$cluster)
```

```
##            cluster_wine$cluster
## wine$color    1    2
##      red   1581   18
##      white 2716 2182
```

It looks like cluster 1 is almost all whites, and cluster 2 almost all reds. Now let's try running a few scatterplots of the data, with different variables on the axes. This will inform us of the ability of these variables to explain some of the difference between red and white wines:

Is there any difference between red and white wines in terms of Fixed Acidity versus volatile acidity?

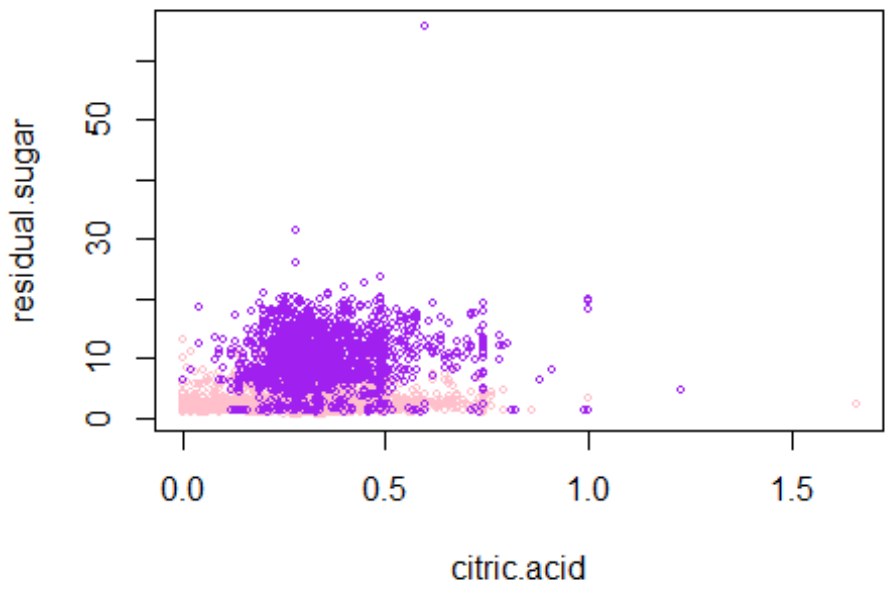
```
plot(wine[,1:2], type='n')
points(wine[cluster_wine$cluster==1,1:2], col='pink', cex=0.5)
points(wine[cluster_wine$cluster==2,1:2], col='purple', cex=0.5)
```



Yes. These two variables appear to tell us something about the differences between red and white wines, although there is significant overlap.

How about Citric acid versus residual sugar?

```
plot(wine[,3:4], type='n')
points(wine[cluster_wine$cluster==1,3:4], col='pink', cex=0.5)
points(wine[cluster_wine$cluster==2,3:4], col='purple', cex=0.5)
```



Here we have a similar story. These two variables appear to tell us a little bit about the differences between red and white wines, but again there is significant overlap.