Initialize the data, and then using the plots shown below make a decision about how many clusters to search for using the k-means algorithm.

> reds=read.csv("winequality-red.csv")

> whites=read.csv("winequality-white.csv")

> min\_itr=1

> max\_itr=30

> # Determine the appropriate number of clusters to use.

>

> n\_clusters\_reds = (nrow(reds)-1)\*sum(apply(reds,2,var))

> n\_clusters\_whites = (nrow(whites)-1)\*sum(apply(whites,2,var))

> for (i in min\_itr:max\_itr) n\_clusters\_reds[i] <- sum(kmeans(reds,centers=i,iter.max =30)$withinss)

> n\_clusters\_reds

[1] 1915121.79 692917.02 396314.53 284696.35 222180.07 194422.50

[7] 178655.53 165132.23 153473.13 133340.56 129228.62 124198.36

[13] 115636.01 110049.71 107950.34 68192.51 69595.94 67898.86

[19] 67101.64 64170.96 53153.37 59649.83 57493.77 54282.46

[25] 57786.41 58596.26 52497.67 56527.27 45446.90 52118.17

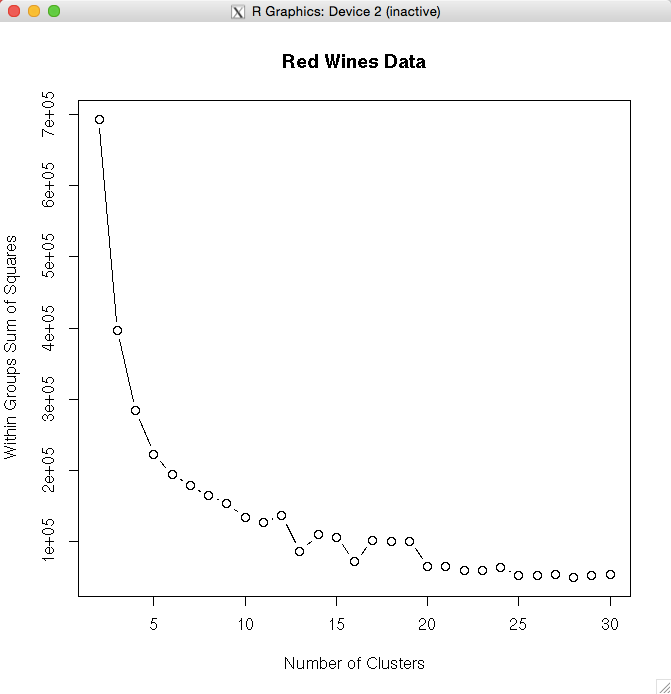
> for (i in min\_itr:max\_itr) n\_clusters\_whites[i] <- sum(kmeans(whites,centers=i,iter.max=30)$withinss)

> lop\_off=2

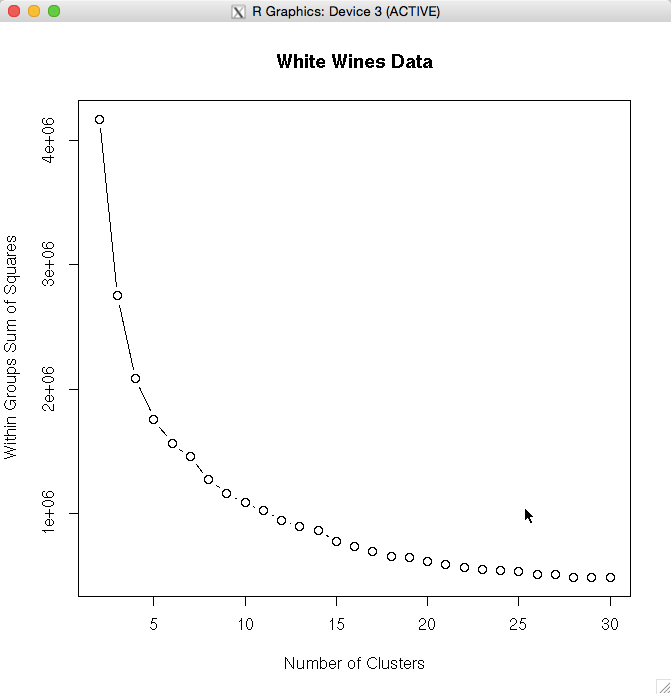
> plot(lop\_off:max\_itr, n\_clusters\_reds[lop\_off:max\_itr], main="Red Wines Data", type="b", xlab="Number of Clusters", ylab="Within Groups Sum of Squares")

> dev.new() # allows for a new plot window.

> plot(lop\_off:max\_itr, n\_clusters\_whites[lop\_off:max\_itr], main="White Wines Data", type="b", xlab="Number of Clusters", ylab="Within Groups Sum of Squares")



Based on this graph, I selected 16 as a reasonable number of clusters. I ran this simulation a number of times, and that dip at 16 is there frequently.



Based on this graph I seemingly arbitrarily decided that 18 is a reasonable number of clusters. Having run the simulation many times it appeared that 18 was

frequently lower than 19, and therefore, not wanting to choose twenty or more, I decided 18 was a good place make the cut-off.

Following this I computed the clusters for the reds and whites:

> k.reds <- readline("Select preferred K for red wines")

> k.whites <- readline("Select preferred K for white wines")

> k.reds <- as.integer(k.reds)

> k.whites <- as.integer(k.whites)

> reds.fit <- kmeans(reds,k.reds, iter.max=30)

> whites.fit <- kmeans(whites,k.whites, iter.max=30)

Below is the cluster data for the red wines:

> reds.fit

K-means clustering with 16 clusters of sizes 48, 102, 74, 206, 60, 126, 2, 80, 39, 153, 100, 87, 200, 151, 41, 130

Cluster means:

**fixed.acidity volatile.acidity citric.acid residual.sugar chlorides**

1 8.127083 0.5689583 0.3285417 3.025000 0.08985417

2 7.692157 0.6122549 0.2287255 2.779412 0.08577451

3 8.786486 0.4775000 0.3516216 2.821622 0.09439189

4 8.510194 0.5294660 0.2841262 2.454612 0.08466990

5 7.966667 0.5610000 0.2803333 3.120000 0.08548333

6 8.773810 0.5291270 0.2853175 2.626190 0.09590476

7 7.900000 0.3000000 0.6800000 8.300000 0.05000000

8 8.053750 0.5797500 0.2873750 2.316250 0.11600000

9 7.876923 0.5223077 0.2797436 4.446154 0.09810256

10 9.200000 0.4954248 0.3132026 2.392157 0.08278431

11 7.850000 0.5020500 0.2212000 2.463500 0.07597000

12 8.024138 0.4907471 0.2713793 2.373563 0.08991954

13 8.793000 0.5062250 0.3035000 2.462750 0.08724500

14 8.252318 0.5302649 0.2435762 2.327483 0.08813245

15 7.719512 0.5163415 0.1882927 2.023171 0.07578049

16 7.403077 0.5458077 0.1750769 2.132308 0.07693077

alcohol quality

1 9.816667 5.125000

2 10.100980 5.441176

3 10.305405 5.554054

4 10.816748 5.684466

5 9.835000 5.133333

6 10.270635 5.579365

7 12.300000 7.000000

8 9.828542 5.275000

9 10.226923 5.461538

10 10.205338 5.725490

11 10.760500 5.930000

12 10.520690 5.712644

13 10.601250 5.800000

14 10.357616 5.688742

15 10.558537 5.634146

16 10.847308 5.792308

free.sulfur.dioxide total.sulfur.dioxide density pH sulphates

1 32.010417 140.02083 0.9970402 3.227083 0.7441667

2 19.568627 89.42157 0.9968629 3.349216 0.6209804

3 29.432432 71.83784 0.9973755 3.300676 0.6714865

4 4.844660 11.82524 0.9963340 3.310146 0.6160194

5 20.166667 111.78333 0.9971158 3.266333 0.6585000

6 13.170635 47.26984 0.9973936 3.289841 0.6643651

7 37.500000 283.50000 0.9931600 3.010000 0.5100000

8 14.337500 66.10000 0.9969004 3.277375 0.6726250

9 43.820513 97.92308 0.9968821 3.280000 0.6166667

10 7.901961 26.51634 0.9976234 3.261765 0.6892810

11 24.810000 42.65000 0.9959376 3.343500 0.6705000

12 24.344828 56.27586 0.9968307 3.349655 0.7067816

13 7.695000 18.73000 0.9967125 3.277150 0.6606000

14 13.072848 36.54967 0.9968950 3.348609 0.6319868

15 38.512195 56.85366 0.9961493 3.405610 0.7282927

16 15.192308 27.36154 0.9955687 3.382000 0.6401538

Considering the red wine cluster data, specifically the clusters with respect to quality, it almost appear these clusters would not be good predictors of quality. The quality means are not strikingly different. Cluster number 7 however looks promising in that it has a much higher average quality rating. Considering the other predictors, cluster seven stands out among the other predictors in alcohol content – it has the highest mean alcohol content, it is also on the high end in both free and total sulfer dioxides, it has the lowest volatile acidity, it has the highest citric acid, it has the highest residual sugar, and it has the lowest total chlorides.

Note however, that group seven only has two wines in the group. Unfortunately that makes any generalizations we might want to make pretty much useless.

Now, considering the white wines cluster data:

Cluster means:

fixed.acidity volatile.acidity citric.acid residual.sugar chlorides

1 6.736628 0.2586919 0.3854070 8.789826 0.05386047

2 6.669198 0.2415616 0.3317192 6.464756 0.04296275

3 6.985556 0.2740694 0.3455556 9.565694 0.04949167

4 6.898592 0.3028169 0.3626056 10.619366 0.05022535

5 7.050000 0.3005540 0.3503693 6.971023 0.04811932

6 7.109390 0.3308685 0.3365728 6.980282 0.05455399

7 6.730310 0.2468974 0.3420286 7.171002 0.04875179

8 7.084615 0.2908654 0.3080769 3.920192 0.03361538

9 6.694603 0.2789048 0.3079048 3.518571 0.03851429

10 7.357143 0.3650000 0.2814286 5.085714 0.05571429

11 7.132468 0.3104870 0.3608442 10.295130 0.05420130

12 6.729003 0.2508308 0.3250151 3.451208 0.03986707

13 6.944444 0.2831790 0.3504012 9.655864 0.04965123

14 6.840095 0.2874463 0.3196659 6.543914 0.04547494

15 6.876856 0.2801965 0.3220961 2.991485 0.04106987

16 6.687156 0.2731422 0.3182798 5.582110 0.04152064

17 6.861398 0.2742857 0.3206687 4.427964 0.04411854

18 6.983051 0.3070508 0.3286780 3.791864 0.04414915

free.sulfur.dioxide total.sulfur.dioxide density pH sulphates

1 69.985465 174.23837 0.9955682 3.179302 0.5277326

2 43.398281 122.52006 0.9933781 3.178797 0.4813754

3 45.465278 170.55278 0.9961701 3.192306 0.4965000

4 68.665493 216.58451 0.9968170 3.152254 0.5074648

5 27.649148 156.29545 0.9948457 3.184915 0.4908239

6 27.253521 184.68075 0.9954800 3.198779 0.5234742

7 51.106205 148.25060 0.9944230 3.194511 0.4844869

8 7.923077 36.32692 0.9920287 3.137885 0.4207692

9 23.287302 84.73333 0.9916113 3.188476 0.4733651

10 123.857143 335.14286 0.9953729 3.275714 0.5757143

11 43.746753 230.73052 0.9971155 3.177013 0.5404545

12 35.341390 100.26586 0.9915151 3.195801 0.4860121

13 50.229938 194.96451 0.9964147 3.178519 0.4961728

14 31.811456 136.72076 0.9940323 3.209928 0.4791169

15 14.242358 68.66376 0.9918294 3.166332 0.4738428

16 27.768349 114.73165 0.9931103 3.205069 0.4790596

17 17.407295 100.68693 0.9929009 3.183070 0.4756535

18 16.738983 126.65763 0.9931482 3.186576 0.5014576

alcohol quality

1 9.886919 5.779070

2 10.791165 6.286533

3 9.820556 5.666667

4 9.449296 5.500000

5 10.243750 5.764205

6 9.968623 5.474178

7 10.311297 6.038186

8 11.090385 5.269231

9 11.407513 6.161905

10 10.214286 4.000000

11 9.577273 5.590909

12 11.500302 6.347432

13 9.613220 5.580247

14 10.597128 5.964200

15 11.170815 5.672489

17 10.816109 5.793313

18 10.681017 5.657627

16 10.917431 6.123853

Again, the unsupervised clustering for White Wines appears to be unhelpful with regard to quality. While some further analysis might bring some key information to light. At this point such information is well obscured.