HW5_q5

Will Scheib

12/3/2021

5

```r
mtcars2 <- mtcars
```

a

```r
head(mtcars2)
```

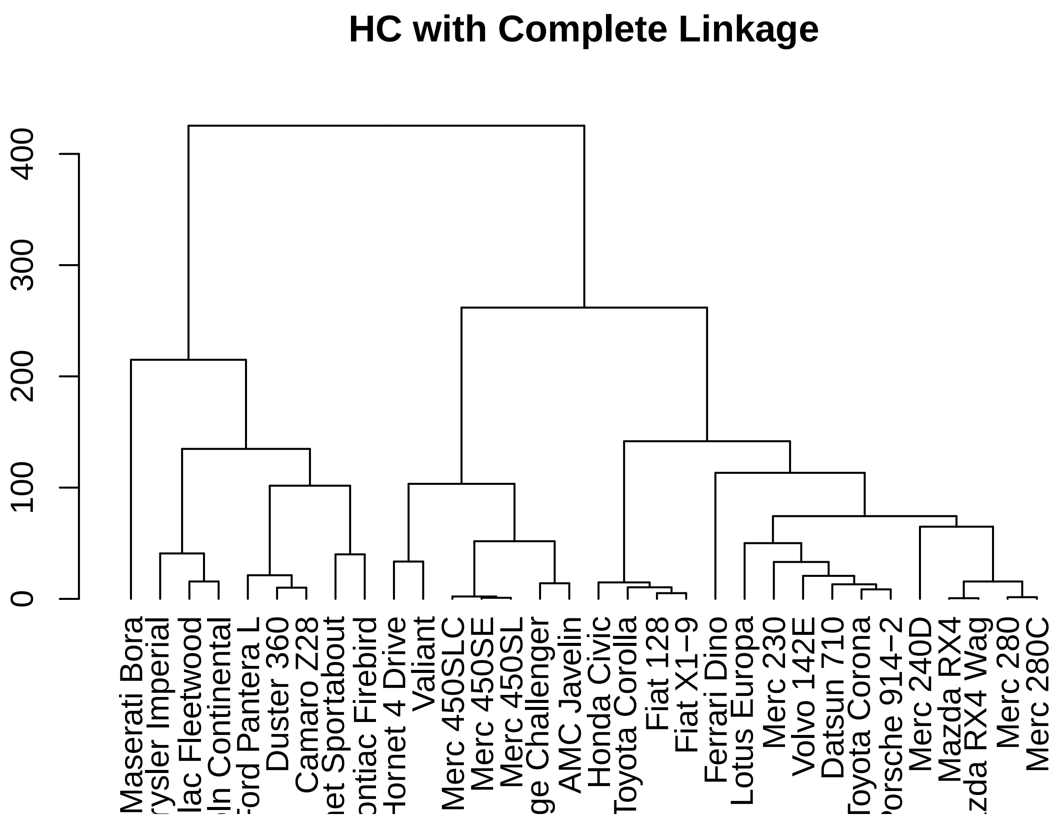
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

We should exclude vs and am because they are categorical, and hierarchical clustering cannot handle categorical variables.

```r
mtcars2 <- mtcars2 %>% select(-c(vs, am))
```

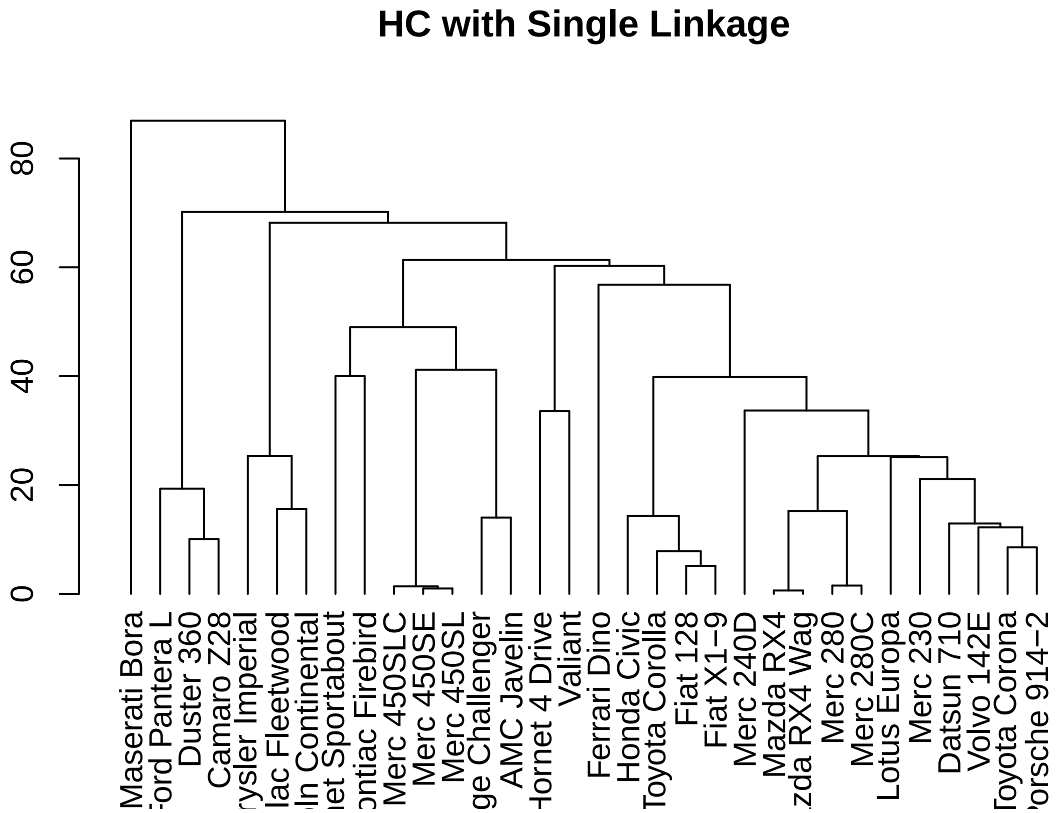
b

```
hc.complete<-hclust(dist(mtcars2), method="complete")

plot(as.dendrogram(hc.complete), main="HC with Complete Linkage")
```



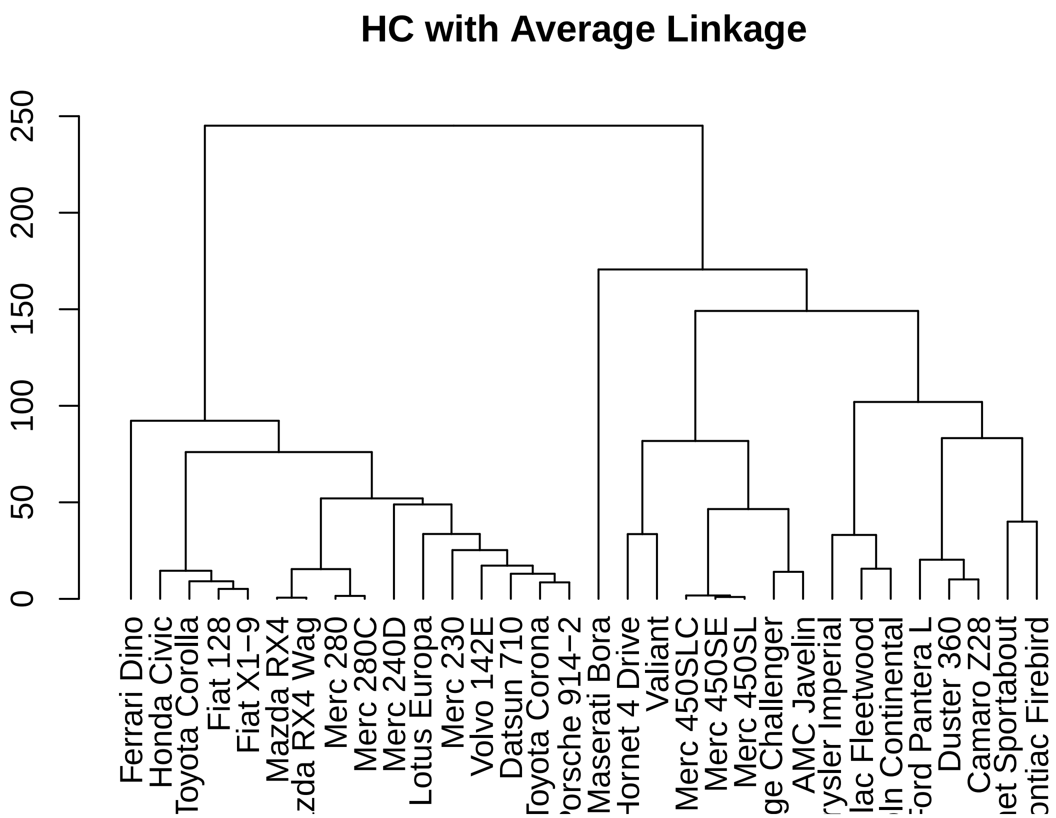
c

```
hc.single<-hclust(dist(mtcars2), method="single")

plot(as.dendrogram(hc.single), main="HC with Single Linkage")
```



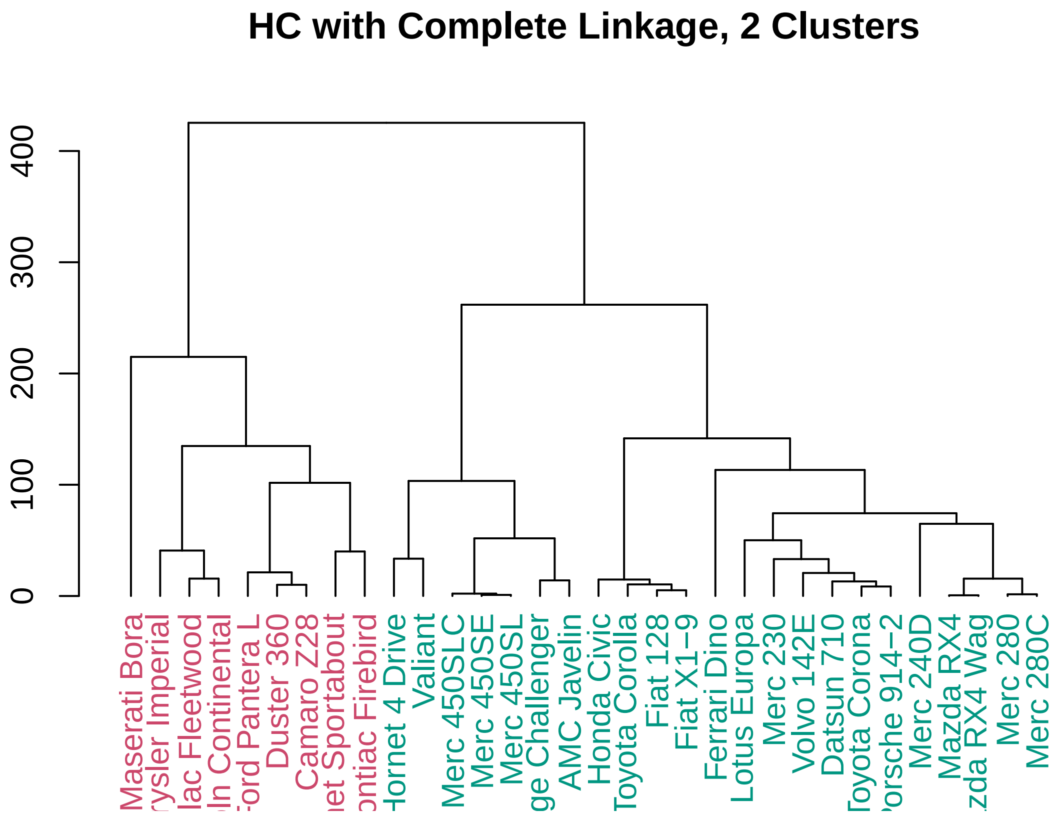
d

```r
hc.average<-hclust(dist(mtcars2), method="average")

plot(as.dendrogram(hc.average), main="HC with Average Linkage")
```

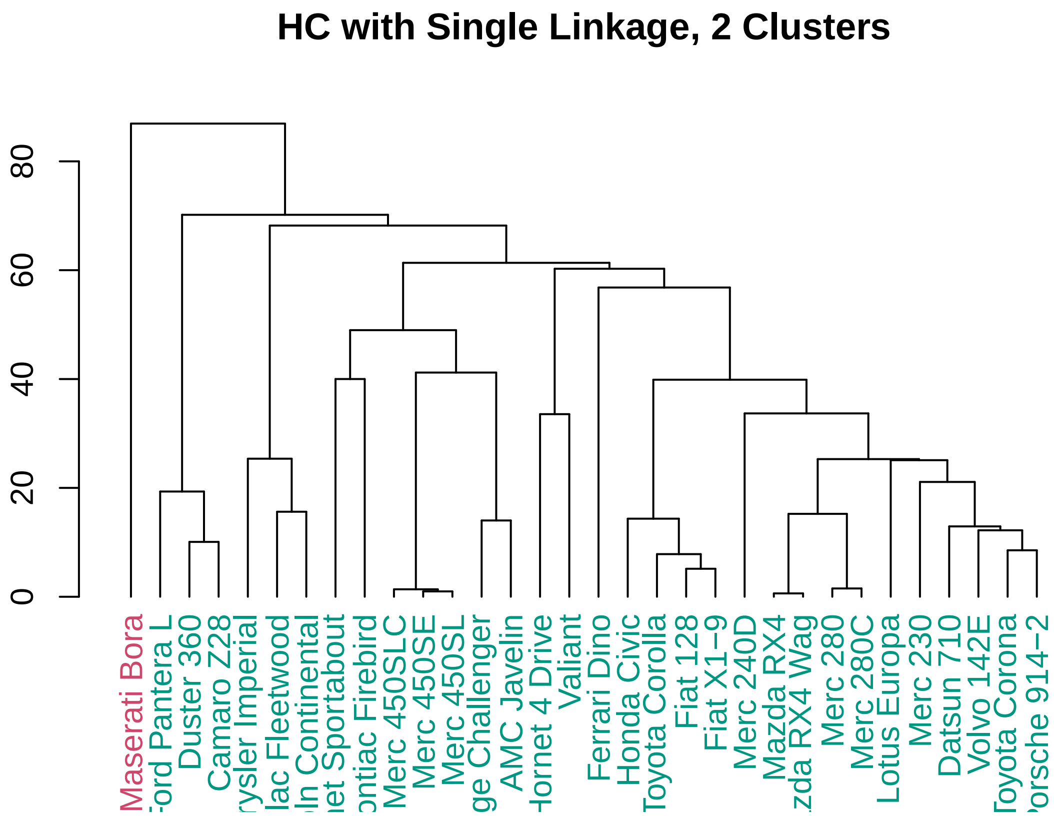


e

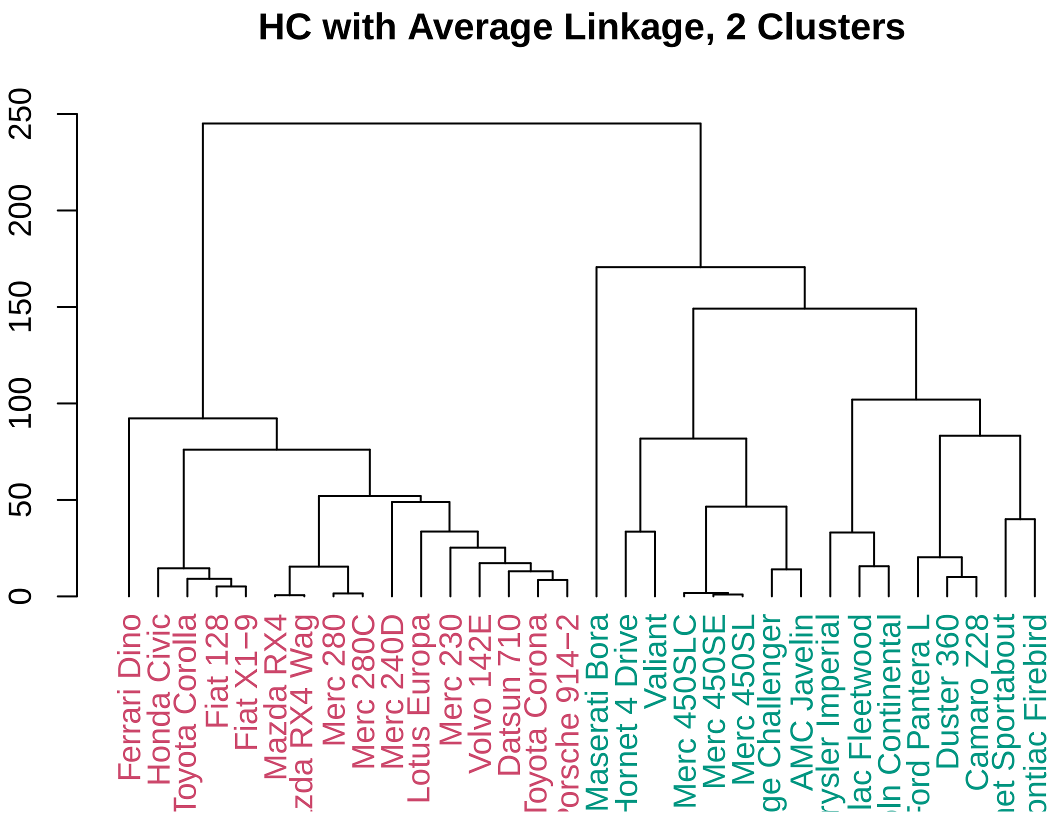
```
dend.complete.col2<-dendextend::color_labels(hc.complete, k=2)
plot(as.dendrogram(dend.complete.col2), main="HC with Complete Linkage, 2 Clusters")
```



```
dend.single.col2<-dendextend::color_labels(hc.single, k=2)
plot(as.dendrogram(dend.single.col2), main="HC with Single Linkage, 2 Clusters")
```



```
dend.average.col2<-dendextend::color_labels(hc.average, k=2)
plot(as.dendrogram(dend.average.col2), main="HC with Average Linkage, 2 Clusters")
```

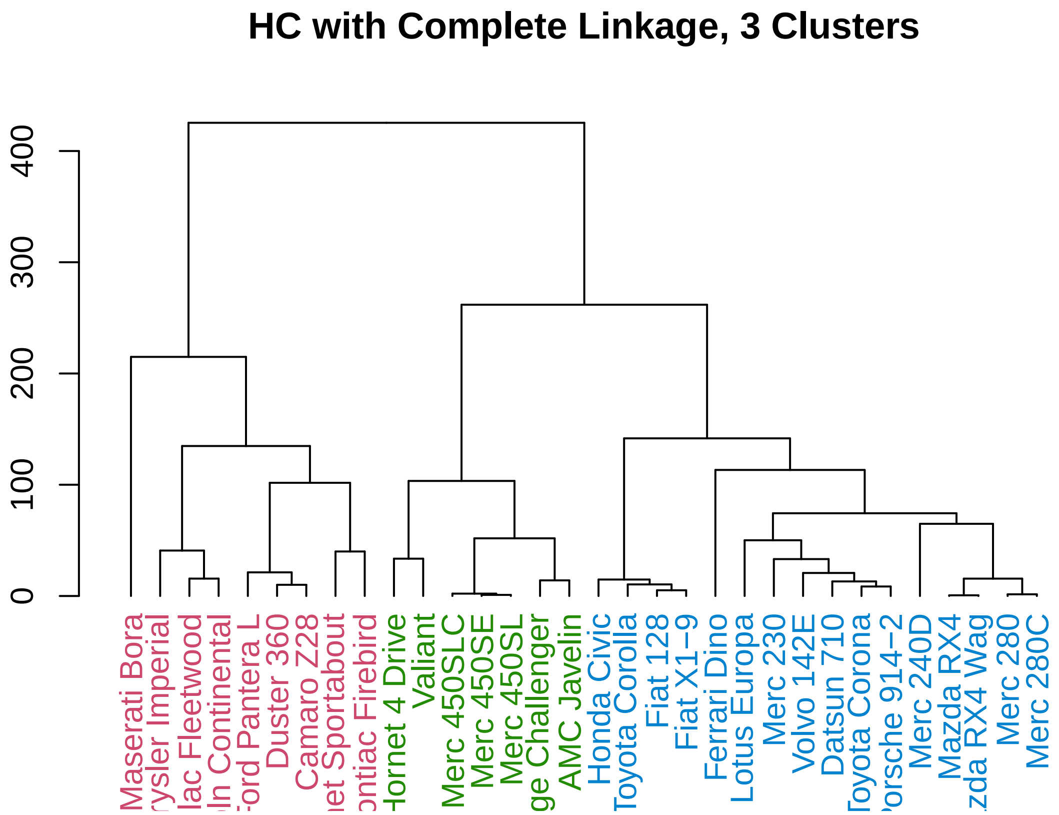


```
list(
  complete=table(cutree(hc.complete,2)),
  single=table(cutree(hc.single,2)),
  average=table(cutree(hc.average,2))
)
```

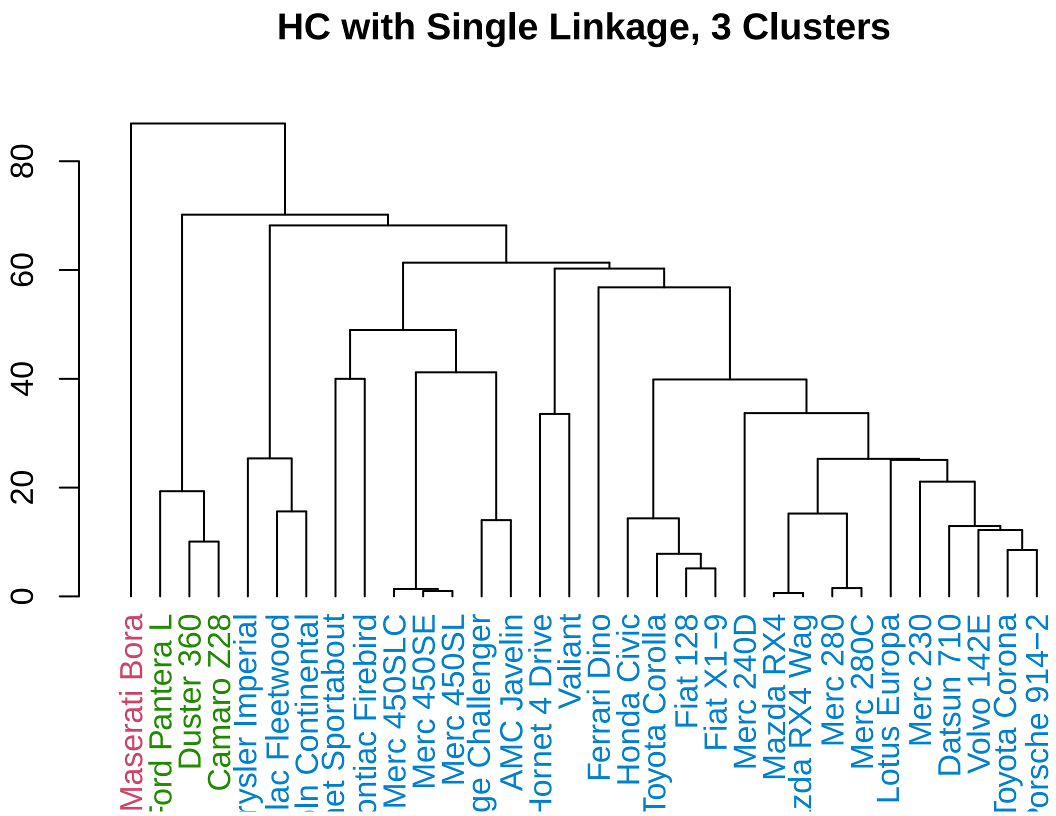
```
## $complete
##
##  1  2
## 23  9
##
## $single
##
##  1  2
## 31  1
##
## $average
##
##  1  2
## 16 16
```

f

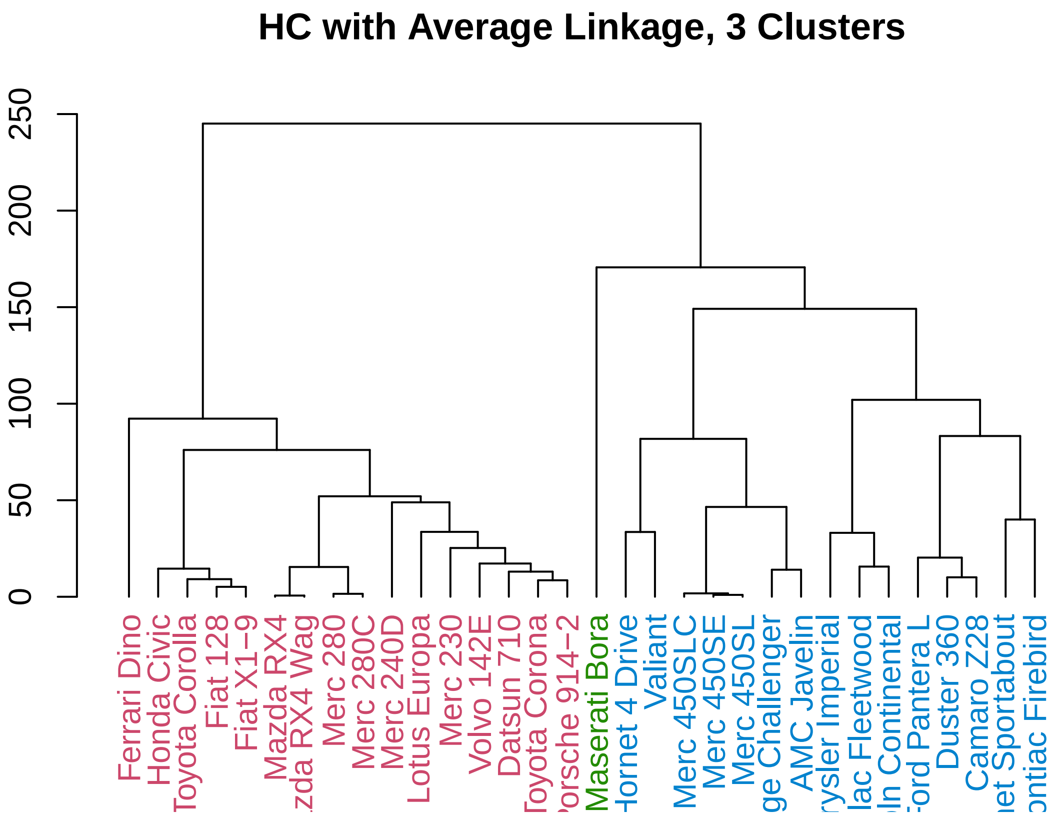
```
dend.complete.col3<-dendextend::color_labels(hc.complete, k=3)
plot(as.dendrogram(dend.complete.col3), main="HC with Complete Linkage, 3 Clusters")
```



```
dend.single.col3<-dendextend::color_labels(hc.single, k=3)
plot(as.dendrogram(dend.single.col3), main="HC with Single Linkage, 3 Clusters")
```



```
dend.average.col3<-dendextend::color_labels(hc.average, k=3)
plot(as.dendrogram(dend.average.col3), main="HC with Average Linkage, 3 Clusters")
```

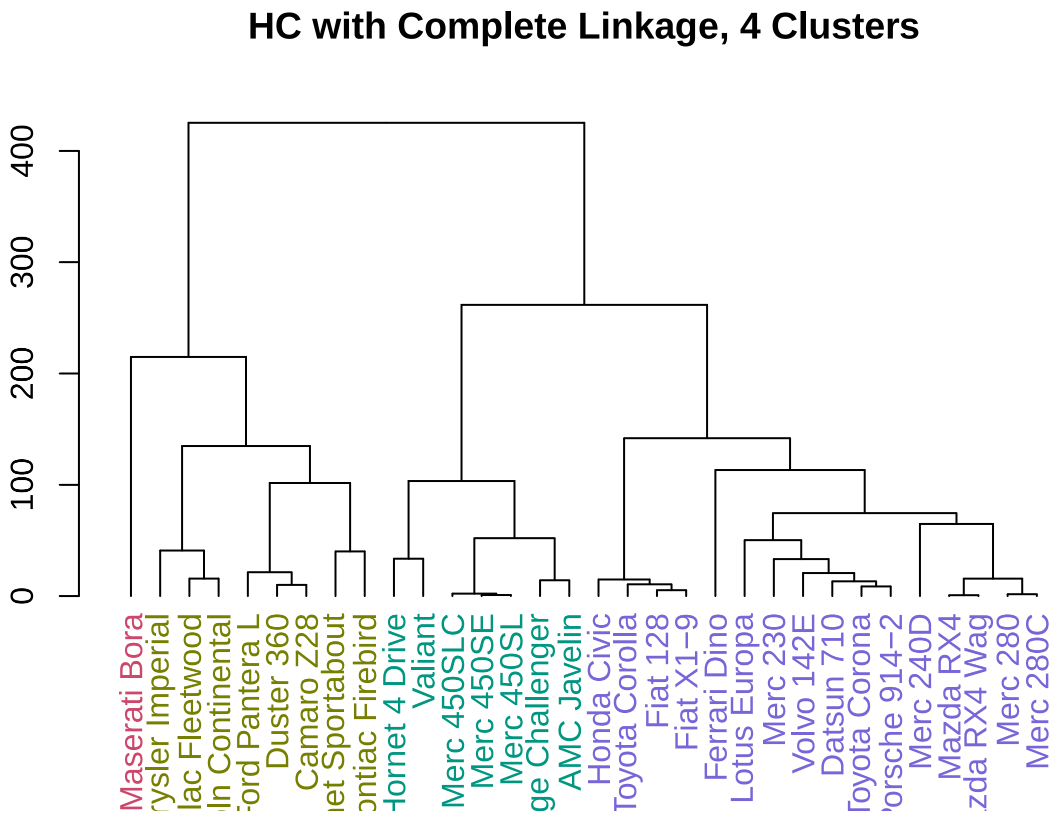


```
list(
  complete=table(cutree(hc.complete,3)),
  single=table(cutree(hc.single,3)),
  average=table(cutree(hc.average,3))
)
```

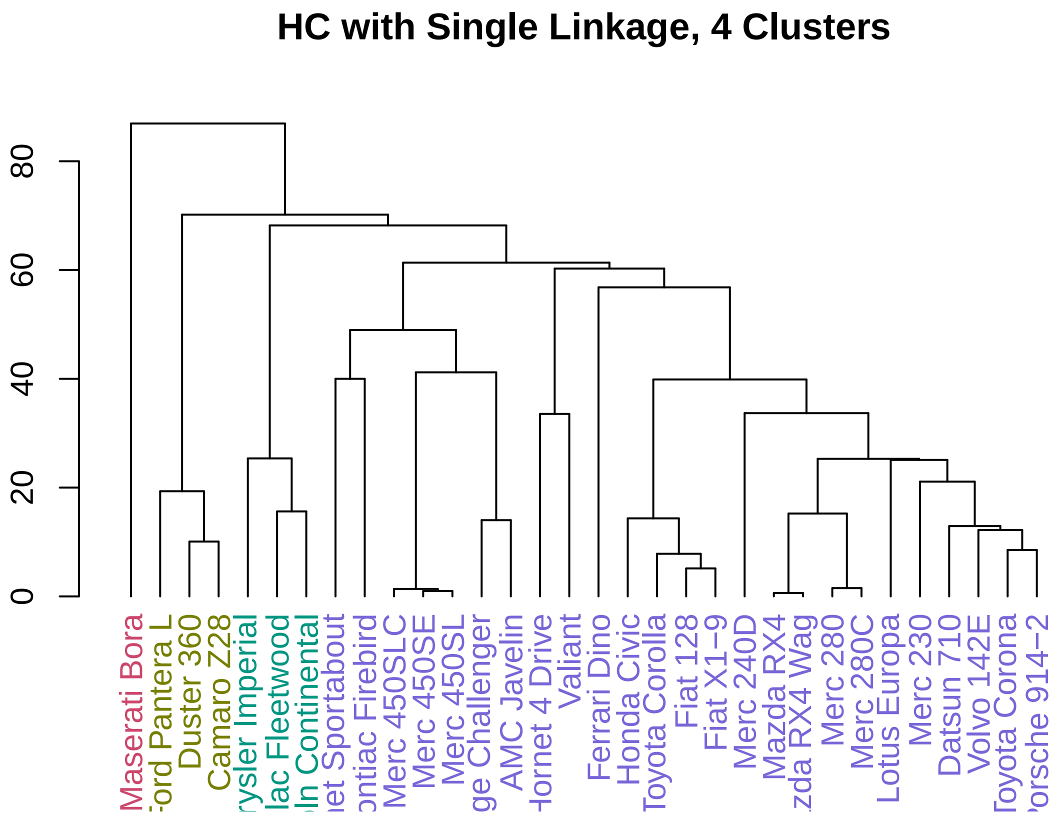
```
## $complete
##
##  1  2  3
## 16  7  9
##
## $single
##
##  1  2  3
## 28  3  1
##
## $average
##
##  1  2  3
## 16 15  1
```

g

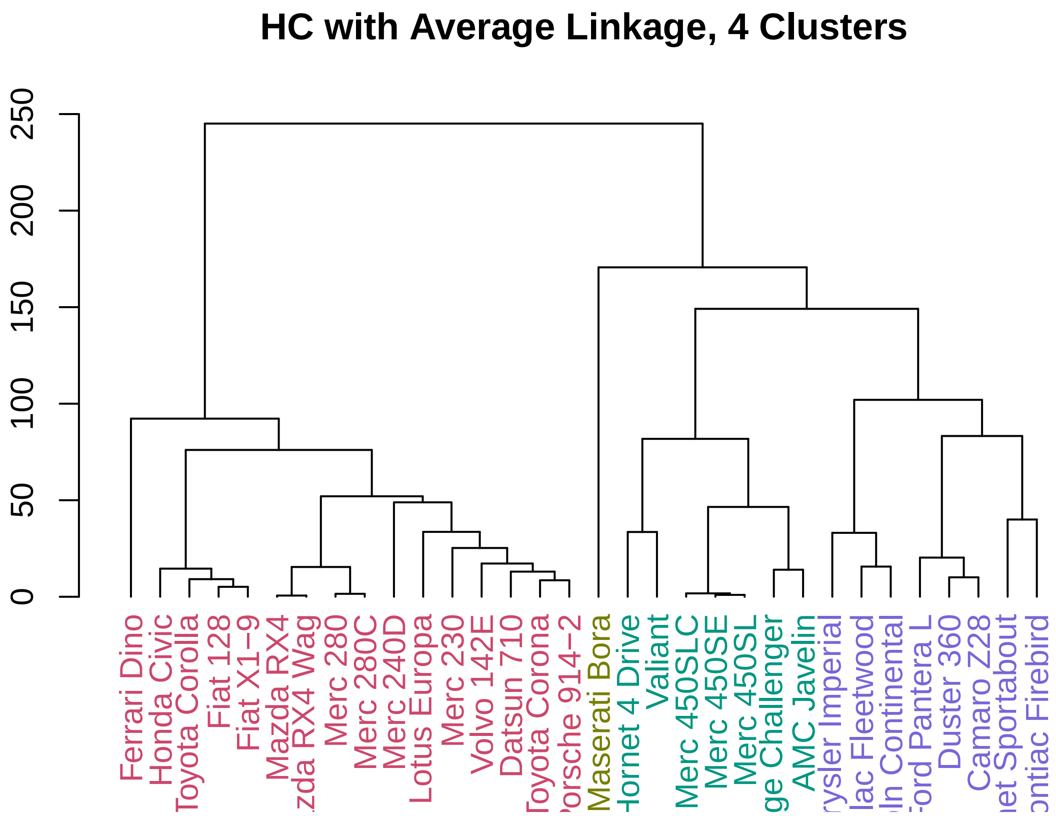
```r
dend.complete.col4<-dendextend::color_labels(hc.complete, k=4)
plot(as.dendrogram(dend.complete.col4), main="HC with Complete Linkage, 4 Clusters")
```



```
dend.single.col4<-dendextend::color_labels(hc.single, k=4)
plot(as.dendrogram(dend.single.col4), main="HC with Single Linkage, 4 Clusters")
```



```
dend.average.col4<-dendextend::color_labels(hc.average, k=4)
plot(as.dendrogram(dend.average.col4), main="HC with Average Linkage, 4 Clusters")
```



```
list(
  complete=table(cutree(hc.complete,4)),
  single=table(cutree(hc.single,4)),
  average=table(cutree(hc.average,4))
)
```

```
## $complete
##
##  1  2  3  4
## 16  7  8  1
##
## $single
##
##  1  2  3  4
## 25  3  3  1
##
## $average
##
##  1  2  3  4
## 16  7  8  1
```

h

(i)

```
table(cutree(hc.complete,3))
```

```
##
##  1  2  3
## 16  7  9
```

(ii)

```
grps.com<-cutree(hc.complete,3)
x<-data.frame(mtcars2,grps.com)

aggregate(x[,c(1,3,6)],by=list(x$grps), mean)
```

```
##   Group.1      mpg     disp       wt
## 1       1 24.50000 122.2938 2.518000
## 2       2 17.01429 276.0571 3.601429
## 3       3 14.64444 388.2222 4.161556
```

(iii)

Cluster 1 has high miles per gallon, low displacement, and low weight.
Cluster 2 has middle miles per gallon, middle displacement, and middle weight.
Cluster 3 has low miles per gallon, high displacement, and high weight.

i

(i)

```
table(cutree(hc.single,3))
```

```
## 
##  1  2  3 
## 28  3  1
```

```


grps.com<-cutree(hc.single,3)
x<-data.frame(mtcars2,grps.com)

aggregate(x[,c(1,3,6)],by=list(x$grps), mean)
```

(ii)

```
##   Group.1      mpg     disp       wt
## 1       1 20.87500 215.0393 3.171500
## 2       2 14.46667 353.6667 3.526667
## 3       3 15.00000 301.0000 3.570000
```

(iii)

Cluster 1 has high miles per gallon, low displacement, and low weight.
Cluster 2 has low miles per gallon, high displacement, and high weight.
Cluster 3 has low miles per gallon, middle displacement, and high weight.

j

(i)

```
table(cutree(hc.average,3))
```

```
##
##  1  2  3
## 16 15  1
```

(ii)

```
grps.com<-cutree(hc.average,3)
x<-data.frame(mtcars2,grps.com)

aggregate(x[,c(1,3,6)],by=list(x$grps), mean)
```

```
##   Group.1      mpg     disp     wt
## 1       1 24.50000 122.2938 2.5180
## 2       2 15.72667 341.6933 3.9396
## 3       3 15.00000 301.0000 3.5700
```

(iii)

Cluster 1 has high miles per gallon, low displacement, and low weight.
Cluster 2 has low miles per gallon, high displacement, and high weight.
Cluster 3 has low miles per gallon, middle/high displacement, and middle/high weight.

k

Our clusters using complete linkage resulted in larger ranges, which is not what we expected because in large data sets, complete linkage tends to result in clusters that are closer together (smaller ranges in means).

In situations where we want more distinct clusters, we should not choose average linkage because it resulted in clusters with similar means, thus making it hard to classify their distinct characteristics.