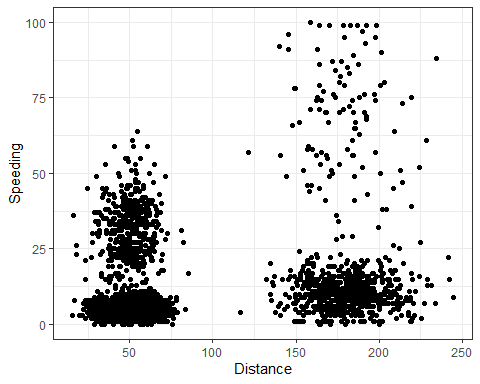
###BAN502 - MOD6 - Assignment 1 ###Unsupervised Learning Assignment ###Schaeffer, Sarah

trucks = read\_csv("trucks.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## Driver\_ID = col\_double(),  
## Distance = col\_double(),  
## Speeding = col\_double()  
## )

###Task 1 - plot of relationship between Distance and Speeding

ggplot(trucks, aes(x = Distance, y = Speeding)) + geom\_point() + theme\_bw()



Based on the plot of distance and speeding there does appear to be some natural clustering. There is clustering where the distance is between 25 - 75 miles and the speeding percentage between 0 - 12.5. There also appears to be some additional clustering around the distance 150 - 200 however the speeding relationship is less defined.

###Task 2 - scaled and centered variables

kmeans\_recipe = recipe(~ Distance + Speeding, trucks)   
  
trucks\_dummy = kmeans\_recipe %>%   
 step\_scale(all\_numeric()) %>%  
 step\_center(all\_numeric())   
  
trucks\_dummy = prep(trucks\_dummy, trucks) #prepares the recipe  
  
trucks\_cleaned = bake(trucks\_dummy, trucks) #applies the recipe and yields a data frame

summary(trucks\_cleaned)

## Distance Speeding   
## Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 3.1560 Max. : 6.5127

summary(trucks)

## Driver\_ID Distance Speeding   
## Min. :3.423e+09 Min. : 15.52 Min. : 0.00   
## 1st Qu.:3.423e+09 1st Qu.: 45.25 1st Qu.: 4.00   
## Median :3.423e+09 Median : 53.33 Median : 6.00   
## Mean :3.423e+09 Mean : 76.04 Mean : 10.72   
## 3rd Qu.:3.423e+09 3rd Qu.: 65.63 3rd Qu.: 9.00   
## Max. :3.423e+09 Max. :244.79 Max. :100.00

###Task 3

set.seed(64)  
clusters = kmeans(trucks\_cleaned, 2)

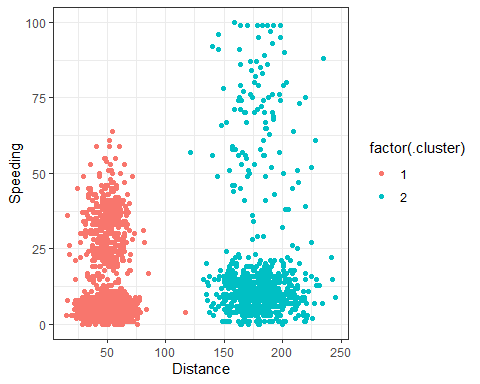
Attach cluster to dataset

trucks = augment(clusters, trucks)  
str(trucks)

## tibble [4,000 x 4] (S3: tbl\_df/tbl/data.frame)  
## $ Driver\_ID: num [1:4000] 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ Distance : num [1:4000] 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding : num [1:4000] 28 25 27 22 25 10 20 8 34 19 ...  
## $ .cluster : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Driver\_ID = col\_double(),  
## .. Distance = col\_double(),  
## .. Speeding = col\_double()  
## .. )

Plots

ggplot(trucks, aes(x=Distance,y=Speeding,color=factor(.cluster))) + geom\_point() + theme\_bw()



The clusters created using k-Means with k=2 resemble the natural clusters that we observed in Task 1. The clusters are tied to distance only and speeding does not have an influence on the clusters.

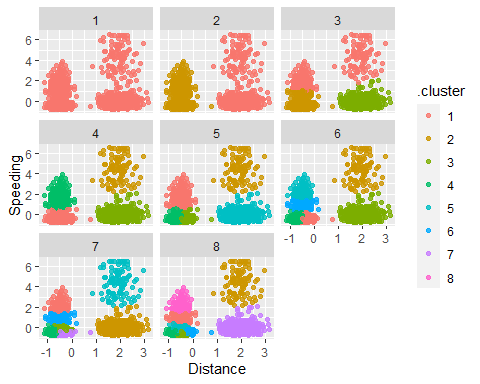
###Task 4

set.seed(412)  
clusts =   
 tibble(k = 1:8) %>%  
 mutate(  
 kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

## # A tibble: 8 x 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 2 2 <kmeans> <tibble [2 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 3 3 <kmeans> <tibble [3 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 4 4 <kmeans> <tibble [4 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 5 5 <kmeans> <tibble [5 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 6 6 <kmeans> <tibble [6 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 7 7 <kmeans> <tibble [7 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 8 8 <kmeans> <tibble [8 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>

clusters =   
 clusts %>%  
 unnest(cols = c(tidied))  
  
assignments =   
 clusts %>%   
 unnest(cols = c(augmented))  
  
clusterings =   
 clusts %>%  
 unnest(cols = c(glanced))

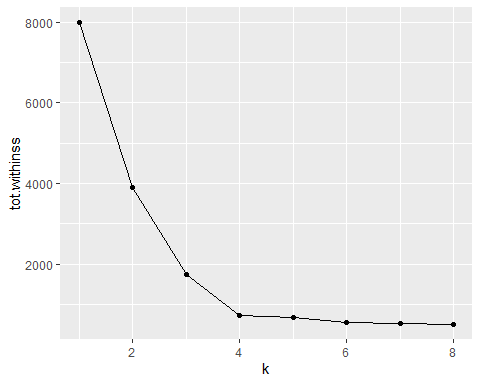
p1 =   
 ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p1



Based on the graphic representation, 4 clusters (k=4) appears to be the most appropriate for this data set as it takes both distance and speeding into account.

###Task 5

ggplot(clusterings, aes(k, tot.withinss)) +  
 geom\_line() +  
 geom\_point()



Based on the elbow chart, K=4 appears to be the best for this data set.

###Task 6

set.seed(64)  
clusters = kmeans(trucks\_cleaned, 4)

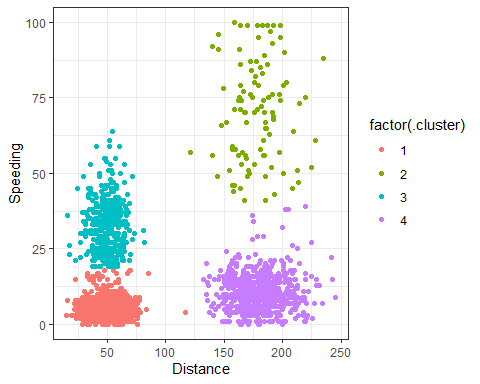
Attach cluster to dataset

trucks = augment(clusters, trucks)  
str(trucks)

## tibble [4,000 x 4] (S3: tbl\_df/tbl/data.frame)  
## $ Driver\_ID: num [1:4000] 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ Distance : num [1:4000] 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding : num [1:4000] 28 25 27 22 25 10 20 8 34 19 ...  
## $ .cluster : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 1 3 1 3 3 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Driver\_ID = col\_double(),  
## .. Distance = col\_double(),  
## .. Speeding = col\_double()  
## .. )

Plots

ggplot(trucks, aes(x=Distance,y=Speeding,color=factor(.cluster))) + geom\_point() + theme\_bw()



The 4 clusters in this final tasks show relationships between both distance and speeding. 2 clusters have a shorter distance with variations in percentage of speeding. The 2 remaining clusters have a longer distance with variations in percentage of speeding. There does appear to be one observation close to 125 miles and 5% of speeding that classified into cluster where it might be more appropriate for it to be part of cluster 4.