# Clustering

## Module 6 - Assignment 1

### Nicole Westrick

str(trucks)

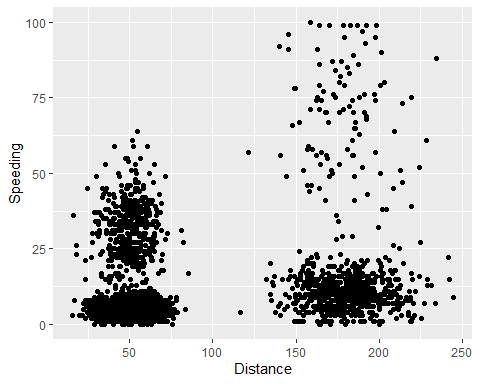
## tibble [4,000 x 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Driver\_ID: num [1:4000] 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ Distance : num [1:4000] 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding : num [1:4000] 28 25 27 22 25 10 20 8 34 19 ...  
## - attr(\*, "spec")=  
## ..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

summary(trucks)

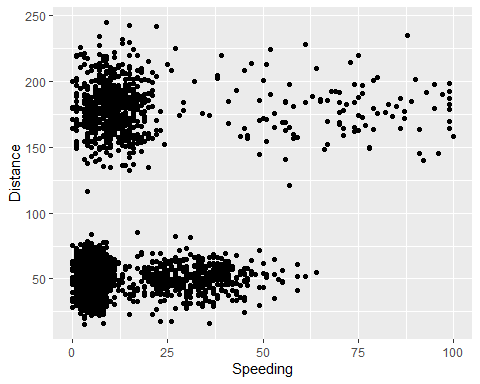
## Driver\_ID Distance Speeding   
## Min. :3.423e+09 Min. : 15.52 Min. : 0.00   
## 1st Qu.:3.423e+09 1st Qu.: 45.25 1st Qu.: 4.00   
## Median :3.423e+09 Median : 53.33 Median : 6.00   
## Mean :3.423e+09 Mean : 76.04 Mean : 10.72   
## 3rd Qu.:3.423e+09 3rd Qu.: 65.63 3rd Qu.: 9.00   
## Max. :3.423e+09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

#### Task 1:

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



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



For distance, there appears to be natural clustering between 25-75, and between 150-200. For speeding, there appears to be natural clustering between 0-25.

#### Task 2:

trucks\_recipe = recipe(~ Distance + Speeding, trucks)  
  
trucks\_cleaned = trucks\_recipe %>%   
 step\_scale(all\_numeric()) %>%  
 step\_center(all\_numeric())   
  
trucks\_cleaned = prep(trucks\_cleaned, trucks)  
  
trucks\_cleaned = bake(trucks\_cleaned, trucks)  
  
summary(trucks\_cleaned)

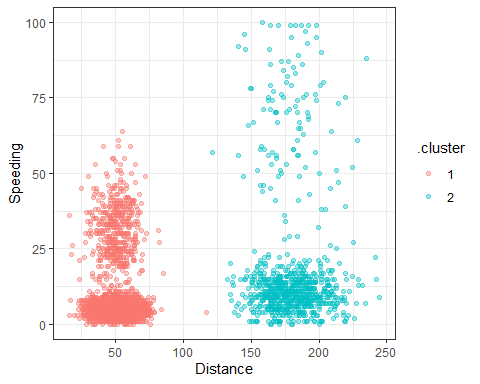
## Distance Speeding   
## Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 3.1560 Max. : 6.5127

#### Task 3:

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

trucks = augment(kclust, trucks)

ggplot(trucks, aes(Distance, Speeding, color = .cluster)) +  
 geom\_point(alpha = 0.4) + theme\_bw()



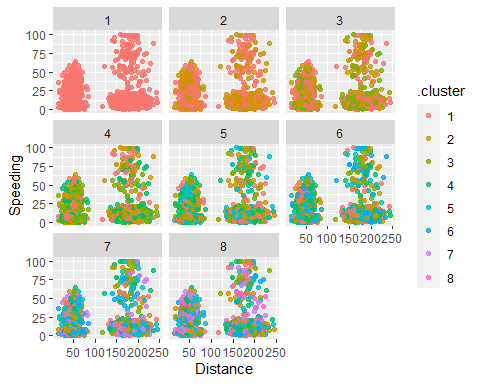
Here, we can see two clusters identified by different colors. Each cluster is more tightly clustered at lower speeding values.

#### Task 4:

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

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

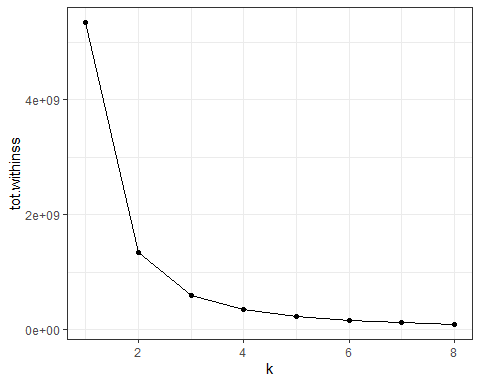
ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)



It is hard to tell which k-value, or number of clusters, would be optimal for this data. An elbow plot would be more helpful in making a determination.

#### Task 5:

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



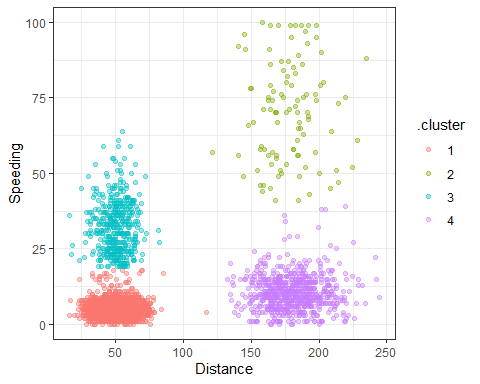
A k-value of 4 appears to be most appropriate for this data.

#### Task 6:

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

trucks = augment(kclust, trucks)

ggplot(trucks, aes(Distance, Speeding, color = .cluster)) +  
 geom\_point(alpha = 0.4) + theme\_bw()



Here, we can clearly see the 4 distinct clusters. Cluster 2 has the most spread, while the others are relatively tightly clustered.