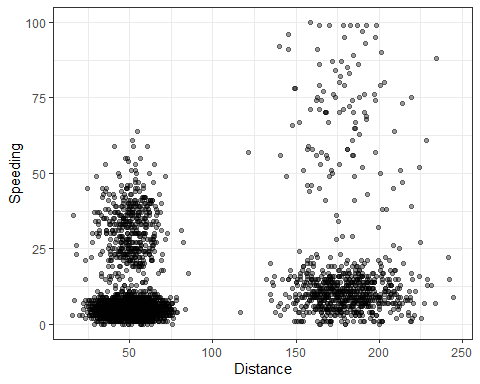
# BAN 502 - Clustering Assignment

## Khayrayyah Haamid-Day

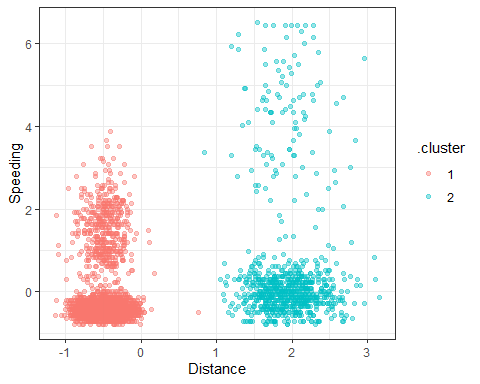
ggplot(trucks, aes(Distance, Speeding)) +  
 geom\_point(alpha = 0.4) + theme\_bw()

 *Task 1*

It does appear that there is natural clustering happening. The first cluster has a stronger concentration of data around speeding of 12. The second cluster is more spread apart with data points extending from 0 to 100 speeding, though there is a strong concentration below 25. Distance for cluster 1 is within a range of more than 0 and less than 75. While the distance for cluster 2 is mainly focused around 150 to 225.

#Task 2  
  
trucks\_cleaned = recipe(~ Distance + Speeding, trucks)  
  
trucks\_dummy = trucks\_cleaned %>%   
 step\_scale(all\_numeric()) %>%  
 step\_center(all\_numeric())  
  
trucks\_dummy = prep(trucks\_dummy, trucks)  
trucks\_cleaned = bake(trucks\_dummy, trucks)  
  
#summary(trucks\_dummy)

#Task 3  
set.seed(64)  
kclust = kmeans(trucks\_cleaned, centers = 2)  
#kclust  
trucks\_cleaned = augment(kclust, trucks\_cleaned)  
#trucks\_cleaned  
  
ggplot(trucks\_cleaned, aes(Distance, Speeding, color = .cluster)) +  
 geom\_point(alpha = 0.4) + theme\_bw()

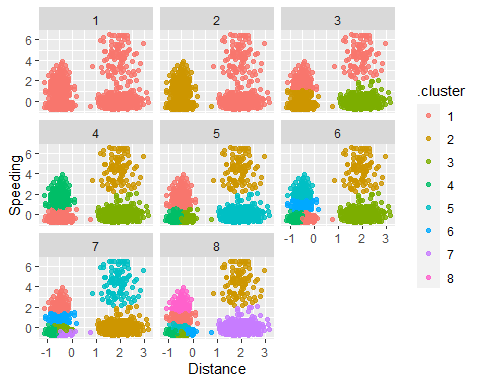


The clusters look the same as the graph done before.

#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

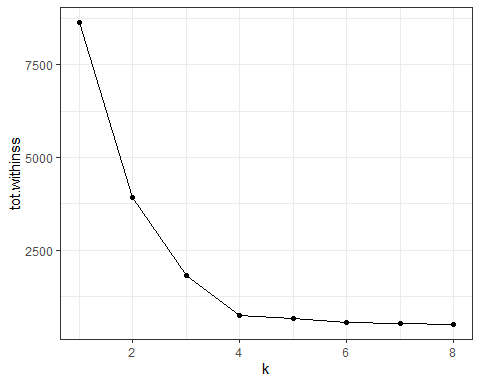
## # A tibble: 8 x 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 x 6]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 2 2 <kmeans> <tibble [2 x 6]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 3 3 <kmeans> <tibble [3 x 6]> <tibble [1 x 4]> <tibble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



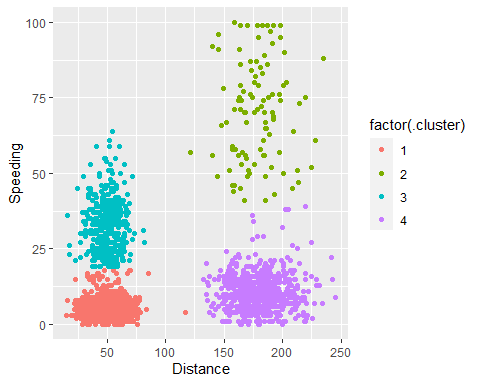
A value of 4 looks to be the most appropriate for this data.

#Task 5   
ggplot(clusterings, aes(k, tot.withinss)) +  
 geom\_line() +  
 geom\_point() + theme\_bw()



It appears 4 is the best.

#Task 6  
set.seed(64)  
clusters = kmeans(trucks\_cleaned, centers = 4)  
#clusters  
trucks = augment(clusters, trucks)  
  
ggplot(trucks, aes(x=Distance,y=Speeding,color=factor(.cluster))) + geom\_point()

 Using 4 clusters looks visually good and separated in a way that makes sense. 1 and 4 are clustered for lower speed while 2 and 3 are clustered with higher speeds. 2 and 4 are separated from 1 and 3 to account for larger distance differences.