Clustering Assignment

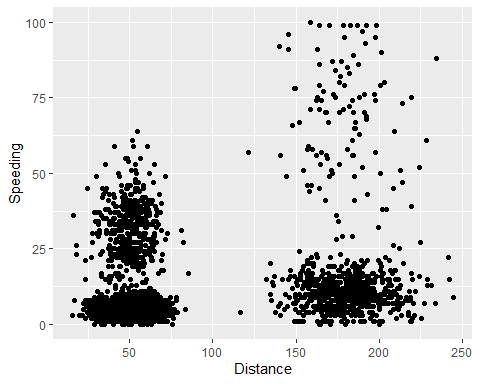
Engel, Alec

2/27/2021

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

#summary(trucks)  
#str(trucks)

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



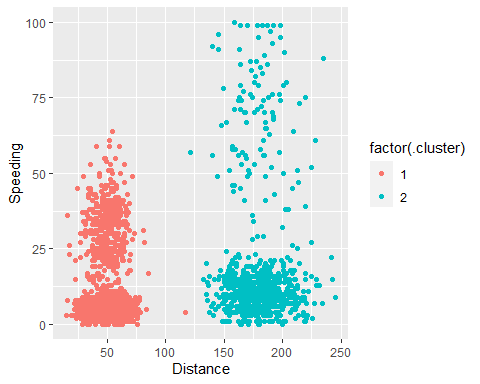
**Yes, there appears to be a natural clustering of drivers. Drivers traveling less than 75 miles in a day look to be less likely to speed. We can see that with these drivers, a majority speed less than 10% of the time while the remaining may speed from between 10% and no more than 65% of the time. With the drivers traveling more than 125 miles a day, we see a majority may speed up to 25% of the time while the remaining speed between 25% and 100% of the time. With the naked eye I can pick out 4 separate clusters.**

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

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

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

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

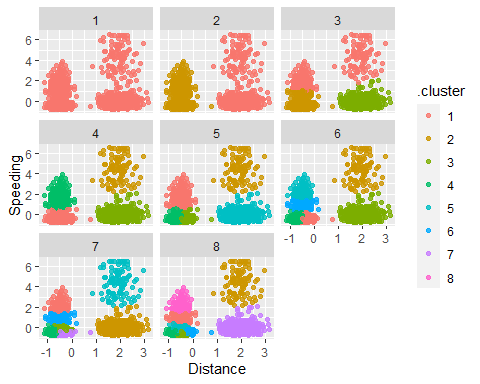


**Since we only divided the data into 2 clusters here, the most similar points per cluster seems to be distance traveled per day. We can see drivers clustered by less than 120 miles per day and drivers who drove more than 120 miles per day. So the speeding percentage was only really clustered with weather a driver traveled a long or short distance.**

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

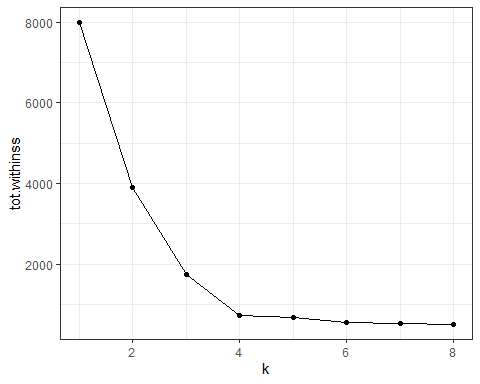
## # A tibble: 8 x 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 2 2 <kmeans> <tibble [2 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 3 3 <kmeans> <tibble [3 x 5]> <tibble [1 x 4]> <tibble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



**I believe a k value of 4 appears to be the most appropriate for this data.**

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

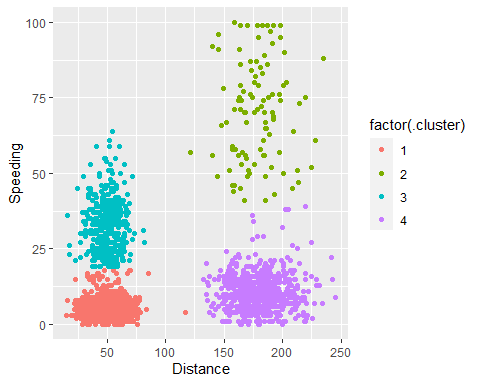


**With this plot of k versus, I am still seeing a k value to 4 to be the elbow and appears to be the best value to chose.**

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

trucks = augment(trucks\_clust, trucks)  
#str(trucks)

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



**These clusters make sense and align for the most part with what I was trying to visualize on my first naked comparison plot between Distance and Speeding. The difference here versus what I originally had observed is the few drivers that are clustered with lower speeding percentiles instead of the higher speeding clusters. I can now observe a cutoff with shorter distance drivers at say roughly 18% speeding vs not and with longer distance drivers that cluster cutoff is closer to 37% speeding vs not. These findings will allow me to send warnings, discipline or possibly implement additional trainings for drivers that fall in the upper speeding percentile clusters of short and long distances**