## Dana Miller

### Clustering Assignment

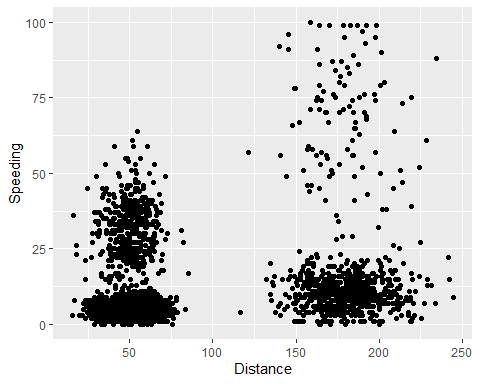
trucks = read\_csv("trucks.csv")

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

#summary(trucks)

**Task 1**

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



**There seems to be 2 distinct clusters just by looking at the graph above. The clusters seems to be based off distances of 0-100 and the second 125+.**

**Task 2**

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

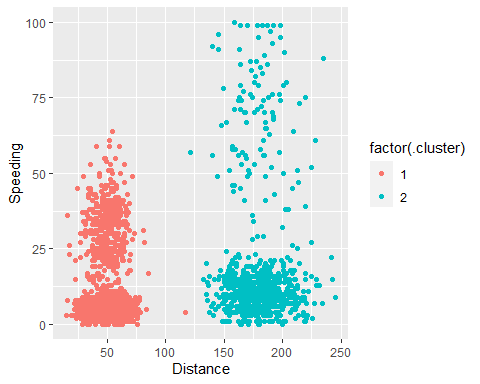
**Task 3**

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

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()  
## .. )

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



**There are 2 clusters that represent drivers who on average drive less than 125 miles a day and drivers who on average drive more than 125 miles a day. The clusters are heavily focused at below 15% of speeding. Therefore, the average driver is speeding 15% of the time.**

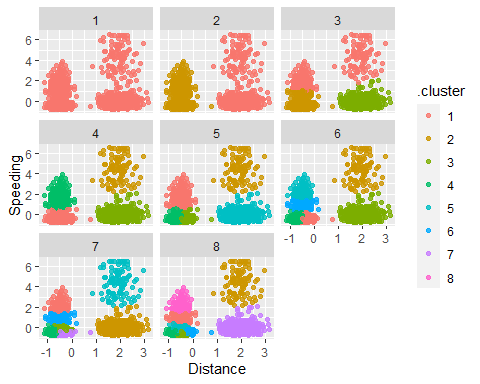
**Task 4**

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]>

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

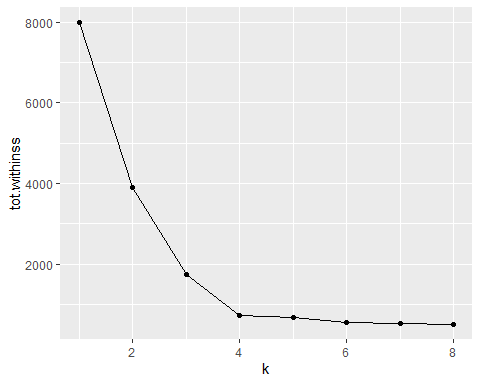
p1 =   
 ggplot(assignments, aes(x= Distance, y= Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +  
 facet\_wrap(~ k)  
  
p1



**Having 4 clusters, would be the most ideal. 4 clusters seems to separate the data nicely without looking too messy. Less than 4, the data seems to not be separated enough.**

**Task 5**

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



**The k value that is most appropriate would be 4.**

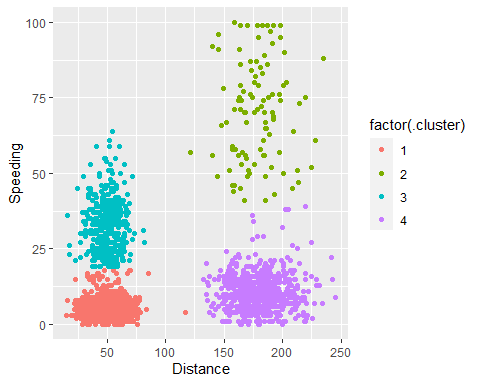
**Task 6**

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

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()  
## .. )

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



**The 4 clusters shown above group the data into like data. the 1st cluster shows drivers who drive on average less than about 100 miles a day and speed 15% or less of the time. CLuster 2 shows drivers who drive about 125-225 miles a day and speed about 50% or more of the time. Cluster 3 shows drivers who drive less than 100 miles a day and speed between 20%-60% of the time. Finally, cluster 4 consists of drivers who drive 125-250 miles a day and speed about less 25% of the time.**