# Module 6 - Assignment 1

## Andrew Jennings

trucks <- read\_csv("trucks.csv")

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

str(trucks)

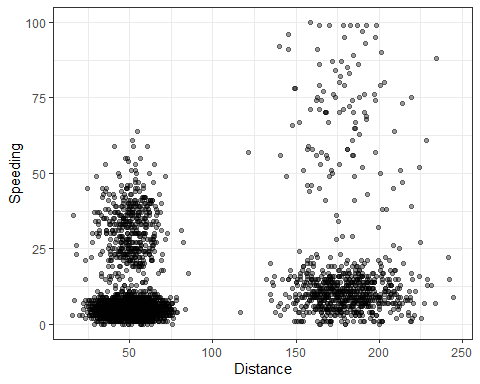
## spec\_tbl\_df [4,000 x 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Driver\_ID: num [1:4000] 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ Distance : num [1:4000] 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding : num [1:4000] 28 25 27 22 25 10 20 8 34 19 ...  
## - attr(\*, "spec")=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

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

 There appears to be 4 different clusters in the Trucks data set. Those who travel shorter distances with very little percentage of speeding, and those that spend roughyl 20%-65% of their time speeding. On the other end there are those that travel over 125 miles and speed less than 25% of the time and and another clustering that travel that distance and speeds over 25% of the time.

### Task 2 - Cleaned/Scaled Data Set

trucks\_scaled = scale(trucks)  
  
trucks\_recipe = recipe(~ Distance + Speeding , trucks\_scaled)   
  
trucks\_recipe = prep(trucks\_recipe, trucks\_scaled)   
  
trucks\_cleaned = bake(trucks\_recipe, trucks\_scaled)

### Task 3 - kMeans Custering

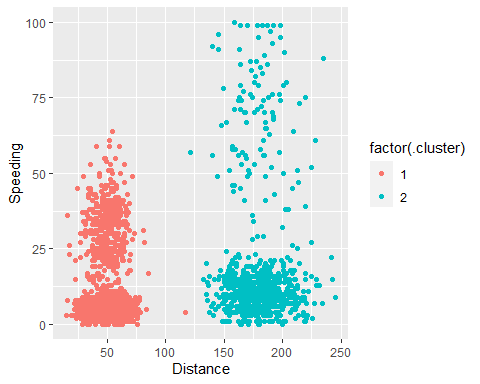
set.seed(64)  
clusts =   
 tibble(k = 2) %>%  
 mutate(kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned))  
clusts

## # A tibble: 1 x 5  
## k kclust tidied glanced augmented   
## <dbl> <list> <list> <list> <list>   
## 1 2 <kmeans> <tibble [2 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))

clusters = kmeans(trucks\_cleaned , 2)  
trucks = augment(clusters, trucks)

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



This model broke the trucks data set into 2 clusters. Those that drove over 120 miles and those that drove under 120 miles.

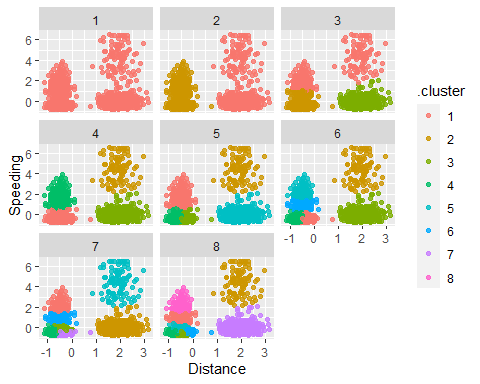
### Task 4 - More K Values

set.seed(412)  
clusts2 =   
 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))  
clusts2

## # 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]>

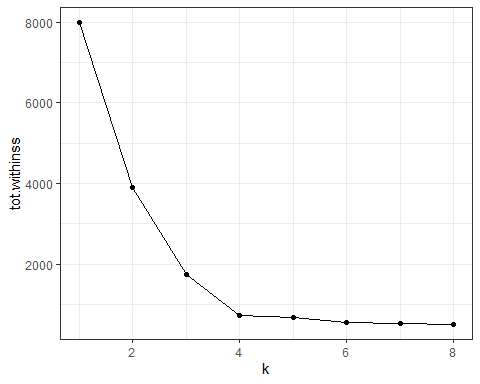
clusters2 =  
 clusts2 %>%  
 unnest(cols = c(tidied))  
  
assignments2 =   
 clusts2 %>%   
 unnest(cols = c(augmented))  
  
clusterings2 =   
 clusts2 %>%  
 unnest(cols = c(glanced))

p1 =   
 ggplot(assignments2, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p1

 Based off of the chart below I believe that 4 clusters is the most appropriate number of clusters because it appears to be the most descirptive without overfitting the data set.

### Task 5 -

ggplot(clusterings2, aes(k, tot.withinss)) +  
 geom\_line() +  
 geom\_point() + theme\_bw()



Based off of the graph above we can see that 4 clusters is the most appropriate number of clusters for the trucks data set.

### Task 6 -

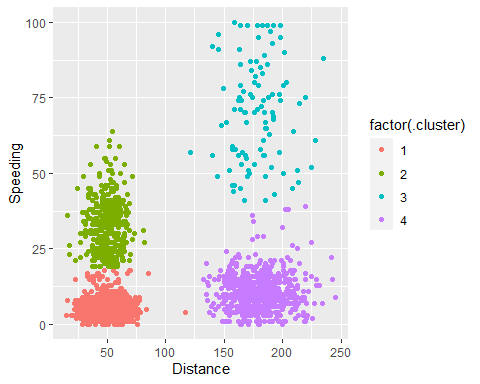
set.seed(64)  
clusts3 =   
 tibble(k = 4) %>%  
 mutate(kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned))  
clusts3

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

clusters3 =   
 clusts %>%  
 unnest(cols = c(tidied))  
  
assignments3 =   
 clusts %>%   
 unnest(cols = c(augmented))  
  
clusterings3 =   
 clusts %>%  
 unnest(cols = c(glanced))

clusters3 = kmeans(trucks\_cleaned , 4)  
trucks = augment(clusters3, trucks)

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

 The clusters shown are great depcitors of 4 categories of different people. Those who drive under 120 miles who rarely speed, those who drive under 120 miles who speed roughly 20%-60% of the time, those who travel over 120 miles and speed less than 40% of the time and those who travel over 120 miles and speed more than 40% of theh time.