## Module 6 Assignment

## Clustering

## Katherine Hewitt

Libraries

Read in Data

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

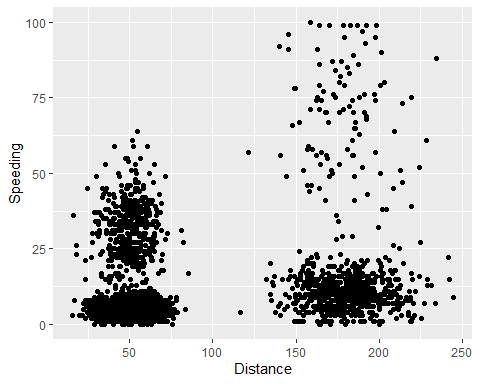
## Rows: 4000 Columns: 3

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (3): Driver\_ID, Distance, Speeding

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#Task 1 The most significant cluster is for drivers who drive an average of 25 - 75 miles per day. These drivers go over the speed limit by at least 5 mph 0 - 12.5% of the time. There is also another cluster of drivers within this distance range who drive at least 5 mph over the speed limit approx. 15 – 50% of the time. Another cluster is for drivers who drive an average of 150 -225 miles per day. These drivers go over the speed limit by at least 5 mph 0 - 25% of the time.

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

 #Task 2

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

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

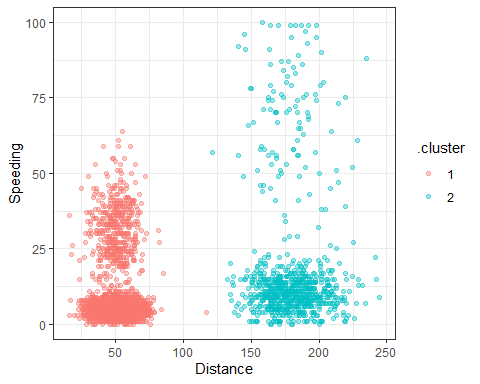
## Distance Speeding   
## Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 3.1560 Max. : 6.5127

#Task 3 The clusters look similar to those in Task 1. The most significant cluster is for drivers who drive an average of 25 - 75 miles per day. There is also another cluster of drivers within this distance range who drive at least 5 mph over the speed limit approx. 15 – 50% of the time. These drivers go over the speed limit by at least 5 mph 0 - 12.5% of the time. Another cluster is for drivers who drive an average of 150 -225 miles per day. These drivers go over the speed limit by at least 5 mph 0 - 25% of the time.

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

## tibble [4,000 x 3] (S3: tbl\_df/tbl/data.frame)  
## $ 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 ...

ggplot(trucks, aes(x = Distance, y = Speeding, color = .cluster))+geom\_point(alpha = .4) + theme\_bw()

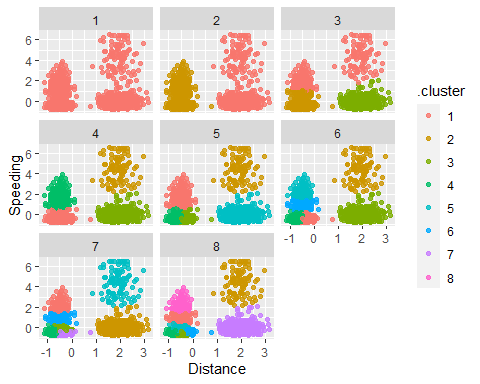


#Task 4 The best value of K appears to be either 3 or 4 based on how the points are clustered together and color coded.

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

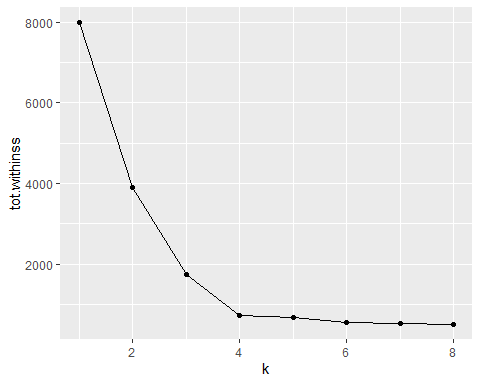
## # A tibble: 8 x 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 2 2 <kmeans> <tibble [2 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 3 3 <kmeans> <tibble [3 x 5]> <tibble [1 x 4]> <tibble [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))  
clustering = clusts %>% unnest(cols = c(glanced))  
  
p1= ggplot(assignments, aes (x=Distance, y= Speeding)) + geom\_point(aes(color = .cluster), alpha = .8) + facet\_wrap(~k)  
p1



## Task 5

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



## Task 6

This plot with k value of 4 has 4 different colored plots. The red plot is as decribed earlier- drivers who drive an average of 25 - 75 miles per day. These drivers go over the speed limit by at least 5 mph 0 - 12.5% of the time. The turquoise cluster drivers is a second set of drivers who drive between approx 27 - 48 miles a day. These drivers go over the speed limit by at least 5 MPH approx 15 - 50% of the time. The purple cluster was also described earlier and is for drivers who drive an average of 150 -225 miles per day. These drivers go over the speed limit by at least 5 mph 0 - 25% of the time. Green is the final color of points. These points are more scattered and not as dense. These drives drive 15 - 250 miles per day and go over the speed limit by at least 5 mph from approx 40 - 100% of the time.

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

## tibble [4,000 x 3] (S3: tbl\_df/tbl/data.frame)  
## $ 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 ...

ggplot(trucks, aes(x = Distance, y = Speeding, color = .cluster))+geom\_point(alpha = .4) + theme\_bw()

