K-Means Clustering

Code

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Hide

```
#K-Means Clustering

library(data.table)
library(ggplot2)
bike = copy(Ch5_bike_station_locations)
setDT(bike)
str(bike)
```

```
Classes ‘data.table’ and 'data.frame':  244 obs. of  2 variables:
 $ latitude : num  39 38.9 39 38.9 38.9 ...
 $ longitude: num  -77 -77 -77.1 -76.9 -77.1 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

Hide

```
#lat is y axis
#long is x axis
```

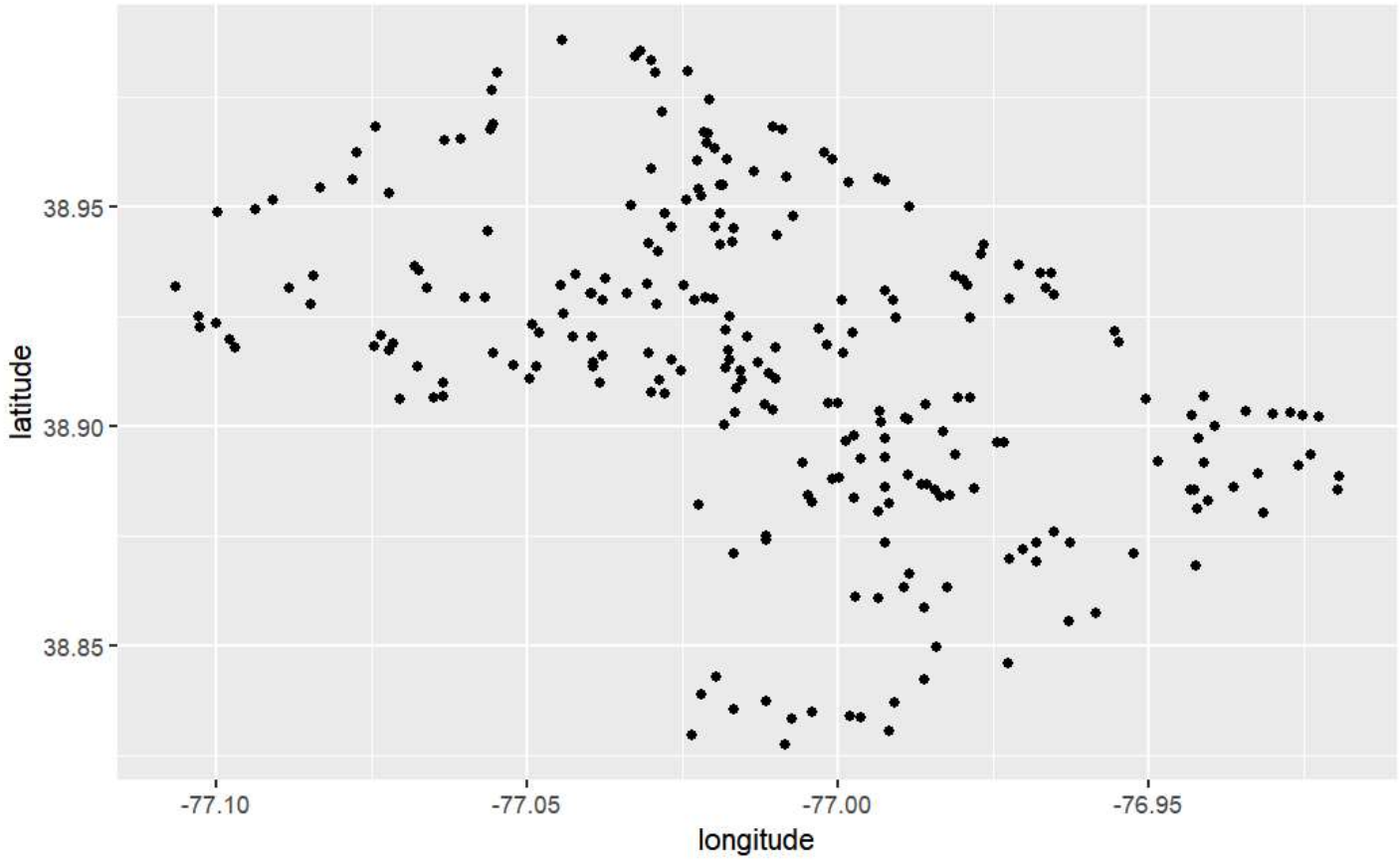
Hide

```
#Exploring the Data
grep('NA',bike)
```

```
integer(0)
```

Hide

```
ggplot(bike ,aes(x=longitude ,y=latitude)) + geom_point()
```



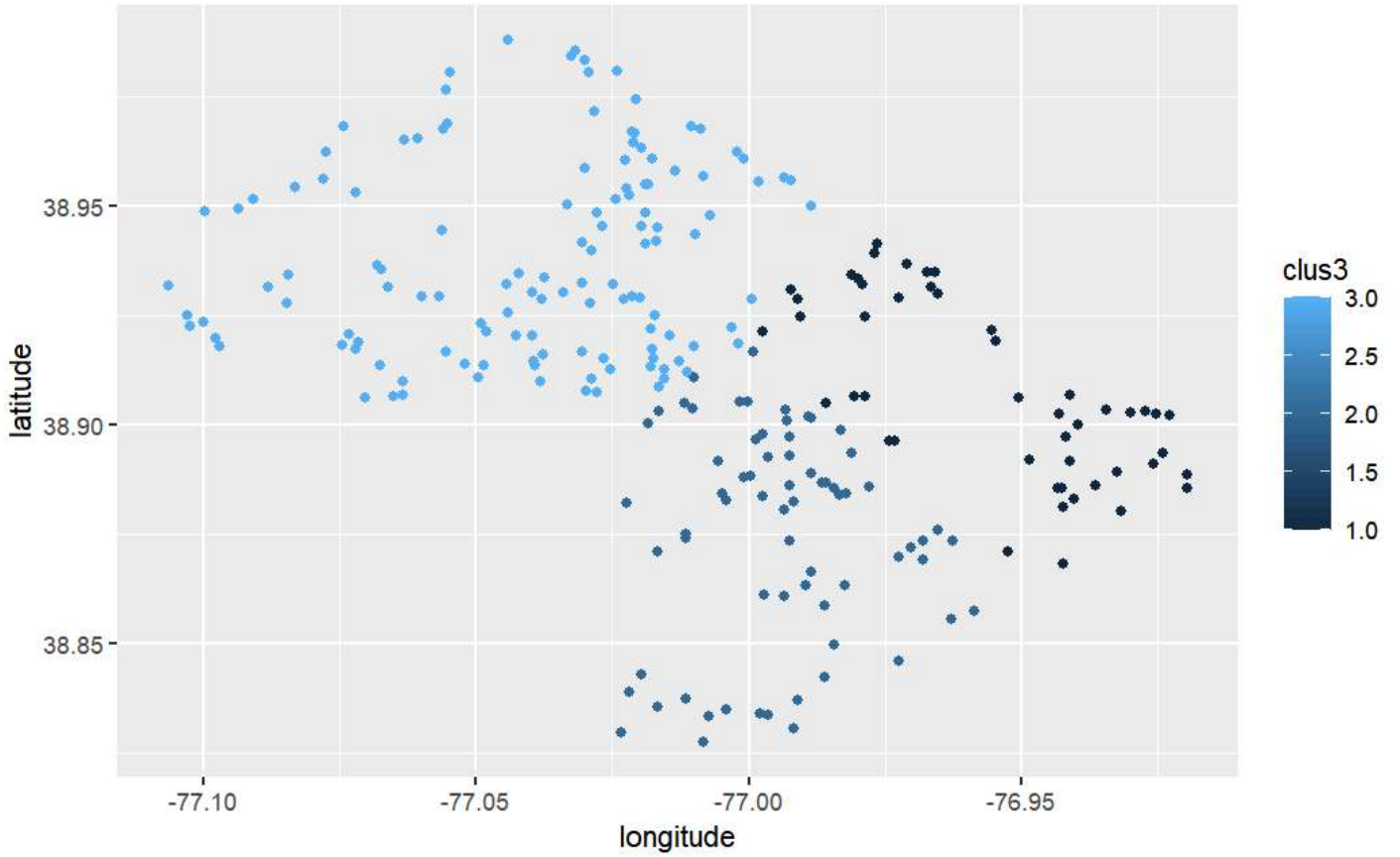
Hide

```
#Running kmeans() Function
set.seed (123)
k3=kmeans(bike ,3) #3 is the number of clusters that you want
k3
```

```
K-means clustering with 3 clusters of sizes 48, 69, 127

Cluster means:
  latitude longitude
1 38.90753 -76.95526
2 38.87463 -76.99426
3 38.93839 -77.03945

Clustering vector:
  [1] 3 2 3 1 3 3 1 2 1 1 3 1 1 1 3 3 3 3 3 1 2 3 3 1 3 1 3 3 3 2 1 2 3 3 3 3 3 1 3 3 2 3
 [43] 3 3 1 3 1 3 2 3 2 3 2 2 1 1 1 2 3 1 3 3 3 3 2 2 2 3 1 3 1 3 3 3 1 3 2 3 2 3 1 3 3 3
 [85] 1 3 2 3 2 2 3 1 3 2 2 3 3 2 3 3 2 2 3 3 1 3 3 3 3 3 3 3 2 3 3 2 3 3 2 1 1 1 3 3 3 2
[127] 2 2 2 2 2 3 3 3 3 2 1 2 3 1 3 3 3 3 3 3 3 2 3 2 3 3 2 3 3 2 3 1 1 1 1 2 2 2 3 1 3 3
[169] 3 2 3 3 3 1 3 2 2 3 3 3 2 2 2 3 2 2 2 3 2 3 2 2 3 3 2 1 2 3 1 3 3 3 3 3 3 1 2 3 3 3
[211] 3 2 2 3 3 1 3 1 2 3 2 3 1 2 3 2 2 3 2 1 3 1 3 3 1 2 1 3 2 3 2 1 2 1

Within cluster sum of squares by cluster:
[1] 0.04361512 0.05663749 0.15939642
 (between_SS / total_SS =  63.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

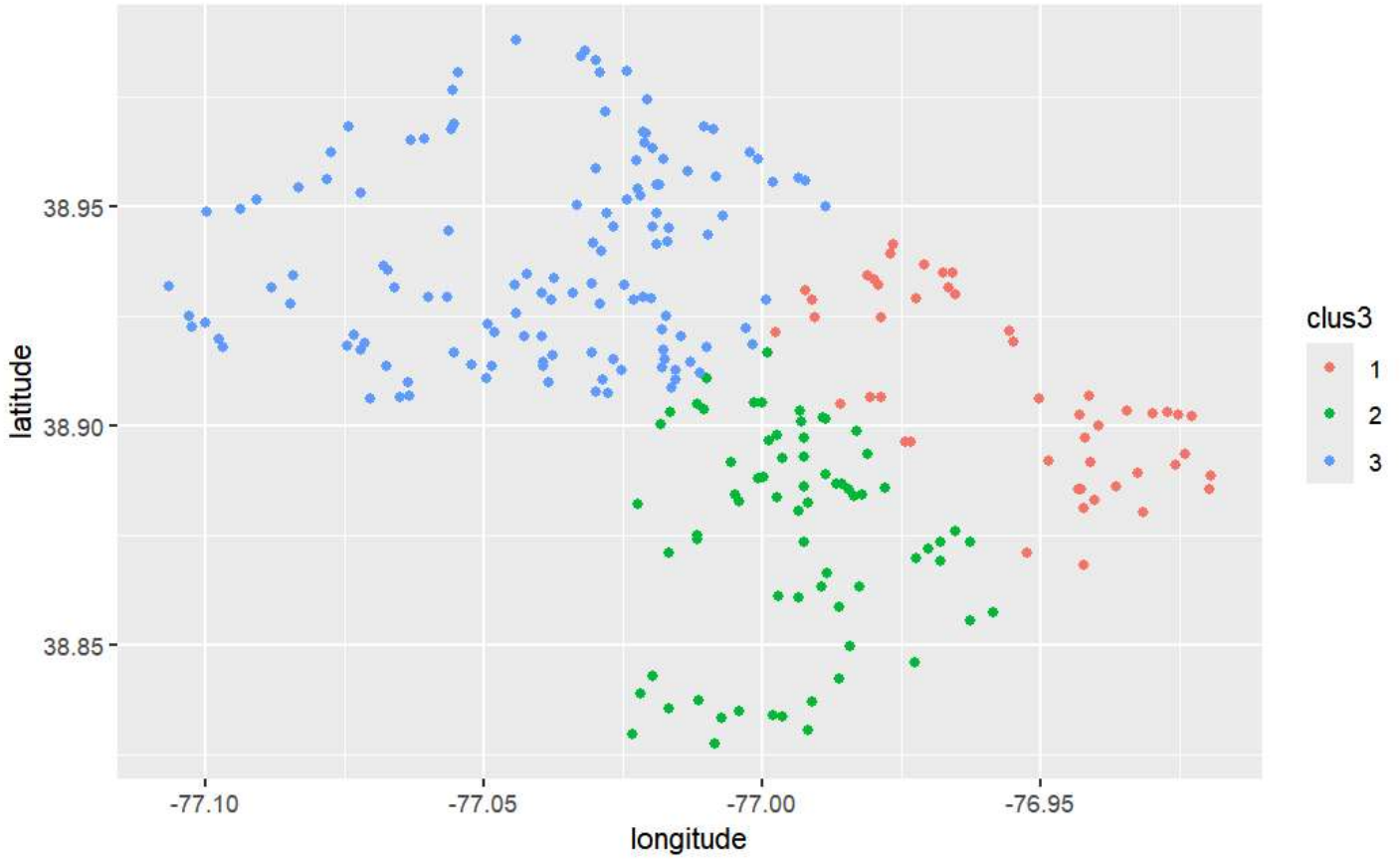
Hide

```
bike[,clus3:=k3$cluster]
ggplot(bike ,aes(x=longitude ,y=latitude ,color=clus3)) + geom_point ()
```



Hide

```
#clus3 is a categorical -> ordinal so it needs to be factored
bike[,clus3:= factor(clus3)]
ggplot(bike ,aes(x=longitude ,y=latitude ,color=clus3)) +
  geom_point ()
```



Hide

```
#tell is the center of each of the locations
k3$centers
```

```
  latitude longitude
1 38.90753 -76.95526
2 38.87463 -76.99426
3 38.93839 -77.03945
```

Hide

```
#makes data into a datatable
centdt=data.table(k3$centers)
centdt
```

latitude <dbl>	longitude <dbl>
38.90753	-76.95526
38.87463	-76.99426
38.93839	-77.03945

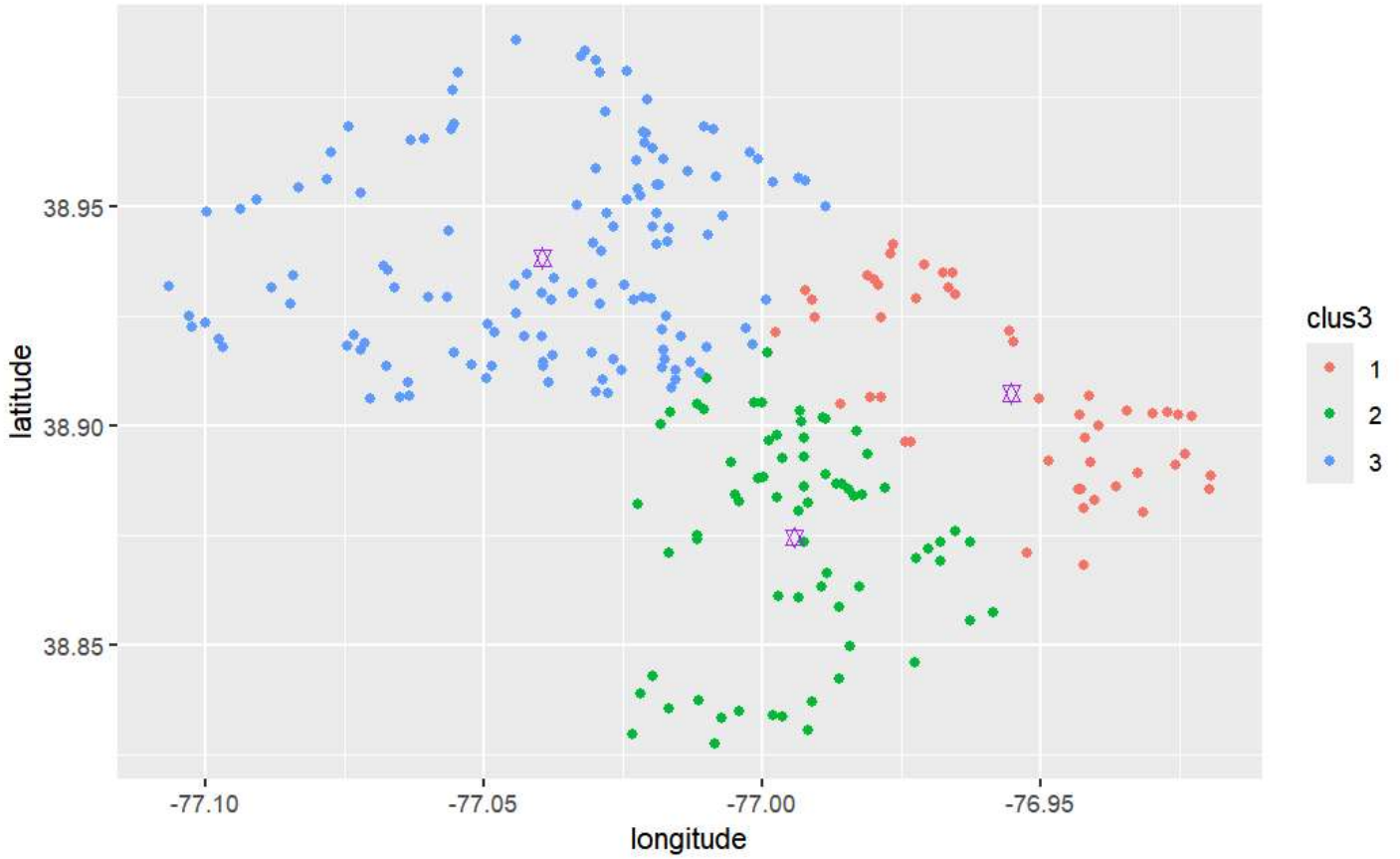
3 rows

Hide

NA

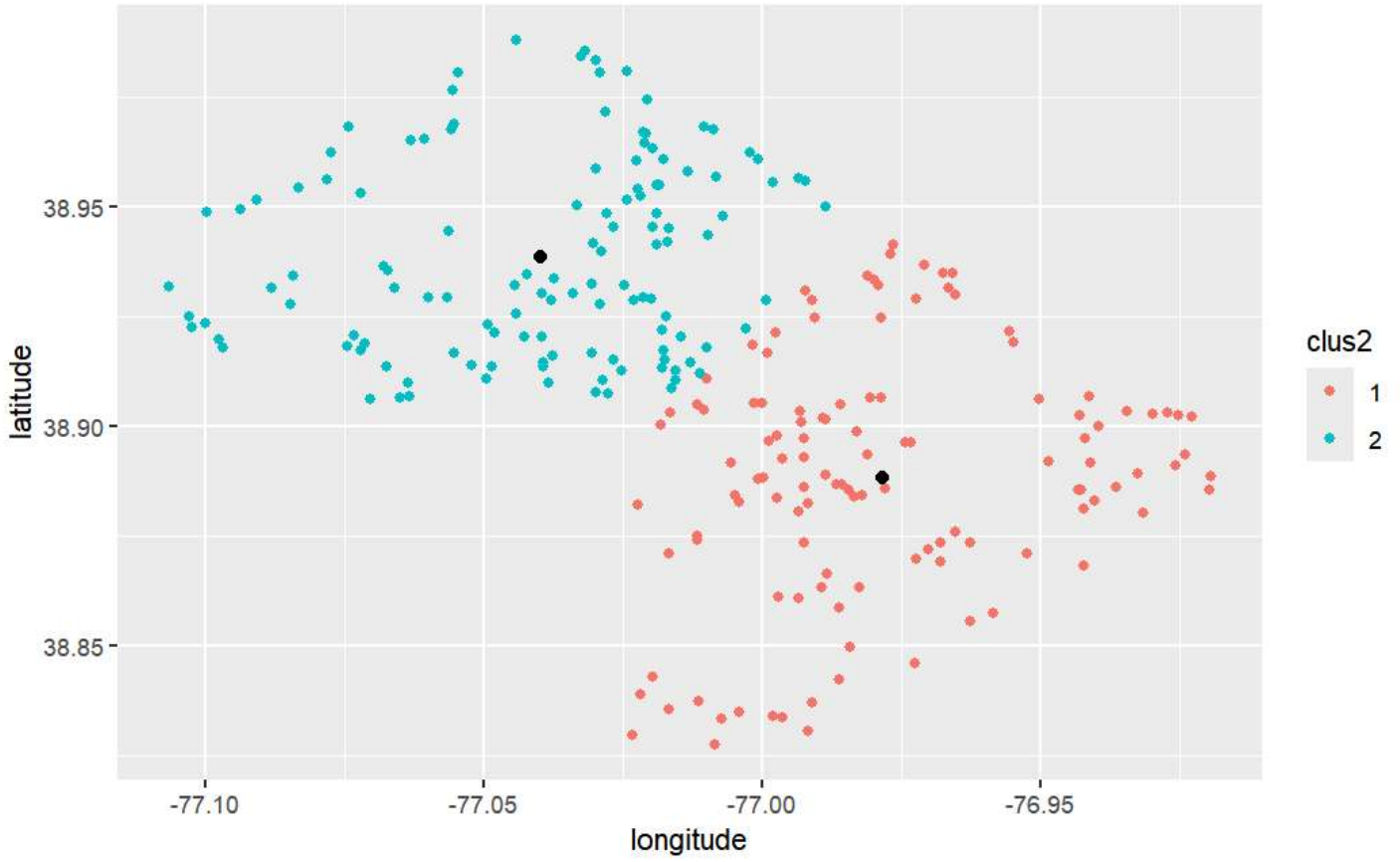
Hide

```
#This shows where the middle each cluster is
ggplot(bike ,aes(x=longitude ,y=latitude ,color=clus3)) +
  geom_point () + geom_point(data=centdt ,aes(x=longitude , y=latitude), colour="purple", shape=
11, size =2)
```



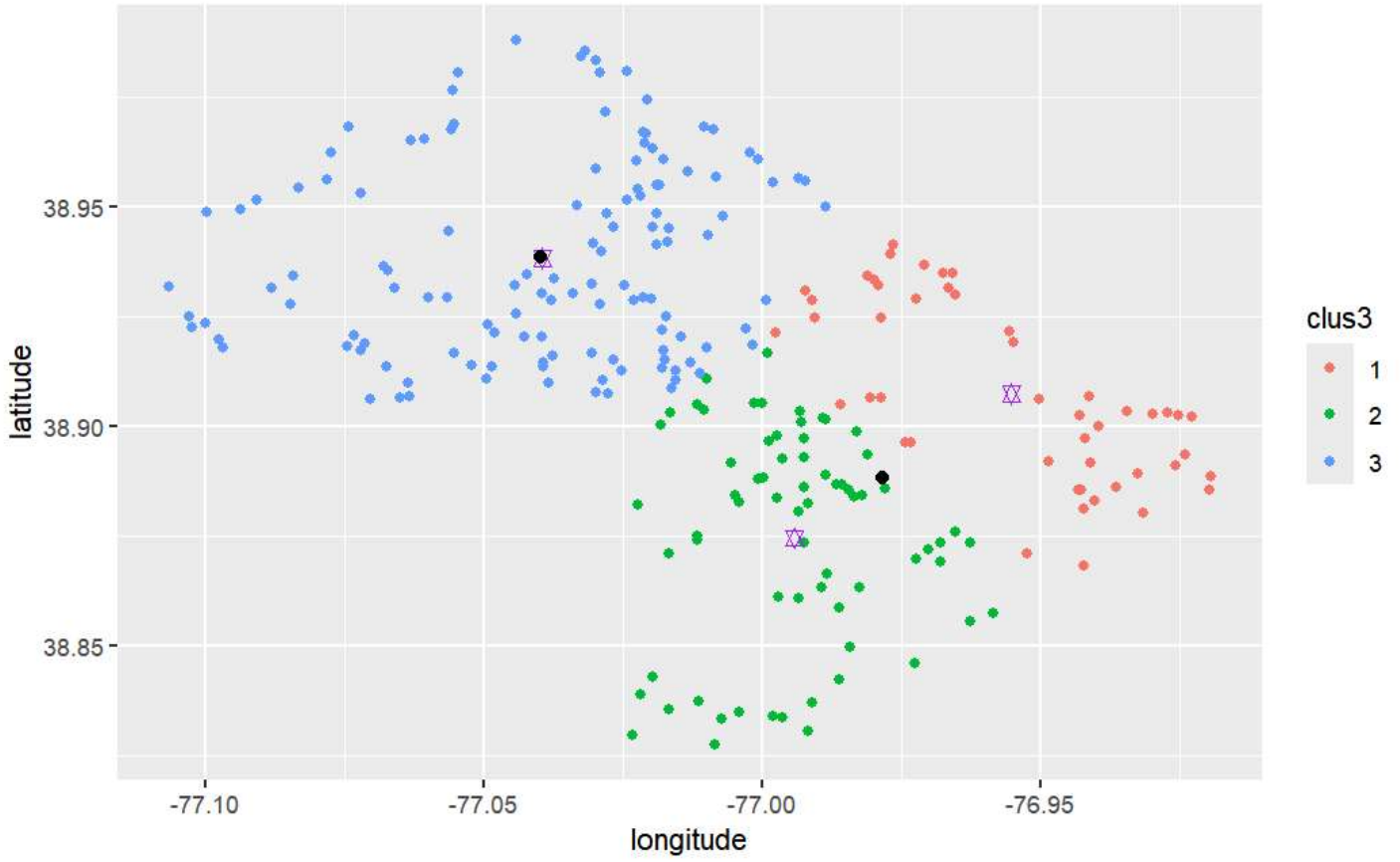
Hide

```
#rerun analysis for 2 stations
k2 = kmeans(bike[,.(latitude ,longitude)],2)
bike[,clus2:=k2$cluster]
bike[,clus2:= factor(clus2)]
centdt2=data.table(k2$centers)
ggplot(bike ,aes(x=longitude ,y=latitude ,color=clus2)) +geom_point () +
  geom_point(data=centdt2 ,aes(x=longitude , y=latitude),colour="black", shape=19,size =2)
```



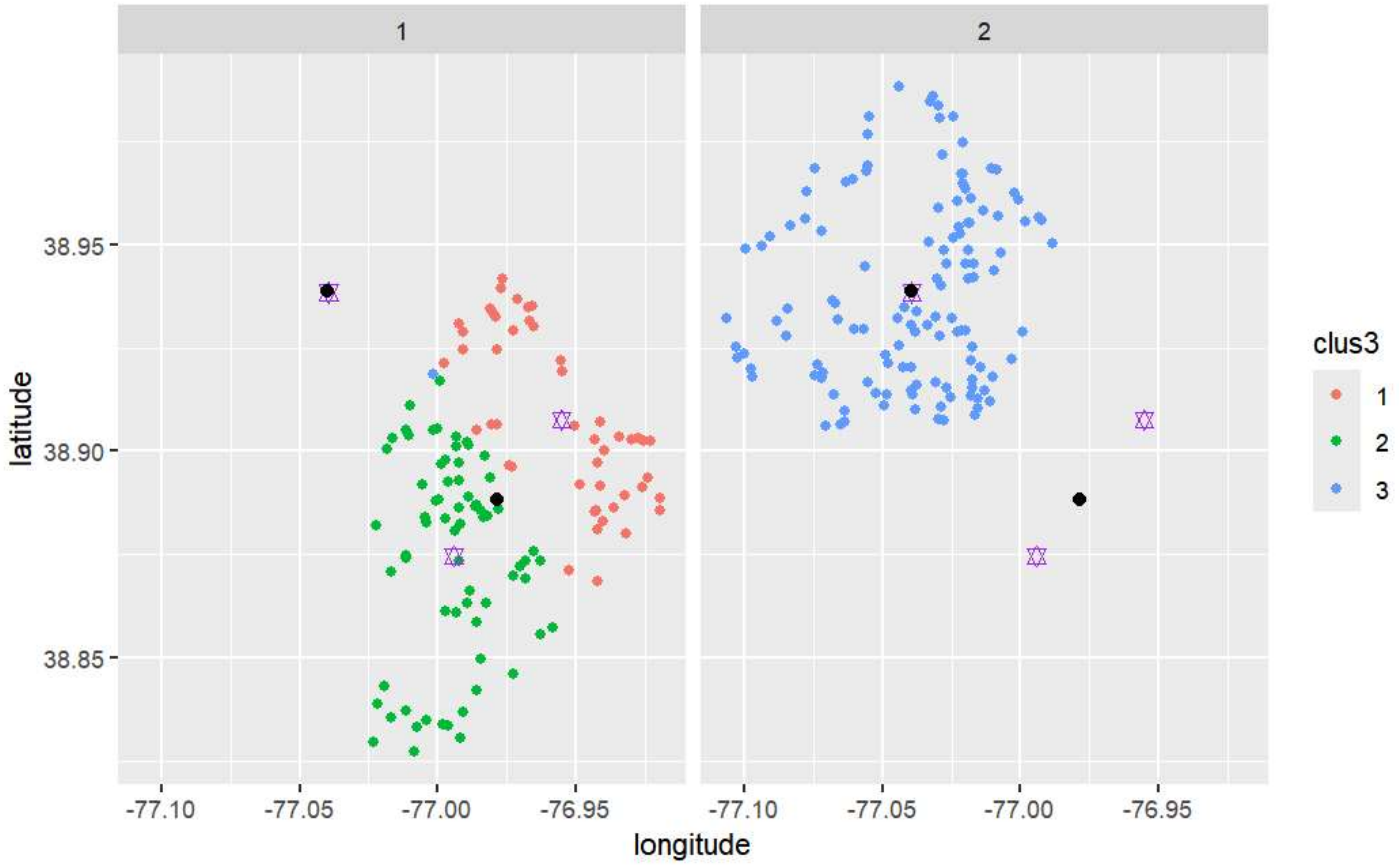
Hide

```
#2 stations vs 3
ggplot(bike ,aes(x=longitude ,y=latitude ,color=clus3)) +
  geom_point ()+geom_point(data=centdt ,aes(x=longitude , y=latitude), colour="purple", shape=1
1,size =2)+geom_point(data=centdt2 ,aes(x=longitude , y=latitude), colour="black", shape =19,siz
e =2)
```



Hide

```
ggplot(bike ,aes(x=longitude ,y=latitude ,color=clus3)) +
  geom_point () + geom_point(data=centdt ,aes(x=longitude , y=latitude), colour="purple", shape=
11,size =2)+geom_point(data=centdt2 ,aes(x=longitude , y=latitude), colour="black", shape =19,si
ze =2) +facet_wrap(~clus2)
```


1
2
38.95
38.90
38.85
latitude
-77.10
-77.05
-77.00
-76.95
longitude
clus3
1
2
3