Daanish Ahsan
Applications of Data Science and Statistical Modelling
Candidate Number: 124070

Assignment: Clustering Power Data

Q1-10

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)

library(MASS)

library(ggplot2)
library(caret)

library(splines)
library(dplyr)
library(mgcv)

library(viridis)

library(Rmisc)

library(chron)
library(kableExtra)

library(factoextra)

library(rdist)
library(ggpubr)

library(flexclust)

library(NbClust)
library(corrplot)

library(scico)
#Apps Project

set.seed(500)

#Loading in the data
load("/Users/Daanish/Downloads/Data 7/Autumn_2012.RData")
load("/Users/Daanish/Downloads/Data 7/HighSummer_2012.RData")
load("/Users/Daanish/Downloads/Data 7/Spring_2013.RData")
load("/Users/Daanish/Downloads/Data 7/Summer_2012.RData")
load("/Users/Daanish/Downloads/Data 7/Winter_2012.RData")
Rawmeasurementsnewstations <- read.csv('Data/NewSubstations.csv', stringsAsFactors = FALSE)

Characteristics <- read.csv("Data/Characteristics.csv", stringsAsFactors=FALSE)
```

Daanish Ahsan
Applications of Data Science and Statistical Modelling
Candidate Number: 124070

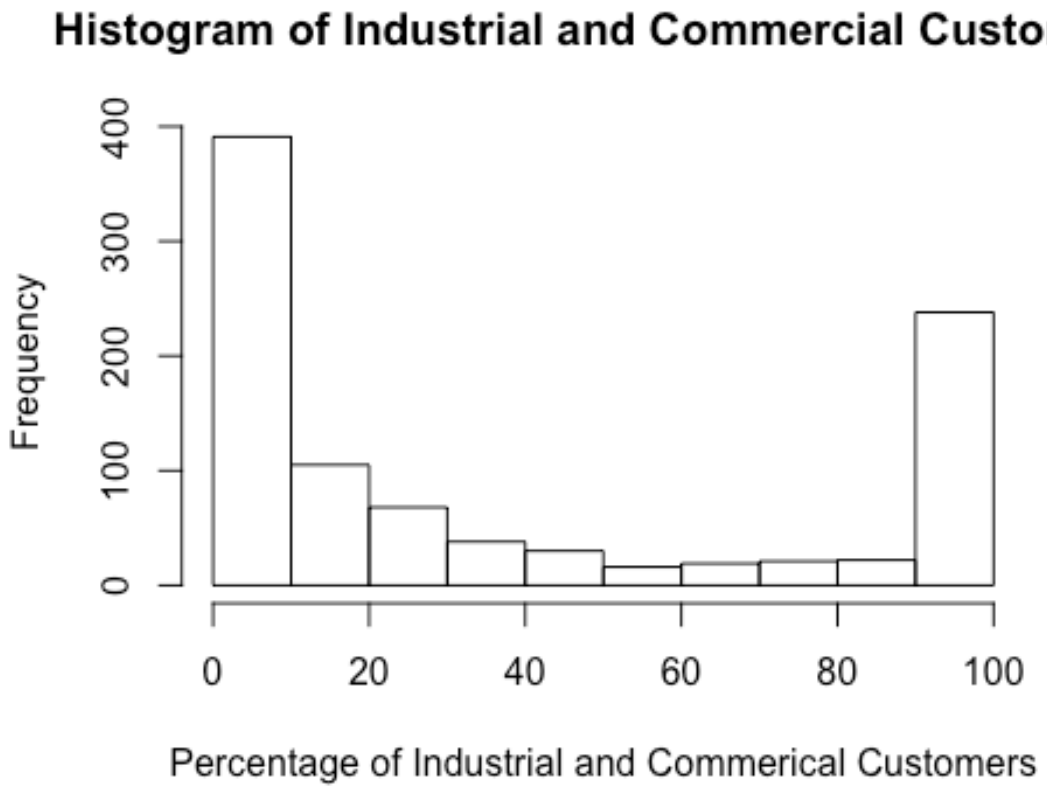
```
#Converting Julian Dates, format of the dates is month, day, year

converted_dates <- dates(Autumn_2012[,2], origin = c(month = 1,day = 1,year = 1970))
Autumn2012convdates <- converted_dates[1:20962]

Autumn_2012$Date <- Autumn2012convdates

converted_dates <- dates(Spring_2013[,2], origin = c(month = 1,day = 1,year = 1970))
Spring2013convdates <- converted_dates[1:23451]

Spring_2013$Date <- Spring2013convdates

converted_dates <- dates(Summer_2012[,2], origin = c(month = 1,day = 1,year = 1970))
Summer2012convdates <- converted_dates[1:17336]

Summer_2012$Date <- Summer2012convdates

converted_dates <- dates(Winter_2012[,2], origin = c(month = 1,day = 1,year = 1970))
Winter2012convdates <- converted_dates[1:57210]

Winter_2012$Date <- Winter2012convdates

converted_dates <- dates(HighSummer_2012[,2], origin = c(month = 1,day = 1,year = 1970))
HighSummer20120convdates <- converted_dates[1:17727]

HighSummer_2012$Date <- HighSummer20120convdates
```

#Q1

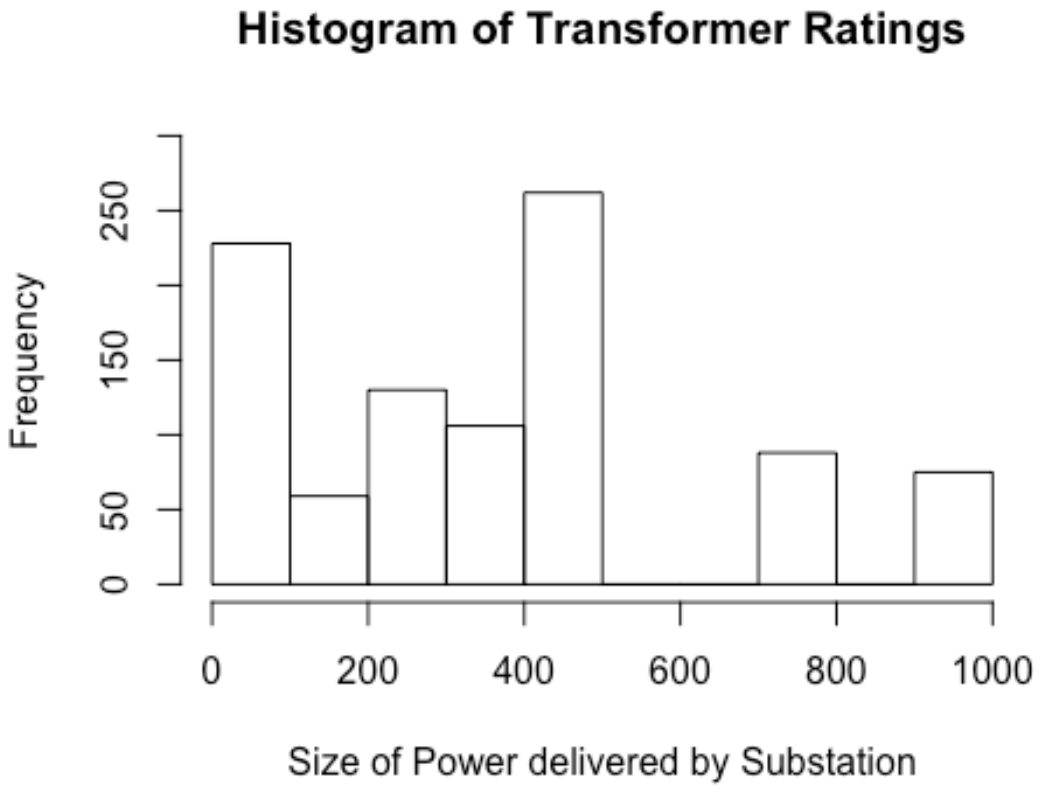
```
summary(Characteristics)

##  SUBSTATION_NUMBER TRANSFORMER_TYPE   TOTAL_CUSTOMERS Transformer_RATIN
G
##  Min.   :511016    Length:948         Min.   :  0.0   Min.   :   0.0
##  1st Qu.:521516    Class :character   1st Qu.:  3.0   1st Qu.: 200.0
##  Median :532652    Mode  :character   Median : 67.5   Median : 315.0
##  Mean   :534344                       Mean   :104.3   Mean   : 389.1
##  3rd Qu.:552386                       3rd Qu.:179.2   3rd Qu.: 500.0
##  Max.   :564512                       Max.   :569.0   Max.   :1000.0
##  Percentage_IC     LV_FEEDER_COUNT  GRID_REFERENCE
##  Min.   :0.00000   Min.   : 0.000   Length:948
##  1st Qu.:0.01048   1st Qu.: 1.000   Class :character
##  Median :0.17849   Median : 3.000   Mode  :character
##  Mean   :0.37982   Mean   : 2.762
##  3rd Qu.:0.90271   3rd Qu.: 4.000
##  Max.   :1.00000   Max.   :16.000
```

Daanish Ahsan
Applications of Data Science and Statistical Modelling
Candidate Number: 124070

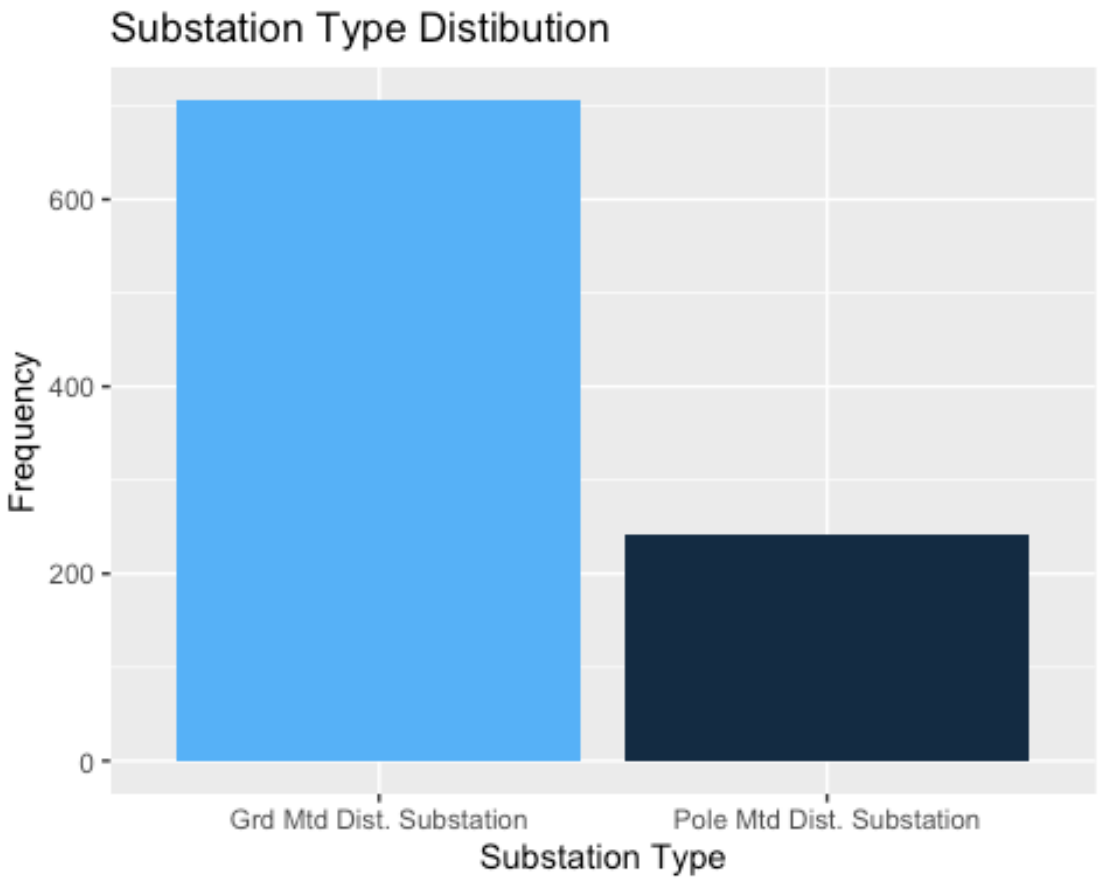
```
#Average number of customers is 104, median amount is 67.5,
#average number of feeders coming from substation is 2.7
#Average % of customers that are industrial and commercial is approx 38%

Characteristics$Percentage_IC <- Characteristics$Percentage_IC*100

hist(Characteristics$Percentage_IC, xlab='Percentage of Industrial and Commerical Customers',
    main ='Histogram of Industrial and Commercial Customers')
```




```
#Can see that nearly half of all stations have no IC customers, with just under 20% having 90-100 %
#IC. Between 10 and 90% IC customers are very low among substations, indicating that
#Generally, they either specialise in serving IC or not at all

hist(Characteristics$Transformer_RATING, ylim = c(0,300), xlab = 'Size of Power delivered by Substati
on',
    main='Histogram of Transformer Ratings')
```




```
#About half of all substations deliver either 0-100 power or 400-500.
#No stations deliver between 500 and 700, or 800-900
#Suggests perhaps that most substations do not have heavy power demands

p <- as.factor(Characteristics$TRANSFORMER_TYPE)

Characteristics <- mutate(Characteristics, TRANSFORMER_TYPE=p )

transformertype <- as.data.frame(table(Characteristics$TRANSFORMER_TYPE))

ggplot(transformertype, aes(x=Var1, y=Freq)) + geom_bar(stat='identity',aes(fill = Freq)) + labs(title="
Substation Type Distibution", x='Substation Type', y='Frequency') +
  theme(legend.position = 'none')
```




```
#About 700 stations are ground mounted and thus urban, with about 240 being pole mounted and rural
#The split suggests that power demands are much higher in urban areas which is why
#there are much more urban types
#This makes sense as populations are much higher in cities compared to the countryside
```

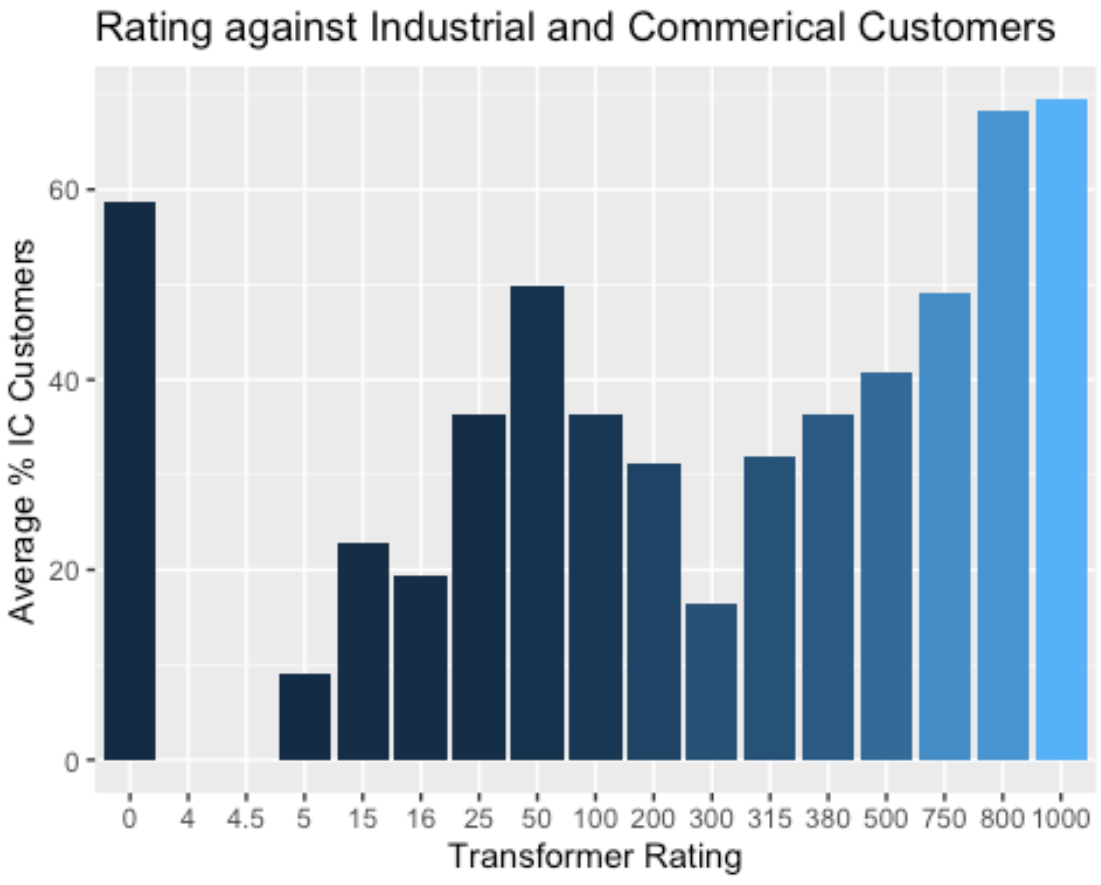
#Q2

```
#Relationship between average total customers and transformer type

custtype <- aggregate(Characteristics$TOTAL_CUSTOMERS, by=list(TransformerType=Characteristic
s$TRANSFORMER_TYPE), FUN=mean)

custtype <- rename(custtype,c('x'='Average Number of Total Customers'))

kable(custtype, 'html') %>%
  cat(., file = 'custtype.html') #table
```

TransformerType	Average Number of Total Customers
Grd Mtd Dist. Substation	135.7776

Pole Mtd Dist. Substation	12.6281

```
#Clear that from previous diagram ground mounted substations are much more prevalent, and also have
on average
#many more customers, at about 136
#Whilst pole mounted, which are present more in rural areas have only about
#13 customers on average

#Relationship between Rating and IC%

RatingIC <- aggregate(Characteristics$Percentage_IC,
by=list(TransformerRating=Characteristics$Transformer_RATING), FUN=mean)

ggplot(RatingIC, aes(x=factor(TransformerRating), y=x, fill=TransformerRating)) + geom_bar(stat =
'identity', width = 0.9) + ggtitle("Rating against Industrial and Commerical Customers") +
  labs(y='Average % IC Customers', x='Transformer Rating') +theme(legend.position = 'none')
```




```
#Interesting trend here in that stations with a rating of 0 seem to have a high %
#of IC customers- perhaps this simply means these stations are not yet active, but have been
#built primarily for IC customers
#Generally, from a rating of 5 to 50, as rating rises, so does the % of IC customers
#Yet the trend then reverses from 100 to 300 rating
#and then shows a positive trend once again from 315 to 1000
```

```
#this is quite odd
#But perhaps in general we can conclude that higher rating stations serve more industrial
#and commercial customers, but that some stations which show this trend reversed are instead used
#for domestic customer needs.

#relationship between feeder count and type

Feedertype <- aggregate(Characteristics$LV_FEEDER_COUNT, by=list(TransformerType=Characteristics$TRANSFORMER_TYPE), FUN=mean)

Feedertype <- rename(Feedertype,c('x'='Average Feeder Count'))

kable(Feedertype, 'html') %>%
  cat(., file = 'Feedertype.html')
```

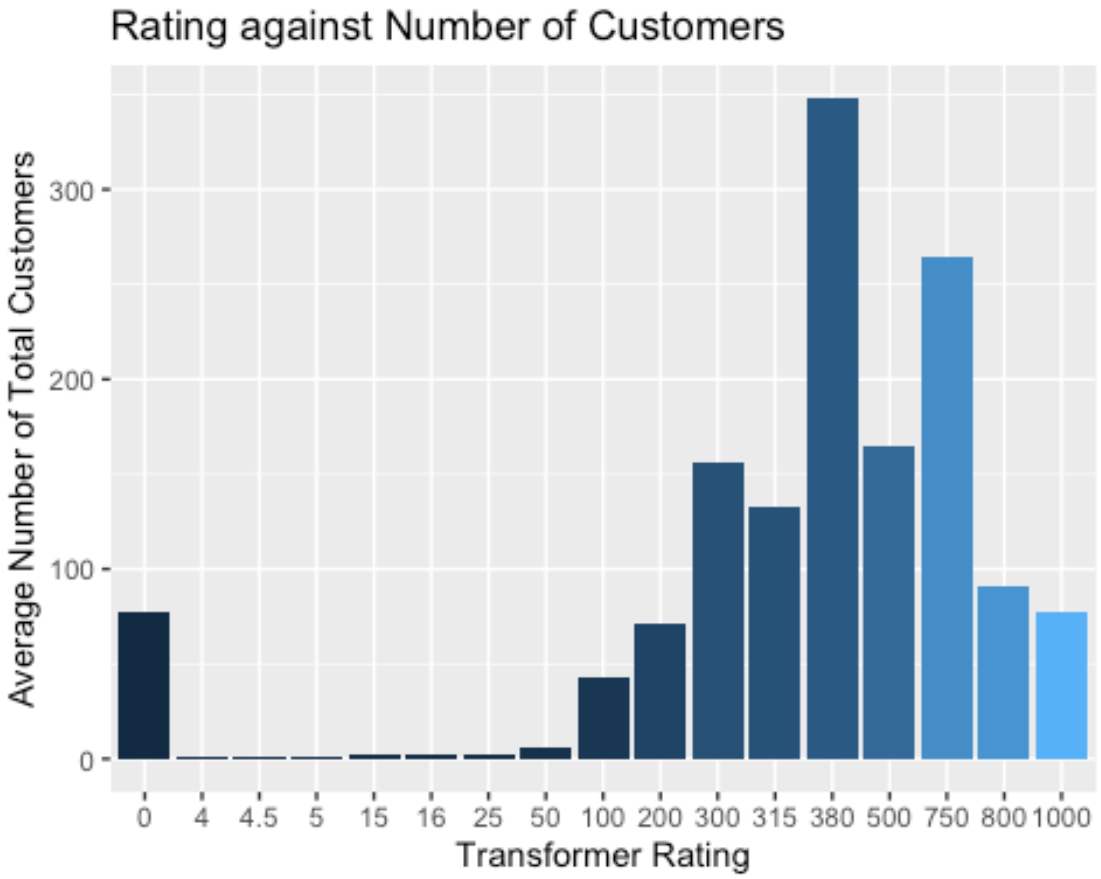
TransformerType	Average Feeder Count
Grd Mtd Dist. Substation	3.355524
Pole Mtd Dist. Substation	1.028926

```
#ground mounted (urban) have 3.4 feeders on average coming from them
#Pole mounted (rural) have 1 feeder on average
#Again this makes sense as urban stations have more customers and so likely provide more power and th
us have more feeders.

#relationship between rating and type

ratingtype <- aggregate(Characteristics$Transformer_RATING, by=list(TransformerType=Characteristic
s$TRANSFORMER_TYPE),FUN=mean)

ratingtype <- rename(ratingtype,c('x'='Average Rating'))



kable(ratingtype, 'html') %>%
  cat(., file = 'ratingtype.html')
```

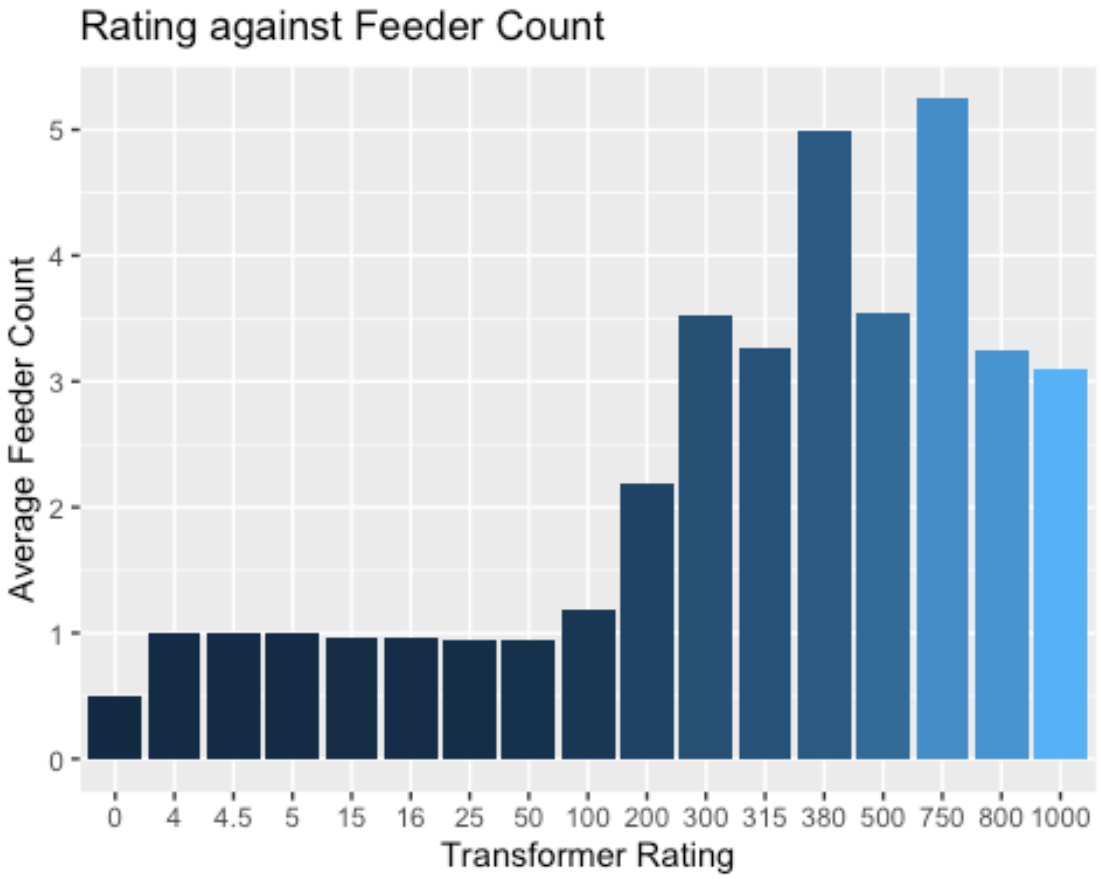
TransformerType	Average Rating
Grd Mtd Dist. Substation	505.15581
Pole Mtd Dist. Substation	50.60124

```
#Table shows that average rating of urban stations is 505, but 10% of this for rural
#at 50.6
#Highlights again that size of power of stations in urban areas is much higher than in rural

#Relationship between customers and rating

custrating <- aggregate(Characteristics$TOTAL_CUSTOMERS, by=list(TransformerRating=Characteris
```

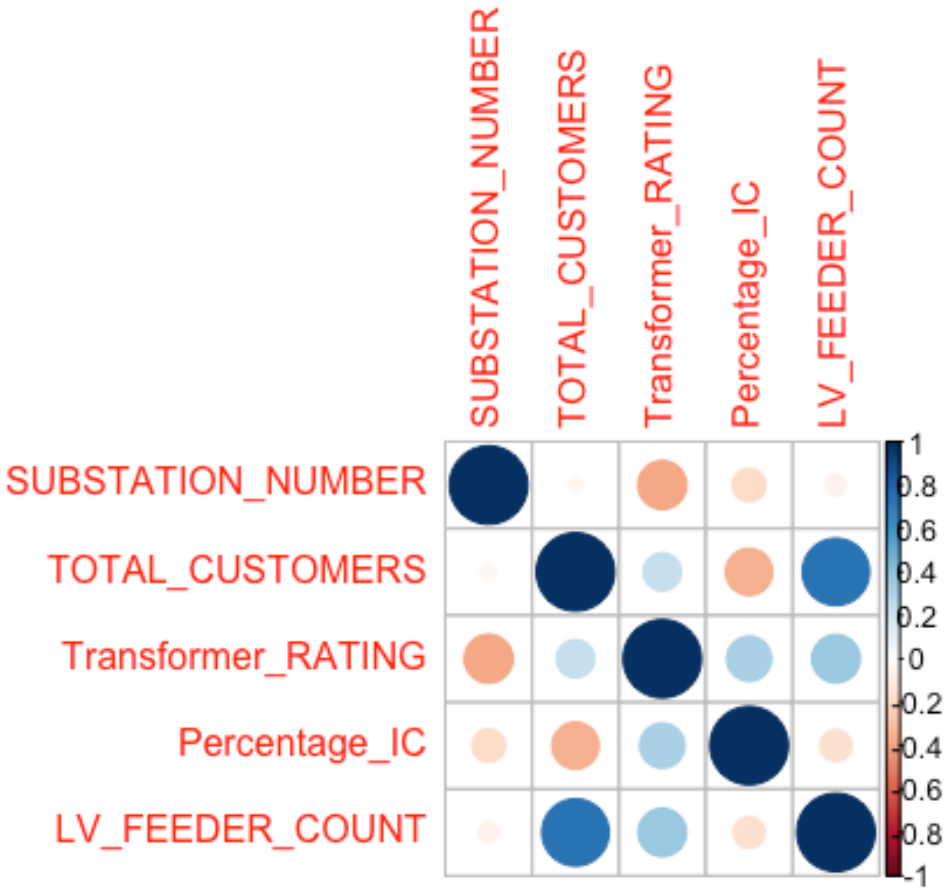
```
tics$Transformer_RATING), FUN=mean)

ggplot(custrating, aes(x=factor(TransformerRating), y=x, fill=TransformerRating)) + geom_bar(stat = 'identity', width = 0.9) + ggtitle("Rating against Number of Customers") +
  labs(y='Average Number of Total Customers', x='Transformer Rating') +theme(legend.position = 'none')
```




```
#As we know that on average, urban stations have higher ratings, we can conclude from this plot
#That most urban stations also have a higher number of total customers on average
#Which again makes sense

#relationship between feeder count and rating

feederrating <- aggregate(Characteristics$LV_FEEDER_COUNT, by=list(TransformerRating=Characteristics$Transformer_RATING), FUN=mean)

ggplot(feederrating, aes(x=factor(TransformerRating), y=x, fill=TransformerRating)) + geom_bar(stat = 'identity', width = 0.9) + ggtitle("Rating against Feeder Count") +
  labs(y='Average Feeder Count', x='Transformer Rating') +theme(legend.position = 'none')
```

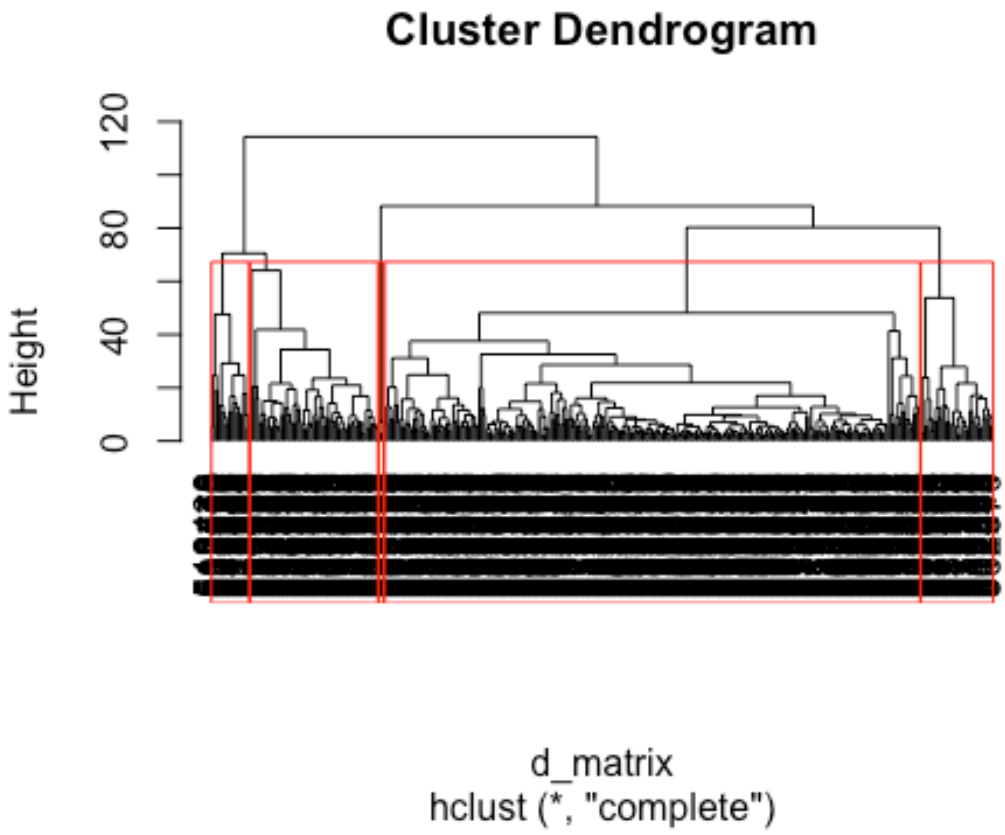



```
#The general trend appears to be that as rating increases, so too does average feeder count
#but clearly there are exceptions, as several stations do not adhere to this trend
#so we can conclude that rating against IC %, number of customers, and feeder count
#do not show a clear linear pattern, but a vague and weakly positive one

Corr <- Characteristics

#Checking which columns are numeric
sapply(Corr, is.numeric)

#Dropping all non-numeric columns
Corr <- Corr[, sapply(Corr, is.numeric)]

#Note that NA values need to be removed here for this cor to work
yaya <- cor(Corr, use = "complete.obs", method = "pearson")

bfgy <- corrplot(yaya, method = 'circle')
```




```
#Printing the corrplot to scale and saving the file
col4 <- scico(100, palette = 'vik')
filetag <- "bfgplsyaya.png"

png(filetag, height = 500, width = 500)

corrplot(yaya, order = "AOE", upper = "ellipse", lower = "number",
         upper.col = col4, lower.col = col4,
         tl.cex = 1, cl.cex = 1, number.cex = 1)

#heat map shows fairly strong positive correlation between total customers and feeder count,
#and weaker but positive correlation transformer rating and feeder count.
#There is also weak but inverse correlation between industrial and commercial customers,
#and total customers, which makes sense as industrial and commerical customers
#exclude domestic customers.

#Q3

Autumn2012v1 <- Autumn_2012

ogbaloo <- aggregate(Autumn2012v1[2:ncol(Autumn2012v1)],(Autumn2012v1['Station']), FUN=mean)
```

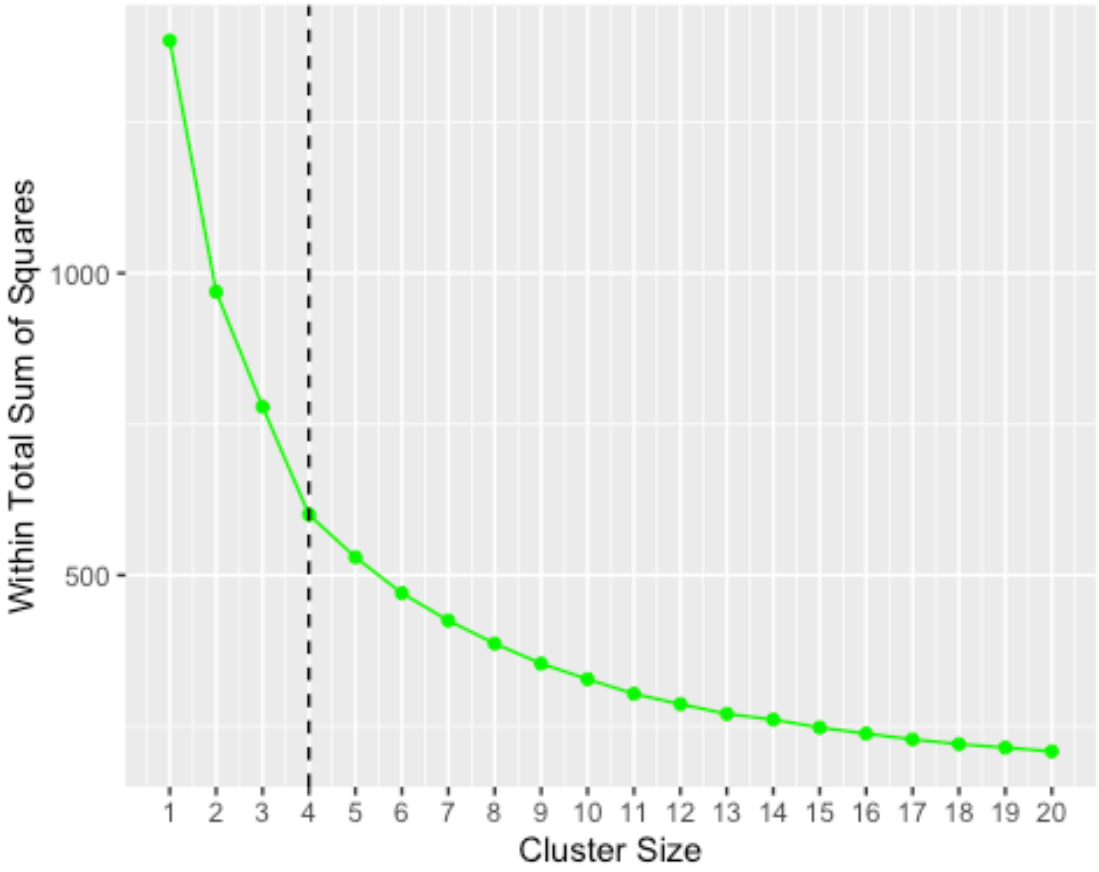
```
ogbaloo1 <- ogbaloo[3:146]

d_matrix <- dist(rbind(ogbaloo1),method="manhattan")

clusterd <- hclust(d_matrix)
#plotting our dendrogram, with hang=-1 to have all labels at same level
plot(clusterd, hang=-1, label=ogbaloo$Station)
rect.hclust(clusterd, k=5, border='RED')
```




```
#4- correct for 4
#picking optimal number of clusters

#do not run NBclust in markdown
NbClust(ogbaloo1, distance = 'manhattan', min.nc = 2, max.nc = 20, method = 'kmeans', index = 'all')
```

```
#NB clust concludes that according to the majority rule, best number of clusters is 4.

#elbow method for optimal k
k.max <- 20
```

```
wss <- sapply(1:k.max,
         function(k){kmeans(ogbaloo1, k, nstart=50,iter.max = 15 )$tot.withinss})
wss

SoSagainstClustersize <- data.frame(1:k.max, wss)

ggplot(SoSagainstClustersize, aes(x = 1:k.max, y = wss)) +
 geom_point(col= 'green') +
 geom_line(col= 'green') +
 scale_x_continuous(breaks = seq(1, 20, by = 1)) +
 labs(x='Cluster Size', y='Within Total Sum of Squares') +
 geom_vline(xintercept=4, linetype=2) + #vertical line here represents point of maximum curvature
 theme(legend.position='none')
```




```
#Optimal K is 4, where curve is starting to have a diminishing return and total within sum of squares is at
a maximum value
#Vertical dotted line at x=4 shows maximum curve where k is optimal

#But I will use 5 clusters

#cutree
Autumnclusterd <- cutree(clusterd,5)
```

```
od <- table(ogbaloo$Station,Autumnclusterd)

od <- as.data.frame(od)

clusterm <- as.data.frame(Autumnclusterd)

clusterm$Station <- ogbaloo$Station

pom <- aggregate(od$Freq, by=list(od$Autumnclusterd),FUN=sum)

#pom is stations in each cluster group

#Adding cluster membership to ogbaloo dataset

ogbaloo$cluster <- clusterm$Autumnclusterd
```

#Q5

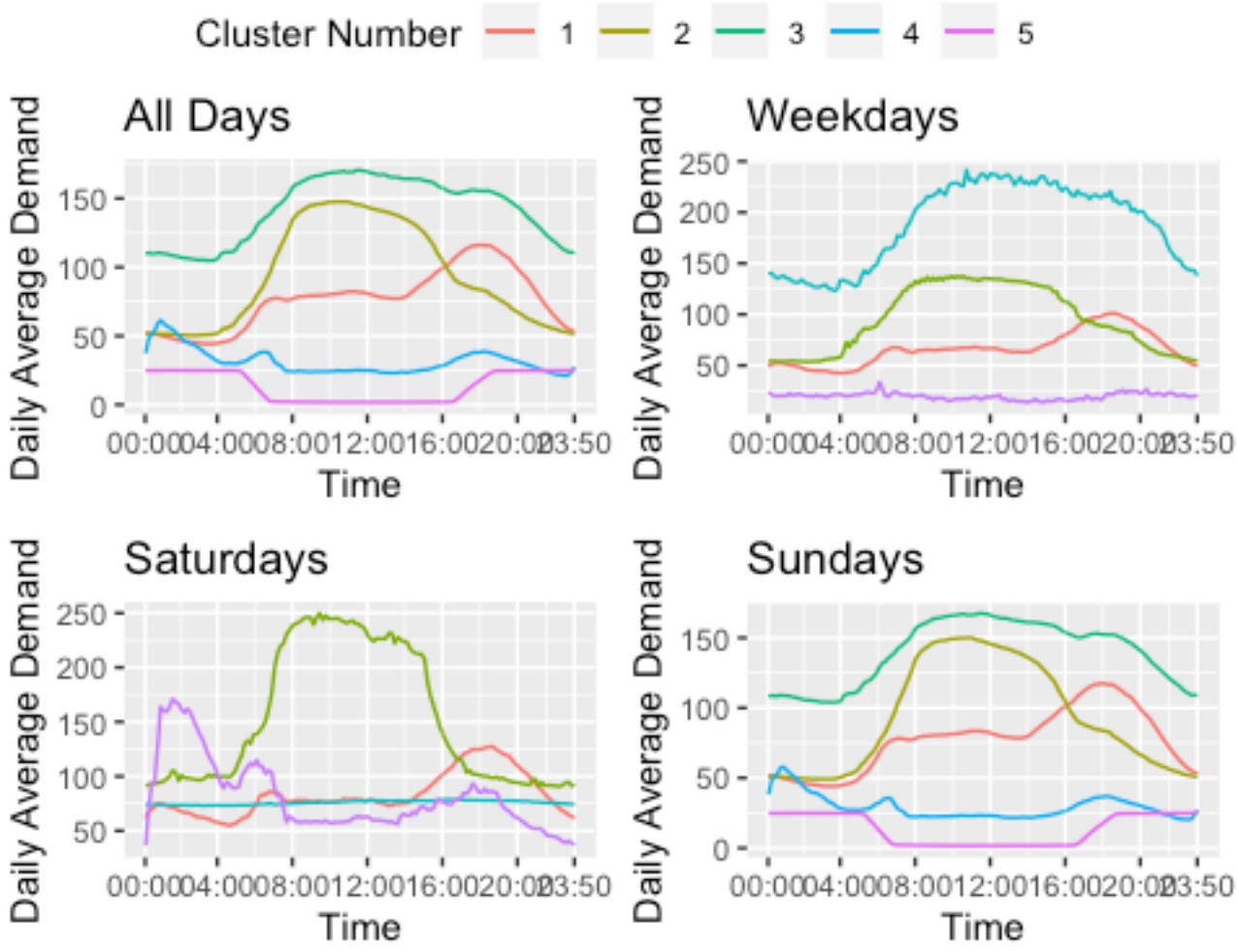
```
#Using non scaled data
t1 <- ogbaloo[148:291]

t1$day <- weekdays(ogbaloo$Date)

t1$cluster <- clusterm$Autumnclusterd

t1$Station <- ogbaloo$Station

t1$Date <- ogbaloo$Date

t1gath <- t1 %>% gather(time, value, -Date, -cluster, -day, -Station)

t1gath$time <- as.numeric(t1gath$time)

#ALL DAYS
tc1ad <- t1gath %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

tc2ad <- t1gath %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

tc3ad <- t1gath %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
```

```
  dplyr::summarise(avgvalue=mean(value))

tc4ad <- t1gath %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

tc5ad <- t1gath %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

tacad <- as.data.frame(c(tc1ad, tc2ad, tc3ad, tc4ad, tc5ad))

tacad <- subset(tacad, select = -c(time.1, time.2, time.3, time.4))

ad <- ggplot(tacad, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='All Days', x='Time', y='Daily Average Demand', colour='Cluster Number')

#WEEKDAYS

t1baloowd <- filter(t1gath, day %in% c('Mon', 'Tue', 'Wed', 'Thur', 'Fri'))

t1c1wd <- t1baloowd %>%
  dplyr::filter(cluster==1) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1c2wd <- t1baloowd %>%
  dplyr::filter(cluster==2) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1c3wd <- t1baloowd %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1c4wd <- t1baloowd %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
```

```
  dplyr::summarise(avgvalue=mean(value))

#cluster 5 has no weekdays

t1c5wd <- t1baloowd %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1acwd <- as.data.frame(c(t1c1wd, t1c2wd, t1c3wd, t1c4wd))

t1acwd <- subset(t1acwd, select = -c(time.1, time.2, time.3))

wd <- ggplot(t1acwd, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='Weekdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Saturdays

t1baloosatu <- filter(t1gath, day %in% ('Sat'))

t1c1sat <- t1baloosatu %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1c2sat <- t1baloosatu %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1c3sat <- t1baloosatu %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1c4sat <- t1baloosatu %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

t1c5sat <- t1baloosatu %>%
  dplyr::filter(cluster==5) %>%
```

```
 group_by(time) %>%
 dplyr::summarise(avgvalue=mean(value))

#cluster 5 has no saturdays

t1acsat <- as.data.frame(c(t1c1sat, t1c2sat, t1c3sat, t1c4sat))

t1acsat <- subset(t1acsat, select = -c(time.1, time.2, time.3))

saturd <- ggplot(t1acsat, aes(x = time, y = 'power demand', colour = 'Cluster')) +
 geom_line(aes(y = avgvalue, col = '1'))  +
 geom_line(aes(y = avgvalue.1, col = '2')) +
 geom_line(aes(y = avgvalue.2, col= '3')) +
 geom_line(aes(y = avgvalue.3, col='4')) +
 scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
 labs(title='Saturdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Sundays

t1baloosun <- filter(t1gath, day %in% ('Sun'))

t1c1sun <- t1baloosun %>%
 dplyr::filter(cluster==1) %>%
 group_by(time) %>%
 dplyr::summarise(avgvalue=mean(value))

t1c2sun <- t1baloosun %>%
 dplyr::filter(cluster==2) %>%
 group_by(time) %>%
 dplyr::summarise(avgvalue=mean(value))

t1c3sun <- t1baloosun %>%
 dplyr::filter(cluster==3) %>%
 group_by(time) %>%
 dplyr::summarise(avgvalue=mean(value))

t1c4sun <- t1baloosun %>%
 dplyr::filter(cluster==4) %>%
 group_by(time) %>%
 dplyr::summarise(avgvalue=mean(value))

t1c5sun <- t1baloosun %>%
 dplyr::filter(cluster==5) %>%
 group_by(time) %>%
 dplyr::summarise(avgvalue=mean(value))

t1acsun <- as.data.frame(c(t1c1sun, t1c2sun, t1c3sun, t1c4sun, t1c5sun))
```

```
t1acsun <- subset(t1acsun, select = -c(time.1, time.2, time.3, time.4))

sun <- ggplot(t1acsun, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='Sundays', x='Time', y='Daily Average Demand', colour='Cluster Number')

ggarrange(ad, wd, saturd, sun, common.legend = T)
```




```
#Q6

Characteristics$TRANSFORMER_TYPE <- as.factor(Characteristics$TRANSFORMER_TYPE)
```

```
t1stc1 <- filter(clusterm, Autumnclusterd=='1')

charclust1 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                     (t1stc1$Station))

t1c2 <- filter(clusterm, Autumnclusterd=='2')

charclust2 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                     (t1c2$Station))

t1c3 <- filter(clusterm, Autumnclusterd=='3')

charclust3 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                     (t1c3$Station))

t1c4 <- filter(clusterm, Autumnclusterd=='4')

charclust4 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                     (t1c4$Station))

t1c5 <- filter(clusterm, Autumnclusterd=='5')

charclust5 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                     (t1c5$Station))

summary(charclust1)

##  SUBSTATION_NUMBER               TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :511029    Grd Mtd Dist. Substation :251     Min.   :  0.00
##  1st Qu.:513516    Pole Mtd Dist. Substation: 29     1st Qu.: 88.75
##  Median :532780                                      Median :152.00
##  Mean   :537190                                      Mean   :156.29
##  3rd Qu.:552834                                      3rd Qu.:216.25
##  Max.   :563737                                      Max.   :485.00
##  Transformer_RATING Percentage_IC     LV_FEEDER_COUNT  GRID_REFERENCE
##  Min.   :   0.0     Min.   :  0.000   Min.   : 0.000   Length:280
##  1st Qu.: 300.0     1st Qu.:  1.055   1st Qu.: 2.750   Class :character
##  Median : 315.0     Median :  8.867   Median : 4.000   Mode  :character
##  Mean   : 394.7     Mean   : 18.100   Mean   : 3.429
##  3rd Qu.: 500.0     3rd Qu.: 23.237   3rd Qu.: 4.000
##  Max.   :1000.0     Max.   :100.000   Max.   :10.000

summary(charclust2)

##  SUBSTATION_NUMBER               TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :511033    Grd Mtd Dist. Substation :65      Min.   :  0.00
##  1st Qu.:513147    Pole Mtd Dist. Substation: 2      1st Qu.:  2.00
##  Median :531313                                      Median : 11.00
```

```
##  Mean   :530842                          Mean   : 31.73
##  3rd Qu.:552034                          3rd Qu.: 37.00
##  Max.   :564285                          Max.   :292.00
##  Transformer_RATING Percentage_IC    LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   :   0.0    Min.   :  0.00   Min.   :0.000   Length:67
##  1st Qu.: 500.0    1st Qu.: 92.55   1st Qu.:1.000   Class :character
##  Median : 500.0    Median : 99.90   Median :2.000   Mode  :character
##  Mean   : 630.3    Mean   : 91.55   Mean   :2.552
##  3rd Qu.: 800.0    3rd Qu.:100.00   3rd Qu.:4.000
##  Max.   :1000.0    Max.   :100.00   Max.   :8.000

summary(charclust3)

##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :511034    Grd Mtd Dist. Substation :35     Min.   :  0.0
##  1st Qu.:512156    Pole Mtd Dist. Substation: 3     1st Qu.:  3.0
##  Median :513298                                     Median : 40.5
##  Mean   :522861                                     Mean   :112.7
##  3rd Qu.:532649                                     3rd Qu.:195.2
##  Max.   :562070                                     Max.   :477.0
##  Transformer_RATING Percentage_IC    LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   : 100.0    Min.   : 16.94   Min.   :0.000   Length:38
##  1st Qu.: 500.0    1st Qu.: 72.82   1st Qu.:1.000   Class :character
##  Median : 500.0    Median : 90.77   Median :4.000   Mode  :character
##  Mean   : 633.2    Mean   : 82.41   Mean   :3.289
##  3rd Qu.:1000.0    3rd Qu.:100.00   3rd Qu.:5.000
##  Max.   :1000.0    Max.   :100.00   Max.   :8.000

summary(charclust4)

##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :512438    Grd Mtd Dist. Substation :12     Min.   :  0.00
##  1st Qu.:513246    Pole Mtd Dist. Substation: 8     1st Qu.:  4.75
##  Median :513422                                     Median : 20.50
##  Mean   :520475                                     Mean   : 41.50
##  3rd Qu.:532227                                     3rd Qu.: 66.50
##  Max.   :535409                                     Max.   :146.00
##  Transformer_RATING Percentage_IC     LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   :  15.0    Min.   :  0.000   Min.   :0.00    Length:20
##  1st Qu.:  50.0    1st Qu.:  0.000   1st Qu.:1.00    Class :character
##  Median : 315.0    Median :  5.383   Median :1.00    Mode  :character
##  Mean   : 410.4    Mean   : 24.445   Mean   :1.75
##  3rd Qu.: 800.0    3rd Qu.: 34.010   3rd Qu.:3.00
##  Max.   :1000.0    Max.   :100.000   Max.   :5.00

summary(charclust5)

##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :531057    Grd Mtd Dist. Substation :0      Min.   :0.0
##  1st Qu.:531644    Pole Mtd Dist. Substation:3      1st Qu.:0.5
```

```
##  Median :532232                          Median :1.0
##  Mean   :531841                          Mean   :1.0
##  3rd Qu.:532234                          3rd Qu.:1.5
##  Max.   :532235                          Max.   :2.0
##  Transformer_RATING Percentage_IC    LV_FEEDER_COUNT  GRID_REFERENCE
##  Min.   : 25.00     Min.   :  0.00   Min.   :0.0000   Length:3
##  1st Qu.: 37.50     1st Qu.: 50.00   1st Qu.:0.5000   Class :character
##  Median : 50.00     Median :100.00   Median :1.0000   Mode  :character
##  Mean   : 58.33     Mean   : 66.67   Mean   :0.6667
##  3rd Qu.: 75.00     3rd Qu.:100.00   3rd Qu.:1.0000
##  Max.   :100.00     Max.   :100.00   Max.   :1.0000
```

#Q7

```
#CLUSTER 1
#According to summary stats cluster 1 has 280 stations
#251 of which are ground mounted, 29 pole mounted so much more urban than rural
#average of 156 customers per station, median of 152
#average transformer rating of 395, median of 315
#On average 18% of customers are industrial and commercial
#average feeders from station is 3.4, max of 10

#pattern from graphs
#all days- power demand is lowest from midnight till 6 am
#it rises from about 50 to 75 by noon and the peaks at around 6pm
# at around 115 before falling quickly until midnight
#weekdays, pattern is very similar as all days but power demand
#does not peak above 100
#saturdays exhibits same pattern but power demand peaks at 125
#sundays pattern is same as all days and same demand values

#Cluster 1 name could be urban domestic customers, as perhaps stations in
#it cater for home electricity
#expaining why demand peaks in evenings

#CLUSTER 2

#summary stats show there are 67 stations in cluster 2, 65 ground mounted, 2 pole
#average of 32 customers per station, average transformer rating of 630- very high
#On average 91.5% of customers are industrial or commercial
#Average feeder count of 2.5

#patterns from graphs
#all days- from about 2am rises from 50 quickly
#to peak at 10 am at about 150 demand
#before falling at slower rate to 50 by midnight
#weekdays shows similar pattern but peaks at 8 am at about 125
#Saturdays similar pattern, but starts from 100
```

```
#rises from 5am and peaks at 8am at 250, falls to 225 at noon
#before rapidly falling back to 100 by 4pm and staying there till next day
#Sunday is same patttern as all days

#Cluster 2 name could be urban industrial and commercial- stations maybe cater to
#city infrastructure and appliances

#CLUSTER 3

#summary stats show 38 stations in cluster 3, 35 ground mounted, 3 pole mounted
#on average, 112 customers per station, average rating of 633 and median of 500
#on average, 82% of customers are industrial and commercial
#average feeder count of 3.3

#graph patterns
#all days- starts from 105, rises at 4 am to peak of 170
#by 10 am, before falling slowly, and then rapidly from 6pm
#to midnight at about 105
#weekdays- similar pattern to all days
#but peaks at about 240 at 11 30 am, and lowest demand
#Saturdays shows a different pattern, w demand largely
#flat and the same during day, at about 75
#sundays is same pattern as all days

#Cluster 3 could also be urban industrial- perhaps stations
#provide electricity for something used on all days but sundays

#CLUSTER 4

#Summary stats show 20 stations in cluster 4, 12 ground mounted, 8 pole mounted
#average of 41.5 customers per station, average transformer rating of 410
#on average 24% of customers are industrial and commercial
#average of 1.75 feeders

#graph patterns
#all days- starts at about 35 at midnight, rises slightly and then falls
#to 28 by 4 am, then rises slightly to 40 by 6 am,
#then falls and stays flat until rising from 4pm to 8 pm
#at 37 before falling again
#weekdays- stays slightly below 25 whole day and demand
#changes regularly but marginally
#Saturday is similar pattern to all days, but higher and more exaggerated demand values
#at same times as all days
#peaks at about 170, but falls to same values as all day
#Sundays is same as all days

#Cluster 4 could be suburban and mainly domestic customers w lower power demands
```

```
#Cluster 5

#Summary stats show only 3 stations in this cluster, all are pole mounted
#average of 1 customer each, average transformer rating of 58
#On average 66% of customers are industrial and commerical
#average feeder count of 0.66

#graph patterns
#all days- starts from 25 til 5 30 am, falls to 0 until 4 30 pm
#rises back to 25
#same for sundays

#Cluster 5 could be rural industrial/commercial
#stations perhaps provide energy for something
#that is automated and takes little energy and only
#turns on on sundays
#
```

#Q8

```
unique(Rawmeasurementsnewstations$Substation)
```

```
## [1] 511079 512457 532697 552863 563729
```

```
Rawmeasurementsnewstations$day <- weekdays(as.Date(Rawmeasurementsnewstations$Date))

lolol <- Rawmeasurementsnewstations %>% gather(time, Value, -Date, -day, -Substation, -X)

lolol$time <- as.numeric(factor(lolol$time))

lolol$time <- as.numeric(lolol$time)

lolol$day <- ordered(lolol$day)

lolol$Date <- as.Date((lolol$Date))

lolol$Substation <- as.numeric(lolol$Substation)

str(lolol)

str(t1gath)

#Station 511079

#ad
ad079 <- lolol %>%
  dplyr::filter(Substation=='511079') %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))
```

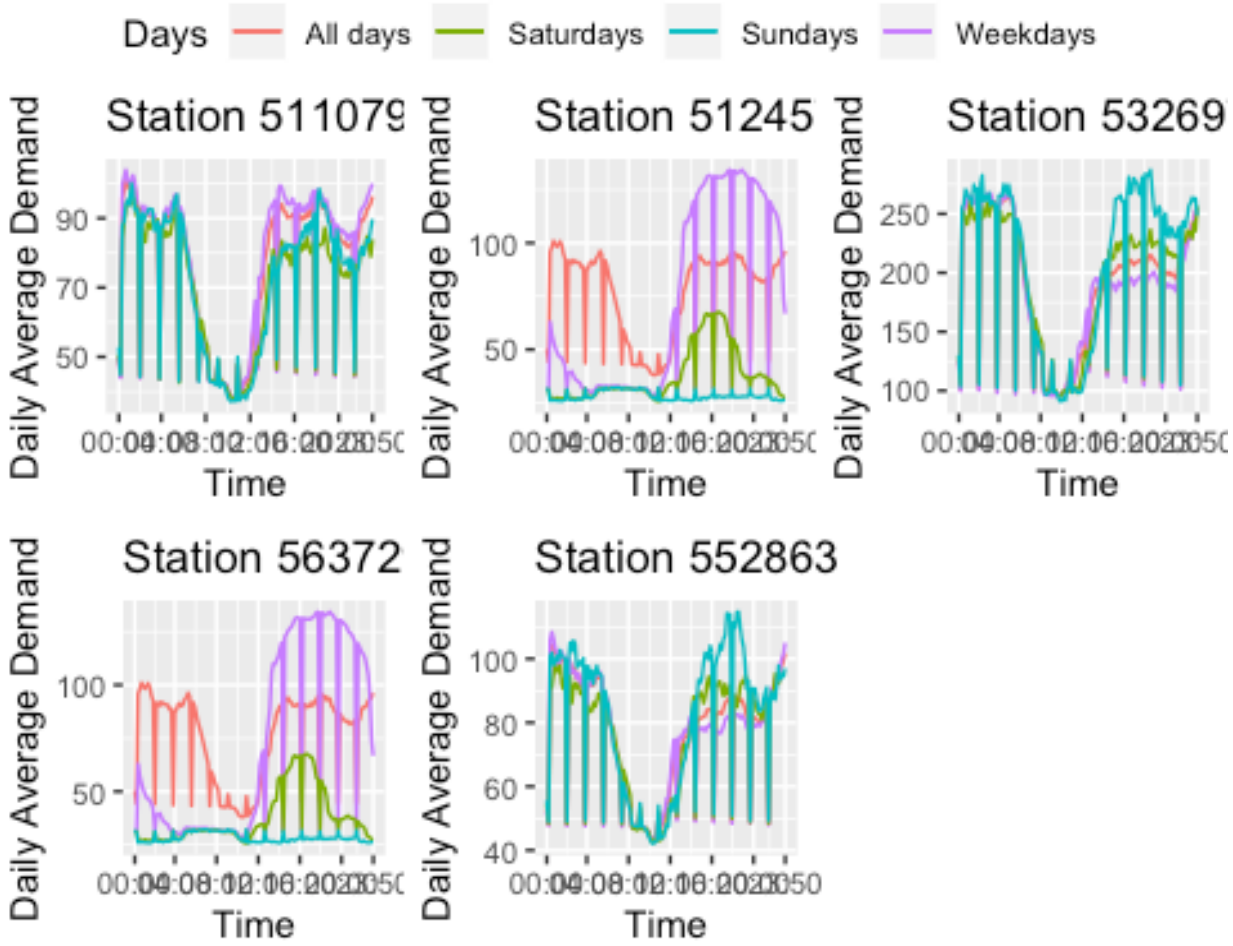
```
#wd
wd079 <- lolol %>%
  dplyr::filter(Substation=='511079') %>%
  dplyr::filter(day %in% c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sat
sat079 <- lolol %>%
  dplyr::filter(Substation=='511079') %>%
  dplyr::filter(day %in% ('Saturday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sun
sun079 <- lolol %>%
  dplyr::filter(Substation=='511079') %>%
  dplyr::filter(day %in% ('Sunday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

statio079 <- as.data.frame(c(ad079, wd079, sat079, sun079))

statio079 <- subset(statio079, select = -c(time.1, time.2, time.3))

Station079 <- ggplot(statio079, aes(x = time, y = 'power demand')) +
  geom_line(aes(y = avgvalue, col = 'All days'))  +
  geom_line(aes(y = avgvalue.1, col = 'Weekdays')) +
  geom_line(aes(y = avgvalue.2, col= 'Saturdays')) +
  geom_line(aes(y = avgvalue.3, col='Sundays')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='Station 511079', x='Time', y='Daily Average Demand', colour='Days')

#Station 512457

#ad
ad457 <- lolol %>%
  dplyr::filter(Substation=='512457') %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#wd
wd457 <- lolol %>%
  dplyr::filter(Substation=='512457') %>%
  dplyr::filter(day %in% c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')) %>%
  group_by(time) %>%
```

```
  dplyr::summarise(avgvalue=mean(Value))

#sat
sat457 <- lolol %>%
  dplyr::filter(Substation=='512457') %>%
  dplyr::filter(day %in% ('Saturday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sun
sun457 <- lolol %>%
  dplyr::filter(Substation=='512457') %>%
  dplyr::filter(day %in% ('Sunday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

stat457 <- as.data.frame(c(ad079, wd457, sat457, sun457))

stat457 <- subset(stat457, select = -c(time.1, time.2, time.3))

Station457 <- ggplot(stat457, aes(x = time, y = 'power demand')) +
  geom_line(aes(y = avgvalue, col = 'All days'))  +
  geom_line(aes(y = avgvalue.1, col = 'Weekdays')) +
  geom_line(aes(y = avgvalue.2, col= 'Saturdays')) +
  geom_line(aes(y = avgvalue.3, col='Sundays')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='Station 512457', x='Time', y='Daily Average Demand', colour='Days')

#Station 532697

#ad
ad697 <- lolol %>%
  dplyr::filter(Substation=='532697') %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#wd
wd697 <- lolol %>%
  dplyr::filter(Substation=='532697') %>%
  dplyr::filter(day %in% c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sat
sat697 <- lolol %>%
  dplyr::filter(Substation=='532697') %>%
  dplyr::filter(day %in% ('Saturday')) %>%
```

```
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sun
sun697 <- lolol %>%
  dplyr::filter(Substation=='532697') %>%
  dplyr::filter(day %in% ('Sunday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

stat697 <- as.data.frame(c(ad697, wd697, sat697, sun697))

stat697 <- subset(stat697, select = -c(time.1, time.2, time.3))

Station697 <- ggplot(stat697, aes(x = time, y = 'power demand')) +
  geom_line(aes(y = avgvalue, col = 'All days'))  +
  geom_line(aes(y = avgvalue.1, col = 'Weekdays')) +
  geom_line(aes(y = avgvalue.2, col= 'Saturdays')) +
  geom_line(aes(y = avgvalue.3, col='Sundays')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='Station 532697', x='Time', y='Daily Average Demand', colour='Days')

#Station 552863

#ad
ad863 <- lolol %>%
  dplyr::filter(Substation=='552863') %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#wd
wd863 <- lolol %>%
  dplyr::filter(Substation=='552863') %>%
  dplyr::filter(day %in% c('Monday', 'Tuesday', 'Wednesdy', 'Thursday', 'Friday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sat
sat863 <- lolol %>%
  dplyr::filter(Substation=='552863') %>%
  dplyr::filter(day %in% ('Saturday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sun
sun863 <- lolol %>%
  dplyr::filter(Substation=='552863') %>%
```

```
  dplyr::filter(day %in% ('Sunday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

stat863 <- as.data.frame(c(ad863, wd863, sat863, sun863))

stat863 <- subset(stat863, select = -c(time.1, time.2, time.3))

Station863 <- ggplot(stat863, aes(x = time, y = 'power demand')) +
  geom_line(aes(y = avgvalue, col = 'All days'))  +
  geom_line(aes(y = avgvalue.1, col = 'Weekdays')) +
  geom_line(aes(y = avgvalue.2, col= 'Saturdays')) +
  geom_line(aes(y = avgvalue.3, col='Sundays')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='Station 552863', x='Time', y='Daily Average Demand', colour='Days')

#Station 563729

#ad
ad729 <- lolol %>%
  dplyr::filter(Substation=='563729') %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#wd
wd729 <- lolol %>%
  dplyr::filter(Substation=='563729') %>%
  dplyr::filter(day %in% c('Monday', 'Tuesday', 'Wednesdy', 'Thursday', 'Friday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sat
sat729 <- lolol %>%
  dplyr::filter(Substation=='563729') %>%
  dplyr::filter(day %in% ('Saturday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

#sun
sun729 <- lolol %>%
  dplyr::filter(Substation=='563729') %>%
  dplyr::filter(day %in% ('Sunday')) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(Value))

stat729 <- as.data.frame(c(ad079, wd457, sat457, sun457))
```

```
stat729 <- subset(stat729, select = -c(time.1, time.2, time.3))

Station729 <- ggplot(stat729, aes(x = time, y = 'power demand')) +
  geom_line(aes(y = avgvalue, col = 'All days'))  +
  geom_line(aes(y = avgvalue.1, col = 'Weekdays')) +
  geom_line(aes(y = avgvalue.2, col= 'Saturdays')) +
  geom_line(aes(y = avgvalue.3, col='Sundays')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00', '
16:00', '20:00', '23:50')) +
  labs(title='Station 563729', x='Time', y='Daily Average Demand', colour='Days')

ggarrange(Station079, Station457, Station697, Station729, Station863, common.legend = T)
```




```
#Q9

#calculating centres of clusters
clust1 <- filter(t1gath, cluster=='1')

clust1centre <- aggregate(clust1$value, list(clust1$time), mean)
```

```
clust2 <- filter(t1gath, cluster=='2')

clust2centre <- aggregate(clust2$value, list(clust2$time), mean)

clust3 <- filter(t1gath, cluster=='3')

clust3centre <- aggregate(clust3$value, list(clust3$time), mean)

clust4 <- filter(t1gath, cluster=='4')

clust4centre <- aggregate(clust4$value, list(clust4$time), mean)

clust5 <- filter(t1gath, cluster=='5')

clust5centre <- aggregate(clust5$value, list(clust5$time), mean)

#Scaling newsubstation data

newdata <- Rawmeasurementsnewstations[4:147]

newdata[, "Daily max"] <- apply(newdata, 1, max)

p <- newdata[,1:144] / newdata[,145]

newstations <- (Rawmeasurementsnewstations$Substation)

p$station <- newstations

pagg <- aggregate(p[,1:144], list(p$station), mean)

unique(pagg$Group.1)

## [1] 511079 512457 532697 552863 563729

stat79 <- pagg[1,]
stat79[,1] <- NULL

stat457 <- pagg[2,]
stat457[,1] <- NULL

stat697 <- pagg[3,]
stat697[,1] <- NULL

stat863 <- pagg[4,]
stat863[,1] <- NULL

stat729 <- pagg[5,]
```

```
stat729[,1] <- NULL

#Distance function
#Station 511079
c1sum79 <- sum((stat729-clust1centre$x)^2)
c2sum79 <- sum((stat729-clust2centre$x)^2)
c3sum79 <- sum((stat729-clust3centre$x)^2)
c4sum79 <- sum((stat729-clust4centre$x)^2)
c5sum79 <- sum((stat729-clust5centre$x)^2)

d79 <- data.frame(c(c1sum79, c2sum79, c3sum79, c4sum79, c5sum79))
clust5centre$x

min(d79)

## [1] 40275.83

#Station 511079 should be allocated to cluster 5

#Repeating dist function for Station 512457

c1sum457 <- sum((stat457-clust1centre$x)^2)
c2sum457 <- sum((stat457-clust2centre$x)^2)
c3sum457 <- sum((stat457-clust3centre$x)^2)
c4sum457 <- sum((stat457-clust4centre$x)^2)
c5sum457 <- sum((stat457-clust5centre$x)^2)

d457 <- data.frame(c(c1sum457, c2sum457, c3sum457, c4sum457, c5sum457))

min(d457)

## [1] 40812.43

#Station 457 should be allocated to cluster 5

#Repeating dist function for Station 532697

c1sum697 <- sum((stat697-clust1centre$x)^2)
c2sum697 <- sum((stat697-clust2centre$x)^2)
c3sum697 <- sum((stat697-clust3centre$x)^2)
c4sum697 <- sum((stat697-clust4centre$x)^2)
c5sum697 <- sum((stat697-clust5centre$x)^2)

d697 <- data.frame(c(c1sum697, c2sum697, c3sum697, c4sum697, c5sum697))

min(d697)

## [1] 40223.26
```

```
#Station 697 should be allocated to cluster 5...

#Repeating dist function for Station 552863

c1sum863 <- sum((stat863-clust1centre$x)^2)
c2sum863 <- sum((stat863-clust2centre$x)^2)
c3sum863 <- sum((stat863-clust3centre$x)^2)
c4sum863 <- sum((stat863-clust4centre$x)^2)
c5sum863 <- sum((stat863-clust5centre$x)^2)

d863 <- data.frame(c(c1sum863, c2sum863, c3sum863, c4sum863, c5sum863))

min(d863)

## [1] 40136.1

#Repeating dist function for Station 563729

c1sum729 <- sum((stat729-clust1centre$x)^2)
c2sum729 <- sum((stat729-clust2centre$x)^2)
c3sum729 <- sum((stat729-clust3centre$x)^2)
c4sum729 <- sum((stat729-clust4centre$x)^2)
c5sum729 <- sum((stat729-clust5centre$x)^2)

d729 <- data.frame(c(c1sum729, c2sum729, c3sum729, c4sum729, c5sum729))

min(d729)

## [1] 40275.83

#This station should also be in cluster 5...

#All new stations should be in cluster 5...

#10

#find new stations in characteristics and compare to summary of cluster 5
newsub1char <- filter(Characteristics, SUBSTATION_NUMBER=='511079')
newsub2char <- filter(Characteristics, SUBSTATION_NUMBER=='512457')
newsub3char <- filter(Characteristics, SUBSTATION_NUMBER=='532697')
newsub4char <- filter(Characteristics, SUBSTATION_NUMBER=='552863')
newsub5char <- filter(Characteristics, SUBSTATION_NUMBER=='563729')

summary(charclust5)

##  SUBSTATION_NUMBER               TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :531057    Grd Mtd Dist. Substation :0      Min.   :0.0
##  1st Qu.:531644    Pole Mtd Dist. Substation:3      1st Qu.:0.5
##  Median :532232                                     Median :1.0
```

```
##  Mean   :531841                          Mean   :1.0
##  3rd Qu.:532234                          3rd Qu.:1.5
##  Max.   :532235                          Max.   :2.0
##  Transformer_RATING Percentage_IC    LV_FEEDER_COUNT  GRID_REFERENCE
##  Min.   : 25.00     Min.   :  0.00   Min.   :0.0000   Length:3
##  1st Qu.: 37.50     1st Qu.: 50.00   1st Qu.:0.5000   Class :character
##  Median : 50.00     Median :100.00   Median :1.0000   Mode  :character
##  Mean   : 58.33     Mean   : 66.67   Mean   :0.6667
##  3rd Qu.: 75.00     3rd Qu.:100.00   3rd Qu.:1.0000
##  Max.   :100.00     Max.   :100.00   Max.   :1.0000
```

```
summary(newsub1char)
```

```
##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :511079    Grd Mtd Dist. Substation :1      Min.   :129
##  1st Qu.:511079    Pole Mtd Dist. Substation:0      1st Qu.:129
##  Median :511079                                     Median :129
##  Mean   :511079                                     Mean   :129
##  3rd Qu.:511079                                     3rd Qu.:129
##  Max.   :511079                                     Max.   :129
##  Transformer_RATING Percentage_IC   LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   :500        Min.   :16.67   Min.   :3       Length:1
##  1st Qu.:500        1st Qu.:16.67   1st Qu.:3       Class :character
##  Median :500        Median :16.67   Median :3       Mode  :character
##  Mean   :500        Mean   :16.67   Mean   :3
##  3rd Qu.:500        3rd Qu.:16.67   3rd Qu.:3
##  Max.   :500        Max.   :16.67   Max.   :3
```

```
summary(newsub2char)
```

```
##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :512457    Grd Mtd Dist. Substation :1      Min.   :15
##  1st Qu.:512457    Pole Mtd Dist. Substation:0      1st Qu.:15
##  Median :512457                                     Median :15
##  Mean   :512457                                     Mean   :15
##  3rd Qu.:512457                                     3rd Qu.:15
##  Max.   :512457                                     Max.   :15
##  Transformer_RATING Percentage_IC LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   :800        Min.   :100   Min.   :3       Length:1
##  1st Qu.:800        1st Qu.:100   1st Qu.:3       Class :character
##  Median :800        Median :100   Median :3       Mode  :character
##  Mean   :800        Mean   :100   Mean   :3
##  3rd Qu.:800        3rd Qu.:100   3rd Qu.:3
##  Max.   :800        Max.   :100   Max.   :3
```

```
summary(newsub3char)
```

```
##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :532697    Grd Mtd Dist. Substation :1      Min.   :313
##  1st Qu.:532697    Pole Mtd Dist. Substation:0      1st Qu.:313
```

```
##  Median :532697                          Median :313
##  Mean   :532697                          Mean   :313
##  3rd Qu.:532697                          3rd Qu.:313
##  Max.   :532697                          Max.   :313
##  Transformer_RATING Percentage_IC   LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   :500        Min.   :1.201   Min.   :5       Length:1
##  1st Qu.:500        1st Qu.:1.201   1st Qu.:5       Class :character
##  Median :500        Median :1.201   Median :5       Mode  :character
##  Mean   :500        Mean   :1.201   Mean   :5
##  3rd Qu.:500        3rd Qu.:1.201   3rd Qu.:5
##  Max.   :500        Max.   :1.201   Max.   :5
```

```
summary(newsub4char)
```

```
##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :552863    Grd Mtd Dist. Substation :0       Min.   :328
##  1st Qu.:552863    Pole Mtd Dist. Substation:1       1st Qu.:328
##  Median :552863                                      Median :328
##  Mean   :552863                                      Mean   :328
##  3rd Qu.:552863                                      3rd Qu.:328
##  Max.   :552863                                      Max.   :328
##  Transformer_RATING Percentage_IC   LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   :200        Min.   :35.29   Min.   :5       Length:1
##  1st Qu.:200        1st Qu.:35.29   1st Qu.:5       Class :character
##  Median :200        Median :35.29   Median :5       Mode  :character
##  Mean   :200        Mean   :35.29   Mean   :5
##  3rd Qu.:200        3rd Qu.:35.29   3rd Qu.:5
##  Max.   :200        Max.   :35.29   Max.   :5
```

```
summary(newsub5char)
```

```
##  SUBSTATION_NUMBER              TRANSFORMER_TYPE TOTAL_CUSTOMERS
##  Min.   :563729    Grd Mtd Dist. Substation :1       Min.   :158
##  1st Qu.:563729    Pole Mtd Dist. Substation:0       1st Qu.:158
##  Median :563729                                      Median :158
##  Mean   :563729                                      Mean   :158
##  3rd Qu.:563729                                      3rd Qu.:158
##  Max.   :563729                                      Max.   :158
##  Transformer_RATING Percentage_IC   LV_FEEDER_COUNT GRID_REFERENCE
##  Min.   :315        Min.   :18.85   Min.   :5       Length:1
##  1st Qu.:315        1st Qu.:18.85   1st Qu.:5       Class :character
##  Median :315        Median :18.85   Median :5       Mode  :character
##  Mean   :315        Mean   :18.85   Mean   :5
##  3rd Qu.:315        3rd Qu.:18.85   3rd Qu.:5
##  Max.   :315        Max.   :18.85   Max.   :5
```

```
#In cluster 5 Summary stats show only 3 stations in this cluster, all are pole mounted
#average of 1 customer each, average transformer rating of 58
#On average 66% of customers are industrial and commerical
#average feeder count of 0.66
```

```
#New Substaion 511079 is ground mounted, and has 129 customers
#with a rating of 500 and and 17% of customers are industrial/commercial
#and a feeder count of 3
#station does not match cluster 5 at all

#New Substation 512457 is also ground mounted, 15 customers, rating of 800
#100% of the customers are industrial and commercial
#Feeder count of 3
#station does not match cluster 5 again really

# New Substation 532697 is ground mounted, 313 customers, rating of 500,
#1.2% of customers are industrial/commercial
#feeder count of 5
#This station also doesnt match cluster 5

#Station 552863 is pole mounted, 328 customers, rating of 200
#35% industrial and commercial
#feeder count of 5
#This station also doesn't seem to match cluster 5, but is at least pole mounted

#Station 563729 is ground mounted, 158 customers, rating of 315
#18% industrial and commercial customers, feeder count of 5
#Station does not match cluster 5

#Cluster Allocation is not as expected at all really
#In summary, my cluster allocation probably hasn't worked but I tried my best...
```

Daanish Ahsan
Applications of Data Science and Statistical Modelling
Candidate Number: 124070

Question 11- Report

Intro

The aim of this analyses of substation power data, was to use clustering on said data and explore if the clusters of stations change over seasons, and if the clusters have common demand profiles. Clustering is a form of unsupervised machine learning, which aims to find meaningful and common features from unlabeled data. Clustering does this by dividing dividing data into groups based on similarity between features. In the case of this report, those features would be the scaled daily power demands per 10 minutes. This scaled data is used so that patterns of demand within days can be determine clusters, rather than the actual data which would result in clusters simply based on magnitude. Agglomerative hierarchical clustering was used throughout this report, as it does not require any prior information, unlike k-means clustering. The analyses in this report uses the "bottom up approach", which merges similar clusters together again and again; this approach is also less sensitive to errors than the alternative "top down approach". The number of clusters chosen throughout this analyses was 5.

Analysis

Figures 1 to 4 are dendrograms of the seasonal scaled substation power data; dendrograms show the hierarchical relationships between data. At first glance they do not seem vastly different between seasons, but the varying heights and distances between clusters does quite clearly change. Spring and High Summer look very similar suggesting that the values they are clustered by are perhaps similar. Interestingly Winter does not look too different from Summer, perhaps reflecting that power demands values are highest at these times. However, it is important not to use dendeograms to assume how many clusters there are. The 5 red rectangles indicate the 5 clusters used from the data, and again it is interesting that High Summer and Winter do not look dissimilar to each other.

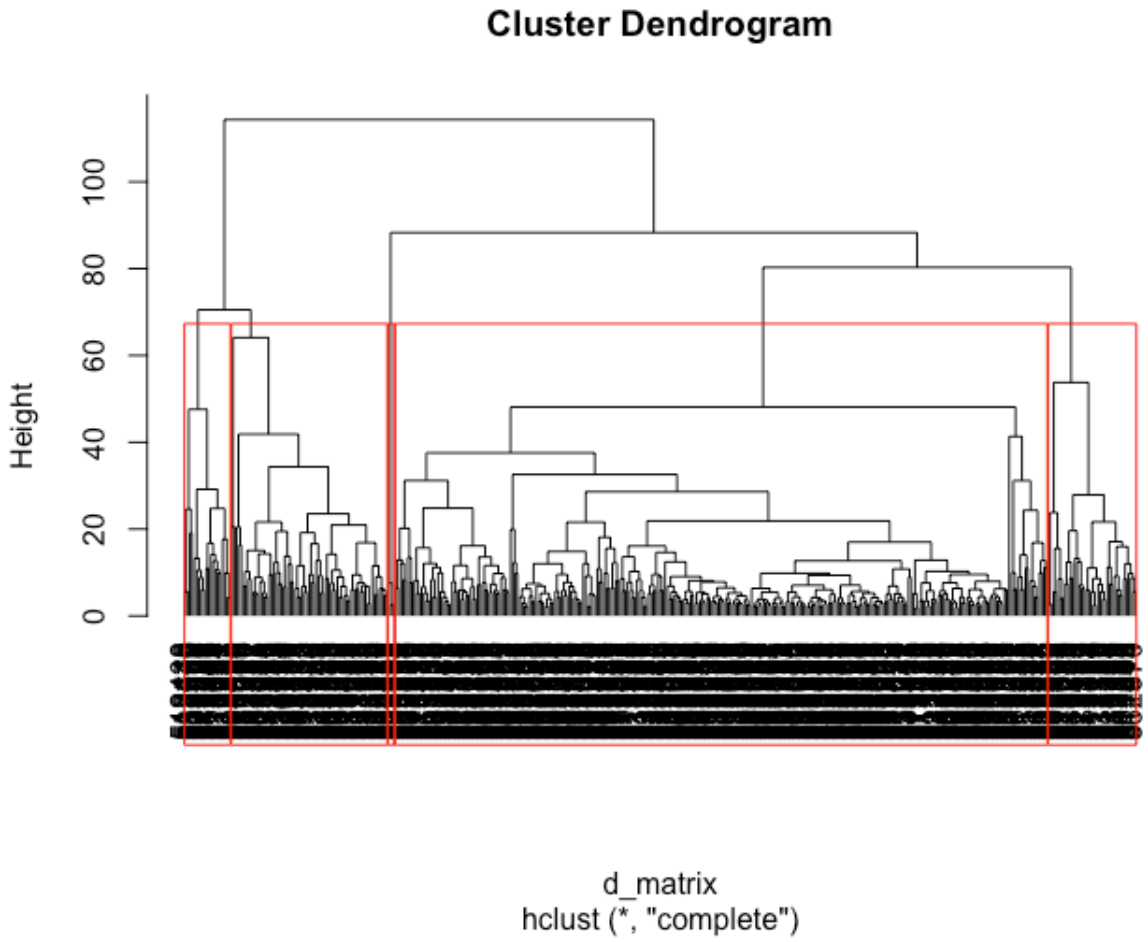


Figure 1, Autumn 2012

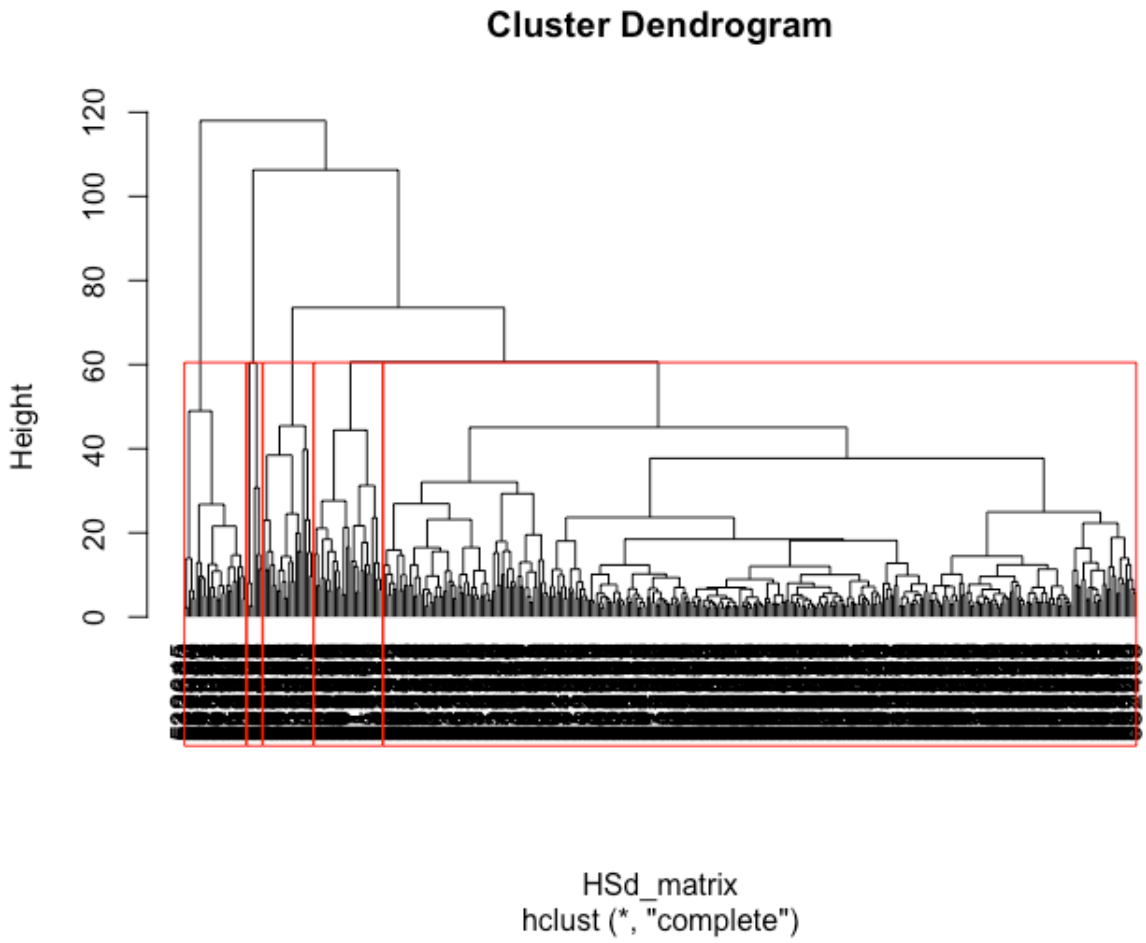


Figure 2, High Summer 2012

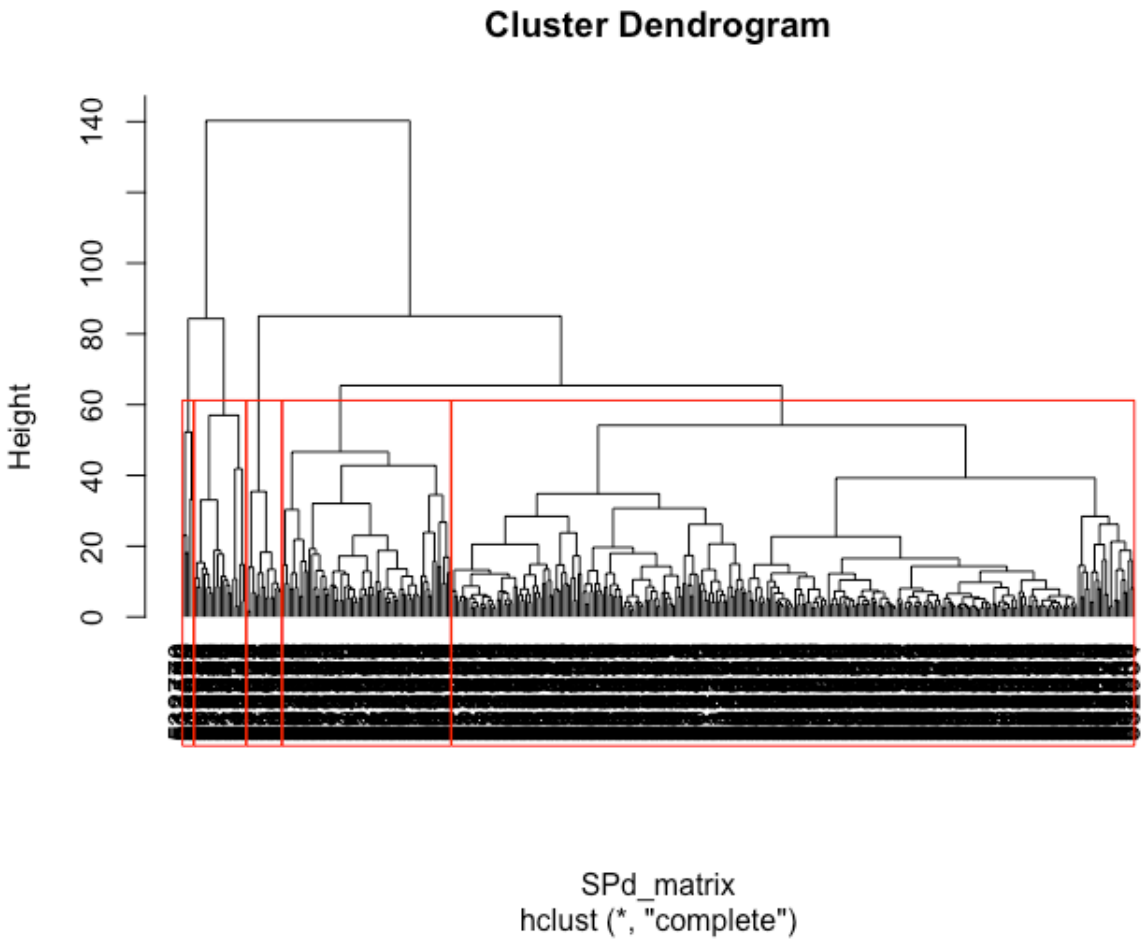


Figure 3, Spring 2013

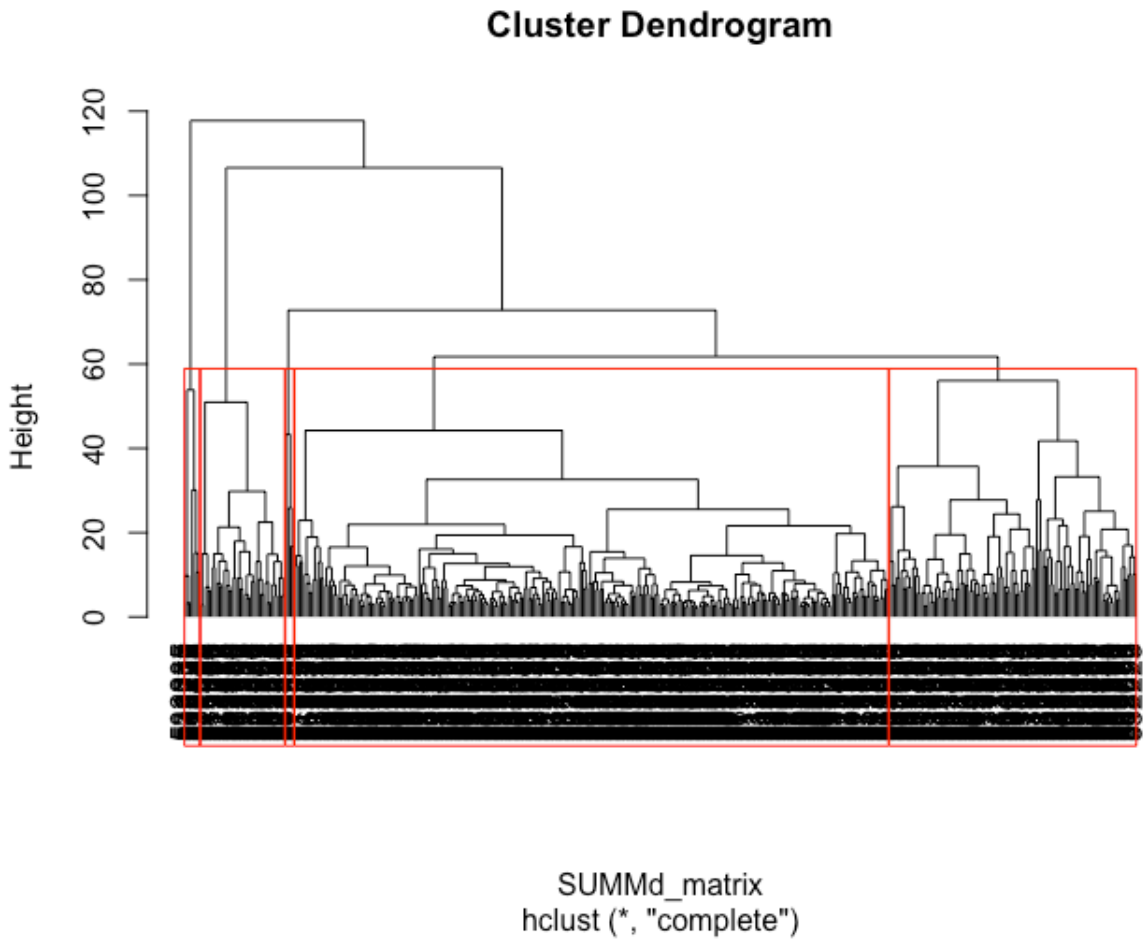


Figure 4, Summer 2012

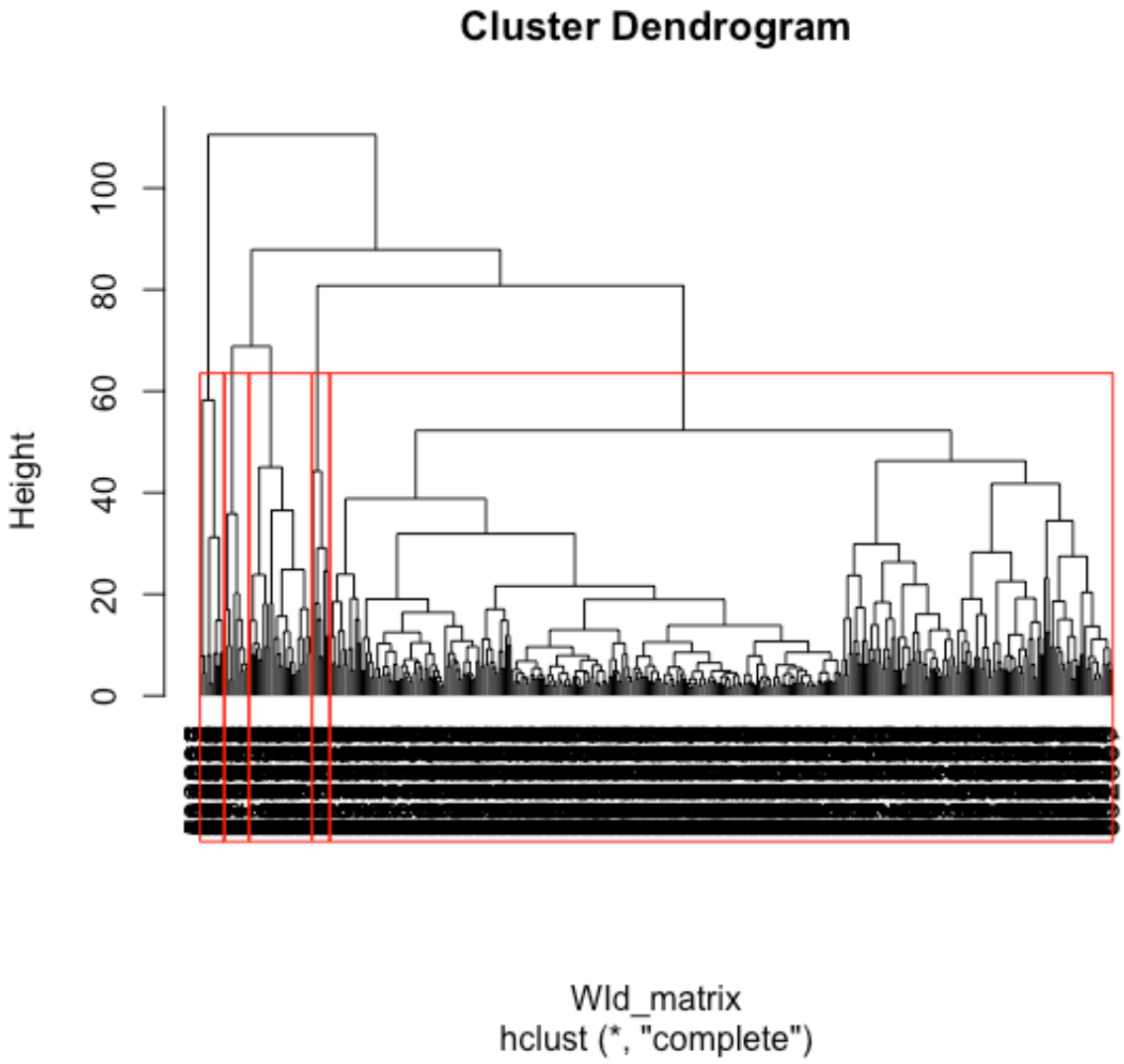


Figure 5, Winter 2012

Cluster Group	Number of Stations
1	287
2	71
3	15
4	22
5	5

Table 1, Spring 2013

Cluster Group	Number of Stations
1	37
2	259
3	108
4	7
5	4

Table 2, Summer 2012

Cluster Group	Number of Stations
1	27
2	326
3	22
4	30
5	7

Table 3, High Summer 2012

Cluster Group	Number of Stations
1	279
2	67
3	38
4	20
5	3

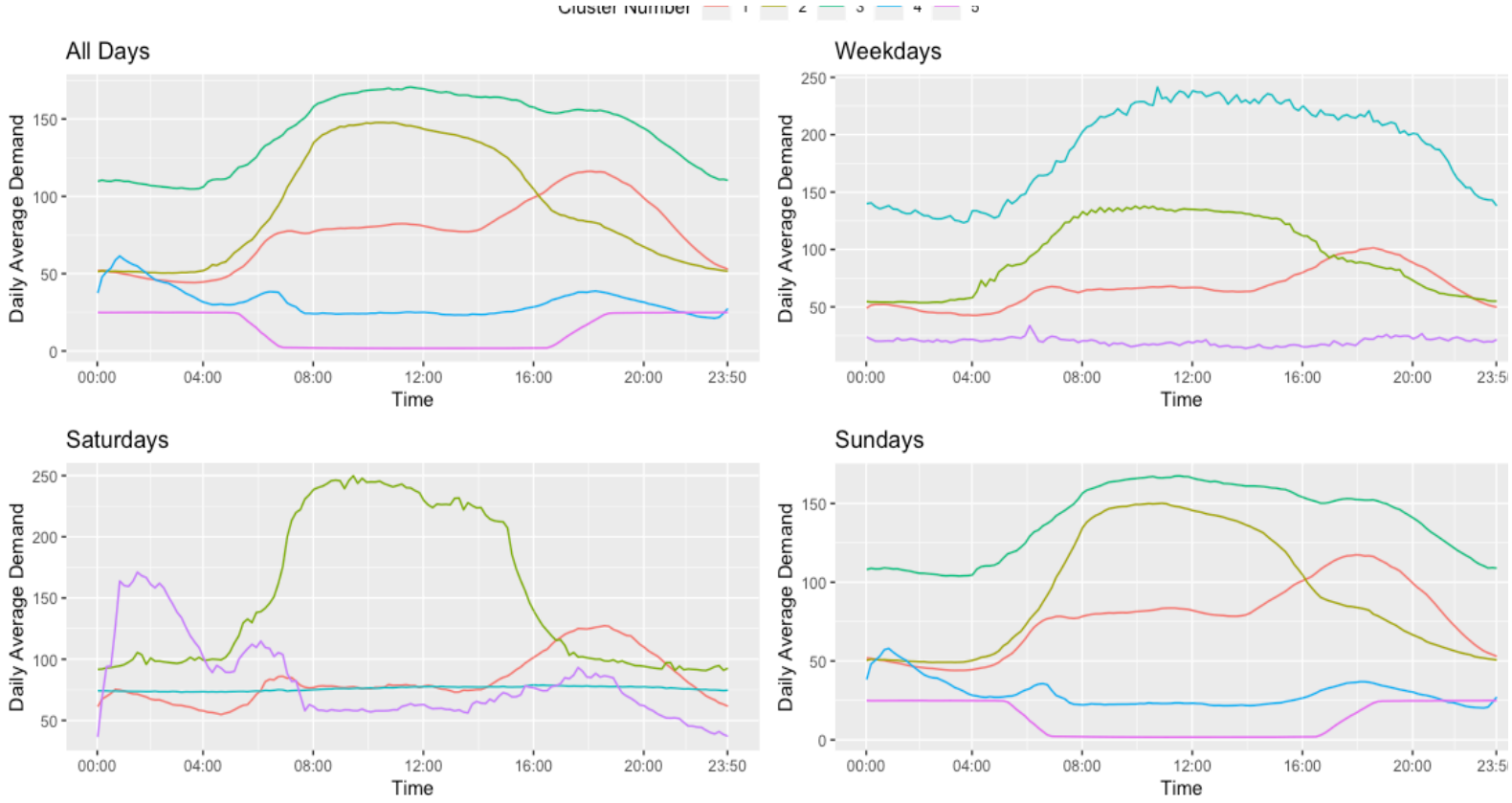
Table 4, Autumn 2012

Cluster Membership	Number of Stations
1	349
2	8
3	11
4	11
5	28

Table 5, Winter 2012

Tables 1 to 5 show how stations change cluster membership over the seasons. There are some interesting trends that are clear. Cluster groups 1 and 2 in every season hold by far the most stations, whilst clusters 4 and 5 always seem to hold the least. This may indicate that the majority of stations between seasons, have similar demand values that place them into the first two clusters. However, it must be taken into account that the hclust function was used on each season's data to create new clusters based off of each seasonal dataset. Thus, cluster 1 in a season is not necessarily representative of cluster 1 in a different season, and the same holds for other equivalent cluster groups across seasons. Nevertheless, it can at least be concluded that in each season the majority of stations are allocated to a certain

Figure 6, Autumn 2012



cluster, which suggests that every season most stations have similar demand

values. In Winter, so many stations being allocated to cluster one could imply how energy demands are generally much higher in Winter, which leads to a vast majority of stations being allocated to the same cluster. Summer shows a closer split of stations between clusters 2 and 3, which suggests that there are two main groups of energy demand; perhaps this split is due to lower energy demands for things like heating, but also higher electricity demand for things like fans in Summer. High Summer similarly to Winter, has the vast majority of stations allocated to a particular cluster, reflecting that power values for most stations are very similar. This could be either due to higher power demand due to usage of fans, ac and sprinklers, or perhaps lower power demand due to a lack of need for heating; it is more likely the latter. Lastly Autumn and Spring show very similar numbers of stations in each of the first two clusters, much like High Summer and Winter do. This is likely because in Autumn and Spring energy needs begin to differ as the lighting and heating are used more in Autumn as days become shorter and temperatures fall, and less in Spring as the opposite happens. As a result, energy use begins to vary among consumers, so the station cluster allocation begins to shift from mainly one cluster, to moving more into another, as seen in comparing Winter or High Summer, with Spring or Autumn.

Figures 6 to 10 show the average daily demand over time, by cluster and season. For the purpose of identifying common clusters across seasons, it may be best to only focus on the all days data so that is what will be done now. Cluster 3 from Figure 6 closely resembles cluster 1 in Figure 7, and cluster 1 in Figure 9, suggesting they may be the same true cluster across the seasons, and hold stations with similar power demands. This cluster could be industrial urban stations, with energy demand fairly high throughout the day. Cluster 2 from Figure 6 resembles cluster 3 from Figure 7 and cluster 3 from Figure 9, and cluster 3 from Figure 8 although it is much exaggerated. This cluster could be representative of stations that power office lights and infrastructure, as demand rises to a peak at 8 am until falling at 4-5pm. Cluster 5 from Figure 6, cluster 5 from Figure 7, cluster 4 from Figure 9 and cluster 5 from figure 10 closely resemble each other- perhaps this cluster could represent stations that power street and motorway lighting, with energy demand being higher at night and near zero during the day. Lastly cluster 4 from Figure 6, cluster 4 from Figure 7, cluster 4 from Figure 9 and cluster 4 but exaggerated from cluster 10 are all fairly similar. This cluster could represent stations which water heating, which is higher at night and in evenings to be ready for the next morning perhaps. By cluster, all days average demand is very similar for Autumn and High Summer, but to a lesser extent for Spring, Summer and Autumn as some clusters show much higher or average daily lower demand.

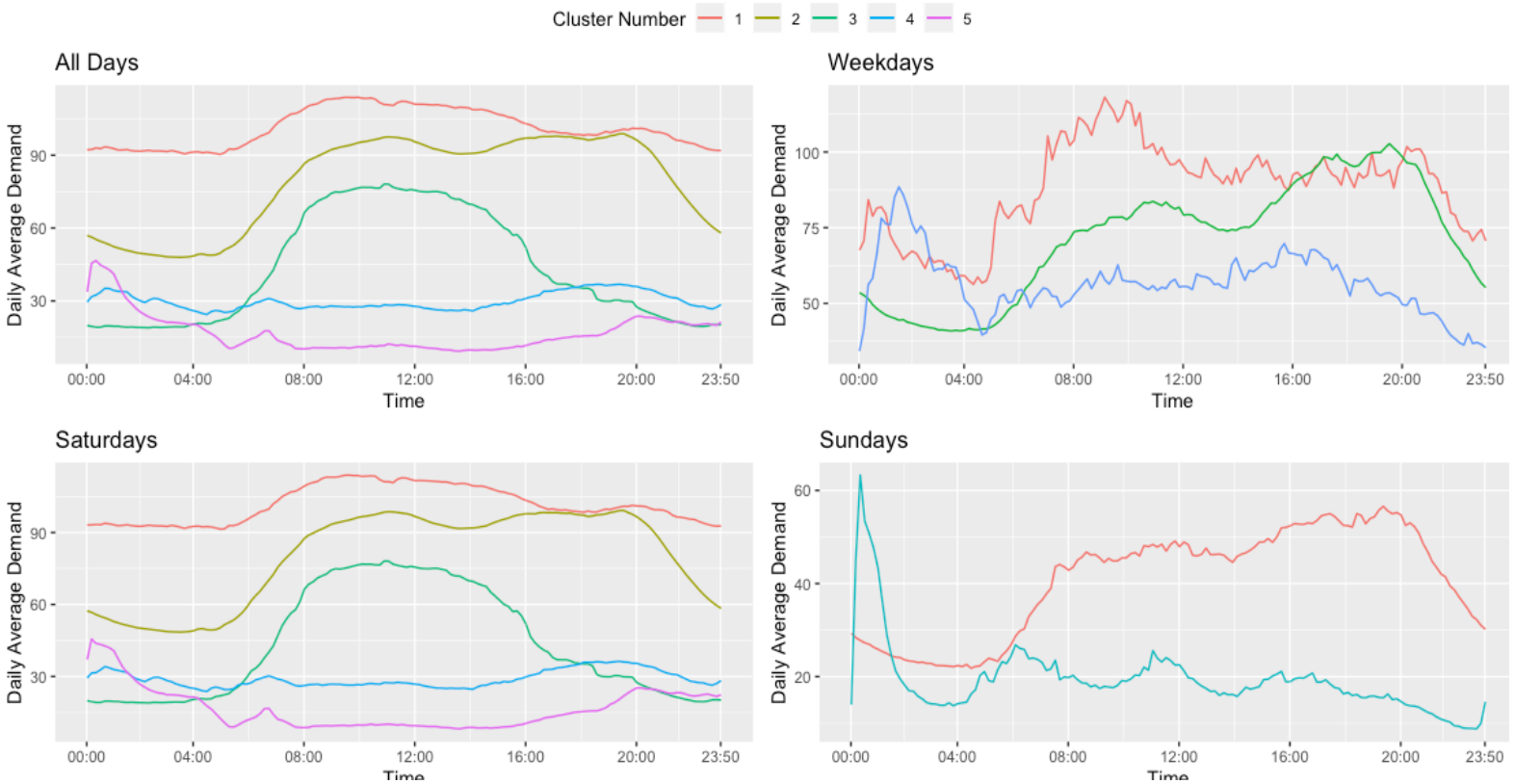


Figure 7, High Summer 2012

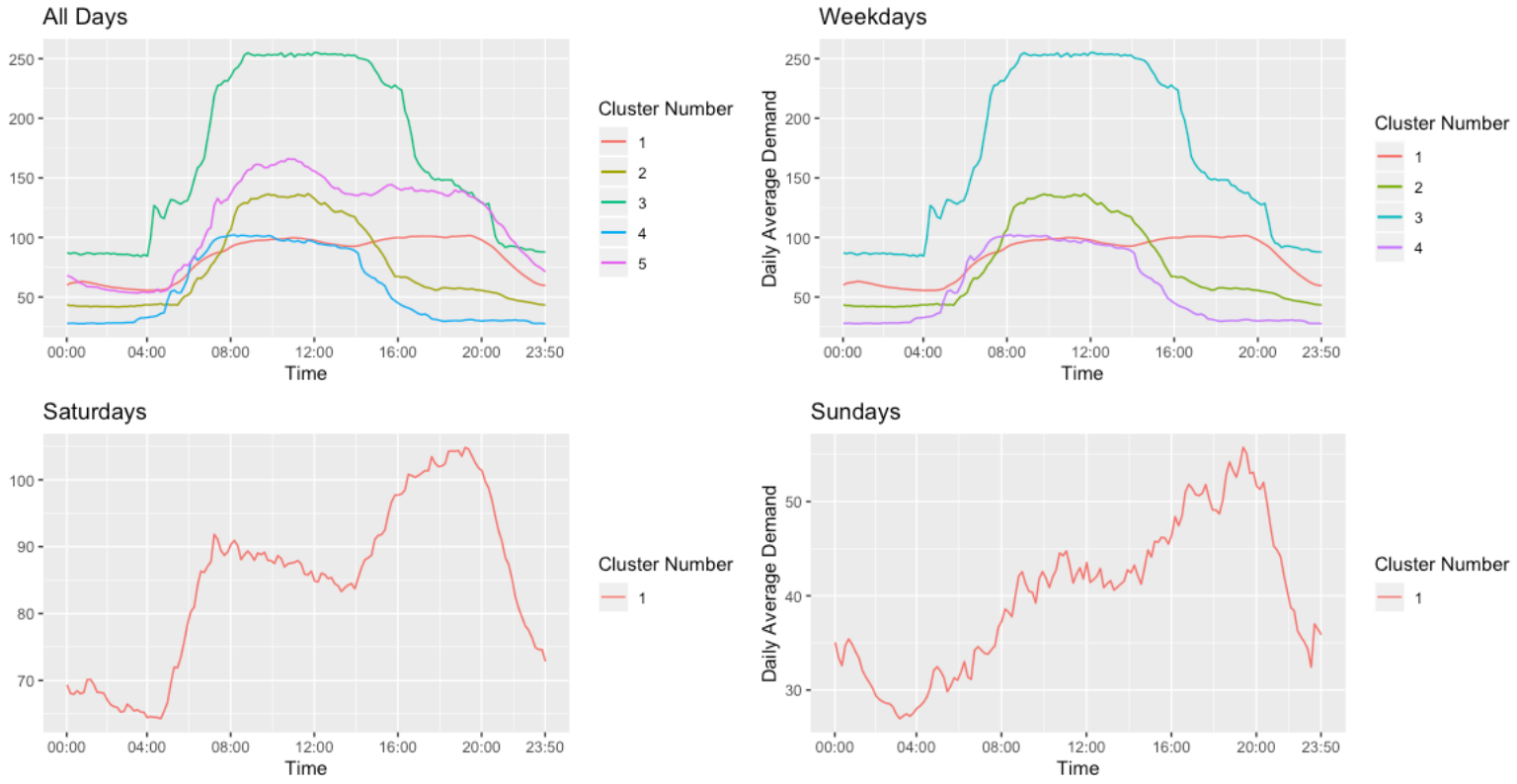


Figure 8, Spring 2013

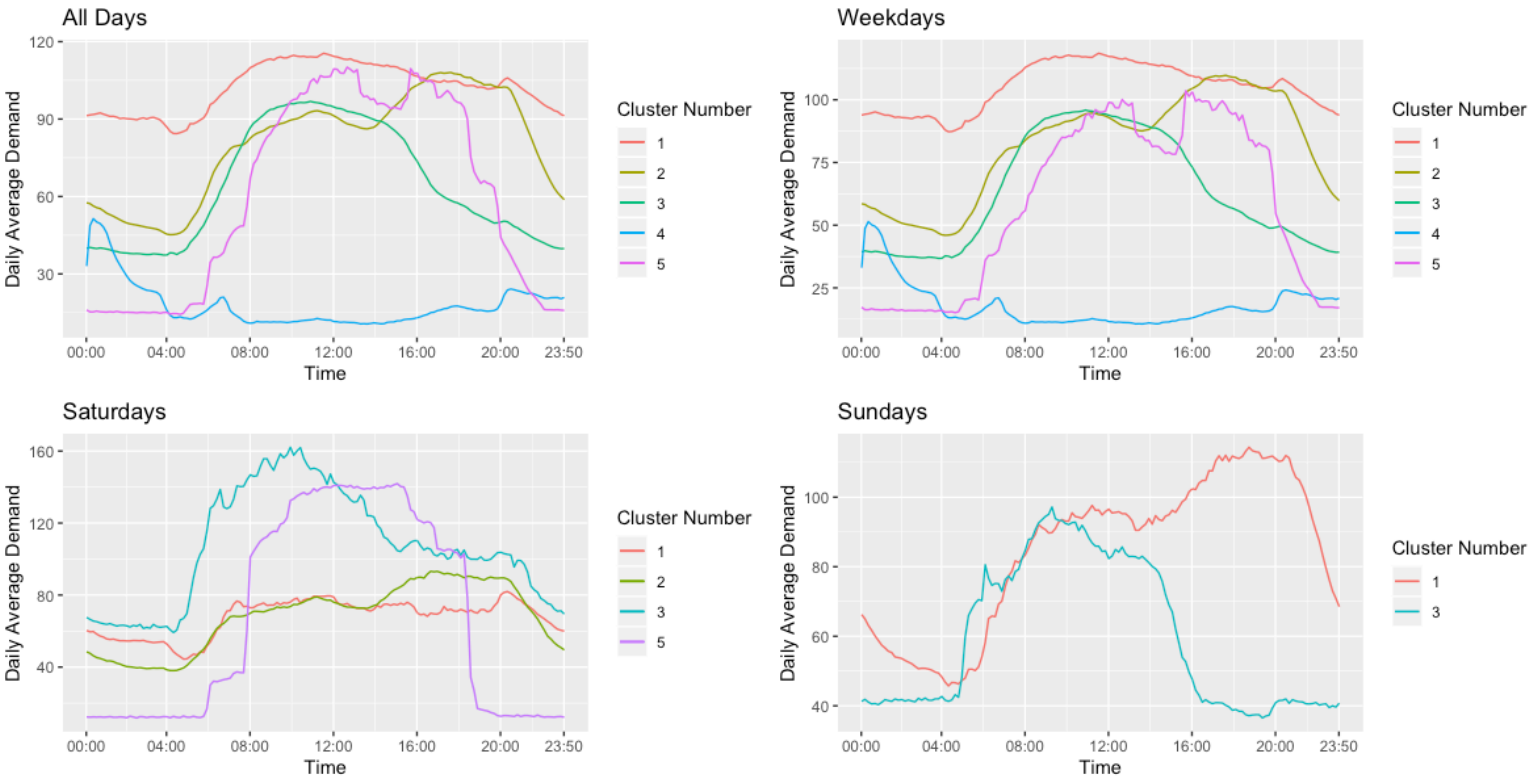


Figure 9, Summer 2012

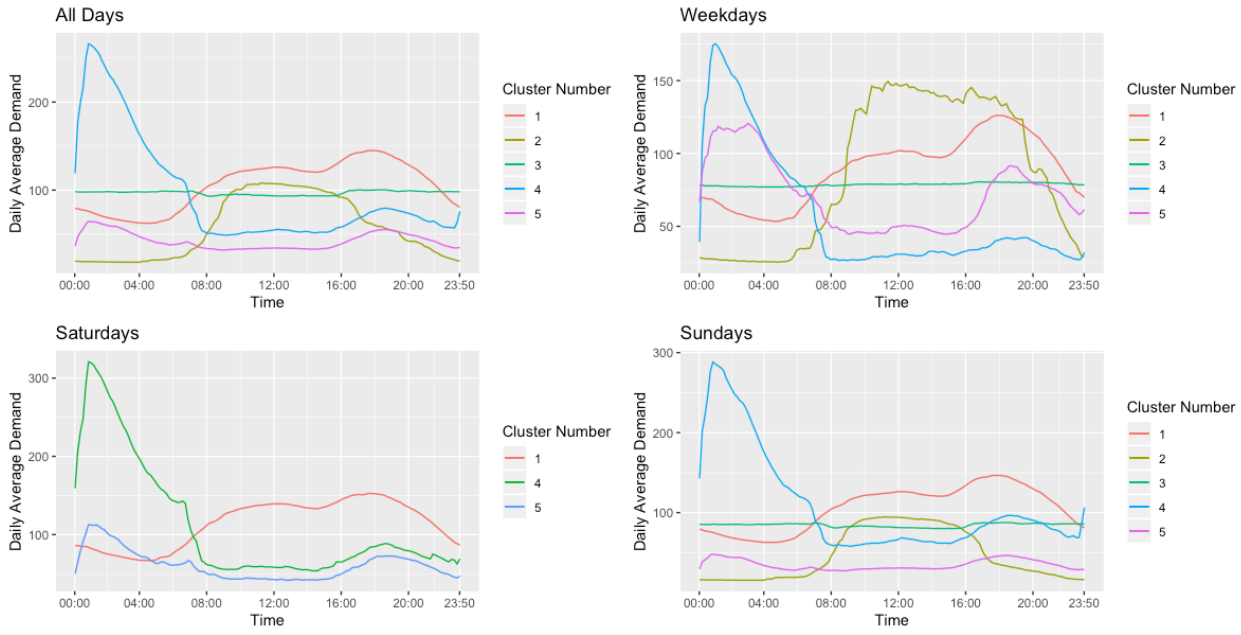


Figure 10, Winter 2012

Conclusion

Overall, average power demands according to figures 6 to 10, do not vary too much by season, although there are some differences by clusters seasonally. Averaging all clusters to compare power demand between seasons may show this more clearly; different patterns of power demand may also vary depending on if scaled or non scaled data is used. It is also clear that stations are allocated into clusters as power demands vary by season, as seen in the tables, but because of new clusters being created for each season's data, it is not easy to tell which exact substations have changed cluster by season. A solution to this may be only clustering once, and then comparing these cluster numbers with those of other seasons but without clustering for them. It may also have been helpful to filter the characteristics dataset by clusters and season to find which station characteristics match, which would imply that the clustering makes sense. In the future, using k-means or perhaps divisive rather than agglomerative hierarchical clustering would make for an interesting comparison to the analyses in this report.

Daanish Ahsan
Applications of Data Science and Statistical Modelling
Candidate Number: 124070

Question 11 Report Code Appendix

```
#Q11- Report
#HIGH SUMMER 2012

HSogbaloo <-
 aggregate(HighSummer_2012[2:ncol(HighSummer_2012)],(HighSummer_2012['Station']), FUN=mean)

HSogbaloo1 <- HSogbaloo[3:146]

HSd_matrix <- dist(rbind(HSogbaloo1),method="manhattan")

HSclusterd <- hclust(HSd_matrix)
#plotting our dendrogram, with hang=-1 to have all labels at same level
plot(HSclusterd, hang=-1, label=HSogbaloo$Station)
rect.hclust(HSclusterd, k=5, border='RED')

HSclusterd <- cutree(HSclusterd,5)
HSod <- table(HSogbaloo$Station,HSclusterd)

HSod <- as.data.frame(HSod)

HSclusterm <- as.data.frame(HSclusterd)

HSclusterm$Station <- HSogbaloo$Station

HSpom <- aggregate(HSod$Freq, by=list(HSod$HSclusterd),FUN=sum)

#pom is stations in each cluster group

#Adding cluster membership to HSogbaloo dataset

HSogbaloo$cluster <- HSclusterm$HSclusterd

#High Summer

HSt1 <- HSogbaloo[148:291]

HSt1$day <- weekdays(HSogbaloo$Date)

HSt1$cluster <- HSclusterm$HSclusterd

HSt1$Station <- HSogbaloo$Station

HSt1$Date <- HSogbaloo$Date

HSt1gath <- HSt1 %>% gather(time, value, -Date, -cluster, -day, -Station)
```

```
HSt1gath$time <- as.numeric(HSt1gath$time)

#ALL DAYS
HStc1ad <- HSt1gath %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HStc2ad <- HSt1gath %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HStc3ad <- HSt1gath %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HStc4ad <- HSt1gath %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HStc5ad <- HSt1gath %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HStacad <- as.data.frame(c(HStc1ad, HStc2ad, HStc3ad, HStc4ad, HStc5ad))

HStacad <- subset(HStacad, select = -c(time.1, time.2, time.3, time.4))

HSad <- ggplot(HStacad, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='All Days', x='Time', y='Daily Average Demand', colour='Cluster Number')

#WEEKDAYS

HSt1baloowd <- filter(HSt1gath, day %in% c('Mon', 'Tue', 'Wed', 'Thur', 'Fri'))

HSt1c1wd <- HSt1baloowd %>%
  dplyr::filter(cluster==1) %>%
```

```
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c2wd <- HSt1baloowd %>%
  dplyr::filter(cluster==2) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c3wd <- HSt1baloowd %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c4wd <- HSt1baloowd %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c5wd <- HSt1baloowd %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

#cluster 3 and 5 have no weekdays

HSt1acwd <- as.data.frame(c(HSt1c1wd, HSt1c2wd, HSt1c4wd))

HSt1acwd <- subset(HSt1acwd, select = -c(time.1, time.2))

HSwd <- ggplot(HSt1acwd, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '4')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Weekdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Saturdays

HSt1baloosatu <- filter(HSt1gath, day %in% ('Sat'))

HSt1c1sat <- HSt1baloosatu %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c2sat <- HSt1baloosatu %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
```

```
  dplyr::summarise(avgvalue=mean(value))

HSt1c3sat <- HSt1baloosatu %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c4sat <- HSt1baloosatu %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c5sat <- HSt1baloosatu %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1acsat <- as.data.frame(c(HSt1c1sat, HSt1c2sat, HSt1c3sat, HSt1c4sat, HSt1c5sat))

HSt1acsat <- subset(HSt1acsat, select = -c(time.1, time.2, time.3, time.4))

HSsaturd <- ggplot(HSt1acsat, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Saturdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Sundays

HSt1baloosun <- filter(HSt1gath, day %in% ('Sun'))

HSt1c1sun <- HSt1baloosun %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c2sun <- HSt1baloosun %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c3sun <- HSt1baloosun %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))
```

```
HSt1c4sun <- HSt1baloosun %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

HSt1c5sun <- HSt1baloosun %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

#Clusters 1, 3, 4 have no sundays

HSt1acsun <- as.data.frame(c(HSt1c2sun, HSt1c5sun))

HSt1acsun <- subset(HSt1acsun, select = -c(time.1))

HSsun <- ggplot(HSt1acsun, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Sundays', x='Time', y='Daily Average Demand', colour='Cluster Number')


ggarrange(HSad, HSwd, HSsaturd, HSsun, common.legend = T)


#SPRING 2013


SPogbaloo <- aggregate(Spring_2013[2:ncol(Spring_2013)],(Spring_2013['Station']), FUN=mean)

SPogbaloo1 <- SPogbaloo[3:146]

SPd_matrix <- dist(rbind(SPogbaloo1),method="manhattan")

SPclusterd <- hclust(SPd_matrix)
#plotting our dendrogram, with hang=-1 to have all labels at same level
plot(SPclusterd, hang=-1, label=SPogbaloo$Station)
rect.hclust(SPclusterd, k=5, border='RED')



SPclusterd <- cutree(SPclusterd,5)
SPod <- table(SPogbaloo$Station,SPclusterd)

SPod <- as.data.frame(SPod)
```

```
SPclusterm <- as.data.frame(SPclusterd)

SPclusterm$Station <- SPogbaloo$Station

SPpom <- aggregate(SPod$Freq, by=list(SPod$SPclusterd),FUN=sum)

#pom is stations in each cluster group

#Adding cluster membership to ogbaloo dataset

SPogbaloo$cluster <- SPclusterm$SPclusterd

#SPRING

SPt1 <- SPogbaloo[148:291]

SPt1$day <- weekdays(SPogbaloo$Date)

SPt1$cluster <- SPclusterm$SPclusterd

SPt1$Station <- SPogbaloo$Station

SPt1$Date <- SPogbaloo$Date

SPt1gath <- SPt1 %>% gather(time, value, -Date, -cluster, -day, -Station)

SPt1gath$time <- as.numeric(SPt1gath$time)

#ALL DAYS
SPtc1ad <- SPt1gath %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPtc2ad <- SPt1gath %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPtc3ad <- SPt1gath %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPtc4ad <- SPt1gath %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))
```

```
SPtc5ad <- SPt1gath %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPtacad <- as.data.frame(c(SPtc1ad, SPtc2ad, SPtc3ad, SPtc4ad, SPtc5ad))

SPtacad <- subset(SPtacad, select = -c(time.1, time.2, time.3, time.4))

SPad <- ggplot(SPtacad, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='All Days', x='Time', y='Daily Average Demand', colour='Cluster Number')

#WEEKDAYS

SPt1baloowd <- filter(SPt1gath, day %in% c('Mon', 'Tue', 'Wed', 'Thur', 'Fri'))

SPt1c1wd <- SPt1baloowd %>%
  dplyr::filter(cluster==1) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c2wd <- SPt1baloowd %>%
  dplyr::filter(cluster==2) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c3wd <- SPt1baloowd %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c4wd <- SPt1baloowd %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c5wd <- SPt1baloowd %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1acwd <- as.data.frame(c(SPt1c1wd, SPt1c2wd, SPt1c3wd, SPt1c4wd))
```

```
SPt1acwd <- subset(SPt1acwd, select = -c(time.1, time.2, time.3))

SPwd <- ggplot(SPt1acwd, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Weekdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Saturdays

SPt1baloosatu <- filter(SPt1gath, day %in% ('Sat'))

SPt1c1sat <- SPt1baloosatu %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c2sat <- SPt1baloosatu %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c3sat <- SPt1baloosatu %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c4sat <- SPt1baloosatu %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c5sat <- SPt1baloosatu %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

#Only cluster 1 has saturdays

SPt1acsat <- as.data.frame(c(SPt1c1sat))

SPsaturd <- ggplot(SPt1acsat, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
```

```
  labs(title='Saturdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Sundays

SPt1baloosun <- filter(SPt1gath, day %in% ('Sun'))

SPt1c1sun <- SPt1baloosun %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c2sun <- SPt1baloosun %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c3sun <- SPt1baloosun %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c4sun <- SPt1baloosun %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SPt1c5sun <- SPt1baloosun %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

#Only cluster one has sundays

SPt1acsun <- as.data.frame(SPt1c1sun)

SPsun <- ggplot(SPt1acsun, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Sundays', x='Time', y='Daily Average Demand', colour='Cluster Number')

ggarrange(SPad, SPwd, SPsaturd, SPsun, common.legend = T)

SUMMogbaloo <- aggregate(Summer_2012[2:ncol(Summer_2012)],(Summer_2012['Station']),
 FUN=mean)

SUMMogbaloo1 <- SUMMogbaloo[3:146]

SUMMd_matrix <- dist(rbind(SUMMogbaloo1),method="manhattan")
```

```
SUMMclusterd <- hclust(SUMMd_matrix)
#plotting our dendrogram, with hang=-1 to have all labels at same level
plot(SUMMclusterd, hang=-1, label=SUMMogbaloo$Station)
rect.hclust(SUMMclusterd, k=5, border='RED')

SUMMclusterd <- cutree(SUMMclusterd,5)
SUMMod <- table(SUMMogbaloo$Station,SUMMclusterd)

SUMMod <- as.data.frame(SUMMod)

SUMMclusterm <- as.data.frame(SUMMclusterd)

SUMMclusterm$Station <- SUMMogbaloo$Station

SUMMpom <- aggregate(SUMMod$Freq, by=list(SUMMod$SUMMclusterd),FUN=sum)

#pom is stations in each cluster group

#Adding cluster membership to ogbaloo dataset

SUMMogbaloo$cluster <- SUMMclusterm$SUMMclusterd

#SUMMER 2012

SUMMt1 <- SUMMogbaloo[148:291]

SUMMt1$day <- weekdays(SUMMogbaloo$Date)

SUMMt1$cluster <- SUMMclusterm$SUMMclusterd

SUMMt1$Station <- SUMMogbaloo$Station

SUMMt1$Date <- SUMMogbaloo$Date

SUMMt1gath <- SUMMt1 %>% gather(time, value, -Date, -cluster, -day, -Station)

SUMMt1gath$time <- as.numeric(SUMMt1gath$time)

#ALL DAYS
SUMMtc1ad <- SUMMt1gath %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMtc2ad <- SUMMt1gath %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))
```

```
SUMMtc3ad <- SUMMt1gath %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMtc4ad <- SUMMt1gath %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMtc5ad <- SUMMt1gath %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMtacad <- as.data.frame(c(SUMMtc1ad, SUMMtc2ad, SUMMtc3ad, SUMMtc4ad, SUMMtc5ad))

SUMMtacad <- subset(SUMMtacad, select = -c(time.1, time.2, time.3, time.4))

SUMMad <- ggplot(SUMMtacad, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='All Days', x='Time', y='Daily Average Demand', colour='Cluster Number')

#WEEKDAYS

SUMMt1baloowd <- filter(SUMMt1gath, day %in% c('Mon', 'Tue', 'Wed', 'Thur', 'Fri'))

SUMMt1c1wd <- SUMMt1baloowd %>%
  dplyr::filter(cluster==1) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c2wd <- SUMMt1baloowd %>%
  dplyr::filter(cluster==2) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c3wd <- SUMMt1baloowd %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))
```

```
SUMMt1c4wd <- SUMMt1baloowd %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c5wd <- SUMMt1baloowd %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1acwd <- as.data.frame(c(SUMMt1c1wd, SUMMt1c2wd, SUMMt1c3wd ,SUMMt1c4wd,
 SUMMt1c5wd))

SUMMt1acwd <- subset(SUMMt1acwd, select = -c(time.1, time.2, time.3, time.4))

SUMMwd <- ggplot(SUMMt1acwd, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Weekdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Saturdays

SUMMt1baloosatu <- filter(SUMMt1gath, day %in% ('Sat'))

SUMMt1c1sat <- SUMMt1baloosatu %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c2sat <- SUMMt1baloosatu %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c3sat <- SUMMt1baloosatu %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c4sat <- SUMMt1baloosatu %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))
```

```
SUMMt1c5sat <- SUMMt1baloosatu %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

#Cluster 4 has no Saturdays

SUMMt1acsat <- as.data.frame(c(SUMMt1c1sat, SUMMt1c2sat, SUMMt1c3sat, SUMMt1c5sat))

SUMMt1acsat <- subset(SUMMt1acsat, select = -c(time.1, time.2, time.3))

SUMMsaturd <- ggplot(SUMMt1acsat, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Saturdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Sundays

SUMMt1baloosun <- filter(SUMMt1gath, day %in% ('Sun'))

SUMMt1c1sun <- SUMMt1baloosun %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c2sun <- SUMMt1baloosun %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c3sun <- SUMMt1baloosun %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c4sun <- SUMMt1baloosun %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

SUMMt1c5sun <- SUMMt1baloosun %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))
```

```
#Clusters 1, 4, 5 have no sundays

SUMMt1acsun <- as.data.frame(c(SUMMt1c2sun, SUMMt1c3sun))

SUMMt1acsun <- subset(SUMMt1acsun, select = -c(time.1))

SUMMsun <- ggplot(SUMMt1acsun, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '3')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Sundays', x='Time', y='Daily Average Demand', colour='Cluster Number')

ggarrange(SUMMad, SUMMwd, SUMMsaturd, SUMMsun, common.legend = T)

#WINTER 2012

WIogbaloo <- aggregate(Winter_2012[2:ncol(Winter_2012)],(Winter_2012['Station']), FUN=mean)

WIogbaloo1 <- WIogbaloo[3:146]

WId_matrix <- dist(rbind(WIogbaloo1),method="manhattan")

WIclusterd <- hclust(WId_matrix)
#plotting our dendrogram, with hang=-1 to have all labels at same level
plot(WIclusterd, hang=-1, label=WIogbaloo$Station)
rect.hclust(WIclusterd, k=5, border='RED')



WIclusterd <- cutree(WIclusterd,5)
WIod <- table(WIogbaloo$Station,WIclusterd)

WIod <- as.data.frame(WIod)

WIclusterm <- as.data.frame(WIclusterd)

WIclusterm$Station <- WIogbaloo$Station

WIpom <- aggregate(WIod$Freq, by=list(WIod$WIclusterd),FUN=sum)

#pom is stations in each cluster group

#Adding cluster membership to ogbaloo dataset

WIogbaloo$cluster <- WIclusterm$WIclusterd

#WINTER 2012
```

```
WIt1 <- WIogbaloo[148:291]

WIt1$day <- weekdays(WIogbaloo$Date)

WIt1$cluster <- WIclusterm$WIclusterd

WIt1$Station <- WIogbaloo$Station

WIt1$Date <- WIogbaloo$Date

WIt1gath <- WIt1 %>% gather(time, value, -Date, -cluster, -day, -Station)

WIt1gath$time <- as.numeric(WIt1gath$time)

#ALL DAYS
WItc1ad <- WIt1gath %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WItc2ad <- WIt1gath %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WItc3ad <- WIt1gath %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WItc4ad <- WIt1gath %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WItc5ad <- WIt1gath %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WItacad <- as.data.frame(c(WItc1ad, WItc2ad, WItc3ad, WItc4ad, WItc5ad))

WItacad <- subset(WItacad, select = -c(time.1, time.2, time.3, time.4))

WIad <- ggplot(WItacad, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
```

```
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='All Days', x='Time', y='Daily Average Demand', colour='Cluster Number')

#WEEKDAYS

WIt1baloowd <- filter(WIt1gath, day %in% c('Mon', 'Tue', 'Wed', 'Thur', 'Fri'))

WIt1c1wd <- WIt1baloowd %>%
  dplyr::filter(cluster==1) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c2wd <- WIt1baloowd %>%
  dplyr::filter(cluster==2) %>%
  dplyr::group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c3wd <- WIt1baloowd %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c4wd <- WIt1baloowd %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c5wd <- WIt1baloowd %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1acwd <- as.data.frame(c(WIt1c1wd, WIt1c2wd, WIt1c3wd ,WIt1c4wd, WIt1c5wd))

WIt1acwd <- subset(WIt1acwd, select = -c(time.1, time.2, time.3, time.4))

WIwd <- ggplot(WIt1acwd, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col= '3')) +
  geom_line(aes(y = avgvalue.3, col='4')) +
  geom_line(aes(y = avgvalue.4, col='5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Weekdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Saturdays
```

```
WIt1baloosatu <- filter(WIt1gath, day %in% ('Sat'))

WIt1c1sat <- WIt1baloosatu %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c2sat <- WIt1baloosatu %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c3sat <- WIt1baloosatu %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c4sat <- WIt1baloosatu %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c5sat <- WIt1baloosatu %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

#Cluster 2, 3 have no Saturdays

WIt1acsat <- as.data.frame(c(WIt1c1sat, WIt1c4sat, WIt1c5sat))

WIt1acsat <- subset(WIt1acsat, select = -c(time.1, time.2))

WIsaturd <- ggplot(WIt1acsat, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '4')) +
  geom_line(aes(y = avgvalue.2, col= '5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Saturdays', x='Time', y='Daily Average Demand', colour='Cluster Number')

#Sundays

WIt1baloosun <- filter(WIt1gath, day %in% ('Sun'))

WIt1c1sun <- WIt1baloosun %>%
  dplyr::filter(cluster==1) %>%
  group_by(time) %>%
```

```
  dplyr::summarise(avgvalue=mean(value))

WIt1c2sun <- WIt1baloosun %>%
  dplyr::filter(cluster==2) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c3sun <- WIt1baloosun %>%
  dplyr::filter(cluster==3) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c4sun <- WIt1baloosun %>%
  dplyr::filter(cluster==4) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1c5sun <- WIt1baloosun %>%
  dplyr::filter(cluster==5) %>%
  group_by(time) %>%
  dplyr::summarise(avgvalue=mean(value))

WIt1acsun <- as.data.frame(c(WIt1c1sun, WIt1c2sun, WIt1c3sun, WIt1c4sun, WIt1c5sun))

WIt1acsun <- subset(WIt1acsun, select = -c(time.1, time.2, time.3, time.4))

WIsun <- ggplot(WIt1acsun, aes(x = time, y = 'power demand', colour = 'Cluster')) +
  geom_line(aes(y = avgvalue, col = '1'))  +
  geom_line(aes(y = avgvalue.1, col = '2')) +
  geom_line(aes(y = avgvalue.2, col = '3')) +
  geom_line(aes(y = avgvalue.3, col = '4')) +
  geom_line(aes(y = avgvalue.4, col = '5')) +
  scale_x_continuous(breaks = c(1, 25, 50, 75, 100, 125, 144), labels = c('00:00', '04:00', '08:00', '12:00',
 '16:00', '20:00', '23:50')) +
  labs(title='Sundays', x='Time', y='Daily Average Demand', colour='Cluster Number')

ggarrange(WIad, WIwd, WIsaturd, WIsun, common.legend = T)

#stuff to compare

imtired <- merge.data.frame(clusterm, HSclusterm, all.x = TRUE)
smh <- merge.data.frame(imtired, WIclusterm, all.x = TRUE)
Donald <- merge.data.frame(smh, SPclusterm, all.x = TRUE)
pie <- merge.data.frame(Donald, SUMMclusterm, all.x = TRUE) #but as new clusters are created hard to
  tell

#compare pom tables

HSpom
```

```
WIpom
SUMMpom
SPpom
Pom

kable(HSpom, 'html') %>%
  cat(., file = 'HSpom.html')

kable(WIpom, 'html') %>%
  cat(., file = 'WIpom.html')

kable(SPpom, 'html') %>%
  cat(., file = 'SPpom.html')

kable(SUMMpom, 'html') %>%
  cat(., file = 'SUMMpom.html')

kable(pom, 'html') %>%
  cat(., file = 'pom.html')

#comparing seasonal stats by cluster number
Characteristics$TRANSFORMER_TYPE <- as.factor(Characteristics$TRANSFORMER_TYPE)

#CLUSTER 1

#autumn
t1stc1 <- filter(clusterm, Autumnclusterd=='1')

charclust1 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                     (t1stc1$Station))

#HS

hstc1 <- filter(HSclusterm, HSclusterd=='1')

HScharclust1 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                       (hstc1$Station))

#summer

summtc1 <- filter(SUMMclusterm, SUMMclusterd=='1')

SUMMcharclust1 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                         (summtc1$Station))

#spring

SPtc1 <- filter(SPclusterm, SPclusterd=='1')
```

```
SPcharclust1 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (SPtc1$Station))

#winter

WItc1 <- filter(WIclusterm, WIclusterd=='1')

WIcharclust1 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (WItc1$Station))

#CLUSTER 2

#autumn
t1stc2 <- filter(clusterm, Autumnclusterd=='2')

charclust2 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (t1stc2$Station))

#HS

hstc2 <- filter(HSclusterm, HSclusterd=='2')

HScharclust2 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (hstc2$Station))

#summer

summtc2 <- filter(SUMMclusterm, SUMMclusterd=='2')

SUMMcharclust2 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (summtc2$Station))

#spring

SPtc2 <- filter(SPclusterm, SPclusterd=='2')

SPcharclust2 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (SPtc2$Station))

#winter

WItc2 <- filter(WIclusterm, WIclusterd=='2')

WIcharclust2 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (WItc2$Station))

#CLUSTER 3

#autumn
```

```
t1stc3 <- filter(clusterm, Autumnclusterd=='3')

charclust3 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                (t1stc3$Station))

#HS

hstc3 <- filter(HSclusterm, HSclusterd=='3')

HScharclust3 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                (hstc3$Station))

#summer

summtc3 <- filter(SUMMclusterm, SUMMclusterd=='3')

SUMMcharclust3 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                (summtc3$Station))

#spring

SPtc3 <- filter(SPclusterm, SPclusterd=='3')

SPcharclust3 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                (SPtc3$Station))

#winter

WItc3 <- filter(WIclusterm, WIclusterd=='3')

WIcharclust3 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                (WItc3$Station))

#CLUSTER 4

#autumn
t1stc4 <- filter(clusterm, Autumnclusterd=='4')

charclust4 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                (t1stc4$Station))

#HS

hstc4 <- filter(HSclusterm, HSclusterd=='4')

HScharclust4 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                (hstc4$Station))

#summer
```

```
summtc4 <- filter(SUMMclusterm, SUMMclusterd=='4')

SUMMcharclust4 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (summtc4$Station))

#spring

SPtc4 <- filter(SPclusterm, SPclusterd=='4')

SPcharclust4 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (SPtc4$Station))

#winter

WItc4 <- filter(WIclusterm, WIclusterd=='4')

WIcharclust4 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (WItc4$Station))

#CLUSTER 5

#autumn
t1stc5 <- filter(clusterm, Autumnclusterd=='5')

charclust5 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (t1stc5$Station))

#HS

hstc5 <- filter(HSclusterm, HSclusterd=='5')

HScharclust5 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (hstc5$Station))

#summer

summtc5 <- filter(SUMMclusterm, SUMMclusterd=='5')

SUMMcharclust5 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (summtc5$Station))

#spring

SPtc5 <- filter(SPclusterm, SPclusterd=='5')

SPcharclust5 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (SPtc5$Station))
```

```
#winter

WItc5 <- filter(WIclusterm, WIclusterd=='5')

WIcharclust5 <- filter(Characteristics, SUBSTATION_NUMBER %in%
                  (WItc5$Station))

#Summary Statistics of characteristics by cluster/season
#CLUSTER 1
summary(charclust1)
summary(HScharclust1)
summary(SUMMcharclust1)
summary(SPcharclust1)
summary(WIcharclust1)

#CLUSTER 2
summary(charclust2)
summary(HScharclust2)
summary(SUMMcharclust2)
summary(SPcharclust2)
summary(WIcharclust2)

#CLUSTER 3
summary(charclust3)
summary(HScharclust3)
summary(SUMMcharclust3)
summary(SPcharclust3)
summary(WIcharclust3)

#CLUSTER 4
summary(charclust4)
summary(HScharclust4)
summary(SUMMcharclust4)
summary(SPcharclust4)
summary(WIcharclust4)

#CLUSter 5
summary(charclust5)
summary(HScharclust5)
summary(SUMMcharclust5)
summary(SPcharclust5)
summary(WIcharclust5)
```