UK Power Station Clustering Analysis

Juan Joy Paul

```
library(tidyverse)
library(ggplot2)
library(GGally)
library(reshape2)
library(clValid)
library(knitr)
library(ggpubr)
library(repr)
library(factoextra)
```

Summary table and visualization of Characteristics data set.

```
char_data = read.csv('Characteristics.csv')
names(char_data) = tolower(names(char_data))
char_data = char_data %>% mutate(grid_reference = factor(grid_reference), substation_number = factor(su

nrow(char_data)
```

```
## [1] 948
```

```
ncol(char_data)
```

```
## [1] 7
```

```
sum_data = char_data %>%
  rename(Sub_No = "substation_number",
         Customers = "total_customers",
         Tranformer = "transformer_type",
         Transformer_rate = "transformer_rating",
         IC = "percentage_ic") %>%
  select(-grid_reference)%>%
  mutate(Transformer_rate = factor(Transformer_rate))
table = summary(sum_data)
kable(table, caption = 'SUMMARY')
```

Table 1: SUMMARY

Sub_No	Tranformer	Customers	Transformer_rate	IC	lv_feeder_count
563225 : 2	Grd Mtd Dist. Substation :706	Min. : 0.0	500 :262	Min. :0.00000	1 :298

Sub_No	Tranformer	Customers	Transformer_rate	IC	lv_feeder_count
511016 : 1	Pole Mtd Dist. Substation:242	1st Qu.: 3.0	300 :130	1st Qu.:0.01048	4 :202
511017 : 1	NA	Median : 67.5	315 :105	Median :0.17849	3 :128
511028 : 1	NA	Mean :104.3	800 : 84	Mean :0.37982	5 :115
511029 : 1	NA	3rd Qu.:179.2	1000 : 75	3rd Qu.:0.90271	2 : 94
511030 : 1	NA	Max. :569.0	16 : 72	Max. :1.00000	0 : 59
(Other):941	NA	NA	(Other):220	NA	(Other): 52

The data set has 948 rows and 7 columns. The summary of the data set shows that the data set has no nonsensical values.

VISUALISATION:

```
p1 =ggplot(char_data) +
 aes(x = transformer_type, y = transformer_rating, fill = transformer_type) +
 geom_boxplot()+
  stat_summary(fun = 'mean', colour = 'black', geom = 'point' )+
  stat_summary(fun = 'mean', colour = 'black', geom = 'text',
               vjust = 1,aes(label = paste("Mean:",round(..y.., digits = 1))))+
  labs(x = "TRANSFORMER TYPE",y = "TRANSFORMER RATING",title = "Mean and Meadian")+
 theme_minimal()



p2=ggplot(char_data) +
 aes(x = transformer_type, y = percentage_ic,  fill = transformer_type) +
 geom_boxplot()+
  stat_summary(fun = 'mean', colour = 'black', geom = 'point' )+
  stat_summary(fun = 'mean', colour = 'black', geom = 'text',
               vjust = 1,aes(label = paste("Mean:",round(..y.., digits = 1))))+
  labs(x = "TRANSFORMER TYPE",y = "PERCENTAGE OF IC",title = "Mean and Meadian")+
 theme_minimal()


ggplot(sum_data) +
 aes(x = Transformer_rate, fill = Tranformer, weight = IC) +
 geom_bar() +
 scale_fill_hue(direction = 1) +
 labs(y = "percentage of I&C") +
 theme_minimal()
```

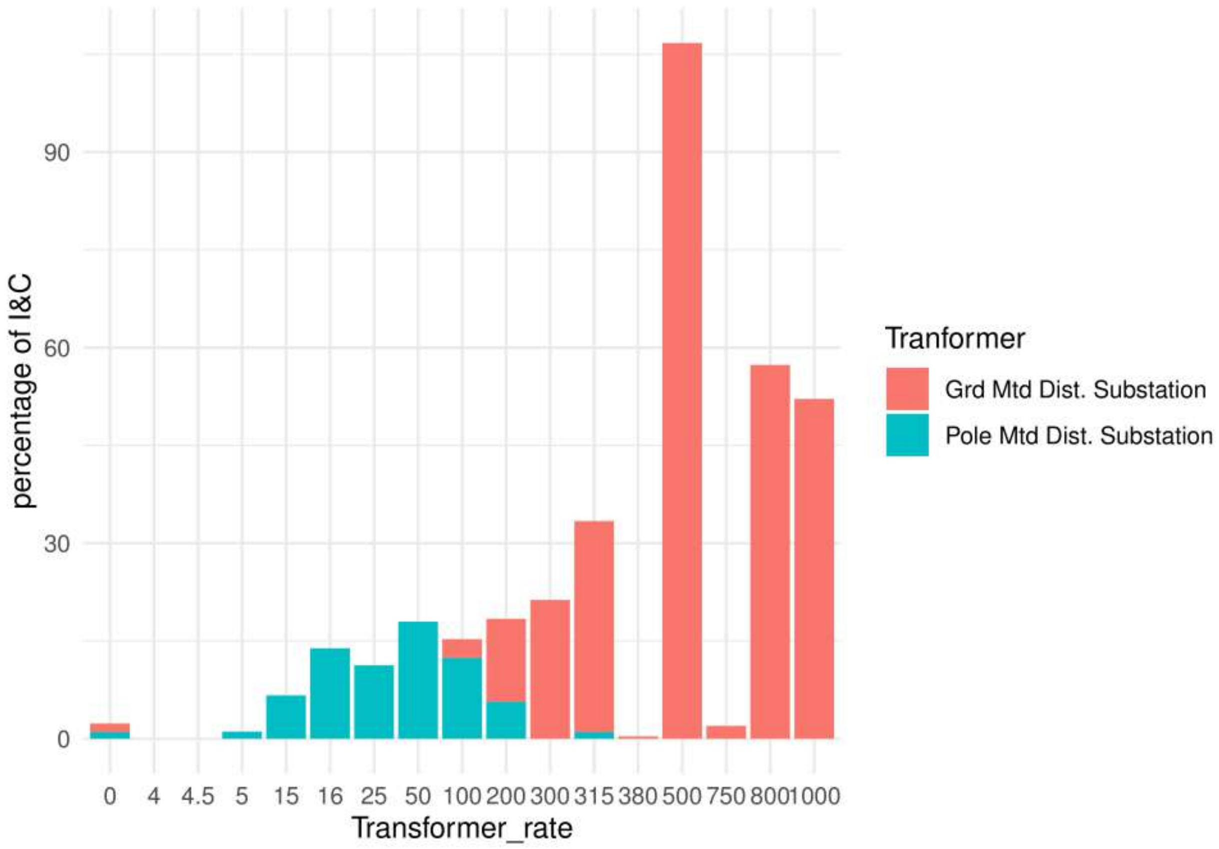

Figure 1: Bar graph

```
plot_1 = ggarrange(p1,p2 ,ncol = 1, nrow = 2)
plot_1
```

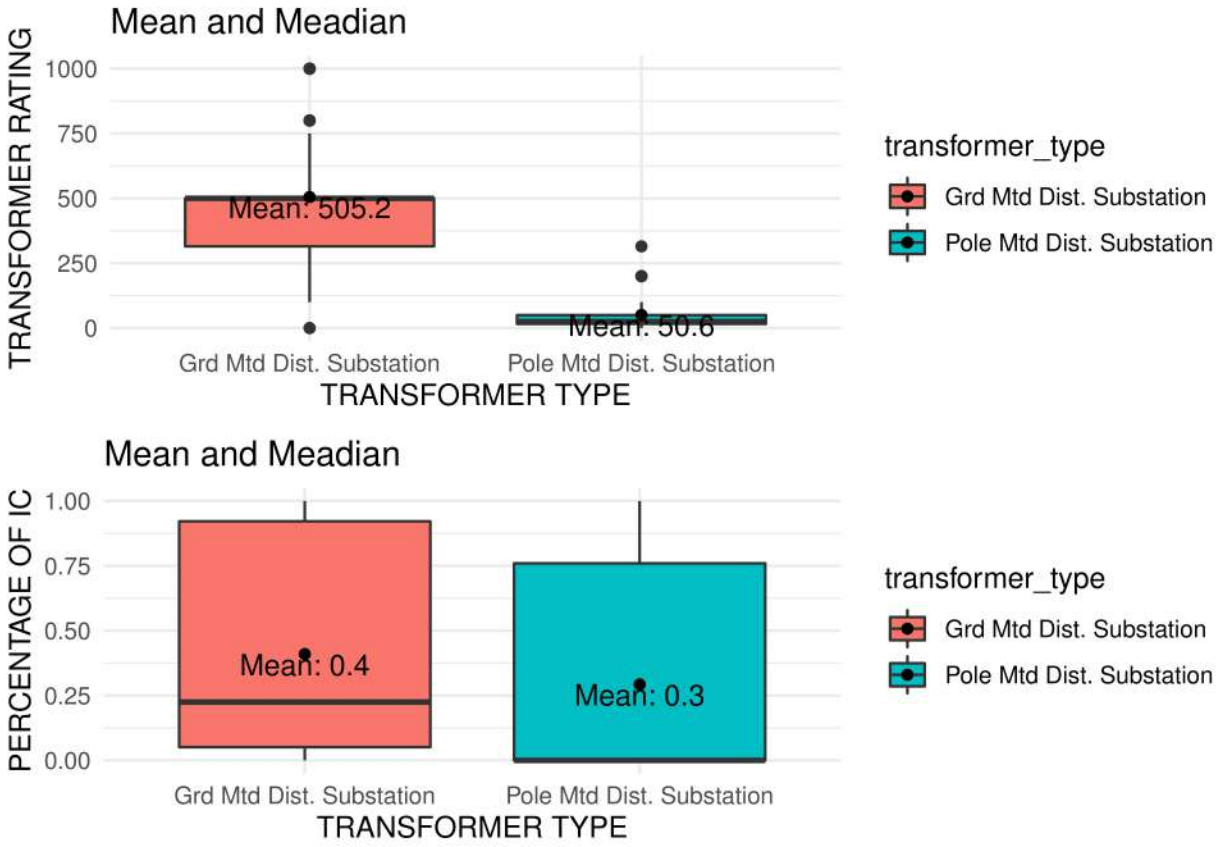

Figure 2: Box plot

Relations between different substation characteristics.

From the box plots, we see a clear relation between Transformer rating and transformer type. In the case of the percentage of industrial and commercial usage, there is a clear difference when looking at the median but when we take the average it looks as if there is no big difference. The reason for this might be due to outliers in the data. The bar graph shows clearly that the industrial and commercial customers demands more power. The grouping show's that there are some local commercial and industrial customers.

```
ggplot(char_data) +
 aes(x = lv_feeder_count, y = total_customers, fill = transformer_type) +
 geom_boxplot() +
 theme_minimal()
```

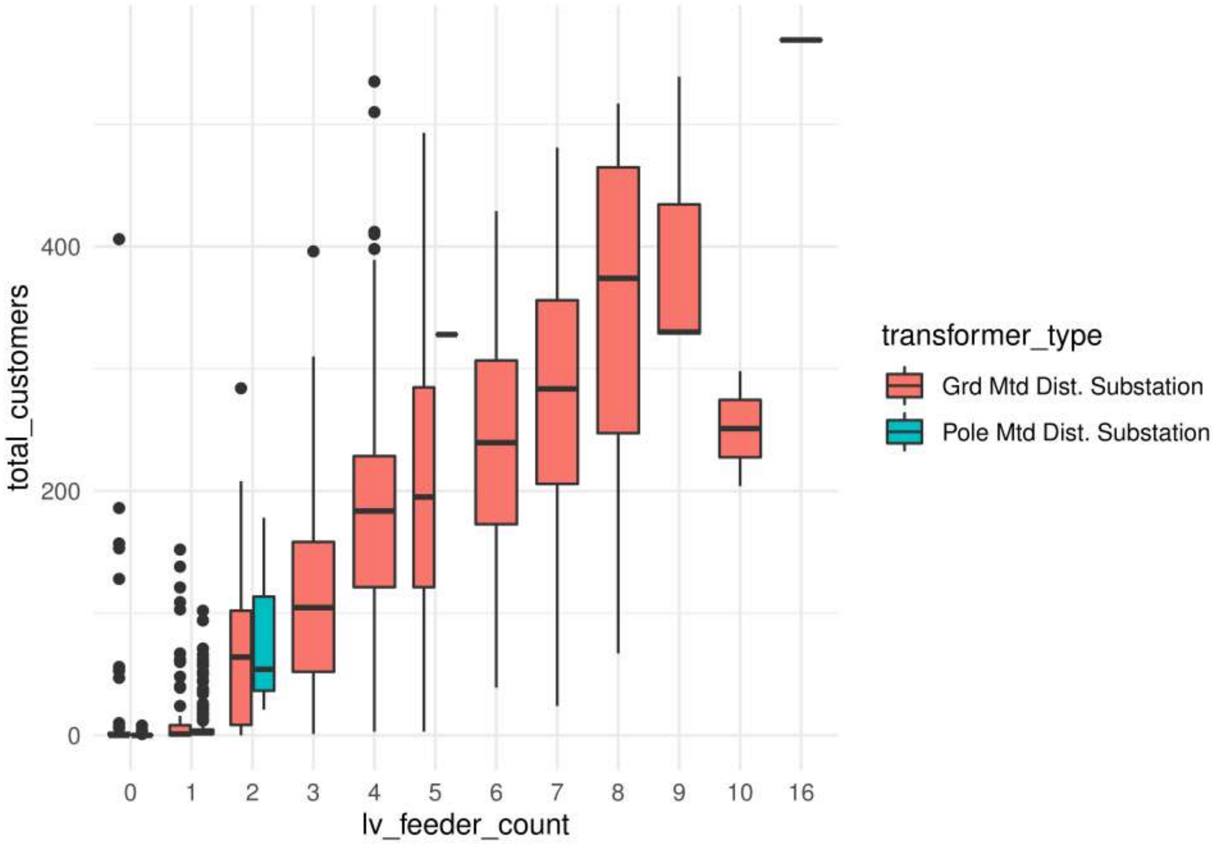

Figure 3: Box plot

From the above box plot, we observe that the population or the number of customers are positively related to the number of low voltage feeders. It is also observed that fewer low voltage feeders are in areas with pole-mounted distribution.

```
load("January_2013.RData")
Modified_January_2013 = January_2013

# making substation as factor
Modified_January_2013 = Modified_January_2013 %>% mutate(Substation = factor(Substation))
```

Dividing power recorded in 10 mins, by daily maximum value.

```
# Creating an max column
Modified_January_2013$Max = apply(Modified_January_2013[3:146], MARGIN =  1, FUN = max, na.rm = T)


# Dividing by max to see patterns of demand.
mod_new = Modified_January_2013[,3:147]
mod_new = mod_new/mod_new$Max
col_1_2_3 = select(Modified_January_2013, Date, Substation, Max  )
new = cbind(col_1_2_3,mod_new)

new = new %>% select(-Max)
new = new %>% rename(substation = 'Substation')
```

Average daily power demand profile

```
# Grouping all substations
mean_jan = aggregate(new[,3:146], list(new$substation), FUN = mean)
mean_jan = mean_jan %>% rename(substation = 'Group.1')
mean_jan[1:2,1:10]
```

```
##   substation     00:00     00:10     00:20     00:30     00:40     00:50
## 1     511016 0.6215941 0.6248997 0.6100203 0.5992056 0.5902452 0.5783236
## 2     511029 0.5677445 0.6417366 0.6750539 0.7016390 0.7621812 0.7493059
##       01:00     01:10     01:20
## 1 0.5661389 0.5534972 0.5450424
## 2 0.7336417 0.7161868 0.7078567
```

Distance matrix and dendrogram

```
# Distance Matrix
distance_M = dist(mean_jan[-1], method = "euclidean")

distance_matrix = as.matrix(distance_M)

cluster_all = hclust(distance_M)
plot(cluster_all,cex = 0.6 , hang = -1)
rect.hclust(cluster_all, k = 5, border = 3:7)
```

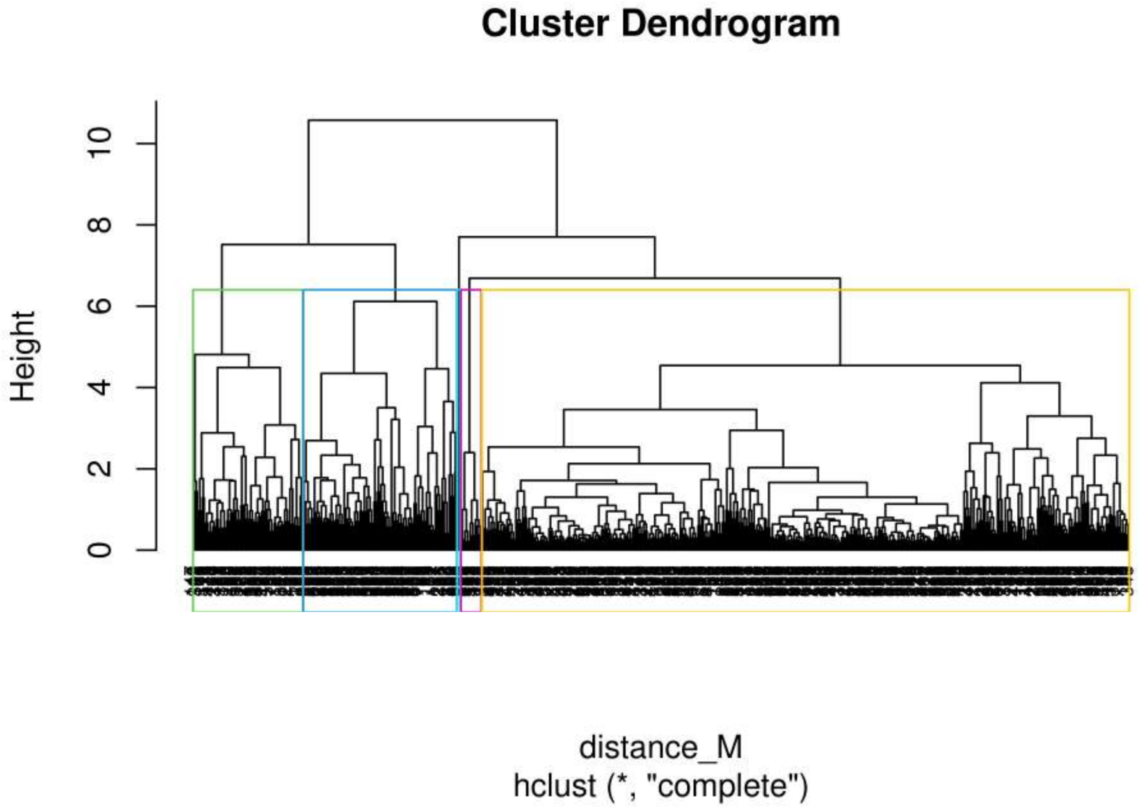

Figure 4: Dendrogram

Finding number of cluster

```
# dunn indexing:

cluster_point = cutree(cluster_all, k = 5)
dunn(distance_M,cluster_point)
```

```
## [1] 0.1169634
```

```
#Elbow Method
pam = fviz_nbclust(mean_jan[-1],hcut,method = "wss")
#Average Silhouette Method
pan = fviz_nbclust(mean_jan[-1],hcut,method = "silhouette")

plot_23 = ggarrange(pam,pan ,ncol = 1, nrow = 2)
plot_23
```

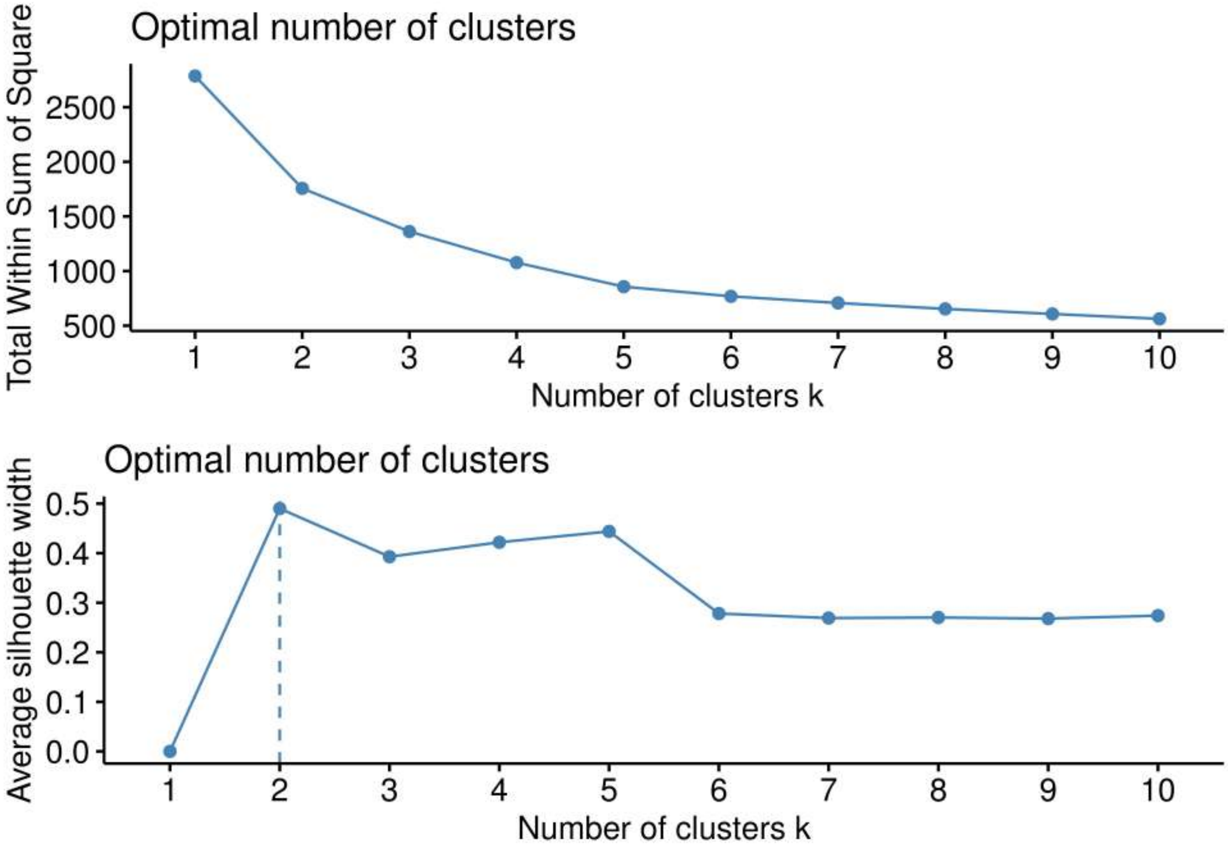

Figure 5: Elbow and average silhouette

Based on different tests the number of clusters is determined. From the elbow method, we can see the effect of cluster plateaus after cluster number 5. From the average silhouette method, it is observed best cluster number is 2 but the number of clusters is low for analysis the best next value is 5. The Dunn index for 5 clusters is relatively good therefore 5 clusters are selected.

```
sil = silhouette(cluster_point,distance_M)
plot(sil)
```

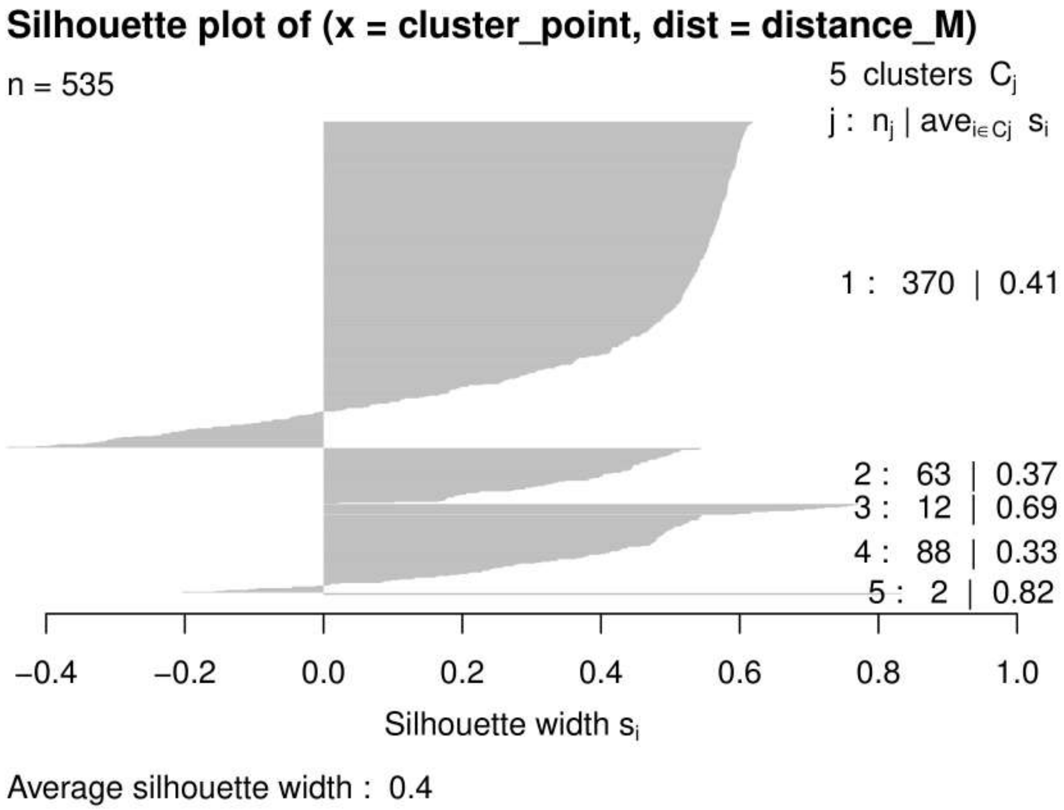

Figure 6: Silhouette plot

From the silhouette plot it is observed that the strongest cluster is cluster 2 other clusters groupings are relatively weak.

```
mean_jan_clust_2013 = cbind(mean_jan,cluster_point)
mean_jan_clust_2013 = mean_jan_clust_2013 %>% mutate(cluster_point = factor(cluster_point))

#number of substations in each cluster
cluster_table = table(mean_jan_clust_2013$cluster_point)
#cluster_table = cluster_table %>% mutate(cluster = "Var1")
cluster_table = as.data.frame(cluster_table)
cluster_table = cluster_table %>% rename(Cluster = "Var1",   Numb_of_substation = "Freq")
kable(cluster_table, caption = 'Number of substation in each cluster')
```

Table 2: Number of substation in each cluster

Cluster	Numb_of_substation
1	370
2	63
3	12
4	88
5	2

```
fviz_cluster(list(data = mean_jan[-1], cluster = cluster_point))
```

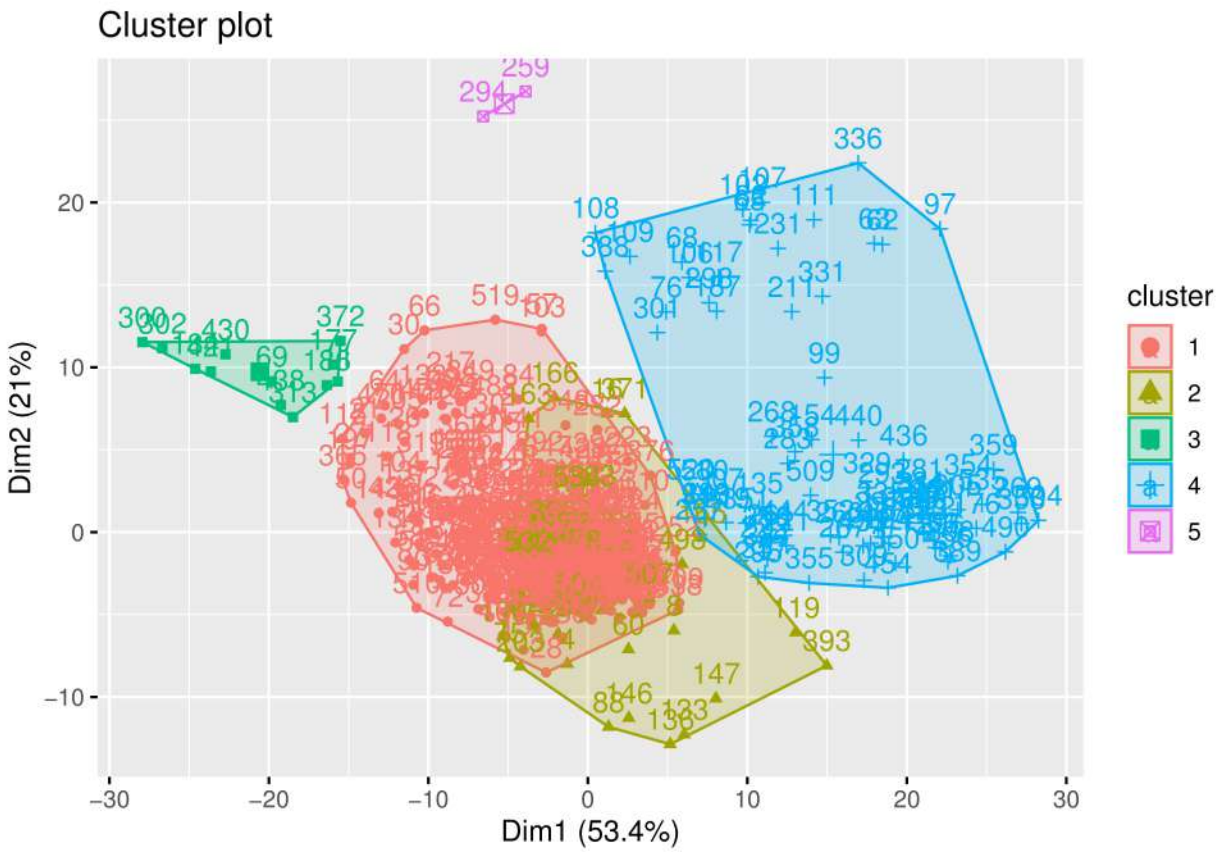

Figure 7: cluster regions

Visualization of average demand profile

```
# (v) LONG FORMAT

data_long_full = mean_jan_clust_2013 %>%
  gather(variable,value,-c(substation,cluster_point))
data_long_full = arrange(data_long_full, substation, variable)
mean_val = aggregate(value ~ variable + cluster_point, data_long_full, mean)
mean_val = mean_val %>% rename(power = 'value', time = 'variable')

# average demand profile for each cluster

data_long_full_1 = arrange(data_long_full, substation, variable)
data_long_full_1 = data_long_full_1 %>% rename(power = 'value', time = 'variable')


ggplot(data_long_full_1,aes(x = time, y = power, colour = cluster_point))+
  geom_point(shape = "circle", size = 1.5)+
  scale_color_manual(values = c(`1` = "#EFA49F", `2` = "#F2FF90", `3` = "#A0FF9D", `4` = "#619CFF", `5`
  geom_point(data = mean_val, colour ='red')+
 facet_wrap(vars(cluster_point))
```

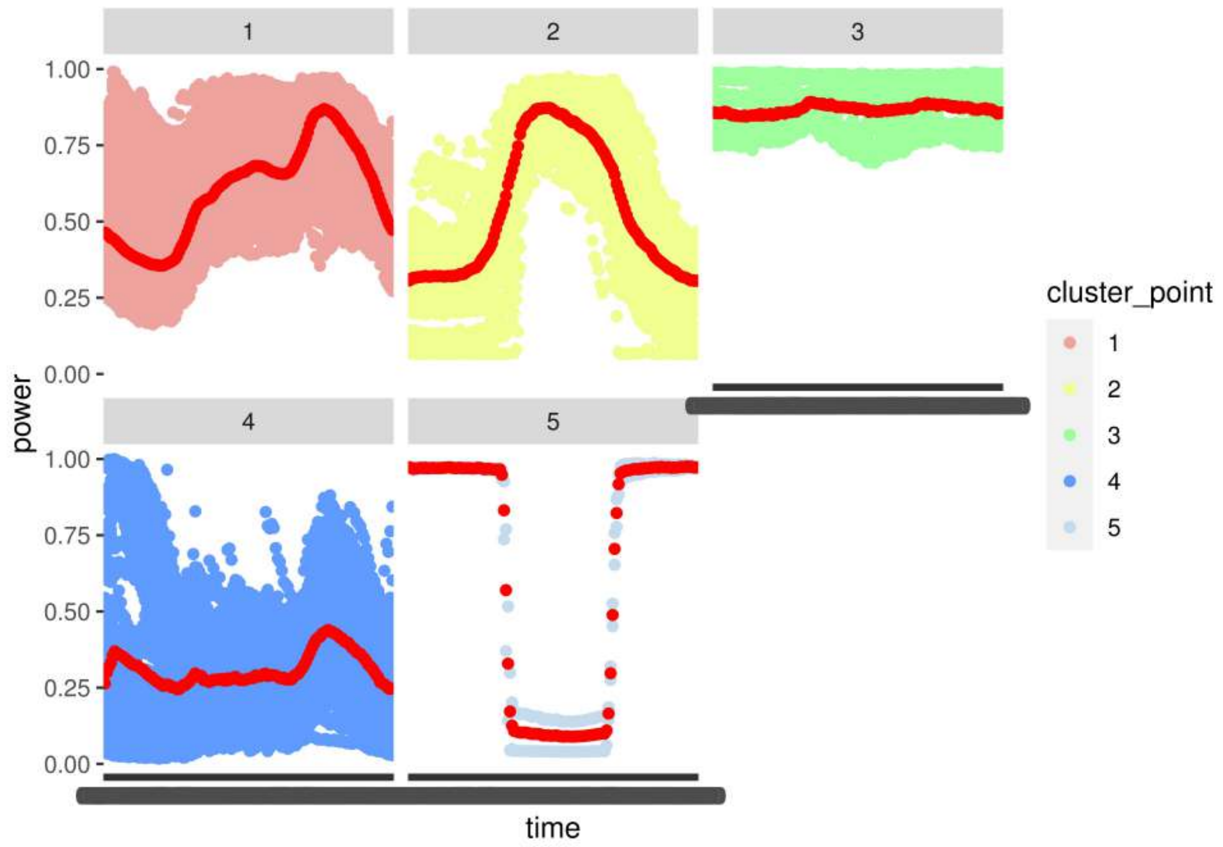

Figure 8: power demand profile (2013 substations)

Weekend and weekday demand profile.

```
df_dummy_1 = mean_jan_clust_2013%>% select(substation,cluster_point)

comb_original = merge(df_dummy_1,new, by = 'substation', all = T)


#weekend and weekday

comb_original$weekday = weekdays(comb_original$Date)

is_weekend = function(n){
  require(lubridate)
  (ifelse(wday(as.Date(n)) == 1, T, F) | ifelse(wday(as.Date(n)) == 7, T, F))
}
comb_original$weekend = is_weekend(comb_original$Date)


comb_original_long = comb_original %>%
  gather(Time,power,-c(substation,cluster_point,Date,weekend,weekday))
```

```
sub13_long_week = aggregate(power ~ Time + cluster_point + weekend, comb_original_long, mean)
sub13_long_week = arrange(sub13_long_week,cluster_point,Time)
sub13_long_week = sub13_long_week %>% mutate(weekend = factor(weekend))
```

```
ggplot(sub13_long_week) +
 aes(x = Time, y = power, colour = weekend) +
 geom_point(shape = "circle", size = 1.5) +
 scale_color_hue(direction = 1) +
 theme_minimal() +
 facet_wrap(vars(cluster_point))
```

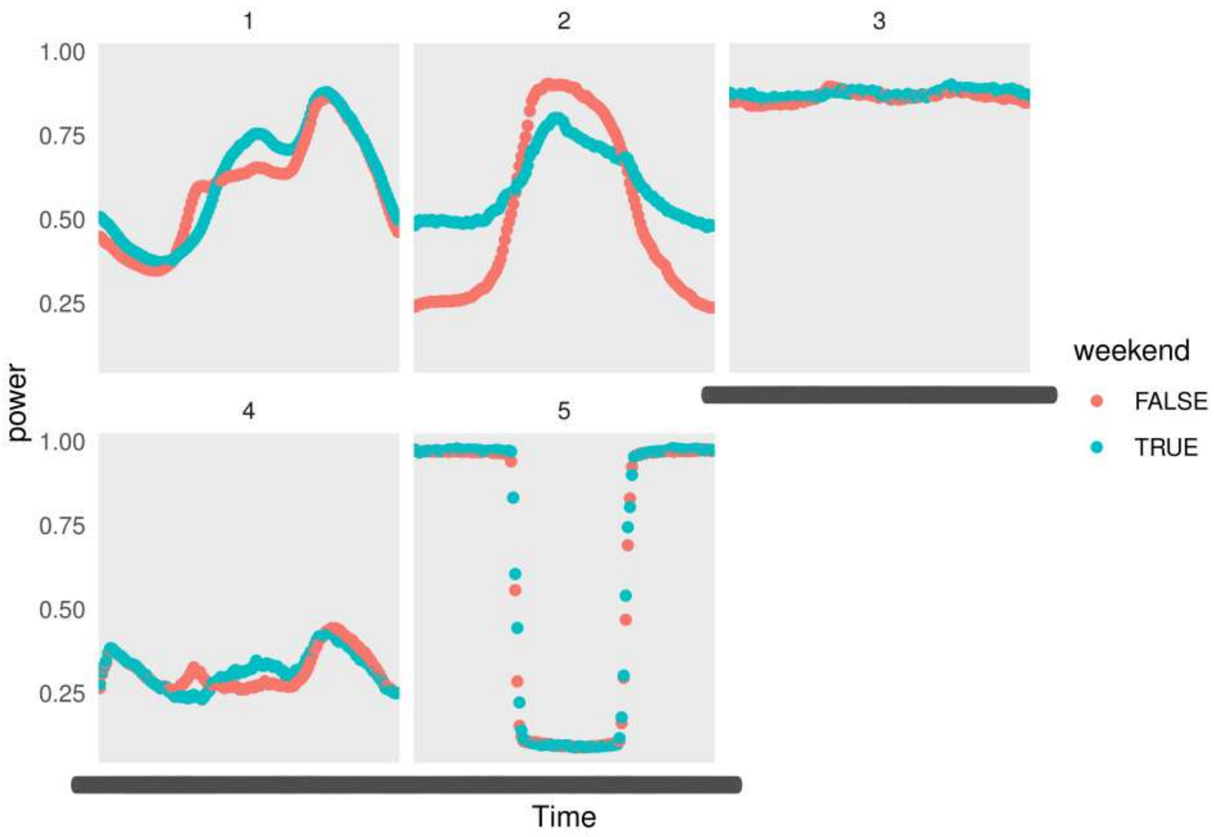

Figure 9: weekend and weekday power demand profile (2013 substations)

We observe some variations for weekday and weekend in cluster 2. All other clusters have almost same demand profile.

Comparing with characteristic Data set

```
req_char = char_data %>% select(-grid_reference)
req_char = req_char %>% rename(substation = 'substation_number')

sub_clust_2013 = mean_jan_clust_2013 %>% select(substation,cluster_point)
characteristic_naming = merge(sub_clust_2013,req_char, by = 'substation', all = T)
characteristic_naming =characteristic_naming%>% arrange(substation)


cluster_1 = filter(characteristic_naming, cluster_point == '1')
summary(cluster_1)
```

```
##    substation  cluster_point                    transformer_type total_customers
##  563225 :  2   1:371         Grd Mtd Dist. Substation :346     Min.   :  0.0
##  511016 :  1   2:  0         Pole Mtd Dist. Substation: 25      1st Qu.: 70.0
##  511029 :  1   3:  0                                            Median :152.0
##  511030 :  1   4:  0                                            Mean   :159.7
##  511034 :  1   5:  0                                            3rd Qu.:232.0
##  511035 :  1                                                    Max.   :569.0
##  (Other):364
##  transformer_rating percentage_ic     lv_feeder_count
##  Min.   :   0.0     Min.   :0.00000   4      :120
##  1st Qu.: 300.0     1st Qu.:0.02601   3      : 72
##  Median : 500.0     Median :0.15466   5      : 64
##  Mean   : 440.2     Mean   :0.30151   1      : 45
##  3rd Qu.: 500.0     3rd Qu.:0.45748   2      : 30
##  Max.   :1000.0     Max.   :1.00000   0      : 15
##                                       (Other): 25
```

Cluster 1:

From Figure 9: average demand profile the power demand at night is low. The power demand increases in the afternoon and in the evening. the peak of power demand is seen in the evening. This observation gives us an idea that it might be a cluster containing domestic or household power demand. From the cluster summary, the number of customers is high and the percentage of industrial and commercial customers is low. Therefore we can conclude and name the cluster as domestic poer demand cluster.

```
cluster_2 = filter(characteristic_naming, cluster_point == '2')
summary(cluster_2)
```

```
##    substation cluster_point                  transformer_type total_customers
##  511033 : 1   1: 0          Grd Mtd Dist. Substation :62      Min.   :  0.00
##  511150 : 1   2:63          Pole Mtd Dist. Substation: 1      1st Qu.:  2.50
##  511151 : 1   3: 0                                            Median : 11.00
##  511188 : 1   4: 0                                            Mean   : 29.03
##  511266 : 1   5: 0                                            3rd Qu.: 26.00
##  511269 : 1                                                   Max.   :292.00
##  (Other):57
##  transformer_rating percentage_ic    lv_feeder_count
##  Min.   : 200.0     Min.   :0.0000   3      :16
##  1st Qu.: 500.0     1st Qu.:0.9423   1      :15
##  Median : 500.0     Median :1.0000   2      :10
##  Mean   : 624.8     Mean   :0.9235   4      : 9
##  3rd Qu.: 800.0     3rd Qu.:1.0000   5      : 7
##  Max.   :1000.0     Max.   :1.0000   0      : 5
##                                      (Other): 1
```

Cluster 2:

The peak power demand is midday. From Figure 9: it might be an industrial and commercial customer. The number of customers is low and the percentage of industrial and commercial customers is high. This cluster can be named the industrial and commercial cluster.

```
cluster_3 = filter(characteristic_naming, cluster_point == '3')
summary(cluster_3)
```

```
##    substation cluster_point                  transformer_type total_customers
##  511191 :1    1: 0          Grd Mtd Dist. Substation :5       Min.   :  0.00
##  512454 :1    2: 0          Pole Mtd Dist. Substation:7       1st Qu.:  0.75
##  521859 :1    3:12                                            Median :  2.00
##  521874 :1    4: 0                                            Mean   : 45.25
##  522239 :1    5: 0                                            3rd Qu.:  3.75
##  532645 :1                                                    Max.   :510.00
##  (Other):6
##  transformer_rating percentage_ic    lv_feeder_count
##  Min.   :  25.0     Min.   :0.9332   1      :9
##  1st Qu.:  87.5     1st Qu.:0.9984   4      :2
##  Median : 150.0     Median :1.0000   5      :1
##  Mean   : 351.2     Mean   :0.9903   0      :0
##  3rd Qu.: 575.0     3rd Qu.:1.0000   2      :0
##  Max.   :1000.0     Max.   :1.0000   3      :0
##                                      (Other):0
```

Cluster 3:

From Figure 9: this cluster shows it requires high power all day. The percentage of industrial and commercial customers is high. This cluster can be named the heavy industrial cluster.

```
cluster_4 = filter(characteristic_naming, cluster_point == '4')
summary(cluster_4)
```

```
##    substation cluster_point                  transformer_type total_customers
##  512438 : 1   1: 0          Grd Mtd Dist. Substation :20      Min.   :  0.00
##  512440 : 1   2: 0          Pole Mtd Dist. Substation:68      1st Qu.:  1.00
##  512443 : 1   3: 0                                            Median :  3.00
##  512448 : 1   4:88                                            Mean   : 25.02
##  512918 : 1   5: 0                                            3rd Qu.: 13.00
##  513044 : 1                                                   Max.   :235.00
##  (Other):82
##  transformer_rating percentage_ic    lv_feeder_count
##  Min.   :   5.0     Min.   :0.0000   1      :73
##  1st Qu.:  16.0     1st Qu.:0.0000   4      : 6
##  Median :  25.0     Median :0.0000   3      : 4
##  Mean   : 189.9     Mean   :0.1845   0      : 3
##  3rd Qu.: 100.0     3rd Qu.:0.2020   5      : 2
##  Max.   :1000.0     Max.   :1.0000   2      : 0
##                                      (Other): 0
```

Cluster 4:

The graph shows that this cluster uses low power throughout the day. The number of pole-mounted distribution is higher in this cluster therefore it is in a rural region. The percentage of industrial and commercial customers is Low. This suggests an all-day functioning rural region device a traffic light or electricity meters. This cluster can be called all-day low power devices.

```
cluster_5 = filter(characteristic_naming, cluster_point == '5')
summary(cluster_5)
```

```
##    substation cluster_point                  transformer_type total_customers
##  531057 :1    1:0           Grd Mtd Dist. Substation :0       Min.   :0.0
##  532235 :1    2:0           Pole Mtd Dist. Substation:2       1st Qu.:0.5
##  511016 :0    3:0                                             Median :1.0
##  511029 :0    4:0                                             Mean   :1.0
##  511030 :0    5:2                                             3rd Qu.:1.5
##  511033 :0                                                    Max.   :2.0
##  (Other):0
##  transformer_rating percentage_ic  lv_feeder_count
##  Min.   : 50.0      Min.   :0.00   0      :1
##  1st Qu.: 62.5      1st Qu.:0.25   1      :1
##  Median : 75.0      Median :0.50   2      :0
##  Mean   : 75.0      Mean   :0.50   3      :0
##  3rd Qu.: 87.5      3rd Qu.:0.75   4      :0
##  Max.   :100.0      Max.   :1.00   5      :0
##                                    (Other):0
```

Cluster 5:

Here the power demand is only at the night and when it is dark outside. The power is drawn from a pole-mounted distributer and the customers are really low. This might be a street light in a rural area. This cluster can be called rural street light.

Week day and Week end pattern

```r
load("new_substations.RData")
added_substation = new_substations
added_substation = added_substation %>% mutate(Substation = factor(Substation))
added_substation[c(-1,-2)] = t(apply(added_substation[c(-1,-2)], 1, function(y) y/max(y)))

#weekend and weekday
#added_substation[['Date']] = strptime(added_substation[['Date']],format = "%Y-%m-%d")

added_substation$Date=as.Date(added_substation$Date)

# add simple date column
added_substation$weekday = weekdays(added_substation$Date)
added_substation$weekend = is_weekend(added_substation$Date)
added_substation = added_substation %>% mutate(weekend  = factor(weekend))


#long format
added_substation = select(added_substation, -Date)
added_substation_long = added_substation %>%
  gather(Time,power,-c(Substation,weekend,weekday))
summary(added_substation_long)
```

```
##   Substation      weekday            weekend          Time
##  513687:4176   Length:20880       FALSE:15120   Length:20880
##  521055:4176   Class :character   TRUE : 5760   Class :character
##  522287:4176   Mode  :character                 Mode  :character
##  525379:4176
##  531475:4176
##
##      power
##  Min.   :0.2044
##  1st Qu.:0.5418
##  Median :0.7198
##  Mean   :0.7004
##  3rd Qu.:0.8876
##  Max.   :1.0000
```

```
mean_add_sub = aggregate(power ~ Substation + Time + weekend,added_substation_long, mean)
mean_add_sub = arrange(mean_add_sub,Substation,Time)


ggplot(mean_add_sub) +
 aes(x = Time, y = power, colour = weekend) +
 geom_point(shape = "circle", size = 1.5) +
 scale_color_hue(direction = 1) +
 theme_minimal() +
 facet_wrap(vars(Substation))
```

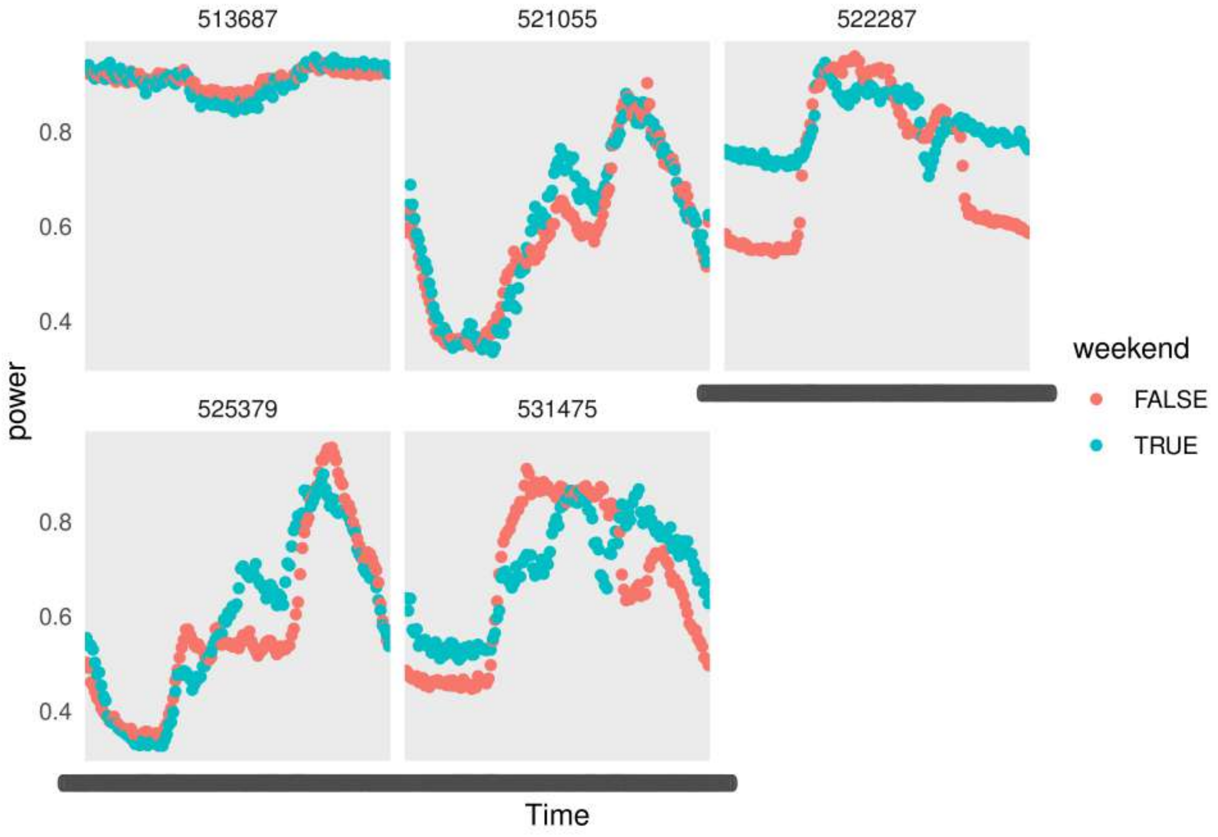

Figure 10: weekend and weekday power demand profile (New substations)

Assigning clusters

```
centroids_org = aggregate(comb_original[,4:147], list(comb_original$cluster_point), FUN = mean)
centroids_org = centroids_org %>% rename(cluster_point = "Group.1")


euclidian_distance <- function(x, y) {
  diff <- x - y
  sqrt(sum(diff^2))
}

avg_new_station = added_substation %>% group_by(Substation) %>%
  select(-weekday,-weekend) %>%
  summarise_all(mean)

i = 1
point = avg_new_station[i,]
centroid_id = 1

assign_point_to_cluster <- function(point, centroids_df) {
  k = nrow(centroids_df)
  m = ncol(centroids_df)
  distances = vector(length = k)
  for(centroid_id in 1:k) {
    centroid_pnt = centroids_df[centroid_id, 2:m]
    distances[centroid_id] = euclidian_distance(point, centroid_pnt)
  }
  which.min(distances)
}

assign_all_points_to_clusters <- function(avg_new_station, centroids_df) {
  npoints = nrow(avg_new_station)
  clusters = vector(length=npoints)
  for(i in 1:npoints) {
    point <- avg_new_station[i, ]
    clusters[i] <- assign_point_to_cluster(point, centroids_df)
  }
  clusters
}

new_clust = assign_all_points_to_clusters(avg_new_station[-1], centroids_org)

avg_new_station = avg_new_station%>% mutate(cluster = new_clust)
avg_new_station = avg_new_station %>% rename(substation = 'Substation', cluster_point = 'cluster')
avg_new_station[,c(1,146)]
```

```
## # A tibble: 5 x 2
##   substation cluster_point
##   <fct>              <int>
## 1 513687                 3
## 2 521055                 1
## 3 522287                 3
## 4 525379                 1
## 5 531475                 1
```

```
# original dataset with average substations values with new substations:

new_sub_data_2013 = rbind(avg_new_station,mean_jan_clust_2013)
```

```
#
avg_new_sub_without_clust = avg_new_station[-146]
new_sub_with_and_without_clust = merge(mean_jan_clust_2013,avg_new_sub_without_clust,on = substation, a
```

```
centroid_with_NA = new_sub_with_and_without_clust %>%
  group_by(cluster_point)%>%
  select(-cluster_point,-substation)%>%
  summarise_all(mean)

ggplot(new_sub_with_and_without_clust, aes(y = `10:00`, x = `20:00`, colour = cluster_point, size = is.
  geom_point() +
  geom_point(data = centroid_with_NA, shape = 8,size = 10)+
  theme_minimal()
```

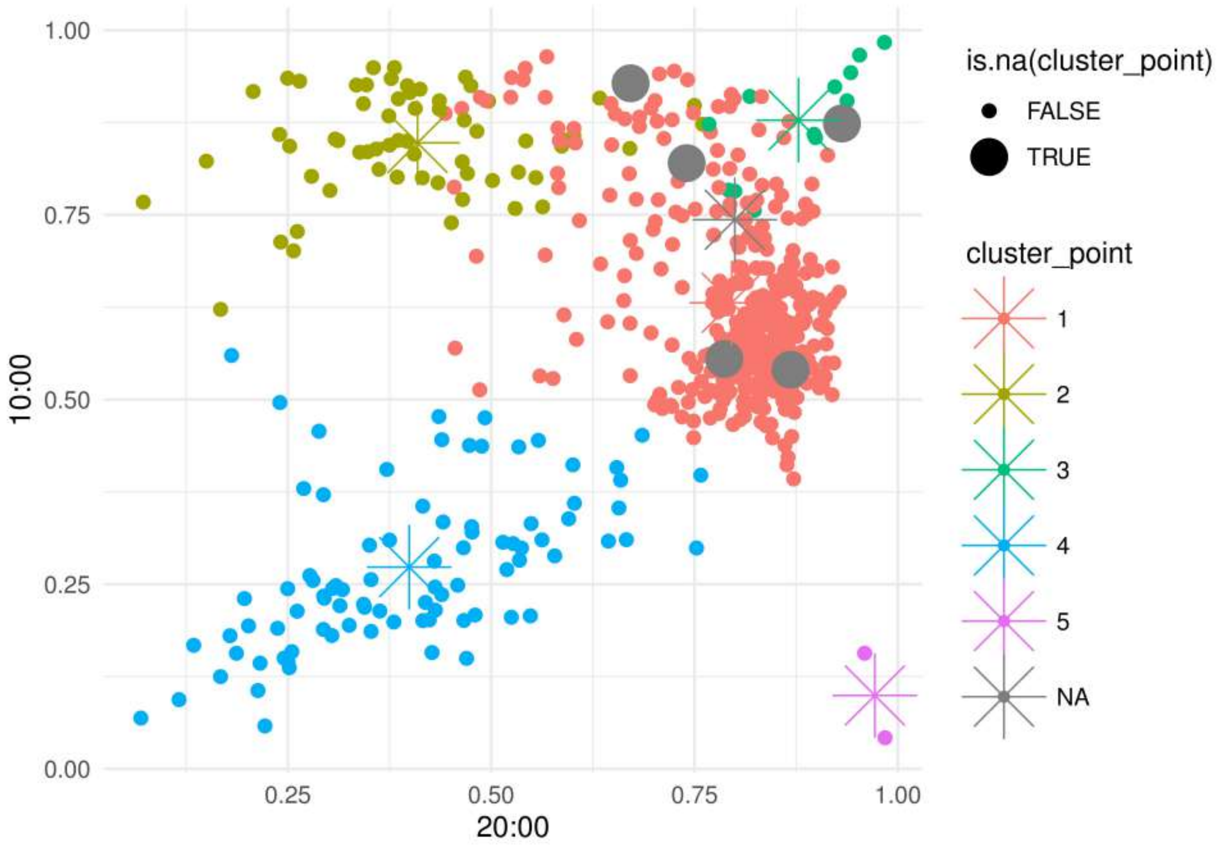

Figure 11: predicting cluster

The plot above visualizes only two dimensions. But from what we observe, we see that 2 new substation centroids are close to cluster 1 and the other 3 are close to cluster 3 this alone is not sufficient. In comparing Figure 10: and Figure 9: we observe that substation 521055(2nd one) and 525379(4th one) looks exactly like it belongs to cluster 1 and substation 513687(1st one) looks like it belongs to cluster 3. These three substation clustering was true on doing distance matrix. The substation 522287 looks like either cluster 2 or 3 and 531475 looks like cluster 2 or 1. when doing the distance matrix we found it belongs to 3 and 1 respectively. Therefore the cluster allocation was as expected.

```
# CENTROID FOR DATA WITH ADDITIONAL NEW SUBSTATION
centroid_with_new_sub = new_sub_data_2013 %>%
  group_by(cluster_point)%>%
  select(-cluster_point,-substation)%>%
  summarise_all(mean)

# CHANGE THE CENTROID
ggplot(new_sub_data_2013, aes(y = `10:00`, x = `20:10`, colour = cluster_point)) +
  geom_point() +
  geom_point(data = centroid_with_new_sub, shape = 15,size = 5)+
  theme_minimal()
```

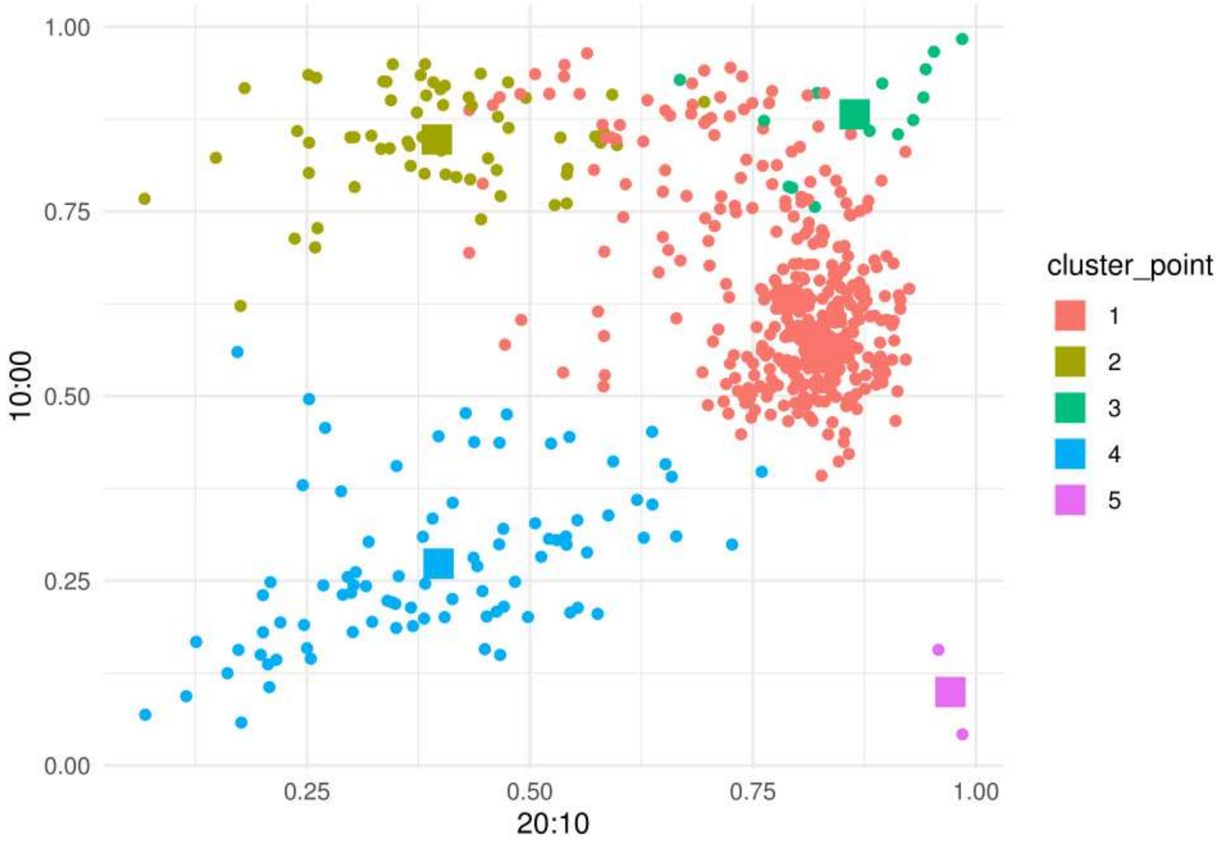

Figure 12: scatter plot with centroid's

```
#2014 data clustering
load("January_2014.RData")
original_2014 = January_2014
original_2014 = original_2014 %>% rename(substation = Substation)
original_2014 = original_2014 %>% mutate(substation = factor(substation))
original_2014[c(-1,-2)] = t(apply(original_2014[c(-1,-2)], 1,function(y) y/max(y)))

avg_sub_2014 = aggregate(original_2014[,3:146], list(original_2014$substation), FUN = mean)

#clustering
dist_2014_no_max = dist(avg_sub_2014[-1], method = "euclidean")
distance_matrix_2014 = as.matrix(dist_2014_no_max)

cluster_2014 = hclust(dist_2014_no_max)
```

```r
cluster_point_2014 = cutree(cluster_2014, k = 5)

#cluster checking
#dunn(distance_M,cluster_point_2014)

mean_jan_clust_2014 = cbind(avg_sub_2014,cluster_point_2014)
mean_jan_clust_2014 = mean_jan_clust_2014 %>% mutate(cluster_point_2014 = factor(cluster_point_2014))
mean_jan_clust_2014 = mean_jan_clust_2014 %>% rename(substation = 'Group.1')


#long format
data_long_full_2014 = mean_jan_clust_2014 %>%
  gather(variable,value,-c(substation,cluster_point))
data_long_full_2014 = arrange(data_long_full, substation, variable)
mean_val_2014 = aggregate(value ~ variable + cluster_point, data_long_full, mean)
mean_val_2014 = mean_val_2014 %>% rename(power = 'value', time = 'variable')

# average demand profile for each cluster

data_long_full_2014  = arrange(data_long_full_2014 , substation, variable)
data_long_full_2014  = data_long_full_2014  %>% rename(power = 'value', time = 'variable')


ggplot(data_long_full_2014 ,aes(x = time, y = power, colour = cluster_point))+
  geom_point(data = mean_val_2014, colour ='orange')+
  geom_point(data = mean_val, colour ='red', size = 0.1)+
  facet_wrap(vars(cluster_point))+
  theme_minimal()
```

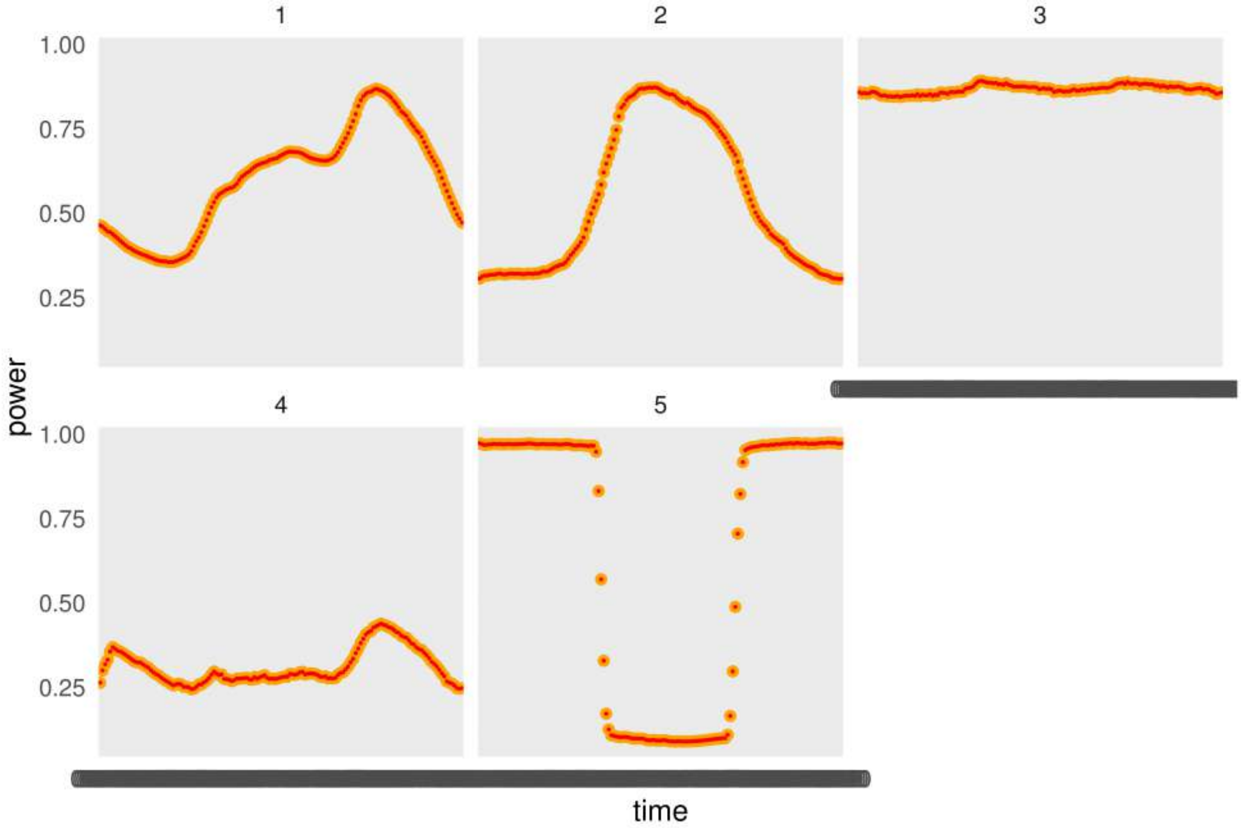

Figure 13: power demand profile 2014 and 2013

After performing hierarchical clustering individually on the 2014 data set, which is done to check if the same pattern of demand occurs again. We check to see if the cluster categories are the same. From figure 13 it is observed that the clustering is exactly the same. The red line shows the power demand for 2013 and the orange for 2014. The clustering shows the same average power demand.

```
# testing if substation is same
sum(mean_jan_clust_2013[,1] == mean_jan_clust_2014[,1])
```

```
## [1] 535
```

```
# testing if cluster is same
#mean_jan_clust_2013[,146] == mean_jan_clust_2014[,146]
sum(mean_jan_clust_2013[,146] == mean_jan_clust_2014[,146])
```

```
## [1] 329
```

The above value 535 shows that all the substations are the same and are still active in 2014. The value 329 shows that only 329 substations belong to the same cluster as in 2013. This shows that there may be a change in power demand for a few substations. But the average power demand for each cluster still remains the same.

```
#2015 data clustering
load("January_2015.RData")
original_2015 = January_2015
```

```
original_2015 = original_2015 %>% rename(substation = Substation)
original_2015 = original_2015 %>% mutate(substation = factor(substation))
original_2015[c(-1,-2)] = t(apply(original_2015[c(-1,-2)], 1,function(y) y/max(y)))

avg_sub_2015 = aggregate(original_2015[,3:146], list(original_2015$substation), FUN = mean)

#clustering
dist_2015_no_max = dist(avg_sub_2015[-1], method = "euclidean")
distance_matrix_2015 = as.matrix(dist_2015_no_max)
cluster_2015 = hclust(dist_2015_no_max)
cluster_point_2015 = cutree(cluster_2015, k = 5)

#cluster checking
#dunn(distance_M,cluster_point_2015)


mean_jan_clust_2015 = cbind(avg_sub_2015,cluster_point_2015)
mean_jan_clust_2015 = mean_jan_clust_2015 %>% mutate(cluster_point_2015 = factor(cluster_point_2015))
mean_jan_clust_2015 = mean_jan_clust_2015 %>% rename(substation = 'Group.1')

# long format
data_long_full_2015 = mean_jan_clust_2015 %>%
  gather(variable,value,-c(substation,cluster_point))
data_long_full_2015 = arrange(data_long_full, substation, variable)
mean_val_2015 = aggregate(value ~ variable + cluster_point, data_long_full, mean)
mean_val_2015 = mean_val_2015 %>% rename(power = 'value', time = 'variable')

data_long_full_2015  = arrange(data_long_full_2015 , substation, variable)
data_long_full_2015  = data_long_full_2015  %>% rename(power = 'value', time = 'variable')


ggplot(data_long_full_2015 ,aes(x = time, y = power, colour = cluster_point))+
  geom_point(data = mean_val_2015, colour ='blue')+
  geom_point(data = mean_val, colour ='red', size = 0.1)+
  facet_wrap(vars(cluster_point))+
  theme_minimal()
```

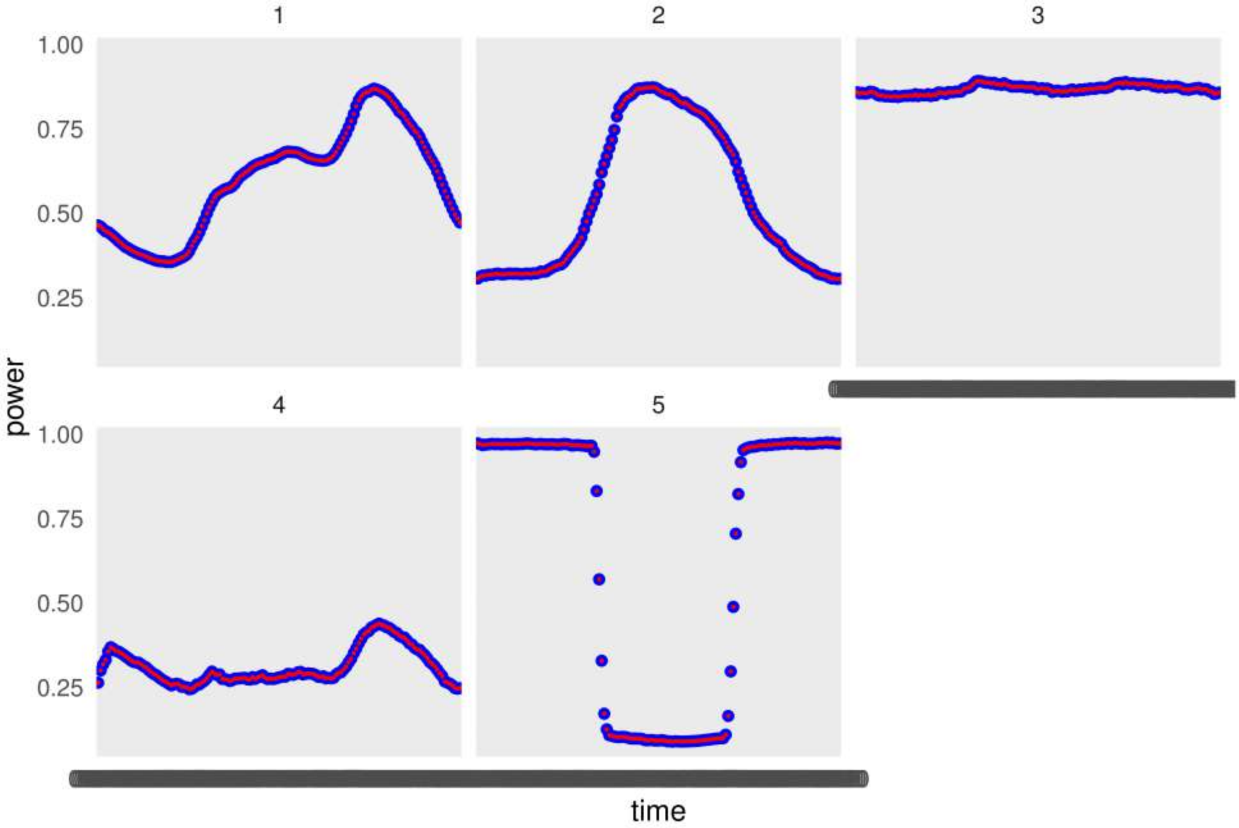

Figure 14: power demand profile 2015 and 2013

```
# testing if substation is same
sum(mean_jan_clust_2013[,1] == mean_jan_clust_2015[,1])
```

```
## [1] 535
```

```
# testing if cluster is same

sum(mean_jan_clust_2013[,146] == mean_jan_clust_2015[,146])
```

```
## [1] 490
```

We do the clustering using hierarchical again for 2015 data. Here we observe the same average power demand across all the clusters. The red line is the 2013 demand profile and the blue is the 2015 demand profile. The number of substations for the year 2015 remains the same as in 2014 and 2013. On comparing the cluster of each substation of the years 2013 and 2015 we see that 490 clusters match. the variation for the average demand profile between the years 2015 and 2013 is less than that between 2014 and 2013. Comparing the years 2015 and 2014 318 cluster points match.

```
#between 2014 and 2015;

sum(mean_jan_clust_2014[,1] == mean_jan_clust_2015[,1])
```

```
## [1] 535
```

```
sum(mean_jan_clust_2014[,146] == mean_jan_clust_2015[,146])
```

```
## [1] 318
```