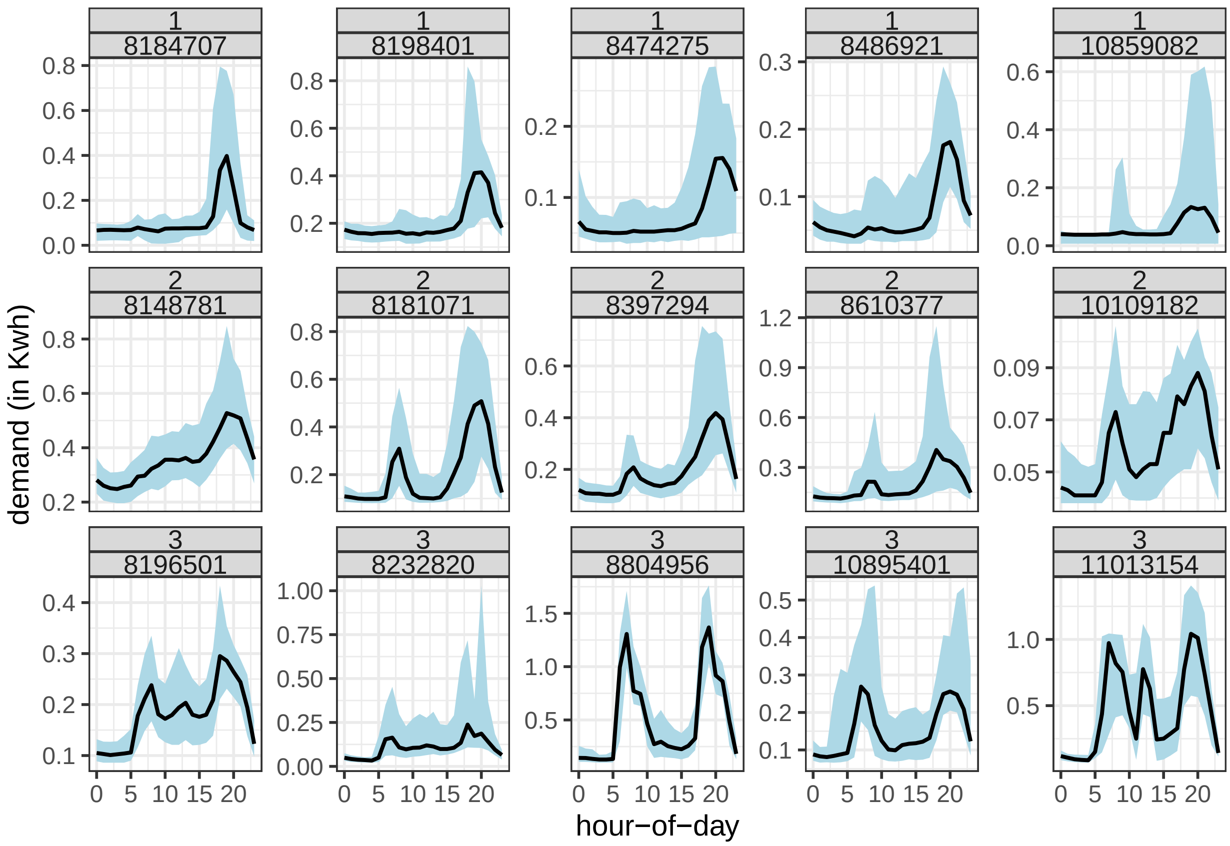
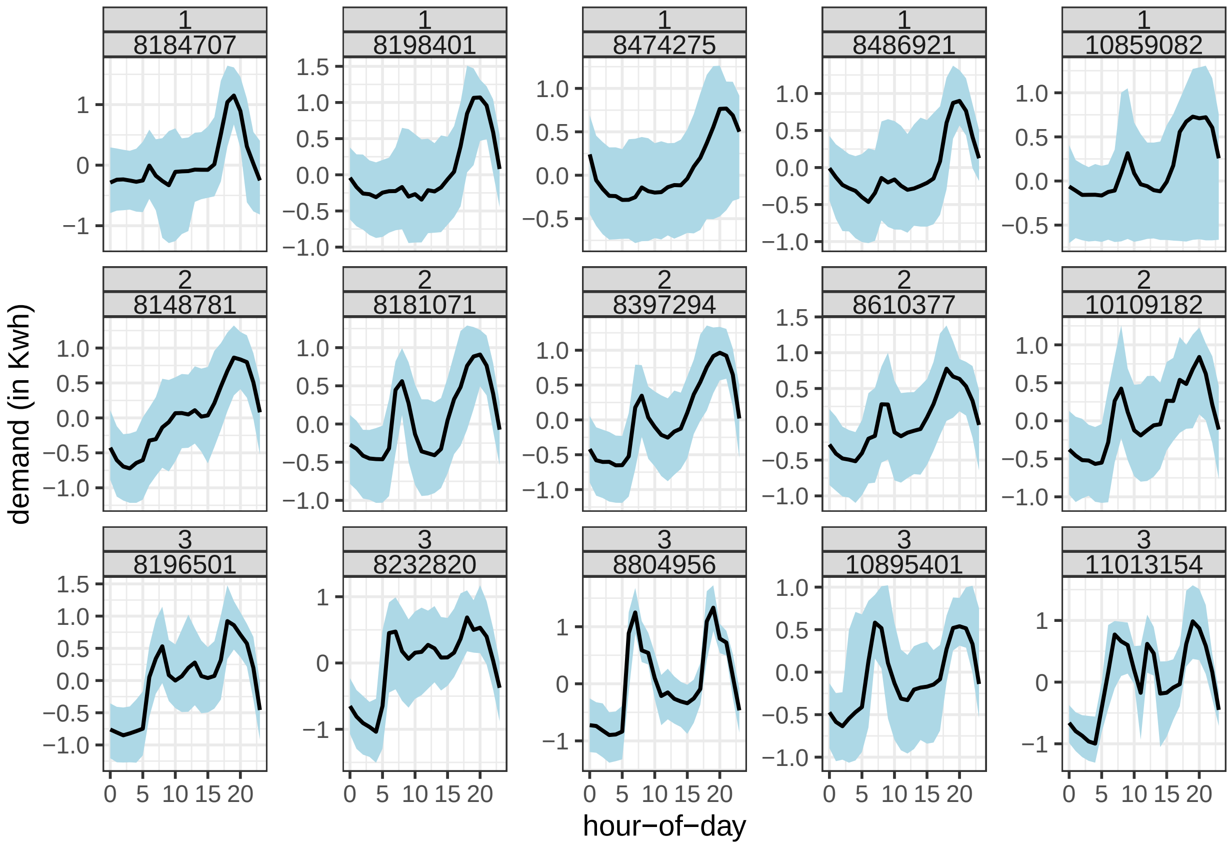
Hand picking similar behaving group of customers to check clustering results

A clean dataset is obtained by choosing minimum sum of JS distances from each typical customer based on $quantiles = seq(0.1, 0.9, 0.1)$. The objective is to see if the clustering algorithm then picks the least distant ones as the group.

only hod



Do they look similar on the transformed scale?



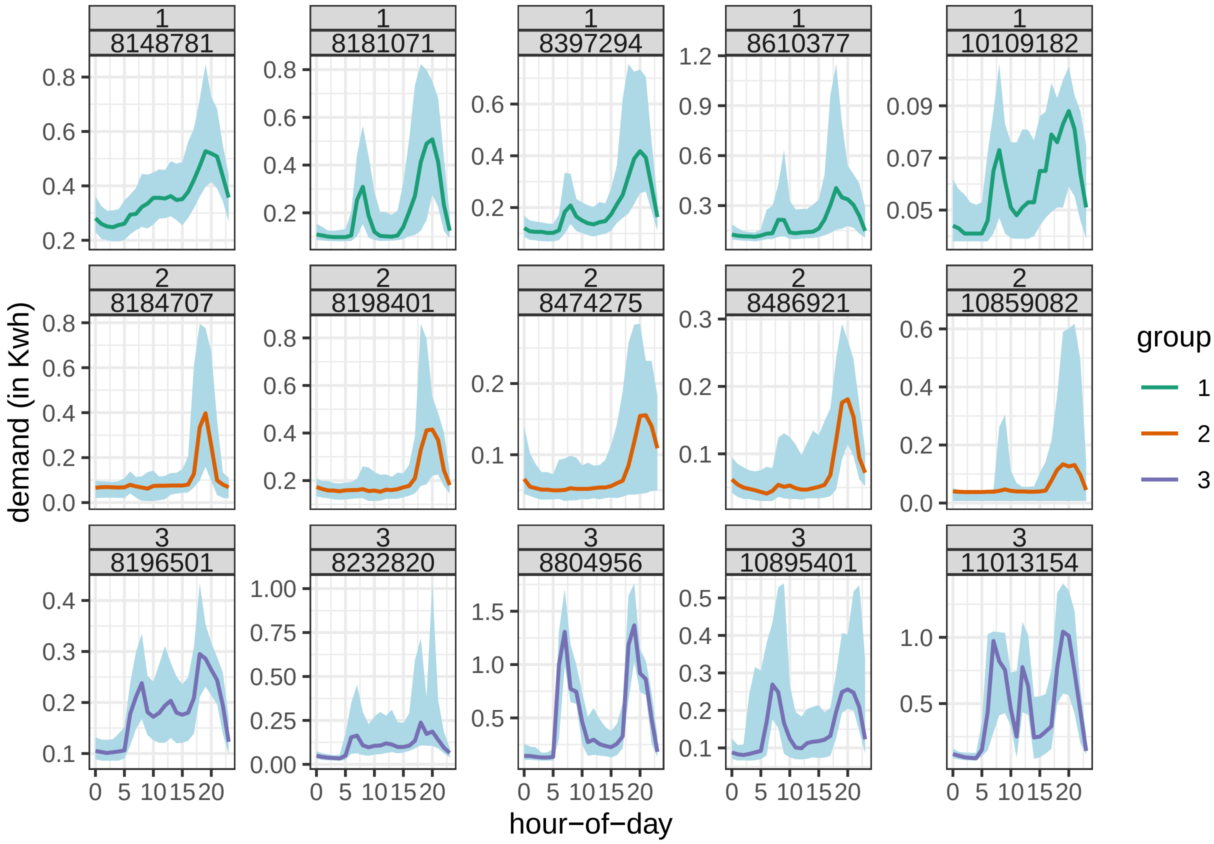
only hod

Does hod as the only variable correctly identifies the groups?

```
#quantile_prob_val = c(0.5, 0.75)
#data_pick <- data_pick %>% filter(!(customer_id %in% c(8485375, 8952846)))
library(gracsr)
v2 <- suppressWarnings(
  scaled_dist_gran(data_pick, "hour_day",
                   response = "general_supply_kwh",
                   quantile_prob_val = quantile_prob_clust)) %>% rename("dist_hod" = "dist")
v3 <- suppressWarnings(
  scaled_dist_gran(data_pick, "day_month",
                   response = "general_supply_kwh",
                   quantile_prob_val = quantile_prob_clust)) %>% rename("dist_dom" = "dist")

data_dist <- v3 %>%
  left_join(v2) %>%
  mutate(dist = dist_hod + dist_dom) %>%
    pivot_wider(-c(3, 4),
                names_from = customer_to,
              values_from = dist) %>%
  rename("customer_id" = "customer_from")
```

```
## # A tibble: 3 x 2
##   group     n
##   <int> <int>
```

```
## 1        1        5
## 2        2        5
## 3        3        5
```

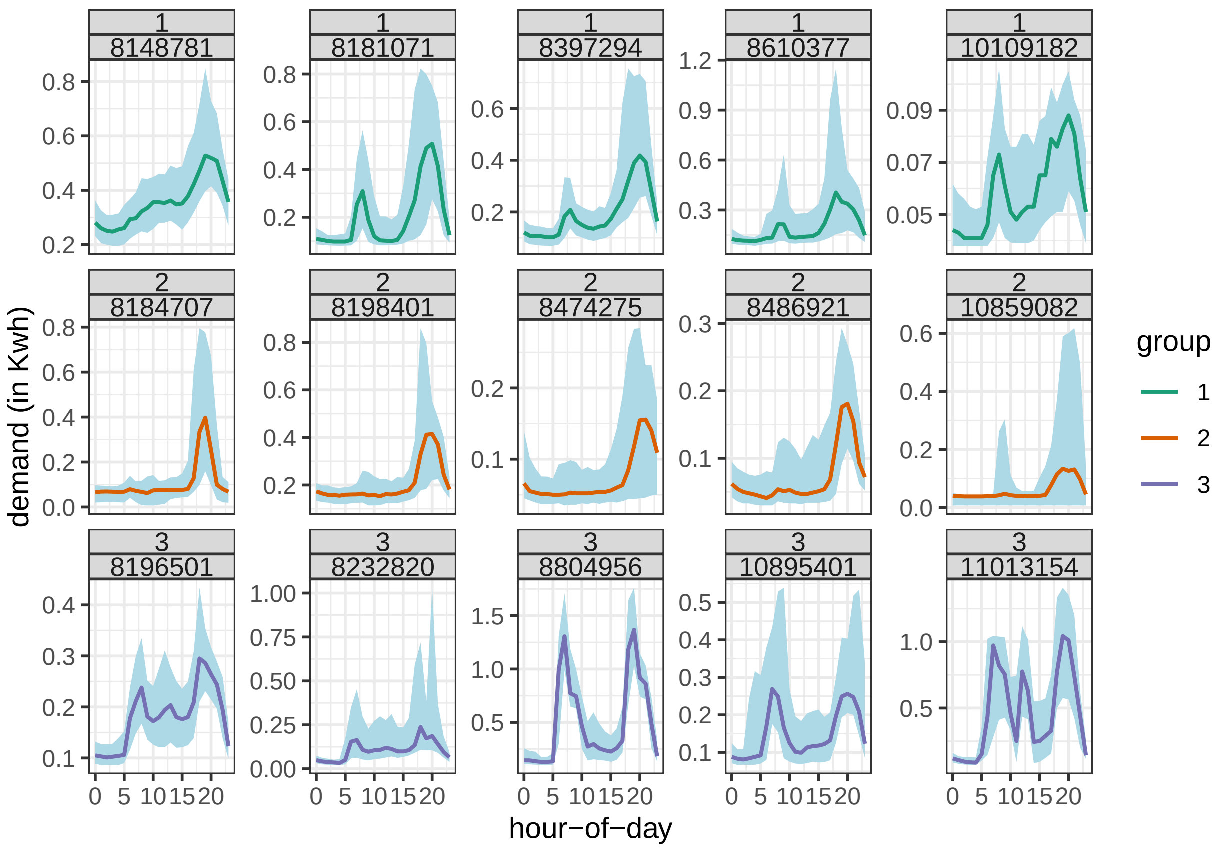


hod + dom

Is the clustering sensitive to nuisance parameter?

```
#quantile_prob_val = c(0.5, 0.75)
#data_pick <- data_pick %>% filter(!(customer_id %in% c(8485375, 8952846)))
library(gracsr)
v2 <- suppressWarnings(
  scaled_dist_gran(data_pick, "hour_day",
                   response = "general_supply_kwh",
                   quantile_prob_val = quantile_prob_clust)) %>% rename("dist_hod" = "dist")
v3 <- suppressWarnings(
  scaled_dist_gran(data_pick, "day_month",
                   response = "general_supply_kwh",
                   quantile_prob_val = quantile_prob_clust)) %>% rename("dist_dom" = "dist")

data_dist <- v3 %>%
  left_join(v2) %>%
  mutate(dist = dist_hod + dist_dom) %>%
    pivot_wider(-c(3, 4),
                names_from = customer_to,
              values_from = dist) %>%
  rename("customer_id" = "customer_from")
```

```
## # A tibble: 3 x 2
##   group     n
##   <int> <int>
## 1     1     5
## 2     2     5
## 3     3     5
```



All 100 customers

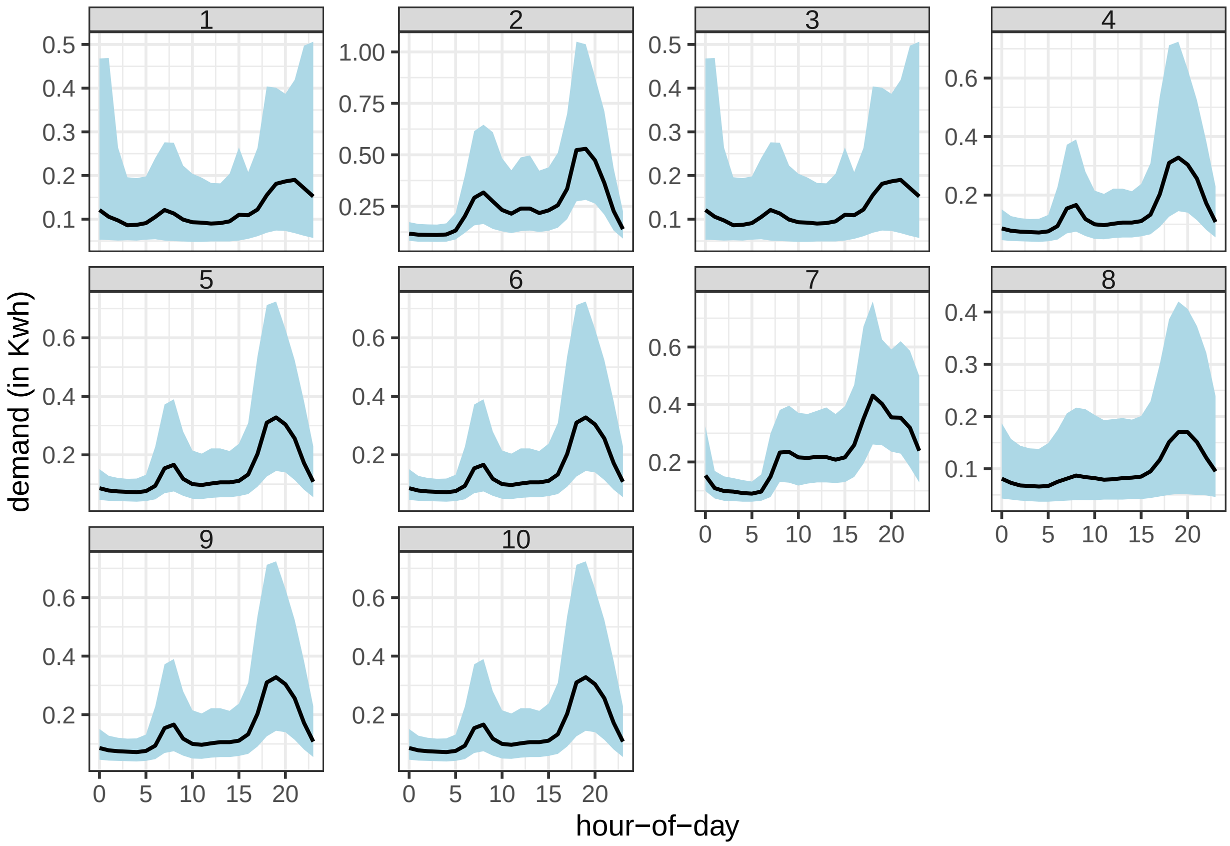
Running it on all 100 and making 10 clusters. Do they have sufficiently different shapes?

Variable importance

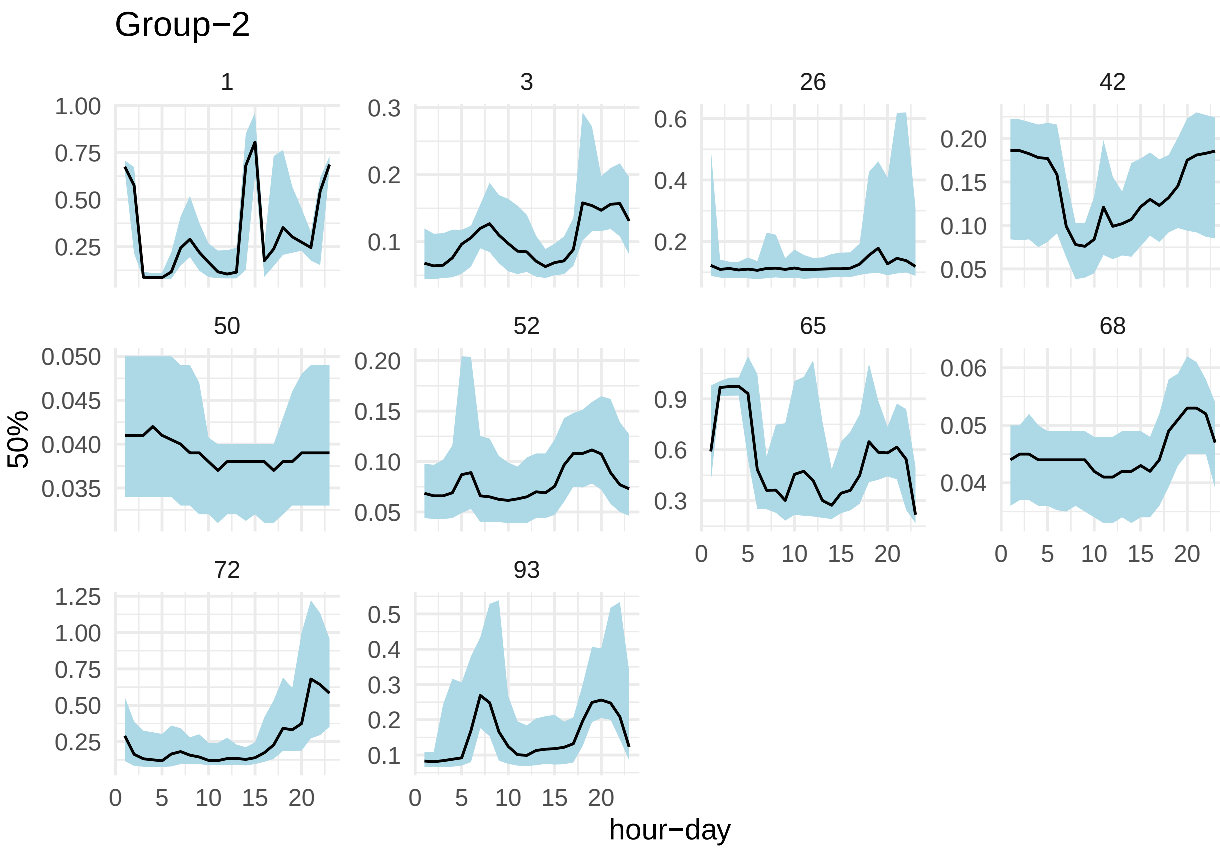
```
# library(gracsr)
# v2 <- suppressWarnings(
#   scaled_dist_gran(data, "hour_day",
#                    response = "general_supply_kwh",
#                    quantile_prob_val = quantile_prob_clust)) %>% rename("dist_hod" = "dist")

#write_rds(v2, "../../../data/scaled_dist_100clust_deciles.rds")
v2 <- read_rds("../../../data/scaled_dist_100clust_deciles.rds")

data_dist <- v2 %>%
  mutate(dist = dist_hod) %>%
    pivot_wider(-c(3, 4),
                names_from = customer_to,
              values_from = dist) %>%
```

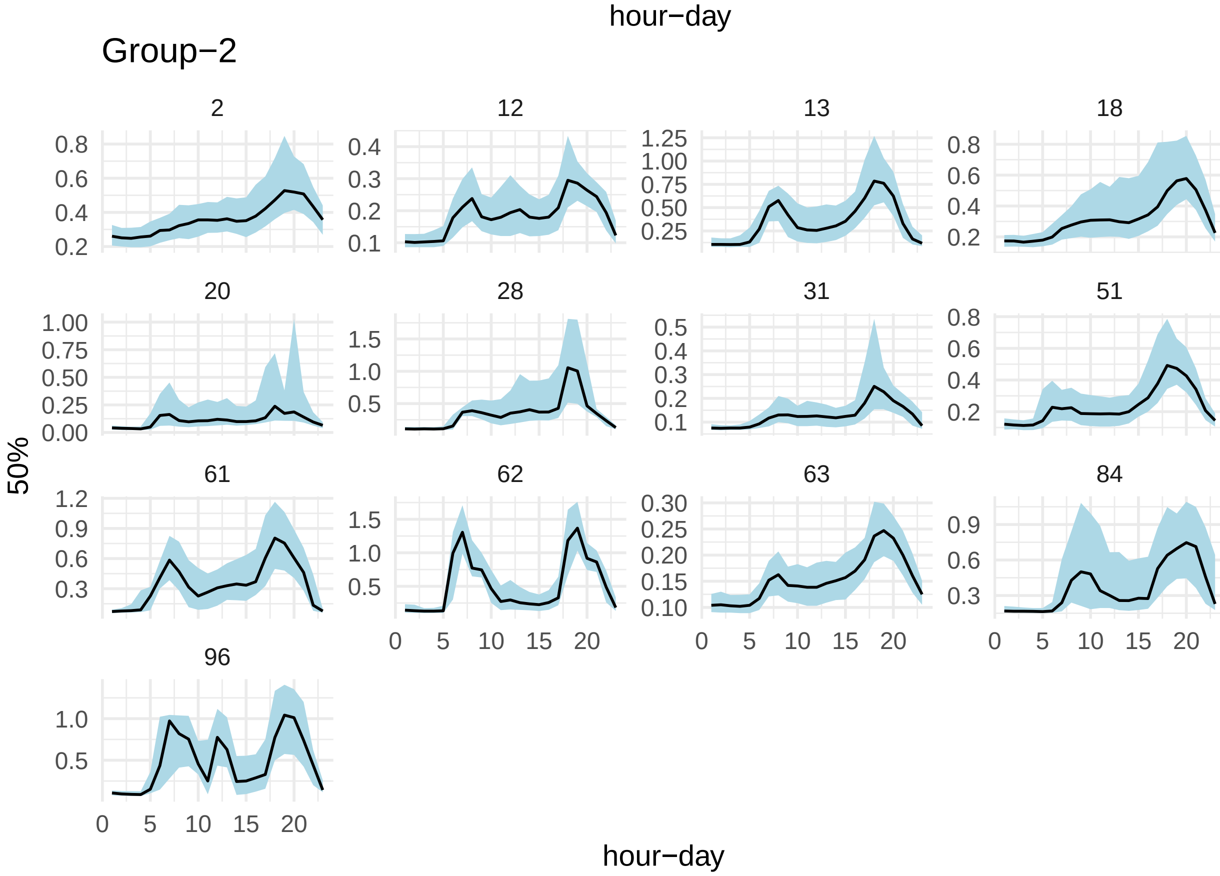
```r
  rename("customer_id" = "customer_from")
```

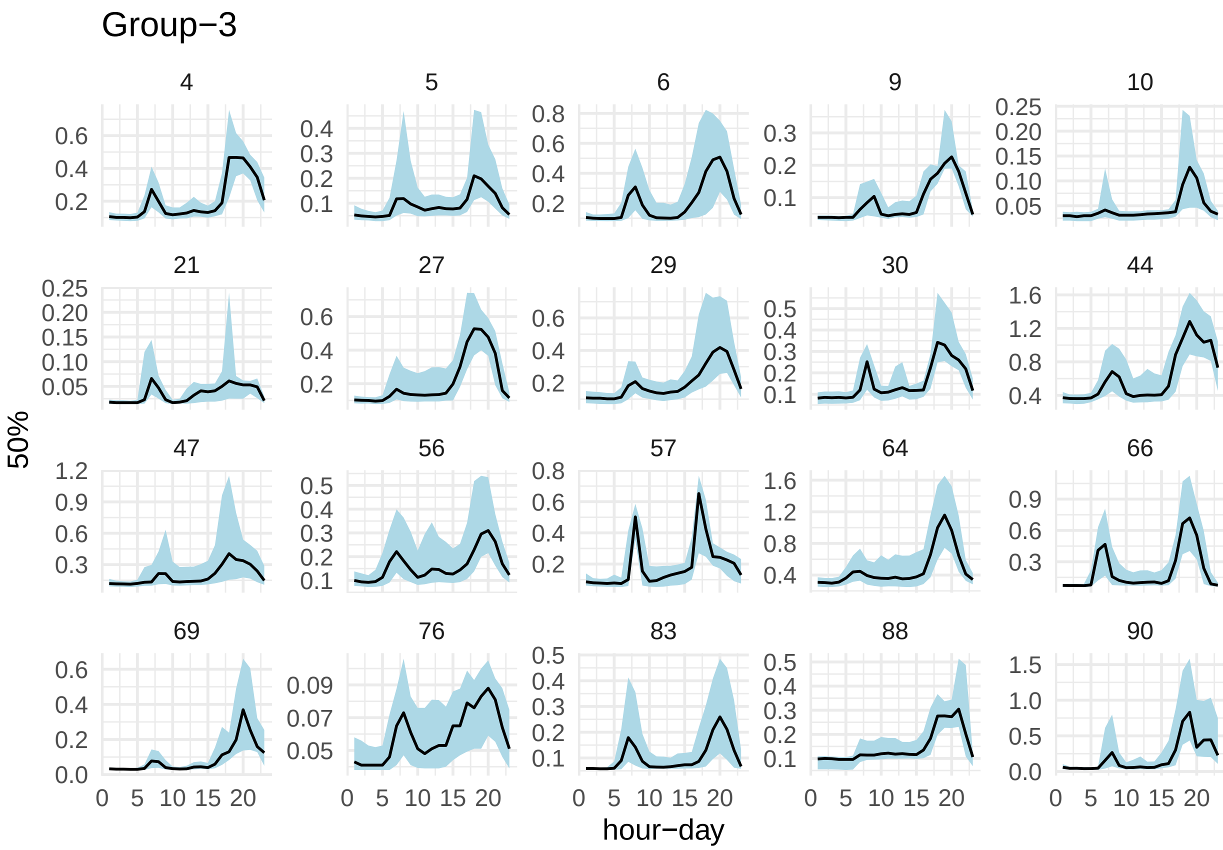
```
## # A tibble: 10 x 2
##    group     n
##    <int> <int>
##  1     1    10
##  2     2    13
##  3     3    20
##  4     4    10
##  5     5    19
##  6     6    10
##  7     7     2
##  8     8     5
##  9     9     5
## 10    10     6
```

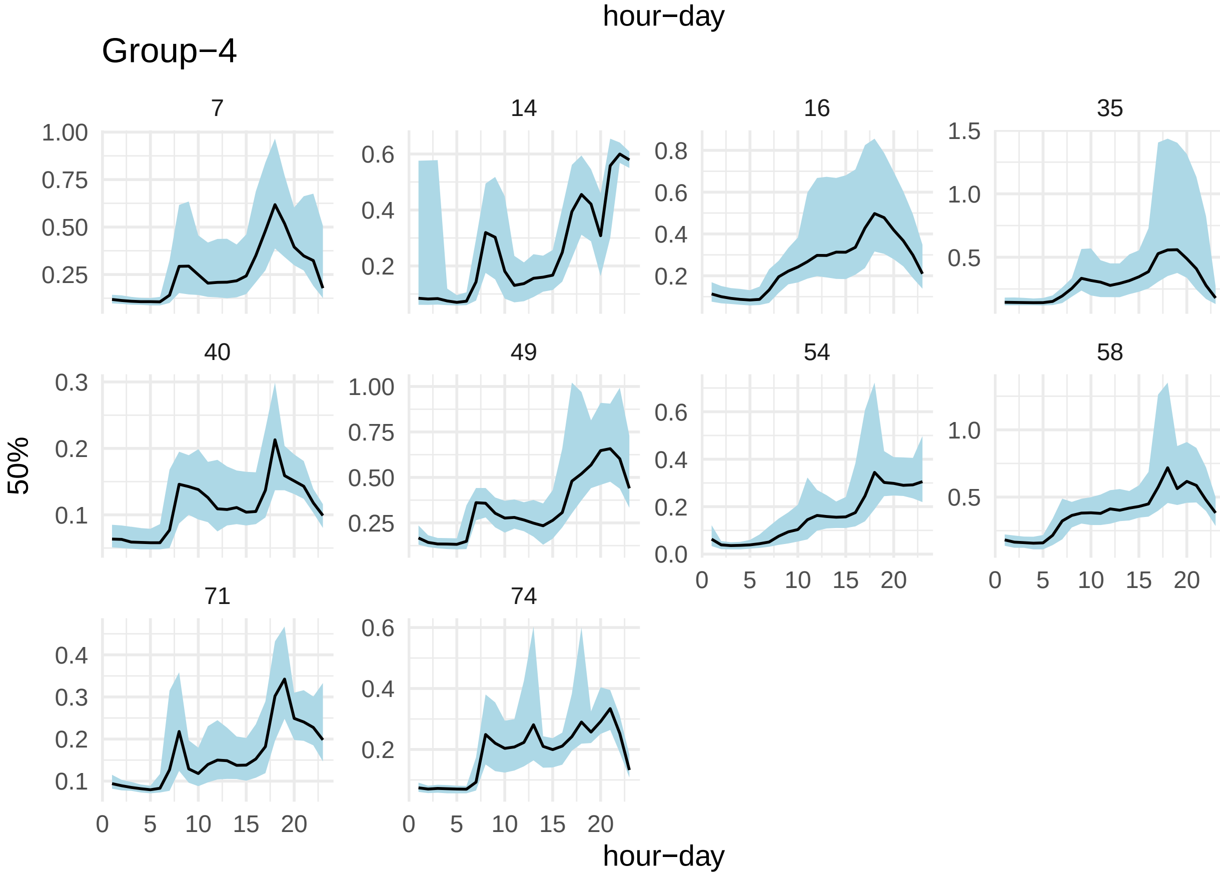


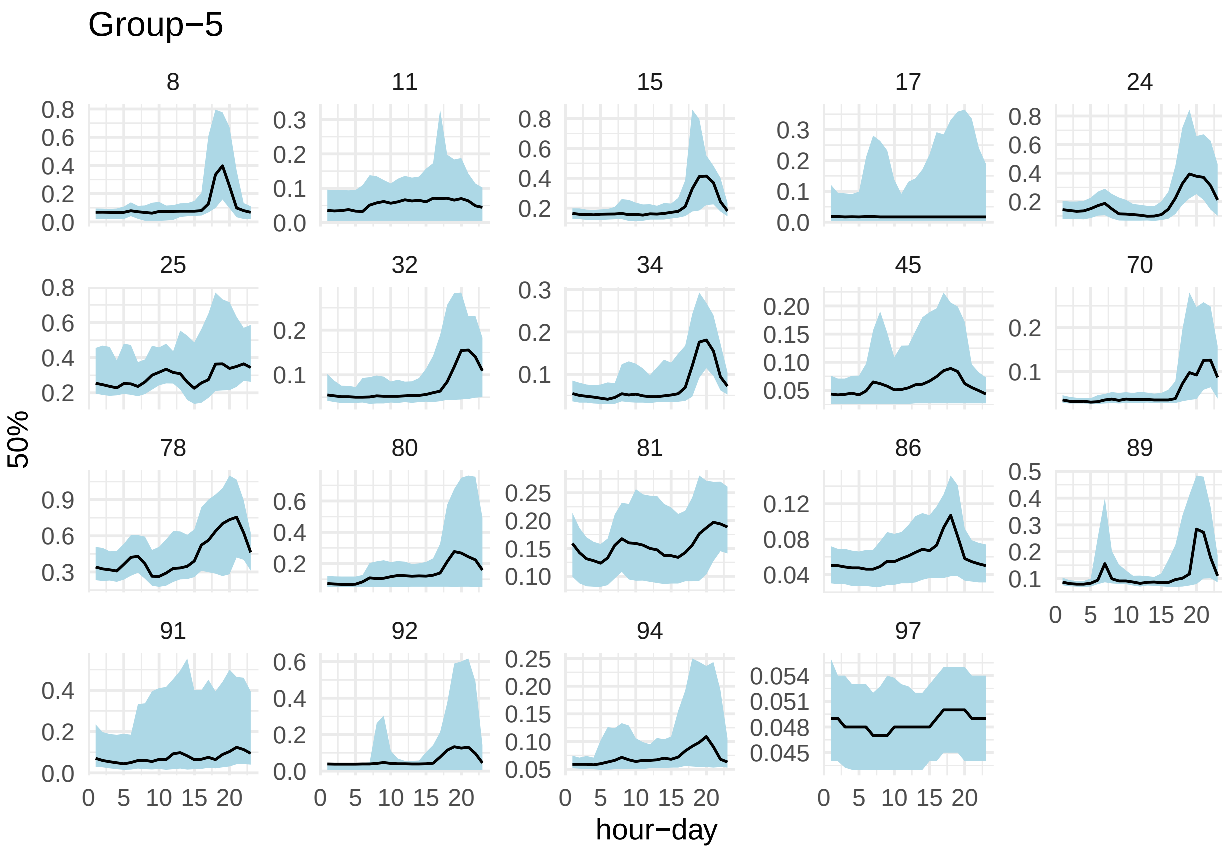
Split these groups to see if the shapes of individual customers in a group is the same

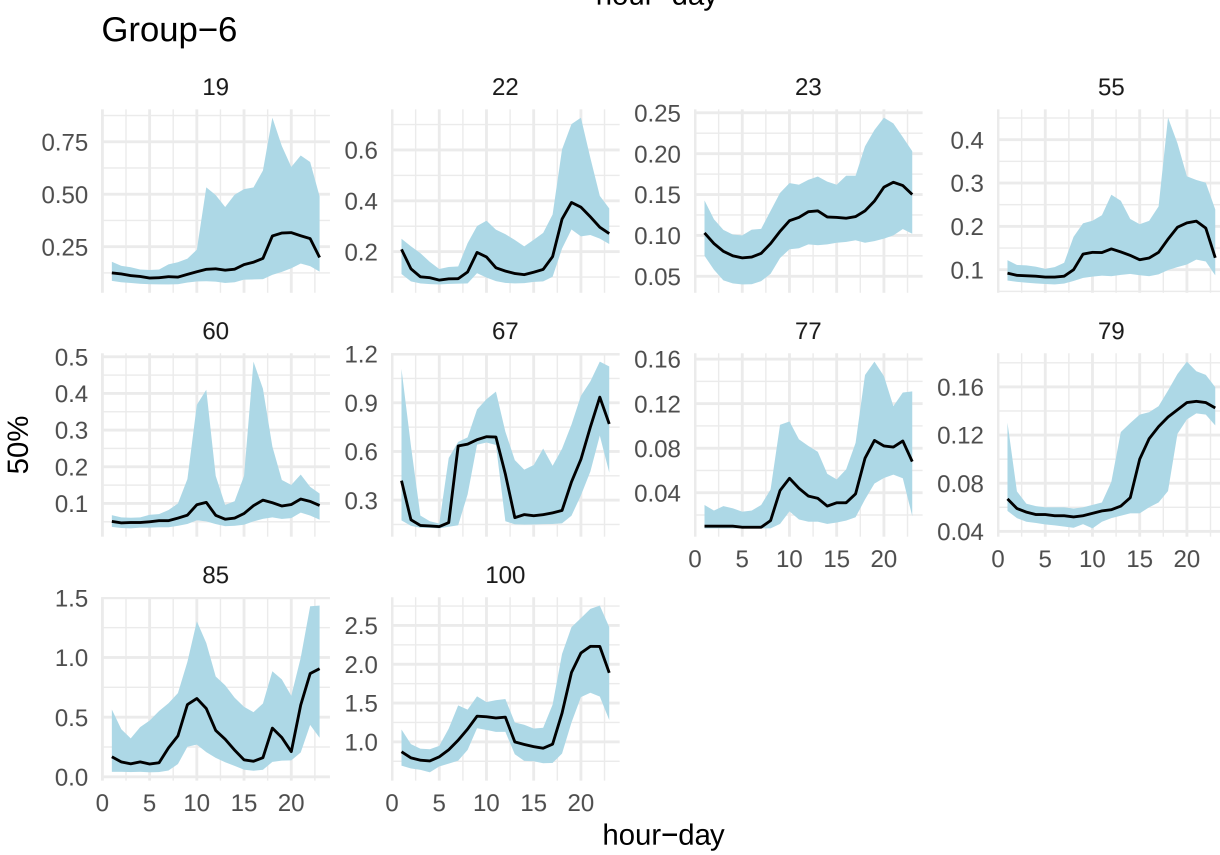

Group−2
1
3
26
42
50
52
65
68
72
93
50%
hour−day

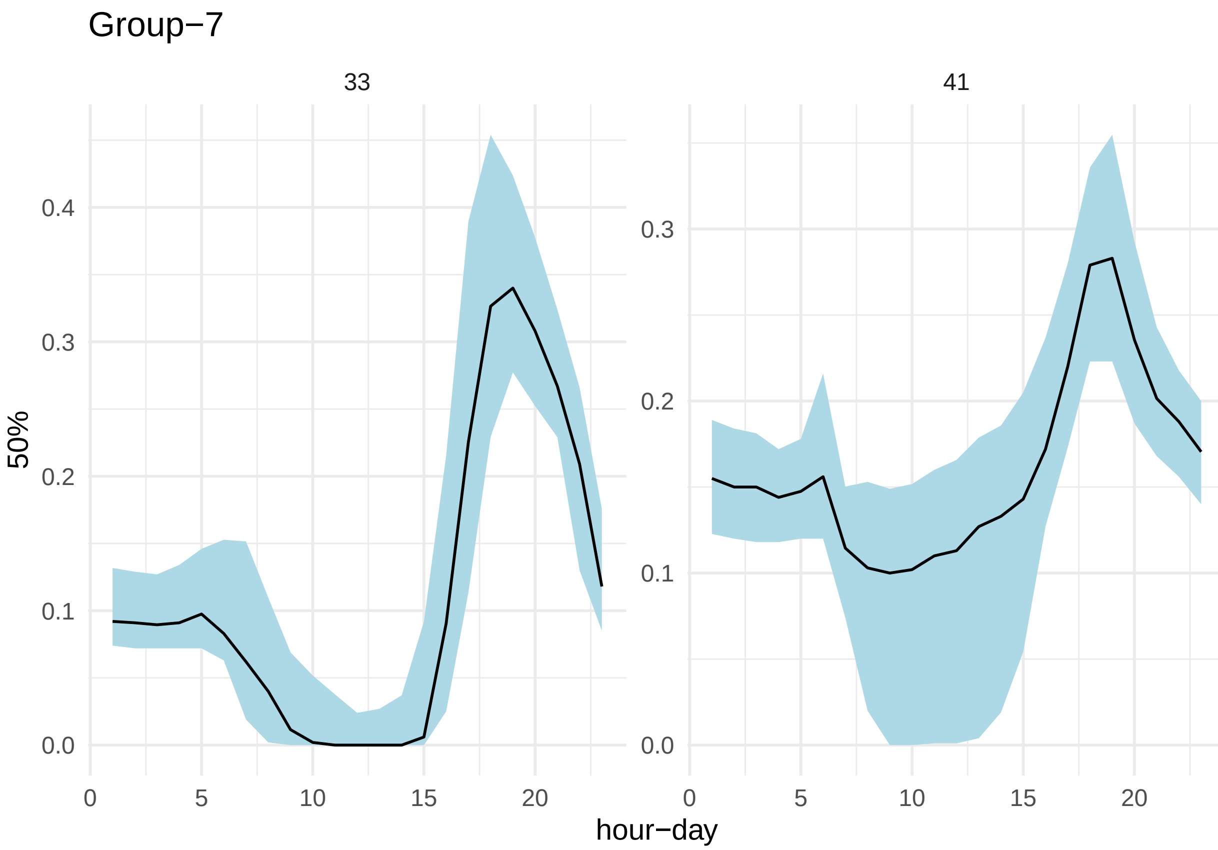

Group−2
2
12
13
18
20
28
31
51
61
62
63
84
96
50%
hour−day

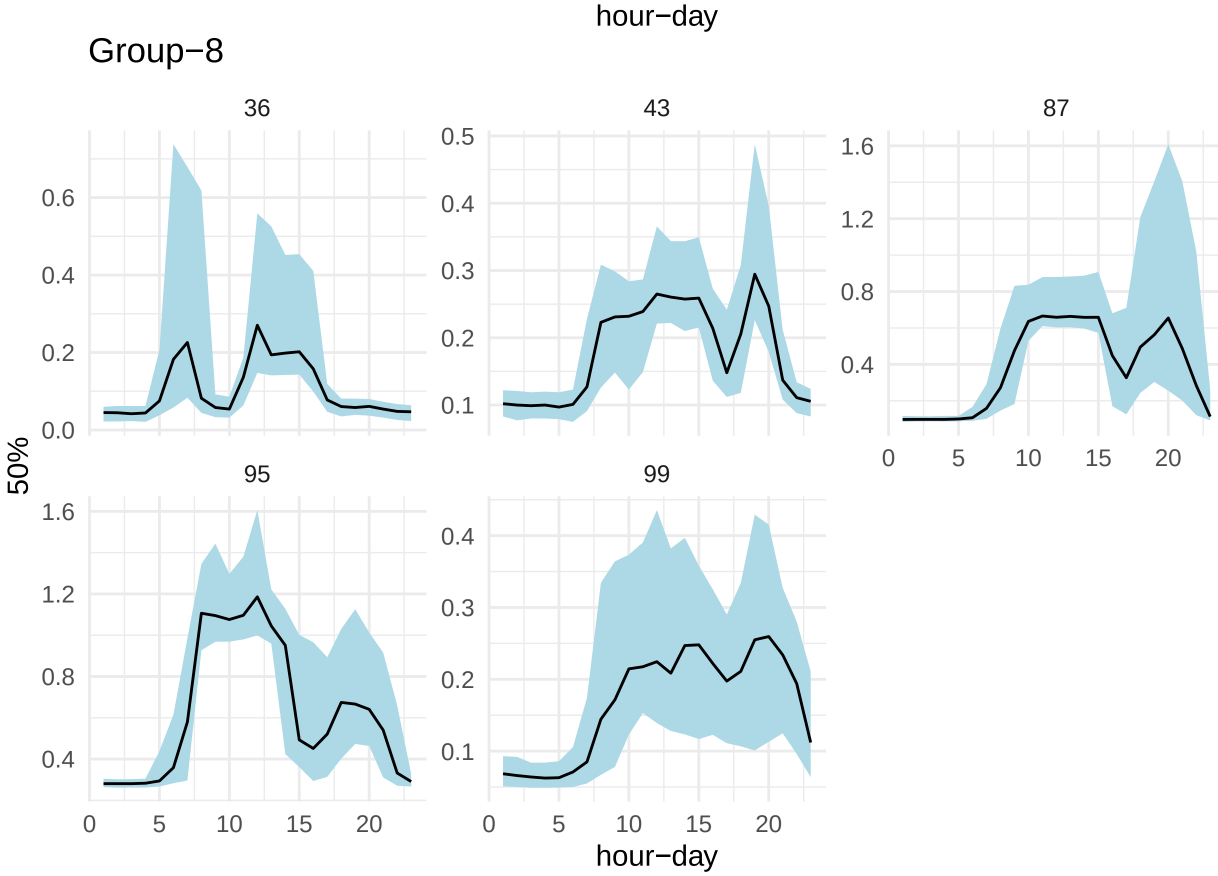

Group−3
4
5
6
9
10
21
27
29
30
44
47
56
57
64
66
69
76
83
88
90
50%
hour−day

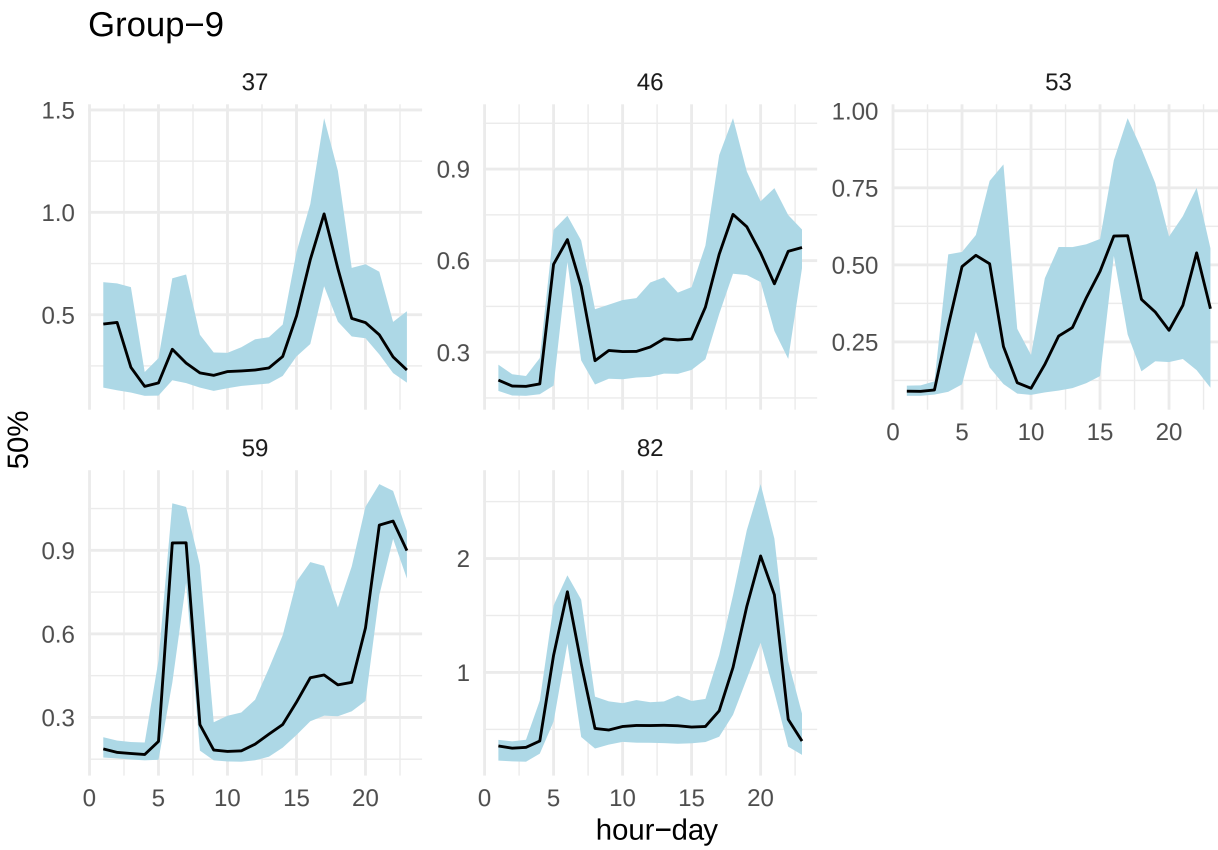

Group−4
7
14
16
35
40
49
54
58
71
74
50%
hour−day

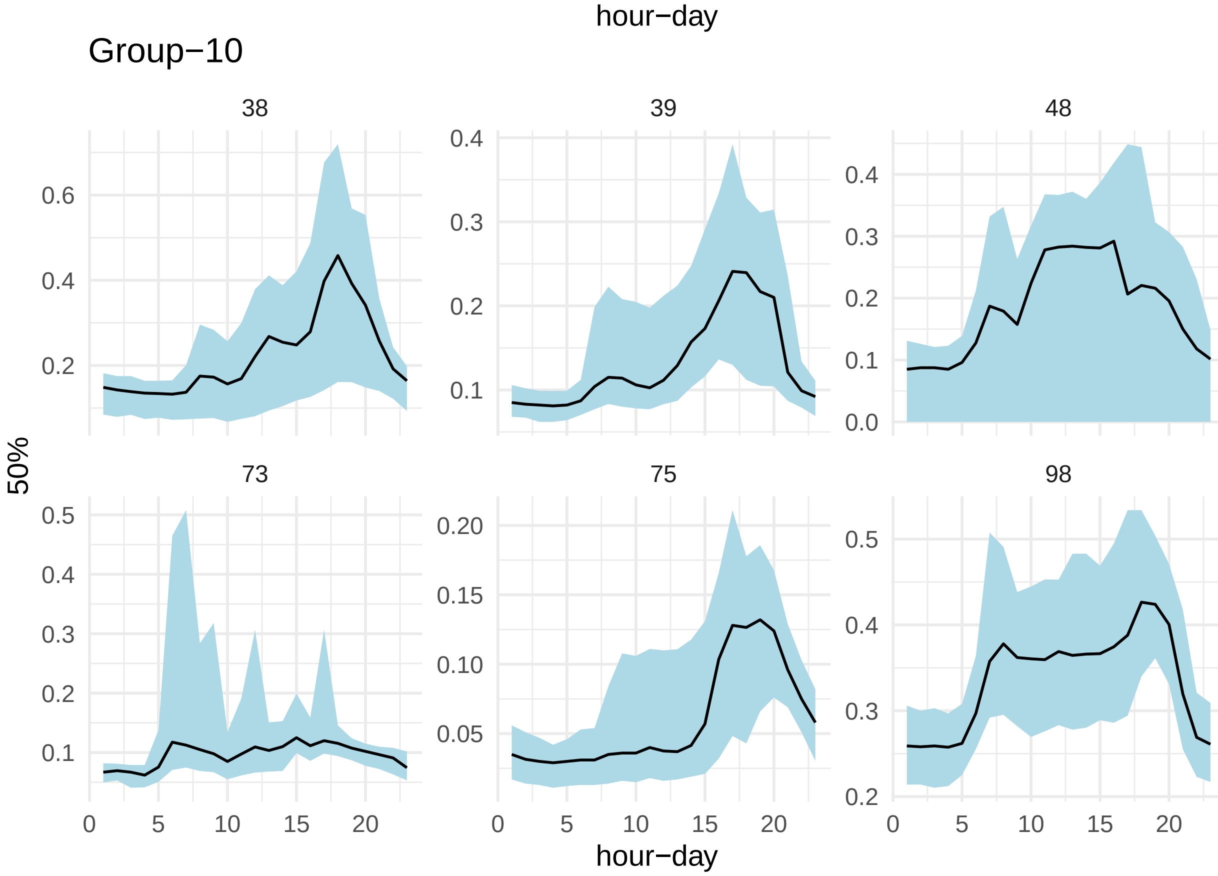

Group−5
8
11
15
17
24
25
32
34
45
70
78
80
81
86
89
91
92
94
97
50%
hour−day


Group−6
19
22
23
55
60
67
77
79
85
100
50%
hour−day

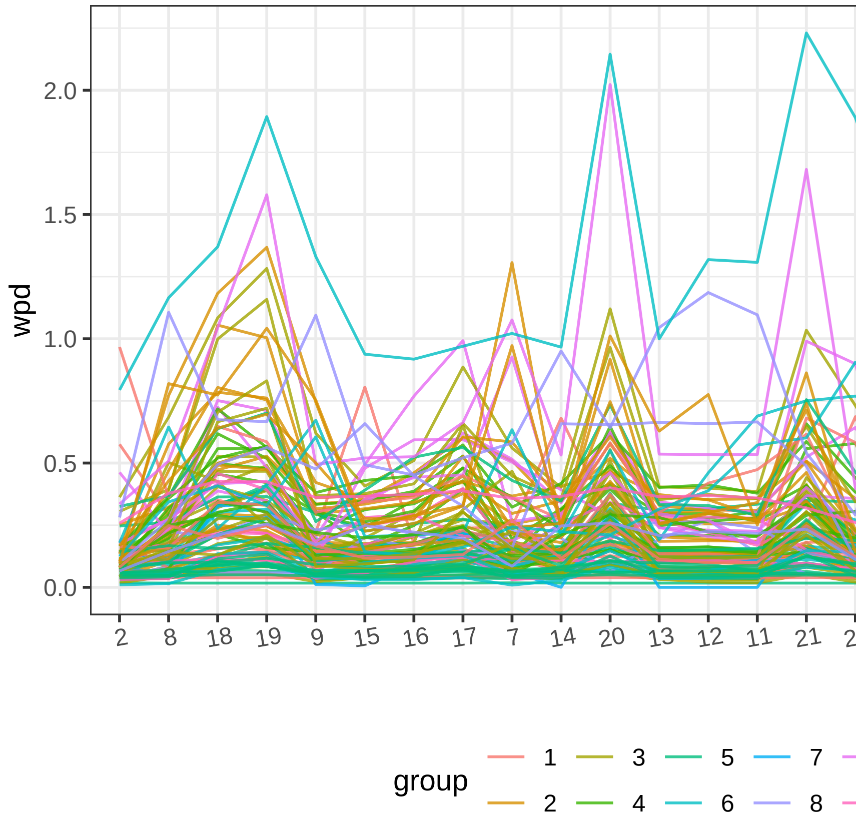

Group−7
33
41
50%
hour−day
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20


Group−8
36
43
87
95
99
50%
hour−day
0.0
0.2
0.4
0.6
0.1
0.2
0.3
0.4
0.5
0.4
0.8
1.2
1.6
0
5
10
15
20

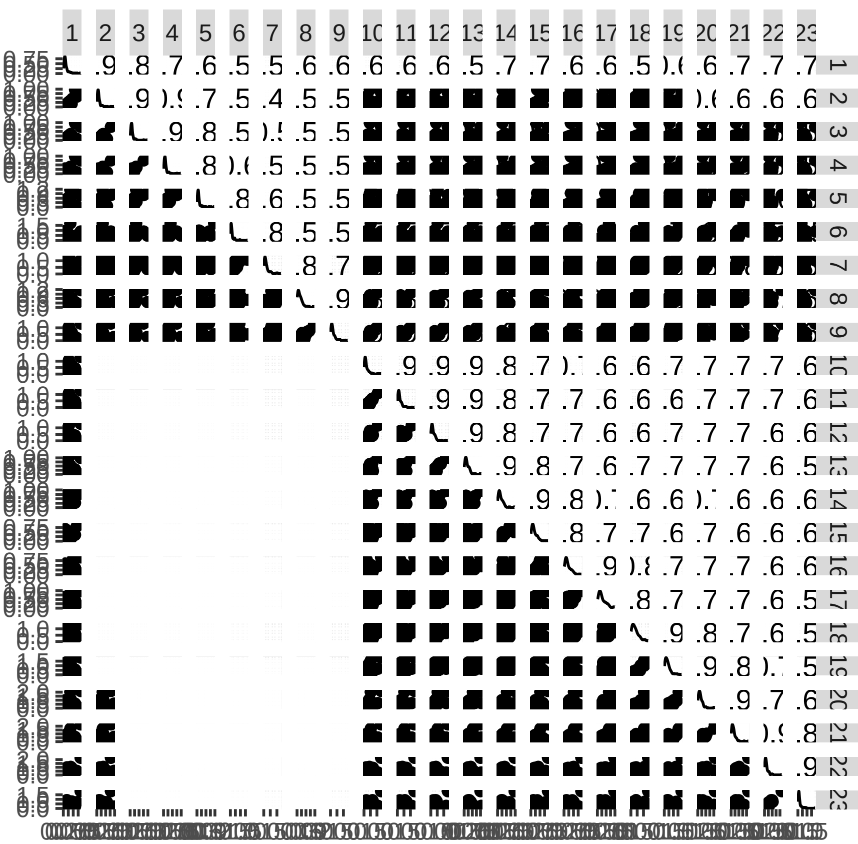

Group−9
37
46
53
59
82
50%
hour−day


Group−10
38
39
48
73
75
98
50%
hour−day

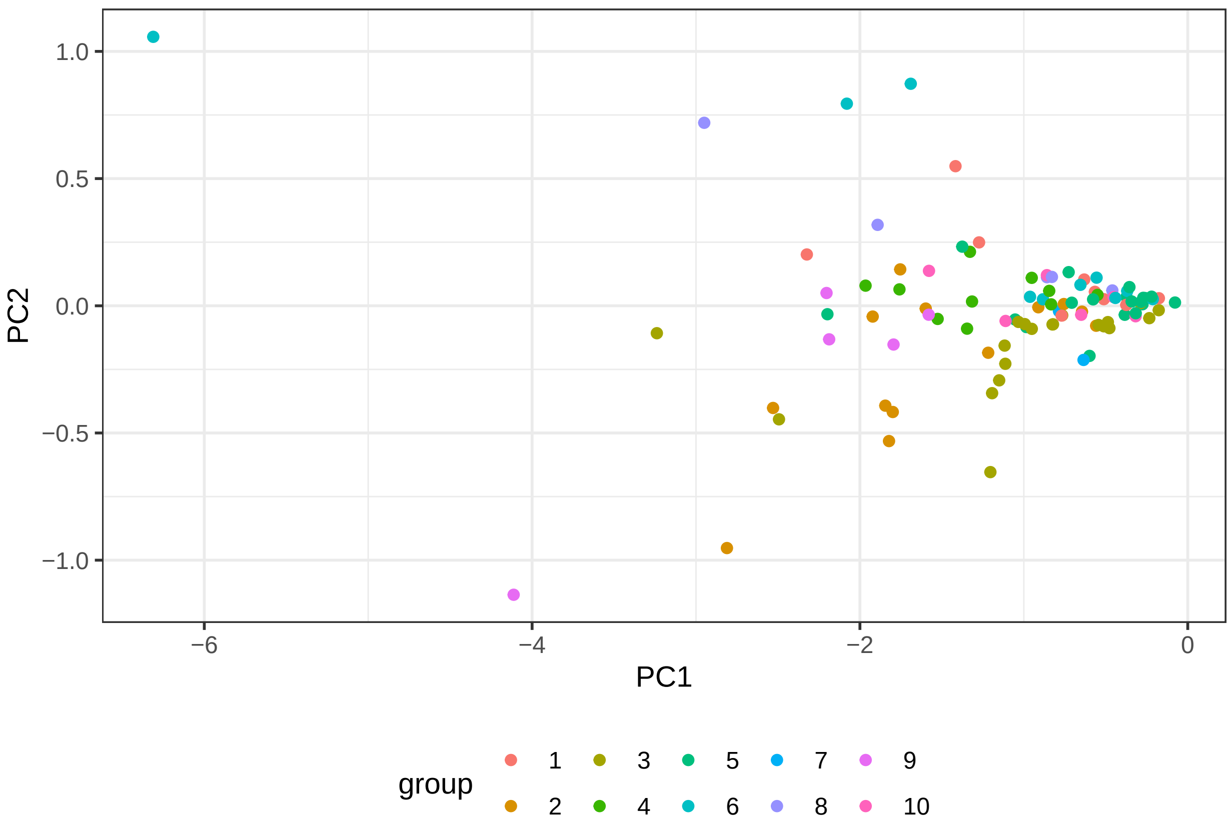
Which variables are important for this clustering



I can do a ggpairs or parallel coordinate plot for this.



```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.4512 0.28362 0.25449 0.21064 0.19584 0.14632 0.12747
## Proportion of Variance 0.8761 0.03346 0.02694 0.01846 0.01596 0.00891 0.00676
## Cumulative Proportion  0.8761 0.90955 0.93649 0.95495 0.97091 0.97981 0.98657
##                            PC8     PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.09963 0.08240 0.06427 0.0581 0.04579 0.04222 0.03271
## Proportion of Variance 0.00413 0.00282 0.00172 0.0014 0.00087 0.00074 0.00045
## Cumulative Proportion  0.99070 0.99353 0.99524 0.9967 0.99752 0.99826 0.99871
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.02875 0.02803 0.02356 0.01698 0.01545 0.01257 0.01177
## Proportion of Variance 0.00034 0.00033 0.00023 0.00012 0.00010 0.00007 0.00006
## Cumulative Proportion  0.99905 0.99938 0.99961 0.99973 0.99983 0.99990 0.99995
##                           PC22     PC23
## Standard deviation     0.00831 0.006646
## Proportion of Variance 0.00003 0.000020
## Cumulative Proportion  0.99998 1.000000
```



See if month_year works for your case

Try with two granularities