Clustering real data with five customers and choosing optimal number of clusters for hours of day and day of week

1 Clutering approach

1. *Compute quantiles of distributions across each combination of hour of day and day of week*
2. *Compute JS distance between households for each combination of categories*
3. *Total distance between households computed as sum of JS distances for all combinations*
4. *Cluster using this distance with hierarchical clustering algorithm (method "complete")*

2 Plots of raw data

Five households are considered for the following analysis from the SGSC data set. The data sets contain three columns `customer_id`, `reading_datetime` and `general_supply_kwh`. They consist of half-hourly data from 2012-2014. Figure 1 shows how the raw time plots for these five households look like. Since all the dataset looks squeezed on this linear time scale, Figure 2 shows how the raw plot looks for 2 months (Sept 2013 and Oct 2013).

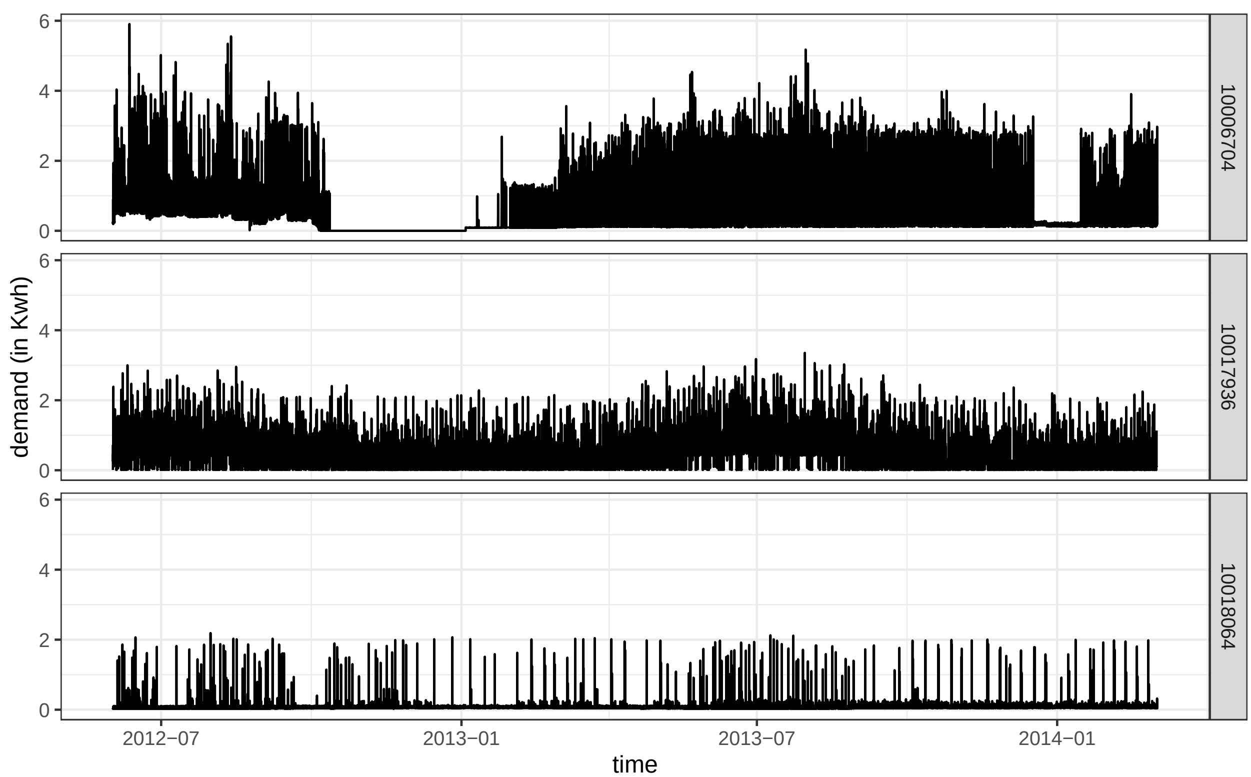


Figure 1: The raw time plots for demand shown for the entire observation period between 2012-2014 for 5 households (facets). The data is squeezed stopping us from seeing any behavioral patterns.

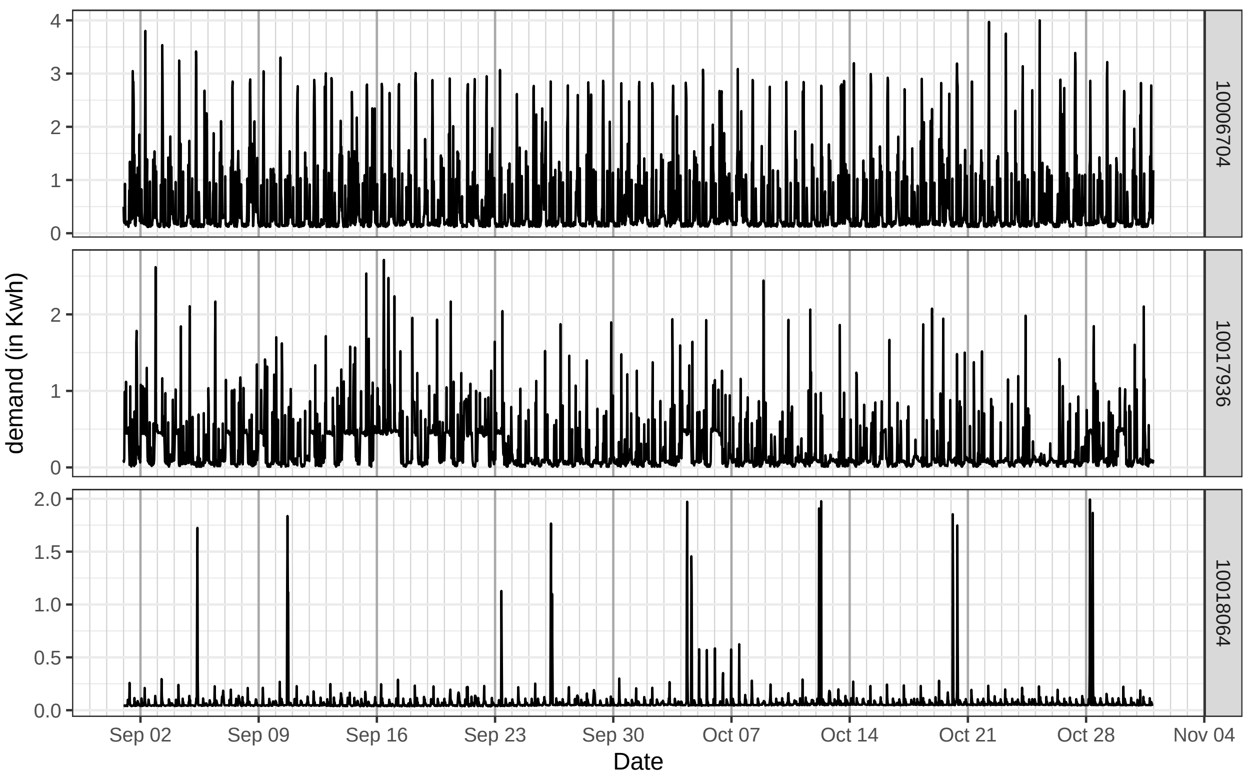


Figure 2: The raw plots for 5 a shown for Sept 2013-Oct 2013. The data is zoomed in and the y-scales made free to emphasize weekly, hourly or any behavior/patterns that they might have. There is some daily and weekly pattern in all households except the 4th one, that has spikes which seem to occur at irregular intervals.

3 Plots of distribution of data across categories of granularities

The distribution of demand across different hours of the day observed for these five households through heatmaps (Figure **??**) and quantile plots (Figure 3). The following characterization is done to validate results of the clustering approach later on.

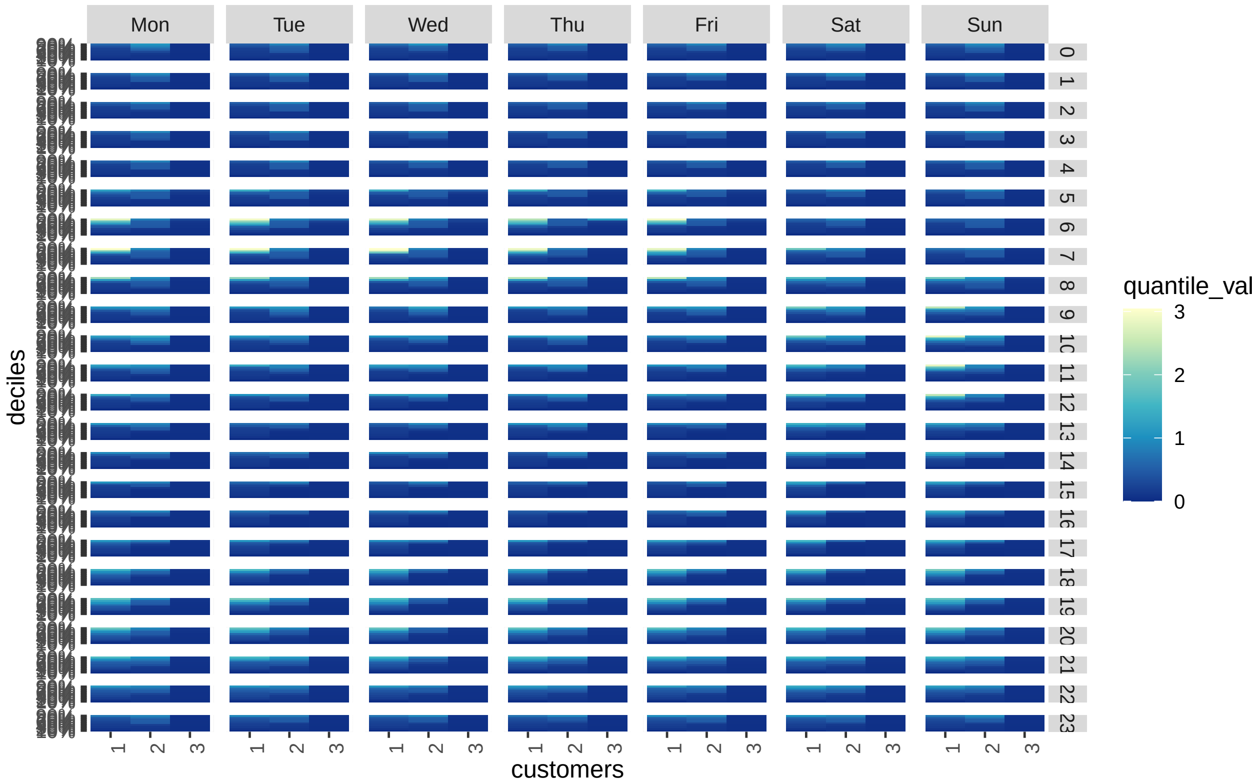
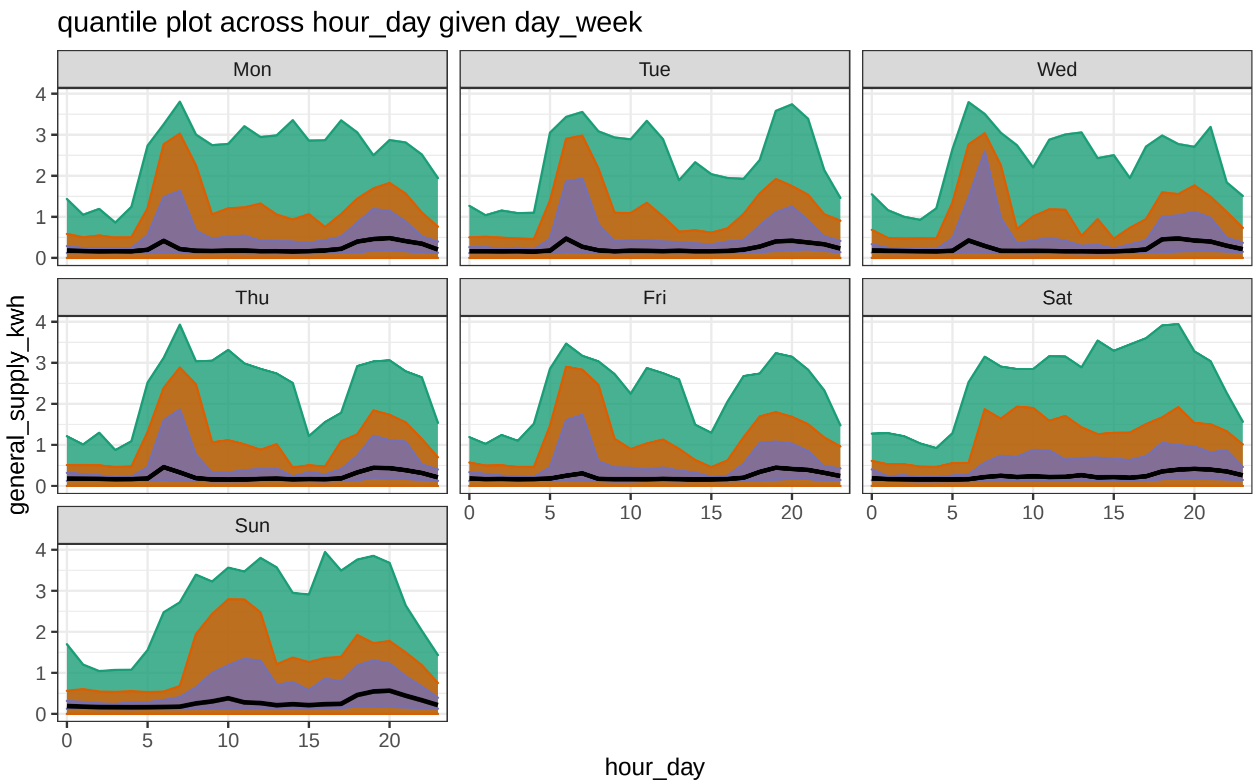
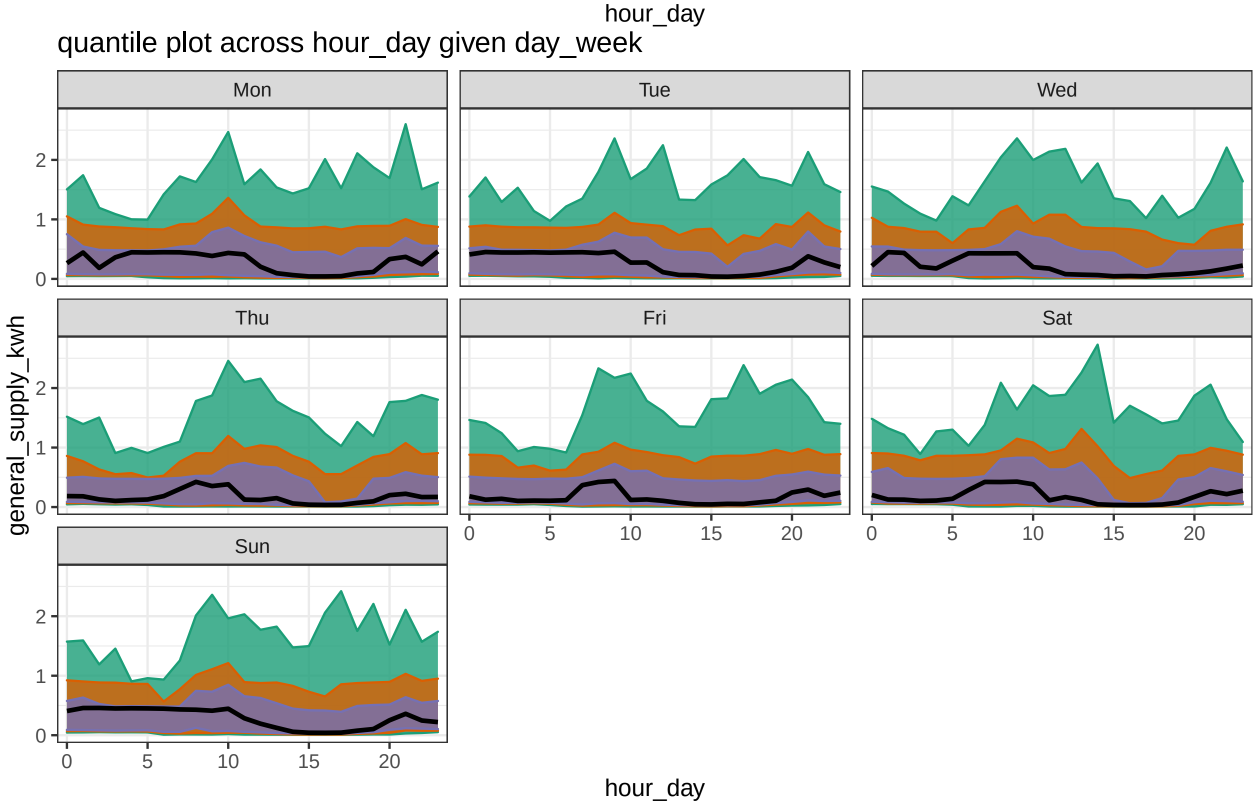
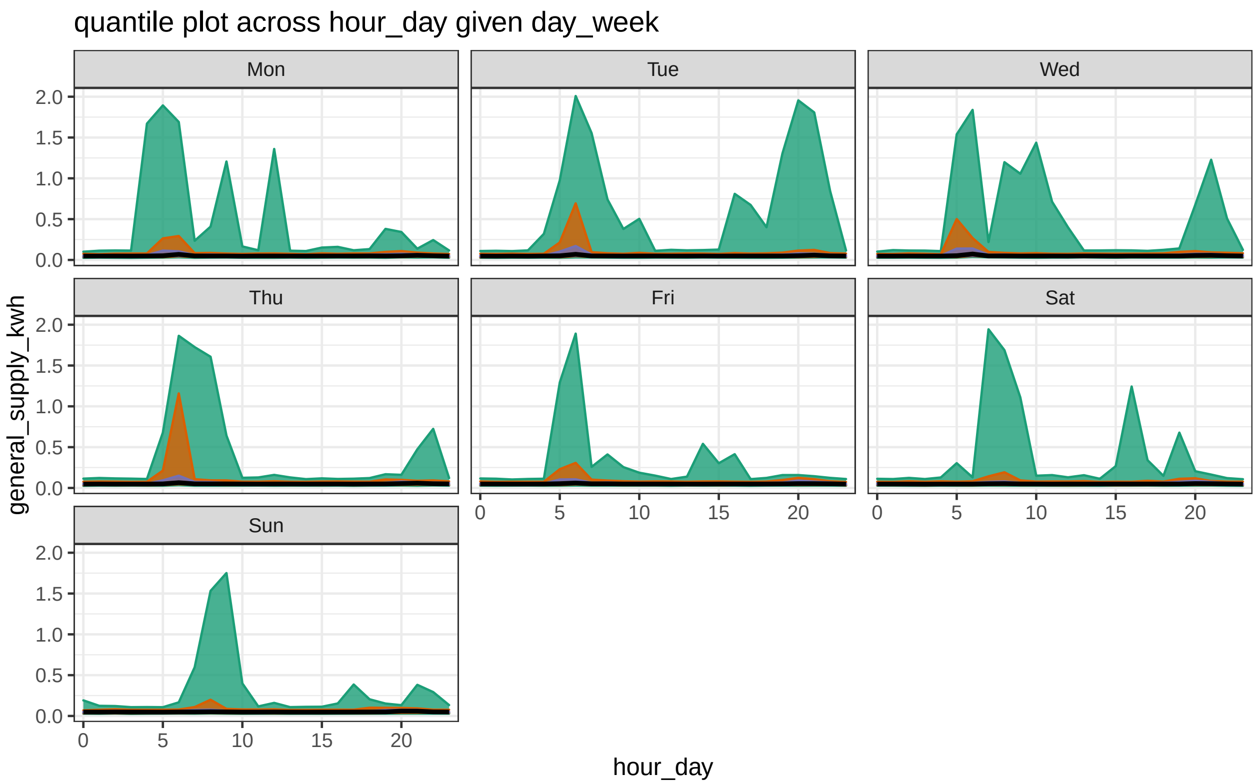


Figure 3: Heatmaps show deciles on the y-axis with customers on the x-axis with colors filled by the value of deciles and faceted by hours of the day. It again seems like all the households have gradual change in colors (almost for all hours) as we move to higher deciles except for the 4th household.


quantile plot across hour_day given day_week
Mon
Tue
Wed
Thu
Fri
Sat
Sun
general_supply_kwh
hour_day


quantile plot across hour_day given day_week
Mon
Tue
Wed
Thu
Fri
Sat
Sun
general_supply_kwh
hour_day



4 Iterations of each customer (adding some random noise for each iteration)

Iterations of each customer are considered by adding random noise ($N(0, \sigma^2)$, where σ^2 is very small relative to the the variance of the data). Figure 4 shows that raw plots of the simulated iterations to show that their structure is same as the parent data set.

5 Dendogram

The following figures show dendograms when we are using optimal number of clusters (k) from k = fpc::nselectboot(), k = 2, 3, 4 and 5. We observe that the 4^{th} household tend to get split as different clusters as increase the number of clusters.

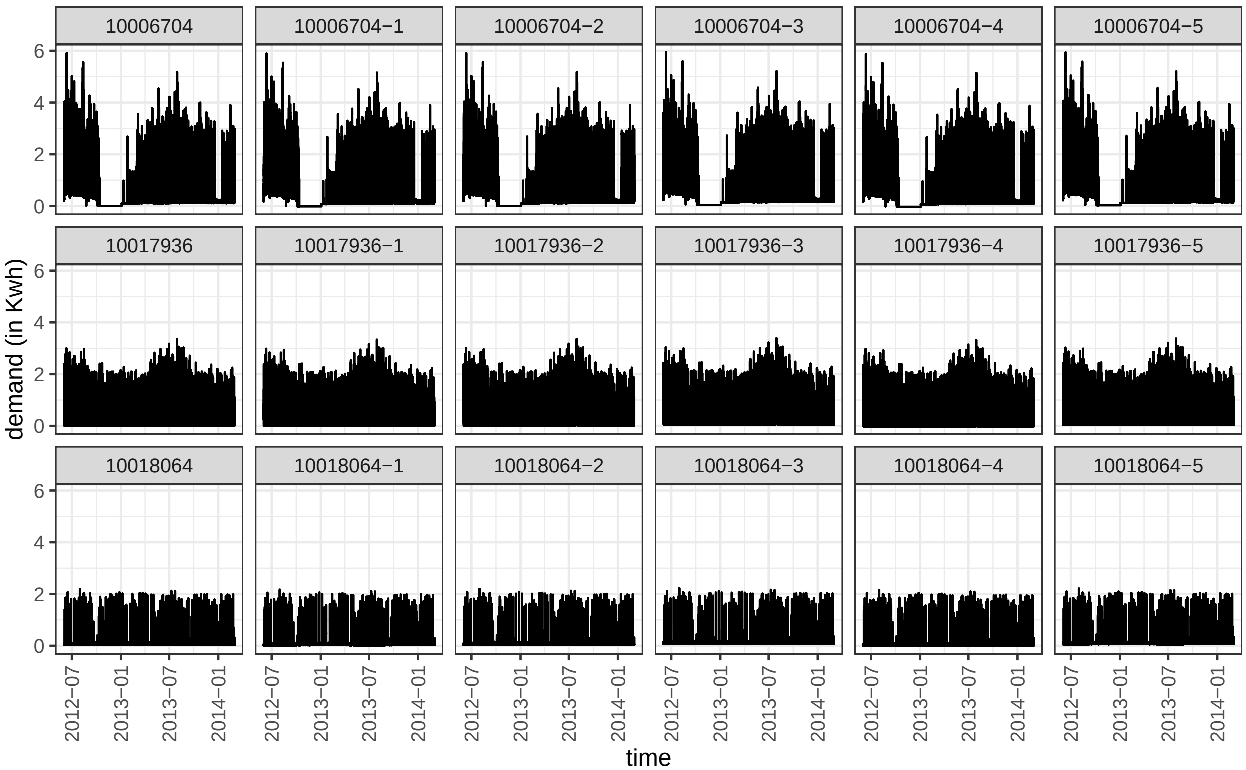
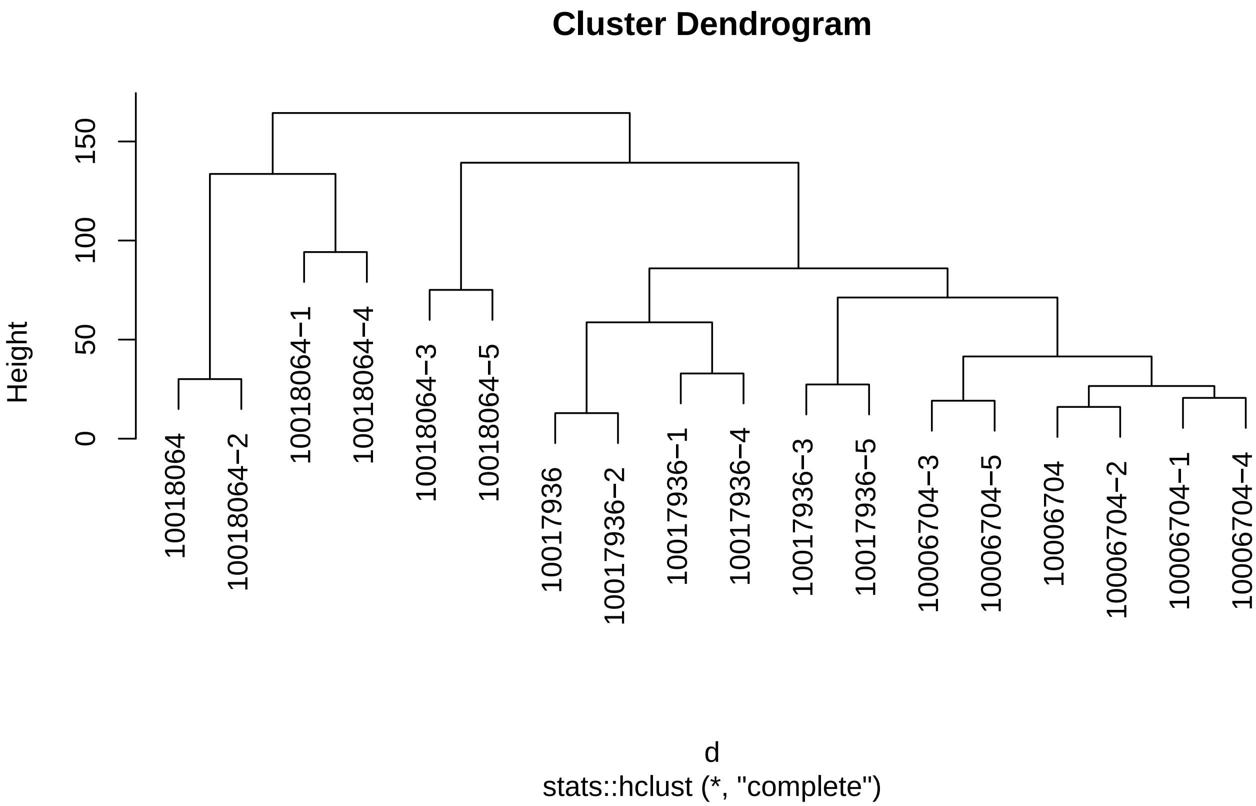
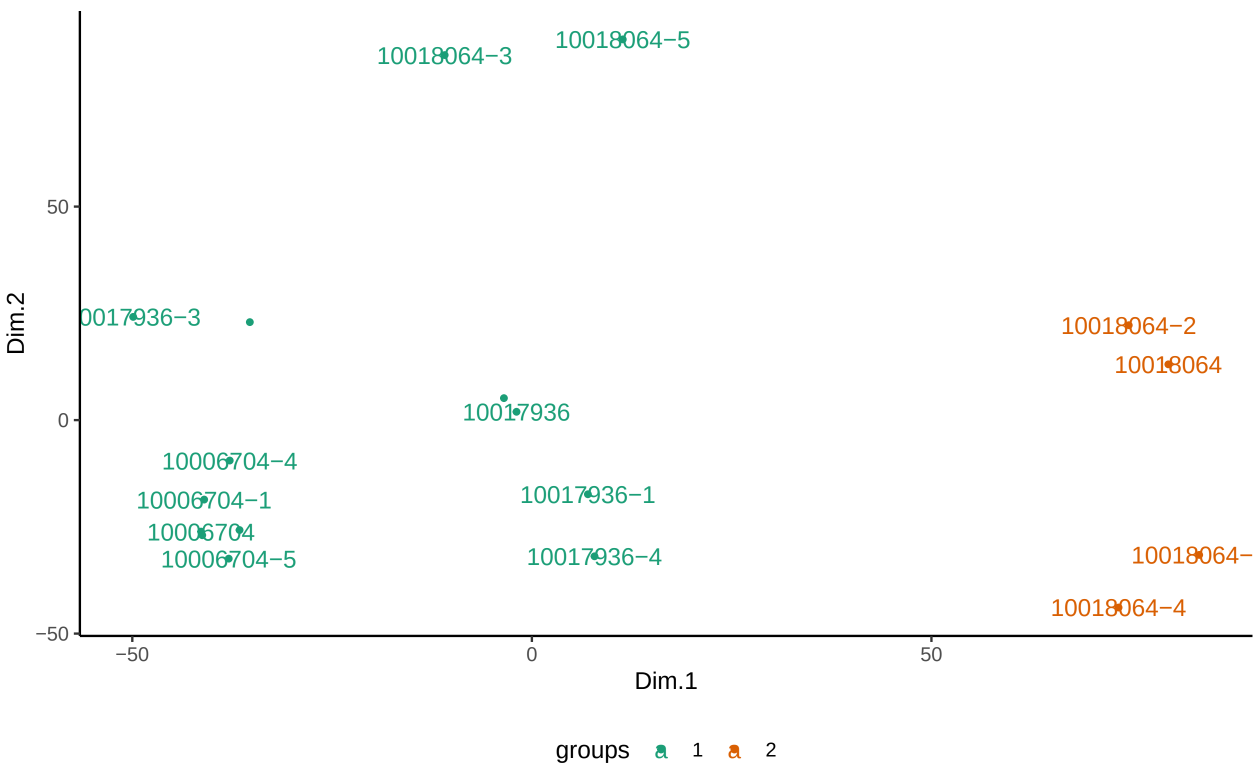


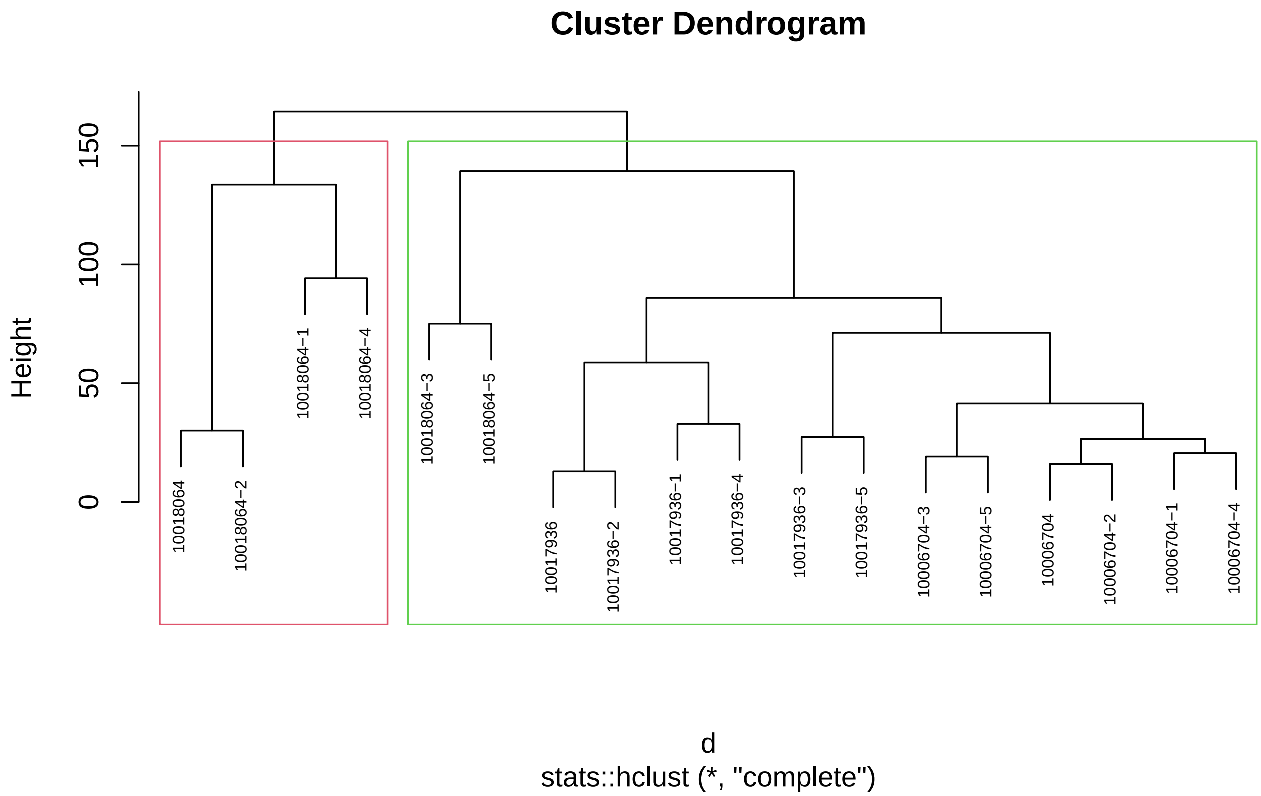
Figure 4: Five iterations for each customer are considered by adding random noise ($N(0, \sigma^2)$), where σ^2 is very small relative to the the variance of the data. The raw plots for all simulated dataset is shown to make sure they look similar to the structure.

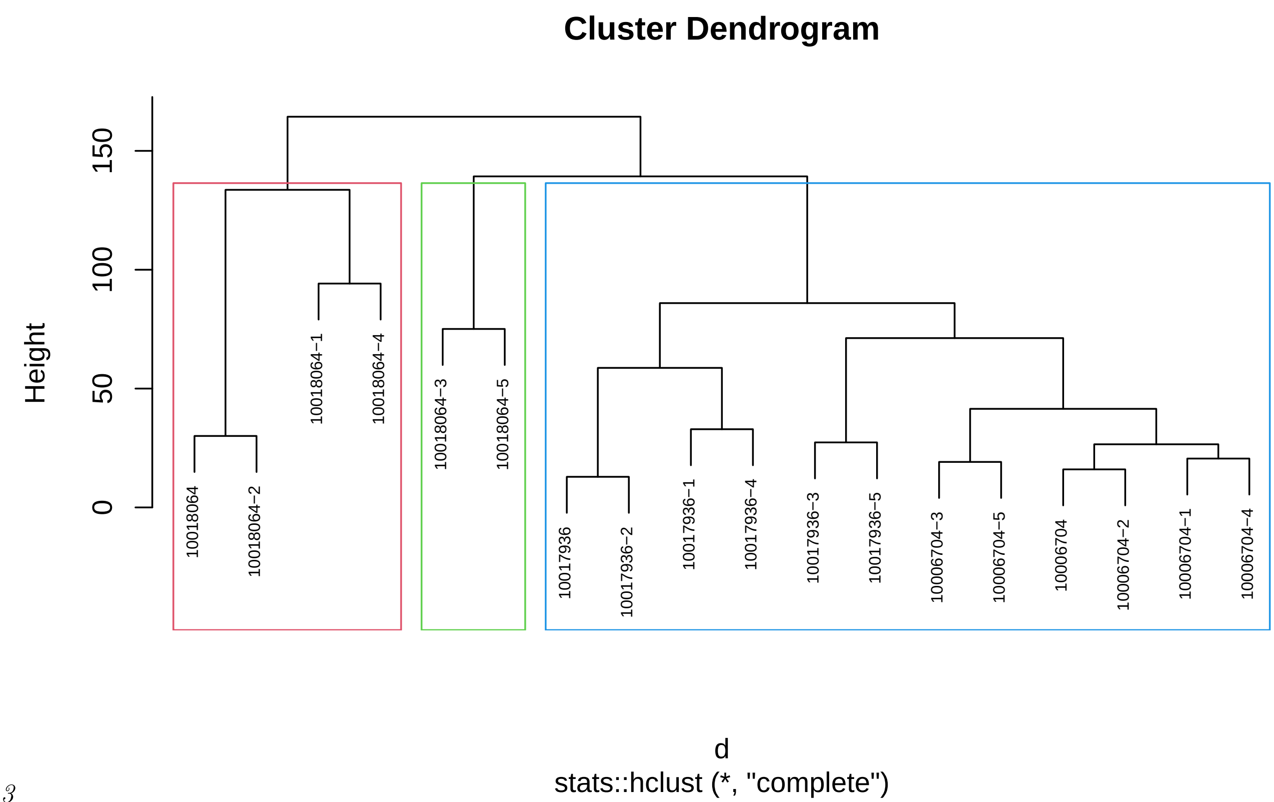


6 Multi-dimensional scaling



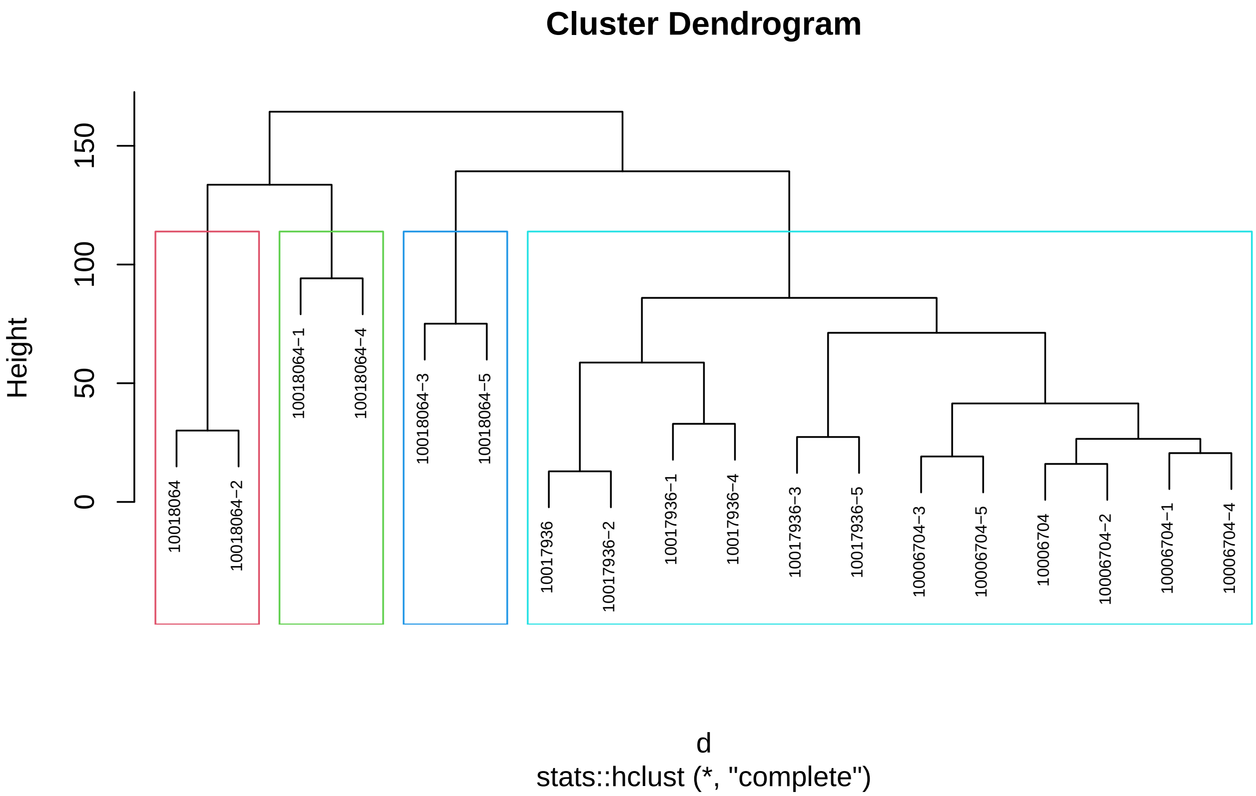
$k = 2$


Cluster Dendrogram
Height
0
50
100
150
10018064
10018064−2
10018064−1
10018064−4
10018064−3
10018064−5
10017936
10017936−2
10017936−1
10017936−4
10017936−3
10017936−5
10006704−3
10006704−5
10006704
10006704−2
10006704−1
10006704−4
d
stats::hclust (*, "complete")

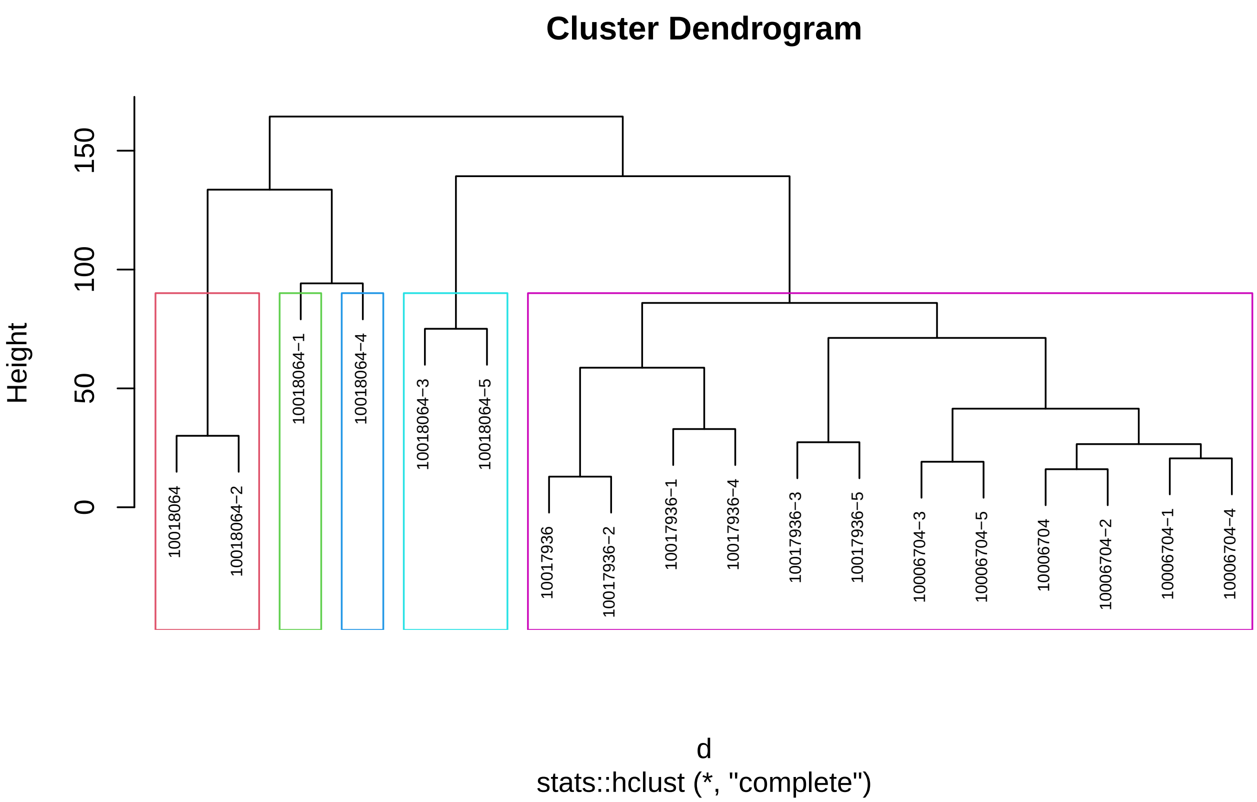

Cluster Dendrogram
Height
0
50
100
150
10018064
10018064−2
10018064−1
10018064−4
10018064−3
10018064−5
10017936
10017936−2
10017936−1
10017936−4
10017936−3
10017936−5
10006704−3
10006704−5
10006704
10006704−2
10006704−1
10006704−4
d
stats::hclust (*, "complete")

$k = 3$

$k = 4$


Cluster Dendrogram
Height
0
50
100
150
10018064
10018064−2
10018064−1
10018064−4
10018064−3
10018064−5
10017936
10017936−2
10017936−1
10017936−4
10017936−3
10017936−5
10006704−3
10006704−5
10006704
10006704−2
10006704−1
10006704−4
d
stats::hclust (*, "complete")

$k = 5$


Cluster Dendrogram
Height
0
50
100
150
10018064
10018064−2
10018064−1
10018064−4
10018064−3
10018064−5
10017936
10017936−2
10017936−1
10017936−4
10017936−3
10017936−5
10006704−3
10006704−5
10006704
10006704−2
10006704−1
10006704−4
d
stats::hclust (*, "complete")