Screening harmonies

Contents

1 Idea

Even after excluding clashes, the list of harmonies left could be large and overwhelming for human consumption. Hence, there is a need to rank the harmonies basis how well they capture the variation in the measured variable and additionally reduce the number of harmonies for further exploration/visualization. Gestalt theory suggests that when items are placed in close proximity, people assume that they are in the same group because they are close to one another and apart from other groups. Hence, displays that capture more variation within different categories in the same group would be important to bring out different patterns of the data. Thus the idea here is to rate a harmony pair higher if this variation between different levels of the x-axis variable is higher on an average across all levels of facet variables.

2 Computing distances

One of the potential ways to evaluate this variation is by computing the pairwise distances between the distributions of the measured variable. We do this through Jensen-Shannon divergence which is based on Kullback-Leibler divergence. Probability distributions are represented through sample quantiles instead of kernel density estimate so that there is minimal dependency on selecting kernel or bandwidth.

We shall call this measure of variation as Median Maximum Pairwise Distances (MMPD)

3 Normalize distances

The harmony pairs could be arranged from highest to lowest average maximum pairwise distances across different levels of the harmonies. But maximum is not robust to the number of levels and is higher for harmonies with higher levels. Thus these maximum pairwise distances need to be normalized for different harmonies in a way that eliminates the effect of different levels. The Fisher–Tippett–Gnedenko theorem in the field of Extreme Value Theory states that the maximum of a sample of iid random variables after proper re- normalization can converge in distribution to only one of Weibull, Gumbel or Freschet distribution, independent of the underlying data or process. The normalizing constants, however, vary depending on the underlying distribution and hence it is important to assume a distribution of distances in our case.

4 Choose thresholds for harmonies

facets/x-axis	A_1	A_2	A_3	..	..	A_K
B_1	p_11	p_12	p_13	..	..	p_1K
B_2						
B_3						
..						
..						
B_L	p_L1	p_L2	p_L3	..	..	p_LK

$H_{01} : p_{11} = p_{21} = \ldots = p_{L1}$
$H_{02} : p_{12} = p_{22} = \ldots = p_{L2}$
$\vdots$
$H_{0K} : p_{1K} = p_{2K} = \ldots = p_{LK}$

$m = \binom{L}{2}$ (unordered)
$m = L - 1$ (ordered)

facets/distances	A_1	A_2	A_3	..	..	A_K
d_1	d_11	d_12	d_13	..	..	d_1K
d_2						
d_3						
..						
..						
d_m	d_m1	d_m2	d_m3	..	..	d_mK

$H_{01} : d_{11} = d_{21} = \ldots = d_{m1} = 0$
$H_{02} : d_{12} = d_{22} = \ldots = d_{m2} = 0$
$\vdots$
$H_{0K} : d_{1K} = d_{2K} = \ldots = d_{mK} = 0$

- can do ANOVA at this stage
- interpretation of results (if interaction of levels significant when testing if means of distributions of distances are equal to zero)

facets/max-dist	A_1	A_2	..	..	A_K
max-dist	max(d_11, .., dm1)	max(d_12, . . . d_m2)	..	..	max(d_1K,. . . d_mK)

$H_{01} : max(d_{11}, .., d_{m1}) = 0$
$H_{02} : max(d_{12}, ...d_{m2}) = 0$
$\vdots H_{0K} : max(d_{1K}, ...d_{mK}) = 0$

- normalised maximum distribution follows standardised Gumbel distribution
- multiple hypothesis testing problem where p-values needs to be adjusted with Fisher's combination test (preferred) or Bonferroni's correction
- What is the test statistic for multiple hypothesis problem?

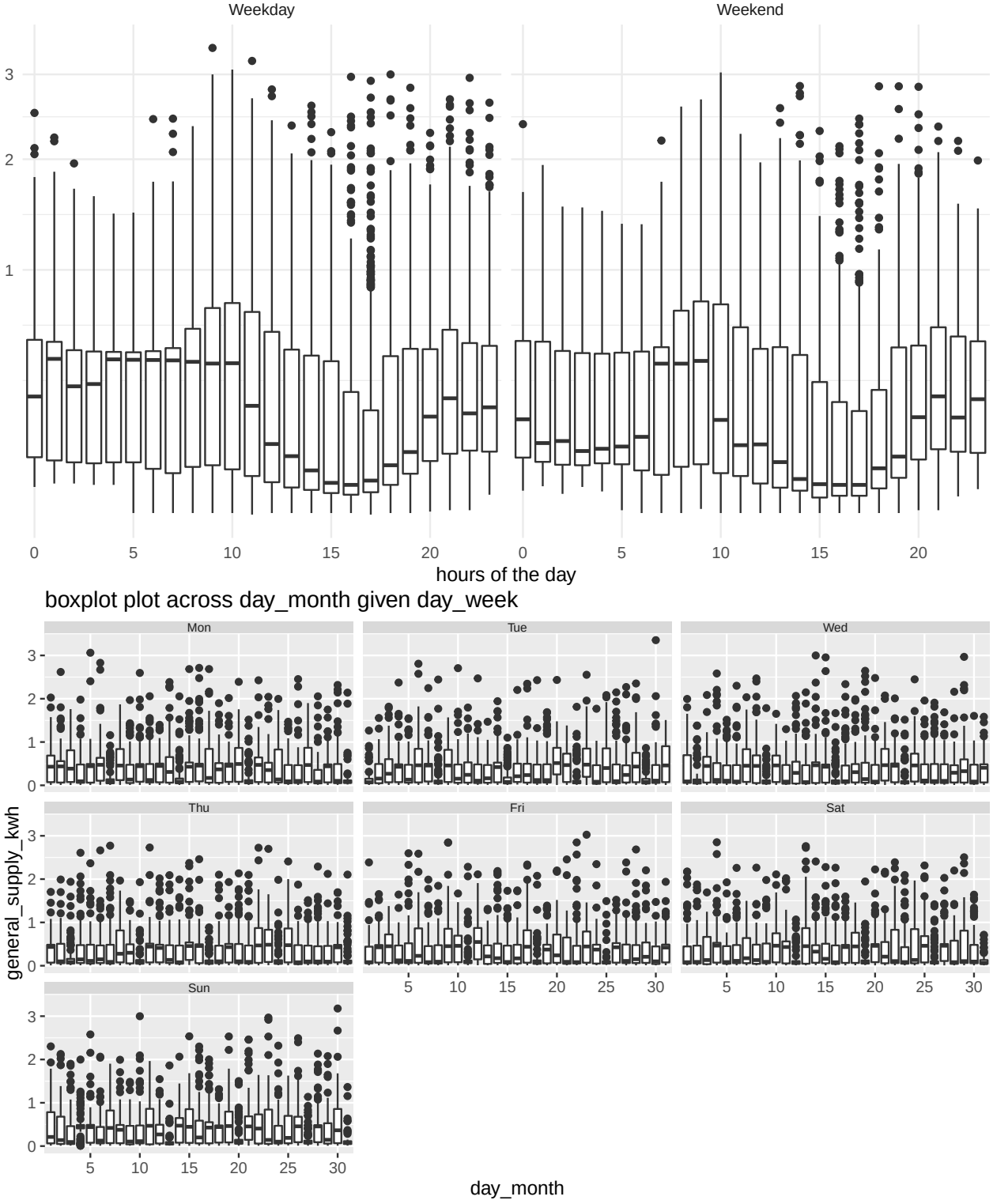
5 Results

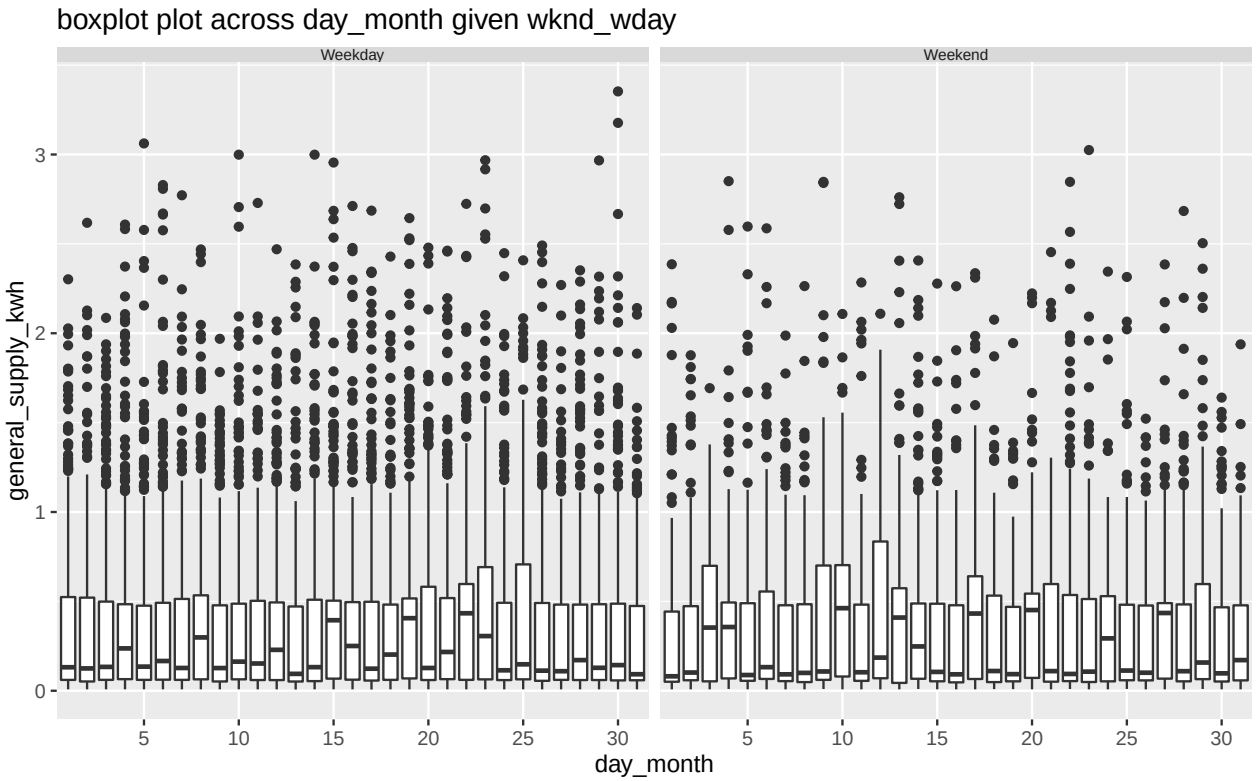
5.1 Smart meter data

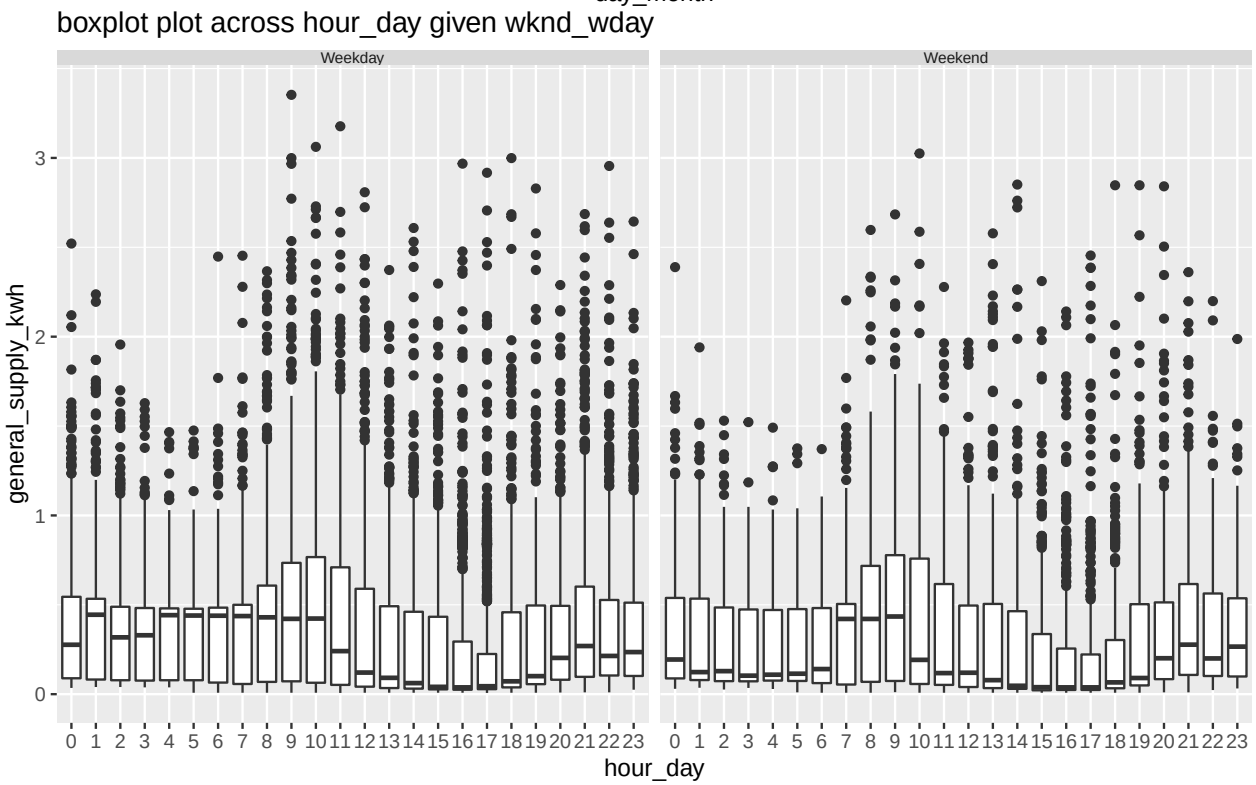
normal: standard normal ordered distances
normal_nonstd: non-standard normal ordered distances
normal_un: standard normal unordered distances
normal_nonstd_un: non-standard normal unordered distances

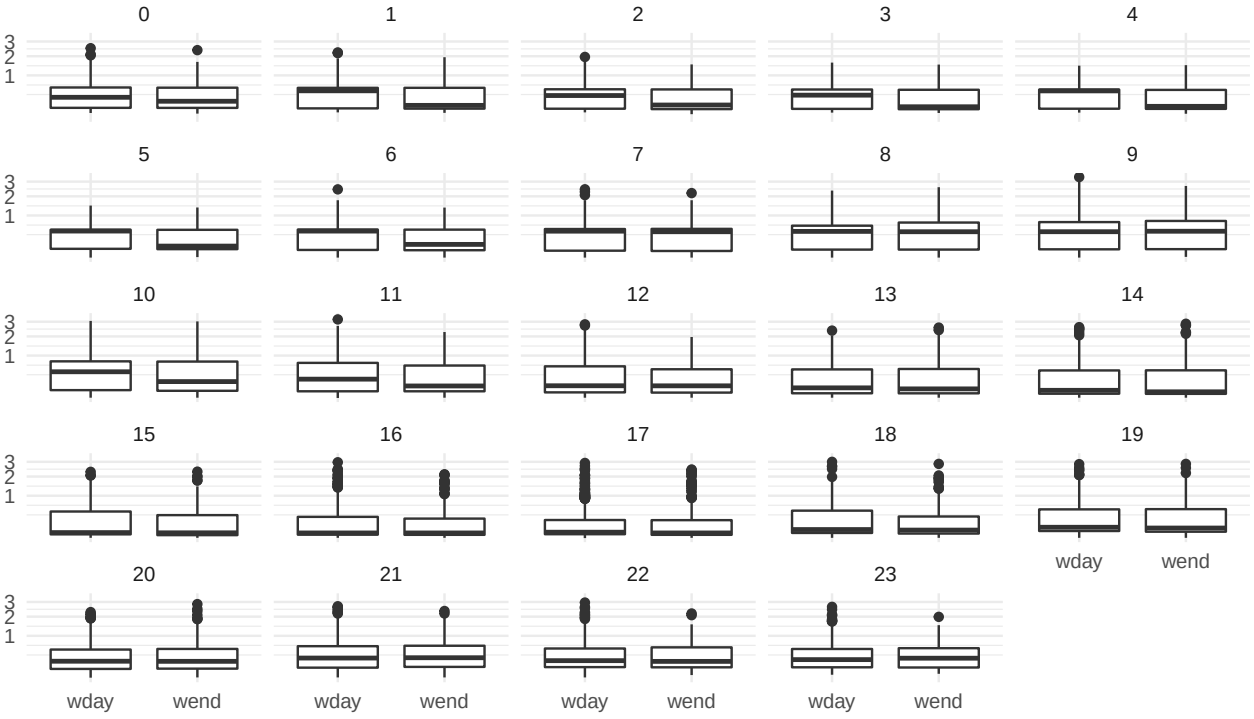
facet_variable	x_variable	facet_levels	x_levels	normal	normal_nonstd	normal_un	normal_nonstd_un
day_week	day_month	7	31	1	1	3	3
wknd_wday	day_month	2	31	2	2	2	5
wknd_wday	hour_day	2	24	3	3	1	1
day_week	hour_day	7	24	4	5	5	4
week_month	hour_day	5	24	5	6	4	2
hour_day	day_month	24	31	6	4	6	7
day_month	day_week	31	7	7	8	8	8
day_month	hour_day	31	24	8	7	7	6
day_week	week_month	7	5	9	11	11	12
wknd_wday	week_month	2	5	10	9	12	11
hour_day	week_month	24	5	11	12	14	13
day_month	wknd_wday	31	2	12	14	13	14
hour_day	wknd_wday	24	2	13	15	15	15
week_month	day_week	5	7	14	13	10	10
hour_day	day_week	24	7	15	10	9	9
week_month	wknd_wday	5	2	16	16	16	16

5.2 Graphical evidence

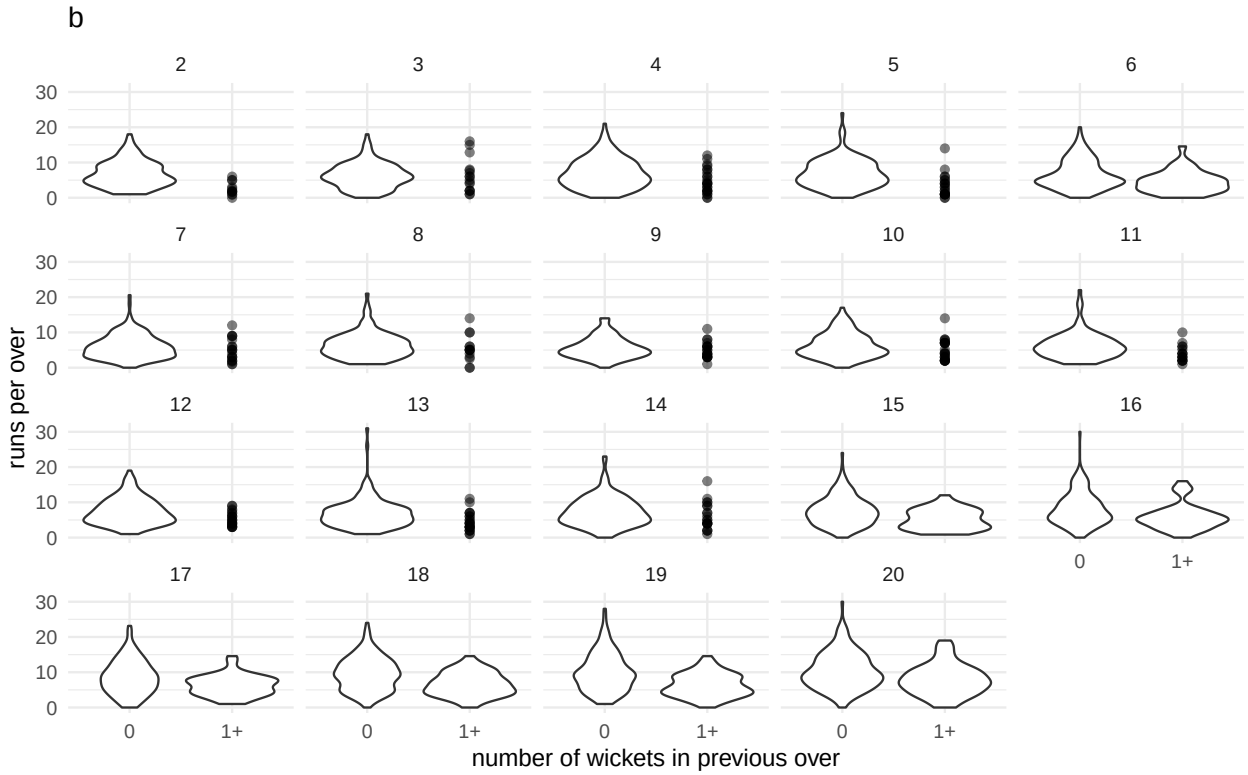



boxplot plot across day_month given wknd_wday
Weekday
Weekend
general_supply_kwh
0
1
2
3
5
10
15
20
25
30
day_month


boxplot plot across hour_day given wknd_wday
Weekday
Weekend
general_supply_kwh
0
1
2
3
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
hour_day



5.3 cricket data



6 Bibliography