1 Multiple Correspondence Analysis (5 pts)

In order to apply Multiple Correspondence Analysis (MCA), we transform all quantitative variables to qualitative ones. More precisely, for Monday consider the loading averaged over the time 7h - 20h. Denote by q_α the quantile of order α of this quantity over all stations. Then, we create a qualitative variable day1 with three levels: 'a' if the loading belongs to the interval $[0, q_{1/3})$, 'b' for the interval $[q_{1/3}, q_{2/3})$ and 'c' for the interval $[q_{2/3}, 1]$. Similarly we create variables day2, ..., day7, night1, ..., night7. Finally, we create dummy variables by creating one variable per level. Using this process for all days and nights of the week, we obtain 42 indicator variables called day1_a, day1_b, day1_c, ..., day7_a, day7_b, day7_c, night1_a, night1_b, night1_c, ..., night7_a, night7_b, night7_c. We further consider the dummy variables corresponding to the presence of stations on a hill of not. Table 1 presents an extraction of the corresponding table, denoted **D**.

	nohill	hill	day1_a	day1_b	day1_c	day2_a	day2_b	day2_c
EURYALE DEHAYNIN	1	0	1	0	0	1	0	0
LEMERCIER	1	0	1	0	0	0	1	0
MEZIERES RENNES	1	0	0	1	0	0	0	1
FARMAN	1	0	0	0	1	0	0	1
QUAI DE LA RAPEE	1	0	0	0	1	0	0	1
CHOISY POINT D'IVRY	1	0	1	0	0	0	1	0

Table 1: Extraction of the table **D** of dummy variables built from the velib data (first 6 rows, and first 8 columns).

1. (1 pt) How is called the table **D**? Consider a particular row. What can you say about the sum over the columns (considering all the 44 columns)? Deduce that the table of row profiles of **D** is proportional to **D**.
   It is the disjonctive table. The sum is equal to the number of qualitative variables, i.e. 15 here. The table of row profiles is equal to $\mathbf{D}/15$.

2. (1 pt) Consider the table $\mathbf{D}'$ extracted from **D** by removing the columns 'nohill' and 'hill'. What can you say about the sum over the rows, considering all the 1189 stations? Deduce that the table of column profiles of $\mathbf{D}'$ is nearly proportional to $\mathbf{D}'$.
   Because we have used the quantiles of order 1/3 and 2/3, the sum over the rows should be nearly equal to $1189/3 \approx 396$. Hence for these columns, the subtable of columns profiles will be nearly proportional to the original subtable (up to a factor 396).

3. (1 pt) Does the same result hold if $\mathbf{D}'$ is the table extracted from **D** by considering only the first two columns 'nohill' and 'hill' ? (Remember that there are 127 stations located on a hill among 1189).
   The column profile for Hill is obtained from Hill by dividing by 127, whereas we have to divide by $1189 - 127 = 1063$ for Nohill. Hence the result does not hold.

4. (2 pts) Figure 1 shows the result of MCA. Recall what PCA is done here (which matrix computed from **D**? which metric?). Interpret briefly the results: groups of levels? Relative position of the level 'hill'?

   The graph is obtained by doing a PCA on the column profiles of **D** with the χ^2 metric, viewed as individuals. We can see clusters of levels corresponding to loading value, days / nights in the week / week-end. Hill is close to the levels corresponding to small loadings.

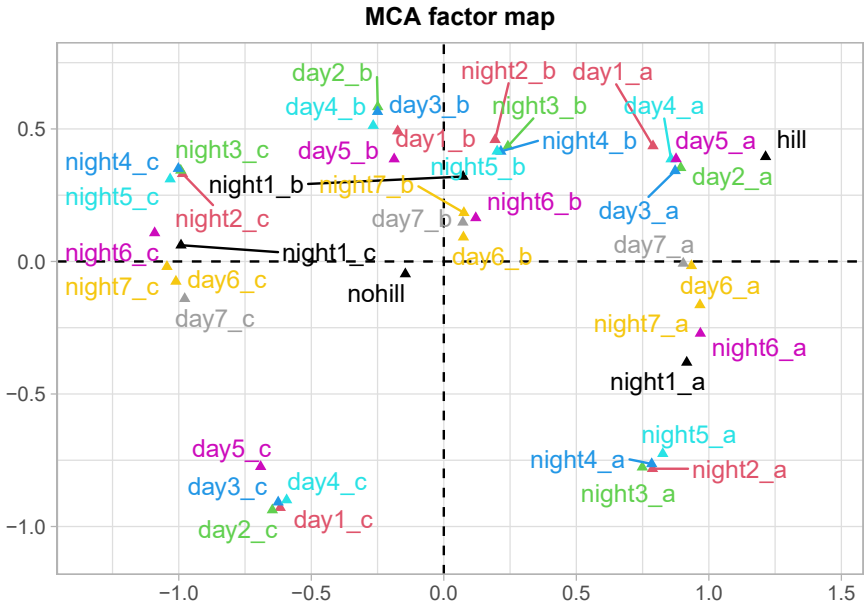

Figure 1: Graph of variable levels for MCA.