Prieskumná analýza

Denisa Mensatorisová a Adam Štuller

```
all_data <- read.csv(file= "../data/all.csv")
```

Vo vybraných dátach sa nachádza 15 atribútov, z toho X označuje poradové číslo merania, STATION je ID stanice Sliač a DATE predstavuje presný dátum a čas merania. Ostatné atribúty predstavujú hodnoty meraní pre:

- TMP - teplota
- WND_ANGLE - uhol vetra
- WND_SPEED - rýchlosť vetra
- VIS - viditeľnosť
- SLP - atmosférický tlak
- CIG - výška oblakov
- DEW - rosný bod
- SNOW_DEPTH - hĺbka snehu
- PWO - pozorované počasie
- GSO - zemský povrch
- LP - zrážky namerané každých 6 hodín
- LP24 - zrážky namerané každých 24 hodín

V dátach sa nachádzajú prevažne numerické spojité atribúty, kategorické sú iba PWO a GSO. Jednotlivé atribúty sú bližšie opísané v časti prieskumnej analýzy spolu s ich distribúciami a grafmi. Kompletný zoznam atribútov, ktoré sa v dátach nachádzajú je uvedený v priloženej dokumentácii k datasetu.

```
str(all_data)
```

```
## 'data.frame':    413336 obs. of  15 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ STATION   : num  1.19e+10 1.19e+10 1.19e+10 1.19e+10 1.19e+10 ...
##  $ DATE      : chr  "2004-05-10T00:00:00" "2004-05-10T01:00:00" "2004-05-10T02:00:00" "2004-05-10T03
##  $ TMP       : num  8 8 7 6 7 7 8 9 14 15 ...
##  $ WND_ANGLE : int  NA NA 330 NA NA NA NA NA 260 270 ...
##  $ WND_SPEED : num  1 0 1 0 1 1 2 1 3 4 ...
##  $ VIS       : int  11265 11265 11265 11265 1500 3500 5000 11265 11265 11265 ...
##  $ SLP       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CIG       : int  3000 3000 22000 22000 120 120 120 22000 22000 1290 ...
##  $ DEW       : num  8 8 7 6 7 7 7 7 8 6 ...
##  $ SNOW_DEPTH: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ PWO       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ GSO       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LP        : chr  NA NA NA NA ...
##  $ LP24      : int  NA NA NA NA NA NA NA NA NA NA ...
```

```
glimpse(all_data)
```

```
## Rows: 413,336
## Columns: 15
## $ X          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
```

```
## $ STATION    <dbl> 11927599999, 11927599999, 11927599999, 11927599999, 1192759~
## $ DATE       <chr> "2004-05-10T00:00:00", "2004-05-10T01:00:00", "2004-05-10T0~
## $ TMP        <dbl> 8, 8, 7, 6, 7, 7, 8, 9, 14, 15, 15, 16, 16, 15, 14, 13, 13,~
## $ WND_ANGLE  <int> NA, NA, 330, NA, NA, NA, NA, NA, 260, 270, 300, 250, 290, 2~
## $ WND_SPEED  <dbl> 1, 0, 1, 0, 1, 1, 2, 1, 3, 4, 3, 5, 6, 5, 5, 5, 1, 2, 2, 1,~
## $ VIS        <int> 11265, 11265, 11265, 11265, 1500, 3500, 5000, 11265, 11265,~
## $ SLP        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ CIG        <int> 3000, 3000, 22000, 22000, 120, 120, 120, 22000, 22000, 1290~
## $ DEW        <dbl> 8, 8, 7, 6, 7, 7, 7, 7, 8, 6, 4, 5, 4, 4, 5, 6, 6, 6, 6, 6,~
## $ SNOW_DEPTH <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ PWO        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ GSO        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ LP         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ LP24       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
status(all_data)
```

```
##              variable q_zeros      p_zeros   q_na         p_na q_inf p_inf
## X                   X       0 0.000000e+00      0 0.0000000000     0     0
## STATION       STATION       0 0.000000e+00      0 0.0000000000     0     0
## DATE             DATE       0 0.000000e+00      0 0.0000000000     0     0
## TMP               TMP    9413 2.277324e-02    341 0.0008249947     0     0
## WND_ANGLE   WND_ANGLE       0 0.000000e+00 205906 0.4981564635     0     0
## WND_SPEED   WND_SPEED   99825 2.415105e-01  18057 0.0436860085     0     0
## VIS               VIS     123 2.975787e-04  89684 0.2169760195     0     0
## SLP               SLP       0 0.000000e+00 224734 0.5437077825     0     0
## CIG               CIG       2 4.838678e-06 120770 0.2921835988     0     0
## DEW               DEW   13161 3.184092e-02    499 0.0012072503     0     0
## SNOW_DEPTH SNOW_DEPTH    2415 5.842704e-03 408068 0.9872549209     0     0
## PWO               PWO   15505 3.751185e-02 323213 0.7819618906     0     0
## GSO               GSO     968 2.341920e-03 397749 0.9622897594     0     0
## LP                 LP       0 0.000000e+00 346083 0.8372921788     0     0
## LP24             LP24    3764 9.106393e-03 406562 0.9836113961     0     0
##                 type unique
## X            integer 413336
## STATION      numeric      2
## DATE       character 315105
## TMP          numeric    628
## WND_ANGLE    integer     36
## WND_SPEED    numeric     58
## VIS          integer     98
## SLP          numeric    686
## CIG          integer    217
## DEW          numeric    486
## SNOW_DEPTH   integer     52
## PWO          integer     10
## GSO          integer     30
## LP         character    352
## LP24         integer    259
```

```
summary(all_data)
```

```
##        X             STATION              DATE                TMP
##  Min.   :     1   Min.   :1.190e+10   Length:413336      Min.   :-28.700
##  1st Qu.:103335   1st Qu.:1.190e+10   Class :character   1st Qu.:  2.000
```

```
##  Median :206669   Median :1.193e+10   Mode  :character   Median :  9.200
##  Mean   :206669   Mean   :1.192e+10                      Mean   :  9.482
##  3rd Qu.:310002   3rd Qu.:1.193e+10                      3rd Qu.: 16.500
##  Max.   :413336   Max.   :1.193e+10                      Max.   : 38.000
##                                                          NA's   :341
##    WND_ANGLE        WND_SPEED           VIS             SLP
##  Min.   : 10.0    Min.   : 0.000   Min.   :    0   Min.   : 970.6
##  1st Qu.: 70.0    1st Qu.: 0.000   1st Qu.: 6000   1st Qu.:1012.0
##  Median :210.0    Median : 1.000   Median :10000   Median :1016.9
##  Mean   :191.2    Mean   : 1.716   Mean   :12841   Mean   :1017.3
##  3rd Qu.:300.0    3rd Qu.: 2.600   3rd Qu.:15000   3rd Qu.:1022.2
##  Max.   :360.0    Max.   :26.000   Max.   :75000   Max.   :1049.4
##  NA's   :205906   NA's   :18057    NA's   :89684   NA's   :224734
##       CIG              DEW            SNOW_DEPTH          PWO
##  Min.   :    0    Min.   :-29.500   Min.   :  0.0    Min.   :0.0
##  1st Qu.:  750    1st Qu.: -1.000   1st Qu.:  0.0    1st Qu.:1.0
##  Median : 1707    Median :  5.000   Median :  1.0    Median :2.0
##  Mean   : 7360    Mean   :  4.574   Mean   :  6.6    Mean   :3.4
##  3rd Qu.:22000    3rd Qu.: 11.000   3rd Qu.: 10.0    3rd Qu.:6.0
##  Max.   :22000    Max.   : 23.000   Max.   :207.0    Max.   :9.0
##  NA's   :120770   NA's   :499       NA's   :408068   NA's   :323213
##       GSO              LP                 LP24
##  Min.   : 0.0     Length:413336      Min.   :   0.0
##  1st Qu.: 1.0     Class :character   1st Qu.:   0.0
##  Median : 9.0     Mode  :character   Median :   0.0
##  Mean   :11.7                        Mean   :  22.7
##  3rd Qu.:20.0                        3rd Qu.:  10.0
##  Max.   :31.0                        Max.   :9901.0
##  NA's   :397749                      NA's   :406562
```

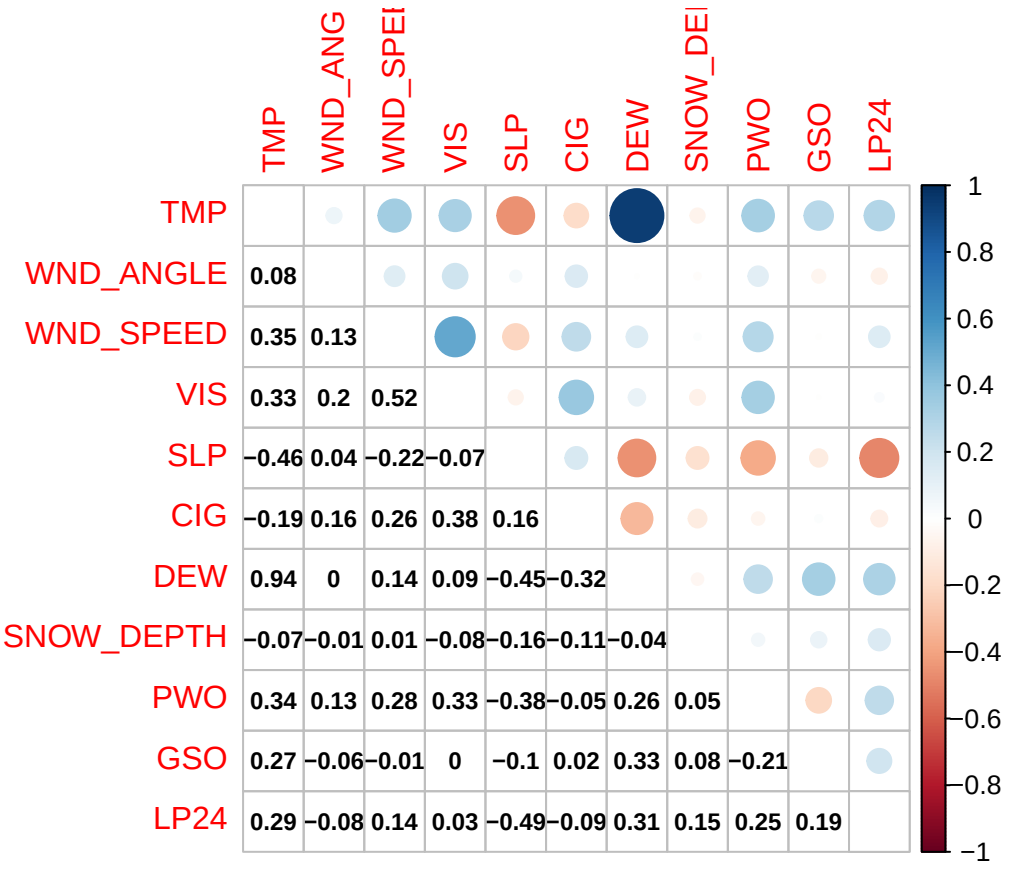
Korelačná matica zobrazuje hodnoty korelácie medzi vybranými atribútmi - TMP, WND_ANGLE, WND_SPEED, VIS, SLP, CIG, DEW, SNOW_DEPTH, PWO, GSO, LP24. Hodnota (výberového) korelačného koeficientu sa pohybuje v intervale od -1 do 1. Ak je toto číslo kladné, vzťah medzi danými dvoma atribútmi je priamy, teda s narastajúcimi hodnotami X narastajú aj hodnoty Y. Ak je číslo záporné, tak medzi X a Y je vzťah nepriamy, čiže s narastajúcimi (klesajúcimi) hodnotami X klesajú (rastú) hodnoty Y. Ak je číslo rovné 0 tak neexistuje lineárna závislosť medzi X a Y, môže však existovať nelineárna závislosť. A teda čím bližšia hodnota k |1|, tým silnejší je vzťah medzi X a Y.

Najvyššia korelácia je medzi TMP a DEW (0.94), ide o kladnú koreláciu teda, čím vyššia je teplota tým vyššia je hodnota rosného bodu. Ďalej je dosť silná kladná korelácia aj medzi VIS a WND_SPEED.

Naopak záporná korelácia je medzi TMP a SLP, SLP a LP24.

```
# korelacie vybranych numerickych atributov - TMP, WND_ANGLE, WND_SPEED, VIS, SLP, CIG, DEW, LP
data_cor <- select(all_data, TMP, WND_ANGLE, WND_SPEED, VIS, SLP, CIG, DEW, SNOW_DEPTH, PWO, GSO, LP24)
cor <- cor(data_cor, use = "na.or.complete")

corrplot.mixed(cor, lower="number", upper="circle", tl.pos = "lt", lower.col = "black",number.cex=0.75)
```




```
pairs(~TMP + SLP + DEW + VIS + CIG + SLP , data = all_data)
```