Лабораторна робота №1. Підготовка даних

2025-03-05

Перший огляд датасету

Одразу бачимо, що є числові колонки, які зчиталися, як текстові. `Warning` пропонує застосувати функцію `problems`, щоб дізнатися, що пішло не так під час спроби розпарсити значення колонок csv файлу.

```
dt <- read_csv("../data/raw/air_quality.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
glimpse(dt)
```

```
## Rows: 5,882,208
## Columns: 25
## $ date       <dttm> 2024-08-31 23:00:00, 2024-08-31 23:00:00, 2024-08-31 23:00:~
## $ sitename   <chr> "Hukou", "Zhongming", "Zhudong", "Hsinchu", "Toufen", "Miaol~
## $ county     <chr> "Hsinchu County", "Taichung City", "Hsinchu County", "Hsinch~
## $ aqi        <dbl> 62, 50, 45, 42, 50, 40, 39, 44, 46, 49, 44, 58, 45, 62, 58, ~
## $ pollutant  <chr> "PM2.5", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "PM2.5", NA~
## $ status     <chr> "Moderate", "Good", "Good", "Good", "Good", "Good", "Good", ~
## $ so2        <dbl> 0.9, 1.6, 0.4, 0.8, 1.0, 1.1, 0.9, 1.3, 2.5, 0.7, 1.9, 0.1, ~
## $ co         <chr> "0.17", "0.32", "0.17", "0.2", "0.16", "0.17", "0.18", "0.24~
## $ o3         <chr> "35.0", "27.9", "25.1", "30.0", "33.5", "35.2", "35.3", "24.~
## $ o3_8hr     <chr> "40.2", "35.1", "40.6", "35.9", "35.9", "35.0", "42.9", "39.~
## $ pm10       <chr> "18.0", "27.0", "21.0", "19.0", "18.0", "15.0", "14.0", "21.~
## $ pm2.5      <chr> "17.0", "14.0", "13.0", "10.0", "14.0", "12.0", "9.0", "12.0~
## $ no2        <dbl> 2.3, 7.6, 2.9, 4.0, 1.8, 4.0, 2.4, 6.8, 7.3, 5.6, 7.3, 9.8, ~
## $ nox        <dbl> 2.6, 9.3, 4.1, 4.8, 3.1, 5.1, 3.1, 7.3, 7.7, 6.3, 7.7, 10.5,~
## $ no         <dbl> 0.3, 1.6, 1.1, 0.7, 1.2, 1.1, 0.7, 0.5, 0.3, 0.7, 0.4, 0.7, ~
## $ windspeed  <chr> "2.3", "1.1", "0.4", "1.9", "1.8", "1.4", "0.6", "1.2", "0.5~
## $ winddirec  <chr> "225", "184", "210", "239", "259", "235", "203", "38", "8", ~
## $ unit       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ co_8hr     <chr> "0.2", "0.2", "0.2", "0.2", "0.1", "0.1", "0.1", "0.2", "0.1~
## $ pm2.5_avg  <dbl> 20.1, 15.3, 13.8, 13.0, 15.3, 12.2, 11.4, 13.5, 14.2, 15.1, ~
## $ pm10_avg   <chr> "26.0", "23.0", "24.0", "26.0", "28.0", "17.0", "16.0", "21.~
## $ so2_avg    <dbl> 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, ~
## $ longitude  <dbl> 121.0389, 120.6411, 121.0890, 120.9724, 120.8987, 120.8201, ~
## $ latitude   <dbl> 24.90010, 24.15196, 24.74091, 24.80564, 24.69691, 24.56499, ~
## $ siteid     <dbl> 22, 31, 23, 24, 25, 26, 27, 28, 29, 30, 32, 20, 34, 35, 36, ~
```

Parsing issues

Застосування функції `problems` показало, що:

- іноді замість порожнього значення використовується “-” або “ND”. Через це відповідні колонки стають текстовими
- трапляється неправильний формат дати (роздільник `/` замість очікуваного `-`).

```
problems(dt)
```

```
## # A tibble: 223 x 5
##       row   col expected        actual              file
##     <int> <int> <chr>           <chr>               <chr>
##  1 274922     7 a double        -                   /Users/artem/KPI/Data_Analy~
##  2 274922    13 a double        -                   /Users/artem/KPI/Data_Analy~
##  3 274922    14 a double        -                   /Users/artem/KPI/Data_Analy~
##  4 274922    15 a double        -                   /Users/artem/KPI/Data_Analy~
##  5 274922    20 a double        -                   /Users/artem/KPI/Data_Analy~
##  6 274922    22 a double        -                   /Users/artem/KPI/Data_Analy~
##  7 496094     1 date in ISO8601 2023/11/13 10:00:00 /Users/artem/KPI/Data_Analy~
##  8 496095     1 date in ISO8601 2023/11/13 10:00:00 /Users/artem/KPI/Data_Analy~
##  9 496096     1 date in ISO8601 2023/11/13 10:00:00 /Users/artem/KPI/Data_Analy~
## 10 496097     1 date in ISO8601 2023/11/13 10:00:00 /Users/artem/KPI/Data_Analy~
## # i 213 more rows
```

До функції `read_csv` додано аргумент `na`, який вказує, що значення “ “,”-” і “ND” треба сприймати як порожні. Виправлено формат дати. Колонки `sitename`, `county`, `pollutant` і `status` перетворено на категорійні. Проблеми із типами даних на цьому вирішено.

```
dt <- read_csv(
  "../data/raw/air_quality.csv",
  na = c("", "-", "ND"),
  col_types = c(date = "character")
)

dt$date <- anytime(dt$date)

dt <- dt %>% mutate(
  sitename = as.factor(sitename),
  county = as.factor(county),
  pollutant = as.factor(pollutant),
  status = factor(status, levels=c("Good", "Moderate", "Unhealthy for Sensitive Groups", "Un
)

glimpse(dt)
```

```
## Rows: 5,882,208
## Columns: 25
## $ date      <dttm> 2024-08-31 22:00:00, 2024-08-31 22:00:00, 2024-08-31 22:00:~
## $ sitename  <fct> Hukou, Zhongming, Zhudong, Hsinchu, Toufen, Miaoli, Sanyi, F~
## $ county    <fct> Hsinchu County, Taichung City, Hsinchu County, Hsinchu City,~
## $ aqi       <dbl> 62, 50, 45, 42, 50, 40, 39, 44, 46, 49, 44, 58, 45, 62, 58, ~
## $ pollutant <fct> PM2.5, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, PM2.5, NA, PM~
## $ status    <fct> Moderate, Good, Good, Good, Good, Good, Good, Good, Good, Go~
## $ so2       <dbl> 0.9, 1.6, 0.4, 0.8, 1.0, 1.1, 0.9, 1.3, 2.5, 0.7, 1.9, 0.1, ~
## $ co        <dbl> 0.17, 0.32, 0.17, 0.20, 0.16, 0.17, 0.18, 0.24, 0.20, 0.24, ~
```

```
## $ o3        <dbl> 35.0, 27.9, 25.1, 30.0, 33.5, 35.2, 35.3, 24.6, 30.3, 29.4, ~
## $ o3_8hr    <dbl> 40.2, 35.1, 40.6, 35.9, 35.9, 35.0, 42.9, 39.7, 40.4, 37.0, ~
## $ pm10      <dbl> 18, 27, 21, 19, 18, 15, 14, 21, 33, 20, 32, 29, 34, 37, 17, ~
## $ pm2.5     <dbl> 17, 14, 13, 10, 14, 12, 9, 12, 16, 12, 17, 18, 12, 21, 13, 1~
## $ no2       <dbl> 2.3, 7.6, 2.9, 4.0, 1.8, 4.0, 2.4, 6.8, 7.3, 5.6, 7.3, 9.8, ~
## $ nox       <dbl> 2.6, 9.3, 4.1, 4.8, 3.1, 5.1, 3.1, 7.3, 7.7, 6.3, 7.7, 10.5,~
## $ no        <dbl> 0.3, 1.6, 1.1, 0.7, 1.2, 1.1, 0.7, 0.5, 0.3, 0.7, 0.4, 0.7, ~
## $ windspeed <dbl> 2.3, 1.1, 0.4, 1.9, 1.8, 1.4, 0.6, 1.2, 0.5, 0.9, 0.9, 1.2, ~
## $ winddirec <dbl> 225, 184, 210, 239, 259, 235, 203, 38, 8, 97, 244, 252, 19, ~
## $ unit      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ co_8hr    <dbl> 0.2, 0.2, 0.2, 0.2, 0.1, 0.1, 0.1, 0.2, 0.1, 0.2, 0.1, 0.3, ~
## $ pm2.5_avg <dbl> 20.1, 15.3, 13.8, 13.0, 15.3, 12.2, 11.4, 13.5, 14.2, 15.1, ~
## $ pm10_avg  <dbl> 26, 23, 24, 26, 28, 17, 16, 21, 26, 21, 26, 32, 25, 32, 29, ~
## $ so2_avg   <dbl> 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, ~
## $ longitude <dbl> 121.0389, 120.6411, 121.0890, 120.9724, 120.8987, 120.8201, ~
## $ latitude  <dbl> 24.90010, 24.15196, 24.74091, 24.80564, 24.69691, 24.56499, ~
## $ siteid    <dbl> 22, 31, 23, 24, 25, 26, 27, 28, 29, 30, 32, 20, 34, 35, 36, ~
```

Виявлення закодованих і неадекватних значень

Дослідження колонки “sitename”

Проблем не виявлено

```
table(dt$sitename)
```

```
##
##                            Annan                          Banqiao
##                            71911                            71915
##                          Cailiao                         Changhua
##                            71913                            71891
##              Changhua (Dacheng)              Changhua (Tianwei)
##                            19357                               41
##              Changhua (Yuanlin)                         Chaozhou
##                            25771                            71236
##                           Chiayi              Chiayi (Shuishang)
##                            71912                             166
##                         Chonglun                          Dacheng
##                                4                            32538
## Dajia (Rinan Elementary School)                             Dali
##                             3336                            71627
##                           Daliao                          Datong
##                            71899                            71934
##                           Dayuan                 Dayuan (Zhubei)
##                            71942                               11
##                         Dongshan                          Douliu
##                            71916                            71884
##  [ reached getOption("max.print") -- omitted 103 entries ]
```

Дослідження колонки “county”

Проблем не виявлено

```
table(dt$county)
```

```
##
##   Changhua County       Chiayi City     Chiayi County      Hsinchu City
##            293423             71912            143986             75850
##    Hsinchu County    Hualien County    Kaohsiung City      Keelung City
##            143846             71912            888497             71913
##     Kinmen County Lienchiang County     Miaoli County     Nantou County
##             71913             71914            219683            216420
##   New Taipei City     Penghu County   Pingtung County     Taichung City
##            898819             71910            305495            367033
##       Tainan City       Taipei City    Taitung County      Taoyuan City
##            366827            503766            143809            448718
##      Yilan County     Yunlin County
##            147071            287491
```

Дослідження колонки “aqi”

Очевидно кодові значення: -1

```r
table(dt$aqi)
```

```
##
##     -1      0      1      2      3      4      5      6      7      8      9
##   7391     36    122     77    111     58    245    577   1258   2518   3732
##     10     11     12     13     14     15     16     17     18
##   7125   9678  15063  20175  23382  32298  35373  43842  49767
##  [ reached getOption("max.print") -- omitted 290 entries ]
```

Рядки, де `aqi == -1`, майже повністю заповнені NA.

```r
dt %>% filter(aqi == -1) %>% select(!c(0:3))
```

```
## # A tibble: 7,391 x 22
##      aqi pollutant status   so2    co    o3 o3_8hr  pm10 pm2.5   no2   nox    no
##    <dbl> <fct>    <fct>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1    -1 <NA>     <NA>     0.9  0.17    NA     NA    14    10   8.2   9.9   1.6
##  2    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
##  3    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
##  4    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
##  5    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
##  6    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
##  7    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
##  8    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
##  9    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
## 10    -1 <NA>     <NA>      NA    NA    NA     NA    NA    NA    NA    NA    NA
## # i 7,381 more rows
## # i 10 more variables: windspeed <dbl>, winddirec <dbl>, unit <lgl>,
## #   co_8hr <dbl>, pm2.5_avg <dbl>, pm10_avg <dbl>, so2_avg <dbl>,
## #   longitude <dbl>, latitude <dbl>, siteid <dbl>
```

Який відсоток, займаються такі рядки? Менше 0.2%.

```r
mean(dt$aqi == -1, na.rm = TRUE)
```

```
## [1] 0.001265758
```

Дослідження колонки “pollutant”

Проблем не виявлено

```r
table(dt$pollutant)
```

```
##
##   Carbon Monoxide (CO) Nitrogen Dioxide (NO2)                  Ozone
##                      2                  23951                    100
##            Ozone (8hr)                   PM10                  PM2.5
##                 250979                  75230                2296437
##   Sulfur Dioxide (SO2)
##                    497
```

Дослідження колонки “status”

Проблем не виявлено

```r
table(dt$status)
```

```
##
##                           Good                       Moderate
##                        3185191                        2159158
## Unhealthy for Sensitive Groups                      Unhealthy
##                         343909                          51008
##                 Very Unhealthy                      Hazardous
##                            173                             51
```

Дослідження колонки “so2”

Очевидно кодові значення: -999. Є підозрілі від’ємні числа. Можна припустити, що від’ємні показники є наслідком неідеальності калібрування датчиків (тобто вони є справжніми, а не кодовими). Також у датасеті зафіксовані значно більші за модулем додатні концентрації, тому зсув показників є незначним.

```r
table(dt$so2)
```

```
##
##   -999   -4.1   -3.4   -0.8  -0.75   -0.7   -0.6   -0.5   -0.4   -0.3   -0.2
##      3      1      1      1      1      1      2    774    971   1422   1868
##  -0.13   -0.1      0    0.1    0.2    0.3    0.4   0.41    0.5    0.6    0.7
##      1   2982  66010  48252  64554  82827 103274      1 221368 145602 162218
##   0.78    0.8    0.9      1   1.06   1.07    1.1   1.11   1.12   1.13   1.14
##      1 176570 188558 218141      1      1 228865      1      1      1      1
##    1.2   1.21   1.26   1.27    1.3  1.313  1.318   1.35   1.39    1.4    1.5
## 233568      1      2      1 235985      1      1      1      1 234208 229847
##   1.55   1.57   1.58   1.59    1.6   1.61   1.63   1.65   1.67    1.7   1.73
##      1      1      1      2 221475      1      1      1      2 212248      1
##   1.76   1.77   1.78    1.8   1.81   1.82   1.84   1.85   1.86   1.87   1.88
##      1      1      1 201877      2      3      1      1      4      1      1
##    1.9   1.91   1.92   1.96   1.98   1.99      2   2.03   2.05   2.06   2.07
## 190522      2      1      1      2      2 177818      1      2      2      1
##   2.09    2.1   2.11   2.16   2.17   2.18   2.19    2.2   2.21   2.22   2.24
##      2 165383      3      1      1      1      1 154496      1      1      2
##   2.25   2.26   2.28   2.29    2.3   2.34   2.35   2.37   2.38    2.4   2.41
##      2      1      1      2 143099      1      2      2      1 130414      3
##   2.45   2.47   2.48   2.49    2.5   2.55   2.56   2.57    2.6   2.62   2.63
##      1      1      2      1 120750      2      1      1 109778      3      1
##   2.64   2.65   2.67   2.69    2.7   2.71   2.73   2.74   2.75   2.76   2.78
##      1      1      1      2 100664      2      1      2      1      2      1
##   2.79    2.8   2.81   2.83   2.85   2.86   2.87   2.89    2.9   2.92   2.96
##      1  92121      1      2      2      1      2      2  83601      2      1
##   2.99      3   3.02   3.05   3.07    3.1   3.14   3.15   3.16   3.18    3.2
##      1  77088      1      1      3  69583      1      1      1      1  63879
##   3.21   3.23   3.24   3.25   3.29    3.3    3.4   3.41   3.42   3.44    3.5
##      1      1      2      2      2  58659  52875      1      1      1  49108
##   3.51   3.52    3.6   3.62   3.69    3.7   3.74   3.78    3.8    3.9   3.91
##      1      1  44475      1      1  40834      1      2  37507  33760      1
##   3.92      4    4.1   4.12    4.2    4.3    4.4   4.43    4.5    4.6    4.7
##      2  31521  28417      1  26729  24615  22475      1  20480  19036  17590
##   4.73   4.79    4.8   4.82    4.9      5    5.1   5.13    5.2    5.3   5.34
##      1      1  16373      1  15238  13639  12972      1  11969  11047      1
##   5.35    5.4    5.5    5.6    5.7    5.8   5.83    5.9      6    6.1    6.2
##      1  10176   9577   8923   8315   7623      1   7267   6792   6358   5973
```

```
##    6.3    6.4    6.5    6.6    6.7    6.8    6.9      7    7.1    7.2    7.3
##   5498   5209   4900   4533   4199   4128   3992   3536   3546   3244   3095
##    7.4    7.5    7.6    7.7    7.8    7.9      8   8.02    8.1    8.2    8.3
##   2954   2643   2631   2426   2377   2273   2199      1   2053   1982   1864
##    8.4    8.5    8.6    8.7    8.8    8.9      9    9.1    9.2    9.3    9.4
##   1770   1641   1590   1562   1419   1465   1227   1274   1254   1204   1153
##    9.5    9.6    9.7    9.8    9.9     10   10.1   10.2   10.3   10.4   10.5
##   1043   1094   1012    968    900   3262    211    177    199    152    193
##   10.6   10.7   10.8   10.9     11   11.1   11.2   11.3   11.4   11.5   11.6
##    185    161    136    132   4949    141    125    114     93    119    112
##   11.7   11.8   11.9     12   12.1   12.2   12.3   12.4   12.5   12.6   12.7
##    105     88    129   3390     95     75     73     73     73     80     70
##   12.8   12.9     13   13.1   13.2   13.3   13.4   13.5   13.6   13.7   13.8
##     79     66   2480     57     64     56     57     46     48     54     62
##   13.9     14   14.1   14.2   14.3   14.4   14.5   14.6   14.7   14.8   14.9
##     60   1863     46     51     36     38     49     53     35     42     43
##     15   15.1   15.2   15.3   15.4   15.5   15.6   15.7   15.8   15.9     16
##   1450     30     37     27     34     34     36     33     46     31   1110
##   16.1   16.2   16.3   16.4   16.5   16.6   16.7   16.8   16.9     17   17.1
##     26     16     38     42     18     31     21     18     18    887     23
##   17.2   17.3   17.4   17.5   17.6   17.7   17.8   17.9     18   18.1   18.2
##     22     16     16     27     19     24     22     25    717     25     19
##   18.3   18.4   18.5   18.6   18.7   18.8   18.9     19   19.1   19.2   19.3
##     13     12     16     18     15     19     14    553     14     13      9
##   19.4   19.5   19.6   19.7   19.8   19.9     20   20.1   20.2   20.3   20.4
##     12     14     15     10      7      8    431      9     14     10      8
##   20.5   20.6   20.7   20.8   20.9     21   21.1   21.2   21.3   21.4   21.5
##      8      8      8      8      8    353     14     14     14      6     11
##   21.6   21.7   21.8   21.9     22   22.1   22.2   22.3   22.4   22.5   22.6
##     10     10      9      7    287      5     11      8      6      4      4
##   22.7   22.8   22.9     23   23.1   23.2   23.3   23.4   23.5   23.6   23.7
##     11     10      8    257      5      7      4      6     10      4      6
##   23.8   23.9     24   24.1   24.2   24.3   24.4   24.5   24.6   24.7   24.8
##      8      7    197     11     11      7      5      2      2      8      4
##   24.9     25   25.1   25.2   25.3   25.4   25.5   25.6   25.7   25.8   25.9
##      8    185      4      3      6      4      3      4      9      5      4
##     26   26.1   26.2   26.3   26.4   26.5   26.6   26.7   26.8   26.9     27
##    138      4      1      5      4      5      3      3      3      2    116
##   27.1   27.2   27.3   27.4   27.5   27.6   27.7   27.8   27.9     28   28.1
##      5      6      7      2      6      2      3      1      2    105      1
##   28.2   28.3   28.4   28.6   28.7   28.8   28.9     29   29.1   29.3   29.4
##      2      1      1      4      2      1      3     74      1      4      1
##   29.5   29.6   29.7   29.9     30   30.1   30.2   30.3   30.4   30.5   30.6
##      3      1      2      3     73      9      2      1      3      2      1
##   30.7   30.8   30.9     31   31.1   31.2   31.3   31.4   31.5   31.6   31.8
##      1      1      4     55      2      1      1      2      5      1      4
##   31.9     32   32.1   32.2   32.3   32.4   32.5   32.6   32.7   32.8     33
##      2     56      3      2      2      3      2      3      2      1     52
##   33.1   33.5   33.6   33.9     34   34.1   34.2   34.3   34.4   34.5   34.6
##      2      5      4      1     46      2      3      1      1      1      2
##   34.7   34.8   34.9     35   35.1   35.4   35.6   35.7   35.8   35.9     36
##      2      1      1     35      1      1      4      1      2      1     34
##   36.1   36.3   36.5   36.6   36.7   36.8   36.9     37   37.2   37.3   37.5
##      3      2      1      2      2      3      1     27      1      3      2
```

```
##   37.6   37.8     38   38.1   38.3   38.6   38.9     39   39.1   39.4   39.6
##      3      1     14      1      1      2      2     21      2      1      2
##     40   40.3   40.6   40.8   40.9     41   41.2   41.3   41.4   41.5   41.6
##     15      1      1      1      3     11      1      1      1      2      2
##   41.8   41.9     42   42.1   42.3   42.4   42.5   42.6   42.7   42.8     43
##      1      3      9      2      4      2      1      1      1      2     14
##   43.1   43.3   43.5   43.8     44   44.1   44.3   44.5   44.7   44.9     45
##      1      1      1      1     14      1      2      1      1      1     13
##   45.1   45.6   45.8     46   46.4   46.6   46.7     47   47.1   47.2   47.5
##      1      1      1     12      2      1      1      6      2      1      3
##   47.8   47.9     48   48.8     49   49.4   49.7     50   50.7   50.8   50.9
##      2      1      9      1      5      2      2      6      3      1      1
##     51   51.3     52   52.1   52.7   52.8   52.9     53   53.1   53.4   53.5
##      4      1      7      1      1      1      2      6      2      1      1
##   53.7   53.9     54   54.7   54.8   54.9     55   55.2   55.3   55.4     56
##      1      2      8      1      1      2      3      4      2      2      5
##     57   57.7   57.8     58   58.1   58.6     59   59.1   59.7   59.8     60
##      4      2      2      2      1      1      4      3      1      1      5
##   60.5   60.8     61   61.3   61.5   61.6   61.7   61.8     62   62.1   62.7
##      1      1      4      1      2      2      1      1      2      2      1
##   62.9     63   63.2   63.6     64   64.2   64.3   65.4   65.8     66   66.1
##      1      4      2      1      1      1      1      1      2      1      3
##   66.3     67   67.7     68     69   69.2   69.4   69.8   69.9     70   70.1
##      2      2      2      1      4      2      1      2      1      2      1
##   70.2   70.7     72   72.3     73   73.4     74   74.8     75     76   76.2
##      1      2      1      1      2      1      3      2      3      2      2
##     77     78   79.5   80.4     81   81.3   81.7     82     83   83.3     84
##      1      2      1      1      2      1      1      5      3      1      1
##   84.6     85     86   87.8     88     91   91.8     93   93.7     94     96
##      1      5      3      2      2      2      2      1      2      1      1
##     97   97.2   98.1   98.2   99.4    100  100.5    101  101.3  102.2  102.7
##      1      1      1      1      2      3      2      1      2      1      1
##    103    109    111  113.3  115.6  115.9    116  117.8  118.7    119    124
##      1      1      2      1      2      1      1      1      1      1      1
##  129.3    131    132  137.6    142  157.5    180  255.4
##      2      2      1      1      1      1      1      1
```

Дослідження колонки “co”

Очевидно кодові значення: -999. Є від’ємні числа.

```
table(filter(dt, co < 0)$co)
```

```
##
##   -999 -0.274 -0.129  -0.03  -0.02  -0.01
##      8      1      1     80    148    174
```

Дослідження колонки “o3”

Очевидно кодові значення: -999. Є від’ємні числа.

```
table(filter(dt, o3 < 0)$o3)
```

```
##
## -999   -1 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1
##    1   10   16   17   29   27   29   39   62  110  138
```

Дослідження колонки “o3_8hr”

Є від’ємне число -1. Хоча воно і схоже на кодове, за аналогією до попередніх колонок, припустимо, що воно справжнє. До того ж, частка таких рядків дуже мала: у цьому можна переконатися, поглянувши на гістограму.

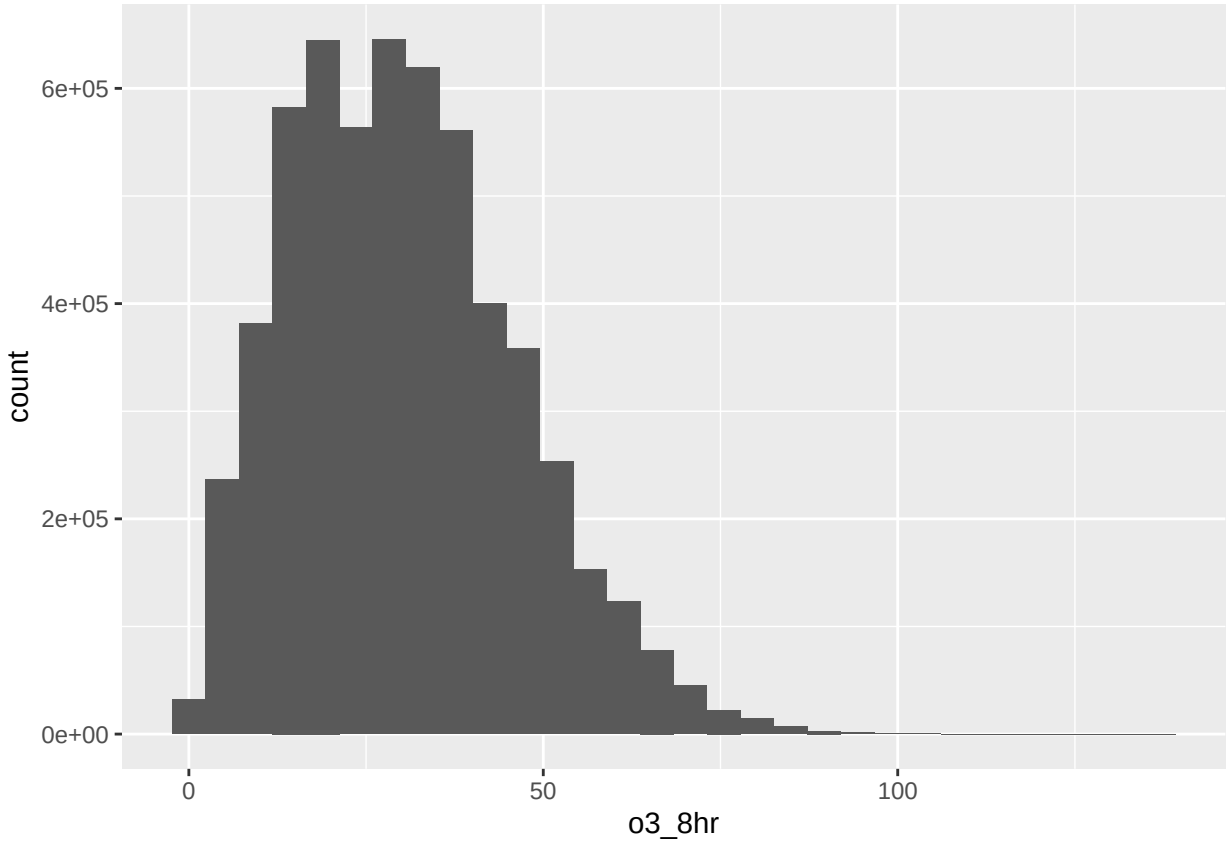
```
table(filter(dt, o3_8hr < 0)$o3_8hr)
```

```
##
## -1
## 50
```

```
ggplot(dt, aes(x = o3_8hr)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 153648 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Дослідження колонки “pm10”

Очевидно кодові значення: -999

```
table(filter(dt, pm10 < 0)$pm10)
```

```
##
## -999
##    2
```

Дослідження колонки “pm2.5”

Очевидно кодові значення: -999

```
table(filter(dt, pm2.5 < 0)$pm2.5)
```

```
##
## -999
##    3
```

Дослідження колонки “no2”

Є від’ємні числа.

```
table(filter(dt, no2 < 0)$no2)
```

```
##
## -27.78 -17.52   -0.5   -0.4  -0.39   -0.3   -0.2   -0.1
##      1      1    101    160      1    213    230    322
```

Дослідження колонки “nox”

Є від’ємні числа.

```
table(filter(dt, nox < 0)$nox)
```

```
##
## -1.6 -0.5 -0.4 -0.3 -0.2 -0.1
##    1    4   21   26   59   49
```

Дослідження колонки “no”

Є від’ємні числа.

```
table(filter(dt, no < 0)$no)
```

```
##
##  -7.2 -2.32 -2.04 -0.91 -0.87 -0.86 -0.74 -0.62  -0.6 -0.57 -0.51  -0.5 -0.46
##     1     1     1     1     1     1     1     2     1     1     1  1370     1
##  -0.4 -0.34 -0.31  -0.3 -0.29 -0.26 -0.22 -0.21  -0.2 -0.19 -0.17 -0.14 -0.13
##  3087     1     1  4478     1     3     1     1  7214     1     1     1     1
## -0.12 -0.11  -0.1 -0.09 -0.08 -0.05 -0.04 -0.03 -0.02 -0.01
##     4     1 13271     1     3     2     4     1     4     5
```

Дослідження колонки “windspeed”

Є від’ємні числа.

```
table(filter(dt, windspeed < 0)$windspeed)
```

```
##
## -0.4 -0.2 -0.1
##   15    1   56
```

Дослідження колонки “winddirec”

Очевидно кодові значення: 990

```r
table(filter(dt, winddirec < 0 | winddirec > 360)$winddirec)
```

```
##
## 990
## 578
```

Рядки, де `winddirec == 990`, у всьому іншому на вигляд цілком адекватні.

```r
dt %>% filter(winddirec == 990) %>% select(!c(0:3))
```

```
## # A tibble: 578 x 22
##      aqi pollutant status   so2    co    o3 o3_8hr  pm10 pm2.5   no2   nox    no
##    <dbl> <fct>    <fct>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1    49 <NA>      Good    NA    NA    42.4   53.4    NA     0   NA    NA    NA
##  2    58 Ozone (8~ Moder~   0.3  0.19  55.8   57.6     1     0   0.7   1.1   0.4
##  3    58 Ozone (8~ Moder~   0.2  0.19  55.4   57.9     2     0   0.9   1.4   0.5
##  4    58 Ozone (8~ Moder~   0.3  0.23  60.2   57.9     6     0   1.1   1.7   0.5
##  5    45 <NA>      Good     0.2  0.25  51.7   49.7     1     0   1.1   1.5   0.4
##  6    44 <NA>      Good     0.3  0.25  49.4   48.1     2     0   1.7   1.9   0.2
##  7    41 <NA>      Good     0.2  0.24  49.2   45       3     0   1.9   2.2   0.3
##  8    38 <NA>      Good     0.3  0.26  49.5   42       7     1   2.3   2.6   0.3
##  9    60 PM2.5     Moder~   0.8  0.55  14.5   24      NA    NA  20.3  20.9   0.5
## 10    80 PM2.5     Moder~   0.9  0.34  29.6   48.5    42    29   8.9   9.3   0.4
## # i 568 more rows
## # i 10 more variables: windspeed <dbl>, winddirec <dbl>, unit <lgl>,
## #   co_8hr <dbl>, pm2.5_avg <dbl>, pm10_avg <dbl>, so2_avg <dbl>,
## #   longitude <dbl>, latitude <dbl>, siteid <dbl>
```

Дослідження колонки “unit”

Усі NA. Колонку можна буде видалити.

```r
table(dt$unit)
```

```
## < table of extent 0 >
```

Дослідження колонки “co_8hr”

Є від’ємні числа.

```r
table(filter(dt, co_8hr < 0)$co_8hr)
```

```
##
## -1
## 35
```

Дослідження колонки “pm2.5_avg”

Є від’ємні числа.

```r
table(filter(dt, pm2.5_avg < 0)$pm2.5_avg)
```

```
##
## -1
##  6
```

Дослідження колонки “pm10_avg”

Є від’ємні числа.

```
table(filter(dt, pm10_avg < 0)$pm10_avg)
```

```
##
## -1
##  6
```

Дослідження колонки “so2_avg”

Є від’ємні числа.

```
table(filter(dt, so2_avg < 0)$so2_avg)
```

```
##
## -1
## 12
```

Дослідження колонки “longitude”

Значення 0 схоже на помилкове. Колонку можна буде видалити, оскільки цю інформацію навряд чи вдасться використати.

```
table(dt$longitude)
```

```
##
##           0   118.312256   119.566158 119.93149378   119.949875     119.9525
##         191        60297        60294        14280        39182           10
##  119.952724 120.12416667 120.12444444     120.1547 120.18339722 120.19933333
##        1455            2         8089            5        17763        41937
##  120.202617   120.202842     120.2175   120.218333 120.21947897    120.22085
##       43744        18364        60285           10         6505         6716
## 120.22238056 120.24205556
##        3309           21
##  [ reached getOption("max.print") -- omitted 169 entries ]
```

Дослідження колонки “latitude”

Значення 0 схоже на помилкове. Колонку можна буде видалити, оскільки цю інформацію навряд чи вдасться використати.

```
table(dt$latitude)
```

```
##
##           0   21.958069   22.260899    22.35222 22.37094722 22.38474167
##         191       60274       16152       41793        6157       13476
##     22.4794     22.4795   22.523108   22.526986   22.544317 22.54779444
##           3       60282       58125           4           3        1497
##  22.560847   22.564136 22.56413611   22.565747   22.565833    22.58564
##           1          10       41941       18332       60294         833
##  22.587069  22.6044507
##           3       15736
##  [ reached getOption("max.print") -- omitted 169 entries ]
```

Дослідження колонки “siteid”

Проблем не виявлено. Колонку можна буде видалити, оскільки цю інформацію навряд чи вдасться використати.

```
table(dt$siteid)
```

```
##
##     0     1     2     3     4     5     6     7     8     9    10    11    12
##    28 48909 48913 48922 48907 48908 48909 48911 48907 48911 49167 48938 48937
##    13    14    15    16    17    18    19    20    21    22    23    24    25
## 48912 48920 48913 48928 48910 48936 48907 48929 48907 48912 48922 48926 48915
##    26    27    28    29    30    31    32    33    34    35    36    37    38
## 48908 48913 48909 48902 48621 48919 48914 48907 48904 48909 48913 48878 48915
##    39    40    41    42    43    44    45    46    47    48    49    50    51
## 48901 48903 48908 48906 48908 48906 48905 48879 48905 48910 48905 44256 48893
##    52    53    54    55    56    57    58    59    60    61    62    63    64
## 48891 47832 48882    21 48727 48890 48901 48881 46733 48884 48907 48906 49174
##    65    66    67    68    69    70    71    72    73    74    75    76    77
## 48907 48910 48895 48907 48912 48895 48897 48907    21    21 48908    21 48886
##    78    80    83    84    85    86    87    90    92    96   201   202   203
## 48883 48869 48895 48907 32531    19    21    19    21    21 33508 40585 48201
##   204   206   310   311   312   313   314
## 48722     1 30426 40148 31521 38758   166
```

Підбиття підсумків, очищення від кодів і видалення непотрібних колонок

Колонки, що містять очевидно кодові значення:

- aqi - кодові: -1
- so2 - кодові: -999
- co - кодові: -999
- o3 - кодові: -999
- pm10 - кодові: -999
- pm2.5 - кодові: -999
- winddirec - кодові: 990

Колонки, що містять від’ємні значення:

- so2
- co
- o3
- o3_8hr
- no2
- nox
- no
- windspeed
- co_8hr
- pm2.5_avg
- pm10_avg
- so2_avg

Колонки, які можна видалити:

- unit - порожня колонка
- longitude - корисність під сумнівом
- latitude - корисність під сумнівом
- siteid - корисність під сумнівом

Припуститимо, що від'ємні показники є справжніми, а не кодовими, і виникли через незначний зсув у калібруванні датчиків. За неохідності (наприклад, для логаритмування) цей зсув можна буде компенсувати додаванням певного числа до всіх значень відповідної колонки.

Кодові значення замінимо на NA і видалимо непотрібні колонки.

```
dt <- dt %>%
  replace_with_na(replace = list(
    aqi = -1,
    so2 = -999,
    co = -999,
    o3 = -999,
    pm10 = -999,
    pm2.5 = -999,
    winddirec = 990
  ))

dt <- dt %>% select(!c(unit, longitude, latitude, siteid))
```

Оцінка кількості пропущених значень

Колонки, у яких NA більше, ніж 5%

```
dt %>%
  summarise(across(everything(), ~ mean(is.na(.)))) %>%
  select(where(~ all(. > 0.05)))
```

```
## # A tibble: 1 x 4
##   pollutant windspeed winddirec so2_avg
##       <dbl>     <dbl>     <dbl>   <dbl>
## 1     0.550    0.0515    0.0516   0.107
```

Повний код для підготовки даних

Повний код знаходиться у папці `code/lab1_cleaning.R`