China Contamination

Data Science: Capstone Proyect

Jorge Haces

16/6/2020

Contents

1 Overview

This is the final project required for **Data Science: Capstone** course offered by *edX HarvardX for Professional Certificate Program in Data Science*. The theme of this proyect is the Contamination, specifically in China and the aim is to predict the air quality in the fastest growing country nowadays.

1.1 Introduction

Contanimation is defined as the presence of materials in the air that cause serious harm or discomfort to people. The contamination has increased since the Industrial Revolution began, in the second half of the 18th century, with production processes in factories, the development of transportation and the use of fuels.

According to the World Health Organization (WHO), the state of the current atmosphere causes, by the simple act of breathing, the death of around seven million people a year (fine particle respiration).

The most common air pollutant gases are carbon monoxide, sulfur dioxide, chlorofluorocarbons, and nitrogen oxides. Photochemicals such as ozone and smog are increased in the air by nitrogen oxides and hydrocarbons reacting with sunlight.

Contaminants are classified into:

- Primaries are those that are emitted directly into the atmosphere such as sulfur dioxide, carbon monoxide
- Secondary are those that are formed by atmospheric chemical processes that act on primary contaminants such as sulfuric acid, which is formed by the oxidation of sulfur dioxide, nitrogen dioxide that is formed by oxidizing the primary pollutant nitric oxide and ozone that is formed from oxygen. [1]

1.2 Project Description

An analysis of the data will be carried out based on the following models: k Nearest Neighbors, Logistic Regression, Support Vector Machines (SVM), Random Forests and Neural Network to help us predict if pollution will grow even more (2.3% in 2018 almost at double compared to 2010). To answer the question, will china be able to comply with the Paris agreement signed in 2015? Whose goal is to reduce the global temperature to 2°C in 2050.

For this we will divide the data into two: training data and test data. Later, we will train the different models in the first set and then will be evaluated in the second set. Finally, we will use the **Root-Mean-Square-Error (RMSE)** and the **"overall accuracy"** to rate the performance of each model and thus identify the best for this project.

1.3 DataSet

The Dataset used in this project is *Beijing Multi-Site Air-Quality Data Data Set*, available at the UCI Machine Learning Repository [2].

This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017. Missing data are denoted as NA.

The Attribute Information is the following:

- No: row number
- year: year of data in this row
- month: month of data in this row
- day: day of data in this row
- hour: hour of data in this row
- PM2.5: PM2.5 concentration (ug/mˆ3)
- PM10: PM10 concentration (ug/mˆ3)
- SO2: SO2 concentration (ug/mˆ3)
- NO2: NO2 concentration (ug/mˆ3)
- CO: CO concentration (ug/mˆ3)
- O3: O3 concentration (ug/mˆ3)
- TEMP: temperature (degree Celsius)
- PRES: pressure (hPa)
- DEWP: dew point temperature (degree Celsius)
- RAIN: precipitation (mm)
- wd: wind direction
- WSPM: wind speed (m/s)
- station: name of the air-quality monitoring site

The data is contained in a Zip file named *PRSA2017_Data_20130301-20170228.zip* containing 12 files (one for each municipality), as follow:

- PRSA_Data_Aotizhongxin_20130301-20170228.csv
- PRSA_Data_Changping_20130301-20170228.csv
- PRSA_Data_Dingling_20130301-20170228.csv
- PRSA_Data_Dongsi_20130301-20170228.csv
- PRSA_Data_Guanyuan_20130301-20170228.csv
- PRSA_Data_Gucheng_20130301-20170228.csv
- PRSA_Data_Huairou_20130301-20170228.csv
- PRSA_Data_Nongzhanguan_20130301-20170228.csv
- PRSA_Data_Shunyi_20130301-20170228.csv
- PRSA_Data_Tiantan_20130301-20170228.csv
- PRSA_Data_Wanliu_20130301-20170228.csv
- PRSA_Data_Wanshouxigong_20130301-20170228.csv

2 Methods and Analysis

2.1 Data Stage

Next, the UCI data will be downloaded in ZIP format to decompress them and load the 12 files in a single variable called PRSA, identifying the data to be analyzed with 420,768 records and 18 attributes (columns).

```
## [1] 420768     18
```

Subsequently, we execute the nearZeroVar function to identify the attributes that have no significant variation in our data set; In this case, the attribute "Rain" does not present significant variations, so we remove that attribute.

```
# Structure of the data (data type, numbers of rows, number of attributes)
str(PRSA)
```

```
## tibble [420,768 x 18] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ No     : num [1:420768] 1 2 3 4 5 6 7 8 9 10 ...
##  $ year   : num [1:420768] 2013 2013 2013 2013 2013 ...
##  $ month  : num [1:420768] 3 3 3 3 3 3 3 3 3 3 ...
##  $ day    : num [1:420768] 1 1 1 1 1 1 1 1 1 1 ...
##  $ hour   : num [1:420768] 0 1 2 3 4 5 6 7 8 9 ...
##  $ PM2.5  : num [1:420768] 4 8 7 6 3 5 3 3 3 3 ...
##  $ PM10   : num [1:420768] 4 8 7 6 3 5 3 6 6 8 ...
##  $ SO2    : num [1:420768] 4 4 5 11 12 18 18 19 16 12 ...
##  $ NO2    : num [1:420768] 7 7 10 11 12 18 32 41 43 28 ...
##  $ CO     : num [1:420768] 300 300 300 300 300 400 500 500 500 400 ...
##  $ O3     : num [1:420768] 77 77 73 72 72 66 50 43 45 59 ...
##  $ TEMP   : num [1:420768] -0.7 -1.1 -1.1 -1.4 -2 -2.2 -2.6 -1.6 0.1 1.2 ...
##  $ PRES   : num [1:420768] 1023 1023 1024 1024 1025 ...
##  $ DEWP   : num [1:420768] -18.8 -18.2 -18.2 -19.4 -19.5 -19.6 -19.1 -19.1 -19.2 -19.3 ...
##  $ RAIN   : num [1:420768] 0 0 0 0 0 0 0 0 0 0 ...
##  $ wd     : chr [1:420768] "NNW" "N" "NNW" "NW" ...
##  $ WSPM   : num [1:420768] 4.4 4.7 5.6 3.1 2 3.7 2.5 3.8 4.1 2.6 ...
##  $ station: chr [1:420768] "Aotizhongxin" "Aotizhongxin" "Aotizhongxin" "Aotizhongxin" ...
```

```
#Near zero variance ( identify the attributes that do not give us valuable data )
nzv <- nearZeroVar(PRSA)

#Remove the nzv columns and save in another variable
data <- PRSA[,-nzv]
```

Therefore, we will continue with 17 attributes.

```
# Dimensions of the valuable data (rows,columns)
dim(data)
```

```
## [1] 420768     17
```

It was identified that the attribute "No" is a consecutive attribute and not contribute anything to the data, therefore it will also be discarded. It is identified that all the attributes are numerical type with the exception of "station" which is character type, therefore, the attributes will be transformed to factor type to make the data set more efficient.

```
# Show the data in a transposed version to see more data
glimpse(data)
```

```
## Rows: 420,768
## Columns: 17
```

```
## $ No      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...
## $ year    <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
## $ month   <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ day     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ hour    <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ PM2.5   <dbl> 4, 8, 7, 6, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 6, 8, 9, 10, 11, ...
## $ PM10    <dbl> 4, 8, 7, 6, 3, 5, 3, 6, 6, 8, 6, 6, 6, 6, 9, 15, 19, 23, 20...
## $ SO2     <dbl> 4, 4, 5, 11, 12, 18, 18, 19, 16, 12, 9, 9, 7, 7, 7, 7, 9, 1...
## $ NO2     <dbl> 7, 7, 10, 11, 12, 18, 32, 41, 43, 28, 12, 14, 13, 12, 11, 1...
## $ CO      <dbl> 300, 300, 300, 300, 300, 400, 500, 500, 500, 400, 400, 400,...
## $ O3      <dbl> 77, 77, 73, 72, 72, 66, 50, 43, 45, 59, 72, 71, 74, 76, 77,...
## $ TEMP    <dbl> -0.7, -1.1, -1.1, -1.4, -2.0, -2.2, -2.6, -1.6, 0.1, 1.2, 1...
## $ PRES    <dbl> 1023.0, 1023.2, 1023.5, 1024.5, 1025.2, 1025.6, 1026.5, 102...
## $ DEWP    <dbl> -18.8, -18.2, -18.2, -19.4, -19.5, -19.6, -19.1, -19.1, -19...
## $ wd      <chr> "NNW", "N", "NNW", "NW", "N", "N", "NNE", "NNW", "NNW", "N"...
## $ WSPM    <dbl> 4.4, 4.7, 5.6, 3.1, 2.0, 3.7, 2.5, 3.8, 4.1, 2.6, 3.6, 3.7,...
## $ station <chr> "Aotizhongxin", "Aotizhongxin", "Aotizhongxin", "Aotizhongx...
```

```r
#Remove the "no" variable because is a consecutive (also do not give us valuable data)
data <- data[,-c(1)]

# Cast the attributes year, month, day, hour, wd and station from "dbl" type to "factor" type
data$year <- as.factor(data$year)
data$month <- as.factor(data$month)
data$day <- as.factor(data$day)
data$hour <- as.factor(data$hour)
data$wd <- as.factor(data$wd)
data$station <- as.factor(data$station)
```

Similarly, it is validated that the PM (2.5 and 10) attributes; in this case, only the attribute "CO" will be transformed to an integer.

```
##   [1]   4.0   8.0   7.0   6.0   3.0   5.0   9.0  10.0  11.0  12.0  15.0  24.0
##  [13]  22.0  14.0  13.0  18.0  26.0  25.0  37.0  44.0  54.0  61.0  67.0  74.0
##  [25]  81.0  93.0 112.0 109.0 110.0 105.0 106.0 101.0  91.0  79.0  77.0  83.0
##  [37]  94.0  96.0  98.0  72.0  48.0  28.0  19.0  60.0 117.0  46.0  42.0  49.0
##  [49]  34.0  16.0  20.0  31.0  40.0  51.0  58.0  71.0 100.0 103.0 115.0 104.0
##  [61] 111.0 114.0 131.0 138.0 142.0 163.0 168.0 165.0 166.0 190.0 203.0 210.0
##  [73] 205.0 215.0 219.0 226.0 284.0 272.0 242.0 212.0 192.0 188.0 186.0 194.0
##  [85] 184.0 175.0 179.0 201.0 230.0 246.0 248.0 239.0 254.0 266.0 260.0 250.0
##  [97] 240.0 238.0 243.0 275.0 306.0 320.0 292.0 251.0 228.0 217.0 255.0 258.0
## [109] 277.0 282.0 267.0 216.0 211.0 227.0 363.0 376.0 344.0 339.0 324.0 322.0
## [121] 315.0 310.0 309.0 326.0 121.0 330.0 172.0  82.0  84.0  90.0  89.0  97.0
## [133] 102.0 108.0 113.0 145.0  69.0  78.0  47.0  21.0  17.0  27.0  30.0  36.0
## [145]  38.0  70.0  64.0  95.0 120.0 141.0 140.0 149.0 150.0 132.0 144.0 156.0
## [157] 174.0 152.0 125.0 146.0 116.0 123.0 130.0 129.0 139.0 135.0 133.0 124.0
## [169] 137.0 160.0  88.0  50.0  23.0  35.0  62.0  65.0  73.0 119.0 162.0 180.0
## [181] 183.0 182.0 173.0 164.0 236.0 256.0 287.0 278.0 269.0 286.0 304.0 312.0
## [193] 316.0 353.0 355.0 328.0 308.0 314.0 276.0 264.0 270.0 185.0 197.0 176.0
## [205] 206.0 257.0 293.0 294.0 332.0 352.0 362.0 356.0 357.0 364.0 392.0 423.0
## [217] 434.0 450.0 463.0  55.0  33.0  29.0  32.0  59.0  45.0  52.0  39.0  68.0
## [229]  41.0  92.0 178.0 181.0 187.0 202.0  63.0  56.0  75.0  66.0  76.0  53.0
## [241]  80.0  86.0  43.0 118.0 209.0 220.0 223.0 214.0 200.0 199.0 204.0 222.0
## [253] 229.0 143.0 128.0 221.0 159.0  57.0 151.0 225.0 235.0 233.0 126.0 161.0
## [265] 136.0 134.0 147.0 154.0 153.0 157.0 208.0 122.0  85.0  99.0 107.0 127.0
```

```
## [277]  87.0 171.0 189.0 198.0 232.0 263.0 237.0 169.0 193.0    NA 195.0 366.0
## [289] 170.0 252.0 262.0 279.0 288.0 273.0 311.0 303.0 167.0 665.0 158.0 148.0
## [301] 155.0 218.0 191.0 177.0 207.0 247.0 368.0 317.0 249.0 464.0 494.0 485.0
## [313] 510.0 430.0 281.0 271.0 245.0 224.0 213.0 253.0 234.0 265.0 283.0 290.0
## [325] 297.0 321.0 295.0 280.0 340.0 349.0 334.0 289.0 305.0 261.0 345.0 327.0
## [337] 313.0 302.0 196.0 268.0 274.0 231.0 244.0 342.0 367.0 307.0 241.0 361.0
## [349] 341.0 285.0 300.0 388.0 385.0 378.0 371.0 413.0 420.0 335.0 350.0 323.0
## [361] 337.0 331.0 296.0 402.0 259.0 436.0 472.0 499.0 511.0 530.0 538.0 535.0
## [373] 521.0 475.0 398.0 383.0 389.0 396.0 418.0 422.0 333.0 298.0 343.0 481.0
## [385] 501.0 584.0 525.0 461.0 414.0 455.0 505.0 466.0 401.0 403.0 412.0 503.0
## [397] 500.0 488.0 478.0 456.0 435.0 425.0 407.0 358.0 379.0 375.0 291.0 299.0
## [409] 325.0 377.0 427.0 416.0 394.0 346.0 483.0 489.0 426.0 445.0 440.0 518.0
## [421] 498.0 319.0 318.0 301.0 374.0 370.0 410.0 354.0  87.8  84.6  85.6 225.6
## [433]  26.8 102.3 113.6  78.3  81.3 381.0 393.0 384.0 336.0 338.0 446.0 351.0
## [445] 329.0 360.0 432.0 431.0 409.0   8.6 404.0 419.0 411.0 347.0 365.0 390.0
## [457] 408.0 497.0 469.0 438.0 399.0 114.1  66.2  67.5 397.0 359.0 380.0 369.0
## [469] 492.0 619.0 618.0 606.0 583.0 598.0 612.0 587.0 565.0 542.0 439.0 444.0
## [481] 453.0 552.0 473.0 529.0 550.0 607.0 604.0 569.0 577.0 470.0 487.0 507.0
## [493] 405.0 586.0 627.0 646.0 644.0 657.0 635.0 546.0 476.0 477.0 480.0 479.0
## [505] 467.0 428.0 527.0 523.0 382.0 519.0 541.0 532.0 547.0 898.0 713.0 615.0
## [517] 585.0 544.0 348.0 433.0 406.0 386.0 448.0 395.0 373.0 400.0 387.0 451.0
## [529] 454.0 421.0 424.0 568.0 543.0 522.0 459.0 417.0 429.0 437.0 447.0 441.0
## [541] 576.0 641.0 651.0 682.0 697.0 462.0 443.0   2.0 391.0 581.0 524.0  74.5
## [553]  75.8  80.7  80.4 154.2 540.0  26.9  51.7  76.7  15.5  18.9 112.4  79.5
## [565] 452.0 474.0 458.0 471.0 882.0 557.0 517.0 596.0 415.0 560.0 662.0 495.0
## [577] 491.0 539.0 442.0 548.0  77.4  83.3  78.7  88.6 139.7   9.6   7.9   7.2
## [589]  15.7 120.4   8.4  82.1  72.4   4.3   8.5 117.9 103.8  61.2  12.6  13.8
## [601] 468.0 515.0 506.0 496.0 614.0 632.0 617.0 647.0 594.0 513.0 509.0 490.0
## [613] 564.0 881.0 610.0 372.0 493.0 536.0 520.0 531.0 553.0 593.0 597.0 603.0
## [625] 590.0 555.0 504.0 684.0 737.0 679.0 449.0 516.0 105.4  92.8  94.1 238.6
## [637]  23.7  42.3  97.9  45.4 120.7  84.7 147.9   4.4  10.7 174.3  74.3 460.0
## [649] 512.0 624.0 626.0 629.0 638.0 666.0 670.0 685.0 640.0 628.0 637.0 680.0
## [661] 678.0 671.0 482.0 625.0 558.0 561.0 575.0 573.0 622.0 660.0 554.0 661.0
## [673] 642.0 695.0 537.0 508.0 457.0 484.0 572.0 580.0 589.0 563.0 533.0 514.0
## [685] 599.0 654.0 663.0 620.0 645.0 681.0 528.0 595.0 578.0 549.0 486.0 502.0
## [697] 116.7  84.4  72.6  80.6  23.8  99.1 111.3  71.4 121.7  14.3 256.9 136.2
## [709]  62.8 556.0 609.0 636.0 592.0 605.0 633.0 465.0 526.0 534.0 639.0 570.0
## [721] 545.0 664.0 567.0 602.0 770.0 110.9 156.7  88.9  78.4  80.5  40.3  15.9
## [733]  38.6  20.6  58.8  70.7 123.7  48.8 114.6 652.0 574.0  70.3 675.0 733.0
## [745] 741.0 588.0 649.0 705.0 559.0 600.0 571.0 551.0 634.0 608.0 677.0 658.0
## [757] 739.0 767.0 566.0  68.9  72.5  77.2 193.3  13.5  42.5  92.6  65.3  53.5
## [769]  12.7 144.6  55.1 659.0 762.0  83.7  81.8  55.7 683.0 844.0 809.0 781.0
## [781] 687.0  91.3  78.2  32.2  23.6  81.7  91.7 100.9 623.0 197.1  71.2 601.0
## [793] 611.0 667.0 579.0 835.0 744.0  41.2  60.5  65.7  66.1  66.5 275.1 208.1
## [805]  13.7  67.4  99.4  92.4  14.7  52.8  83.9  67.8  11.2   4.6 613.0 127.2
## [817]  38.1  20.8 689.0 631.0 816.0 941.0 707.0 582.0 650.0  18.3  87.3  87.9
## [829] 235.3  28.1  92.1 116.9  86.8 145.8  68.6 591.0 821.0 801.0 758.0 720.0
## [841] 712.0 743.0 808.0 691.0 125.7  81.2 197.2  19.3 104.1 150.8 115.5  71.5
## [853]  85.2  20.7  11.5 105.6  64.9 718.0 708.0 957.0 791.0 692.0  12.5 704.0
## [865] 616.0 655.0 111.8  98.5  89.1 224.5  13.4  92.9 113.3  77.9 125.3   8.8
## [877] 106.4 153.8 669.0 690.0 826.0 999.0 857.0 748.0 630.0 621.0 804.0 730.0
## [889] 823.0

##    [1]   4.0   8.0   7.0   6.0   3.0   5.0   9.0  15.0  19.0  23.0  20.0  14.0
```

```
##   [13]  17.0  18.0  24.0  13.0  11.0  10.0  29.0  30.0  33.0  35.0  40.0  46.0
##   [25]  58.0  79.0  86.0  96.0 103.0 113.0 120.0 130.0 132.0 129.0 136.0 135.0
##   [37] 142.0 116.0 110.0 119.0 122.0 117.0 108.0 134.0 106.0 114.0 100.0  82.0
##   [49]  71.0 175.0 181.0 105.0  94.0  83.0  80.0  62.0  34.0  32.0  36.0  28.0
##   [61]  12.0  44.0  63.0  84.0  85.0  91.0 127.0 151.0 153.0 145.0 146.0 147.0
##   [73] 159.0 177.0 184.0 193.0 182.0 171.0 186.0 218.0 248.0 255.0 244.0 252.0
##   [85] 253.0 269.0 315.0 300.0 265.0 229.0 203.0 194.0 196.0 205.0 198.0 212.0
##   [97] 219.0 233.0 277.0 285.0 276.0 291.0 338.0 396.0 380.0 335.0 360.0 319.0
##  [109] 297.0 294.0 293.0 304.0 337.0 366.0 374.0 344.0 282.0 311.0 283.0 284.0
##  [121] 326.0 310.0 257.0 287.0 298.0 377.0 452.0 426.0 389.0 370.0 400.0 373.0
##  [133] 345.0 346.0 327.0 318.0 371.0 844.0 362.0 225.0 348.0 508.0 128.0 123.0
##  [145] 104.0  87.0 102.0  67.0  97.0 125.0 118.0 143.0 139.0 166.0 272.0 587.0
##  [157]    NA 628.0 662.0 443.0 263.0 251.0 107.0  16.0  25.0  21.0  74.0  75.0
##  [169]  59.0  56.0  52.0  54.0 115.0 140.0 156.0 138.0 155.0 207.0 189.0 133.0
##  [181] 162.0 214.0 246.0 124.0 161.0 237.0 165.0 121.0  90.0  93.0 131.0 101.0
##  [193]  31.0  50.0  22.0  38.0  47.0  39.0  45.0  88.0  66.0  53.0  77.0  81.0
##  [205]  95.0 112.0 126.0 137.0 158.0 179.0 176.0 211.0 200.0 170.0 174.0 185.0
##  [217] 331.0 261.0 279.0 305.0 343.0 320.0 324.0 367.0 430.0 411.0 357.0 330.0
##  [229] 332.0 316.0 289.0 273.0 243.0 169.0 167.0 204.0 217.0 223.0 242.0 192.0
##  [241] 195.0 188.0 210.0 259.0 321.0 347.0 372.0 392.0 383.0 369.0 395.0 390.0
##  [253] 442.0 462.0 455.0 494.0 476.0  68.0  76.0  42.0  41.0  57.0  61.0  49.0
##  [265]  78.0  27.0  51.0  48.0  89.0 141.0 163.0 172.0 187.0 202.0 221.0 240.0
##  [277] 236.0 234.0 209.0 168.0 152.0 157.0  70.0  69.0 111.0  26.0 160.0 260.0
##  [289] 292.0 264.0 256.0 241.0 274.0 302.0 230.0 216.0 206.0  73.0  99.0  98.0
##  [301]  92.0 231.0 222.0 247.0 250.0 245.0 239.0 280.0 268.0 267.0 232.0 249.0
##  [313] 178.0 227.0 238.0 197.0 275.0 351.0 358.0 191.0 144.0 190.0 149.0  43.0
##  [325]  72.0  55.0  60.0 215.0 150.0 109.0 164.0 173.0 208.0 224.0  65.0 312.0
##  [337]  37.0 154.0  64.0 183.0 213.0 199.0 254.0 148.0 309.0 299.0 303.0 308.0
##  [349] 359.0 354.0 336.0 365.0 306.0 228.0 325.0 258.0 180.0 201.0 235.0 655.0
##  [361] 353.0 341.0 340.0 296.0 262.0 290.0 388.0 278.0 328.0 364.0 407.0 271.0
##  [373] 226.0 270.0 333.0 281.0 220.0 339.0 507.0 322.0 528.0 544.0 527.0 564.0
##  [385] 375.0 317.0   2.0 286.0 266.0 314.0 323.0 295.0 313.0 382.0 301.0 409.0
##  [397] 406.0 397.0 334.0 288.0  10.5 307.0 416.0 408.0 419.0 350.0 460.0 539.0
##  [409] 423.0 594.0 470.0 516.0 536.0 540.0 549.0 555.0 640.0 610.0 612.0 598.0
##  [421] 499.0 427.0 431.0 445.0 401.0 356.0 483.0 329.0 415.0 386.0 561.0 654.0
##  [433] 488.0 422.0 393.0 492.0 510.0 501.0 481.0 468.0 435.0 403.0 379.0 349.0
##  [445] 352.0 434.0 428.0 418.0 479.0 475.0 413.0 446.0 421.0 458.0 520.0 498.0
##  [457] 466.0 453.0 438.0 518.0 478.0 342.0 417.0 471.0 429.0 486.0 469.0 424.0
##  [469] 477.0 368.0 402.0 463.0 447.0 436.0 485.0 521.0 554.0 603.0 502.0 404.0
##  [481] 530.0 948.0 548.0 524.0 114.6 131.9 125.3  98.8 399.0 341.5 391.0 385.0
##  [493] 601.0 526.0 725.0  26.8 170.1 112.2  81.3 381.0 384.0 394.0 448.0 441.0
##  [505] 432.0 433.0 378.0 414.0 376.0  37.8 440.0 363.0 405.0 450.0 613.0 439.0
##  [517] 568.0 629.0 609.0 500.0 482.0 437.0 361.0 633.0 634.0 506.0 538.0 355.0
##  [529] 387.0 578.0 497.0 489.0 464.0 472.0 465.0 137.6 570.0 533.0 585.0 721.0
##  [541] 845.0 862.0 757.0 637.0 412.0 812.0 874.0 504.0 456.0 588.0 552.0 514.0
##  [553] 474.0 420.0 398.0  36.9  64.7   6.4  92.8  33.9   6.6 984.0 944.0 525.0
##  [565] 576.0 511.0 630.0 661.0 671.0 664.0 620.0 631.0 567.0 535.0 542.0 444.0
##  [577] 473.0 529.0 550.0 619.0 581.0 600.0 410.0 534.0 509.0 512.0 451.0 646.0
##  [589] 644.0 675.0 647.0 557.0 537.0 573.0 571.0 580.0 558.0 577.0 599.0 590.0
##  [601] 604.0 523.0 513.0 582.0 565.0 884.0 762.0 722.0 693.0 666.0 873.0 638.0
##  [613] 777.0 625.0 627.0 563.0 642.0 827.0 834.0 575.0 449.0 496.0 425.0 572.0
##  [625] 543.0 491.0 467.0 484.0 493.0 815.0 773.0 799.0 858.0 754.0 653.0 579.0
##  [637] 595.0 999.0 584.0 531.0 546.0 195.5 169.1  90.4  15.5 522.0 770.0 652.0
##  [649] 487.0 112.4 547.0 720.0 793.0 930.0 980.0 992.0 747.0 608.0 480.0 976.0
```

```
##  [661]  79.8 714.0 641.0 503.0 551.0 562.0 665.0 702.0 691.0 688.0 674.0 683.0
##  [673] 657.0 663.0 686.0 685.0 677.0 660.0 632.0 692.0 454.0 541.0 635.0 706.0
##  [685] 517.0 933.0 490.0 596.0 457.0 597.0 775.0 559.0 717.0 895.0 801.0  99.5
##  [697]  91.5  96.6 152.7  26.9   7.9  21.1  15.7 171.2   8.4 131.5   9.6 623.0
##  [709] 117.9 616.0 731.0 842.0 904.0 905.0 782.0 776.0 611.0 556.0  91.9  60.9
##  [721]  14.5   5.6 138.5  61.4  49.3 545.0 676.0  36.5  17.1  30.4  42.4  45.5
##  [733] 645.0 650.0 461.0 515.0 734.0 771.0 737.0 699.0 690.0 673.0 532.0 864.0
##  [745] 678.0 792.0 811.0 828.0 553.0 602.0 495.0 589.0 560.0 760.0 726.0 566.0
##  [757] 459.0 319.8  23.7  42.3  45.4 201.1 114.1 147.9  22.2 659.0 672.0 574.0
##  [769] 794.0 185.9 164.9  16.3 622.0 593.0 669.0 649.0 816.0 680.0 606.0 583.0
##  [781]  84.1 221.6 643.0 624.0 626.0 807.0 876.0 787.0 758.0 708.0 888.0 891.0
##  [793] 830.0 848.0 800.0 586.0 710.0 715.0 701.0 618.0 605.0 607.0 695.0 703.0
##  [805] 591.0 694.0 748.0 759.0 819.0 857.0 856.0 847.0 735.0 955.0 907.0 915.0
##  [817] 796.0 825.0 987.0 743.0 519.0 505.0 814.0  79.5 322.1  23.8  25.4 145.4
##  [829] 179.9  71.4 121.7 305.4 157.7 136.2 117.6 124.4  57.8  62.9 961.0 656.0
##  [841] 658.0 592.0 614.0 802.0 744.0 636.0 711.0 806.0 769.0 906.0 617.0 813.0
##  [853] 783.0 746.0 651.0 917.0 820.0 947.0 983.0 957.0 639.0 804.0 697.0 849.0
##  [865] 195.2 186.6  88.9 126.5 134.1 315.1  40.3  15.9  38.6  58.8 123.7 142.6
##  [877] 116.4 112.3 707.0 817.0 890.0 941.0  47.3  53.7 733.0 741.0 705.0 689.0
##  [889] 826.0 772.0 986.0 883.0 681.0 700.0 728.0 805.0 724.0 679.0 682.0 994.0
##  [901] 738.0 835.0 887.0 193.3  42.5  99.9  53.5  34.2 993.0 180.2  84.7  55.1
##  [913] 991.0 922.0 973.0  60.3  50.3  15.8  11.8 104.4   9.8 107.4 111.5 125.6
##  [925] 569.0 750.0 648.0 740.0 615.0 687.0  32.2  85.6  23.6 139.4 146.1 117.2
##  [937] 100.9  30.7 742.0 668.0 899.0  79.3  16.4   9.5 126.3 169.7 153.3  93.3
##  [949]  81.6  55.3  35.9  70.6 103.9  72.5 214.3 667.0 909.0 995.0 716.0 939.0
##  [961] 878.0 766.0 778.0 764.0  41.2 307.6 208.1  67.4  99.4  92.4  14.7  52.8
##  [973]  74.9  28.4 903.0 127.2  38.1 914.0 912.0 920.0  41.8  84.4  42.7  26.6
##  [985]  33.5  59.3 106.1 194.6 118.8  51.8 790.0  87.8  20.8  54.5 684.0 823.0
##  [997] 951.0 786.0 789.0 839.0 894.0  87.9  28.1 175.1 135.2  30.1 108.9 159.4
## [1009] 172.6  45.6   8.7   7.7   5.4   8.2  67.9  68.9  68.3 121.5 198.3 144.5
## [1021] 242.7 229.6 180.6 169.5 156.4 135.9 108.6  74.1  28.8  31.7  99.2  78.4
## [1033] 102.8 215.2 988.0 927.0 870.0 781.0 893.0 709.0 863.0 829.0 107.5 923.0
## [1045] 335.1 158.8 170.3 161.1  81.4  85.2  77.5  11.5  73.4 207.7 718.0 712.0
## [1057] 919.0 751.0 736.0 785.0 704.0 621.0 784.0 902.0  24.5 732.0  90.5 670.0
## [1069] 768.0 886.0 730.0 836.0 145.5  89.1 176.4 133.1  28.9 150.4 931.0 929.0
## [1081] 745.0 713.0 763.0 952.0 950.0
```

```
##   [1]   300   400   500   600   700   800   900  1000  1200  1100  1300  1399
##  [13]  1500  1700  1899  2200  2399  2500  2799  2100  2000  2299  2600  1800
##  [25]  1600    NA  2899  2700  3100  3200  4400  4000  3500  3700  4200  3799
##  [37]  4099  5000  5599  5700  4599  3299  3399  3899  4900  5200  3600  3000
##  [49]   200  1400   100  1900  2300  2400  2800  3300  2900  3400  5400  5800
##  [61]  5900  5300  3800  3900  4100  4500  5100  5600  4800  4600  6300  6400
##  [73]  4300  4700  6000  6700  7300  7000  7100  7700  8300  8900  8100  7600
##  [85]  6800  5500  6500  7400  6900  6200  6600  6100  7800  7200  7900  8000
##  [97]  9000  7500  8800  9300  9100  8400  9600 10000  9500  9900  9400  8500
## [109]  8200  8600  8700  9700  9200  4799  5099  5299  6299  9800   350   950
## [121]  1150   150  4299  6599  6799  6099  7299  8199  8099  7099  5799  7599
## [133]  8599
```

```
# Review the first 6 rows
head(data)
```

```
## # A tibble: 6 x 16
##    year  month day   hour  PM2.5  PM10   SO2   NO2    CO    O3  TEMP  PRES  DEWP
```

```
##   <fct> <fct> <fct> <fct> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl>
## 1 2013  3     1     0         4     4     4     7   300    77  -0.7 1023  -18.8
## 2 2013  3     1     1         8     8     4     7   300    77  -1.1 1023. -18.2
## 3 2013  3     1     2         7     7     5    10   300    73  -1.1 1024. -18.2
## 4 2013  3     1     3         6     6    11    11   300    72  -1.4 1024. -19.4
## 5 2013  3     1     4         3     3    12    12   300    72  -2   1025. -19.5
## 6 2013  3     1     5         5     5    18    18   400    66  -2.2 1026. -19.6
## # ... with 3 more variables: wd <fct>, WSPM <dbl>, station <fct>
```

We visualize the data set with the changes made

```
# Review the previous changes in a transposed version to see more data
glimpse(data)
```

```
## Rows: 420,768
## Columns: 16
## $ year    <fct> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
## $ month   <fct> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,...
## $ day     <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ hour    <fct> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ PM2.5   <dbl> 4, 8, 7, 6, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 6, 8, 9, 10, 11, ...
## $ PM10    <dbl> 4, 8, 7, 6, 3, 5, 3, 6, 6, 8, 6, 6, 6, 6, 9, 15, 19, 23, 20...
## $ SO2     <dbl> 4, 4, 5, 11, 12, 18, 18, 19, 16, 12, 9, 9, 7, 7, 7, 7, 9, 1...
## $ NO2     <dbl> 7, 7, 10, 11, 12, 18, 32, 41, 43, 28, 12, 14, 13, 12, 11, 1...
## $ CO      <int> 300, 300, 300, 300, 300, 400, 500, 500, 500, 400, 400, 400,...
## $ O3      <dbl> 77, 77, 73, 72, 72, 66, 50, 43, 45, 59, 72, 71, 74, 76, 77,...
## $ TEMP    <dbl> -0.7, -1.1, -1.1, -1.4, -2.0, -2.2, -2.6, -1.6, 0.1, 1.2, 1...
## $ PRES    <dbl> 1023.0, 1023.2, 1023.5, 1024.5, 1025.2, 1025.6, 1026.5, 102...
## $ DEWP    <dbl> -18.8, -18.2, -18.2, -19.4, -19.5, -19.6, -19.1, -19.1, -19...
## $ wd      <fct> NNW, N, NNW, NW, N, N, NNE, NNW, NNW, N, NNW, N, NNW, NW, N...
## $ WSPM    <dbl> 4.4, 4.7, 5.6, 3.1, 2.0, 3.7, 2.5, 3.8, 4.1, 2.6, 3.6, 3.7,...
## $ station <fct> Aotizhongxin, Aotizhongxin, Aotizhongxin, Aotizhongxin, Aot...
```

Finally, we run the display of the attribute statistics. It is important to note that the mean and median values, in every attribute, are not very far from each other, therefore, there are not many scattered data.

```
# Review the statistics of each attribute
summary(data)
```

```
##    year            month             day              hour
##  2013: 88128   1      : 35712   1      : 13824   0      : 17532
##  2014:105120   3      : 35712   2      : 13824   1      : 17532
##  2015:105120   5      : 35712   3      : 13824   2      : 17532
##  2016:105408   7      : 35712   4      : 13824   3      : 17532
##  2017: 16992   8      : 35712   5      : 13824   4      : 17532
##                10     : 35712   6      : 13824   5      : 17532
##                (Other):206496   (Other):337824   (Other):315576
##      PM2.5             PM10            SO2              NO2
##  Min.   :  2.00   Min.   :  2.0   Min.   :  0.286   Min.   :  1.026
##  1st Qu.: 20.00   1st Qu.: 36.0   1st Qu.:  3.000   1st Qu.: 23.000
##  Median : 55.00   Median : 82.0   Median :  7.000   Median : 43.000
##  Mean   : 79.79   Mean   :104.6   Mean   : 15.831   Mean   : 50.639
##  3rd Qu.:111.00   3rd Qu.:145.0   3rd Qu.: 20.000   3rd Qu.: 71.000
##  Max.   :999.00   Max.   :999.0   Max.   :500.000   Max.   :290.000
##  NA's   :8739     NA's   :6449    NA's   :9021      NA's   :12116
##        CO              O3               TEMP             PRES
##  Min.   :  100   Min.   :   0.214   Min.   :-19.90   Min.   : 982.4
```

```
##  1st Qu.:  500   1st Qu.:  11.000   1st Qu.:  3.10   1st Qu.:1002.3
##  Median :  900   Median :  45.000   Median : 14.50   Median :1010.4
##  Mean   : 1231   Mean   :  57.372   Mean   : 13.54   Mean   :1010.7
##  3rd Qu.: 1500   3rd Qu.:  82.000   3rd Qu.: 23.30   3rd Qu.:1019.0
##  Max.   :10000   Max.   :1071.000   Max.   : 41.60   Max.   :1042.8
##  NA's   :20701   NA's   :13277      NA's   :398      NA's   :393
##       DEWP               wd             WSPM              station
##  Min.   :-43.400   NE     : 43335   Min.   : 0.00   Aotizhongxin: 35064
##  1st Qu.: -8.900   ENE    : 34142   1st Qu.: 0.90   Changping   : 35064
##  Median :  3.100   NW     : 32600   Median : 1.40   Dingling    : 35064
##  Mean   :  2.491   N      : 30869   Mean   : 1.73   Dongsi      : 35064
##  3rd Qu.: 15.100   E      : 29752   3rd Qu.: 2.20   Guanyuan    : 35064
##  Max.   : 29.100   (Other):248248   Max.   :13.20   Gucheng     : 35064
##  NA's   :403       NA's   :  1822   NA's   :318     (Other)     :210384
```

We can verify in the following histogram in this case with the attribute CO

```
# Histogram of Carbon Monoxide
data %>%
  ggplot(aes(CO)) + geom_histogram() +
  labs(title = "China Contamination", x = "Carbon Monoxide")
```