Chicago Data

Samuel Dummer

8/31/2021

Abstract
This document summarizes a dataset about the weather data in Chicago from January 1987 and December 2005. This data includes Temperature (Celsius), Dew point (Celsius), PM10 (m), PM25 (m), and levels of O3 and NO3.

Model Setup

Here setup the script by cleaning up the environment, setting the directory, loading the data frame, and loading in any functions or libraries needed.

```
rm(list=ls())
setwd("C:/Users/isabe/Desktop/RFLoder")
source("myfunctions.R")
library(tidyverse)
library(lubridate)
chicago <- read.csv("chicago.csv", header = T)
```

Characterzation of Data

This part of the script includes the overall characterization of the dataset. This includes names of the columns, dimensions of the data frame, structure, head (a.k.a. first 6 rows of data), summary of the data.

```
names(chicago)
```

```
## [1] "indx"       "city"       "tmpd"       "dptp"       "date"
## [6] "pm25tmean2" "pm10tmean2" "o3tmean2"   "no2tmean2"
```

```
dim(chicago)
```

```
## [1] 6940    9
```

```
str(chicago)
```

```
## 'data.frame':    6940 obs. of  9 variables:
##  $ indx      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ city      : chr  "chic" "chic" "chic" "chic" ...
##  $ tmpd      : num  31.5 33 33 29 32 40 34.5 29 26.5 32.5 ...
##  $ dptp      : num  31.5 29.9 27.4 28.6 28.9 ...
```

```
##  $ date      : chr  "1/1/87" "1/2/87" "1/3/87" "1/4/87" ...
##  $ pm25tmean2: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm10tmean2: num  34 NA 34.2 47 NA ...
##  $ o3tmean2  : num  4.25 3.3 3.33 4.38 4.75 ...
##  $ no2tmean2 : num  20 23.2 23.8 30.4 30.3 ...
```

```
head(chicago)
```

```
##   indx city tmpd   dptp   date pm25tmean2 pm10tmean2 o3tmean2 no2tmean2
## 1    1 chic 31.5 31.500 1/1/87         NA   34.00000 4.250000  19.98810
## 2    2 chic 33.0 29.875 1/2/87         NA         NA 3.304348  23.19099
## 3    3 chic 33.0 27.375 1/3/87         NA   34.16667 3.333333  23.81548
## 4    4 chic 29.0 28.625 1/4/87         NA   47.00000 4.375000  30.43452
## 5    5 chic 32.0 28.875 1/5/87         NA         NA 4.750000  30.33333
## 6    6 chic 40.0 35.125 1/6/87         NA   48.00000 5.833333  25.77233
```

```
summary(chicago)
```

```
##       indx          city                tmpd             dptp
##  Min.   :   1   Length:6940        Min.   :-16.00   Min.   :-25.62
##  1st Qu.:1736   Class :character   1st Qu.: 35.00   1st Qu.: 27.00
##  Median :3470   Mode  :character   Median : 51.00   Median : 39.88
##  Mean   :3470                      Mean   : 50.31   Mean   : 40.34
##  3rd Qu.:5205                      3rd Qu.: 67.00   3rd Qu.: 55.75
##  Max.   :6940                      Max.   : 92.00   Max.   : 78.25
##                                    NA's   :1        NA's   :2
##      date             pm25tmean2      pm10tmean2        o3tmean2
##  Length:6940        Min.   : 1.70   Min.   :  2.00   Min.   : 0.1528
##  Class :character   1st Qu.: 9.70   1st Qu.: 21.50   1st Qu.:10.0729
##  Mode  :character   Median :14.66   Median : 30.28   Median :18.5218
##                     Mean   :16.23   Mean   : 33.90   Mean   :19.4355
##                     3rd Qu.:20.60   3rd Qu.: 42.00   3rd Qu.:27.0010
##                     Max.   :61.50   Max.   :365.00   Max.   :66.5875
##                     NA's   :4447    NA's   :242
##    no2tmean2
##  Min.   : 6.158
##  1st Qu.:19.654
##  Median :24.556
##  Mean   :25.232
##  3rd Qu.:30.139
##  Max.   :62.480
##
```

Removing Columns

Here, we removed some columns that were useless. This includes the index and city column which is unhelpful since all the data is taken in Chicago.

```
chicago <- chicago[(3:9)]
head(chicago)
```

```
##   tmpd   dptp   date pm25tmean2 pm10tmean2 o3tmean2 no2tmean2
## 1 31.5 31.500 1/1/87         NA   34.00000 4.250000  19.98810
## 2 33.0 29.875 1/2/87         NA         NA 3.304348  23.19099
## 3 33.0 27.375 1/3/87         NA   34.16667 3.333333  23.81548
## 4 29.0 28.625 1/4/87         NA   47.00000 4.375000  30.43452
## 5 32.0 28.875 1/5/87         NA         NA 4.750000  30.33333
## 6 40.0 35.125 1/6/87         NA   48.00000 5.833333  25.77233
```

Changing Column Names/Data Munging'

This manipulation helps to create column names are easier to understand to the reader. We changed tmpd to Temp, dptp to Dew point, pm25mean2 to PM25, pm10mean2 to PM10, o3mean2 to O3, and no2mean2 to NO3.

```
names(chicago) <- c("Temp", "Dewpoint", "date", "PM25", "PM10", "O3", "NO3")
head(chicago)
```

```
##   Temp Dewpoint   date PM25     PM10       O3      NO3
## 1 31.5   31.500 1/1/87   NA 34.00000 4.250000 19.98810
## 2 33.0   29.875 1/2/87   NA       NA 3.304348 23.19099
## 3 33.0   27.375 1/3/87   NA 34.16667 3.333333 23.81548
## 4 29.0   28.625 1/4/87   NA 47.00000 4.375000 30.43452
## 5 32.0   28.875 1/5/87   NA       NA 4.750000 30.33333
## 6 40.0   35.125 1/6/87   NA 48.00000 5.833333 25.77233
```

Changing the date

Here, we were asked to change the format of the date. The format was changed from month/day/year to year-month-day.

```
chicago$date <- as.Date(chicago$date , format = "%d/%m/%y")
head(chicago)
```

```
##   Temp Dewpoint       date PM25     PM10       O3      NO3
## 1 31.5   31.500 1987-01-01   NA 34.00000 4.250000 19.98810
## 2 33.0   29.875 1987-02-01   NA       NA 3.304348 23.19099
## 3 33.0   27.375 1987-03-01   NA 34.16667 3.333333 23.81548
## 4 29.0   28.625 1987-04-01   NA 47.00000 4.375000 30.43452
## 5 32.0   28.875 1987-05-01   NA       NA 4.750000 30.33333
## 6 40.0   35.125 1987-06-01   NA 48.00000 5.833333 25.77233
```

Normalizing the Data

To help further clean the data, we normalized the PM25 and PM10. This mean't we divided all the data by the largest value to help keep all the values between 0 and 1.

```
chicago$PM10 <- chicago$PM10/max(chicago$PM10, na.rm = T)
chicago$PM25 <- chicago$PM25/max(chicago$PM25, na.rm = T)
head(chicago)
```

```
##   Temp Dewpoint       date PM25       PM10       O3      NO3
## 1 31.5   31.500 1987-01-01   NA 0.09315068 4.250000 19.98810
## 2 33.0   29.875 1987-02-01   NA         NA 3.304348 23.19099
## 3 33.0   27.375 1987-03-01   NA 0.09360731 3.333333 23.81548
## 4 29.0   28.625 1987-04-01   NA 0.12876712 4.375000 30.43452
## 5 32.0   28.875 1987-05-01   NA         NA 4.750000 30.33333
## 6 40.0   35.125 1987-06-01   NA 0.13150685 5.833333 25.77233
```

Pairwise correlation/Histogram Plot

Finally, we were asked to create a pairwise correlation plot using all the remaining numeric data. Using a built-in funtion from "my.function.R" we created a plot that shows the correlation, histograms, and scatter plots of all the numeric data.

```
numval <- chicago[c(1:2, 4:7)]
view(numval)
pairs(numval, upper.panel = panel.cor, diag.panel = panel.hist)
```