Chicago weather data manipulation

Andrew Wang
North Carolina School of Science and Mathematics

9/6/2020

Model Setup

In this section, we established the basic parameters of the model, including cleaning up old data and setting up the working directory. Furthermore, we are setting up the `myfunctions.R` file so we can reference it for a pair plot.

```
#data cleanup and setup
rm(list=ls())
setwd("C:/Users/hyper/OneDrive/Desktop/Desktop Folders/Programming/R")
source("C:/Users/hyper/OneDrive/Desktop/Desktop Folders/Programming/R/myfunctions.R")
```

Loading Libraries

We are using the lubridate package for date formatting. So we'll need to install and set up the package. Remember that you only need to install the package once.

```
#install the lubridate package
install.packages("lubridate", repos="http://cran.rstudio.com")
```

```
## Warning: package 'lubridate' is in use and will not be installed
```

```
#set up lubridate as library
library(lubridate)
```

Read the CSV file

Here, we're going to read the CSV file so that we can create a dataframe from it. Make sure your file is in the location of your working directory (defined above)

```
#read a csv file
weather <- read.csv("chicago.csv")
```

Data Demographics

We want to print out some basic information about the data. However, the `View(weather)` command here does not output the table, but it will in a normal R script. The following code will show the class, the dimensions, and the structure.

```
#opens up the data (in an R script)
View(weather)
#shows the class of the data
class(weather)
```

```
## [1] "data.frame"
```

```
#shows the dimensions of the weather
dim(weather)
```

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

```
#gives a structural overview of the data
str(weather)
```

```
## '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 ...
```

Data Munging

We are deleting the `index` and `city` columns, as they are not needed.

```
#delete index column
weather$indx <- NULL
#delete city column
weather$city <- NULL
```

Change Column Names

We want to change the column names to something more appropriate. Here we refer to all the columns together instead of doing it separately because there's only one unchanged column name.

```
#output column names
names(weather)
```

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

```
#change column names
names(weather) <- c("Temp","DewPoint","Date","PM25","PM10","O3","NO3")
#output column names again
names(weather)
```

```
## [1] "Temp"     "DewPoint" "Date"     "PM25"     "PM10"     "O3"       "NO3"
```

Change the Date Format

We want to change the date format using the `lubridate` package. Here, we change the date format from `m/d/yy` to `yyyy-mm-dd`

```
#change the data format
weather$Date <- mdy(weather$Date)
```

Data Normalization

We want to change the N/A data values to be a value on a 0 to 1 scale based on the maximum value for each column. Essentially we are setting a percentage. However, before you do that you need to replace the N/A data points, so we're replacing them with the mean values of the column (calculated without the empty data points).

```
#normalization of the PM25 column
weather$PM25 <- ifelse(is.na(weather$PM25),mean(weather$PM25,na.rm=TRUE),weather$PM25)
weather$PM25 <- weather$PM25/max(weather$PM25)
#normalization of the PM10 column
weather$PM10 <- ifelse(is.na(weather$PM10),mean(weather$PM10,na.rm=TRUE),weather$PM10)
weather$PM10 <- weather$PM10/max(weather$PM10)
```

Pairwise Plot

We make a pair plot for correlation and histogram numeric data. First, we create a subset of the data for the numeric columns, and we make sure to include all rows. After this we call the pair function for the numeric data subset and use the `panel.cor` and `panel.hist` functions from `myfunctions.R`.

```
# creates a data subset with numeric columns
numerics <- weather[,c(1,2,4:7)]
# displays data subset as pair plot
pairs(numerics, upper.panel = panel.cor, diag.panel = panel.hist)
```

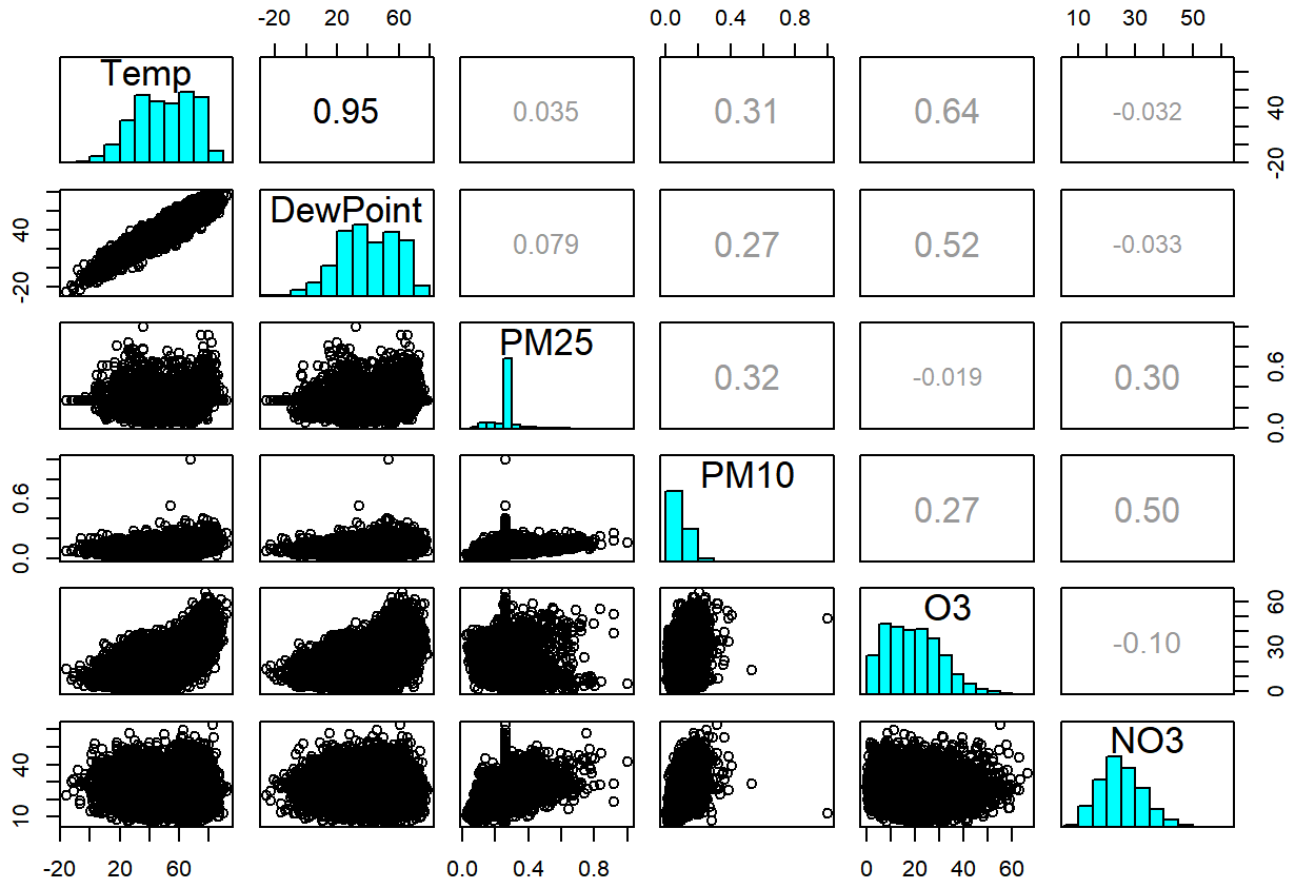

Table of weather data

Below is the first 50 lines of data from the dataframe. There are close to 7,000 rows, so we're only going to display the first 50 rows.

Chicago weather data (first 50 lines)

Temp	DewPoint	Date	PM25	PM10	O3	NO3
31.5	31.500	1987-01-01	0.263918	0.0931507	4.250000	19.98810
33.0	29.875	1987-01-02	0.263918	0.0928636	3.304348	23.19099
33.0	27.375	1987-01-03	0.263918	0.0936073	3.333333	23.81548
29.0	28.625	1987-01-04	0.263918	0.1287671	4.375000	30.43452
32.0	28.875	1987-01-05	0.263918	0.0928636	4.750000	30.33333

Temp	DewPoint	Date	PM25	PM10	O3	NO3
40.0	35.125	1987-01-06	0.263918	0.1315068	5.833333	25.77233
34.5	26.750	1987-01-07	0.263918	0.1123288	9.291667	20.58171
29.0	22.000	1987-01-08	0.263918	0.0986301	11.291667	17.03723
26.5	29.000	1987-01-09	0.263918	0.0911937	4.500000	23.38889
32.5	27.750	1987-01-10	0.263918	0.0928636	4.958333	19.54167
29.5	20.125	1987-01-11	0.263918	0.0602740	17.541667	13.70139
34.5	26.000	1987-01-12	0.263918	0.0712329	8.000000	33.02083
34.0	32.250	1987-01-13	0.263918	0.1452055	4.958333	38.06142
37.5	36.375	1987-01-14	0.263918	0.1178082	4.208333	32.19444
32.5	24.250	1987-01-15	0.263918	0.0789954	4.458333	18.87131
25.0	21.500	1987-01-16	0.263918	0.0520548	7.916667	19.46667
27.0	24.750	1987-01-17	0.263918	0.0928636	5.833333	20.70833
17.5	11.125	1987-01-18	0.263918	0.1068493	6.375000	21.03333
23.0	15.750	1987-01-19	0.263918	0.0876712	14.875000	17.17409
20.5	11.500	1987-01-20	0.263918	0.1041096	7.250000	21.61021
22.0	20.625	1987-01-21	0.263918	0.0900196	8.913044	24.52083
19.5	7.375	1987-01-22	0.263918	0.1424658	10.500000	16.98798
2.5	-12.250	1987-01-23	0.263918	0.1506849	14.625000	14.66250
2.0	-5.625	1987-01-24	0.263918	0.1041096	10.083333	18.69167
9.5	-5.250	1987-01-25	0.263918	0.0928636	6.666667	26.30417
16.0	4.750	1987-01-26	0.263918	0.1945205	4.583333	32.42143
17.5	17.750	1987-01-27	0.263918	0.1077626	6.000000	30.69306
29.5	18.250	1987-01-28	0.263918	0.1287671	6.875000	29.12943
29.5	32.875	1987-01-29	0.263918	0.0958904	2.916667	28.14529
32.5	24.125	1987-01-30	0.263918	0.1616438	8.791667	19.79861
27.5	26.500	1987-01-31	0.263918	0.0986301	10.375000	25.26736
41.0	32.250	1987-02-01	0.263918	0.0928636	8.041667	21.70139

Temp	DewPoint	Date	PM25	PM10	O3	NO3
36.5	34.000	1987-02-02	0.263918	0.1127854	6.041667	30.78472
34.0	26.250	1987-02-03	0.263918	0.1616438	8.041667	20.14615
31.5	24.250	1987-02-04	0.263918	0.1013699	10.500000	29.99928
29.5	27.250	1987-02-05	0.263918	0.1917808	5.729167	38.90909
37.0	30.375	1987-02-06	0.263918	0.1726027	6.414855	27.99471
40.5	35.750	1987-02-07	0.263918	0.1205479	5.750000	25.09028
32.5	3.500	1987-02-08	0.263918	0.0500000	24.937500	10.64683
24.5	17.375	1987-02-09	0.263918	0.1150685	9.208333	25.57986
35.0	27.375	1987-02-10	0.263918	0.0986301	8.312500	31.32746
39.5	29.375	1987-02-11	0.263918	0.2273973	8.121541	42.00113
34.0	24.750	1987-02-12	0.263918	0.2301370	7.687500	34.99306
34.5	30.875	1987-02-13	0.263918	0.1068493	10.369565	33.43086
31.5	22.000	1987-02-14	0.263918	0.0681018	11.166667	21.24306
23.5	10.125	1987-02-15	0.263918	0.0328767	28.104167	10.49097
24.0	18.125	1987-02-16	0.263918	0.0383562	24.187500	15.85417
31.5	20.250	1987-02-17	0.263918	0.0630137	24.937500	19.18056
31.5	20.000	1987-02-18	0.263918	0.0739726	24.170290	29.86012
34.0	19.625	1987-02-19	0.263918	0.1342466	15.501812	44.48068