Week 1 Data Prepration

Kevin McBeth

January 22, 2018

Introduction

mean imputed (y) vs KNN-imputed (x) plot to finish. NO2 column

```
library(data.table)
library(DMwR)
library(corrplot)
library(raster)
```

```
fn <- "C:\\Users\\Kevin\\Desktop\\Workspaces\\R\\MSDS680\\Data\\AirQualityUCI.csv"

df <- fread(fn)
airQuality <- as.data.table(df)
```

Basic Exploratory Data Analysis

Basic exploratory data analysis, including looking at the structure, summary statistics, and the partiuclar 'missing' values of -9 are located.

```
str(airQuality)
```

```
## Classes 'data.table' and 'data.frame':   9471 obs. of  17 variables:
##  $ Date         : chr  "10/03/2004" "10/03/2004" "10/03/2004" "10/03/2004" ...
##  $ Time         : chr  "18.00.00" "19.00.00" "20.00.00" "21.00.00" ...
##  $ CO(GT)       : chr  "2,6" "2" "2,2" "2,2" ...
##  $ PT08.S1(CO)  : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
##  $ NMHC(GT)     : int  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6(GT)     : chr  "11,9" "9,4" "9,0" "9,2" ...
##  $ PT08.S2(NMHC): int  1046 955 939 948 836 750 690 672 609 561 ...
##  $ NOx(GT)      : int  166 103 131 172 131 89 62 62 45 -200 ...
##  $ PT08.S3(NOx) : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2(GT)      : int  113 92 114 122 116 96 77 76 60 -200 ...
##  $ PT08.S4(NO2) : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5(O3)  : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ T            : chr  "13,6" "13,3" "11,9" "11,0" ...
##  $ RH           : chr  "48,9" "47,7" "54,0" "60,0" ...
##  $ AH           : chr  "0,7578" "0,7255" "0,7502" "0,7867" ...
##  $ V16          : logi  NA NA NA NA NA NA ...
##  $ V17          : logi  NA NA NA NA NA NA ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(airQuality)
```

```
##      Date               Time              CO(GT)           PT08.S1(CO)
##  Length:9471        Length:9471        Length:9471        Min.   :-200
##  Class :character   Class :character   Class :character   1st Qu.: 921
##  Mode  :character   Mode  :character   Mode  :character   Median :1053
##                                                           Mean   :1049
```

```
##                                                          3rd Qu.:1221
##                                                          Max.   :2040
##                                                          NA's   :114
##     NMHC(GT)        C6H6(GT)         PT08.S2(NMHC)       NOx(GT)
##  Min.   :-200.0   Length:9471        Min.   :-200.0   Min.   :-200.0
##  1st Qu.:-200.0   Class :character   1st Qu.: 711.0   1st Qu.:  50.0
##  Median :-200.0   Mode  :character   Median : 895.0   Median : 141.0
##  Mean   :-159.1                      Mean   : 894.6   Mean   : 168.6
##  3rd Qu.:-200.0                      3rd Qu.:1105.0   3rd Qu.: 284.0
##  Max.   :1189.0                      Max.   :2214.0   Max.   :1479.0
##  NA's   :114                         NA's   :114      NA's   :114
##   PT08.S3(NOx)     NO2(GT)         PT08.S4(NO2)   PT08.S5(O3)
##  Min.   :-200   Min.   :-200.00   Min.   :-200   Min.   :-200.0
##  1st Qu.: 637   1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 700.0
##  Median : 794   Median :  96.00   Median :1446   Median : 942.0
##  Mean   : 795   Mean   :  58.15   Mean   :1391   Mean   : 975.1
##  3rd Qu.: 960   3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.0
##  Max.   :2683   Max.   : 340.00   Max.   :2775   Max.   :2523.0
##  NA's   :114    NA's   :114       NA's   :114    NA's   :114
##       T                RH                AH              V16
##  Length:9471        Length:9471        Length:9471        Mode:logical
##  Class :character   Class :character   Class :character   NA's:9471
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    V17
##  Mode:logical
##  NA's:9471
##
##
##
##
##
```

Data Cleaning

Cleaning the data by removing the dm column due to too many missing data portions, setting all -9 values to NA, looking at the summary statistics, imputing the data with missing values, and then changing to numeric to run corrplot.

```
airQuality[, Time:=NULL]
airQuality[, Date:=NULL]
airQuality[,V16:=NULL]
airQuality[,V17:=NULL]
str(airQuality)
```

```
## Classes 'data.table' and 'data.frame':   9471 obs. of  13 variables:
##  $ CO(GT)       : chr  "2,6" "2" "2,2" "2,2" ...
##  $ PT08.S1(CO)  : int  1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...
##  $ NMHC(GT)     : int  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6(GT)     : chr  "11,9" "9,4" "9,0" "9,2" ...
##  $ PT08.S2(NMHC): int  1046 955 939 948 836 750 690 672 609 561 ...
```

```
##  $ NOx(GT)      : int  166 103 131 172 131 89 62 62 45 -200 ...
##  $ PT08.S3(NOx) : int  1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...
##  $ NO2(GT)      : int  113 92 114 122 116 96 77 76 60 -200 ...
##  $ PT08.S4(NO2) : int  1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...
##  $ PT08.S5(O3)  : int  1268 972 1074 1203 1110 949 733 730 620 501 ...
##  $ T            : chr  "13,6" "13,3" "11,9" "11,0" ...
##  $ RH           : chr  "48,9" "47,7" "54,0" "60,0" ...
##  $ AH           : chr  "0,7578" "0,7255" "0,7502" "0,7867" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
removeCommas <- function(x) gsub(',', '.', x)
airQuality <- airQuality[,lapply(.SD, removeCommas)]
airQuality <- airQuality[,lapply(.SD, as.numeric)]


str(airQuality)
```

```
## Classes 'data.table' and 'data.frame':   9471 obs. of  13 variables:
##  $ CO(GT)       : num  2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...
##  $ PT08.S1(CO)  : num  1360 1292 1402 1376 1272 ...
##  $ NMHC(GT)     : num  150 112 88 80 51 38 31 31 24 19 ...
##  $ C6H6(GT)     : num  11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.7 ...
##  $ PT08.S2(NMHC): num  1046 955 939 948 836 ...
##  $ NOx(GT)      : num  166 103 131 172 131 89 62 62 45 -200 ...
##  $ PT08.S3(NOx) : num  1056 1174 1140 1092 1205 ...
##  $ NO2(GT)      : num  113 92 114 122 116 96 77 76 60 -200 ...
##  $ PT08.S4(NO2) : num  1692 1559 1555 1584 1490 ...
##  $ PT08.S5(O3)  : num  1268 972 1074 1203 1110 ...
##  $ T            : num  13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 10.3 ...
##  $ RH           : num  48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 60.2 ...
##  $ AH           : num  0.758 0.726 0.75 0.787 0.789 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
airQualityKNN <- airQuality

airQualityNO2mean = mean(airQuality$`NO2(GT)`, na.rm=T)
airQualityMeanImputed <- airQuality
sum(is.na(airQuality$`NO2(GT)`))
```

```
## [1] 114
```

```
airQualityMeanImputed[is.na(airQuality$`NO2(GT)`), ]$`NO2(GT)` <-airQualityNO2mean

summary(airQualityMeanImputed$`NO2(GT)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -200.00   54.00   95.00   58.15  132.00  340.00
```

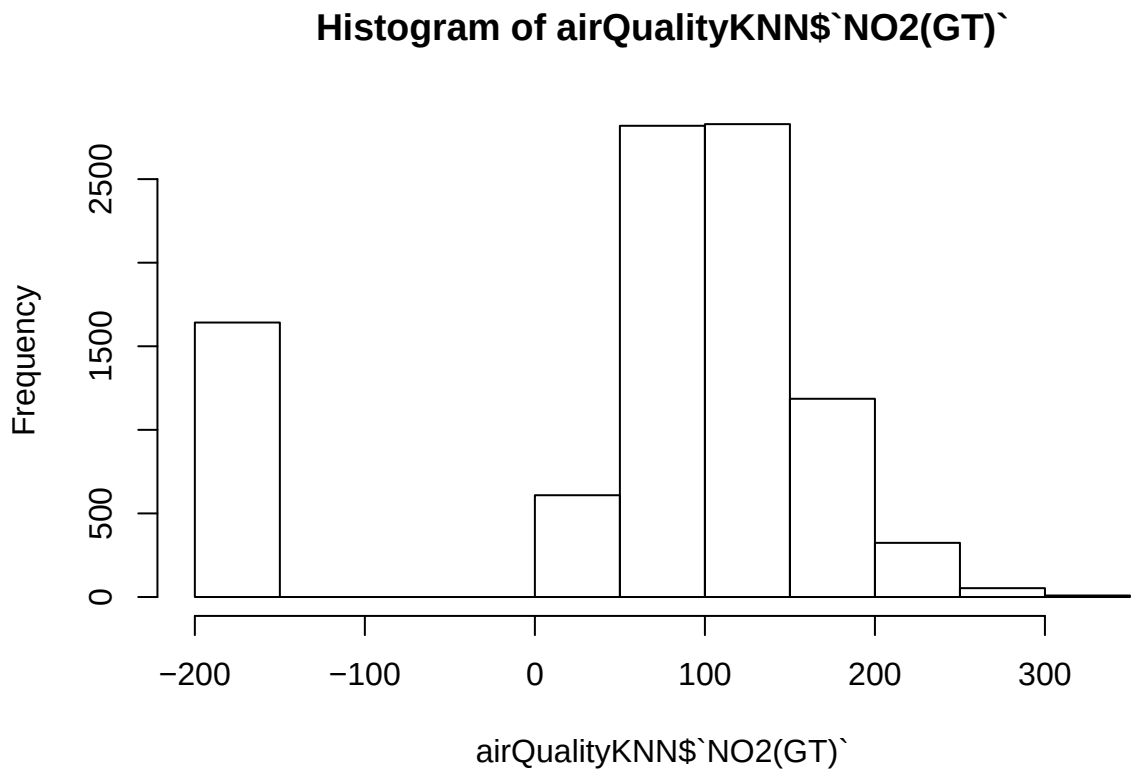
```
airQualityKNN <- knnImputation(airQualityKNN)

summary(airQualityKNN)
```

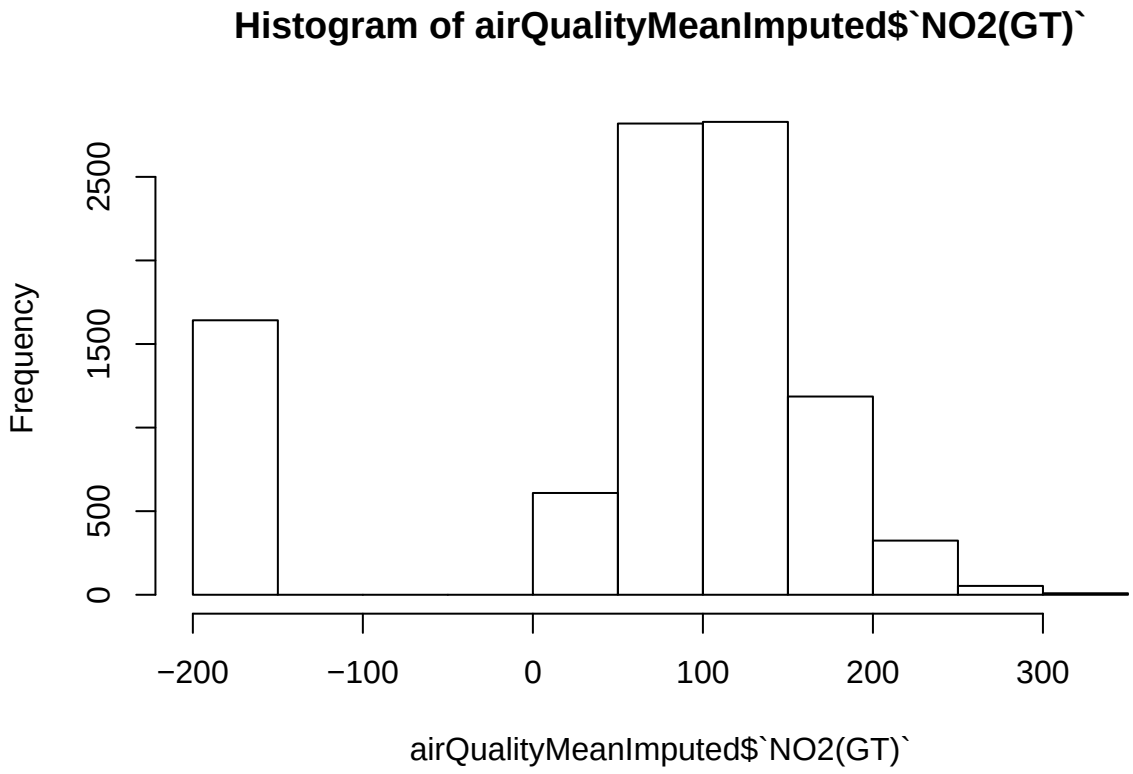
```
##      CO(GT)          PT08.S1(CO)      NMHC(GT)          C6H6(GT)
##  Min.   :-200.00   Min.   :-200   Min.   :-200.0   Min.   :-200.000
##  1st Qu.:   0.60   1st Qu.: 923   1st Qu.:-200.0   1st Qu.:   4.000
##  Median :   1.50   Median :1055   Median :-200.0   Median :   7.800
##  Mean   : -33.78   Mean   :1051   Mean   :-156.4   Mean   :   1.917
```

```
##  3rd Qu.:   2.60   3rd Qu.:1230   3rd Qu.:-200.0   3rd Qu.:  13.500
##  Max.   :  11.90   Max.   :2040   Max.   :1189.0   Max.   :  63.700
##  PT08.S2(NMHC)       NOx(GT)        PT08.S3(NOx)       NO2(GT)
##  Min.   :-200.0   Min.   :-200.0   Min.   :-200.0   Min.   :-200.00
##  1st Qu.: 713.0   1st Qu.:  51.0   1st Qu.: 639.0   1st Qu.:  54.00
##  Median : 890.0   Median : 139.0   Median : 798.0   Median :  95.00
##  Mean   : 893.5   Mean   : 167.5   Mean   : 801.3   Mean   :  58.25
##  3rd Qu.:1102.0   3rd Qu.: 281.5   3rd Qu.: 969.0   3rd Qu.: 132.00
##  Max.   :2214.0   Max.   :1479.0   Max.   :2683.0   Max.   : 340.00
##   PT08.S4(NO2)   PT08.S5(O3)           T                RH
##  Min.   :-200   Min.   :-200.0   Min.   :-200.000   Min.   :-200.00
##  1st Qu.:1189   1st Qu.: 703.0   1st Qu.:  11.000   1st Qu.:  34.25
##  Median :1445   Median : 936.0   Median :  17.000   Median :  48.90
##  Mean   :1392   Mean   : 974.4   Mean   :   9.799   Mean   :  39.69
##  3rd Qu.:1658   3rd Qu.:1250.0   3rd Qu.:  24.000   3rd Qu.:  61.70
##  Max.   :2775   Max.   :2523.0   Max.   :  44.600   Max.   :  88.70
##        AH
##  Min.   :-200.0000
##  1st Qu.:   0.6967
##  Median :   0.9711
##  Mean   :  -6.7461
##  3rd Qu.:   1.2915
##  Max.   :   2.2310
```

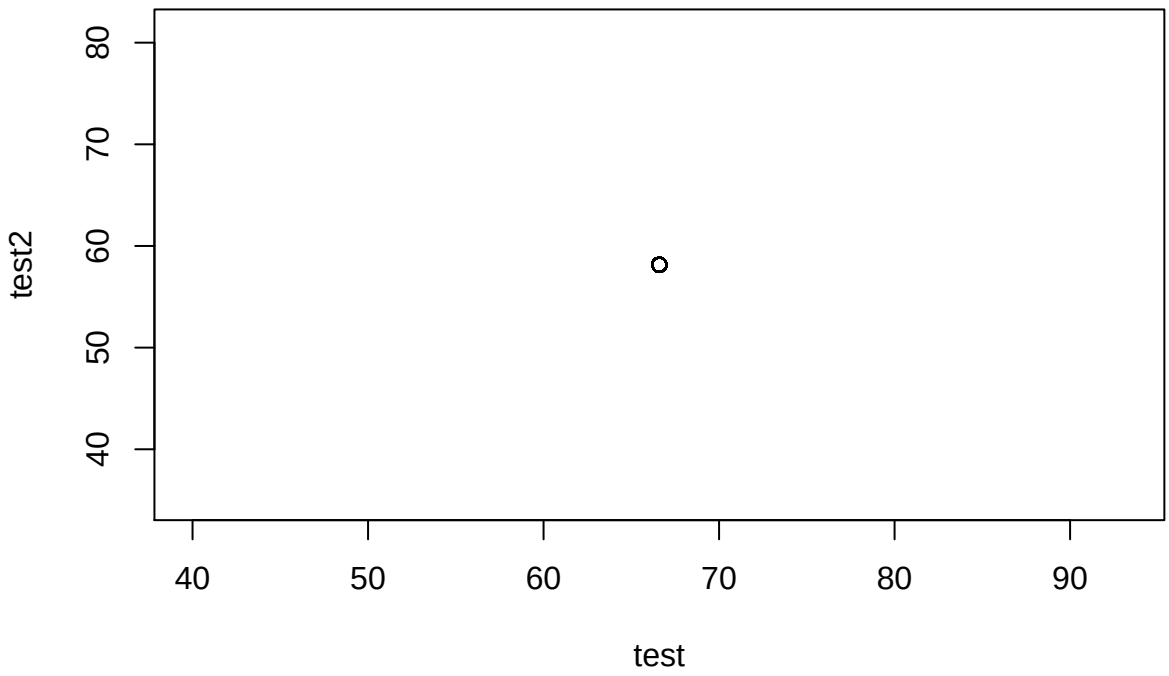
```
hist(airQualityKNN$`NO2(GT)`)
```



```r
hist(airQualityMeanImputed$`NO2(GT)`)
```




```r
test <- airQualityKNN[is.na(airQuality$`NO2(GT)`)]$`NO2(GT)`
test2 <- airQualityMeanImputed[is.na(airQuality$`NO2(GT)`)]$`NO2(GT)`
plot(test, test2)
```


test2
40
50
60
70
80
40
50
60
70
80
90
test