Data Wrangling

Divya Krishnan

Wednesday, October 14, 2015

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `dataWrangling.Rmd` file from Canvas. Open `dataWrangling.Rmd` in RStudio and supply your solutions to the assignment by editing `dataWrangling.Rmd`. You will also want to download the `weather.txt` data file, containing a dataset capturing daily temperatures in Cuernavaca, Mexico during 2010.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
3. Be sure to include code chucks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.
4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit HTML`, rename the R Markdown file to `YourLastName_YourFirstName.Rmd`, and submit on Canvas.

```
# Stardard libraries
library(plyr)
library(dplyr)
library(ggplot2)
library(reshape2)
library(babynames)
```

Data Cleaning In this problem we will use the `weather.txt` data. Import the data in **R** and answer the following questions.

```
# Importing the weather dataset
weather<-read.delim(
  file = "weather.txt",
  header = TRUE,
  check.names = F
)
```

```
# Variables in the dataset
colnames(weather)
```

(a) What are the variables in this dataset? Describe what each variable measures.

```
##  [1] "id"      "year"    "month"   "element" "d1"      "d2"      "d3"
##  [8] "d4"      "d5"      "d6"      "d7"      "d8"      "d9"      "d10"
## [15] "d11"     "d12"     "d13"     "d14"     "d15"     "d16"     "d17"
## [22] "d18"     "d19"     "d20"     "d21"     "d22"     "d23"     "d24"
## [29] "d25"     "d26"     "d27"     "d28"     "d29"     "d30"     "d31"
```

```
summary(weather)
```

```
##            id          year          month        element
##  MX000017004:22   Min.   :2010   Min.   : 1.000   TMAX:11
##                   1st Qu.:2010   1st Qu.: 3.250   TMIN:11
##                   Median :2010   Median : 6.000
##                   Mean   :2010   Mean   : 6.273
##                   3rd Qu.:2010   3rd Qu.: 9.500
##                   Max.   :2010   Max.   :12.000
##
##        d1              d2              d3              d4
##  Min.   :138.0   Min.   :144.0   Min.   :144.0   Min.   :120
##  1st Qu.:178.2   1st Qu.:158.2   1st Qu.:167.2   1st Qu.:158
##  Median :218.5   Median :218.0   Median :208.0   Median :196
##  Mean   :218.5   Mean   :223.2   Mean   :211.5   Mean   :196
##  3rd Qu.:258.8   3rd Qu.:283.0   3rd Qu.:252.2   3rd Qu.:234
##  Max.   :299.0   Max.   :313.0   Max.   :286.0   Max.   :272
##  NA's   :20      NA's   :18      NA's   :18      NA's   :20
##        d5              d6              d7              d8
##  Min.   : 79.0   Min.   :105.0   Min.   :129   Min.   :173.0
##  1st Qu.:141.5   1st Qu.:148.2   1st Qu.:167   1st Qu.:202.2
##  Median :210.5   Median :191.5   Median :205   Median :231.5
##  Mean   :208.6   Mean   :191.5   Mean   :205   Mean   :231.5
##  3rd Qu.:276.5   3rd Qu.:234.8   3rd Qu.:243   3rd Qu.:260.8
##  Max.   :321.0   Max.   :278.0   Max.   :281   Max.   :290.0
##  NA's   :14      NA's   :20      NA's   :20    NA's   :20
##     d9              d10             d11           d12
##  Mode:logical   Min.   :168.0   Min.   :134.0   Mode:logical
##  NA's:22        1st Qu.:212.2   1st Qu.:174.8   NA's:22
##                 Median :256.5   Median :215.5
##                 Mean   :256.5   Mean   :215.5
##                 3rd Qu.:300.8   3rd Qu.:256.2
##                 Max.   :345.0   Max.   :297.0
##                 NA's   :20      NA's   :20
##       d13             d14             d15             d16
##  Min.   :165.0   Min.   :130.0   Min.   :105.0   Min.   :176.0
##  1st Qu.:198.2   1st Qu.:156.2   1st Qu.:150.5   1st Qu.:209.8
##  Median :231.5   Median :230.0   Median :196.0   Median :243.5
##  Mean   :231.5   Mean   :222.2   Mean   :196.0   Mean   :243.5
##  3rd Qu.:264.8   3rd Qu.:296.0   3rd Qu.:241.5   3rd Qu.:277.2
##  Max.   :298.0   Max.   :299.0   Max.   :287.0   Max.   :311.0
##  NA's   :20      NA's   :18      NA's   :20      NA's   :20
##       d17           d18             d19             d20
##  Min.   :175.0   Mode:logical   Mode:logical   Mode:logical
##  1st Qu.:201.2   NA's:22        NA's:22        NA's:22
##  Median :227.5
##  Mean   :227.5
##  3rd Qu.:253.8
##  Max.   :280.0
##  NA's   :20
##    d21             d22             d23           d24
##  Mode:logical   Mode:logical   Min.   :107.0   Mode:logical
##  NA's:22        NA's:22        1st Qu.:139.2   NA's:22
```

```
##                                Median :207.0
##                                Mean   :205.0
##                                3rd Qu.:272.8
##                                Max.   :299.0
##                                NA's   :18
##       d25             d26           d27             d28
##  Min.   :156.0   Min.   :121   Min.   :142.0   Min.   :150.0
##  1st Qu.:191.2   1st Qu.:161   1st Qu.:170.8   1st Qu.:190.5
##  Median :226.5   Median :201   Median :229.5   Median :231.0
##  Mean   :226.5   Mean   :201   Mean   :243.8   Mean   :231.0
##  3rd Qu.:261.8   3rd Qu.:241   3rd Qu.:318.2   3rd Qu.:271.5
##  Max.   :297.0   Max.   :281   Max.   :363.0   Max.   :312.0
##  NA's   :20      NA's   :20    NA's   :16      NA's   :20
##       d29             d30             d31
##  Min.   :153.0   Min.   :145.0   Min.   :154
##  1st Qu.:173.2   1st Qu.:178.2   1st Qu.:179
##  Median :230.0   Median :211.5   Median :204
##  Mean   :228.5   Mean   :211.5   Mean   :204
##  3rd Qu.:285.2   3rd Qu.:244.8   3rd Qu.:229
##  Max.   :301.0   Max.   :278.0   Max.   :254
##  NA's   :18      NA's   :20      NA's   :20
```

```
#View(weather)
```

(b) Tidy up the weather data such that each observation forms a row and each variable forms a column. You might find the following functions helpful:

- melt
- mutate
- dcast

```
# Tidying weather data using melt function
weather.tidy <- melt(
  data = weather,
  id = c("id","year","month","element"),
  variable.name=c("day"),
  value.name = "temperature"
)
#View(weather.tidy)
# Removing d prefix for day column
weather.tidy$day<-sub("d","",weather.tidy$day)
#Removing NAs
weather.tidy<-weather.tidy[complete.cases(weather.tidy[,6:6]),]

head(weather.tidy)
```

```
##             id year month element day temperature
## 21 MX000017004 2010    12    TMAX   1         299
## 22 MX000017004 2010    12    TMIN   1         138
## 25 MX000017004 2010     2    TMAX   2         273
## 26 MX000017004 2010     2    TMIN   2         144
## 41 MX000017004 2010    11    TMAX   2         313
## 42 MX000017004 2010    11    TMIN   2         163
```