INFX 573 Lab: Data Wrangling

Pierre Augustamar

October 13th, 2016

Collaborators: Terry Kim

Don't forget to list the full names of your collaborators!

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio.

1. Download the `week3a_lab.Rmd` file from Canvas. Open `week3a_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week3a_lab.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`, rename the R Markdown file to `YourLastName_YourFirstName_lab3a.Rmd`, and knit it into a PDF. Submit the compiled PDF on Canvas.

   In this lab, you will need access to the following R packages:

```
# Load some helpful libraries
library(tidyverse)
library(babynames)
```

Problem 1: Data Cleaning

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

Hint: You might find the function `read.table()` useful here.

```
# import weather data
read.table("weather.txt", stringsAsFactors = FALSE,
    header = TRUE)
```

```
##             id year month element  d1  d2
## 1  MX000017004 2010     1    TMAX  NA  NA
## 2  MX000017004 2010     1    TMIN  NA  NA
## 3  MX000017004 2010     2    TMAX  NA 273
## 4  MX000017004 2010     2    TMIN  NA 144
## 5  MX000017004 2010     3    TMAX  NA  NA
## 6  MX000017004 2010     3    TMIN  NA  NA
## 7  MX000017004 2010     4    TMAX  NA  NA
## 8  MX000017004 2010     4    TMIN  NA  NA
## 9  MX000017004 2010     5    TMAX  NA  NA
## 10 MX000017004 2010     5    TMIN  NA  NA
## 11 MX000017004 2010     6    TMAX  NA  NA
## 12 MX000017004 2010     6    TMIN  NA  NA
## 13 MX000017004 2010     7    TMAX  NA  NA
## 14 MX000017004 2010     7    TMIN  NA  NA
## 15 MX000017004 2010     8    TMAX  NA  NA
## 16 MX000017004 2010     8    TMIN  NA  NA
## 17 MX000017004 2010    10    TMAX  NA  NA
## 18 MX000017004 2010    10    TMIN  NA  NA
## 19 MX000017004 2010    11    TMAX  NA 313
## 20 MX000017004 2010    11    TMIN  NA 163
## 21 MX000017004 2010    12    TMAX 299  NA
## 22 MX000017004 2010    12    TMIN 138  NA
##     d3  d4  d5  d6  d7  d8 d9 d10 d11 d12
## 1   NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 2   NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 3  241  NA  NA  NA  NA  NA NA  NA 297  NA
## 4  144  NA  NA  NA  NA  NA NA  NA 134  NA
## 5   NA  NA 321  NA  NA  NA NA 345  NA  NA
## 6   NA  NA 142  NA  NA  NA NA 168  NA  NA
## 7   NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 8   NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 9   NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 10  NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 11  NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 12  NA  NA  NA  NA  NA  NA NA  NA  NA  NA
## 13 286  NA  NA  NA  NA  NA NA  NA  NA  NA
## 14 175  NA  NA  NA  NA  NA NA  NA  NA  NA
## 15  NA  NA 296  NA  NA 290 NA  NA  NA  NA
## 16  NA  NA 158  NA  NA 173 NA  NA  NA  NA
## 17  NA  NA 270  NA 281  NA NA  NA  NA  NA
## 18  NA  NA 140  NA 129  NA NA  NA  NA  NA
## 19  NA 272 263  NA  NA  NA NA  NA  NA  NA
## 20  NA 120  79  NA  NA  NA NA  NA  NA  NA
## 21  NA  NA  NA 278  NA  NA NA  NA  NA  NA
## 22  NA  NA  NA 105  NA  NA NA  NA  NA  NA
##    d13 d14 d15 d16 d17 d18 d19 d20 d21 d22
## 1   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 2   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 3   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 4   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 5   NA  NA  NA 311  NA  NA  NA  NA  NA  NA
## 6   NA  NA  NA 176  NA  NA  NA  NA  NA  NA
## 7   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 8   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 9   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 10  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 11  NA  NA  NA  NA 280  NA  NA  NA  NA  NA
## 12  NA  NA  NA  NA 175  NA  NA  NA  NA  NA
## 13  NA 299  NA  NA  NA  NA  NA  NA  NA  NA
## 14  NA 165  NA  NA  NA  NA  NA  NA  NA  NA
## 15 298  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 16 165  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 17  NA 295 287  NA  NA  NA  NA  NA  NA  NA
## 18  NA 130 105  NA  NA  NA  NA  NA  NA  NA
## 19  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 20  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 21  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 22  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
##    d23 d24 d25 d26 d27 d28 d29 d30 d31
## 1   NA  NA  NA  NA  NA  NA  NA 278  NA
## 2   NA  NA  NA  NA  NA  NA  NA 145  NA
## 3  299  NA  NA  NA  NA  NA  NA  NA  NA
## 4  107  NA  NA  NA  NA  NA  NA  NA  NA
## 5   NA  NA  NA  NA  NA  NA  NA  NA  NA
## 6   NA  NA  NA  NA  NA  NA  NA  NA  NA
## 7   NA  NA  NA  NA 363  NA  NA  NA  NA
## 8   NA  NA  NA  NA 167  NA  NA  NA  NA
## 9   NA  NA  NA  NA 332  NA  NA  NA  NA
## 10  NA  NA  NA  NA 182  NA  NA  NA  NA
## 11  NA  NA  NA  NA  NA  NA 301  NA  NA
## 12  NA  NA  NA  NA  NA  NA 180  NA  NA
## 13  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 14  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 15 264  NA 297  NA  NA  NA 280  NA 254
## 16 150  NA 156  NA  NA  NA 153  NA 154
## 17  NA  NA  NA  NA  NA 312  NA  NA  NA
## 18  NA  NA  NA  NA  NA 150  NA  NA  NA
## 19  NA  NA  NA 281 277  NA  NA  NA  NA
## 20  NA  NA  NA 121 142  NA  NA  NA  NA
## 21  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 22  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

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

```
# variables from the weather.txt dataset
str(read.table("weather.txt", stringsAsFactors = FALSE,
    header = TRUE))
```

```
## ’data.frame’:    22 obs. of  35 variables:
##  $ id     : chr  "MX000017004" "MX000017004" "MX000017004" "MX000017004" ...
##  $ year   : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ month  : int  1 1 2 2 3 3 4 4 5 5 ...
##  $ element: chr  "TMAX" "TMIN" "TMAX" "TMIN" ...
##  $ d1     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d2     : int  NA NA 273 144 NA NA NA NA NA NA ...
##  $ d3     : int  NA NA 241 144 NA NA NA NA NA NA ...
##  $ d4     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d5     : int  NA NA NA NA 321 142 NA NA NA NA ...
##  $ d6     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d7     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d8     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d9     : logi  NA NA NA NA NA NA ...
##  $ d10    : int  NA NA NA NA 345 168 NA NA NA NA ...
##  $ d11    : int  NA NA 297 134 NA NA NA NA NA NA ...
##  $ d12    : logi  NA NA NA NA NA NA ...
##  $ d13    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d14    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d15    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d16    : int  NA NA NA NA 311 176 NA NA NA NA ...
##  $ d17    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d18    : logi  NA NA NA NA NA NA ...
##  $ d19    : logi  NA NA NA NA NA NA ...
##  $ d20    : logi  NA NA NA NA NA NA ...
##  $ d21    : logi  NA NA NA NA NA NA ...
##  $ d22    : logi  NA NA NA NA NA NA ...
##  $ d23    : int  NA NA 299 107 NA NA NA NA NA NA ...
##  $ d24    : logi  NA NA NA NA NA NA ...
##  $ d25    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d26    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d27    : int  NA NA NA NA NA NA 363 167 332 182 ...
##  $ d28    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d29    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ d30    : int  278 145 NA NA NA NA NA NA NA NA ...
##  $ d31    : int  NA NA NA NA NA NA NA NA NA NA ...
```

Hint: There are five variables of interest here.

There are actually 35 variables based on the report generated by str, but the following variables are the main one of interest:

id represents a unique identifier for the weather that's being measured. Only one identifier was used for this reading.

year represents the year that the weather's reading was taken. The reading was taken in 2010.

month reprensents the month for the year that the weather's reading was taken.

element represents maximum or minimum temperature for the weather reading.

d1...d31 represents days of the month that the temperature was recorded. For days where there no readings, the value was set to NA

(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

```
# Tidy weather data
library(reshape2)  #library needed for melt if not present
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
melt(read.table("weather.txt", stringsAsFactors = FALSE,
    header = TRUE), id.vars = c("id", "year",
    "month", "element"), measure.vars = c("d1",
    "d2", "d3", "d4", "d5", "d6", "d7", "d8",
    "d9", "d10", "d11", "d12", "d13", "d14", "d15",
    "d16", "d17", "d18", "d19", "d20", "d21",
    "d22", "d23", "d24", "d25", "d26", "d27",
    "d28", "d29", "d30", "d31"), variable.name = "weather_variable",
    value.name = "weather_value", na.rm = TRUE)  # reshape the data by setting all the days into one colum
```

```
##              id year month element
## 21  MX000017004 2010    12    TMAX
## 22  MX000017004 2010    12    TMIN
## 25  MX000017004 2010     2    TMAX
## 26  MX000017004 2010     2    TMIN
## 41  MX000017004 2010    11    TMAX
## 42  MX000017004 2010    11    TMIN
## 47  MX000017004 2010     2    TMAX
## 48  MX000017004 2010     2    TMIN
## 57  MX000017004 2010     7    TMAX
## 58  MX000017004 2010     7    TMIN
## 85  MX000017004 2010    11    TMAX
## 86  MX000017004 2010    11    TMIN
## 93  MX000017004 2010     3    TMAX
## 94  MX000017004 2010     3    TMIN
## 103 MX000017004 2010     8    TMAX
## 104 MX000017004 2010     8    TMIN
## 105 MX000017004 2010    10    TMAX
## 106 MX000017004 2010    10    TMIN
## 107 MX000017004 2010    11    TMAX
## 108 MX000017004 2010    11    TMIN
## 131 MX000017004 2010    12    TMAX
## 132 MX000017004 2010    12    TMIN
## 149 MX000017004 2010    10    TMAX
## 150 MX000017004 2010    10    TMIN
## 169 MX000017004 2010     8    TMAX
## 170 MX000017004 2010     8    TMIN
## 203 MX000017004 2010     3    TMAX
## 204 MX000017004 2010     3    TMIN
## 223 MX000017004 2010     2    TMAX
## 224 MX000017004 2010     2    TMIN
## 279 MX000017004 2010     8    TMAX
## 280 MX000017004 2010     8    TMIN
## 299 MX000017004 2010     7    TMAX
## 300 MX000017004 2010     7    TMIN
## 303 MX000017004 2010    10    TMAX
## 304 MX000017004 2010    10    TMIN
## 325 MX000017004 2010    10    TMAX
## 326 MX000017004 2010    10    TMIN
## 335 MX000017004 2010     3    TMAX
## 336 MX000017004 2010     3    TMIN
## 363 MX000017004 2010     6    TMAX
## 364 MX000017004 2010     6    TMIN
## 487 MX000017004 2010     2    TMAX
## 488 MX000017004 2010     2    TMIN
## 499 MX000017004 2010     8    TMAX
## 500 MX000017004 2010     8    TMIN
## 543 MX000017004 2010     8    TMAX
## 544 MX000017004 2010     8    TMIN
## 569 MX000017004 2010    11    TMAX
## 570 MX000017004 2010    11    TMIN
## 579 MX000017004 2010     4    TMAX
## 580 MX000017004 2010     4    TMIN
## 581 MX000017004 2010     5    TMAX
## 582 MX000017004 2010     5    TMIN
## 591 MX000017004 2010    11    TMAX
## 592 MX000017004 2010    11    TMIN
## 611 MX000017004 2010    10    TMAX
## 612 MX000017004 2010    10    TMIN
## 627 MX000017004 2010     6    TMAX
## 628 MX000017004 2010     6    TMIN
## 631 MX000017004 2010     8    TMAX
## 632 MX000017004 2010     8    TMIN
## 639 MX000017004 2010     1    TMAX
## 640 MX000017004 2010     1    TMIN
## 675 MX000017004 2010     8    TMAX
## 676 MX000017004 2010     8    TMIN
##     weather_variable weather_value
## 21                d1           299
## 22                d1           138
## 25                d2           273
## 26                d2           144
## 41                d2           313
## 42                d2           163
## 47                d3           241
## 48                d3           144
## 57                d3           286
## 58                d3           175
## 85                d4           272
## 86                d4           120
## 93                d5           321
## 94                d5           142
## 103               d5           296
## 104               d5           158
## 105               d5           270
## 106               d5           140
## 107               d5           263
## 108               d5            79
## 131               d6           278
## 132               d6           105
## 149               d7           281
## 150               d7           129
## 169               d8           290
## 170               d8           173
## 203              d10           345
## 204              d10           168
## 223              d11           297
## 224              d11           134
## 279              d13           298
## 280              d13           165
## 299              d14           299
## 300              d14           165
## 303              d14           295
## 304              d14           130
## 325              d15           287
## 326              d15           105
## 335              d16           311
## 336              d16           176
## 363              d17           280
## 364              d17           175
## 487              d23           299
## 488              d23           107
## 499              d23           264
## 500              d23           150
## 543              d25           297
## 544              d25           156
## 569              d26           281
## 570              d26           121
## 579              d27           363
## 580              d27           167
## 581              d27           332
## 582              d27           182
## 591              d27           277
## 592              d27           142
## 611              d28           312
## 612              d28           150
## 627              d29           301
## 628              d29           180
## 631              d29           280
## 632              d29           153
## 639              d30           278
## 640              d30           145
## 675              d31         254
## 676              d31         154
```

```
# dcast to get the lenght of each readings
aswer = melt(read.table("weather.txt", stringsAsFactors = FALSE,
    header = TRUE), id.vars = c("id", "year",
    "month", "element"), measure.vars = c("d1",
    "d2", "d3", "d4", "d5", "d6", "d7", "d8",
    "d9", "d10", "d11", "d12", "d13", "d14", "d15",
    "d16", "d17", "d18", "d19", "d20", "d21",
    "d22", "d23", "d24", "d25", "d26", "d27",
    "d28", "d29", "d30", "d31"), na.rm = TRUE)
dcast(aswer, year + month ~ variable)
```

```
## Aggregation function missing: defaulting to length
```

```
##    year month d1 d2 d3 d4 d5 d6 d7 d8 d10
## 1  2010     1  0  0  0  0  0  0  0  0   0
## 2  2010     2  0  2  2  0  0  0  0  0   0
## 3  2010     3  0  0  0  0  2  0  0  0   2
## 4  2010     4  0  0  0  0  0  0  0  0   0
## 5  2010     5  0  0  0  0  0  0  0  0   0
## 6  2010     6  0  0  0  0  0  0  0  0   0
## 7  2010     7  0  0  2  0  0  0  0  0   0
## 8  2010     8  0  0  0  0  2  0  0  2   0
## 9  2010    10  0  0  0  0  2  0  2  0   0
## 10 2010    11  0  2  0  2  2  0  0  0   0
## 11 2010    12  2  0  0  0  0  2  0  0   0
##    d11 d13 d14 d15 d16 d17 d23 d25 d26 d27
## 1    0   0   0   0   0   0   0   0   0   0
## 2    2   0   0   0   0   0   2   0   0   0
## 3    0   0   0   0   2   0   0   0   0   0
## 4    0   0   0   0   0   0   0   0   0   2
## 5    0   0   0   0   0   0   0   0   0   2
## 6    0   0   0   0   0   2   0   0   0   0
## 7    0   0   2   0   0   0   0   0   0   0
## 8    0   2   0   0   0   0   2   2   0   0
## 9    0   0   2   2   0   0   0   0   0   0
## 10   0   0   0   0   0   0   0   0   2   2
## 11   0   0   0   0   0   0   0   0   0   0
##    d28 d29 d30 d31
## 1    0   0   2   0
## 2    0   0   0   0
## 3    0   0   0   0
## 4    0   0   0   0
## 5    0   0   0   0
## 6    0   2   0   0
## 7    0   0   0   0
## 8    0   2   0   2
## 9    2   0   0   0
## 10   0   0   0   0
## 11   0   0   0   0
```

```
# mutate
```

Problem 2: Data Manipulation

In this problem we will use the `babynames` data. Use the data to answer the following questions.

```
# baby names data structure
str(babynames)
```

```
## Classes ’tbl_df’, ’tbl’ and ’data.frame’:    1825433 obs. of  5 variables:
##  $ year: num  1880 1880 1880 1880 1880 1880 1880 1880 1880 1880 ...
##  $ sex : chr  "F" "F" "F" "F" ...
##  $ name: chr  "Mary" "Anna" "Emma" "Elizabeth" ...
##  $ n   : int  7065 2604 2003 1939 1746 1578 1472 1414 1320 1288 ...
##  $ prop: num  0.0724 0.0267 0.0205 0.0199 0.0179 ...
```

```
# baby names top 6 results
head(babynames)
```

```
## # A tibble: 6 × 5
##    year   sex      name     n       prop
##   <dbl> <chr>     <chr> <int>      <dbl>
## 1  1880     F      Mary  7065 0.07238359
## 2  1880     F      Anna  2604 0.02667896
## 3  1880     F      Emma  2003 0.02052149
## 4  1880     F Elizabeth  1939 0.01986579
## 5  1880     F    Minnie  1746 0.01788843
## 6  1880     F  Margaret  1578 0.01616720
```

(a) What name has been used for the most number of years (when used for a single gender)?

```
# finding baby names used
names = babynames %>% tbl_df() %>% select(year,
    sex, name, n) %>% arrange(year, sex, desc(n))
print(names)
```

```
## # A tibble: 1,825,433 × 4
##     year   sex      name     n
##    <dbl> <chr>     <chr> <int>
## 1   1880     F      Mary  7065
## 2   1880     F      Anna  2604
## 3   1880     F      Emma  2003
## 4   1880     F Elizabeth  1939
## 5   1880     F    Minnie  1746
## 6   1880     F  Margaret  1578
## 7   1880     F       Ida  1472
## 8   1880     F     Alice  1414
## 9   1880     F    Bertha  1320
## 10  1880     F     Sarah  1288
## # ... with 1,825,423 more rows
```

(b) Which name received the largest percentage of any name for any year (consider boy and girl names as distinct)?

```
# finding most popular name
names = babynames %>% tbl_df() %>% select(year,
    sex, name, n) %>% arrange(year, sex, desc(n)) %>%
    mutate(percentage = (n/nrow(babynames)) *
        100)
head(names$name, 1)  #Take the top result from the generated records
```

```
## [1] "Mary"
```

(c) Which name recorded in the data set has been out of use for the longest time?

```
# unused name for the longest

babynames %>% group_by(name) %>% summarize(old = max(year)) %>%
    ungroup %>% head(1)
```

```
## # A tibble: 1 × 2
##    name   old
##   <chr> <dbl>
## 1 Aaban  2014
```

(d) For each year, what is the total number of names that were recorded? Treat boy and girl versions of the same name as two separate names. Did you need to look at the data to answer this question?

```
babynames %>% group_by(year, sex) %>% summarise(uniqueName = n_distinct(name))
```

```
## Source: local data frame [270 x 3]
## Groups: year [?]
##
##     year   sex uniqueName
##    <dbl> <chr>      <int>
## 1   1880     F        942
## 2   1880     M       1058
## 3   1881     F        938
## 4   1881     M        997
## 5   1882     F       1028
## 6   1882     M       1099
## 7   1883     F       1054
## 8   1883     M       1030
## 9   1884     F       1172
## 10  1884     M       1125
## # ... with 260 more rows
```