weather\_report

## Loading libraries

library(tidyr)  
library(dplyr)  
library(stringr)  
library(magrittr)  
library(lubridate)  
library(purrr)

## Loading data

weather <- readRDS("~/R/rclass/weather.rds")

## Tidying data

Let’s look at the data.

str(weather)

## 'data.frame': 286 obs. of 35 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ year : int 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...  
## $ month : int 12 12 12 12 12 12 12 12 12 12 ...  
## $ measure: chr "Max.TemperatureF" "Mean.TemperatureF" "Min.TemperatureF" "Max.Dew.PointF" ...  
## $ X1 : chr "64" "52" "39" "46" ...  
## $ X2 : chr "42" "38" "33" "40" ...  
## $ X3 : chr "51" "44" "37" "49" ...  
## $ X4 : chr "43" "37" "30" "24" ...  
## $ X5 : chr "42" "34" "26" "37" ...  
## $ X6 : chr "45" "42" "38" "45" ...  
## $ X7 : chr "38" "30" "21" "36" ...  
## $ X8 : chr "29" "24" "18" "28" ...  
## $ X9 : chr "49" "39" "29" "49" ...  
## $ X10 : chr "48" "43" "38" "45" ...  
## $ X11 : chr "39" "36" "32" "37" ...  
## $ X12 : chr "39" "35" "31" "28" ...  
## $ X13 : chr "42" "37" "32" "28" ...  
## $ X14 : chr "45" "39" "33" "29" ...  
## $ X15 : chr "42" "37" "32" "33" ...  
## $ X16 : chr "44" "40" "35" "42" ...  
## $ X17 : chr "49" "45" "41" "46" ...  
## $ X18 : chr "44" "40" "36" "34" ...  
## $ X19 : chr "37" "33" "29" "25" ...  
## $ X20 : chr "36" "32" "27" "30" ...  
## $ X21 : chr "36" "33" "30" "30" ...  
## $ X22 : chr "44" "39" "33" "39" ...  
## $ X23 : chr "47" "45" "42" "45" ...  
## $ X24 : chr "46" "44" "41" "46" ...  
## $ X25 : chr "59" "52" "44" "58" ...  
## $ X26 : chr "50" "44" "37" "31" ...  
## $ X27 : chr "52" "45" "38" "34" ...  
## $ X28 : chr "52" "46" "40" "42" ...  
## $ X29 : chr "41" "36" "30" "26" ...  
## $ X30 : chr "30" "26" "22" "10" ...  
## $ X31 : chr "30" "25" "20" "8" ...

We can see that the names of the columns are contained as rows in the ‘Measure’ column, and the columns that we have right now seem to be days of the month. So, our rows are our columns and our columns are our rows.

### Rotating the columns

weather1 <- weather %>% select(-1) %>%   
 pivot\_longer(cols = starts\_with("X"), names\_to = "day", values\_to = "value") %>%   
 pivot\_wider(names\_from = "measure", values\_from = "value")  
  
weather1$day <- str\_replace\_all(weather1$day, "X", "")

head(weather1)

## # A tibble: 6 x 25  
## year month day Max.TemperatureF Mean.Temperatur~ Min.TemperatureF  
## <int> <int> <chr> <chr> <chr> <chr>   
## 1 2014 12 1 64 52 39   
## 2 2014 12 2 42 38 33   
## 3 2014 12 3 51 44 37   
## 4 2014 12 4 43 37 30   
## 5 2014 12 5 42 34 26   
## 6 2014 12 6 45 42 38   
## # ... with 19 more variables: Max.Dew.PointF <chr>, MeanDew.PointF <chr>,  
## # Min.DewpointF <chr>, Max.Humidity <chr>, Mean.Humidity <chr>,  
## # Min.Humidity <chr>, Max.Sea.Level.PressureIn <chr>,  
## # Mean.Sea.Level.PressureIn <chr>, Min.Sea.Level.PressureIn <chr>,  
## # Max.VisibilityMiles <chr>, Mean.VisibilityMiles <chr>,  
## # Min.VisibilityMiles <chr>, Max.Wind.SpeedMPH <chr>,  
## # Mean.Wind.SpeedMPH <chr>, Max.Gust.SpeedMPH <chr>,  
## # PrecipitationIn <chr>, CloudCover <chr>, Events <chr>,  
## # WindDirDegrees <chr>

tail(weather1)

## # A tibble: 6 x 25  
## year month day Max.TemperatureF Mean.Temperatur~ Min.TemperatureF  
## <int> <int> <chr> <chr> <chr> <chr>   
## 1 2015 12 26 <NA> <NA> <NA>   
## 2 2015 12 27 <NA> <NA> <NA>   
## 3 2015 12 28 <NA> <NA> <NA>   
## 4 2015 12 29 <NA> <NA> <NA>   
## 5 2015 12 30 <NA> <NA> <NA>   
## 6 2015 12 31 <NA> <NA> <NA>   
## # ... with 19 more variables: Max.Dew.PointF <chr>, MeanDew.PointF <chr>,  
## # Min.DewpointF <chr>, Max.Humidity <chr>, Mean.Humidity <chr>,  
## # Min.Humidity <chr>, Max.Sea.Level.PressureIn <chr>,  
## # Mean.Sea.Level.PressureIn <chr>, Min.Sea.Level.PressureIn <chr>,  
## # Max.VisibilityMiles <chr>, Mean.VisibilityMiles <chr>,  
## # Min.VisibilityMiles <chr>, Max.Wind.SpeedMPH <chr>,  
## # Mean.Wind.SpeedMPH <chr>, Max.Gust.SpeedMPH <chr>,  
## # PrecipitationIn <chr>, CloudCover <chr>, Events <chr>,  
## # WindDirDegrees <chr>

str(weather1)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 403 obs. of 25 variables:  
## $ year : int 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...  
## $ month : int 12 12 12 12 12 12 12 12 12 12 ...  
## $ day : chr "1" "2" "3" "4" ...  
## $ Max.TemperatureF : chr "64" "42" "51" "43" ...  
## $ Mean.TemperatureF : chr "52" "38" "44" "37" ...  
## $ Min.TemperatureF : chr "39" "33" "37" "30" ...  
## $ Max.Dew.PointF : chr "46" "40" "49" "24" ...  
## $ MeanDew.PointF : chr "40" "27" "42" "21" ...  
## $ Min.DewpointF : chr "26" "17" "24" "13" ...  
## $ Max.Humidity : chr "74" "92" "100" "69" ...  
## $ Mean.Humidity : chr "63" "72" "79" "54" ...  
## $ Min.Humidity : chr "52" "51" "57" "39" ...  
## $ Max.Sea.Level.PressureIn : chr "30.45" "30.71" "30.4" "30.56" ...  
## $ Mean.Sea.Level.PressureIn: chr "30.13" "30.59" "30.07" "30.33" ...  
## $ Min.Sea.Level.PressureIn : chr "30.01" "30.4" "29.87" "30.09" ...  
## $ Max.VisibilityMiles : chr "10" "10" "10" "10" ...  
## $ Mean.VisibilityMiles : chr "10" "8" "5" "10" ...  
## $ Min.VisibilityMiles : chr "10" "2" "1" "10" ...  
## $ Max.Wind.SpeedMPH : chr "22" "24" "29" "25" ...  
## $ Mean.Wind.SpeedMPH : chr "13" "15" "12" "12" ...  
## $ Max.Gust.SpeedMPH : chr "29" "29" "38" "33" ...  
## $ PrecipitationIn : chr "0.01" "0.10" "0.44" "0.00" ...  
## $ CloudCover : chr "6" "7" "8" "3" ...  
## $ Events : chr "Rain" "Rain-Snow" "Rain" "" ...  
## $ WindDirDegrees : chr "268" "62" "254" "292" ...

Now rows and columns are where they are supposed to be, and so the data it tidy, but still not clean. We have a bunch of NA’s in our data, and the columns that should be numeric (‘num’ or ‘int’) are actually strings (‘chr’). Also some of the values in the ‘PrecipitationIn’ column are ‘T’, even though it is not a logical variable, but a numeric one. Some of the cells in the ‘Events’ column are empty, we can assume that those denote the days when nothing happened.

We need to remove the missing values and values that are of wrong type, and change the ‘empty’ events to ‘Nothing’, so that it is clearer and easier to read and filter. Then we are going to change the data types of or columns.

### Removing missing values

weather2 <- na.omit(weather1)   
weather2 %<>% filter(PrecipitationIn != "T")   
weather2$Events <- ifelse(str\_length(weather2$Events) > 0,   
 weather2$Events,   
 "Nothing")  
  
#str\_replace\_all would have been better, but it cannot use empty string as a pattern to replace  
  
str(weather2)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 311 obs. of 25 variables:  
## $ year : int 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...  
## $ month : int 12 12 12 12 12 12 12 12 12 12 ...  
## $ day : chr "1" "2" "3" "4" ...  
## $ Max.TemperatureF : chr "64" "42" "51" "43" ...  
## $ Mean.TemperatureF : chr "52" "38" "44" "37" ...  
## $ Min.TemperatureF : chr "39" "33" "37" "30" ...  
## $ Max.Dew.PointF : chr "46" "40" "49" "24" ...  
## $ MeanDew.PointF : chr "40" "27" "42" "21" ...  
## $ Min.DewpointF : chr "26" "17" "24" "13" ...  
## $ Max.Humidity : chr "74" "92" "100" "69" ...  
## $ Mean.Humidity : chr "63" "72" "79" "54" ...  
## $ Min.Humidity : chr "52" "51" "57" "39" ...  
## $ Max.Sea.Level.PressureIn : chr "30.45" "30.71" "30.4" "30.56" ...  
## $ Mean.Sea.Level.PressureIn: chr "30.13" "30.59" "30.07" "30.33" ...  
## $ Min.Sea.Level.PressureIn : chr "30.01" "30.4" "29.87" "30.09" ...  
## $ Max.VisibilityMiles : chr "10" "10" "10" "10" ...  
## $ Mean.VisibilityMiles : chr "10" "8" "5" "10" ...  
## $ Min.VisibilityMiles : chr "10" "2" "1" "10" ...  
## $ Max.Wind.SpeedMPH : chr "22" "24" "29" "25" ...  
## $ Mean.Wind.SpeedMPH : chr "13" "15" "12" "12" ...  
## $ Max.Gust.SpeedMPH : chr "29" "29" "38" "33" ...  
## $ PrecipitationIn : chr "0.01" "0.10" "0.44" "0.00" ...  
## $ CloudCover : chr "6" "7" "8" "3" ...  
## $ Events : chr "Rain" "Rain-Snow" "Rain" "Nothing" ...  
## $ WindDirDegrees : chr "268" "62" "254" "292" ...  
## - attr(\*, "na.action")= 'omit' Named int 91 92 93 155 173 189 217 256 280 310 ...  
## ..- attr(\*, "names")= chr "91" "92" "93" "155" ...

## Fixing data types

### Fixing dates

First, we are going to fix the date. Right now it’s in three separate columns, two of which are ‘int’ and one is ‘chr’. We are going to clip them together and transform into a ‘Date’ type.

weather2$Date = str\_c(weather2$year, weather2$month, weather2$day, sep = ".") %>% ymd()  
weather2 %<>% select(-c(1:3))  
str(weather2)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 311 obs. of 23 variables:  
## $ Max.TemperatureF : chr "64" "42" "51" "43" ...  
## $ Mean.TemperatureF : chr "52" "38" "44" "37" ...  
## $ Min.TemperatureF : chr "39" "33" "37" "30" ...  
## $ Max.Dew.PointF : chr "46" "40" "49" "24" ...  
## $ MeanDew.PointF : chr "40" "27" "42" "21" ...  
## $ Min.DewpointF : chr "26" "17" "24" "13" ...  
## $ Max.Humidity : chr "74" "92" "100" "69" ...  
## $ Mean.Humidity : chr "63" "72" "79" "54" ...  
## $ Min.Humidity : chr "52" "51" "57" "39" ...  
## $ Max.Sea.Level.PressureIn : chr "30.45" "30.71" "30.4" "30.56" ...  
## $ Mean.Sea.Level.PressureIn: chr "30.13" "30.59" "30.07" "30.33" ...  
## $ Min.Sea.Level.PressureIn : chr "30.01" "30.4" "29.87" "30.09" ...  
## $ Max.VisibilityMiles : chr "10" "10" "10" "10" ...  
## $ Mean.VisibilityMiles : chr "10" "8" "5" "10" ...  
## $ Min.VisibilityMiles : chr "10" "2" "1" "10" ...  
## $ Max.Wind.SpeedMPH : chr "22" "24" "29" "25" ...  
## $ Mean.Wind.SpeedMPH : chr "13" "15" "12" "12" ...  
## $ Max.Gust.SpeedMPH : chr "29" "29" "38" "33" ...  
## $ PrecipitationIn : chr "0.01" "0.10" "0.44" "0.00" ...  
## $ CloudCover : chr "6" "7" "8" "3" ...  
## $ Events : chr "Rain" "Rain-Snow" "Rain" "Nothing" ...  
## $ WindDirDegrees : chr "268" "62" "254" "292" ...  
## $ Date : Date, format: "2014-12-01" "2014-12-02" ...  
## - attr(\*, "na.action")= 'omit' Named int 91 92 93 155 173 189 217 256 280 310 ...  
## ..- attr(\*, "names")= chr "91" "92" "93" "155" ...

### Fixing data types

We now will change the type of all columns except ‘Events’ and ‘Date’ to numeric. Events will be transformed into a factor, and Date is already good enough.

weather3 = purrr::map\_dfc(weather2[c(1:20, 22)], .f = as.numeric)  
weather3$Date = weather2$Date  
weather3$Events = as.factor(weather2$Events)  
str(weather3)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 311 obs. of 23 variables:  
## $ Max.TemperatureF : num 64 42 51 43 42 45 38 29 49 48 ...  
## $ Mean.TemperatureF : num 52 38 44 37 34 42 30 24 39 43 ...  
## $ Min.TemperatureF : num 39 33 37 30 26 38 21 18 29 38 ...  
## $ Max.Dew.PointF : num 46 40 49 24 37 45 36 28 49 45 ...  
## $ MeanDew.PointF : num 40 27 42 21 25 40 20 16 41 39 ...  
## $ Min.DewpointF : num 26 17 24 13 12 36 -3 3 28 37 ...  
## $ Max.Humidity : num 74 92 100 69 85 100 92 92 100 100 ...  
## $ Mean.Humidity : num 63 72 79 54 66 93 61 70 93 95 ...  
## $ Min.Humidity : num 52 51 57 39 47 85 29 47 86 89 ...  
## $ Max.Sea.Level.PressureIn : num 30.4 30.7 30.4 30.6 30.7 ...  
## $ Mean.Sea.Level.PressureIn: num 30.1 30.6 30.1 30.3 30.6 ...  
## $ Min.Sea.Level.PressureIn : num 30 30.4 29.9 30.1 30.4 ...  
## $ Max.VisibilityMiles : num 10 10 10 10 10 10 10 10 10 10 ...  
## $ Mean.VisibilityMiles : num 10 8 5 10 10 4 10 8 2 3 ...  
## $ Min.VisibilityMiles : num 10 2 1 10 5 0 5 2 1 1 ...  
## $ Max.Wind.SpeedMPH : num 22 24 29 25 22 22 25 21 38 23 ...  
## $ Mean.Wind.SpeedMPH : num 13 15 12 12 10 8 15 13 20 13 ...  
## $ Max.Gust.SpeedMPH : num 29 29 38 33 26 25 32 28 52 29 ...  
## $ PrecipitationIn : num 0.01 0.1 0.44 0 0.11 1.09 0.13 0.03 2.9 0.28 ...  
## $ CloudCover : num 6 7 8 3 5 8 6 8 8 8 ...  
## $ WindDirDegrees : num 268 62 254 292 61 313 350 354 38 357 ...  
## $ Date : Date, format: "2014-12-01" "2014-12-02" ...  
## $ Events : Factor w/ 12 levels "Fog","Fog-Rain",..: 8 9 8 7 8 8 8 11 8 8 ...

That’s better.

## Printing the result

summary(weather3)

## Max.TemperatureF Mean.TemperatureF Min.TemperatureF Max.Dew.PointF   
## Min. :18.00 Min. : 8.0 Min. :-3.00 Min. :-6.00   
## 1st Qu.:42.00 1st Qu.:36.0 1st Qu.:30.00 1st Qu.:32.00   
## Median :59.00 Median :52.0 Median :44.00 Median :46.00   
## Mean :58.24 Mean :50.7 Mean :42.59 Mean :44.79   
## 3rd Qu.:75.00 3rd Qu.:67.5 3rd Qu.:59.50 3rd Qu.:61.00   
## Max. :96.00 Max. :84.0 Max. :74.00 Max. :75.00   
##   
## MeanDew.PointF Min.DewpointF Max.Humidity Mean.Humidity   
## Min. :-11.00 Min. :-18.00 Min. : 39.00 Min. :28.00   
## 1st Qu.: 23.00 1st Qu.: 14.50 1st Qu.: 73.00 1st Qu.:55.50   
## Median : 40.00 Median : 33.00 Median : 86.00 Median :66.00   
## Mean : 38.14 Mean : 31.35 Mean : 85.86 Mean :65.68   
## 3rd Qu.: 55.00 3rd Qu.: 51.00 3rd Qu.: 93.00 3rd Qu.:76.00   
## Max. : 70.00 Max. : 68.00 Max. :1000.00 Max. :98.00   
##   
## Min.Humidity Max.Sea.Level.PressureIn Mean.Sea.Level.PressureIn  
## Min. :16.00 Min. :29.58 Min. :29.49   
## 1st Qu.:34.00 1st Qu.:30.00 1st Qu.:29.88   
## Median :45.00 Median :30.14 Median :30.04   
## Mean :47.87 Mean :30.17 Mean :30.05   
## 3rd Qu.:59.00 3rd Qu.:30.32 3rd Qu.:30.20   
## Max. :96.00 Max. :30.88 Max. :30.77   
##   
## Min.Sea.Level.PressureIn Max.VisibilityMiles Mean.VisibilityMiles  
## Min. :29.16 Min. : 2.000 Min. :-1.000   
## 1st Qu.:29.75 1st Qu.:10.000 1st Qu.: 8.000   
## Median :29.94 Median :10.000 Median :10.000   
## Mean :29.93 Mean : 9.891 Mean : 8.727   
## 3rd Qu.:30.09 3rd Qu.:10.000 3rd Qu.:10.000   
## Max. :30.64 Max. :10.000 Max. :10.000   
##   
## Min.VisibilityMiles Max.Wind.SpeedMPH Mean.Wind.SpeedMPH  
## Min. : 0.000 Min. :10.00 Min. : 4.00   
## 1st Qu.: 2.000 1st Qu.:16.00 1st Qu.: 8.00   
## Median :10.000 Median :20.00 Median :10.00   
## Mean : 6.595 Mean :20.85 Mean :10.83   
## 3rd Qu.:10.000 3rd Qu.:24.00 3rd Qu.:13.00   
## Max. :10.000 Max. :38.00 Max. :22.00   
##   
## Max.Gust.SpeedMPH PrecipitationIn CloudCover WindDirDegrees   
## Min. : 0.00 Min. :0.0000 Min. :0.000 Min. : 1.0   
## 1st Qu.:21.00 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:113.5   
## Median :25.00 Median :0.0000 Median :5.000 Median :223.0   
## Mean :27.17 Mean :0.1195 Mean :4.598 Mean :201.4   
## 3rd Qu.:32.00 3rd Qu.:0.0700 3rd Qu.:7.000 3rd Qu.:278.0   
## Max. :94.00 Max. :2.9000 Max. :8.000 Max. :360.0   
##   
## Date Events   
## Min. :2014-12-01 Nothing :180   
## 1st Qu.:2015-02-27 Rain : 70   
## Median :2015-05-24 Snow : 21   
## Mean :2015-05-30 Rain-Snow: 9   
## 3rd Qu.:2015-09-02 Fog-Rain : 8   
## Max. :2015-12-01 Fog-Snow : 7   
## (Other) : 16