Preparing the weather data

Georgie Knight

19 August, 2016

We'll now prepare the weather data to be combined with the status data

Dplyr and tidyr

Load the *dplyr* and *tidyr* packages which will help us wrangle the data:

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("tidyr")
```

Loading up the data

Load up our weather data and station information:

```
weather_data <- read.csv("201508_weather_data.csv")
weather_df   <- data.frame(weather_data)
weather      <- dplyr::tbl_df(weather_df)
glimpse(weather)
```

```
## Observations: 1,825
## Variables: 24
## $ PDT                       <fctr> 9/1/2014, 9/2/2014, 9/3/2014, 9/4/2...
## $ Max.TemperatureF          <int> 83, 72, 76, 74, 72, 72, 72, 68, 72, ...
## $ Mean.TemperatureF         <int> 70, 66, 69, 68, 66, 66, 66, 64, 65, ...
## $ Min.TemperatureF          <int> 57, 60, 61, 61, 60, 60, 60, 59, 57, ...
## $ Max.Dew.PointF            <int> 58, 58, 57, 57, 57, 55, 54, 52, 56, ...
## $ MeanDew.PointF            <int> 56, 57, 56, 57, 56, 54, 53, 51, 53, ...
## $ Min.DewpointF             <int> 52, 55, 55, 56, 54, 52, 45, 50, 52, ...
## $ Max.Humidity              <int> 86, 84, 84, 84, 84, 84, 78, 72, 87, ...
## $ Mean.Humidity             <int> 64, 73, 69, 71, 71, 69, 66, 64, 72, ...
## $ Min.Humidity              <int> 42, 61, 53, 57, 57, 53, 53, 55, 57, ...
## $ Max.Sea.Level.PressureIn  <dbl> 29.86, 29.87, 29.81, 29.81, 29.92, 2...
```

```
## $ Mean.Sea.Level.PressureIn <dbl> 29.82, 29.82, 29.76, 29.76, 29.87, 2...
## $ Min.Sea.Level.PressureIn  <dbl> 29.76, 29.79, 29.72, 29.72, 29.81, 2...
## $ Max.VisibilityMiles       <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Mean.VisibilityMiles      <int> 10, 10, 10, 10, 9, 9, 10, 10, 10, 9,...
## $ Min.VisibilityMiles       <int> 8, 7, 10, 8, 7, 7, 10, 10, 10, 5, 2,...
## $ Max.Wind.SpeedMPH         <int> 16, 21, 21, 22, 18, 17, 18, 18, 17, ...
## $ Mean.Wind.SpeedMPH        <int> 7, 8, 8, 8, 8, 9, 10, 12, 7, 5, 7, 8...
## $ Max.Gust.SpeedMPH         <int> 20, NA, 24, 25, 32, 30, 28, 22, 21, ...
## $ PrecipitationIn           <fctr> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CloudCover                <int> 0, 5, 4, 5, 4, 4, 3, 4, 5, 1, 2, 1, ...
## $ Events                    <fctr> , , , , , , , , , , , , , , , , Rai...
## $ WindDirDegrees            <int> 290, 290, 276, 301, 309, 290, 293, 2...
## $ Zip                       <int> 94107, 94107, 94107, 94107, 94107, 9...
```

We note the following information

Daily weather information per service area, provided from Weather Underground in PST. Weather is listed from north to south (San Francisco, Redwood City, Palo Alto, Mountain View, San Jose). -Precipitation In "numeric, in form x.xx but alpha""T""= trace when amount less than .01 inch" -Cloud Cover "scale of 0-8, 0=clear"
-Zip: 94107=San Francisco, 94063=Redwood City, 94301=Palo Alto, 94041=Mountain View, 95113= San Jose" -No data recorded on 8/8/2015 for 94301

```
weather <- mutate(weather,
           landmark = ifelse(Zip == 94107, 'San Francisco',
                      ifelse(Zip == 94063, 'Redwood City',
                      ifelse(Zip == 94301, 'Palo Alto',
                      ifelse(Zip == 94041, 'Mountain View',
                      ifelse(Zip == 95113, 'San Jose',NA
                             )))))
                 )
```

Take a look at some random rows:

```
randomRows = sample(1:length(weather$PDT), 10, replace=T)
v <- slice(weather, randomRows)
select(v, PDT, Zip, landmark)
```

```
## # A tibble: 10 x 3
##           PDT   Zip      landmark
##        <fctr> <int>         <chr>
## 1   4/29/2015 95113      San Jose
## 2   11/6/2014 94301     Palo Alto
## 3    9/2/2014 94301     Palo Alto
## 4   3/12/2015 94041 Mountain View
## 5  12/25/2014 94107 San Francisco
## 6   7/24/2015 94107 San Francisco
## 7   3/27/2015 94107 San Francisco
## 8  12/26/2014 94107 San Francisco
## 9    4/2/2015 94301     Palo Alto
## 10  4/13/2015 94063  Redwood City
```

Create a Date column

```
weather <- weather %>%
  mutate(Date = as.Date(PDT, format ="%m/%d/%Y")) %>%
  select(-PDT)
```

We'll now be able to add the weather data to the status data using the landmark and Date column.

```
write.csv(weather, file="weather.csv")
```