Combining the data

Georgie Knight

19 August, 2016

We'll now combine our data file on station data with the station information

Dplyr and tidyr

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

```r
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
```

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

Loading up the data

Load up our status data and station information:

```r
status_read  <- read.csv("status.csv")
station_read <- read.csv("station.csv")
weather_read <- read.csv("weather.csv")

status_df    <- data.frame(status_read)
station_df   <- data.frame(station_read)
weather_df   <- data.frame(weather_read)

status       <- dplyr::tbl_df(status_df)
station      <- dplyr::tbl_df(station_df)
weather      <- dplyr::tbl_df(weather_df)

status       <- select(status, -X)
station      <- select(station, -X)
weather      <- select(weather, -X)

status <- left_join(status, station)
```

```
## Joining, by = "station_id"
```

```
glimpse(status)
```

```
## Observations: 1,135,974
## Variables: 10
## $ station_id      <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ bikes_available <int> 15, 14, 15, 14, 15, 14, 15, 14, 15, 14, 15, 14...
## $ docks_available <int> 12, 13, 12, 13, 12, 13, 12, 13, 12, 13, 12, 13...
## $ time            <fctr> 2014-09-01 00:00:03, 2014-09-01 02:57:02, 201...
## $ name            <fctr> San Jose Diridon Caltrain Station, San Jose D...
## $ lat             <dbl> 37.32973, 37.32973, 37.32973, 37.32973, 37.329...
## $ long            <dbl> -121.9018, -121.9018, -121.9018, -121.9018, -1...
## $ dockcount       <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...
## $ landmark        <fctr> San Jose, San Jose, San Jose, San Jose, San J...
## $ installation    <fctr> 2013-08-29, 2013-08-29, 2013-08-29, 2013-08-2...
```

The dock count column is now redundant

```
status <- status %>%
  select(-dockcount) %>%
  mutate(installation = as.Date(installation))
glimpse(status)
```

```
## Observations: 1,135,974
## Variables: 9
## $ station_id      <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ bikes_available <int> 15, 14, 15, 14, 15, 14, 15, 14, 15, 14, 15, 14...
## $ docks_available <int> 12, 13, 12, 13, 12, 13, 12, 13, 12, 13, 12, 13...
## $ time            <fctr> 2014-09-01 00:00:03, 2014-09-01 02:57:02, 201...
## $ name            <fctr> San Jose Diridon Caltrain Station, San Jose D...
## $ lat             <dbl> 37.32973, 37.32973, 37.32973, 37.32973, 37.329...
## $ long            <dbl> -121.9018, -121.9018, -121.9018, -121.9018, -1...
## $ landmark        <fctr> San Jose, San Jose, San Jose, San Jose, San J...
## $ installation    <date> 2013-08-29, 2013-08-29, 2013-08-29, 2013-08-2...
```

Take a look at some random rows:

```
randomRows = sample(1:length(status$time), 10, replace=T)
slice(status, randomRows)
```

```
## # A tibble: 10 x 9
##    station_id bikes_available docks_available                time
##         <int>           <int>           <int>              <fctr>
## 1          64              11               4 2015-08-18 11:41:02
## 2           3              10               5 2014-11-15 00:26:03
## 3          57              12               3 2015-01-16 09:10:02
## 4          76               2              17 2015-05-15 10:20:02
## 5          66               6              13 2014-09-15 08:46:02
## 6          71               9              10 2014-12-03 13:35:03
## 7          39              18               1 2015-04-17 17:37:02
## 8          74              15                8 2015-04-11 17:37:02
## 9           6               7                8 2014-09-07 13:36:03
## 10         11              12                7 2014-11-24 02:47:02
## # ... with 5 more variables: name <fctr>, lat <dbl>, long <dbl>,
## #   landmark <fctr>, installation <date>
```

we'll now create a Date column and convert the landmark to character

```
status <- status %>%
  mutate(Date = as.Date(time)) %>%
  mutate(landmark = as.character(landmark))

weather <- mutate(weather, Date = as.Date(Date))
```

We're now ready to add the weather data

```
status <- left_join(status, weather)
```

```
## Joining, by = c("landmark", "Date")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factor and character vector, coercing into character vector
```

```
write.csv(status, file="status_full.csv")
```

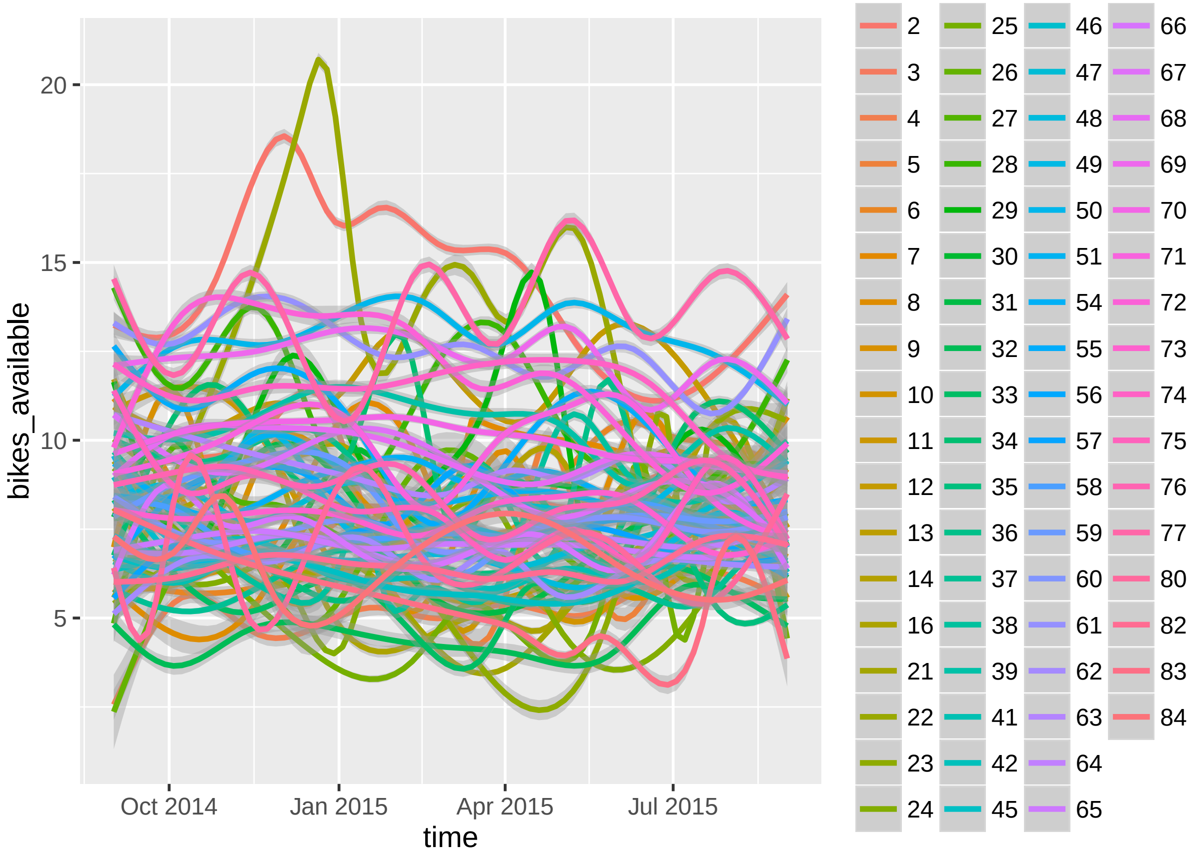
Let's make a quick visual

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
status <- mutate(status, time = ymd_hms(time))
ggplot(status, aes(x= time, y = bikes_available,
                   col = as.factor(station_id)))+
                                   geom_smooth()
```


bikes_available
20
15
10
5
Oct 2014
Jan 2015
Apr 2015
Jul 2015
time
2
3
4
5
6
7
8
9
10
11
12
13
14
16
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
41
42
45
46
47
48
49
50
51
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
80
82
83
84