Preparing the station data

Georgie Knight

18 August, 2016

We'll now prepare the station 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")
```

We note the following important information about the stations:

Although stations were installed prior to 8/29/13 (system launch), no station was active until launch date. Therefore, to accurately capture station popularity, we recommend adjusting all pre-launch installation dates to 8/29/13.

Station names and locations listed on 201508stationdata.csv represent data that was collected on 8/31/15. However, please note that during 9/1/14 and 8/31/15, 5 stations were moved and 1 station stayed in the same location but changed name.

Station 23: From 9/1/14 - 10/22/14: This station was located at (37.488501, -122.231061).

Station 25: From 9/1/14 - 10/22/14: This station was located at (37.486725, -122.225551). It was previously named "Broadway at Main."

Station 49: From 9/1/14 - 2/5/15: This station was located at (37.789625, -122.390264).

Station 69: From 9/1/14 - 3/11/15: This station was located at (37.776377,-122.39607).

Station 72: Moved twice. From 9/1/14 - 2/12/15, this station was located at (37.780356, -122.412919). From 2/13/15 to 6/3/15, the station was located at (37.780353, -122.41226).

Station 80: On 9/1/14, this station changed names from "San Jose Government Center" to "Santa Clara County Civic Center." It did not move.

We'll take the station data and update the installation column.

Loading up the data

Load up our status data and station information:

```
station_data <-
  read.csv("C:/Users/Georgie/Desktop/BABS_Data/201508_station_data.csv")
station_df  <- data.frame(station_data)
station  <- dplyr::tbl_df(station_df)
```

We update all the dates prior to system launch to the system launch date. Then convert the data to date format

```
station <-station %>%
          mutate(
            installation = as.Date(installation, format ="%m/%d/%Y")
                ) %>%
          mutate(
            installation =
            ifelse(installation < "2013-08-29","2013-08-29",as.character(installation))
                ) %>%
          mutate(installation =as.Date(installation))

glimpse(station)
```

```
## Observations: 70
## Variables: 7
## $ station_id   <int> 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 2...
## $ name         <fctr> San Jose Diridon Caltrain Station, San Jose Civi...
## $ lat          <dbl> 37.32973, 37.33070, 37.33399, 37.33141, 37.33672,...
## $ long         <dbl> -121.9018, -121.8890, -121.8949, -121.8932, -121....
## $ dockcount    <int> 27, 15, 11, 19, 15, 15, 15, 15, 15, 19, 19, 15, 1...
## $ landmark     <fctr> San Jose, San Jose, San Jose, San Jose, San Jose...
## $ installation <date> 2013-08-29, 2013-08-29, 2013-08-29, 2013-08-29, ...
```

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