Analysing the BABS data

Georgie Knight

20/08/2016

Dplyr and tidyr

Load the packages and the data. We'll now use the *readr* package so that we don't have to convert the dates and times.

```
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")
library("lubridate")
```

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

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

```
library("readr")
library("ggplot2")
library("ggmap")
trip_read    <- read_csv("trip_full_updated.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Start_trip = col_datetime(format = ""),
##   Start.Station = col_character(),
##   End_trip = col_datetime(format = ""),
##   End.Station = col_character(),
##   Subscriber.Type = col_character(),
##   Zip.Code = col_character(),
##   Date = col_date(format = ""),
##   landmark = col_character(),
```

```
##   start_lat = col_double(),
##   start_long = col_double(),
##   end_lat = col_double(),
##   end_long = col_double(),
##   Max.Sea.Level.PressureIn = col_double(),
##   Mean.Sea.Level.PressureIn = col_double(),
##   Min.Sea.Level.PressureIn = col_double(),
##   Events = col_character()
## )

## See spec(...) for full column specifications.
```

```
status_read  <- read_csv("status_full_updated.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   time = col_datetime(format = ""),
##   name = col_character(),
##   lat = col_double(),
##   long = col_double(),
##   landmark = col_character(),
##   installation = col_date(format = ""),
##   Date = col_date(format = ""),
##   Max.Sea.Level.PressureIn = col_double(),
##   Mean.Sea.Level.PressureIn = col_double(),
##   Min.Sea.Level.PressureIn = col_double(),
##   PrecipitationIn = col_character(),
##   Events = col_character()
## )
## See spec(...) for full column specifications.
```

```
trip         <- dplyr::tbl_df(trip_read)
status       <- dplyr::tbl_df(status_read)

glimpse(trip)
```

```
## Observations: 354,152
## Variables: 40
## $ Trip.ID                  <int> 913460, 913459, 913455, 913454, 9134...
## $ Duration                 <int> 765, 1036, 307, 409, 789, 293, 896, ...
## $ Start_trip               <time> 2015-08-31 23:26:00, 2015-08-31 23:...
## $ Start.Station            <chr> "Harry Bridges Plaza (Ferry Building...
## $ Start.Terminal           <int> 50, 31, 47, 10, 51, 68, 51, 60, 56, ...
## $ End_trip                 <time> 2015-08-31 23:39:00, 2015-08-31 23:...
## $ End.Station              <chr> "San Francisco Caltrain (Townsend at...
## $ End.Terminal             <int> 70, 27, 64, 8, 60, 70, 60, 74, 55, 6...
## $ Bike..                   <int> 288, 35, 468, 68, 487, 538, 363, 470...
## $ Subscriber.Type          <chr> "Subscriber", "Subscriber", "Subscri...
## $ Zip.Code                 <chr> "2139", "95032", "94107", "95113", "...
## $ Date                     <date> 2015-08-31, 2015-08-31, 2015-08-31,...
## $ landmark                 <chr> "San Francisco", "Mountain View", "S...
```

```
## $ start_lat                <dbl> 37.79539, 37.40044, 37.78898, 37.337...
## $ start_long               <dbl> -122.3942, -122.1083, -122.4035, -12...
## $ end_lat                  <dbl> 37.77662, 37.38922, 37.78226, 37.330...
## $ end_long                 <dbl> -122.3953, -122.0819, -122.3927, -12...
## $ Max.TemperatureF         <int> 78, 82, 78, 85, 78, 78, 78, 78, 78, ...
## $ Mean.TemperatureF        <int> 69, 72, 69, 72, 69, 69, 69, 69, 69, ...
## $ Min.TemperatureF         <int> 60, 61, 60, 59, 60, 60, 60, 60, 60, ...
## $ Max.Dew.PointF           <int> 58, 62, 58, 59, 58, 58, 58, 58, 58, ...
## $ MeanDew.PointF           <int> 57, 56, 57, 55, 57, 57, 57, 57, 57, ...
## $ Min.DewpointF            <int> 54, 52, 54, 51, 54, 54, 54, 54, 54, ...
## $ Max.Humidity             <int> 84, 84, 84, 84, 84, 84, 84, 84, 84, ...
## $ Mean.Humidity            <int> 67, 63, 67, 58, 67, 67, 67, 67, 67, ...
## $ Min.Humidity             <int> 50, 42, 50, 32, 50, 50, 50, 50, 50, ...
## $ Max.Sea.Level.PressureIn <dbl> 29.95, 29.97, 29.95, 29.95, 29.95, 2...
## $ Mean.Sea.Level.PressureIn <dbl> 29.91, 29.92, 29.91, 29.90, 29.91, 2...
## $ Min.Sea.Level.PressureIn <dbl> 29.87, 29.86, 29.87, 29.85, 29.87, 2...
## $ Max.VisibilityMiles      <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Mean.VisibilityMiles     <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Min.VisibilityMiles      <int> 9, 10, 9, 10, 9, 9, 9, 9, 9, 9, 9, 9...
## $ Max.Wind.SpeedMPH        <int> 18, 22, 18, 20, 18, 18, 18, 18, 18, ...
## $ Mean.Wind.SpeedMPH       <int> 9, 6, 9, 6, 9, 9, 9, 9, 9, 9, 9, 9, ...
## $ Max.Gust.SpeedMPH        <int> 21, 25, 21, 24, 21, 21, 21, 21, 21, ...
## $ PrecipitationIn          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ CloudCover               <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ Events                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ WindDirDegrees           <int> 246, 6, 246, 308, 246, 246, 246, 246...
## $ Zip                      <int> 94107, 94041, 94107, 95113, 94107, 9...
```

```
glimpse(status)
```

```
## Observations: 1,135,974
## Variables: 33
## $ station_id               <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ bikes_available          <int> 15, 14, 15, 14, 15, 14, 15, 14, 15, ...
## $ docks_available          <int> 12, 13, 12, 13, 12, 13, 12, 13, 12, ...
## $ time                     <time> 2014-09-01 00:00:03, 2014-09-01 02:...
## $ name                     <chr> "San Jose Diridon Caltrain Station",...
## $ lat                      <dbl> 37.32973, 37.32973, 37.32973, 37.329...
## $ long                     <dbl> -121.9018, -121.9018, -121.9018, -12...
## $ landmark                 <chr> "San Jose", "San Jose", "San Jose", ...
## $ installation             <date> 2013-08-29, 2013-08-29, 2013-08-29,...
## $ Date                     <date> 2014-09-01, 2014-09-01, 2014-09-01,...
## $ Max.TemperatureF         <int> 86, 86, 86, 86, 86, 86, 86, 86, 86, ...
## $ Mean.TemperatureF        <int> 72, 72, 72, 72, 72, 72, 72, 72, 72, ...
## $ Min.TemperatureF         <int> 58, 58, 58, 58, 58, 58, 58, 58, 58, ...
## $ Max.Dew.PointF           <int> 60, 60, 60, 60, 60, 60, 60, 60, 60, ...
## $ MeanDew.PointF           <int> 54, 54, 54, 54, 54, 54, 54, 54, 54, ...
## $ Min.DewpointF            <int> 50, 50, 50, 50, 50, 50, 50, 50, 50, ...
## $ Max.Humidity             <int> 86, 86, 86, 86, 86, 86, 86, 86, 86, ...
## $ Mean.Humidity            <int> 59, 59, 59, 59, 59, 59, 59, 59, 59, ...
## $ Min.Humidity             <int> 31, 31, 31, 31, 31, 31, 31, 31, 31, ...
## $ Max.Sea.Level.PressureIn <dbl> 29.86, 29.86, 29.86, 29.86, 29.86, 2...
## $ Mean.Sea.Level.PressureIn <dbl> 29.81, 29.81, 29.81, 29.81, 29.81, 2...
## $ Min.Sea.Level.PressureIn <dbl> 29.75, 29.75, 29.75, 29.75, 29.75, 2...
```

```
## $ Max.VisibilityMiles       <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Mean.VisibilityMiles      <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Min.VisibilityMiles       <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Max.Wind.SpeedMPH         <int> 17, 17, 17, 17, 17, 17, 17, 17, 17, ...
## $ Mean.Wind.SpeedMPH        <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...
## $ Max.Gust.SpeedMPH         <int> 22, 22, 22, 22, 22, 22, 22, 22, 22, ...
## $ PrecipitationIn           <chr> "0", "0", "0", "0", "0", "0", "0", "...
## $ CloudCover                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Events                    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ WindDirDegrees            <int> 296, 296, 296, 296, 296, 296, 296, 2...
## $ Zip                       <int> 95113, 95113, 95113, 95113, 95113, 9...
```

So now we have our data ready we can play with it.

```
trip %>% select(Duration) %>% summary()
```

```
##     Duration
##  Min.   :      60
##  1st Qu.:     342
##  Median :     511
##  Mean   :    1046
##  3rd Qu.:     739
##  Max.   :17270400
```

We see that the average trip length is $1046/60 = 17.4$ minutes, the minimum trip length was 1 minute and the max was nearly 200 days. That may be a mistake... Does this vary by location?

```
distinct(trip, landmark)
```

```
## # A tibble: 5 x 1
##        landmark
##           <chr>
## 1 San Francisco
## 2 Mountain View
## 3      San Jose
## 4     Palo Alto
## 5  Redwood City
```

```
trip %>% filter(landmark == "San Francisco") %>%
        select(Duration) %>%
        summary()
```

```
##     Duration
##  Min.   :      60
##  1st Qu.:     352
##  Median :     519
##  Mean   :     976
##  3rd Qu.:     740
##  Max.   :17270400
```

```
trip %>% filter(landmark == "Mountain View") %>%
        select(Duration) %>%
        summary()
```

```
##     Duration
##  Min.   :     61
##  1st Qu.:    238
##  Median :    294
##  Mean   :   1430
##  3rd Qu.:    457
##  Max.   :1852590
```

```
trip %>% filter(landmark == "San Jose") %>%
        select(Duration) %>%
        summary()
```

```
##     Duration
##  Min.   :    62
##  1st Qu.:   312
##  Median :   466
##  Mean   :  1401
##  3rd Qu.:   663
##  Max.   :655939
```

```
trip %>% filter(landmark == "Palo Alto") %>%
        select(Duration) %>%
        summary()
```

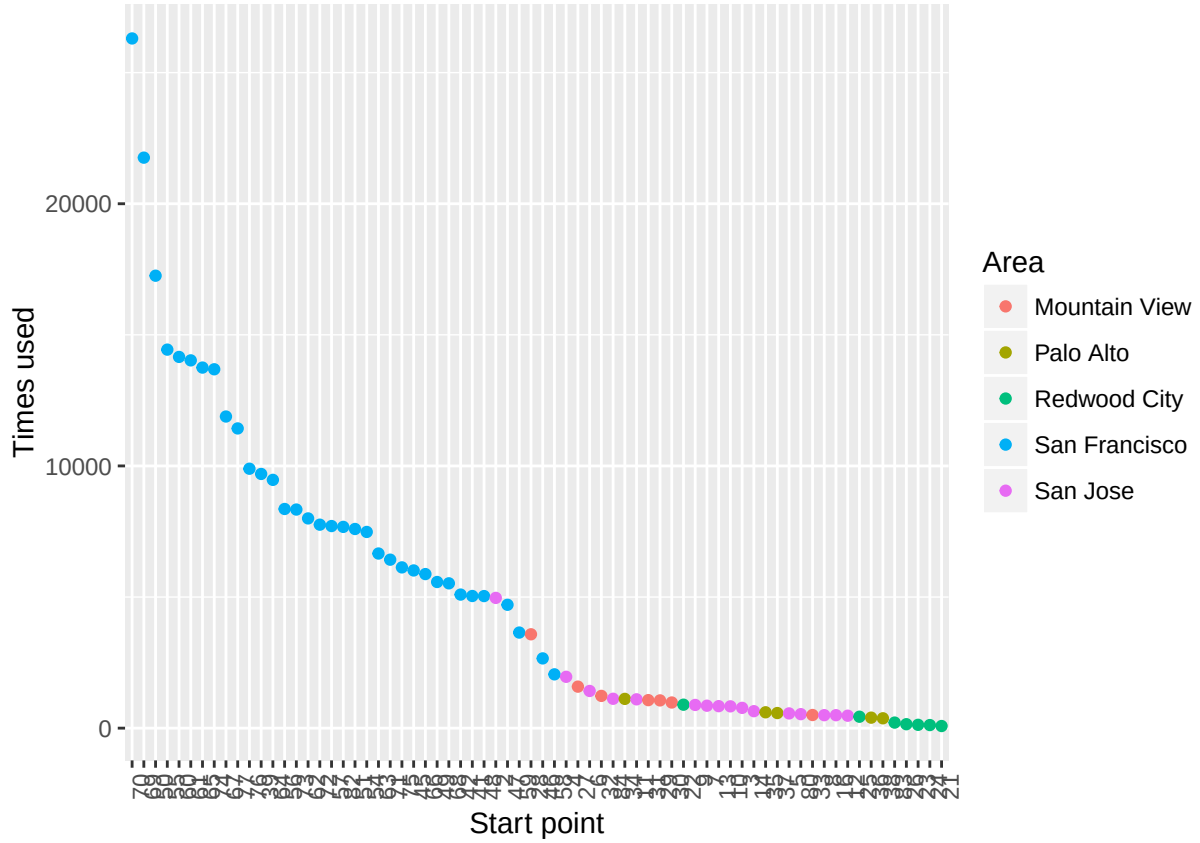
```
##     Duration
##  Min.   :     66
##  1st Qu.:    288
##  Median :    863
##  Mean   :   4230
##  3rd Qu.:   2018
##  Max.   :1133540
```

```
trip %>% filter(landmark == "Redwood City") %>%
        select(Duration) %>%
        summary()
```

```
##     Duration
##  Min.   :    68.0
##  1st Qu.:   273.5
##  Median :   621.0
##  Mean   :  2287.6
##  3rd Qu.:   863.5
##  Max.   :720454.0
```

We can find the most used starting points

```
Count_starting_point <- trip %>%
  count(Start.Terminal) %>%
  arrange(desc(n))

Terminal_landmark <- trip %>%
  select(Start.Terminal, landmark,  start_lat, start_long) %>%
  distinct( Start.Terminal, .keep_all = TRUE)

Count_starting_point <-
left_join(Count_starting_point , Terminal_landmark, by= c("Start.Terminal"= "Start.Terminal"))

ggplot(Count_starting_point, aes(x = reorder(Start.Terminal, -n), y = n,
                                 col = as.factor(landmark))) +
  geom_point()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(col='Area', x = "Start point", y = "Times used")
```



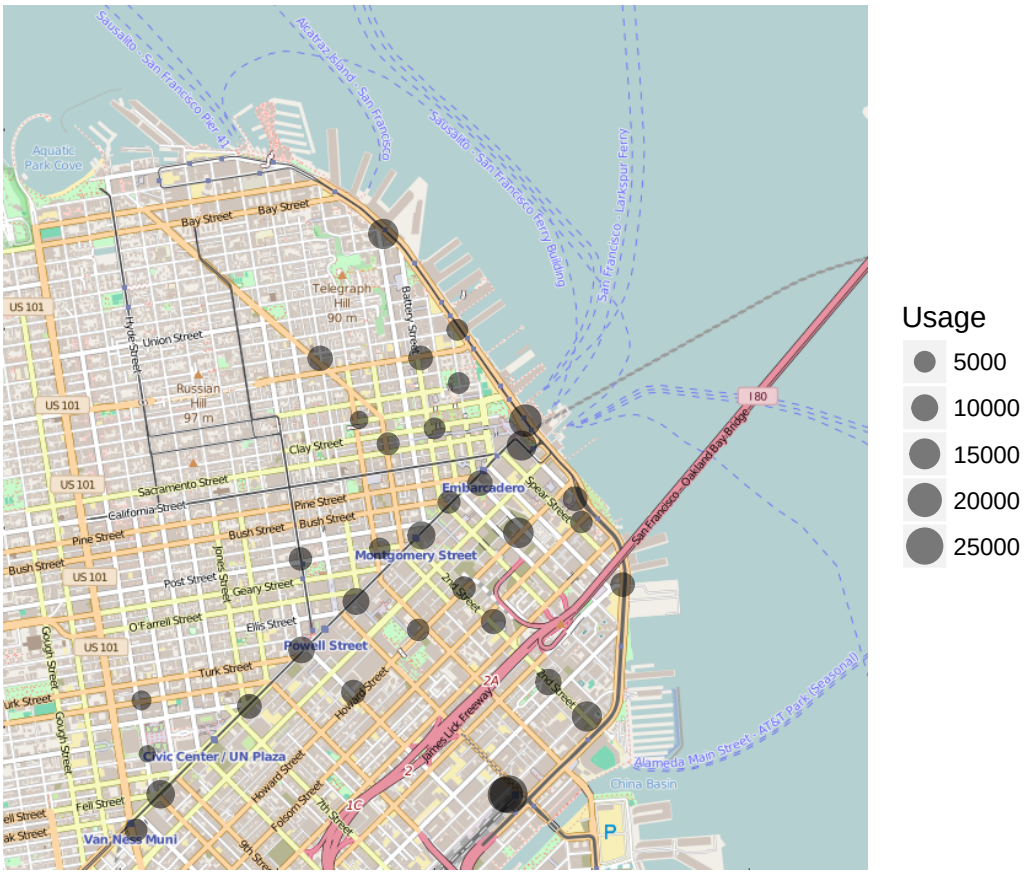

```
San_Fran_map <-qmap(
  location = 'san francisco financial district',
  source = "osm",
  zoom =14)
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=san+francisco+financial+district&
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=san%20francisco%20fi
```

```
San_Fran_map +
  geom_point(data =Count_starting_point, aes( y= start_lat, x =start_long, size = n), alpha = 0.5)+
  labs(size = 'Usage')
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```



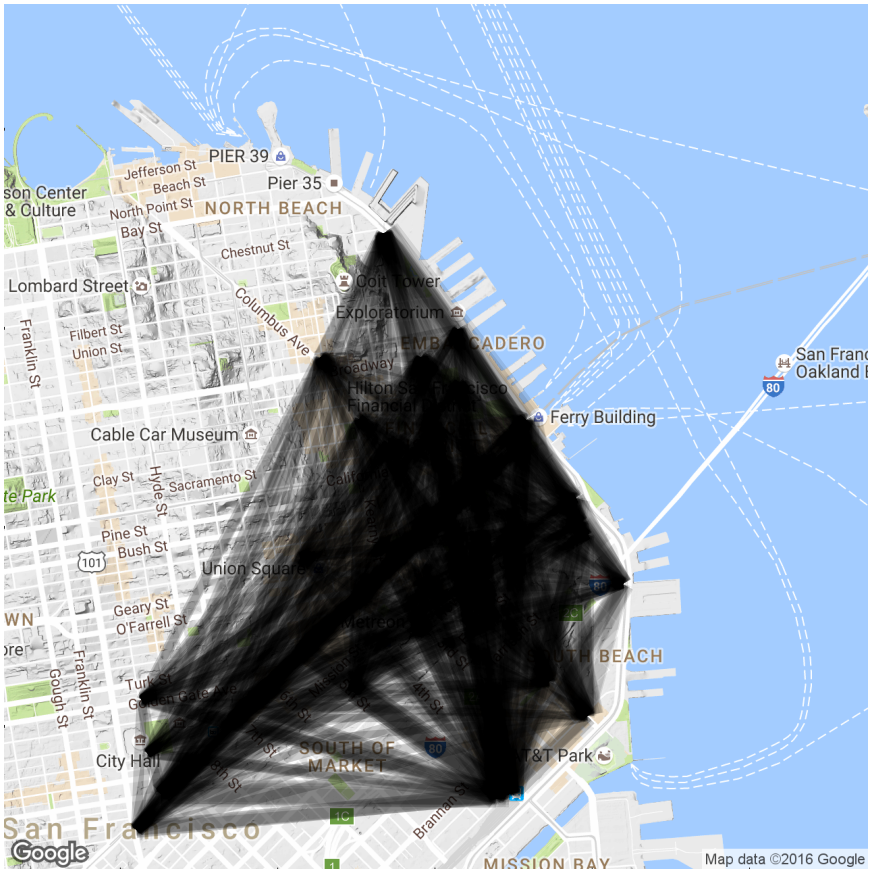

```
San_Fran_trips <-trip %>%
  filter(landmark == "San Francisco") %>%
  select(Start.Station, End.Station, start_lat, start_long,end_lat, end_long) %>%
  count(Start.Station, End.Station,start_lat, start_long, end_lat, end_long)

San_Fran_map <-qmap(location = 'san francisco financial district', zoom =14)
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=san+francisco+financial+district&
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=san%20francisco%20fi
```

```
San_Fran_map +
  geom_segment(data = San_Fran_trips, aes(y= start_lat, x = start_long, yend =  end_lat, xend = end_long
  theme(legend.position="none")
```

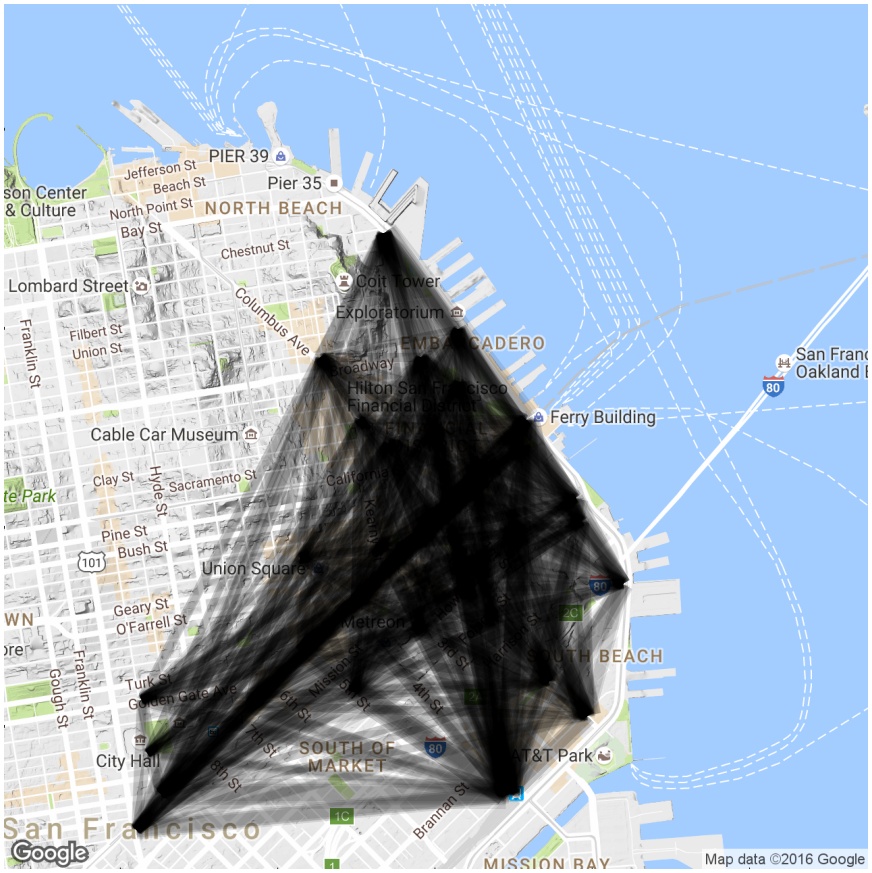
```
## Warning: Removed 79 rows containing missing values (geom_segment).
```



Let's look at all the trips in a given week

```
San_Fran_trip_select <-trip %>%
  filter(landmark == "San Francisco") %>%
  filter(between(Date, as.Date("2015-06-01"), as.Date("2015-06-30"))) %>%
  select(Start.Station, End.Station, start_lat, start_long,end_lat, end_long) %>%
  count(Start.Station, End.Station,start_lat, start_long, end_lat, end_long)


San_Fran_map +
  geom_segment(data = San_Fran_trip_select, aes(y= start_lat, x = start_long, yend =  end_lat, xend = en
  theme(legend.position="none")
```

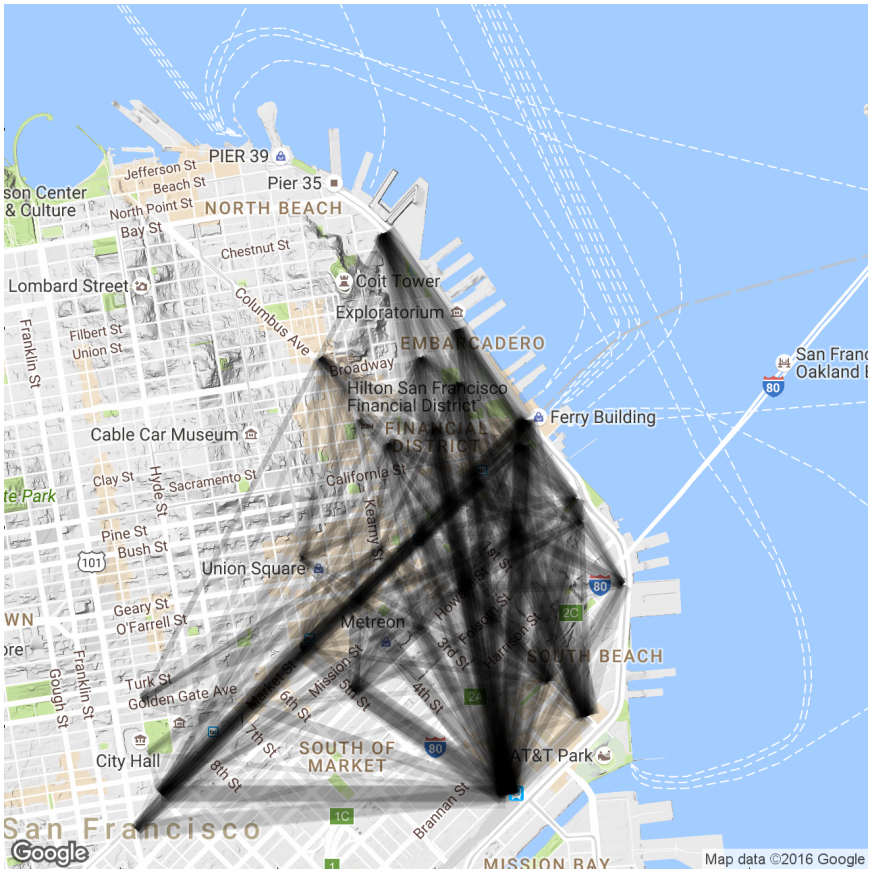
```
## Warning: Removed 69 rows containing missing values (geom_segment).
```



Or on a particular day

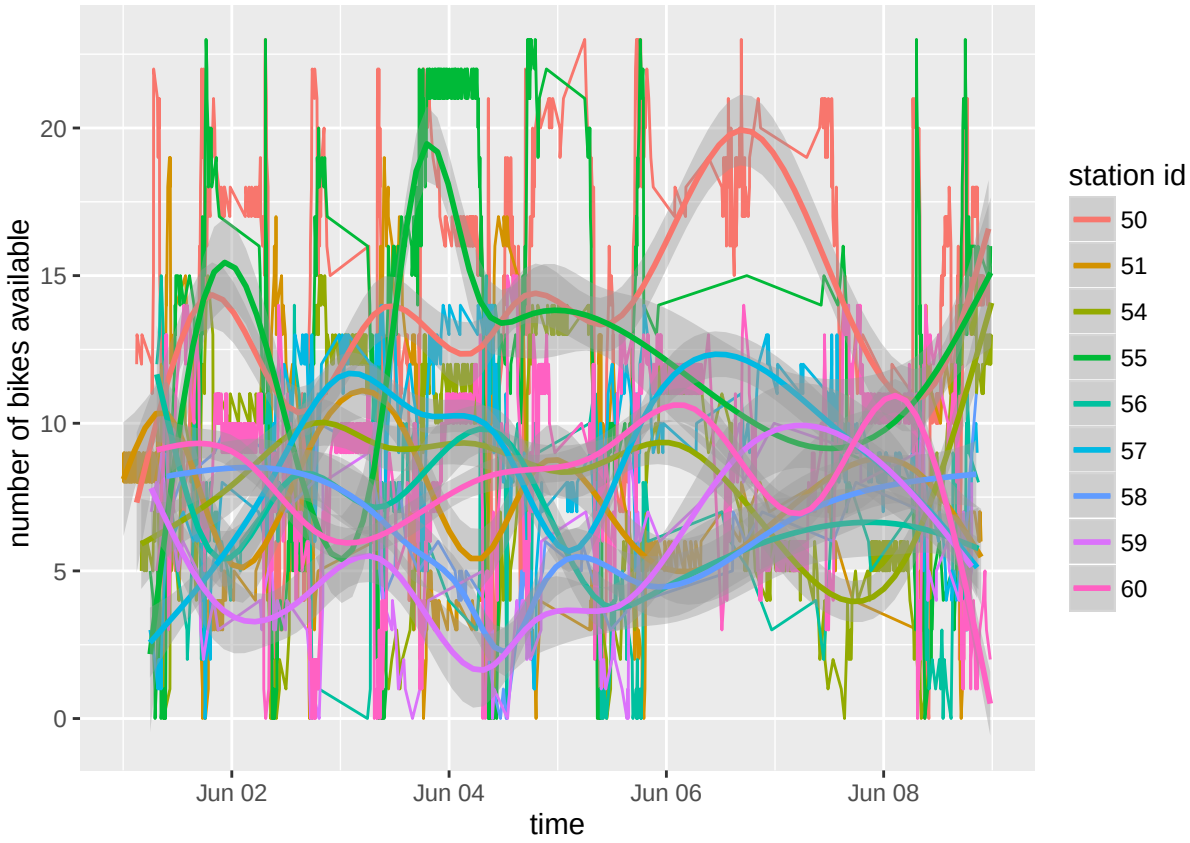
```
San_Fran_trip_select <-trip %>%
  filter(landmark == "San Francisco") %>%
  filter(Date == as.Date("2015-06-02")) %>%
  select(Start.Station, End.Station, start_lat, start_long,end_lat, end_long) %>%
  count(Start.Station, End.Station,start_lat, start_long, end_lat, end_long)

San_Fran_map +
  geom_segment(data = San_Fran_trip_select, aes(y= start_lat, x = start_long, yend =  end_lat, xend = en
  theme(legend.position="none")
```

```
## Warning: Removed 43 rows containing missing values (geom_segment).
```



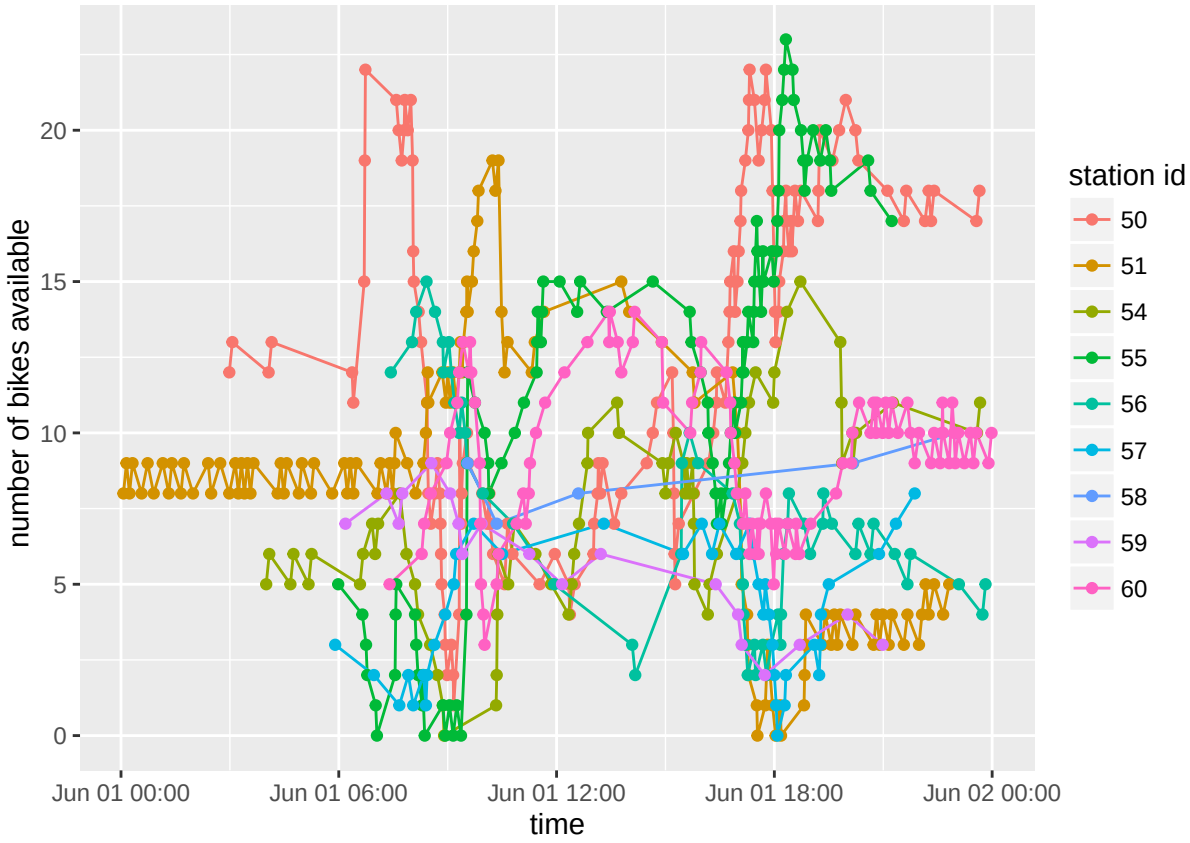
Let's take a look at how bikes available varies in some select stations in San Francisco in a particular week.

```
status_sf_select <- status %>%
  filter(landmark == "San Francisco") %>%
  filter(between(station_id, 50, 60)) %>%
  filter(between(Date, as.Date("2015-06-01"), as.Date("2015-06-08")))

ggplot(status_sf_select, aes(x= time, y = bikes_available,
                  col = as.factor(station_id)))+
  geom_line()+
  geom_smooth()+
  labs(col = 'station id', x = 'time', y = ' number of bikes available' )
```



Or on a particular day

```
status_sf_select <- status %>%
  filter(landmark == "San Francisco") %>%
  filter(between(station_id, 50, 60)) %>%
  filter(Date == as.Date("2015-06-01"))

ggplot(status_sf_select, aes(x= time, y = bikes_available,
                   col = as.factor(station_id)))+
  geom_line()+
  geom_point()+
  labs(col = 'station id', x = 'time', y = ' number of bikes available' )
```


number of bikes available
20
15
10
5
0
Jun 01 00:00
Jun 01 06:00
Jun 01 12:00
Jun 01 18:00
Jun 02 00:00
time
station id
50
51
54
55
56
57
58
59
60