Analyse_5: Docking stations:present and future.

Georgie Knight

20 September, 2016

```
library("dplyr")
library("tidyr")
library("lubridate")
library("readr")
library("ggplot2")
library("ggmap")
library("igraph")
library("popgraph")
library("zipcode")
trip_read    <- read_csv("trip_full_updated2.csv")
status_read  <- read_csv("status_full_updated.csv")
trip         <- dplyr::tbl_df(trip_read)
status       <- dplyr::tbl_df(status_read)
station <-read_csv("201508_station_data.csv")
station <- dplyr::tbl_df(station)
station$name[36] = "Washington at Kearny" #correct the misspellings!
station$name[37] = "Post at Kearny"
```

Introduction

We're going to look at the docking stations. When are they empty? When are they full? Which ones have an imbalance (more bikes taken than deposited) ? We will then look at external data which to suggest locations for other stations.

Zipcodes

We firstly want to take another look at the zip codes of users to get an idea of where the users are coming from

```
data("zipcode")
zips <- trip %>%
  group_by(Zip.Code) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  mutate(zip = clean.zipcodes(Zip.Code)) %>%
  merge(zipcode, by.x='zip', by.y='zip') %>%
  arrange(desc(count))

head(zips)
```

```
##     zip Zip.Code count          city state latitude longitude
## 1 94107    94107 46622 San Francisco    CA 37.76653 -122.3958
## 2 94105    94105 20311 San Francisco    CA 37.78923 -122.3957
## 3 94133    94133 16246 San Francisco    CA 37.80188 -122.4102
```

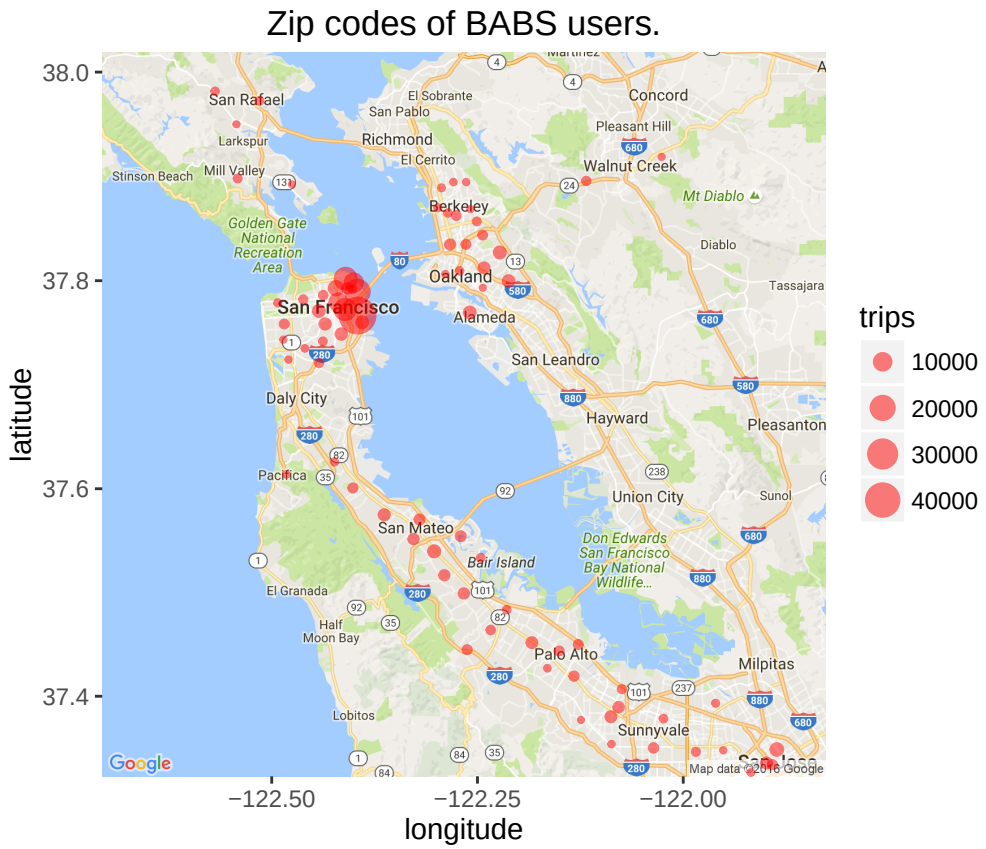
```
## 4 94103    94103 14959 San Francisco    CA 37.77233 -122.4109
## 5 94111    94111 10960 San Francisco    CA 37.79823 -122.4003
## 6 94102    94102 10150 San Francisco    CA 37.77933 -122.4192
```

```
unique(zips$state)
```

```
##  [1] "CA" "AE" "MA" "OR" "NV" "NY" "IL" "WA" "DC" "UT" "SC" "CO" "FL" "MD"
## [15] "AZ" "PA" "VA" "NJ" "WI" "TX" "GA" "MN" "NC" "MT" "HI" "MI" "ID" "TN"
## [29] "IN" "WY" "NE" "PR" "LA" "CT" "SD" "OH" "MO" "IA" "ME" "AL" "KS" "OK"
## [43] "AK" "RI" "NH" "KY" "NM" "AR" "WV" "VT" "ND" "DE" "MS"
```

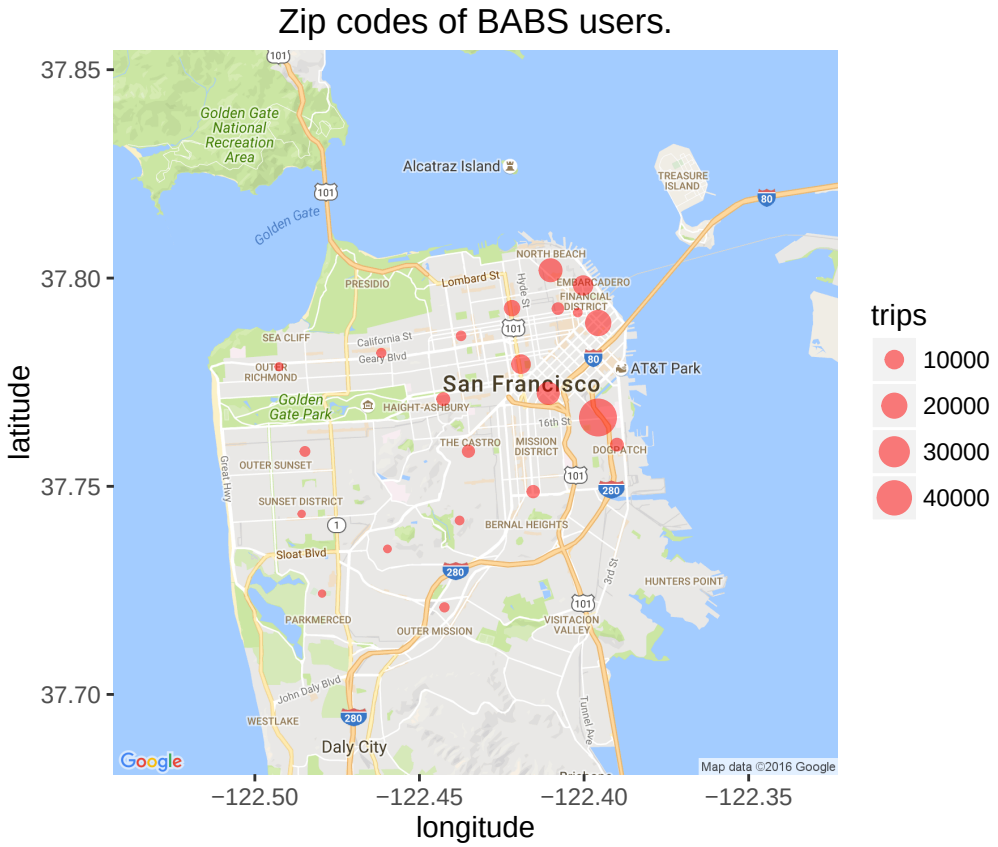
Let's look at the zipcodes which contribute more than 750 trips.

```
zips2 <- filter(zips, count>750)
location <- c( mean(zips2$longitude, na.rm = TRUE), mean(zips2$latitude,na.rm = TRUE))
BABSmap <- get_map(location,maptype ="roadmap", zoom=10)
ggmap(BABSmap)+
  geom_point(data = zips2,
             aes(x=longitude, y=latitude, size = count),
             alpha = 0.5, color ="red")+
  labs(x = "longitude", y = "latitude", size ="trips",
       title = "Zip codes of BABS users.")+scale_size_area()
```



We see there are large contributions from Oakland and all along the caltrain line from san jose to san francsico. In San Francisco itself:

```
zipsCA <- filter(zips, count>750, city == "San Francisco")
location <- c( mean(zipsCA$longitude, na.rm = TRUE), mean(zipsCA$latitude,na.rm = TRUE))
BABSmap <- get_map(location,maptype ="roadmap", zoom=12)
ggmap(BABSmap)+
  geom_point(data = zipsCA,
             aes(x=longitude, y=latitude, size = count),
             alpha = 0.5, color ="red")+
  labs(x = "longitude", y = "latitude", size ="trips",
       title = "Zip codes of BABS users.")+scale_size_area()
```



Interestingly there are many users in the West and the North where no bike stations are located. Perhaps new bike stations here will be used as there is already a base of users living here. Which stations do they use? The one in the north is 94133

```
trip94133 <- trip %>% filter(Zip.Code == 94133) %>%
  group_by(Start.Station) %>%
  summarise(count =n()) %>%
  ungroup() %>%
  arrange(desc(count))

trip94133
```

```
## # A tibble: 54 x 2
##                                Start.Station count
##                                        <chr> <int>
```

```
## 1           Grant Avenue at Columbus Avenue  5017
## 2                     Embarcadero at Sansome  1584
## 3                 Broadway St at Battery St   837
## 4  San Francisco Caltrain (Townsend at 4th)   806
## 5                      Washington at Kearny   628
## 6                         Market at Sansome   512
## 7                           Clay at Battery   415
## 8      Harry Bridges Plaza (Ferry Building)   401
## 9                     Embarcadero at Folsom   395
## 10                        Steuart at Market   381
## # ... with 44 more rows
```

```
tripsGA <- trip %>%
  filter(Zip.Code == 94133, Start.Station =="Grant Avenue at Columbus Avenue") %>%
  group_by(End.Station) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  arrange(desc(count))

tripsGA
```

```
## # A tibble: 35 x 2
##                                    End.Station count
##                                          <chr> <int>
## 1                            Market at Sansome   840
## 2       San Francisco Caltrain (Townsend at 4th)   533
## 3                              Beale at Market   386
## 4                      Commercial at Montgomery   374
## 5  Temporary Transbay Terminal (Howard at Beale)   346
## 6                         Embarcadero at Folsom   276
## 7                               2nd at Townsend   244
## 8                                 Market at 4th   240
## 9                            Powell Street BART   203
## 10                              Townsend at 7th   129
## # ... with 25 more rows
```

They overwhelmingly start at Grant Avenue at Columbus Avenue (37.79852 -122.4072) which is their nearest station, expanding further into this district would be welcome by these and potentially more users.

SF OpenData

Instead of only looking from within the data set, we can look outside by combining with other data sets about bike usage in San Francisco. Avilable from SF OpenData is information on bike volume manual counts.Which is described as

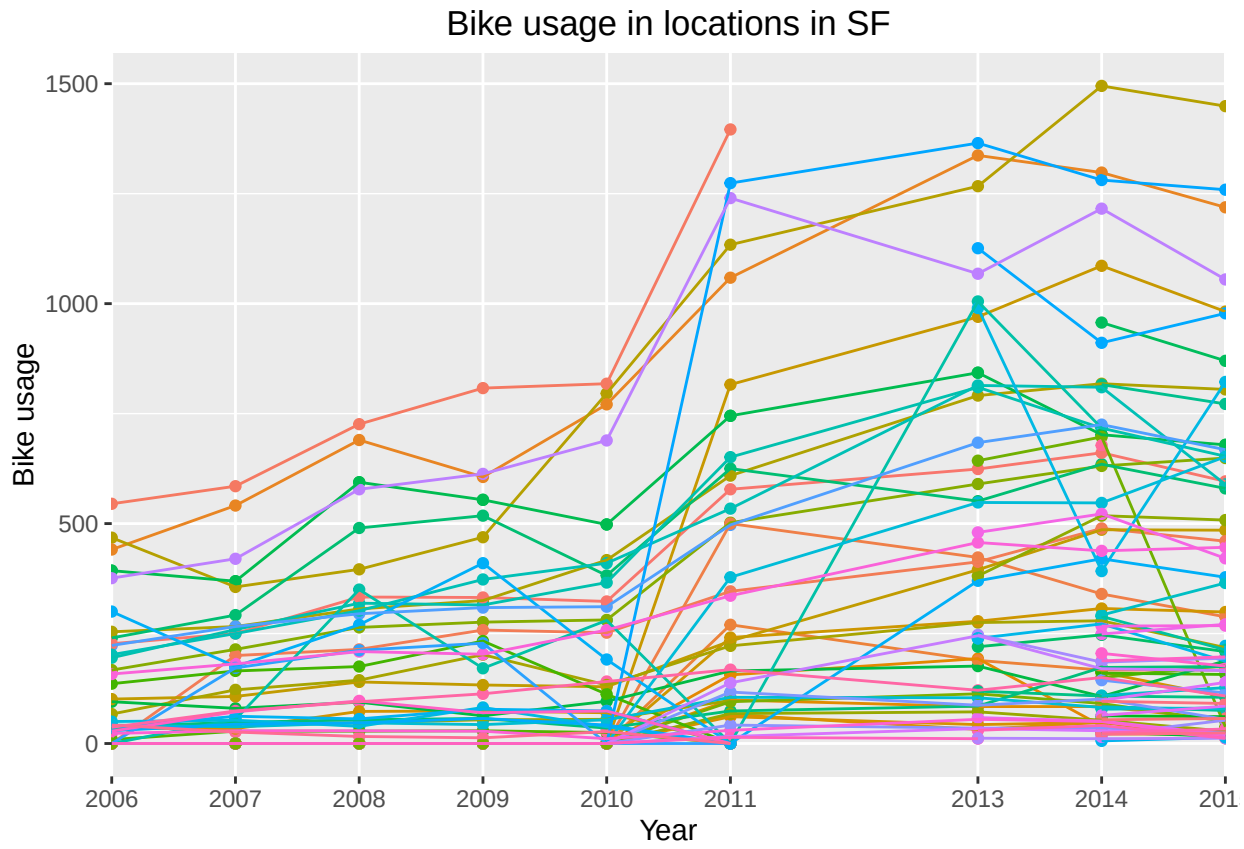
This dataset was created to show the bike counts from 2009-2014 by observation location (not including the year of 2012). This dataset is manually updated annually. Note: A bicycle count value of "-1" indicates a null value (bicycle count data was not collected at this location for that year). Bike counts for 2012 are not included in this file, as they are not available/unable to be located.

We combine it with the data for 2015 found here https://www.sfmta.com/sites/default/files/reports/2016/Annual%20Bicycle%20Count%20Report%202015_04152016.pdf. Let's take a look. We firstly load the files and tidy them up.

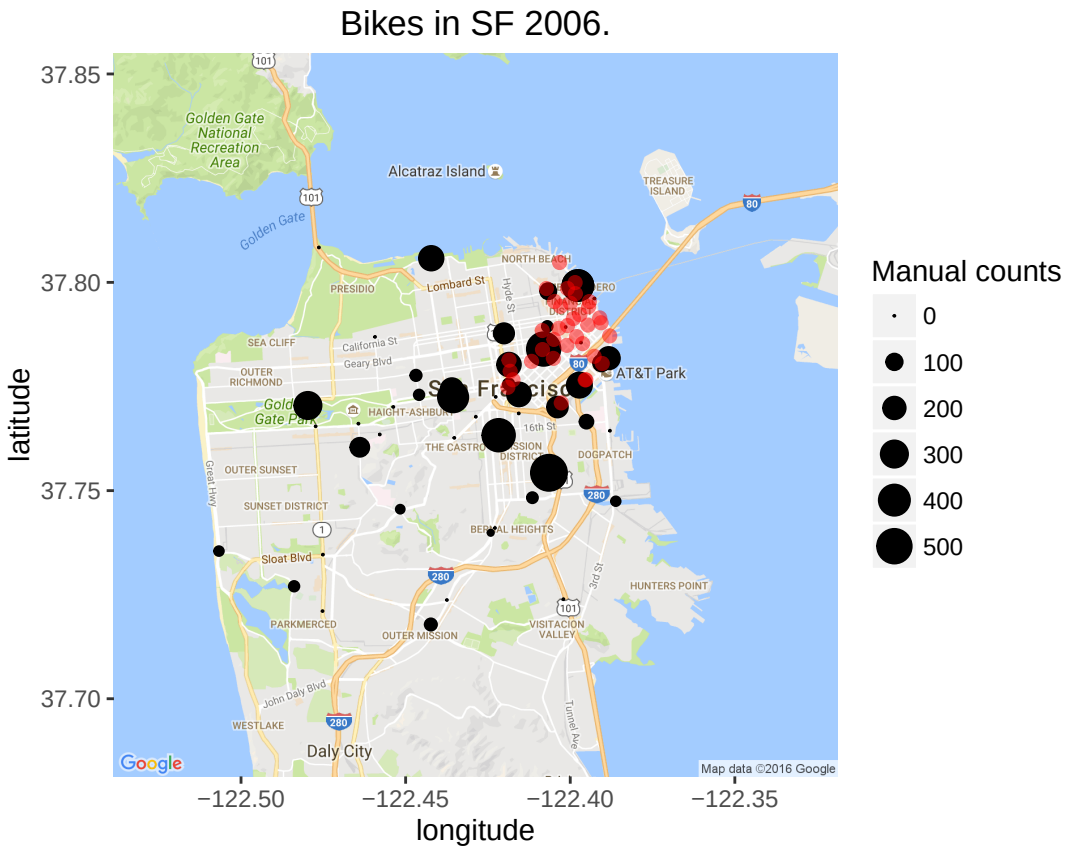
```
bikeCounts <-read_csv("Bike_Volume_Manual_Counts.csv")
bikeCount15 <-read.csv("2015bikecount.txt", sep = " ")
bikeCounts <- dplyr::tbl_df(bikeCounts)
bikeCount15 <- dplyr::tbl_df(bikeCount15)

names(bikeCounts) <- make.names(names(bikeCounts), unique=TRUE)
names(bikeCount15) <- c("ID","Locations","X2014","Bike.Count.2015.Afternoon"  )

bikeCounts <- arrange(bikeCounts, Location.ID)
bikeCount15 <- arrange(bikeCount15, ID)
bikeCounts <-left_join(bikeCounts, bikeCount15, by = c("Location.ID" = "ID"))

bikeCounts <- bikeCounts %>%
  separate(Geom, c("latitude", "longitude"), sep = ", ") %>%
  mutate(latitude =
           as.numeric(gsub("\\(|\\)", "",latitude)),
         longitude =
           as.numeric(gsub("\\(|\\)", "",longitude)))
names(bikeCounts) <- make.names(names(bikeCounts), unique=TRUE)
bikeCounts <- select(bikeCounts,c(2,3,5,6,7,8,9,10,11,12,22,18,19))
bikeCounts <- bikeCounts %>%
  mutate(Bike.Count.2006.Afternoon =
           ifelse(Bike.Count.2006.Afternoon<0, NA,Bike.Count.2006.Afternoon)) %>%
  mutate(Bike.Count.2007.Afternoon =
           ifelse(Bike.Count.2007.Afternoon<0, NA,Bike.Count.2007.Afternoon)) %>%
  mutate(Bike.Count.2008.Afternoon =
           ifelse(Bike.Count.2008.Afternoon<0, NA,Bike.Count.2008.Afternoon)) %>%
  mutate(Bike.Count.2009.Afternoon =
           ifelse(Bike.Count.2009.Afternoon<0, NA,Bike.Count.2009.Afternoon)) %>%
  mutate(Bike.Count.2010.Afternoon =
           ifelse(Bike.Count.2010.Afternoon<0, NA,Bike.Count.2010.Afternoon)) %>%
  mutate(Bike.Count.2011.Afternoon =
           ifelse(Bike.Count.2011.Afternoon<0, NA,Bike.Count.2011.Afternoon)) %>%
  mutate(Bike.Count.2013.Afternoon =
           ifelse(Bike.Count.2013.Afternoon<0, NA,Bike.Count.2013.Afternoon)) %>%
  mutate(Bike.Count.2014.Afternoon =
           ifelse(Bike.Count.2014.Afternoon<0, NA,Bike.Count.2014.Afternoon))
```

Let's visualise our data

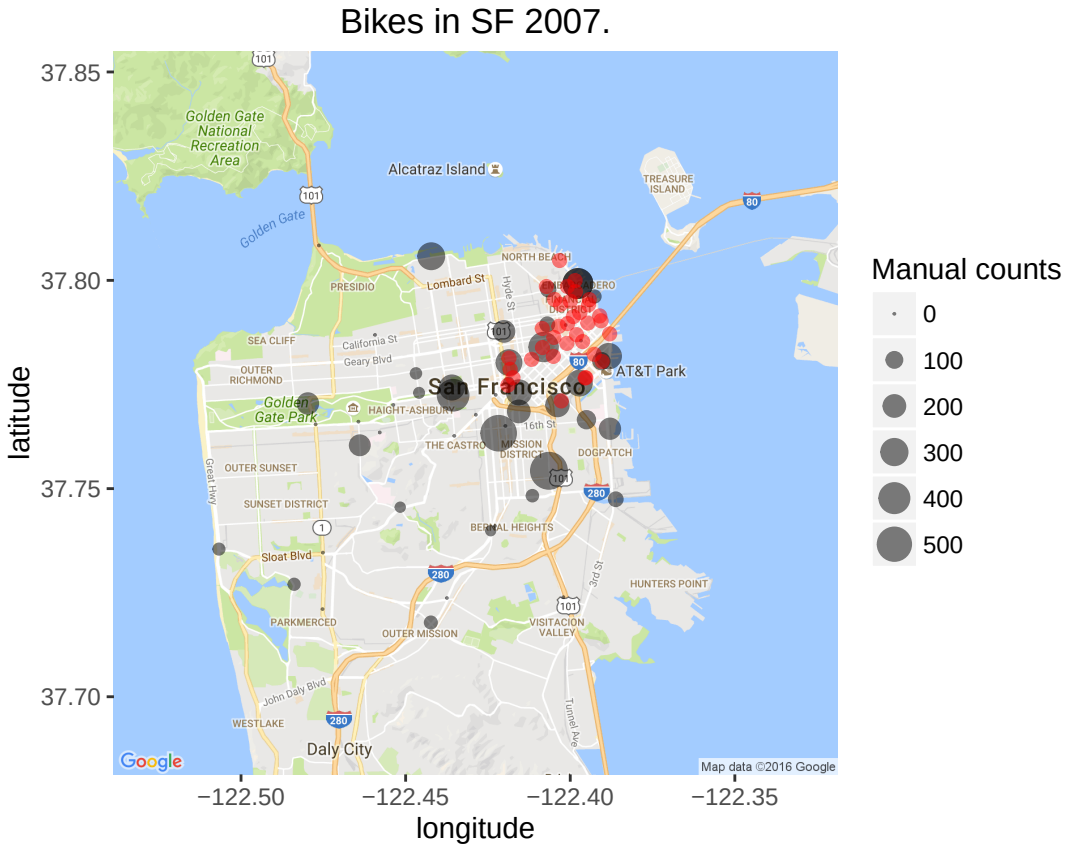
```
bikeCountsNew <- bikeCounts %>%
  gather(Year, value, 3:11 ) %>%
  mutate(Year = extract_numeric(gsub("\\.", "",Year)))

ggplot(bikeCountsNew, aes(x=Year, y = value, col = Location))+
  geom_point()+
  geom_line()+
  theme(legend.position="none")+
  labs(x="Year", y= "Bike usage", title = "Bike usage in locations in SF") +
  scale_x_discrete(limits = unique(bikeCountsNew$Year))
```



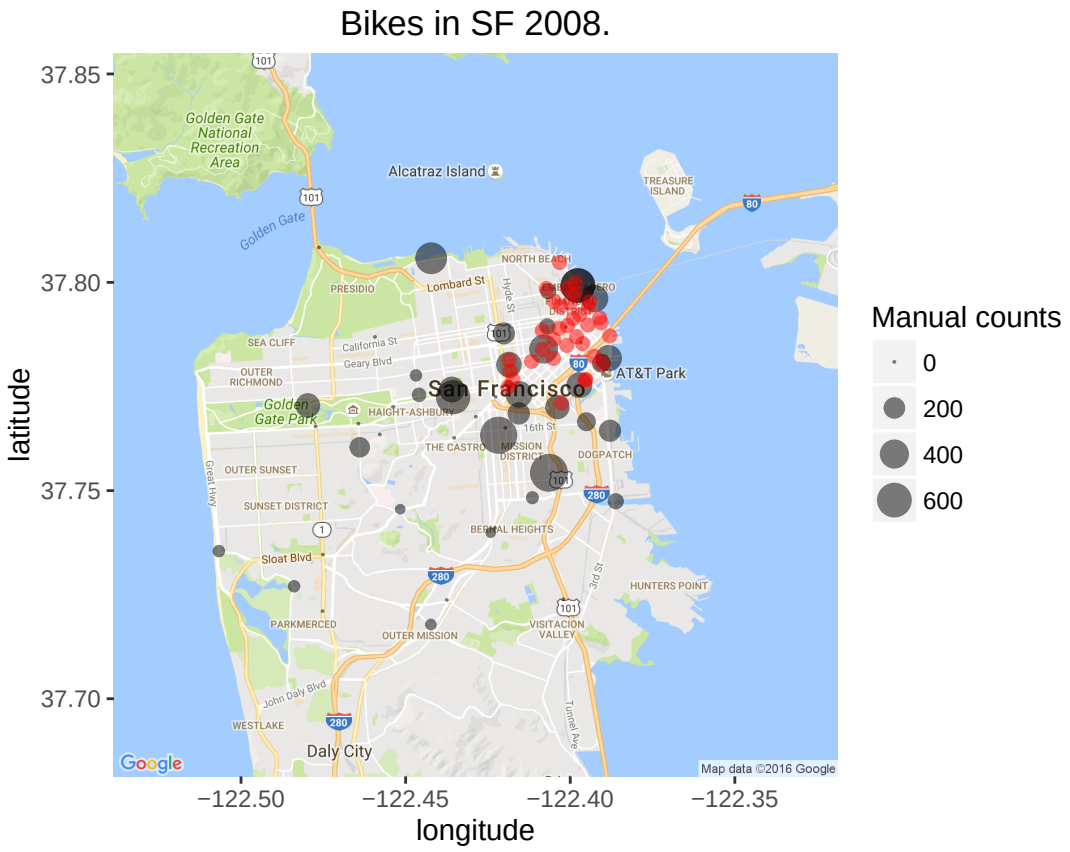

```
location <- c( mean(bikeCountsNew$longitude, na.rm = TRUE),
               mean(bikeCountsNew$latitude,na.rm = TRUE))
BABSmap <- get_map(location,maptype ="roadmap", zoom=12)

ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2006),
             aes(x=longitude, y=latitude, size = value), alpha = 1.0)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2006.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```



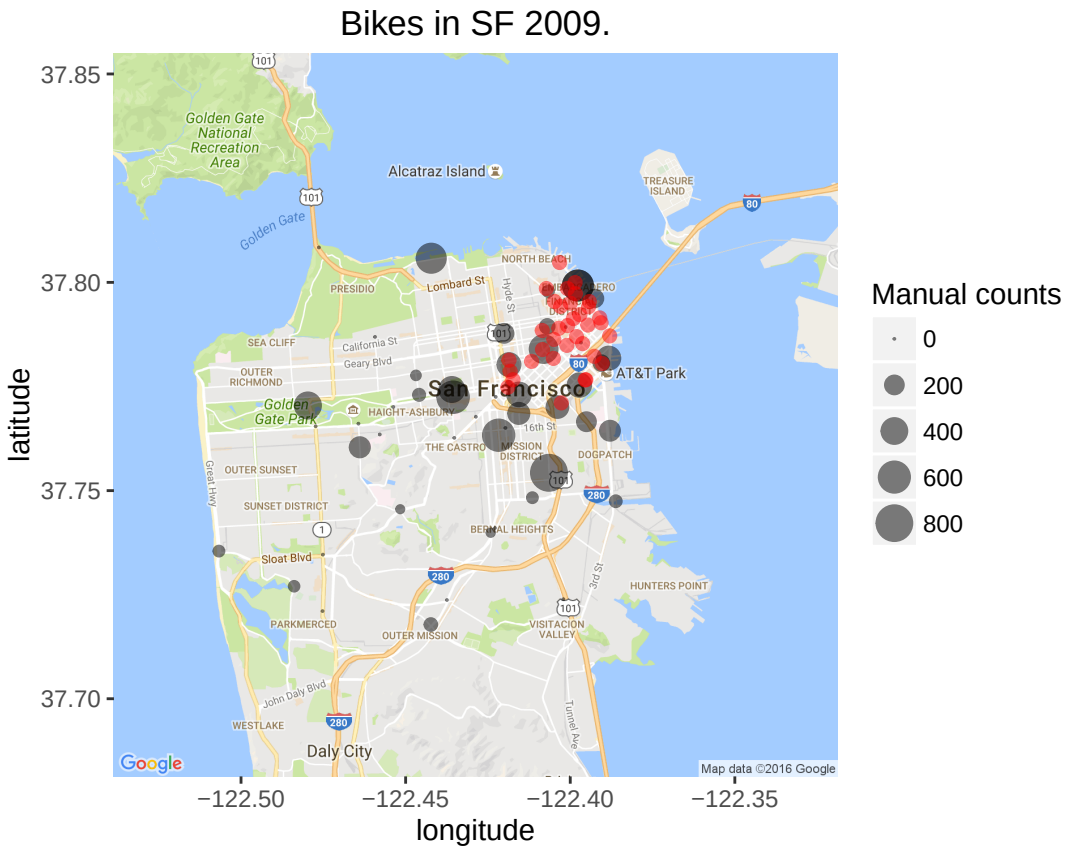

```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2007),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2007.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```



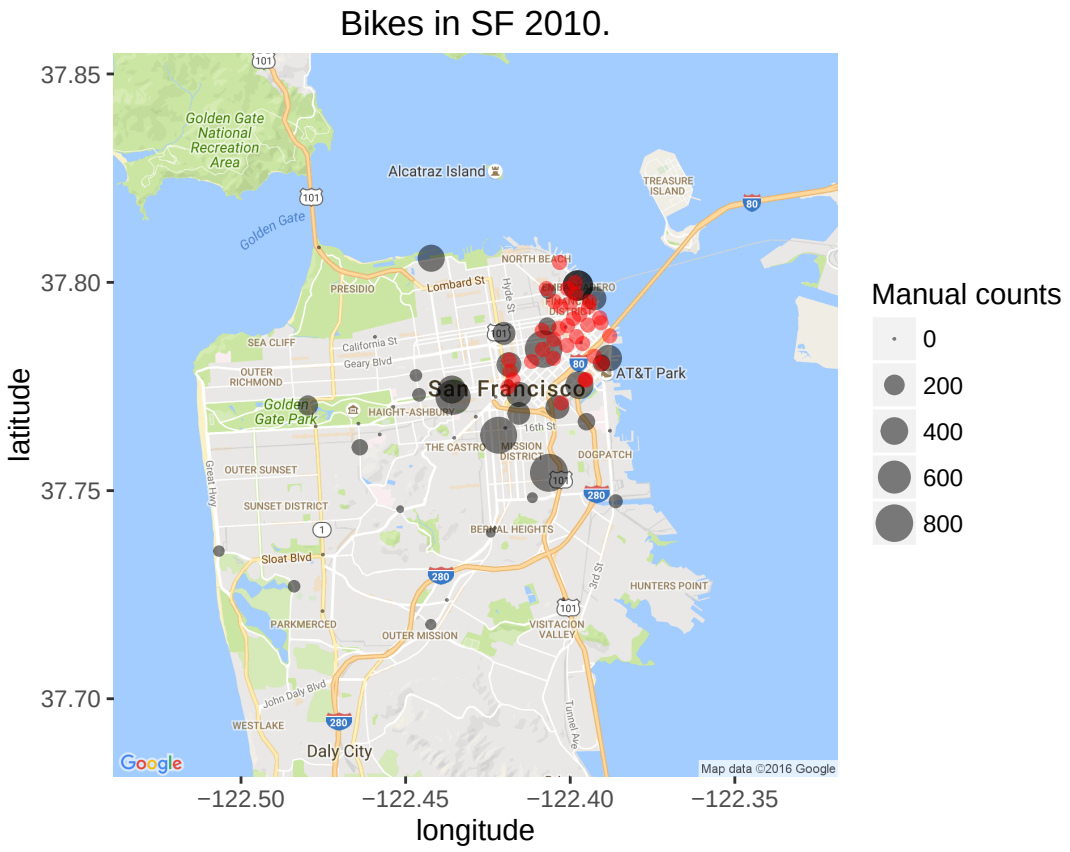

```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2008),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2008.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```



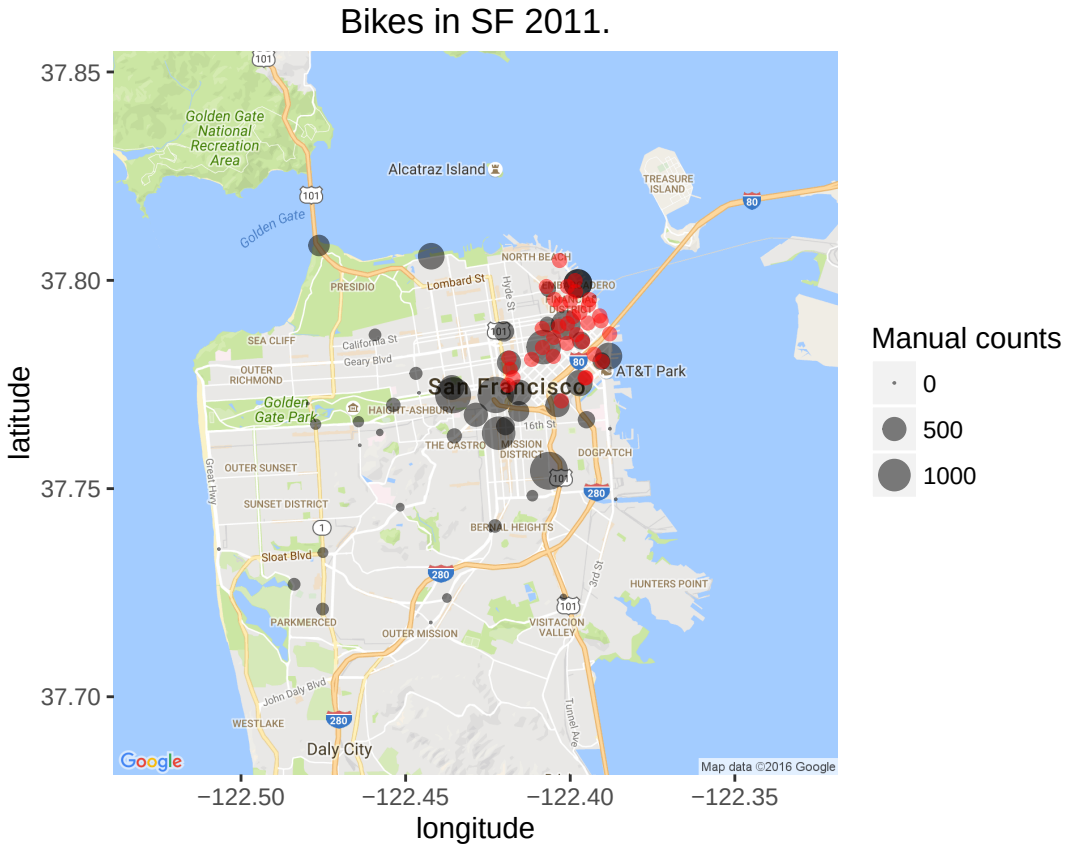

```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2009),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2009.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```



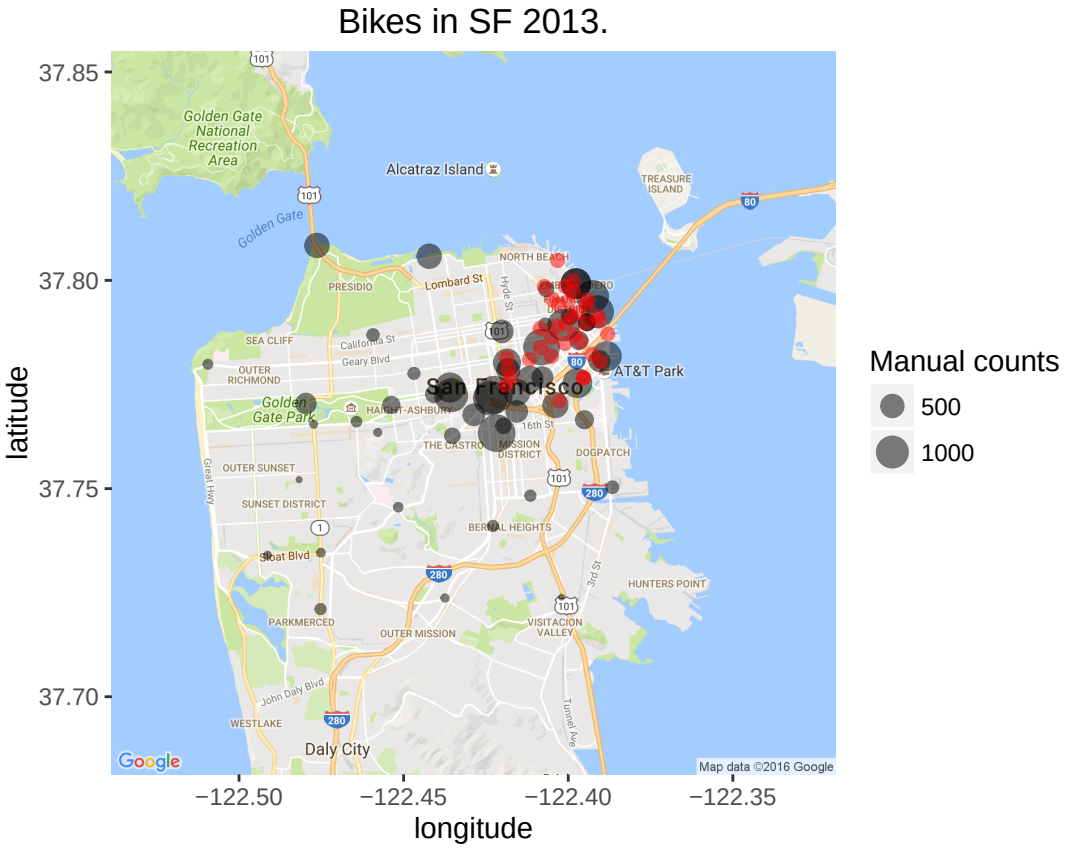

```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2010),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2010.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```



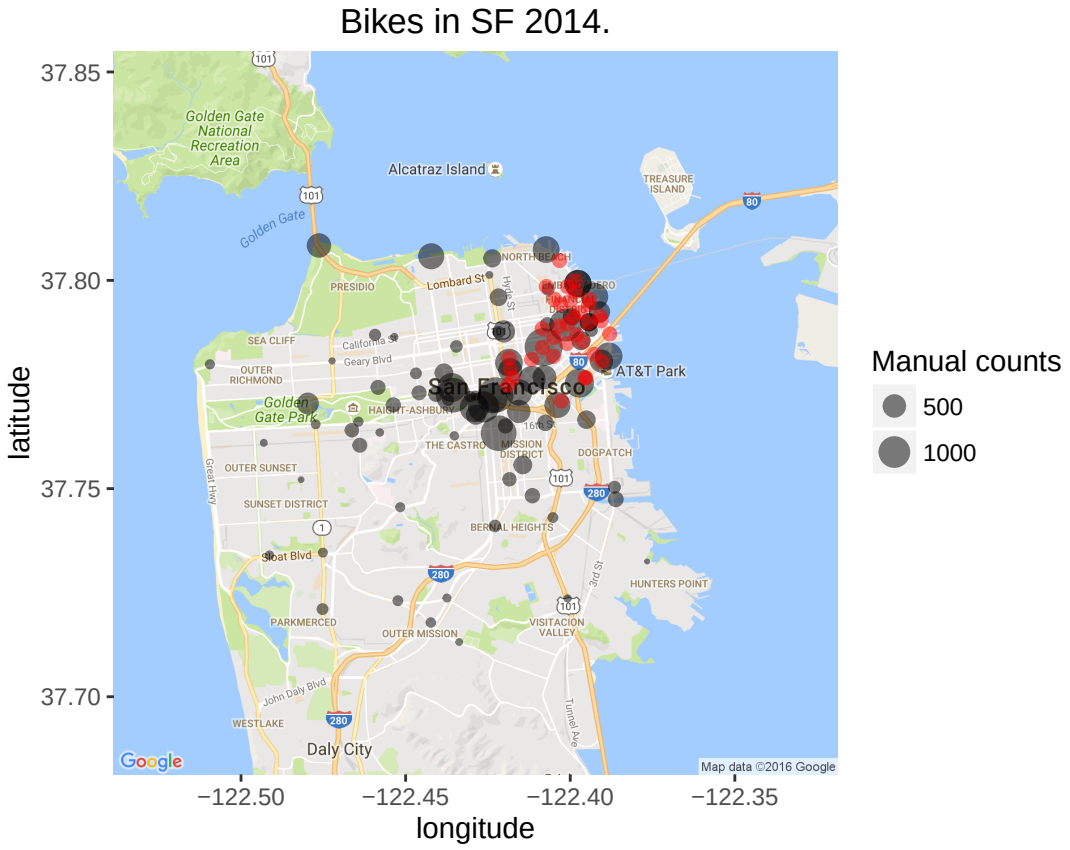

```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2011),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2011.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```



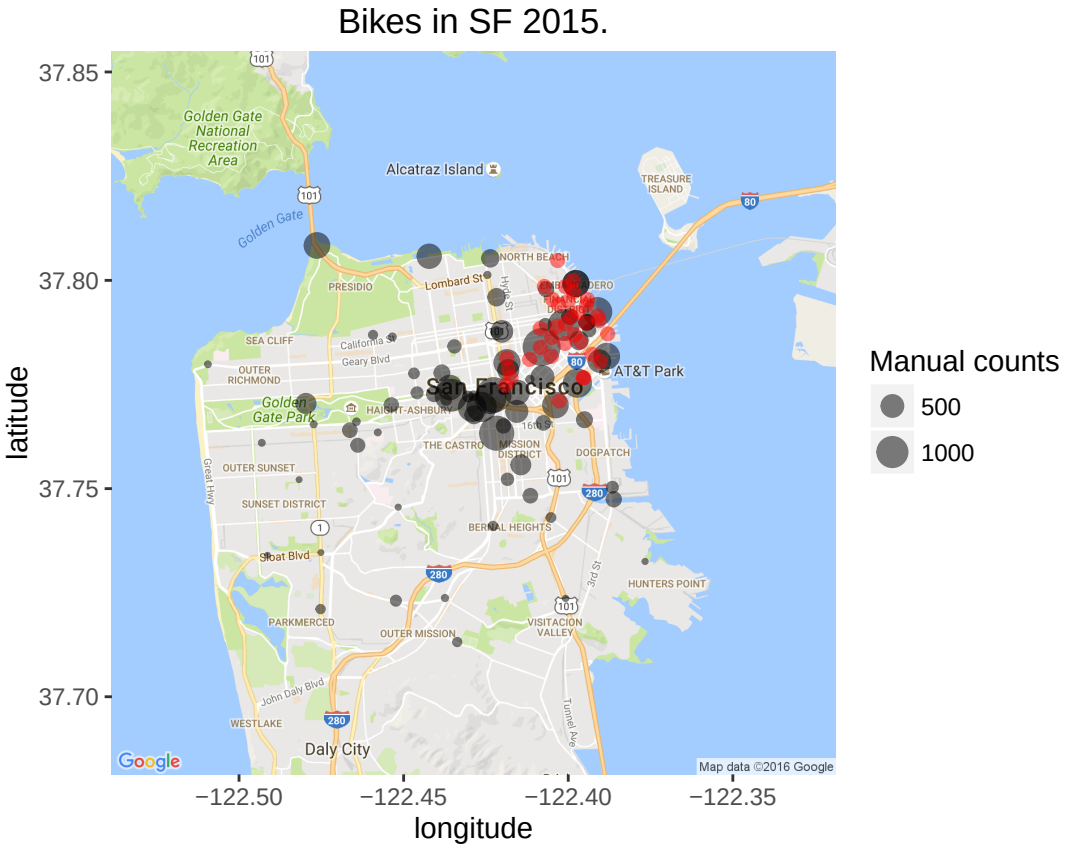

```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2013),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2013.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```




```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2014),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2014.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```




```
ggmap(BABSmap)+
  geom_point(data = filter(bikeCountsNew, Year == 2015),
             aes(x=longitude, y=latitude, size = value), alpha = 0.5)+
  labs(x = "longitude", y = "latitude", size ="Manual counts",
       title = "Bikes in SF 2015.")+scale_size_area()+
  geom_point(data = station, aes(x=long, y=lat), size =2,col = "red", alpha = 0.5)
```



Interestingly we see plenty of bike usage outside of the area covered by the stations (red dots). This points towards potential regions of expansion as we now know that bikes are being used here. Furthermore in recreational/tourist areas such as the parks and the bridge there is plenty of bike usage. Having stations in these areas will promote the use of the bike share system as a recreational tool for residents of and visitors to San Francisco.

Imbalance

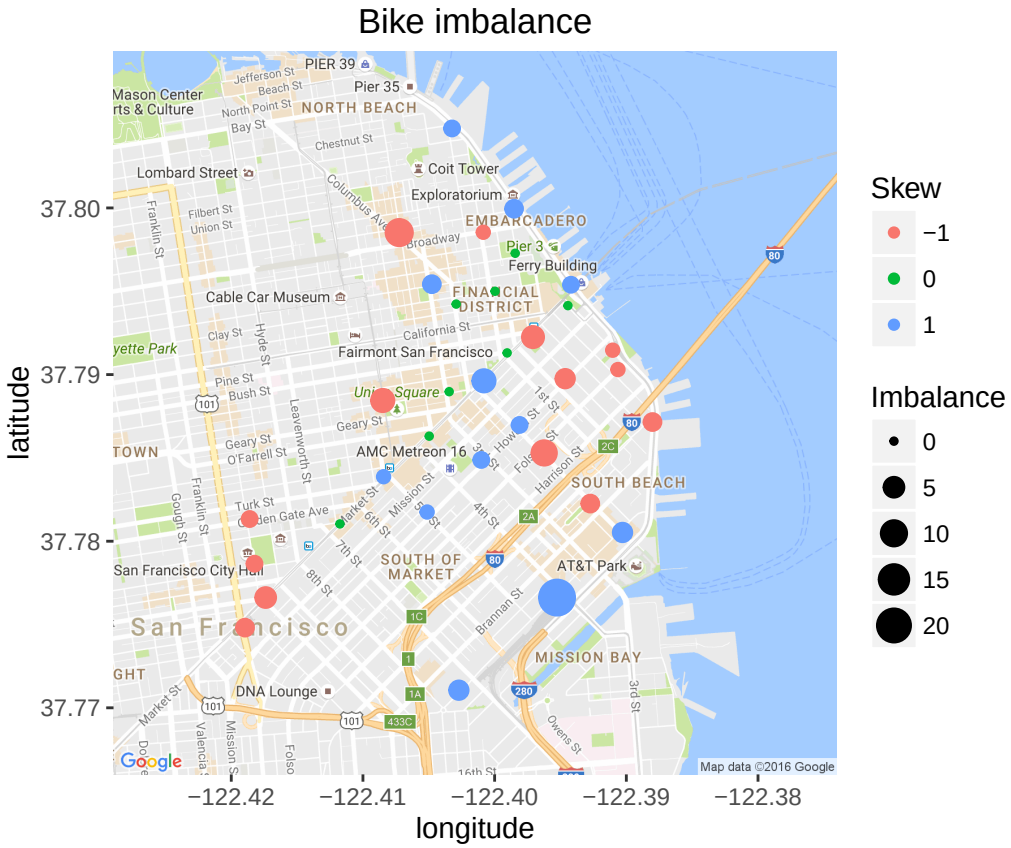
We will now quantify the imbalance in the system. If more cycles are hired than deposited from a station then the station requires re-stocking. Whilst if less cycles are hired than deposited the station fills up and depositing a bike can be frustrating. We'll calculate the averages per day for each station

```
hires <- trip  %>%
  group_by(Start.Station) %>%
  summarise(meanHires = round(n()/365))
deposits <- trip %>%
  group_by(End.Station) %>%
  summarise(meanDeposits = round(n()/365))

inOut <- left_join(hires, deposits, by = c("Start.Station" = "End.Station"))
inOut <- inOut %>%
  mutate(diff = meanDeposits-meanHires) %>%
  mutate(absDiff = abs(diff), loss = ifelse(diff > 0 , 1, ifelse(diff< 0, -1, 0))) %>%
  arrange(desc(absDiff))
names(inOut) <- c("Station", "meanHires","meanDeposits","diff","absDiff", "difference")
```

```
inOut <- left_join(inOut, station, by = c("Station" = "name"))
inOut
```

```
## # A tibble: 70 x 12
##                                       Station meanHires meanDeposits
##                                         <chr>     <dbl>        <dbl>
## 1       San Francisco Caltrain (Townsend at 4th)        72           95
## 2               Grant Avenue at Columbus Avenue        23           12
## 3                                 2nd at Folsom        22           13
## 4                              Market at Sansome        31           38
## 5                 Powell at Post (Union Square)        18           11
## 6                                Beale at Market        23           17
## 7                                 Market at 10th        33           28
## 8                                2nd at Townsend        38           42
## 9  Temporary Transbay Terminal (Howard at Beale)        40           36
## 10                               Townsend at 7th        38           42
## # ... with 60 more rows, and 9 more variables: diff <dbl>, absDiff <dbl>,
## #   difference <dbl>, station_id <int>, lat <dbl>, long <dbl>,
## #   dockcount <int>, landmark <chr>, installation <chr>
```

```
inOutSF <- filter(inOut, landmark == "San Francisco")
location <- c( mean(inOutSF$long), mean(inOutSF$lat))
BABSmap <- get_map(location,maptype ="roadmap", zoom=14)

ggmap(BABSmap)+
  geom_point(data = inOutSF,
             aes(x=long, y=lat, size = absDiff, col = as.factor(difference)))+
  labs(x = "longitude", y = "latitude",
       size ="Imbalance",col ="Skew",title = "Bike imbalance")
```



We see that in general there isn't too much imbalance in the system. It is mainly in the busy San Francisco. During expansion, increasing the size of these stations will help alleviate the strain on the restocking procedure whilst saving money. Also of note is that in the north more bikes are being dropped off at the periphery of the system. This would indicate that in fact people would like to go further into the north beach area.