Project 4

Tyler Netherly

October 2, 2017

**1.** Consider the data available from: <https://bikeshare.metro.net/about/data/>

**1a.** Download the 2017 Q2 data using the wget command in the bash shell (i.e., in the terminal).

system("wget \"https://11ka1d3b35pv1aah0c3m9ced-wpengine.netdna-ssl.com/wp-content/uploads/2017/07/la\_metro\_gbfs\_trips\_Q2\_2017.csv.zip\"")

**1b.** Use the unzip command in the bash shell (i.e., in the terminal) to extract the csv file.

system("unzip la\_metro\_gbfs\_trips\_Q2\_2017.csv.zip")

**1c.** Import the data into R using the read.csv command.

myDF <- read.csv("~/la\_metro\_gbfs\_trips\_Q2\_2017.csv")

**1d.** How many unique bike ID's are found in this file?

length(table(myDF$bike\_id))

## [1] 738

**1e.** Which bike was used for the largest number of trips?

tail(sort(table(myDF$bike\_id)),1)

## 4727   
## 139

**1f.** Which type of passholder is the most common?

tail(sort(table(myDF$passholder\_type)),1)

## Monthly Pass   
## 35737

**2.** The starting and ending times are given, but they are easier to work with in R, if we put them into a date context, for instance, using the as.POSIXlt function. This allows us, for instance, to subtract the times to find the differences.

**2a.** Compare the duration column to the difference of the end time minus the start time (where we use the as.POSIXlt command on each of the end time and start time columns beforehand). Why is the duration not always equal to the end time minus the start time?

timedistance <- as.POSIXct(myDF$end\_time) - as.POSIXct(myDF$start\_time)  
head(timedistance == myDF$duration)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE

There is a possibility that the time in myDF$duration started towards the end of a minute and ended towards the beginning of the end time minute which means that the time might be a minute less. Something similar could happen with the time being greater.

**2b.** How many times are the duraction values in 2a different from the end time minus the start time?

sum(timedistance != myDF$duration)

## [1] 78

**2c.** Create a new column in the data.frame that contains the end time minus the start time.

myDF$time\_difference <- timedistance

**2d.** Find an average of the values in this new column, for each of the bike ID's.

head(tapply(myDF$time\_difference,myDF$bike\_id,mean))

## 4727 4728 5715 5716 5718 5719   
## 18.20863 23.80519 18.39394 10.58824 25.57812 20.78947

**3.** Make a map that displays the locations of the stations. (You can use either the starting or the ending locations, or both.)

library(ggmap)

## Loading required package: ggplot2

myDFlocations <- data.frame(lat=myDF$start\_lat,lon=myDF$start\_lon)  
losangeles <- as.numeric(geocode("Los Angeles"))

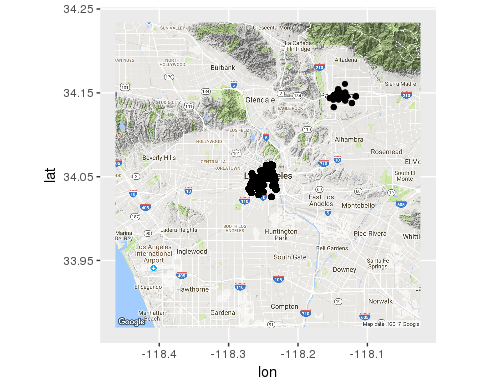
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Los%20Angeles&sensor=false

LAmap <- ggmap(get\_googlemap(losangeles, zoom = 11), extent = "Device")

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=34.052234,-118.243685&zoom=11&size=640x640&scale=2&maptype=terrain&sensor=false

LAmap <- LAmap + geom\_point(data=myDFlocations[myDFlocations$lon < 0,])  
LAmap

## Warning: Removed 64 rows containing missing values (geom\_point).



**4,5,6.** Please solve questions 1, 2, 3 again, but this time use the Q1 2017 data, and then the Q4 2016 data, and then the Q3 2016 data.

Please note that for these three data sets, the seconds are missing, and the data is given in a nonstandard format, so you will need to use the option format='%m/%d/%Y %H:%M' inside each of your functions calls to the as.POSIXlt function. Also please be careful, when comparing the duration to the end time minus the start time, whether the times are given in minutes or seconds.

**4.**

#1.  
# 1a. Download Q1 2017 Data  
system("wget \"https://11ka1d3b35pv1aah0c3m9ced-wpengine.netdna-ssl.com/wp-content/uploads/2017/04/la\_metro\_gbfs\_trips\_Q1\_2017.zip\"")  
# 1b. Unzip data  
system("unzip la\_metro\_gbfs\_trips\_Q1\_2017.zip")

# 1c. Import data into R  
myDF <- read.csv("la\_metro\_gbfs\_trips\_Q1\_2017.csv")

# 1d. Number of unique bike ids  
length(table(myDF$bike\_id))

## [1] 751

# 1e. bikeid with greatest number of trips   
tail(sort(table(myDF$bike\_id)),1)

## 6344   
## 98

# 1f. Most common type of passholder  
tail(sort(table(myDF$passholder\_type)),1)

## Monthly Pass   
## 21007

#2.   
# 2a.   
timedistance <- as.POSIXct(myDF$end\_time,format='%m/%d/%Y %H:%M') - as.POSIXct(myDF$start\_time,format='%m/%d/%Y %H:%M')  
head(timedistance == myDF$duration / 60)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE

#2b.  
sum(timedistance != myDF$duration / 60)

## [1] 89

#2c.  
myDF$time\_difference <- timedistance

#2d.   
head(tapply(myDF$time\_difference,myDF$bike\_id,mean))

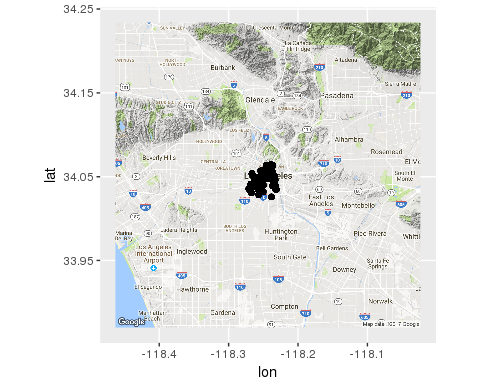
## 1349 4727 4728 5715 5716 5718   
## 311.00000 41.05435 14.73585 23.02857 27.00000 11.23529

#3. Map  
library(ggmap)  
myDFlocations <- data.frame(lat=myDF$start\_lat,lon=myDF$start\_lon)  
#losangeles <- as.numeric(geocode("Los Angeles"))  
LAmap <- ggmap(get\_googlemap(losangeles, zoom = 11), extent = "Device")

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=34.052234,-118.243685&zoom=11&size=640x640&scale=2&maptype=terrain&sensor=false

LAmap <- LAmap + geom\_point(data=myDFlocations[myDFlocations$lon < 0 & myDFlocations$lon > -118.3,])  
LAmap

## Warning: Removed 29 rows containing missing values (geom\_point).



**5.**

##4. Q1 2017  
#1.  
# Download Q1 2017 Data  
system("wget \"https://11ka1d3b35pv1aah0c3m9ced-wpengine.netdna-ssl.com/wp-content/uploads/2017/04/la\_metro\_gbfs\_trips\_Q1\_2017.zip\"")  
# Unzip data  
system("unzip la\_metro\_gbfs\_trips\_Q1\_2017.zip")

# Import data into R  
myDF <- read.csv("la\_metro\_gbfs\_trips\_Q1\_2017.csv")  
# Number of unique bike ids  
length(table(myDF$bike\_id))

## [1] 751

# bikeid with greatest number of trips   
tail(sort(table(myDF$bike\_id)),1)

## 6344   
## 98

# Most common type of passholder  
tail(sort(table(myDF$passholder\_type)),1)

## Monthly Pass   
## 21007

#2.   
# 2a.   
timedistance <- as.POSIXct(myDF$end\_time,format='%m/%d/%Y %H:%M') - as.POSIXct(myDF$start\_time,format='%m/%d/%Y %H:%M')  
head(timedistance == myDF$duration / 60)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE

#2b.  
sum(timedistance != myDF$duration / 60)

## [1] 89

#2c.  
myDF$time\_difference <- timedistance

#2d.   
head(tapply(myDF$time\_difference,myDF$bike\_id,mean))

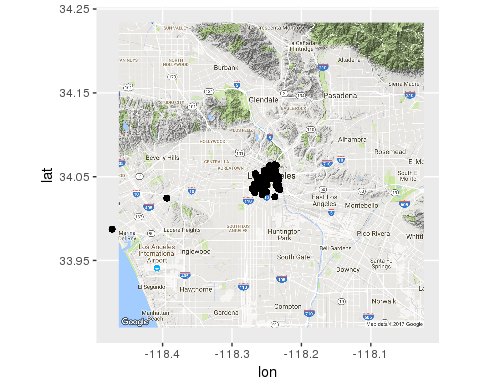
## 1349 4727 4728 5715 5716 5718   
## 311.00000 41.05435 14.73585 23.02857 27.00000 11.23529

#3. Map  
library(ggmap)  
myDFlocations <- data.frame(lat=myDF$start\_lat,lon=myDF$start\_lon)  
##losangeles <- as.numeric(geocode("Los Angeles"))  
LAmap <- ggmap(get\_googlemap(losangeles, zoom = 11), extent = "Device")

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=34.052234,-118.243685&zoom=11&size=640x640&scale=2&maptype=terrain&sensor=false

LAmap <- LAmap + geom\_point(data=myDFlocations[myDFlocations$lon < 0,])  
LAmap

## Warning: Removed 29 rows containing missing values (geom\_point).



**6.**

#6. Q3 2016  
#1.  
# Download Q3 2016 Data  
system("wget \"https://11ka1d3b35pv1aah0c3m9ced-wpengine.netdna-ssl.com/wp-content/uploads/2016/10/MetroBikeShare\_2016\_Q3\_trips.zip\"")  
# Unzip data  
system("unzip MetroBikeShare\_2016\_Q3\_trips.zip")

# Import data into R  
myDF <- read.csv("~/MetroBikeShare\_2016\_Q3\_trips.csv")  
# Number of unique bike ids  
length(table(myDF$bike\_id))

## [1] 761

# bikeid with greatest number of trips   
tail(sort(table(myDF$bike\_id)),1)

## 6608   
## 135

# Most common type of passholder  
tail(sort(table(myDF$passholder\_type)),1)

## Monthly Pass   
## 33216

#2.   
# 2a.   
timedistance <- as.POSIXct(myDF$end\_time,format='%m/%d/%Y %H:%M') - as.POSIXct(myDF$start\_time,format='%m/%d/%Y %H:%M')  
head(timedistance == myDF$duration / 60)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE

#2b.  
sum(timedistance != myDF$duration / 60)

## [1] 89

#2c.  
myDF$time\_difference <- timedistance

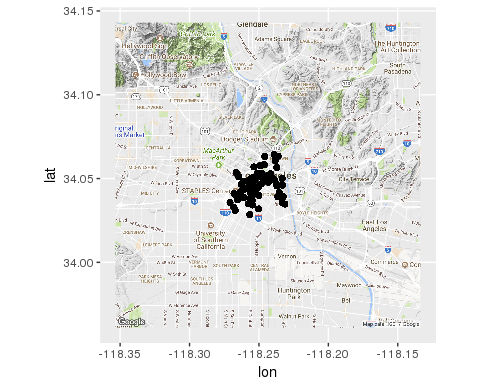
#2d.   
head(tapply(myDF$time\_difference,myDF$bike\_id,mean))

## 4727 4728 5715 5716 5718 5719   
## 18.30833 20.71951 15.87143 12.88750 39.88596 25.48649

#3. Map  
library(ggmap)  
myDFlocations <- data.frame(lat=myDF$start\_lat,lon=myDF$start\_lon)  
##losangeles <- as.numeric(geocode("Los Angeles"))  
LAmap <- ggmap(get\_googlemap(losangeles, zoom = 12), extent = "Device")

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=34.052234,-118.243685&zoom=12&size=640x640&scale=2&maptype=terrain&sensor=false

LAmap <- LAmap + geom\_point(data=myDFlocations[myDFlocations$lon < 0,])  
LAmap



**7a.** In each of the 4 data.frames, carefully give a new defintion to the start\_time and end\_time columns (but keep the same column names), by converting each of these columns using the as.POSIXlt function. After doing so, then the dates and times for all four data.frames should be in the same format.

##7. a  
myDF1 <- read.csv("~/la\_metro\_gbfs\_trips\_Q2\_2017.csv")  
myDF2 <- read.csv("~/la\_metro\_gbfs\_trips\_Q1\_2017.csv")  
myDF3 <- read.csv("~/Metro\_trips\_Q4\_2016.csv")  
myDF4 <- read.csv("~/MetroBikeShare\_2016\_Q3\_trips.csv")  
  
myDF1$start\_time <- as.POSIXct(myDF1$start\_time)  
myDF1$end\_time <- as.POSIXct(myDF1$end\_time)  
  
myDF2$start\_time <- as.POSIXct(myDF2$start\_time,format='%m/%d/%Y %H:%M')  
myDF2$end\_time <- as.POSIXct(myDF2$end\_time,format='%m/%d/%Y %H:%M')  
  
myDF3$start\_time <- as.POSIXct(myDF3$start\_time,format='%m/%d/%Y %H:%M')  
myDF3$end\_time <- as.POSIXct(myDF3$end\_time,format='%m/%d/%Y %H:%M')  
  
myDF4$start\_time <- as.POSIXct(myDF4$start\_time,format='%m/%d/%Y %H:%M')  
myDF4$end\_time <- as.POSIXct(myDF4$end\_time,format='%m/%d/%Y %H:%M')

**7b.** Rename the 5th and 8th columns of the 2017 Q2 data.frame to be "start\_station\_id" and "end\_station\_id" respectively. You can use names(myDF)[5] and names(myDF)[8] to access and change these names.

## 7b.  
names(myDF1)[5] <- "start\_station\_id"  
names(myDF1)[8] <- "end\_station\_id"  
names(myDF1)

## [1] "trip\_id" "duration" "start\_time"   
## [4] "end\_time" "start\_station\_id" "start\_lat"   
## [7] "start\_lon" "end\_station\_id" "end\_lat"   
## [10] "end\_lon" "bike\_id" "plan\_duration"   
## [13] "trip\_route\_category" "passholder\_type"

**7c.** Convert the duration column of the other three data.frames (i.e., 2017 Q1, 2016 Q4, and 2016 Q3) from seconds into minutes by dividing by 60 and saving the new values into the duration column.

##7c.   
myDF2$duration <- myDF2$duration / 60  
myDF3$duration <- myDF3$duration / 60  
myDF4$duration <- myDF4$duration / 60

**7d.** Build a new data.frame, using the rbind function, which contains all of the data from all four data.frames.

##7d.  
myDF <- rbind(myDF1,myDF2,myDF3,myDF4)

**8.**Now repeat questions 1, 2, 3 using the new data.frame that was created in 7d.

##1d. Number of unique bike ID's?  
length(table(myDF$bike\_id))

## [1] 766

##1e. bikeid with greatest number of trips   
tail(sort(table(myDF$bike\_id)),1)

## 4727   
## 451

##1f.   
tail(sort(table(myDF$passholder\_type)),1)

## Monthly Pass   
## 117041

##2a.   
timedistance <- myDF$end\_time - myDF$start\_time  
head(timedistance == myDF$duration)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE

##2b.  
sum(timedistance != myDF$duration)

## [1] 363

##2c.  
myDF$time\_difference <- timedistance

##2d.   
head(tapply(myDF$time\_difference,myDF$bike\_id,mean))

## 1349 4727 4728 5715 5716 5718   
## 311.00000 26.88027 19.77021 19.34914 15.07339 32.55612

##3.   
library(ggmap)  
myDFlocations <- data.frame(lat=myDF$start\_lat,lon=myDF$start\_lon)  
##losangeles <- as.numeric(geocode("Los Angeles"))  
LAmap <- ggmap(get\_googlemap(losangeles, zoom = 11), extent = "Device")

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=34.052234,-118.243685&zoom=11&size=640x640&scale=2&maptype=terrain&sensor=false

LAmap <- LAmap + geom\_point(data=myDFlocations[myDFlocations$lon < 0,])  
LAmap

## Warning: Removed 112 rows containing missing values (geom\_point).

