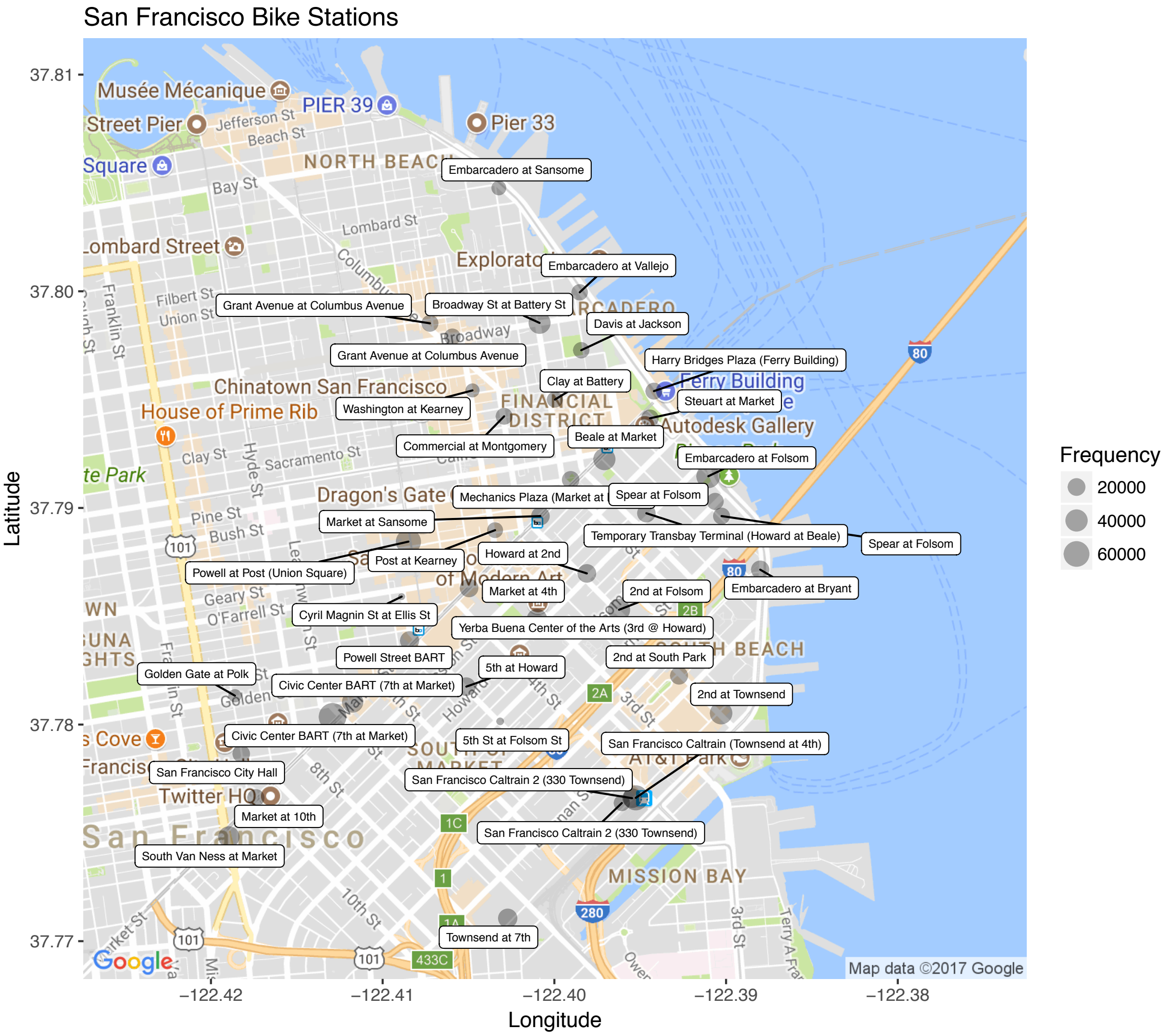
Ricardo Rendon

Sta 141 Hw 3

1 in code

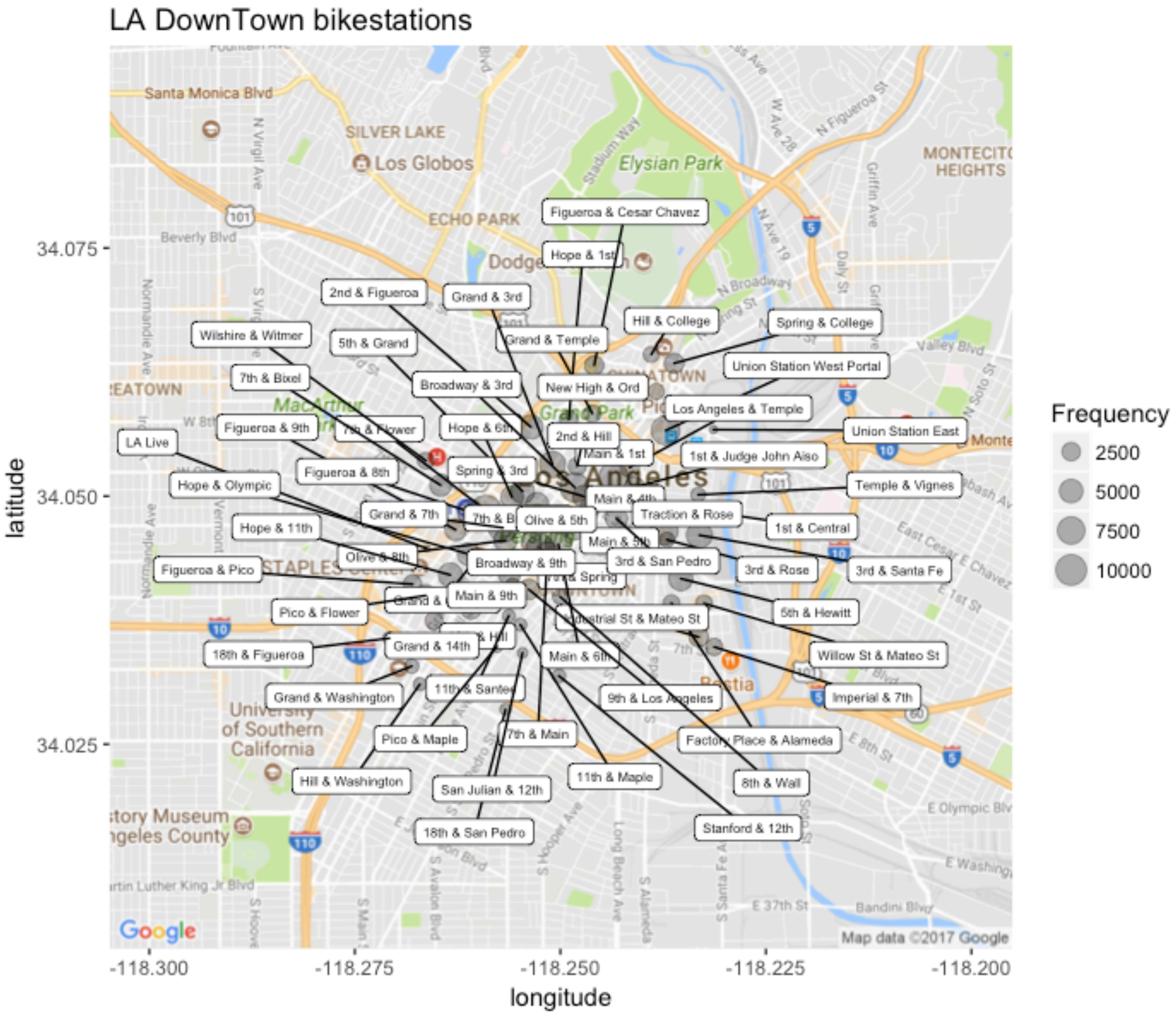
2



We can observe in the map that most of the bike stations are a located in a specific area in sanfranciso. This area is where most of tourist attractions are situated. This data make sense since tourists make up most of the market in this type of industry. Also, the stores with most frequency (amount of circulation in the store,trips started form that station) are the ones with the best locations. For example "Civic Center BART(7 at market)" its at the start of a big street so people will encounter this store before other ones. Another example is "2nd and townsand" which is right next to the at&t park, a very popular spot in SF. One the other hand the stores with less movement are situated in the borders of this bike store area.

3
In code

4



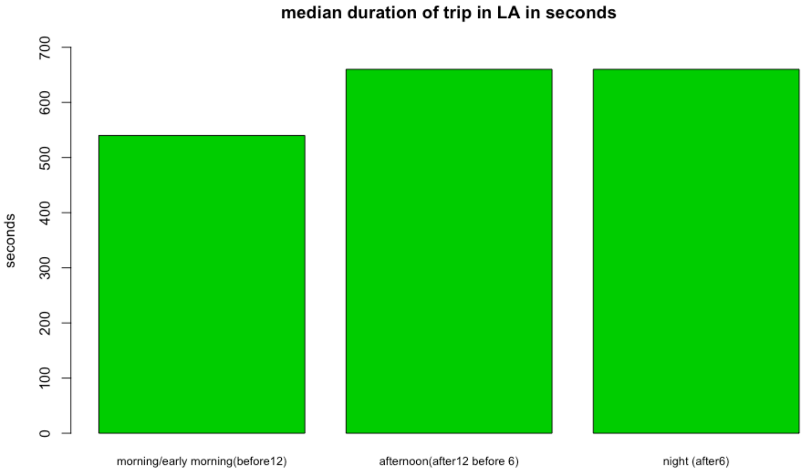
Explanation:

Los angels is a very big city with lots of movement everywhere that is why the frequency (number of started trips for each stations) doesn't follow a clear pattern. Usually these are base more on popularity than location. One might be able to see how in the middle the stores have high frequency but that is just an observation, which can be kind of subjective since there are stores with big frequency outside of this big main central circle.

Ricardo Rendon

5

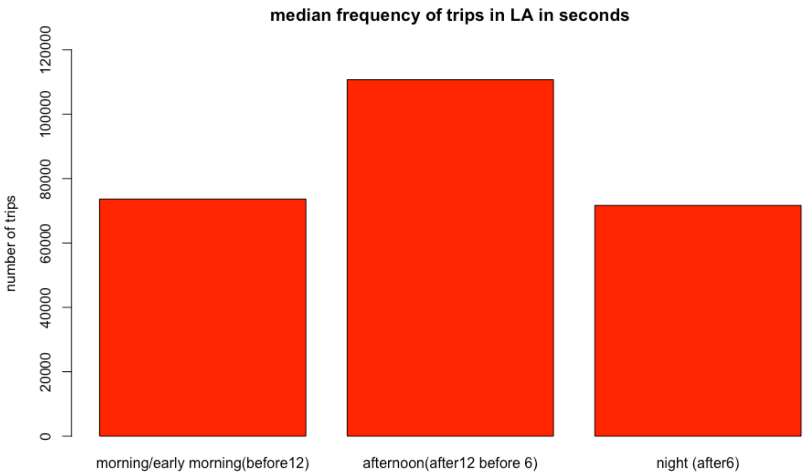
For LA:



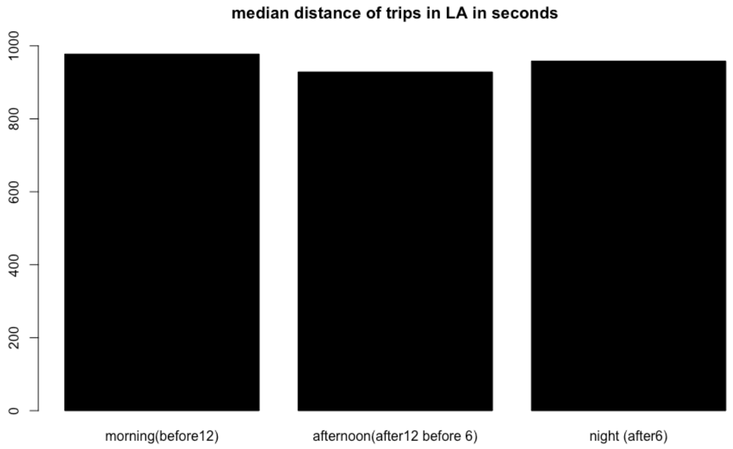
These plots show that people in the morning tend to do short duration trips with a distance of 900m on average. Also, the stores have has low costumers movement in this time of the day.

For the afternoon the duration is above the morning based on duration of trips. The frequency is the highest and the distance is the same (around 900m)

For night it has the lowest frequency and same amount of duration and distance as afternoon.

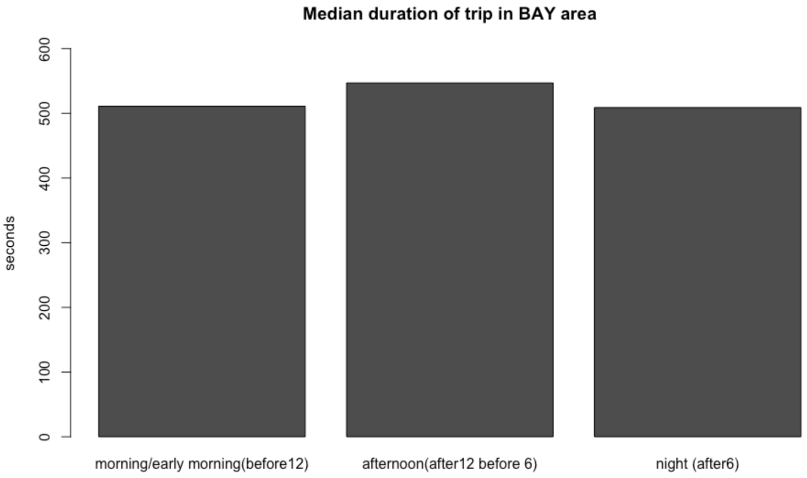


The data repressented in this plots can be explained analyzing the behavior of people. In the morning people don't go outside since its cold and if they do they keep the trip short since they probably have to go to work or its just for a "wake up exercise". On afternoons is when tourists start wondering around the city so it makes sense how the frequency is the highest around this time. Regarding the duration it's more than morning because it is usually warmer and more pleasant to outside and get to know the city for longer periods. Finally, for the night it is normal that it gets a bit colder therefore; people will try to go out to party or dinners (inside a building) so less people go to on bike trips so it's understandable that the frequency is low. Duration stays the same as afternoon because people that go on trips have the goal to explore the city at night, they know its going to be cold so they get ready for the low temperatures and it usually takes the same amount of time and distance to explore the city at afternoon or night.

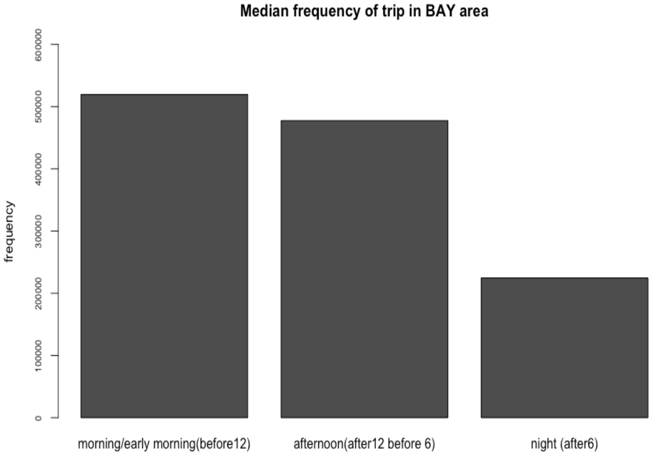


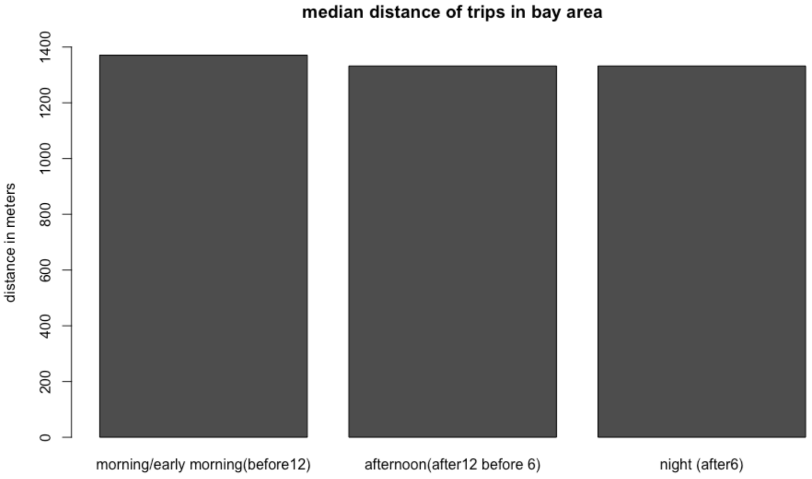
Ricardo Rendon

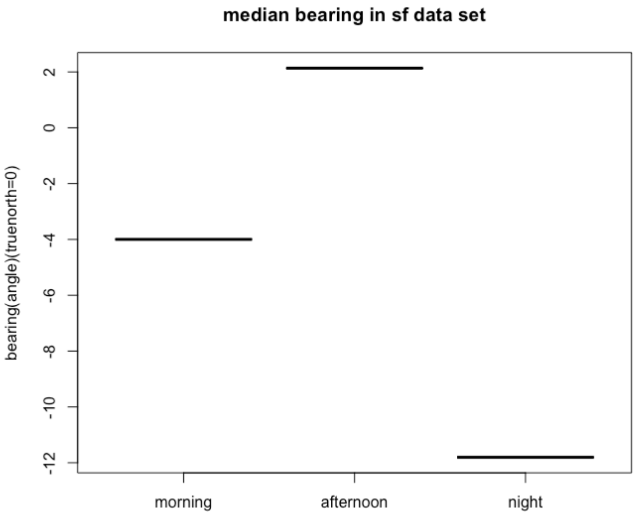
For bay area:

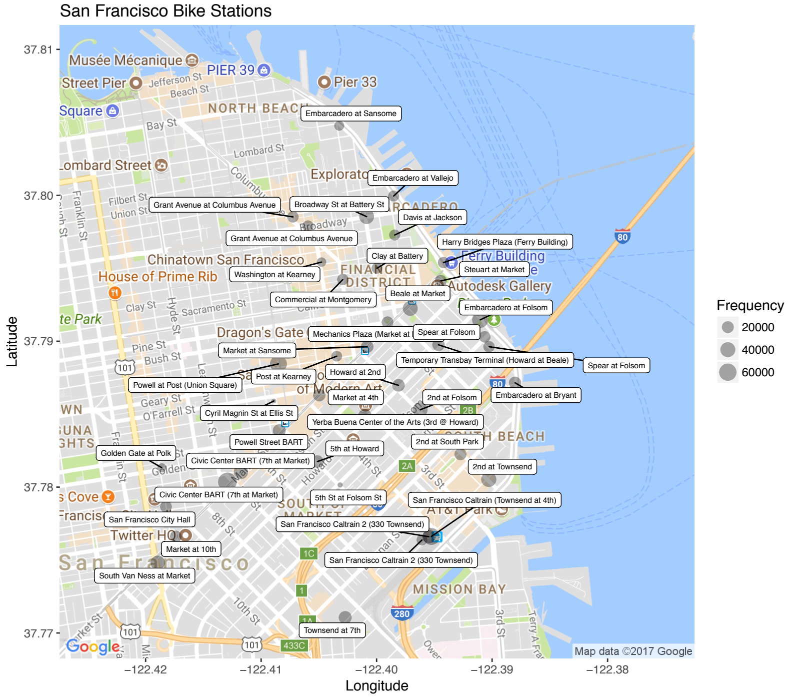


As we can see in this plots the duration of trip and distance of trip are not affected so much by the time of day. However the frequency changes drastically and this is due that in the bay area nights are very cold and people try not to bike with such amount of cold wind and low temperature.









We can see that on average around the day people move towards -4.5 which is north with some degree to the west. This tells us that people after getting their bike they are trying to head towards the ocean and go along it towards pier 39. We can observe the biggest change in bearing from afternoon to night, this could be explained due to that in the afternoon people try to head to pier 39 and get there as fast as possible (since the tourist activities close around five) and going north is the fastest way to get there. At night people want to explore the city so they just take the long way following the cost towards pier 39.

Appendix

Ricardo Rendon

```
##1 Done .Write a function that loads the Bay Area bike share trip data from a CSV file,
##converts the columns toappropriate data types, and then saves the tidied data
##frame to an RDS file. Your function should havearguments to set the path for
##the input CSV file and the output RDS file.Write a second function that does
##the same thing for the Bay Area bike share station data.
install.packages("lubridate")
install.packages("maps")
install.packages("ggmap")
install.packages("ggplot2")
install.packages("geosphere")
install.packages("ggrepel")
install.packages("")
library(maps)
library(ggmap)
library(ggplot2)
library(geosphere)
library(ggrepel)
library(lubridate)


getwd()
setwd("/Users/rire948/Downloads/bikes")


name="sf_bikeshare_trips.csv"


fix.file = function (path=name,output = "/Users/rire948/Downloads/bikes/sf_bikeshare_trips.rds") {
  trips1= read.csv(name)


  trips1$trip_id = as.factor(trips1$trip_id)


  trips1$start_date = ymd_hms(trips1$start_date)##transform to poPOSIX


  trips1$start_station_id = as.factor(trips1$start_station_id)


  trips1$end_date = ymd_hms(trips1$end_date)
  trips1$bike_number = as.factor(trips1$bike_number)##transform to poPOSIX


  saveRDS(trips1, output)

}

fix.file()


head(readRDS( "/Users/rire948/Downloads/bikes/sf_bikeshare_trips.rds"))


name2="sf_bike_share_stations.csv"


fix.file2 = function (path=name2,output = "/Users/rire948/Downloads/bikes/sf_bike_share_stations.rds") {

   share_station=read.csv(name2)


   share_station$station_id = as.factor(share_station$station_id)


   share_station$installation_date = ymd(share_station$installation_date)##transform to poPOSIX


   saveRDS(share_station,output)


}

fix.file2()


head(readRDS( "/Users/rire948/Downloads/bikes/sf_bike_share_stations.rds"))





##2 Done/ Create a map that shows the locations of the Bay Area bike share stations in San Francisco (only).

##Label each station with its name. Make the size of each point correspond to the number of trips startedfrom

##that station. Discuss what you can conclude from the map.


setwd("/Users/rire948/Downloads/bikes")


bikestations <- read.csv(file="sf_bike_share_stations.csv", header=TRUE, sep=",")

travel <- read.csv(file="sf_bikeshare_trips.csv", header=TRUE, sep=",")

head(bikestations)
```

Ricardo Rendon

```
head(travel)


#here is were i separate sf #so we need only bay area so:
levels(bikestations$landmark)
BayArea_BikeStation=subset(bikestations,bikestations$landmark =="San Francisco")
BayArea_BikeStation = BayArea_BikeStation[!duplicated(BayArea_BikeStation, by = "station_id"), ]


head(BayArea_BikeStation)


ab=as.data.frame(matrix(table(travel$start_station_id)))
ab
names((table(travel$start_station_id)))
ab$station_id=names((table(travel$start_station_id)))
ab
ab1=merge(BayArea_BikeStation,ab,by="station_id")
ab1
head(ab1)
#now lets get the map for SF


sf_loc <- c(lon = -122.400, lat = 37.79)
SFO = get_map(location = sf_loc,zoom=14, maptype="roadmap")  # a ggmap object is created, but nothing plotted
ggmap(SFO) +
  labs(size = "Frequency", x = "Longitude", y = "Latitude", title = "San Francisco Bike Stations") +
  geom_point(aes(longitude,latitude,size =ab1$V1 ),data = BayArea_BikeStation, alpha=I(1/3)) +
  geom_label_repel(aes(longitude,latitude,label=name),data = BayArea_BikeStation,size=2)
```

```
####3Write a function that loads the Los Angeles bike share trip data from the 5 provided CSV files,
#bindsthem into one data frame, converts the columns to appropriate data types, and saves the tidied
#dataframe to an RDS file. Your function should have arguments to set the path for the input directoryand
#the output RDS file. Keep your function short and simple by using an apply function rather thanrepeating
#code.Write a second function that loads, tidies, and saves the Los Angeles bike share station data.
```

Ricardo Rendon

```
##https://stackoverflow.com/questions/9564489/opening-all-files-in-a-folder-and-applying-a-function
path = "/Users/rire948/Downloads/bikes/question3"
function3 = function(path = "/Users/rire948/Downloads/bikes/question3", output =
"/Users/rire948/Downloads/bikes/question3/laShareTrips.rds"){

filenames <- list.files(path, pattern="*.csv", full.names=TRUE)
datas <- lapply(filenames, read.csv)

a1=as.data.frame(datas[1])
a2=as.data.frame(datas[2])
a3=as.data.frame(datas[3])
a4=as.data.frame(datas[4])
a5=as.data.frame(datas[5])

#chekinf if data is structure the same way in all files
#duration change
a4$duration=a4$duration*60
a5$duration=a5$duration*60

## bike id,end_station_id and start_station_id are in diferent type(we will fix this after mergin)
## we need to fix name of column
names(a4)[5]="start_station_id"
names(a5)[5]="start_station_id"

names(a4)[8]="end_station_id"
names(a5)[8]="end_station_id"
## now change the time that is expresed diferently in some
head(a1$start_time)
a1$start_time=mdy_hm(a1$start_time)
a2$start_time=mdy_hm(a2$start_time)
a3$start_time=mdy_hm(a3$start_time)
a5$start_time=mdy_hm(a5$start_time)
head(a1$start_time)

a1$end_time=mdy_hm(a1$end_time)
```

```
a2$end_time=mdy_hm(a2$end_time)

a3$end_time=mdy_hm(a3$end_time)

a5$end_time=mdy_hm(a5$end_time)


table(a1$start_station_id)

table(a2$start_station_id)

table(a3$start_station_id)

table(a4$start_station_id)

table(a5$start_station_id)



##now lets merge all dataframes by trip_id


list=list(a1,a2,a3,a4,a5)

new=do.call(rbind,list)

#str(new)

#head(new)


saveRDS(new,output)

}


function3()


a8797=(readRDS("/Users/rire948/Downloads/bikes/question3/laShareTrips.rds"))

table(a8797$start_station_id)


#####part2 Write a second function that loads, tidies, and saves the Los Angeles bike share station data.


functionPart2=function(path="/Users/rire948/Downloads/bikes/metro-bike-share-stations-2017-10-20.csv",output =
"/Users/rire948/Downloads/bikes/laTripsStations.rds"){
  df=read.csv(path)


  df$Go_live_date[[1]]="7/7/2016"

  df$Go_live_date= mdy(df$Go_live_date)

  df$Region=factor(df$Region)
  df$Region[df$Region=="N/A"]= NA
  df$Region=droplevels(df$Region)

  df$Status=factor(df$Status)
  saveRDS(df, output)

}

functionPart2()
readRDS( "/Users/rire948/Downloads/bikes/laTripsStations.rds")


#4.Create a map that shows the locations of the Los Angeles bike share stations near downtown LosAngeles
#(only). Label each station with its name. Make the size of each point correspond to the numberof trips
#started from that station. Discuss what you can conclude from the map.

stationsLA=readRDS("/Users/rire948/Downloads/bikes/laTripsStations.rds")
head(stationsLA)
names(stationsLA)[1]="start_station_id"

tripsLa=readRDS("/Users/rire948/Downloads/bikes/question3/laShareTrips.rds")
head(tripsLa)
names(tripsLa)[6]="latitude"
names(tripsLa)[7]="longitude"


dtlaTrips= subset(tripsLa, tripsLa$start_station_id %in% stationsLA$start_station_id)

table(dtlaTrips$start_station_id)

dtlaTrips=as.data.frame(dtlaTrips)
head(dtlaTrips)


dtlaTrips1 = dtlaTrips[!duplicated(dtlaTrips$start_station_id), ]
```

```
dtlaTrips2=merge(dtlaTrips1,stationsLA,by="start_station_id")
head(dtlaTrips2)
table(dtlaTrips2$Region)



ab=as.data.frame(table(tripsLa$start_station_id))
ab=ab[-1,]
ab

colnames(ab)=c("start_station_id","Freq")
Final=merge(dtlaTrips2,ab,by="start_station_id")
head(Final)

Final=subset(Final,Final$Region=="DTLA")
sum(is.na(Final))

la <- c(lon = -118.250, lat = 34.050)
La = get_map(location = la,zoom=13, maptype="roadmap")  # a ggmap object is created, but nothing plotted
ggmap(La) +
  labs( size = "Frequency",x = "longitude", y = "latitude", title = "LA DownTown bikestations") +
  geom_point(aes(longitude,latitude,size =Final$Freq), data = Final, alpha=I(1/3)) +
  geom_label_repel(aes(longitude,latitude,label=Station_Name),data = Final,size=2)

###5.How do trip frequency, distance, and duration change at different times of day? Investigate for both theBay Area bike
share
###and the Los Angeles bike share. Compare your findings. Thegeosphere::distGeo()1function can compute distances for
longitude and
###latitude coordinates

install.packages("geosphere")
library("geosphere")
library("lubridate")

#for la++++++++ laturude and long ar ealready in so just find distance
la_bikeshareStat=readRDS("/Users/rire948/Downloads/bikes/laTripsStations.rds")
```

```
la_bikeshareStat##dont need
la_bikeshartripse=readRDS("/Users/rire948/Downloads/bikes/question3/laShareTrips.rds")
head(la_bikeshartripse)



##ok so we have have duration and distance

la_bikeshartripse$distance=distGeo(cbind(la_bikeshartripse$start_lon,la_bikeshartripse$start_lat),cbind(la_bikeshartripse$end_lon,la_bikeshartripse$end_lat))
head(la_bikeshartripse)


target_afternoon=as.POSIXct("12:00", format =  "%H:%M")
target_afternoon <- hour(target_afternoon) + minute(target_afternoon)/60
target_afternoon

target_night=as.POSIXct("18:00", format =  "%H:%M")
target_night <- hour(target_night) + minute(target_night)/60
target_night

time_start=hour(la_bikeshartripse$start_time) +minute(la_bikeshartripse$start_time)/60

la_bikeshartripse$date_compare=time_start
head(la_bikeshartripse)

#using start time to determine where it should go:

morning=subset(la_bikeshartripse,la_bikeshartripse$date_compare<target_afternoon)
afternoon=subset(la_bikeshartripse,target_afternoon<la_bikeshartripse$date_compare &
la_bikeshartripse$date_compare<target_night)
night=subset(la_bikeshartripse,target_night<la_bikeshartripse$date_compare)
head(morning)

duration_LA=matrix(nrow = 1,ncol = 3)
```

```
duration_LA[1,]=c(median(morning$duration),median(afternoon$duration),median(night$duration))
colnames(duration_LA)=c("morning/early morning(before12)","afternoon(after12 before 6)","night (after6)")
duration_LA
barplot(duration_LA,ylim = c(0,700), col = 3,main = "median duration of trip in LA in seconds", ylab = "seconds",cex.names = 0.8)



frequancy_LA=matrix(nrow = 1,ncol = 3)
frequancy_LA[1,]=c(nrow(morning),nrow(afternoon),nrow(night))
colnames(frequancy_LA)=c("morning/early morning(before12)","afternoon(after12 before 6)","night (after6)")
frequancy_LA
barplot(frequancy_LA,ylim = c(0,120000), col = 2,main = "median frequency of trips in LA in seconds", ylab = "number of trips")
nrow(night)

distance_LA=matrix(nrow = 1,ncol = 3)
distance_LA[1,]=c(median(morning$distance,na.rm = T),median(afternoon$distance,na.rm = T),median(night$distance,na.rm = T))
colnames(distance_LA)=c("morning(before12)","afternoon(after12 before 6)","night (after6)")
distance_LA
barplot(distance_LA,ylim = c(0,1000), col = 9,main = "median distance of trips in LA in seconds", ylab = "distance in meters")

###done. now for byarea



#============note: sf_bikeshareStat is far all bay

sf_bikeshareStat=readRDS( "/Users/rire948/Downloads/bikes/sf_bikeshare_trips.rds")
head(sf_bikeshareStat)
## so we need longitudes and latuudes for start and end so yeee

share_station=readRDS("/Users/rire948/Downloads/bikes/sf_bike_share_stations.rds")
share_station

share_station=share_station[,-c(2,5,6,7)]
```

```
share_station
share_station=unique(share_station,by="station_id")


sf_bikeshareStat_start=share_station
colnames(sf_bikeshareStat_start)=c("start_station_id","startLatitude","startLongitude")


sf_bikeshareStat=merge(sf_bikeshareStat,sf_bikeshareStat_start,by="start_station_id")
head(sf_bikeshareStat)
## now lets add the end lat and end long


colnames(sf_bikeshareStat_start)=c("end_station_id","end_latitude","end_longitude")


sf_bikeshareStat=merge(sf_bikeshareStat,sf_bikeshareStat_start,by="end_station_id")
head(sf_bikeshareStat)


sf_bikeshareStat$distance=distGeo(cbind(sf_bikeshareStat$startLongitude,sf_bikeshareStat$startLatitude),cbind(sf_bikeshareStat$end_longitude,sf_bikeshareStat$end_latitude))
head(sf_bikeshareStat)


## ok so i hve the starrt lat and long  and distance
## now to separater it into mornign and night


target_afternoon=as.POSIXct("12:00", format =  "%H:%M")
target_afternoon <- hour(target_afternoon) + minute(target_afternoon)/60
target_afternoon


target_night=as.POSIXct("18:00", format =  "%H:%M")
target_night <- hour(target_night) + minute(target_night)/60
target_night


time_start=hour(sf_bikeshareStat$start_date) +minute(sf_bikeshareStat$start_date)/60


sf_bikeshareStat$date_compare=time_start
head(head(sf_bikeshareStat))
#using start time to determine where it should go:
```

```
morningbay=subset(sf_bikeshareStat,sf_bikeshareStat$date_compare<target_afternoon)
afternoonbay=subset(sf_bikeshareStat,target_afternoon<sf_bikeshareStat$date_compare &
sf_bikeshareStat$date_compare<target_night)
nightbay=subset(sf_bikeshareStat,target_night<sf_bikeshareStat$date_compare)
nrow(morningbay)
head(afternoonbay)
head(nightbay)


baymatrix_duration=matrix(ncol = 3,nrow = 1)
baymatrix_duration[1,]=c(median(morningbay$duration_sec),median(afternoonbay$duration_sec),median(nightbay$duration_sec))
baymatrix_duration
colnames(baymatrix_duration)=c("morning/early morning(before12)","afternoon(after12 before 6)","night (after6)")
baymatrix_duration
barplot(baymatrix_duration,ylim = c(0,600),main = "Median duration of trip in BAY area",ylab = "seconds")

baymatrix=matrix(ncol = 3,nrow = 1)
baymatrix[1,]=c(nrow(morningbay),nrow(afternoonbay),nrow(nightbay))
baymatrix
colnames(baymatrix)=c("morning/early morning(before12)","afternoon(after12 before 6)","night (after6)")
options("scipen" = 20)

barplot(baymatrix,ylim = c(0,600000),main = "Median frequency of trip in BAY area",ylab = "frequency",cex.axis = 0.7)



baymatrix_distance=matrix(ncol = 3,nrow = 1)
baymatrix_distance[1,]=c(median(morningbay$distance),median(afternoonbay$distance),median(nightbay$distance))
baymatrix_distance
colnames(baymatrix_distance)=c("morning/early morning(before12)","afternoon(after12 before 6)","night (after6)")
barplot(baymatrix_distance,main = "median distance of trips in bay area",ylab = "distance in meters",ylim = c(0,1400))

###.6.For Bay Area bike share trips in San Francisco, how does bearing (angle) change at different times ofday?
```

Ricardo Rendon

```
###What can you conclude about traffic patterns in the city? Thegeosphere::bearing()functioncan compute bearings for longitude and latitude coordinates.
install.packages("geosphere")
library("geosphere")
library("lubridate")

#for LA:######
#head(la_bikeshartripse)
#morning$bearing=bearing(cbind(morning$start_lon,morning$start_lat),cbind(morning$end_lon,morning$end_lat))
#afternoon$bearing=bearing(cbind(afternoon$start_lon,afternoon$start_lat),cbind(afternoon$end_lon,afternoon$end_lat))
#night$bearing=bearing(cbind(night$start_lon,night$start_lat),cbind(night$end_lon,night$end_lat))

#median(morning$bearing,na.rm = T)
#median(afternoon$bearing,na.rm = T)
#median(night$bearing,na.rm = T)

#sf
head(sf_bikeshareStat)
share_station=readRDS("/Users/rire948/Downloads/bikes/sf_bike_share_stations.rds")
share_station=unique(share_station)
share_station
share_station=subset(share_station,share_station$landmark=="San Francisco")
share_station=share_station[,-c(2,5,7)]
share_station
colnames(share_station)=c("start_station_id","latitude", "longitude"  ,"landmark")
Sf6=merge(sf_bikeshareStat,share_station, by="start_station_id")
head(Sf6)
nrow(Sf6)
nrow(unique(Sf6,by="trip_id"))
Sf6$bearing=bearing(cbind(Sf6$startLongitude,Sf6$startLatitude),cbind(Sf6$end_longitude,Sf6$end_latitude))
## change to morning and afte and night
morningSF=subset(Sf6,Sf6$date_compare<target_afternoon)
afternoonSF=subset(Sf6,target_afternoon<Sf6$date_compare & Sf6$date_compare<target_night)
nightSF=subset(Sf6,target_night<Sf6$date_compare)
```

Ricardo Rendon

```
matrix_bearing_SF=matrix(nrow = 1,ncol = 3)

matrix_bearing_SF[1,]=c(median(morningSF$bearing),median(afternoonSF$bearing),median(nightSF$bearing))

colnames(matrix_bearing_SF)=c("morning","afternoon","night")

matrix_bearing_SF

boxplot(matrix_bearing_SF,main = "median bearing in sf data set",ylab = "bearing(angle)(truenorth=0)")

mean(matrix_bearing_SF)
```