# STA141HW3

*Sam Tsoi*

*11/12/2017*

## 1.

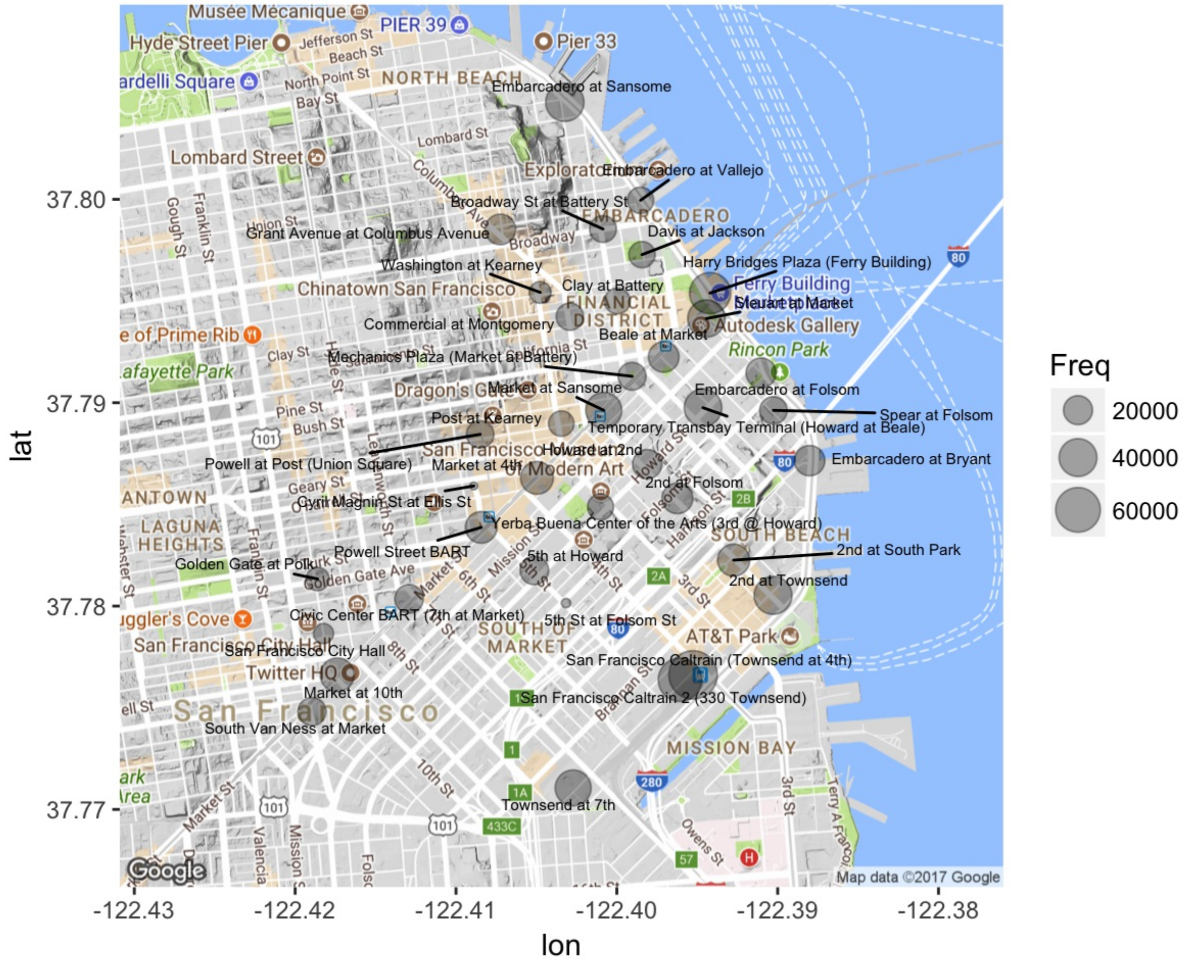
Write a function that loads the Bay Area bike share trip data from a CSV file, converts the columns to appropriate data types, and then saves the tidied data frame to an RDS file. Your function should have arguments to set the path for the input CSV file and the output RDS file. Write a second function that does the same thing for the Bay Area bike share station data.

Answer is attached in the code.

## 2.

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

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=37.787914,-122.40
3483&zoom=14&size=640x640&scale=2&maptype=terrain&language=en-EN
```



More stations are further away from the bart line, which makes sense because they built those stations for people to bike from one location to the BART station to catch the BART. BART is a pretty robust transportation system that people can rely on and since many people don't have cars in SF, people might rely on getting a bike from a bike station and then biking to the BART station to travel outside of SF. There also seems to be a lot of bike stations around the financial district, which might mean that people would need to bike to travel to and from work. I deduce that this might be because people take the BART, and then bike to work.

## 3.

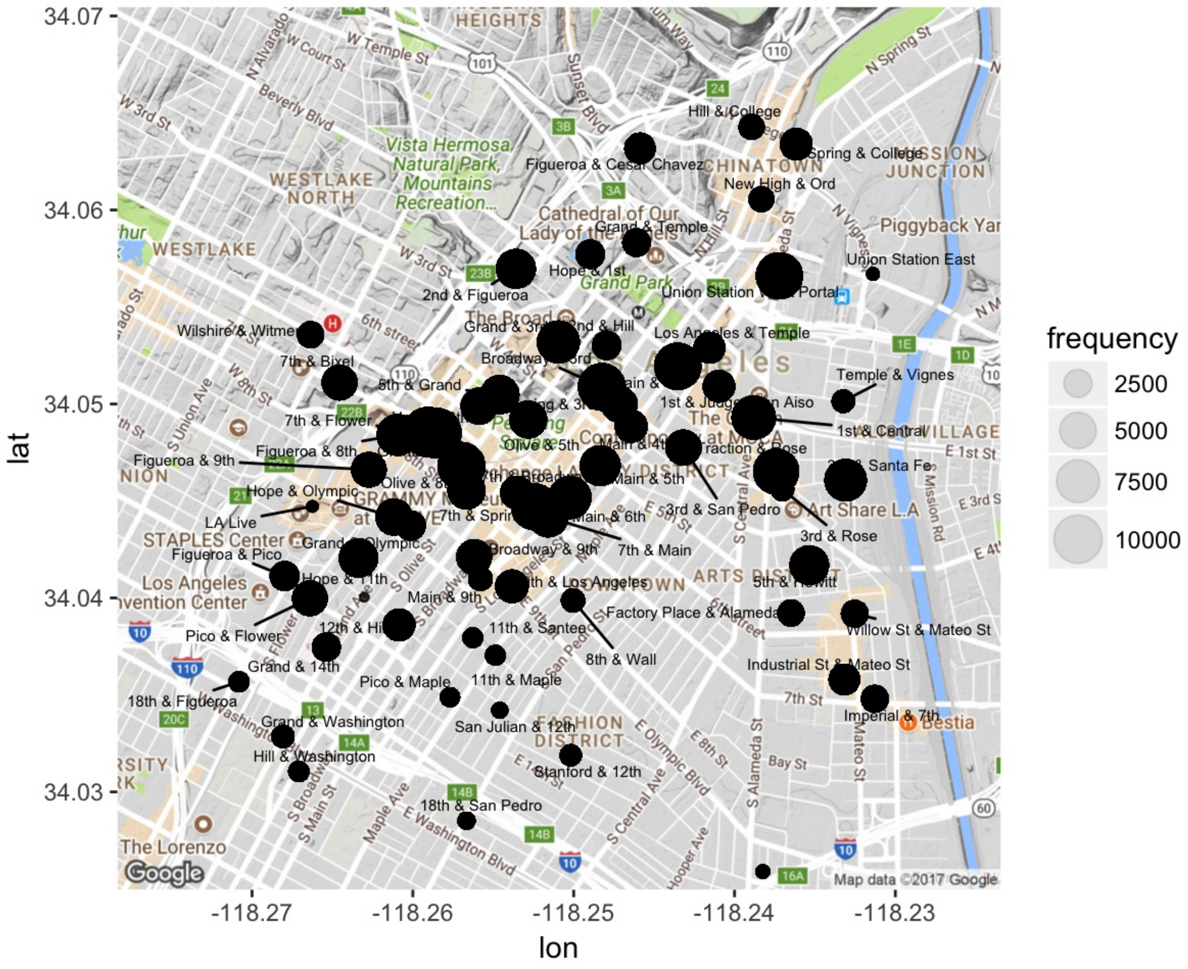
Write a function that loads the Los Angeles bike share trip data from the 5 provided CSV files, binds them into one data frame, converts the columns to appropriate data types, and saves the tidied data frame to an RDS file. Your function should have arguments to set the path for the input directory and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

Answer is attached in the code.

## 4.

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

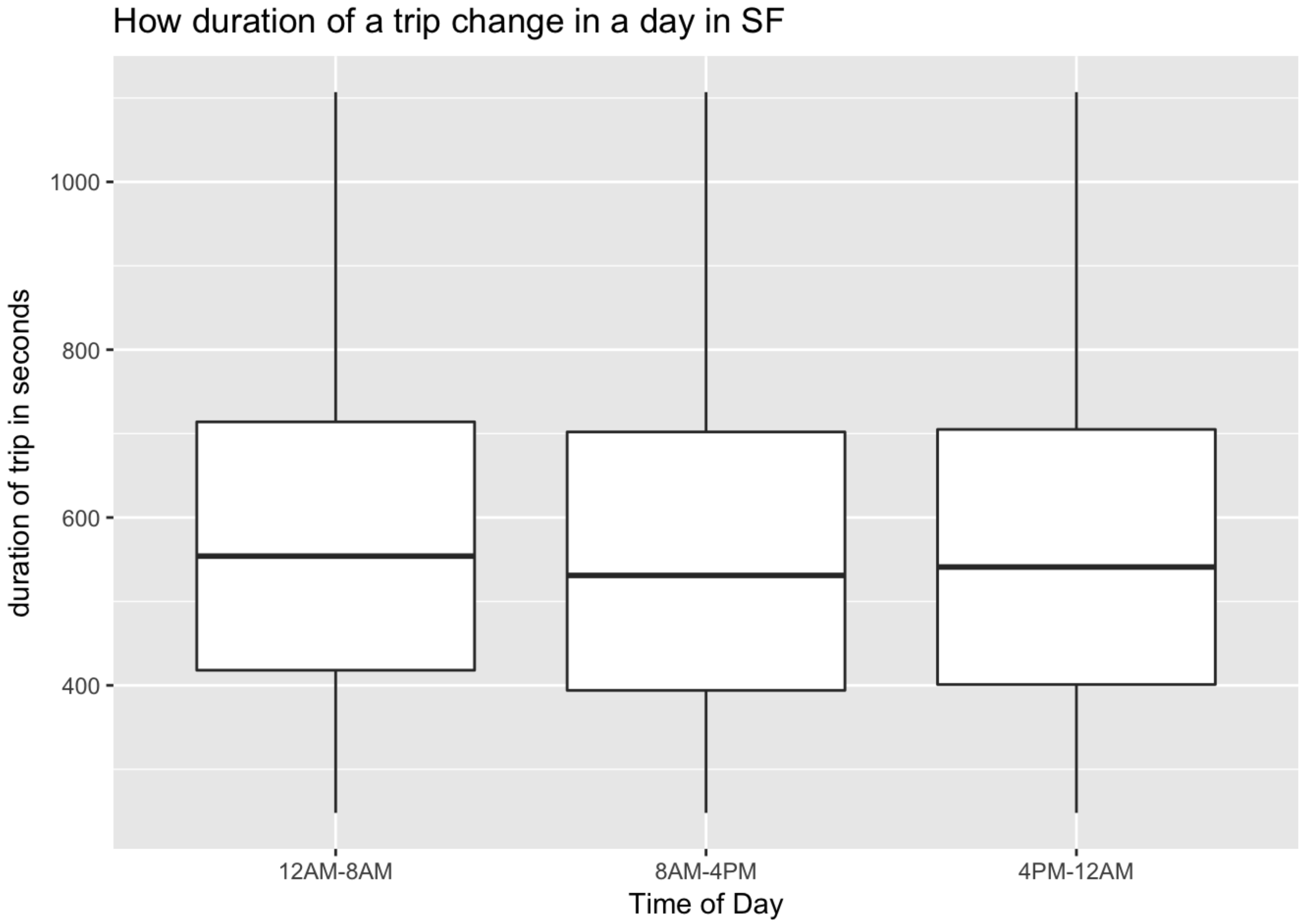
```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=34.047749,-118.25
095&zoom=14&size=640x640&scale=2&maptype=terrain&language=en-EN
```

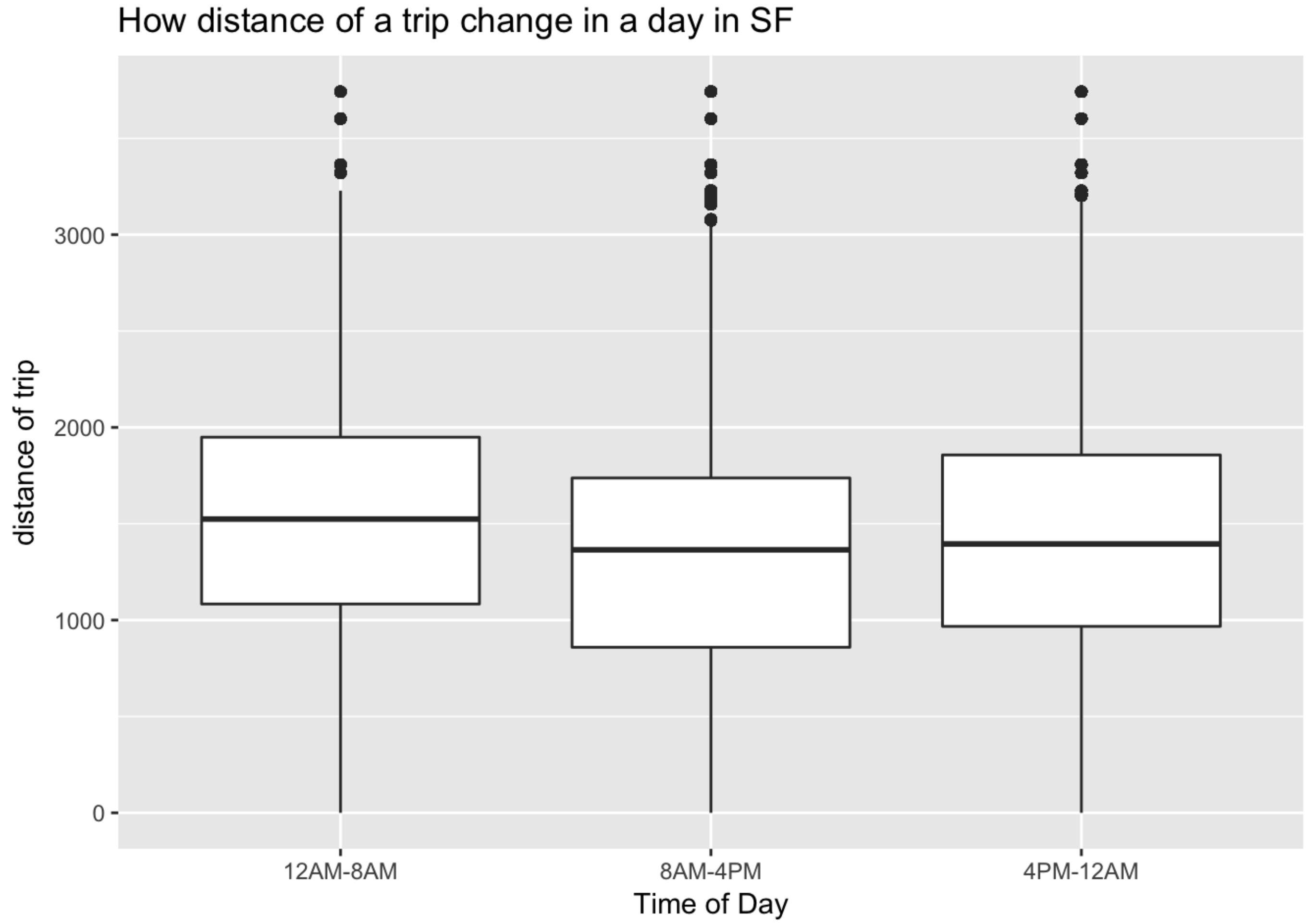


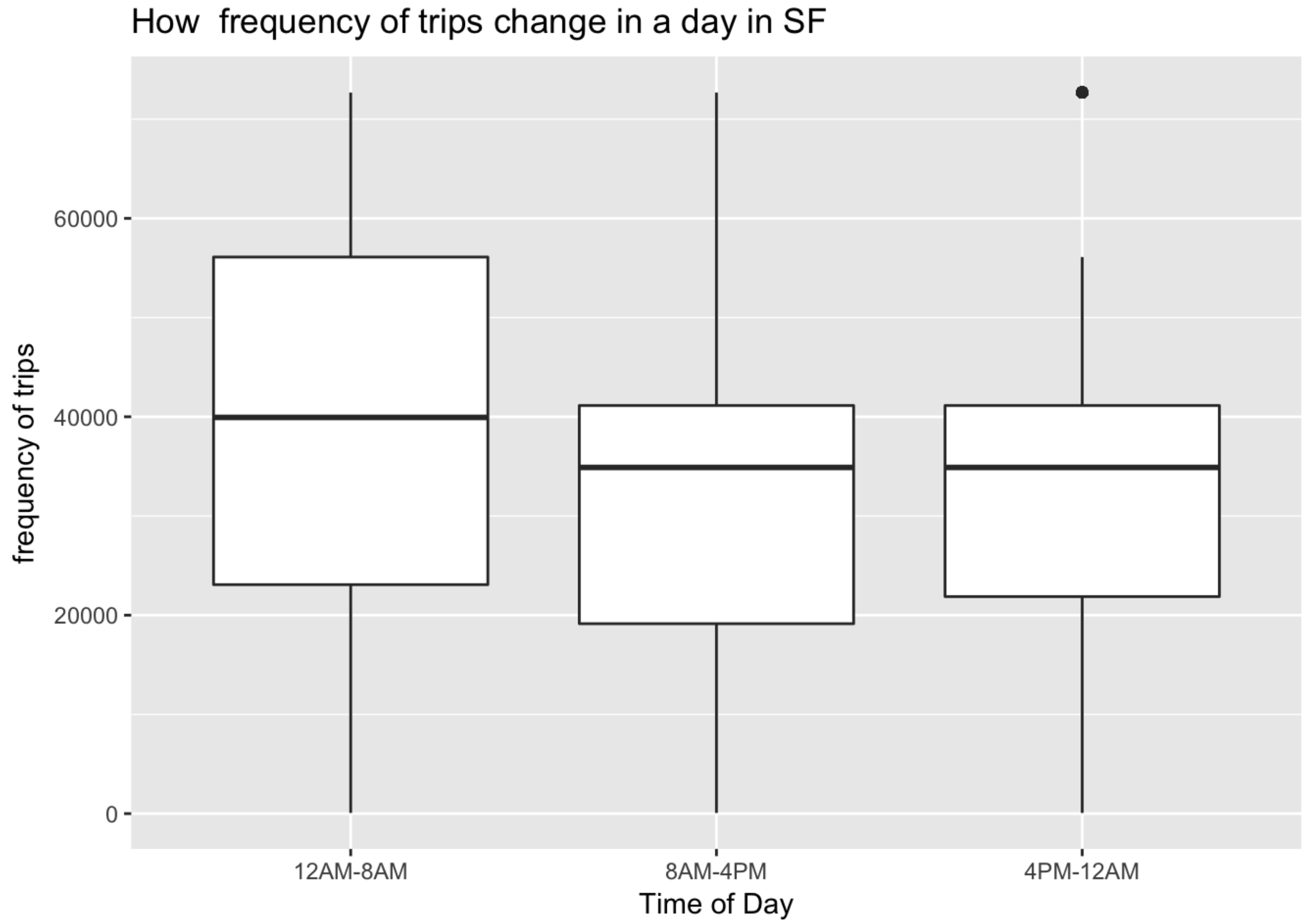
It seems like less people bike in LA compared with that of SF, with the frequencies way lower in LA than in SF. However, we also need to keep in mind that we are only looking at Downtown LA. This might be influenced by the very strong bike culture in the bay area, especially since it has been shown that traffic has increased over the years and that living prices has increased as well. Biking is definitely more cost efficient than driving, as one will not need gas for biking. Additionally, the BART system in the bay is pretty robust, and it makes sense that there are so many bike stations surrounding the BART stations in SF.

# 5.

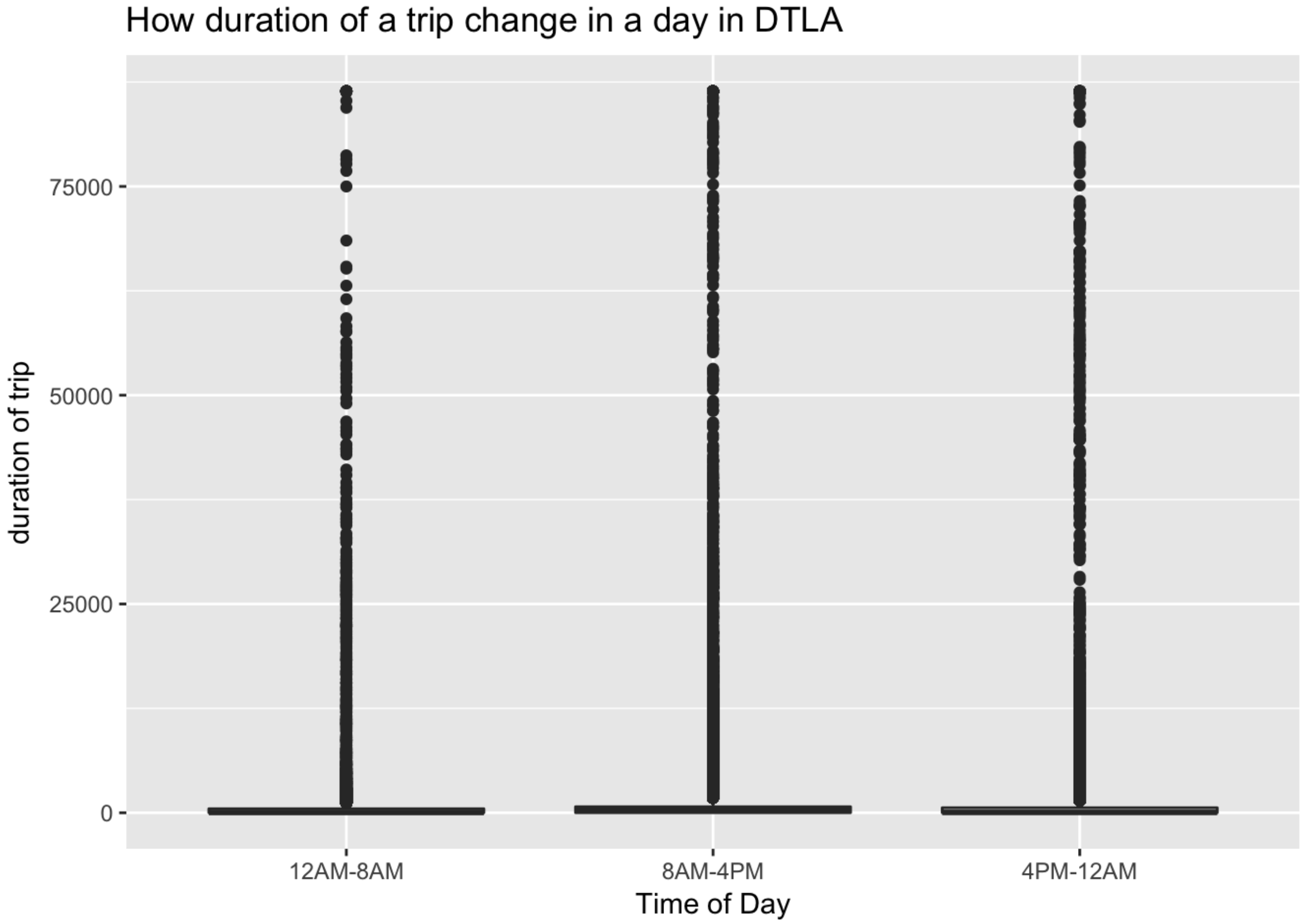
How do trip frequency, distance, and duration change at different times of day? Investigate for both the Bay Area bike share and the Los Angeles bike share. Compare your findings. The geosphere::distGeo()1 function can compute distances for longitude and latitude coordinates.





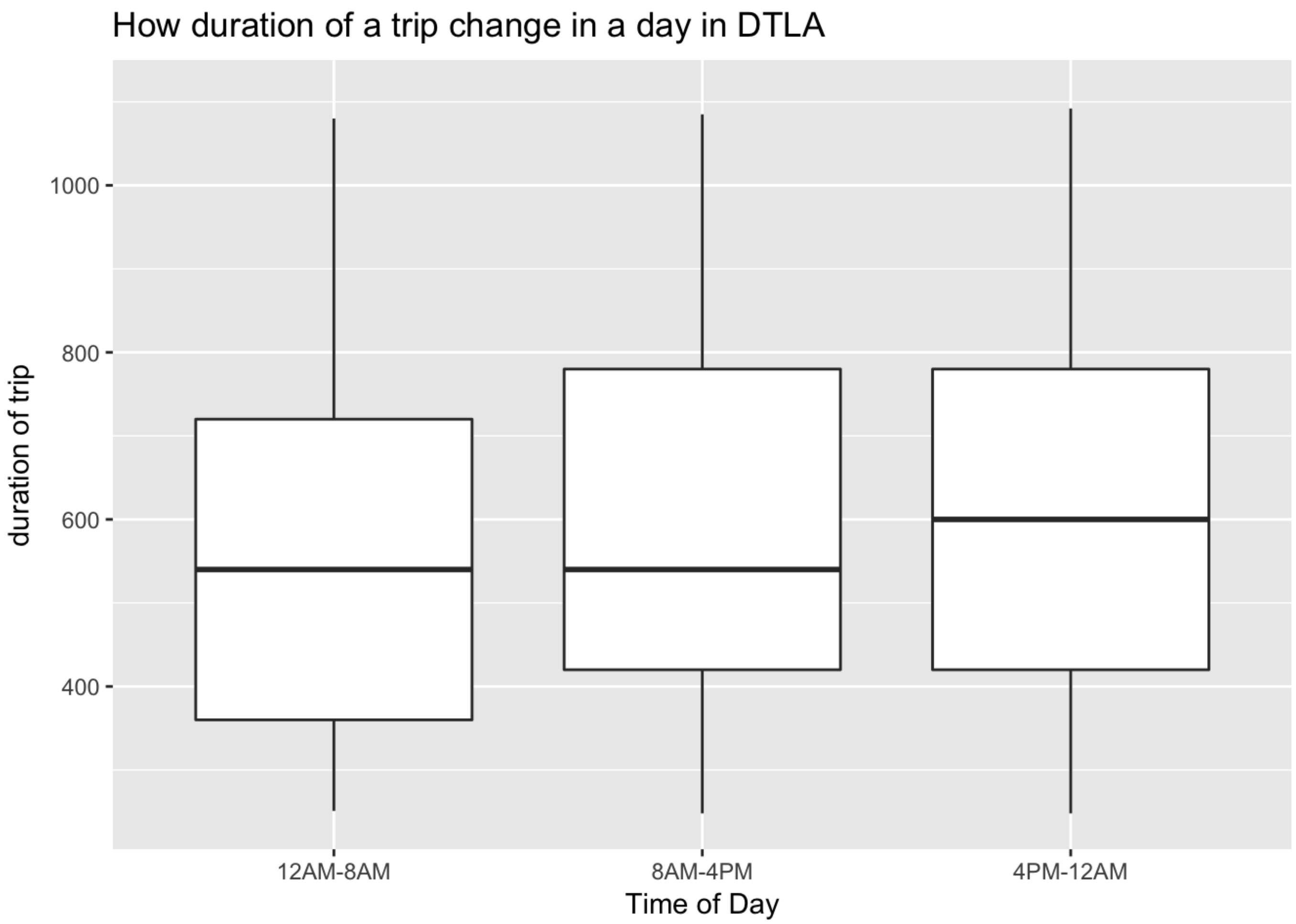


It seems like the duration of the trip in the bay generally stay the same in a day, separated by early morning, day, and night. All generally hover around 580 seconds. The distance of trip in the bay is drastically longer early in the morning, between 12AM-8AM, and then decreases during the day and night. The same pattern is observed for the frequency of trips in the early morning. These might be influenced by people needing to go to work in the morning. People might need to bike to work, but the duration doesn’t change. This is a pretty similar trend to that in DTLA, but as seen in the later plots, the duration increases at night in LA. Since duration doesn’t increase at night in SF (compared to that in DTLA), it might mean that people stroll more on bikes in DTLA later at night after work. Bikes might be a medium of transportation more in SF than in LA, and people in LA might use bikes more for leisure. More testing is needed to make a more definite conclusion.

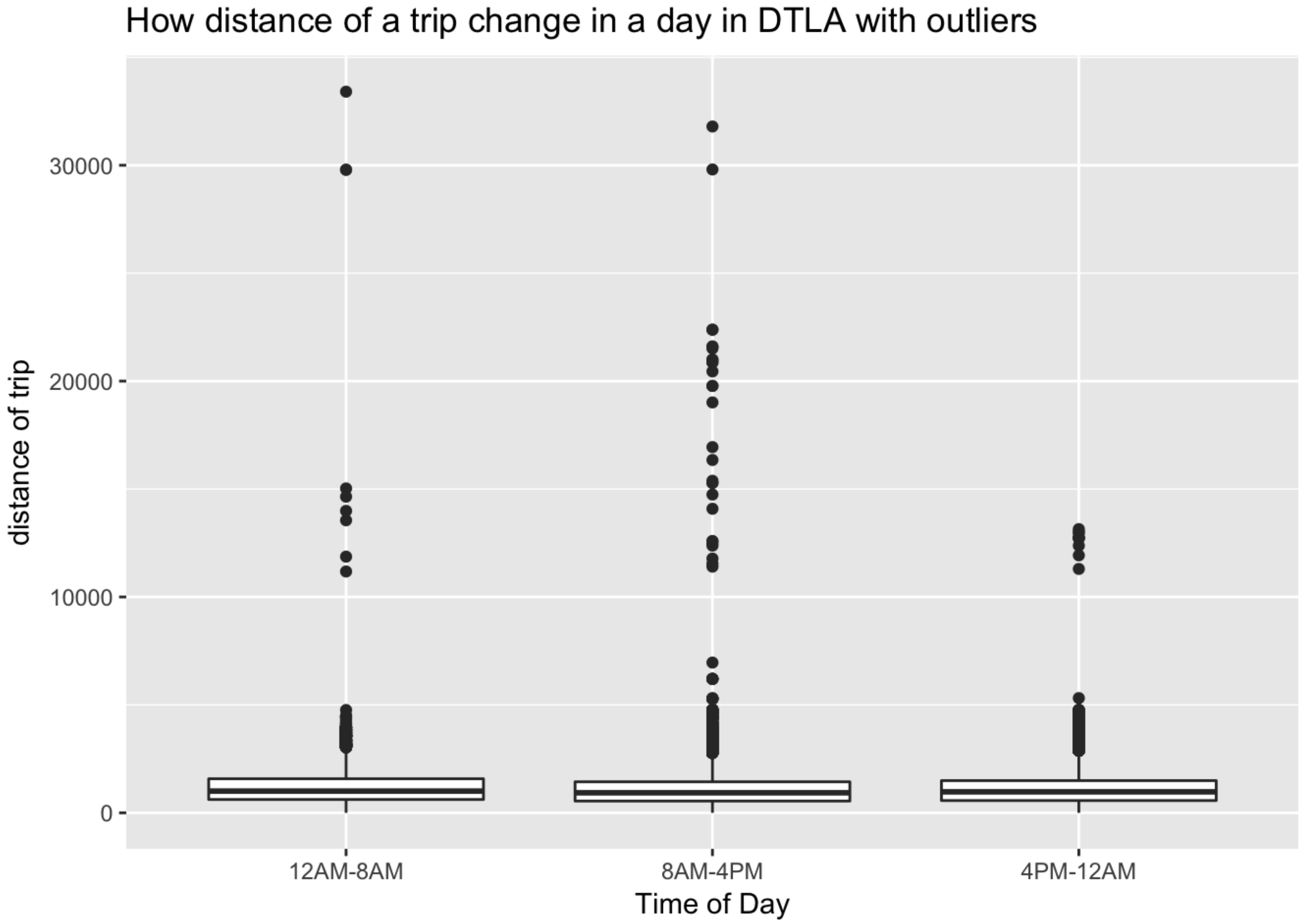


Removing the outliers to see the bulk of the data more clearly,

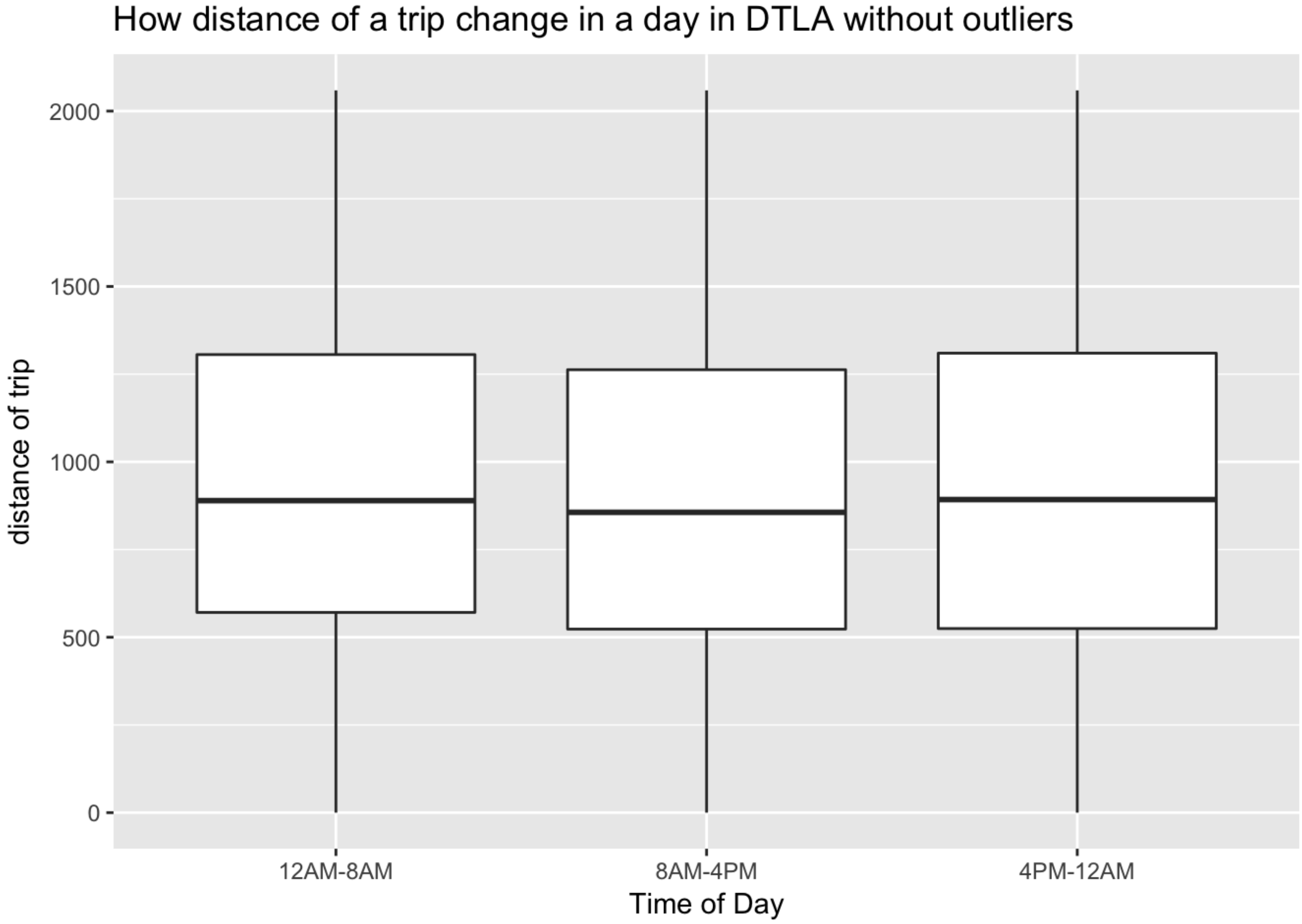
```
## Warning: Removed 167503 rows containing non-finite values (stat_boxplot).
```



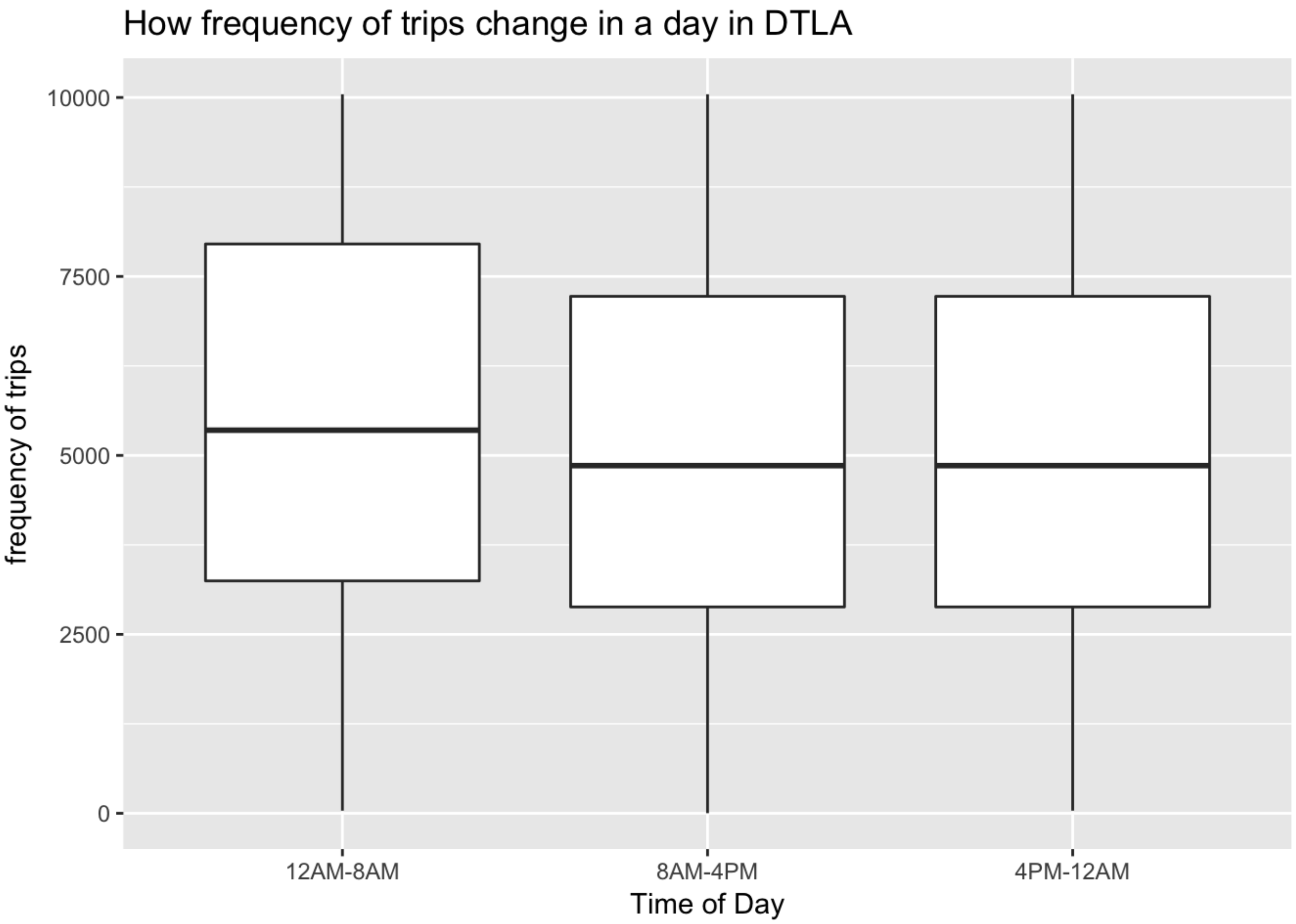
The duration of trips seem to drastically increase at night time, between 4PM-12AM. This might mean that people are using the bikes for longer when they are off work or traveling back home from work. Comparing 12AM-8AM and 8AM-4PM, the duration of trips for 8AM-4PM seems to be more skewed, so it means that the mean of duration of trips between 8AM-4PM is higher than that of 12AM-8AM.



Removing the outliers to see the bulk of the data more clearly,



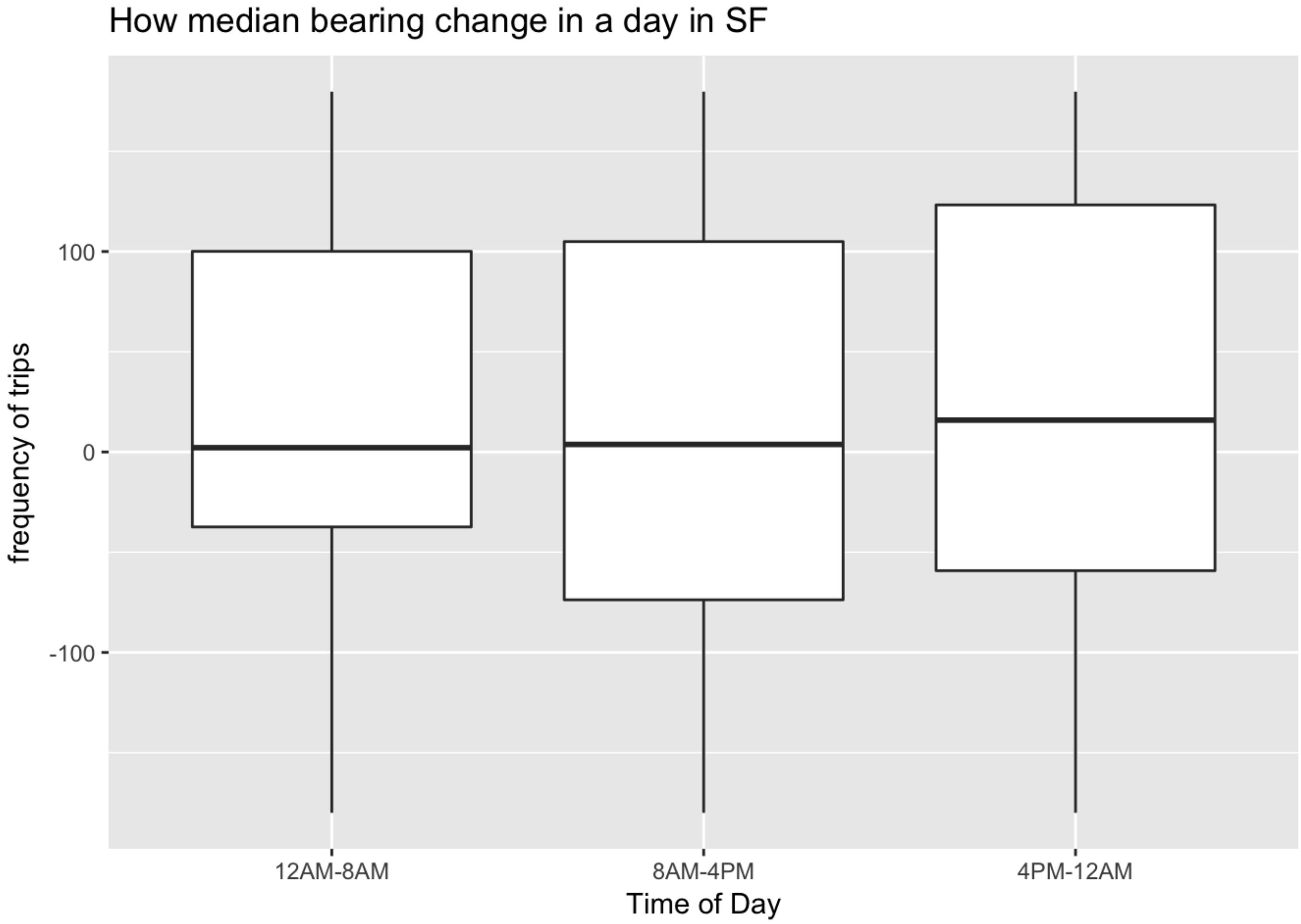
It seems like the distance of trips are pretty similar throughout the day. People in LA seem to use the bike station pretty equally throughout the day.



Based on the plot, the frequnecy of trips is drastically larger in the early morning, between 12AM and 8AM. This is pretty interesting, because the frequency of trips is drastically larger in the early morning but the duration of trips is drastically longer at night, between 4PM-12AM. I predict that this might be due to more people needing to use the bikes in the early morning to get to work, as it might be the fastest transportation medium for short distances. However, people might be using bikes more leisurely at night after work, and for longer, so it might not be to get to one location to another location quickly, but rather to stroll on a bike.

# 6.

For Bay Area bike share trips in San Francisco, how does bearing (angle) change at different times of day? What can you conclude about traffic patterns in the city? The geosphere::bearing() function can compute bearings for longitude and latitude coordinates.



The angle in SF seem to increase at night. This probably means that the traffic around the time when work ends goes on a more upward, positive angle, which is the direction of Highway 80, towards Oakland. I deduce that this might be because many people who work in SF live elsewhere because it is generally cheaper in other areas other than the city. Conversely, it seems like the bearing is lower in the morning time, possibly when people arrive at work, all for the reverse reason. I think it is pretty interesting that the quartile is drastically smaller for the early morning, between 12AM-8AM, and also that it is skewed. This also seems to mean that a lot of traffic is from the northwest to southeast, so there is not a lot of traffic going to and from SF and Marin county area. The median bearings are all positive numbers, which is more in the direction of Highway80 towards Oakland rather than other parts of SF. This is pretty interesting because this would mean that Highway 80, as seen on the map, would get a lot of traffic since it is the only bridge.

# Citations:

Looked on Piazza for tips on ggmap, get_map Worked with Brody Lowry on #3, #5 (for do.call and function for lapply) Consulted dicussion notes

# Code Appendix

```
#1
baybiketrip <- read.csv("~/Desktop/Downloads/bikes/sf_bikeshare_trips.csv")
#summary(baybiketrip)
baybiketrip$start_date = as.POSIXlt(baybiketrip$start_date)
baybiketrip$end_date = as.POSIXlt(baybiketrip$end_date)
baybiketrip$trip_id = as.factor(baybiketrip$trip_id)
baybiketrip$bike_number = as.factor(baybiketrip$bike_number)
baybiketrip$start_station_id = as.factor(baybiketrip$start_station_id)
saveRDS(baybiketrip, "sf_bikeshare_trips.rds")

baybikeshare <- read.csv("~/Desktop/Downloads/bikes/sf_bike_share_stations.csv")
#summary(baybikeshare)

baybikeshare$station_id = as.factor(baybikeshare$station_id)
baybikeshare$name = as.factor(baybikeshare$name)
baybikeshare$installation_date = as.POSIXlt(baybikeshare$installation_date)


saveRDS(baybikeshare, "sf_bike_share_stations.rds")

#2
#install.packages("ggmap")
#install.packages("ggplot2")
#install.packages("readr")
#install.packages("sf")
#install.packages("lubridate")
library(ggmap)
library(ggplot2)

library(lubridate)
library(readr)
library(sf)
#install.packages("ggrepel")
library(ggrepel)
#install.packages("devtools")
devtools::install_github("dkahle/ggmap")
#devtools::install_github("hadley/ggplot2")
#install.packages("geosphere")
library("geosphere")
baystation <- readRDS("sf_bike_share_stations.rds")
baytripdf <-readRDS("sf_bikeshare_trips.rds")
baystationdf <- subset(baystation, baystation$landmark=="San Francisco" &duplicated
(baystation$station_id)==FALSE)
#summary(baytripdf)
#summary(baystationdf)
#ggplot(baystationdf, aes(longitude, latitude))
startTab <- table(baytripdf$start_station_id)
startFrame <- as.data.frame(startTab)
baystationdf <- merge(baystationdf, startFrame, by.x = "station_id", by.y = c("Var
1"))
#library(ggrepel)
loc = sapply(baystationdf[c("longitude", "latitude")], function(longitude) mean(ran
ge(longitude)))
m = get_map(loc, zoom = 14)
```

```r
agg.data <- aggregate(cbind(longitude,latitude) ~ name, data = baystationdf, mean)
#agg.data only save the data that is unique
ggmap(m, xlab = "Longitude", ylab="Latitude", legend="right") + geom_point(data=baystationdf, aes(x=longitude, y=latitude, size=Freq), alpha =  I(1/3)) + geom_text_repel(data = agg.data, aes(x = longitude, y = latitude, label = name),size=2) + scale_size(range = c(1, 8))
#geom_density_2d(aes(x = longitude, y = latitude), baysharedf)
#3
#q316 <- read.csv("~/Desktop/Downloads/bikes/2016_q3_la_metro_trips.csv")
#q416 <- read.csv("~/Desktop/Downloads/bikes/2016_q4_la_metro_trips.csv")
#q117 <- read.csv("~/Desktop/Downloads/bikes/2017_q1_la_metro_trips.csv")
#q217 <- read.csv("~/Desktop/Downloads/bikes/2017_q2_la_metro_trips.csv")
#q317 <- read.csv("~/Desktop/Downloads/bikes/2017_q3_la_metro_trips.csv")
#names(q316) <- gsub("station_id", "station", names(q316))
#q316$start_time <- parse_date_time(q316$start_time, orders = c("y-m-d H:M:S", "m/d/y H:M"))
#q316$end_time <- parse_date_time(q316$end_time, orders = c("y-m-d H:M:S", "m/d/y H:M"))
#q316$trip_id = as.factor(q316$trip_id)
#q316df <- as.data.frame(q316)
#saveRDS(q316, "q316.rds")
#path = c("~/Desktop/Downloads/bikes/2016_q3_la_metro_trips.csv","~/Desktop/Downloads/bikes/2016_q4_la_metro_trips.csv", "~/Desktop/Downloads/bikes/2017_q1_la_metro_trips.csv", "~/Desktop/Downloads/bikes/2017_q2_la_metro_trips.csv","~/Desktop/Downloads/bikes/2017_q3_la_metro_trips.csv")

download_dir<-"~/Desktop/Downloads/bikes/"
pathLAIndex <- grep("la_metro",list.files(download_dir))
pathLAIndex <- grep("\\.csv$",list.files(download_dir)[pathLAIndex])
relPathLA <- list.files(download_dir)[pathLAIndex]
fullPathLA <- paste0(download_dir,list.files(download_dir)[pathLAIndex])

labikes <- lapply(1:5, function(x) {
labike <- read.csv(fullPathLA[x])
names(labike) <- gsub("station_id", "station", names(labike))
labike$start_time<-parse_date_time(labike$start_time,c("m/d/y H:M","y-m-d H:M:S"))
labike$end_time<-parse_date_time(labike$end_time,c("m/d/y H:M","y-m-d H:M:S"))
#labike$start_time <- parse_date_time(labike$start_time, orders = c("y-m-d H:M:S", "m/d/y H:M"))
#labike$end_time <- parse_date_time(labike$end_time, orders = c("y-m-d H:M:S", "m/d/y H:M"))
labike$trip_id = as.factor(labike$trip_id)
labike$bike_id = as.factor(labike$bike_id)
labike <- as.data.frame(labike)
#relPatLA <- gsub("(.+bikes/|\\.csv$)","",fullPathLA) #parsing through the full path name to give simple file name when saving to RDS
#saveRDS(labikedf, file = paste0(relPatLA[x],".rds"))
})

labikestrip <- do.call(rbind, labikes)

# Write a second function that loads, tidies, and saves the Los Angeles bike share station data.
labikeshare<- read.csv("~/Desktop/Downloads/bikes/metro-bike-share-stations-2017-10-20.csv")
labikeshare$Station_ID <- as.factor(labikeshare$Station_ID)
labikeshare$Go_live_date <- parse_date_time(labikeshare$Go_live_date, orders = c("
y-m-d", "m/d/y"))
saveRDS(labikeshare, "metro-bike-share-stations-2017-10-20.rds")
lastation <- readRDS("~/Desktop/Downloads/bikes/metro-bike-share-stations-2017-10-20.rds")
lastationdf <- subset(lastation, lastation$Region=="DTLA" &duplicated(lastation$Station_ID)==FALSE)
#making a new data frame that has all the data of trip plus counts the frequency of start station
startTabLA <- table(labikestrip$start_station)
startFrameLA <- as.data.frame(startTabLA)
names(startFrameLA) <- c("start_station","frequency")
labikedf <- merge(labikestrip, startFrameLA, by = "start_station", na.rm=TRUE)

labikedf <- na.omit(labikedf)

m = get_map(location = c(lon = median(labikedf$start_lon), lat = median(labikedf$start_lat)), zoom = 14)
agg.data <- aggregate(cbind(start_lon,start_lat) ~ start_station, data = labikedf, mean) #agg.data only save the data that is unique

agg.datamerge <- merge(agg.data, lastationdf, by.x = "start_station", by.y = "Station_ID")
ggmap(m, xlab = "Longitude", ylab="Latitude", legend="right") + geom_point(data=labikedf, aes(x=start_lon, y=start_lat, size=frequency), alpha =(1/10))  + geom_text_repel(data = agg.datamerge, aes(x=start_lon, y=start_lat, label = Station_Name),size=2) + scale_size(range = c(1, 8))



##5
sfstation <- readRDS("sf_bike_share_stations.rds")

sfstationdf <- subset(sfstation, sfstation$landmark=="San Francisco")
sfids<-unique(sfstationdf$station_id)
sftripsid <-as.numeric(baytripdf$start_station_id) %in% sfids & as.numeric(baytripdf$end_station_id) %in% sfids
sftrips <- baytripdf[sftripsid,]
station_lon<-aggregate(longitude~station_id,sfstationdf,median)
station_lat<-aggregate(latitude~station_id,sfstationdf,median)

sfstartlonlat <- merge(station_lon,station_lat, by = "station_id")
names(sfstartlonlat) <- c("start_station_id", "startmedlon", "startmedlat")
sftrips <- merge(sftrips, sfstartlonlat, by = "start_station_id")

sfendlonlat <- merge(station_lon,station_lat, by = "station_id")
names(sfendlonlat) <- c("end_station_id", "endmedlon", "endmedlat")
sftrips <- merge(sftrips, sfendlonlat, by = "end_station_id")
#frequency of start station
sftrips <- merge(sftrips, data.frame(table(sftrips$start_station_id)), by.x = "start_station_id", by.y = "Var1")

#distance
sftrips$distance<-distGeo(data.frame(sftrips$startmedlon, sftrips$startmedlat),data.frame(sftrips$endmedlon, sftrips$endmedlat))
# parsing by time of day
sftrips$start_date <- as.POSIXct(sftrips$start_date)
sftrips$timeofday<-cut(hour(sftrips$start_date),c(-1,8,16,25),c("12AM-8AM","8AM-4PM","4PM-12AM"))
```

```
ggplot(data = sftrips, aes (x= timeofday, y=duration_sec)) + geom_boxplot(outlier.
shape=NA) + labs(x="Time of Day", y= "duration of trip in seconds", title = "How du
ration of a trip change in a day in SF") + scale_y_continuous(limits = quantile(sft
rips$duration_sec, c(0.1, 0.9)))
ggplot(data = sftrips, aes (x= timeofday, y=distance)) + geom_boxplot() + labs(x="
Time of Day", y= "distance of trip", title = "How distance of a trip change in a da
y in SF")
ggplot(data = sftrips, aes (x= timeofday, y=Freq)) + geom_boxplot() + labs(x="Time
of Day", y= "frequency of trips", title = "How  frequency of trips change in a day
in SF")

####LA

lastation <-readRDS("metro-bike-share-stations-2017-10-20.rds")
lastationdf <- subset(lastation, lastation$Region=="DTLA")
dtla_ids<-unique(lastation$Station_ID)

intra_metro_trips<-as.numeric(labikedf$start_station) %in% dtla_ids & as.numeric(la
bikedf$end_station) %in% dtla_ids
latrips<-labikedf[intra_metro_trips,]
lastation$Station_ID=factor(lastation$Station_ID)
#Make these numeric
latrips$start_lon<-as.numeric(latrips$start_lon)
latrips$start_lat<-as.numeric(latrips$start_lat)
latrips$end_lon<-as.numeric(latrips$end_lon)
latrips$end_lat<-as.numeric(latrips$end_lat)

#Get the station locations, use median because some long/lat differ for a station
station_lon<-aggregate(start_lon~start_station,latrips,median)
station_lat<-aggregate(start_lat~start_station,latrips,median)
end_lon<-aggregate(end_lon~end_station,latrips,median)
end_lat<-aggregate(end_lat~end_station,latrips,median)
station_loc<-data.frame(Station_ID=station_lon[,1],startmedlon=station_lon[,2],sta
rtmedlat=station_lat[,2])
end_loc <- data.frame(Station_ID=end_lon[,1],endmedlon=end_lon[,2],endmedlat=end_l
at[,2])
#add location to df
latrips <- merge(latrips, station_loc, by.x = "start_station", by.y = "Station_ID"
)
latrips <- merge(latrips, end_loc, by.x = "end_station", by.y = "Station_ID")

#frequency of start station
lastation <- merge(lastation, data.frame(table(latrips$start_station)), by.x = "Sta
tion_ID", by.y = "Var1")

#distance
latrips$distance<-distGeo(data.frame(latrips$startmedlon, latrips$startmedlat),dat
a.frame(latrips$endmedlon, latrips$endmedlat))
# parsing by time of day
latrips$timeofday<-cut(hour(latrips$start_time),c(-1,8,16,25),c("12AM-8AM","8AM-4P
M","4PM-12AM"))
ggplot(data = latrips, aes (x= timeofday, y=duration)) + geom_boxplot() + labs(x="
Time of Day", y= "duration of trip", title = "How duration of a trip change in a da
y in DTLA")
ggplot(data = latrips, aes (x= timeofday, y=duration)) + geom_boxplot(outlier.shap
e=NA) + labs(x="Time of Day", y= "duration of trip", title = "How duration of a tri
p change in a day in DTLA") + scale_y_continuous(limits = quantile(sftrips$duration
p change in a day in DTLA") + scale_y_continuous(limits = quantile(sftrips$duration
_sec, c(0.1, 0.9)))
ggplot(data = latrips, aes (x= timeofday, y=distance)) + geom_boxplot() + labs(x="
Time of Day", y= "distance of trip", title = "How distance of a trip change in a da
y in DTLA with outliers")
ggplot(data = latrips, aes (x= timeofday, y=distance)) + geom_boxplot(outlier.shap
e=NA) + labs(x="Time of Day", y= "distance of trip", title = "How distance of a tri
p change in a day in DTLA without outliers") + scale_y_continuous(limits = quantile
(latrips$distance, c(0.1, 0.9)))
ggplot(data = latrips, aes (x= timeofday, y=frequency)) + geom_boxplot() + labs(x=
"Time of Day", y= "frequency of trips", title = "How frequency of trips change in a
day in DTLA")

###6
sftrips$bearing <- bearing(data.frame(sftrips$startmedlon,sftrips$startmedlat), da
ta.frame(sftrips$endmedlon, sftrips$endmedlat))

ggplot(data = sftrips, aes (x= timeofday, y=bearing)) + geom_boxplot() + labs(x="T
ime of Day", y= "frequency of trips", title = "How median bearing change in a day i
n SF")
```