Untitled

Are citibikes actually used for commuting ?

In this part, we will try to identify geographical trends and see whether they can constitute evidence that citibikes are used by new yorkers as a way of commuting.To do so, we will mainly focus on geographical features (stations locations for the start and end of the trip) while discarding aspects such as seasonnality.

```
library(readr)
library(sp)
library(plyr)
library(dplyr)
library(ggmap)
library(ggplot2)
concise_trips <- read_csv("../concise_trips.csv")
colnames(concise_trips)[4] <- 'start_station_id'
colnames(concise_trips)[5] <- 'end_station_id'
concise_trips <- subset(concise_trips, select=-c(starttime,stoptime,bikeid,gender))
```

```
stations_info <- read_csv("../stations_info.csv")
colnames(stations_info)[1] = 'start_station_id'
concise_trips_merged <- merge(x = concise_trips, y = stations_info, by = 'start_station_id')

colnames(stations_info)[1] = 'end_station_id'
concise_trips_merged <- merge(x = concise_trips_merged, y = stations_info, by = 'end_station_id')
```

Including Plots

```
#COMPUTING THE MOST STARTED STATIONS
colnames(stations_info)[1] = 'start_station_id'
most_started <- count(concise_trips_merged, c(concise_trips_merged$start_station_id))
colnames(most_started)[1] = 'start_station_id'
most_started_merged <- merge(x = most_started, y = stations_info, by = 'start_station_id')
most_started_merged <- most_started_merged[order(-most_started_merged$n),]

#COMPUTING THE MOST ENDED STATIONS
colnames(stations_info)[1] = 'end_station_id'
most_ended <- count(concise_trips_merged, c(concise_trips_merged$end_station_id))
colnames(most_ended)[1] = 'end_station_id'
most_ended_merged <- merge(x = most_ended, y = stations_info, by = 'end_station_id')
most_ended_merged  <- most_ended_merged[order(-most_ended_merged$n),]
```

```
#AGGREGATING THE TWO TO SEE THE MOST ACTIVE STATIONS
colnames(most_started_merged)[1] = 'station_id'
colnames(most_started_merged)[2] = 'n_s'
colnames(most_ended_merged)[1] = 'station_id'
colnames(most_ended_merged)[2] = 'n_e'
most_active <- merge(x = most_started_merged, y = most_ended_merged, by = 'station_id')
most_active$n <- most_active$n_e + most_active$n_s
most_active <- subset(most_active, select=-c(n_s,n_e,longitude.y,latitude.y,name.y))
most_active <- most_active[order(-most_active$n),]
```

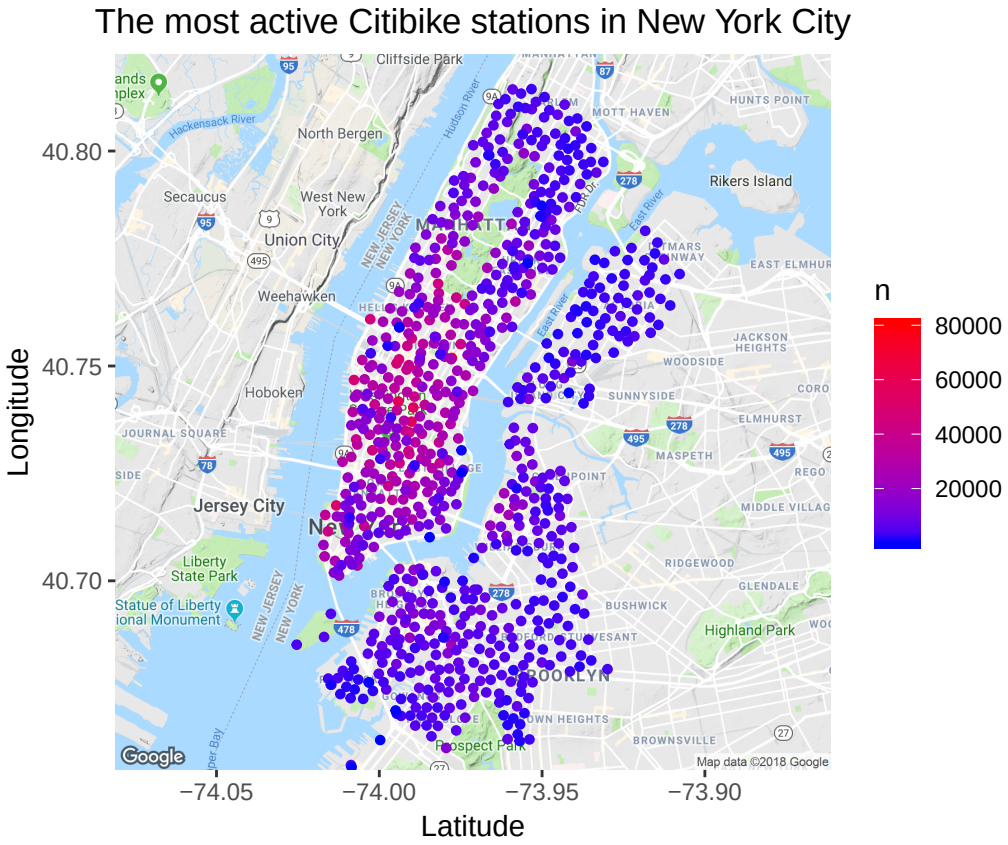
```
colnames(most_active)[2] <- 'names'
colnames(most_active)[3] <- 'latitude'
colnames(most_active)[4] <- 'longitude'
```

First of all, we can visualize the stations that are the most 'active': those are the stations that have the highest traffic of citibikes (in terms of start and end for each station)

```
nyc_base <- ggmap::get_map(location = c(lon = mean(most_active$longitude), lat = mean(most_active$latitu
```

```
FALSE Source : https://maps.googleapis.com/maps/api/staticmap?center=40.73938,-73.971375&zoom=12&size=6
```

```
ggmap(nyc_base) + geom_point(data=most_active, aes(x=longitude, y=latitude,color=n), size=1.2)+
  scale_colour_gradient(low = "blue", high = "red")+
  xlab("Latitude")+
  ylab("Longitude")+
  ggtitle("The most active Citibike stations in New York City")+
  theme(plot.title = element_text(hjust = 0.5))+ scale_fill_discrete(name = "Activity (in number of cust
```



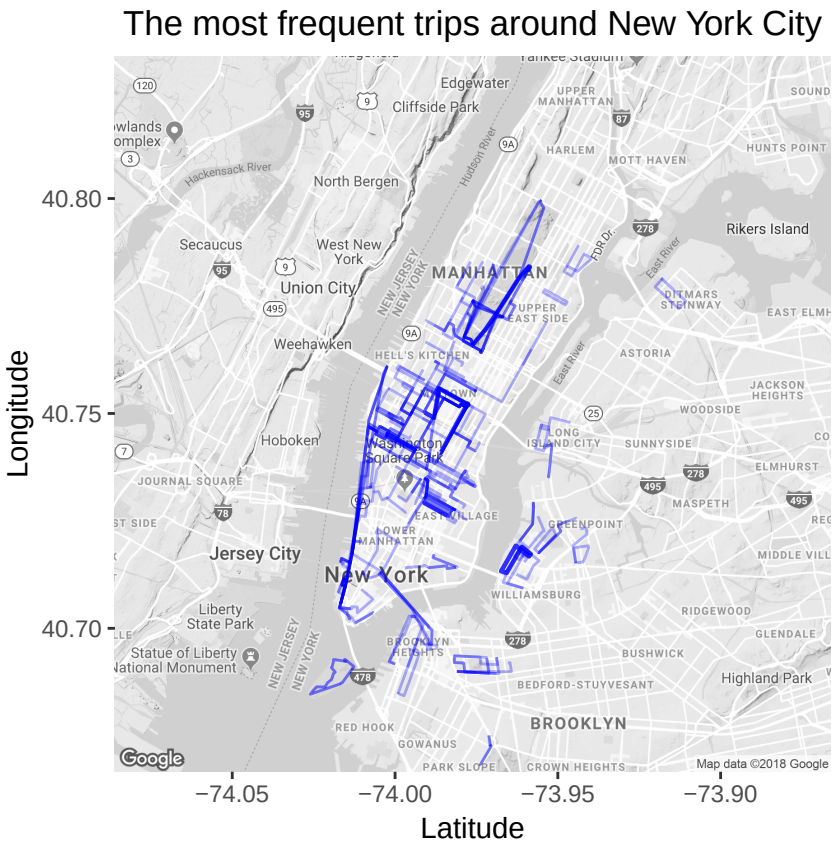
From the map, we can see that the there are multiple 'active' zones but the general trend is that Midtown is the most active part of Manhattan. There is also quite a bit of activity in Downtown, along Central Park and in the neighborhoods of Brooklyn that are close to Manhattan. Areas close to Central Park, or in Brooklyn are zones mostly visited by tourists. However, the high activity of citibikes in Midtown could be evidence that citibikes are used for commuting since Midtown is a working area of Manhattan. Furthermore, the activity in Downtown also seems more likely to be the result of commuting then visiting.

To go further in our analysis, we can visualize the most common trips to see whether we can identify trips indicative of either commuting or not.

```
concise_trips_merged_shorted <- count_(concise_trips_merged, vars=c('start_station_id','end_station_id'
concise_trips_merged_shorted <- concise_trips_merged_shorted[order(-concise_trips_merged_shorted$n),]
concise_trips_merged_shorted <- concise_trips_merged_shorted[1:1000,]

colnames(stations_info)[1] = 'start_station_id'
concise_trips_merged_shorted <- merge(x = concise_trips_merged_shorted, y = stations_info, by = 'start_


colnames(stations_info)[1] = 'end_station_id'
concise_trips_merged_shorted <- merge(x =concise_trips_merged_shorted, y = stations_info, by = 'end_sta
concise_trips_merged_shorted <- concise_trips_merged_shorted[order(-concise_trips_merged_shorted$n),]

concise_trips_merged_shorted['from'] <- paste(concise_trips_merged_shorted$latitude.x,',', concise_trip
concise_trips_merged_shorted['to'] <- paste(concise_trips_merged_shorted$latitude.y,',', concise_trips_
```

```
library(stringr)
options(expressions=10000)

nyc_base <- ggmap::get_map(location = c(long = -73.97634,lat = 40.75000),zoom=12,color='bw',maptype = '
nyc_base <- ggmap(nyc_base)


leg <-function(start, dest,n){
    if(n< 10){
      size_value <- 0.8
      alpha_value <- 0.8
      color <- 'blue'}
    else{
      size_value <- 0.5
      alpha_value <- 0.3
      color <- 'blue'
    }

    r<- route(from=start,to=dest,mode = c("bicycling"),structure = c("route"))
    c<- geom_path(aes(x = lon, y = lat),
              alpha = alpha_value,size = size_value,  data = r, colour = color)
    return (c)
}

base <- nyc_base

 for (n in 1:300){

     l<-leg(str_replace_all(concise_trips_merged_shorted$from[n], " ", ""), str_replace_all(concise_tri

     base<-base+l
 }



base+
  xlab("Latitude")+
  ylab("Longitude")+
  ggtitle("The most frequent trips around New York City")+
```

```
  theme(plot.title = element_text(hjust = 0.5))
```



The above map provides some interesting information:

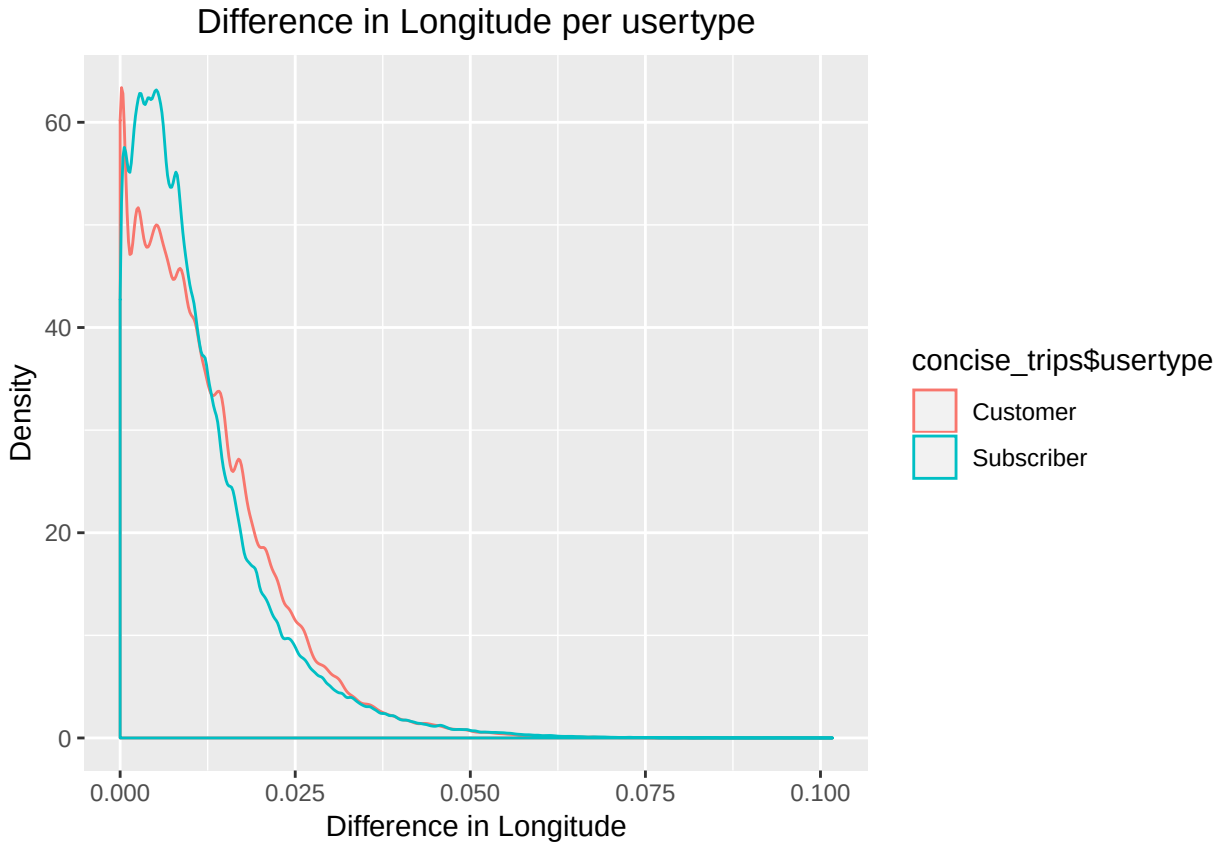
- There are some main zones where the trips are done: around Central Park, in Midtown as well as along the Hudson River
- Touristic zones can be identified: Central Park, Governor's island for instance
- However, for instance the bike lane along the Hudson River can be either commuting or visiting: this is the fastest way of going north/south
- Finally, it seems that an important part of the trips that are done in Midtown are mostly longitudinal: this might be a clue that people are using citibikes to compensate for the lack of subways in the East-West direction.

Finally, from latitude/longitude data from each trip, we can extract useful information to see if there are some patterns: maybe subscribers use citibikes

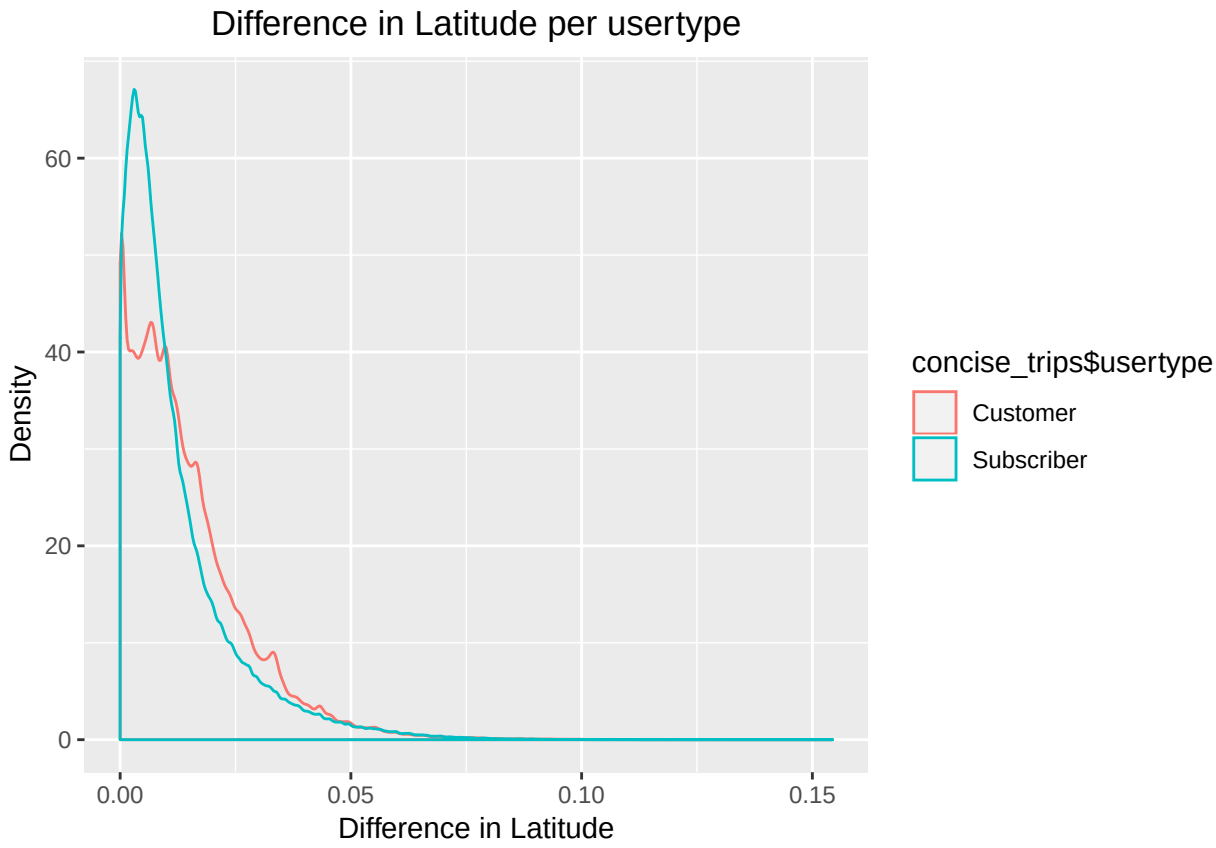
```
colnames(stations_info)[1] = 'start_station_id'
concise_trips <- merge(x = concise_trips, y = stations_info, by = 'start_station_id')


colnames(stations_info)[1] = 'end_station_id'
concise_trips <- merge(x = concise_trips, y = stations_info, by = 'end_station_id')

concise_trips$dirNS <- abs(concise_trips$longitude.y - concise_trips$longitude.x)
concise_trips$dirWE <- abs(concise_trips$latitude.y - concise_trips$latitude.x)
```

```
p <- ggplot(concise_trips,aes(x=concise_trips$dirNS,color=concise_trips$usertype)) +geom_density()
p+
  xlab("Difference in Longitude")+
```

```
  ylab("Density")+
  ggtitle("Difference in Longitude per usertype")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
p <- ggplot(concise_trips,aes(x=concise_trips$dirWE,color=concise_trips$usertype)) +geom_density()
p+
  xlab("Difference in Latitude")+
  ylab("Density")+
  ggtitle("Difference in Latitude per usertype")+
  theme(plot.title = element_text(hjust = 0.5))
```

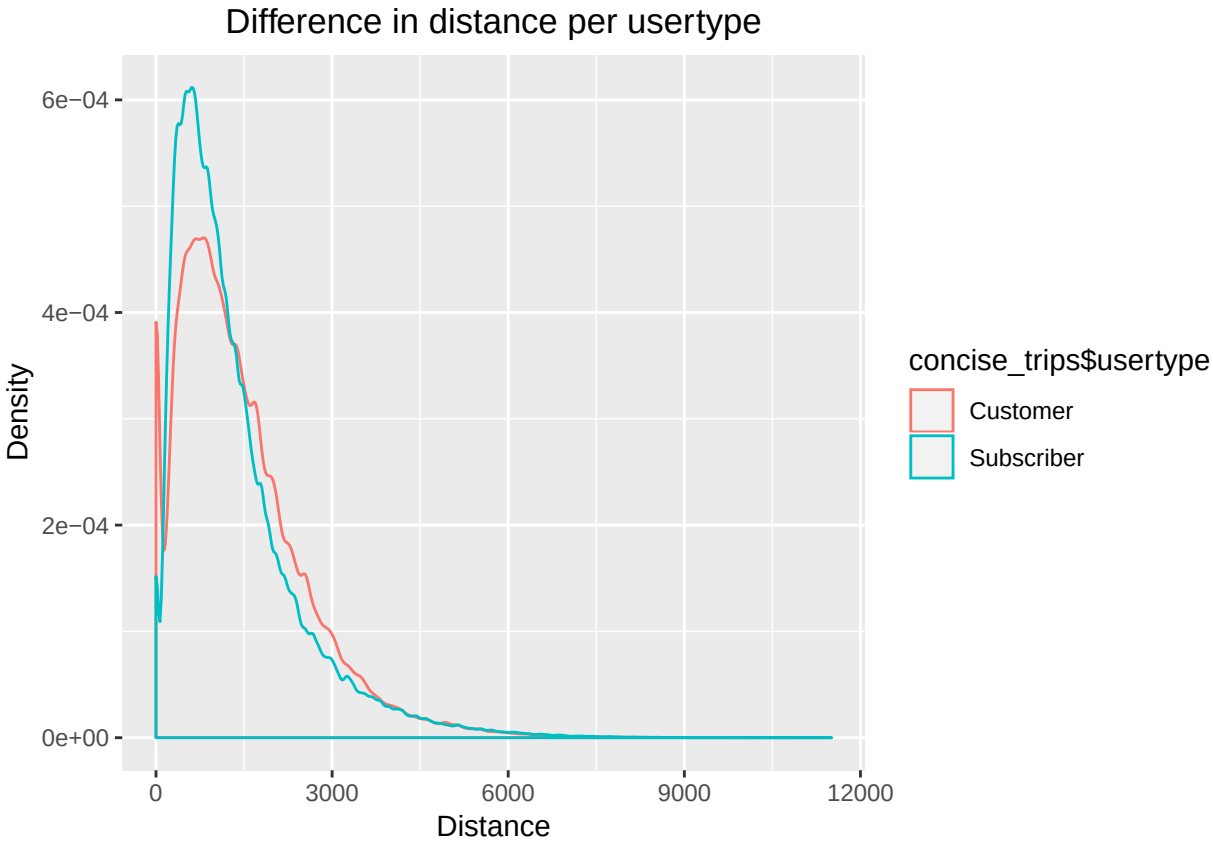


Here we have plotted the absolute difference in longitude between start and finish. Because of the layout fo New York, we can approximate longitude difference as a North/South difference and latitude difference as a East/West difference.We can see that for both latitude and longitude, subscribers have a much higher peak close to low values. This is more visible for latitude.It reveals that subscribers and customers have a different use of the citibikes: subscribers usually do shorter rides (in terms of distance) specially if it is for a North/South difference. This seems to follow the logic of commuting: commuters will not do lengthy North/South rides since it is slower than taking the subway. However, for short rides, it can be quicker to take a bike.

```
library(geosphere)
```

```
## Warning: package 'geosphere' was built under R version 3.4.2
```

```
concise_trips$distance <- distGeo (matrix(c(concise_trips$latitude.x,concise_trips$longitude.x),ncol = 
concise_trips <- subset(concise_trips,
                        select=-c(name.y,name.x,longitude.x,longitude.y,latitude.x,latitude.y))
p <- ggplot(concise_trips,aes(x=concise_trips$distance,color=concise_trips$usertype)) +geom_density()
p+
  xlab("Distance")+
  ylab("Density")+
  ggtitle("Difference in distance per usertype")+
  theme(plot.title = element_text(hjust = 0.5))
```



If we plot the distribution of distance for both customers and subscribers, it confirms what we previously saw with latitude and longitude: subscribers tend to do shorter rides.

Therefore, this discrepancy between subscribers and customers shows that there are two different ways of using citibikes: small short rides or more length rides, and it confirms that citibikes are not only used by tourists, but also by new yorkers as a way of commuting.