CitiBike rental system: data exploration

Wei-Chun Chu

2019-10-01

Abstract: We study how riders used the CitiBike bike rental system in New York City in July 2014. In the data exploration part, we find the basic facts of how people used this system. A shiny app visualizes the data and this doc summarizes it. They provide an intuitive understanding of the project.

1. Data structure and profile

First we load the file and rename the columns for convenience.

```
suppressPackageStartupMessages(library(tidyverse))
set.seed(123456)
data <- read.csv(unz("citibike_2014-07.csv.zip", "citibike_2014-07.csv"),
                 header = T, stringsAsFactors = F)
colnames(data) = c("dur", "ti", "tf",
                   "si", "namei", "lati", "loni",
                   "sf", "namef", "latf", "lonf",
                   "bike", "type", "birth", "gender")
str(data)
```

```
## 'data.frame':    968842 obs. of  15 variables:
##  $ dur   : int  404 850 1550 397 609 2245 1323 320 2430 700 ...
##  $ ti    : chr  "2014-07-01 00:00:04" "2014-07-01 00:00:06" "2014-07-01 00:00:21" "2014-07-01 00:00:1
##  $ tf    : chr  "2014-07-01 00:06:48" "2014-07-01 00:14:16" "2014-07-01 00:26:11" "2014-07-01 00:07:0
##  $ si    : int  545 238 223 224 346 416 501 475 469 320 ...
##  $ namei : chr  "E 23 St & 1 Ave" "Bank St & Washington St" "W 13 St & 7 Ave" "Spruce St & Nassau St
##  $ lati  : num  40.7 40.7 40.7 40.7 40.7 ...
##  $ loni  : num  -74 -74 -74 -74 -74 ...
##  $ sf    : int  402 458 539 2008 521 473 501 116 445 393 ...
##  $ namef : chr  "Broadway & E 22 St" "11 Ave & W 27 St" "Metropolitan Ave & Bedford Ave" "Little Wes
##  $ latf  : num  40.7 40.8 40.7 40.7 40.8 ...
##  $ lonf  : num  -74 -74 -74 -74 -74 ...
##  $ bike  : int  19578 19224 17627 15304 20062 20653 21460 16746 19441 17267 ...
##  $ type  : chr  "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ birth : chr  "1987" "1987" "1973" "1982" ...
##  $ gender: int  2 1 2 1 2 1 1 1 1 1 ...
```

Each observation (row) is a ride, with start time/location and end time/location, bike id, user type (subscriber or customer), and if a subscriber, the gender and birthyear. From the CitiBike website, we know that the "subscribers" pay annual fees and the "customers" rent bikes with the duration from 24 hours to 7 days each time. It is most likely that subscribers are daily commuters and the customers are visitors. In such a case it will be very interesting to differentiate the behaviours of these two groups, since NYC residents and short-term visitors represent very different business opportunities for CitiBike. The number of rides by subscribers and by customers are:

```
table(data$type)
```

```
##
##   Customer Subscriber
##     119064     849778
```

About 12% of the rides are by customers. We found that the percentage of customers was very low (about 2.5%) in winter (January 2014) and it would be harder there to provide reliable information for customer behaviours. It is for this reason we analyze the July data in this work.

A few basic fact of the system can be easily drawn:

```
n_bikes = n_distinct(data$bike) # total number of bikes
n_sta = n_distinct(data$si) # total number of stations
print(c(n_bikes, n_sta))
```

```
## [1] 6204  328
```

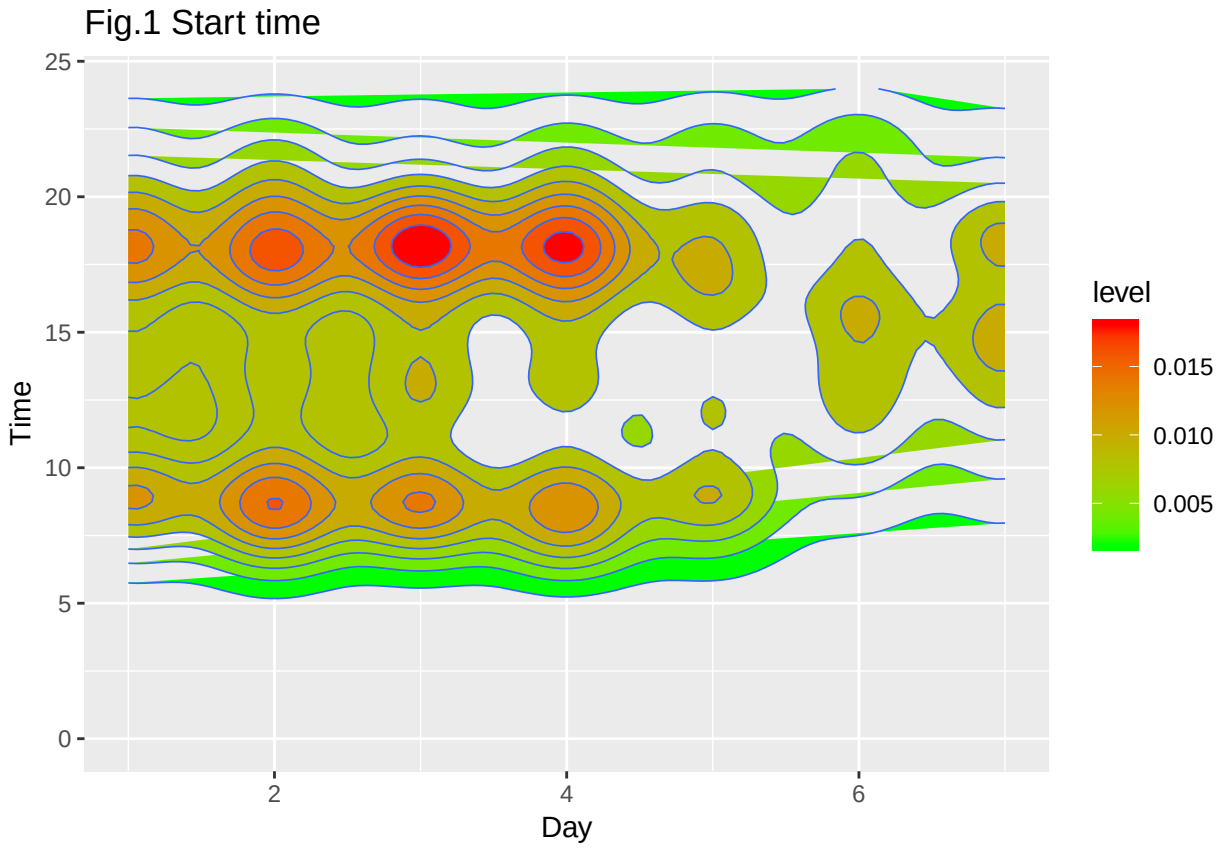
There are totally 6204 bikes and 328 stations in the system, where we assume all bikes have been used at least once. In average a station hosts about 20 bikes.

2. Distribution of rides in time and space

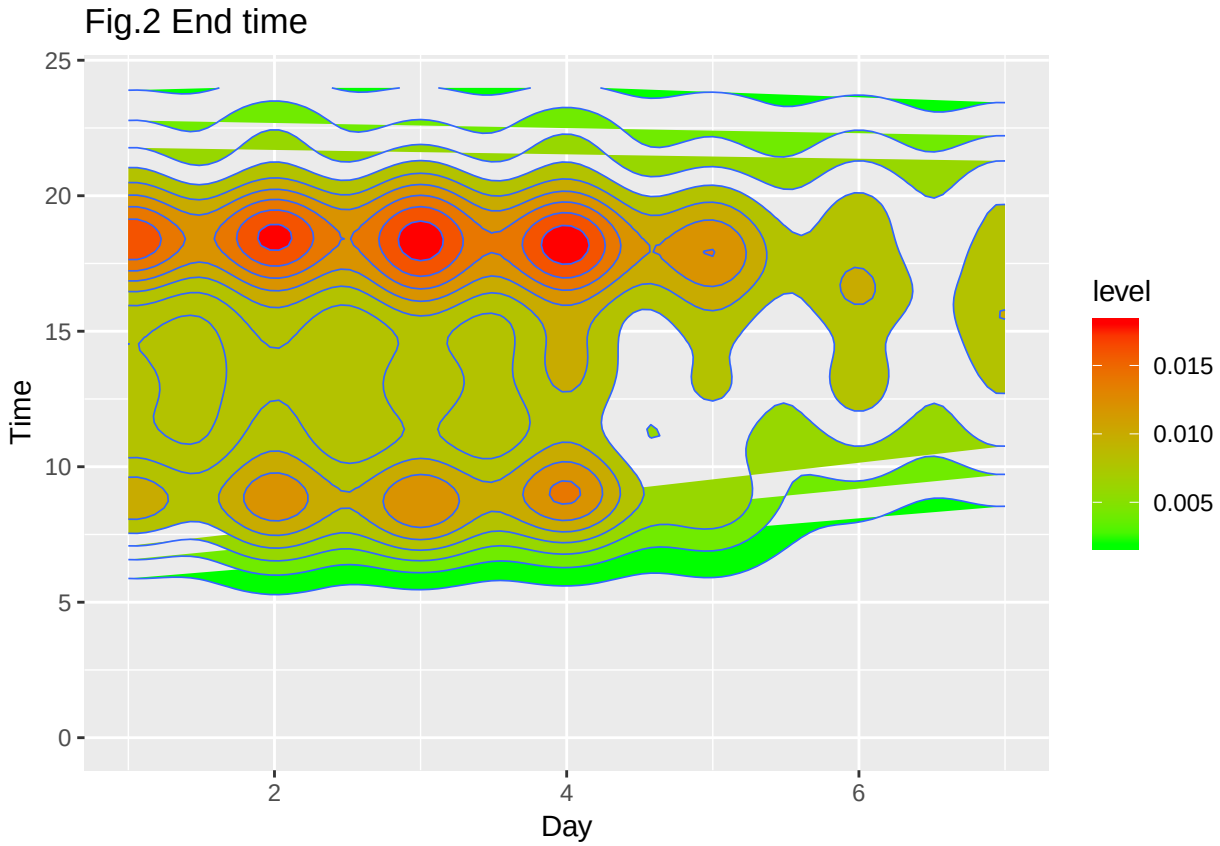
To have the first impression of how people use the system, we want to know when and where people ride most frequently. In each of Fig.1 to Fig.4, we draw 5000 random samples and plot the distribution in the 2D map of day-time or longitude-latitude. The sampled datasets can be plotted much faster than using the complete dataset without losing quality.

```
library(ggplot2)
```

The temporal distribution of starting biking is



Similarly, the end time is plotted in the following figure:



From Fig.1 and Fig.2, it is easy to see that most people ride on weekdays starting roughly around 8 am and 18 pm, and ending just slightly later (much less than an hour) than the start time. This is in line with the time most people go to work and return from work. Since we assume that the NYC commuters are the majority of the riders, this result is not surprising. It is also worth pointing out that people ride bicycles on Saturday and Sunday from early afternoon to evening.

Now we create the density plots for the start and end locations.

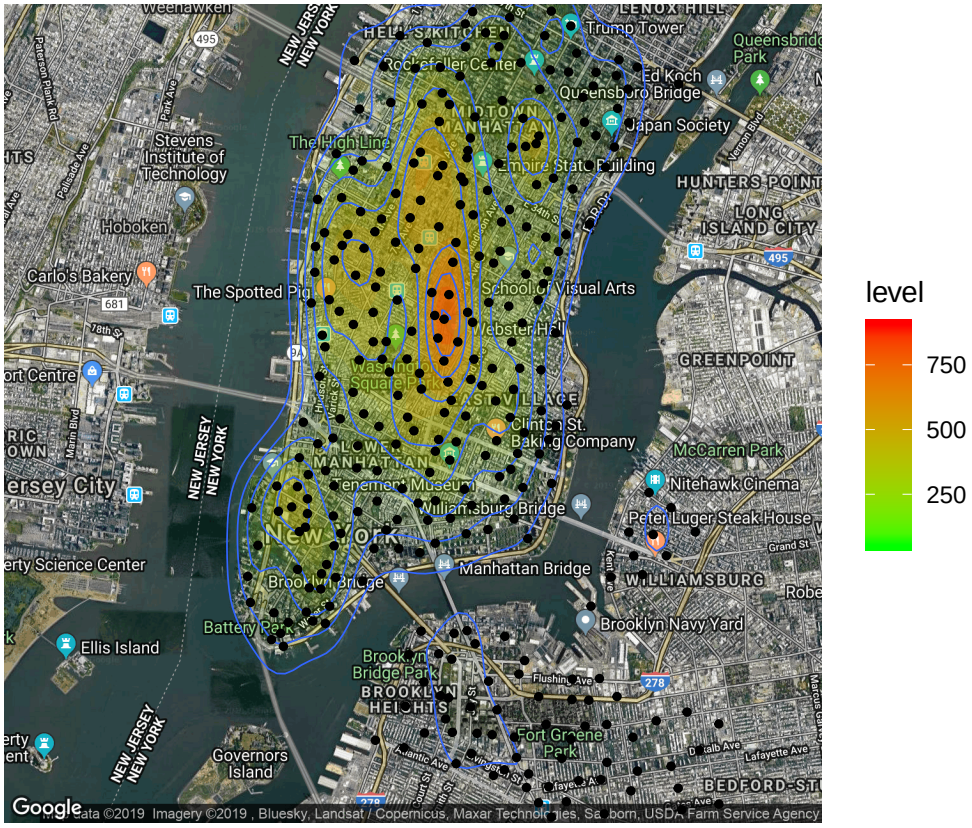
```
# list of stations with locations which can help their labelling on the map
stations = data %>%
  group_by(id = si) %>%
  summarise(name = namei[1],
            lat = lati[1],
            lon = loni[1]) %>%
  arrange(id)
head(stations)
```

```
## # A tibble: 6 x 4
##      id name                              lat   lon
##   <int> <chr>                           <dbl> <dbl>
## 1    72 W 52 St & 11 Ave                 40.8 -74.0
## 2    79 Franklin St & W Broadway         40.7 -74.0
## 3    82 St James Pl & Pearl St           40.7 -74.0
## 4    83 Atlantic Ave & Fort Greene Pl    40.7 -74.0
## 5   116 W 17 St & 8 Ave                  40.7 -74.0
## 6   119 Park Ave & St Edwards St         40.7 -74.0
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.725,-73.995&zoom=13&size=640x640&s
## Warning: Removed 162 rows containing non-finite values (stat_density2d).
## Warning: Removed 162 rows containing non-finite values (stat_density2d).
## Warning: Removed 13 rows containing missing values (geom_point).
```

Fig.3 Start location




```
## Warning: Removed 160 rows containing non-finite values (stat_density2d).
## Warning: Removed 160 rows containing non-finite values (stat_density2d).
## Warning: Removed 13 rows containing missing values (geom_point).
```

Fig.4 End location

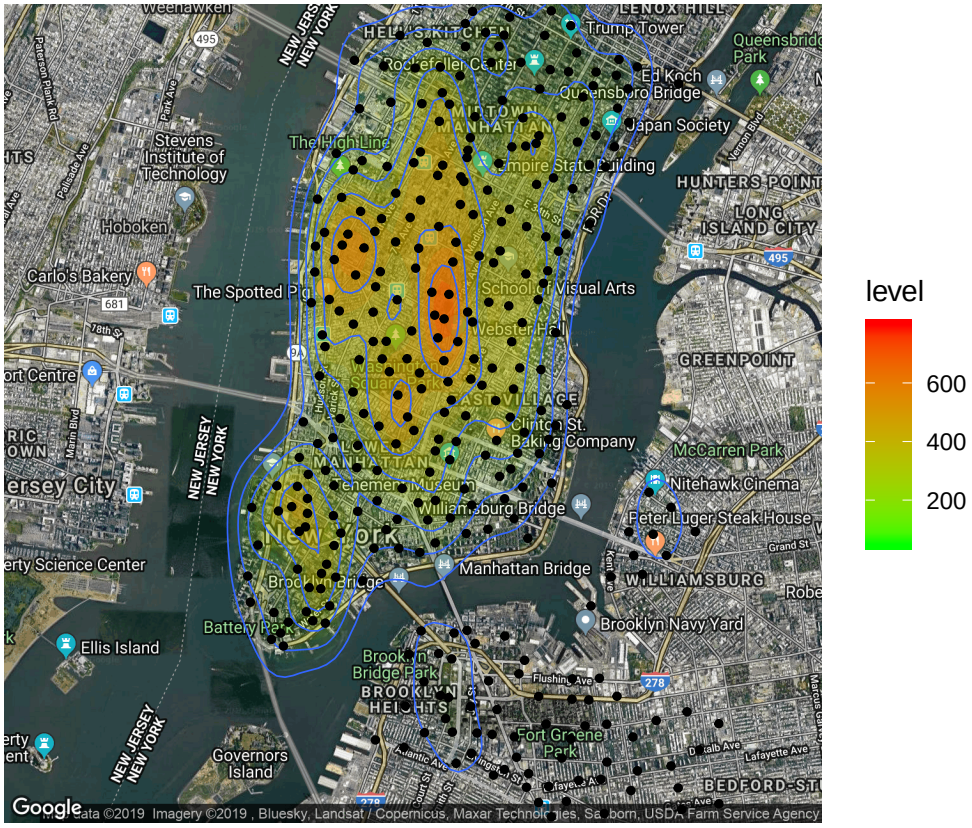


Fig.3 and Fig.4 have almost indistinguishable patterns. It means that the overall start locations balance the overall end locations very well and there should be minimal need to ship the bikes between stations. We notice that despite the even spread of stations, the popularity is highly concentrated in just a few districts. For example the cross section at Park Avenue and 14th St is extremely popular. A question naturally arising from here is: are the locations of these stations optimized for the most popular biking areas? We especially question the usefulness of the stations in Brooklyn, in the southeast corner of the map. The station locations are important when we want to 1. improve the user experience by splitting overcrowded stations and diverting traffic, and to 2. remove useless stations to bring down the cost.

3. Manual relocation of bikes

As mentioned earlier, when borrowing and returning bikes in certain stations are not in equilibrium, the company might have to ship the bikes in bulks from fully parked stations to empty stations. We can trace this by looking at the bike ids: without such manual relocation, a bike's borrow location in a ride must have coincided with its return location in the previous ride.

First of all we create a table to record the riding info for each bike:

```
rides = data.frame(bike = data$bike,
                   ti = as.integer(as.POSIXct(data$ti)),
                   date = as.integer(strftime(data$ti, format = "%d")),
                   wday = as.integer(strftime(data$ti, format = "%u")),
                   dur = data$dur / 60, # in minutes
                   si = data$si,
                   sf = data$sf) %>%
  arrange(bike, ti)
```

```
head(rides)
```

```
##    bike         ti date wday       dur  si  sf
## 1 14529 1405053707   11    5  8.066667 490 536
## 2 14529 1405144379   12    6  3.366667 434 459
## 3 14529 1405154937   12    6 10.900000 459 379
## 4 14529 1405164517   12    6 13.600000 379 438
## 5 14529 1405165660   12    6 14.150000 438 463
## 6 14529 1405177443   12    6 11.233333 463 514
```

From this table it is clear that most of the time, a bike is returned to a certain station before it is borrowed from the same station. But there are exceptions, e.g. between the 1st and 2nd rides in the above table, the bike was shipped from station #536 to #434. We pick out these "disontinuities" and count how often this happened.

```
reloc = rides %>%
  group_by(bike) %>%
  summarise(nrides = n(),
            nreloc = length(si[si != c(si[1], sf[-n()])]),
            freq = round(nreloc/nrides, 2))
head(reloc)
```

```
## # A tibble: 6 x 4
##    bike nrides nreloc  freq
##   <int>  <int>  <int> <dbl>
## 1 14529    115      9  0.08
## 2 14530    118     16  0.14
## 3 14531    149     16  0.11
## 4 14532     78     13  0.17
## 5 14533    215     31  0.14
## 6 14534    181     19  0.1
```

```
sum(reloc$nreloc) / sum(reloc$nrides)
```

```
## [1] 0.1206141
```

In average, a bike is shipped 12 times for every 100 rides. Such manual relocations cost the company extra money and are only necessary if certain stations have insufficient bikes or insufficient parking places.

4. Usage of bikes

When a bunch of bikes are shipped from one station to another, it suggested that these bikes are idle at the first station for some time. For such a "waste" of bikes, it is interesting to know how many "working hours" each bike has per day, and whether the usage changes along a week. The average working time per bike per day is simply the total riding time divided by the number of bikes and 31 days in July.

```
sum(rides$dur) / n_bikes / 31 # average time per bike per day in minutes
```

```
## [1] 72.21196
```

The answer is 72 minutes. We have assumed that the bikes are the same during the whole time.

Now we look deeper into the daily statistics of how people use the bikes. In the following we create a table for the statistics of each of the 31 days.

```
usage_month = rides %>%
  group_by(date) %>%
  summarise(day = wday[1],
```

```
            nrides = n(),
            nbikes = n_distinct(bike),
            dur_tot = sum(dur) / n_bikes, # total duration in minutes per bike on this date
            dur_ave = mean(dur)) # average duration in minutes per ride
head(usage_month, 10)
```

```
## # A tibble: 10 x 6
##     date   day nrides nbikes dur_tot dur_ave
##    <int> <int>  <int>  <int>   <dbl>   <dbl>
##  1     1     2  34854   4375    74.3    13.2
##  2     2     3  26582   4212    54.0    12.6
##  3     3     4  27587   4281    59.1    13.3
##  4     4     5  13612   3415    34.0    15.5
##  5     5     6  22913   3927    67.5    18.3
##  6     6     7  23822   3867    77.6    20.2
##  7     7     1  31863   4557    72.0    14.0
##  8     8     2  32713   4475    70.2    13.3
##  9     9     3  34426   4498    74.0    13.3
## 10    10     4  36288   4550    80.2    13.7
```

Note that in this table, dur_tot is the total working time per bike on that day, which is averaged over all bikes in the system instead of all bikes on the road on that day, since we want to take into account those idle bikes with 0 working time on them. It seems that people use the system quite differently on weekdays and during the weekends. To more clearly see the trend, we further accumulate the statistics for a typical week:

```
usage_week = usage_month %>%
  group_by(day) %>%
  summarise(nrides = mean(nrides),
            nbikes = mean(nbikes),
            dur_tot = mean(dur_tot),
            dur_ave = mean(dur_ave))
show(usage_week)
```

```
## # A tibble: 7 x 5
##     day nrides nbikes dur_tot dur_ave
##   <int>  <dbl>  <dbl>   <dbl>   <dbl>
## 1     1 31637.  4479.    68.8    13.4
## 2     2 32902.  4469     70.4    13.2
## 3     3 33590.  4505.    71.8    13.2
## 4     4 34929   4549     76.8    13.6
## 5     5 30275.  4331.    70.1    14.5
## 6     6 27025.  3904     73.8    17.0
## 7     7 26498.  3842     73.2    17.2
```

While the average daily use of each bike is not too different between weekdays and weekends, there are more bikes borrowed and more daily rides, but shorter trip each ride on weekdays. It seems reasonable that a lot of people ride for short trips to commute on weekdays, and a relatively smaller amount of people ride during the weekends leisurely. To judge the behavior on different days in a week, the sample size (number of observations) is large enough only for the ride durations because each observation is each ride, but is not large enough for number of rides per day or number of bikes borrowed per day because each observation is on each day, and we have only 31 days totally. The t-tests can be done, for example, on the number of rides on Sundays versus on Mondays,

```
t.test(usage_month$nrides[usage_month$day == 7], usage_month$nrides[usage_month$day == 1])
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  usage_month$nrides[usage_month$day == 7] and usage_month$nrides[usage_month$day == 1]
## t = -2.0287, df = 5.2989, p-value = 0.09507
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11541.63   1263.63
## sample estimates:
## mean of x mean of y
##  26497.75  31636.75
```

and the result does not reject the hypothesis that weekdays and weekends are the same. It is not enough to say that Sundays have less rides than Mondays. We can include more months to test this if needed. On the other hand, the t-test on the ride durations on Sundays versus on Mondays can be done by

```
t.test(rides$dur[rides$wday == 7], rides$dur[rides$wday == 1])
```

```
##
##  Welch Two Sample t-test
##
## data:  rides$dur[rides$wday == 7] and rides$dur[rides$wday == 1]
## t = 52.9, df = 180560, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.501449 3.770896
## sample estimates:
## mean of x mean of y
##  17.12698  13.49080
```

with the clear conclusion that the durations on Sundays were longer.

A meaningful relevant question is whether there are "cursed" stations where once a bike is parked there, it spend much longer time before being borrowed again. If there are such cases, where are these stations, and how to remove the curse? To answer this we could trace each bike, list its idle time (time between two consecutive rides) with corresponding stations, and average the idle time for each bike for each station. I think this question is very interesting and can be included in an extended study, but it is not as urgent as the main issues we focus on in this project.

5. Summary

Section 2 shows that the popular stations concentrated sharply in certain areas in NYC and in certain times in a week. Because of this we suspect the station locations could be arranged better to reduce waste and to improve user experiences. In Section 3 we extract the occurances of manual relocation of bikes, and find a bike is shipped 12 times in every 100 times it served. Section 4 shows that a bike works roughly 1.2 hours per day, and the average ride durations on weekends are definitely longer than those on weekdays.