CitiBike rental system Part 1: Exploratory data analysis

Wei-Chun Chu

8 November 2015

Abstract: We study how riders used the CitiBike bike rental system in New York City in July 2014. In Part 1: Exploratory data analysis, we find the basic facts of how people used this system. The data is visually explored and summarized with basic time and location information. This analysis provides an intuitive understanding of the details in Part 2 and Part 3. Possible extension of this work is also discussed.

1. Creating tables

First we load the csv file into R as the initial dataset 'dat0'. The dataset records all CitiBike rides in July 2014. Its column names are renamed for simplicity.

```
set.seed(123456)
library(dplyr) # in the whole project we use dplyr package to manipulate datasets.
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dat0 = read.csv('citibike_2014-07.csv', header = T)
colnames(dat0) = c("dur", "ti", "tf",
                   "si", "namei", "lati", "loni",
                   "sf", "namef", "latf", "lonf",
                   "bike", "type", "birth", "sex")
str(dat0)
```

```
## 'data.frame':    968842 obs. of  15 variables:
##  $ dur  : int  404 850 1550 397 609 2245 1323 320 2430 700 ...
##  $ ti   : Factor w/ 735537 levels "2014-07-01 00:00:04",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ tf   : Factor w/ 738767 levels "2014-07-01 00:06:01",..: 3 32 91 5 17 148 75 2 157 28 ...
##  $ si   : int  545 238 223 224 346 416 501 475 469 320 ...
##  $ namei: Factor w/ 328 levels "10 Ave & W 28 St",..: 111 36 264 251 35 87 157 106 57 203 ...
##  $ 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: Factor w/ 328 levels "10 Ave & W 28 St",..: 47 2 219 209 17 240 157 268 97 150 ...
##  $ 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 : Factor w/ 2 levels "Customer","Subscriber": 2 2 2 2 2 2 2 2 2 2 ...
##  $ birth: Factor w/ 79 levels "1899","1900",..: 67 67 53 62 52 56 73 65 63 63 ...
##  $ sex  : 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" paid annual fees and the "customers" rented bikes from 24 hours to 7 days each time. It is most likely that subscribers were daily commuters and the customers were tourist or 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. The number of rides by subscribers and by customers are:

```
table(dat0$type)
```

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

About 12% of the rides were by customers. We checked the data in winter (January 2014) and found that the percentage of customers was very low (about 2.5%) and it would be harder to provide reliable information for customer behaviours. It is for this reason we analyze the July data in this work. The winter data could be interesting as well and can be studied in the future.

Now we further organize the data. The 'master' dataset and the 'station' dataset together hold the collection of all important but not redundant information. In 'master', the time information is only saved in unix time format, and the location information is only in station ids. 'Station' is a lookup table for stations, where we can find the name and geographical location of each station. These two tables and 'dat0' are the only tables preserved through out the project, which can generate any tables for specific purposes.

```
master = dat0
master = master %>%
  select(
    ti, tf, # time info
    si, sf, # location info: latitude & longitude
    bike, # bike id
    type, birth, sex) # user info
# "Subscriber" & "Customer" changed to 1 and -1:
master$type = as.integer(ifelse(master$type == "Subscriber", 1, -1))
# set birthyear = 0 for unknown customers:
master$birth = as.integer(ifelse(master$birth == '\\N', '0', as.character(master$birth)))
stations = dat0 %>% select(si, namei, lati, loni)
stations = stations %>%
  group_by(id = si) %>%
  summarise(
    name = namei[1],
    lat = lati[1],
    lon = loni[1])
stations = data.frame(stations)
stations = stations %>% arrange(id)
```

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

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

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

There were totally 6204 bikes and 328 stations in the system, where we have assumed all bikes have been used at least once so that they appear in this dataset. In average a station hosted about 20 bikes.

2. Distribution of rides in time and space

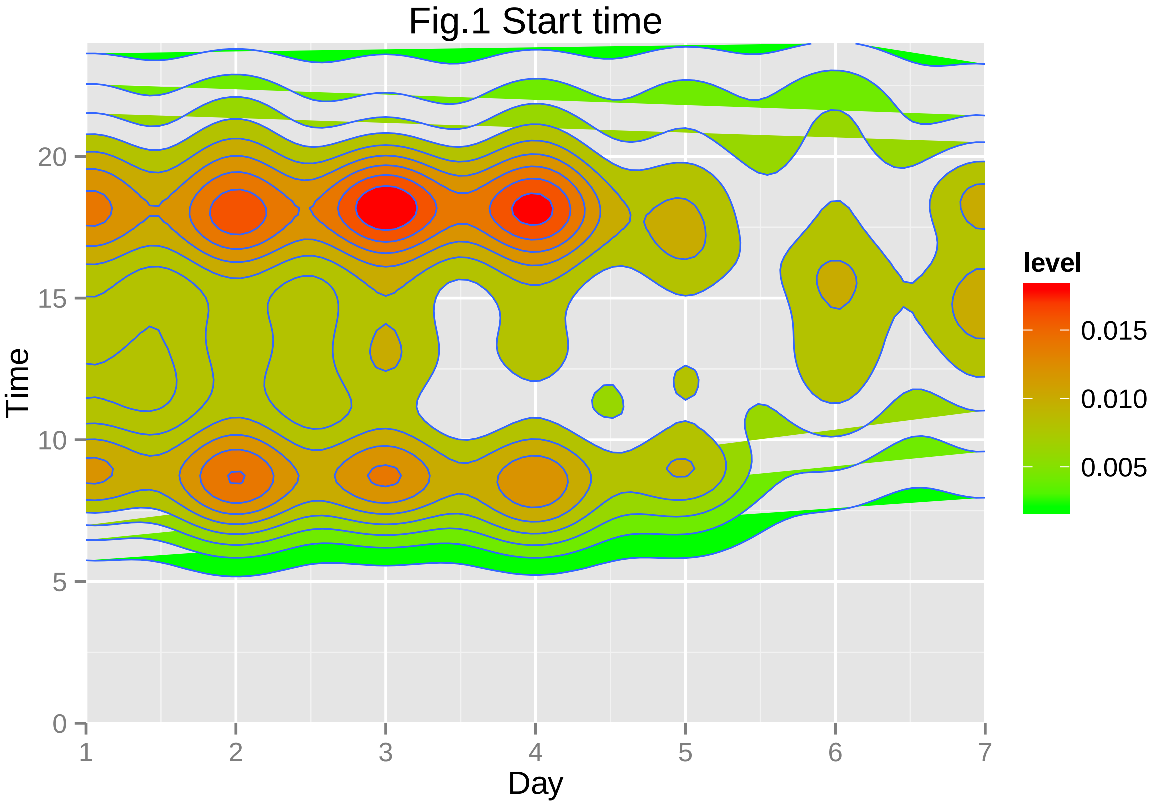
To have the first impression of how people used the system, we want to know when and where people rode most frequently. In each of Fig.1 to Fig.4, we drawn 5000 random samples from 'master' and plot the distribution in the 2D map of day-time or longitude-latitude. The sampled datasets can be plotted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

```
library(ggmap)
```

```
## Loading required package: ggplot2
```

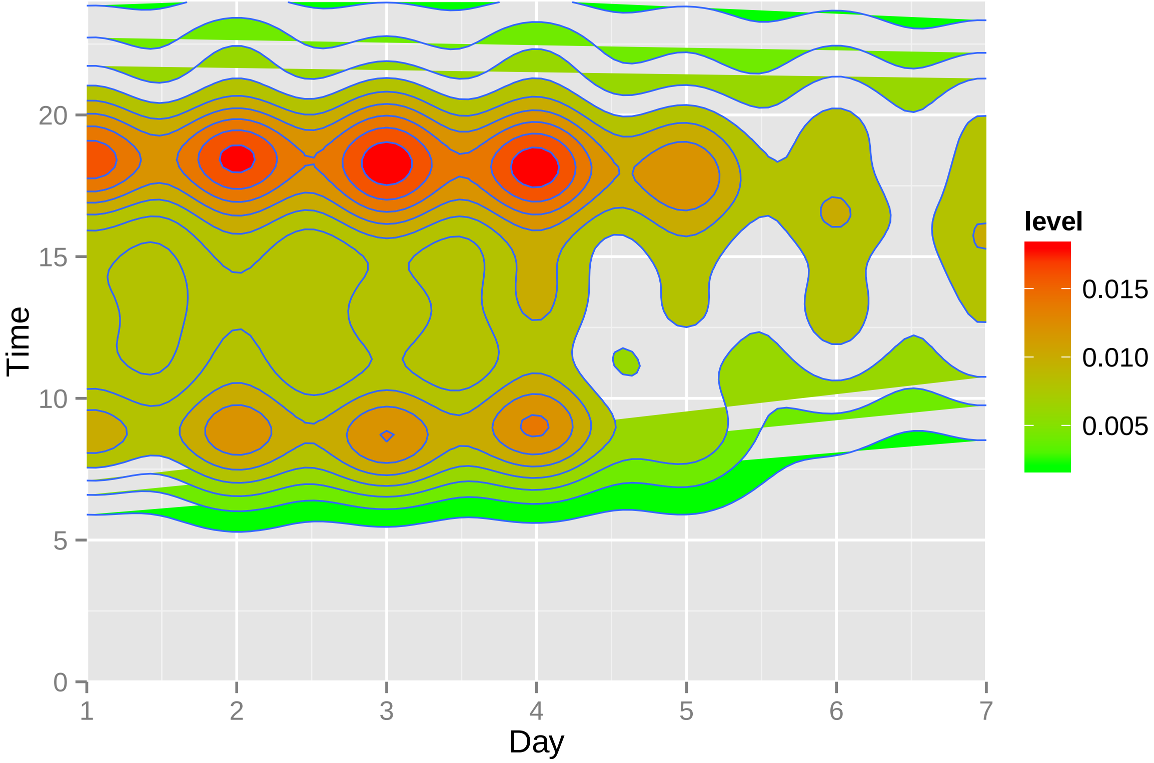
```
library(ggplot2)
```

The temporal distribution of starting biking is



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

Fig.2 End time



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

Now we create the density plots for the start and end locations. A google map with matching range is called by ggmap as the background of the figures.

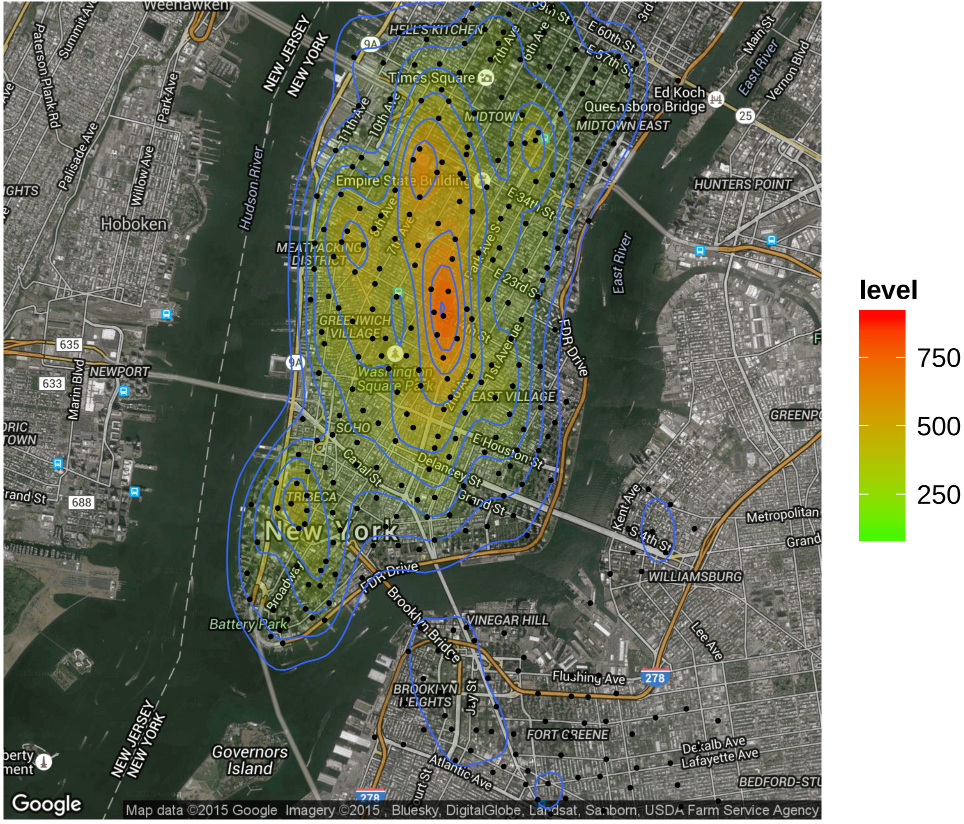
```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=40.725,-73.995&zoom=13&size=640x
```

```
## 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 156 rows containing non-finite values (stat_density2d).

## Warning: Removed 156 rows containing non-finite values (stat_density2d).

## Warning: Removed 13 rows containing missing values (geom_point).
```

Fig.4 End location

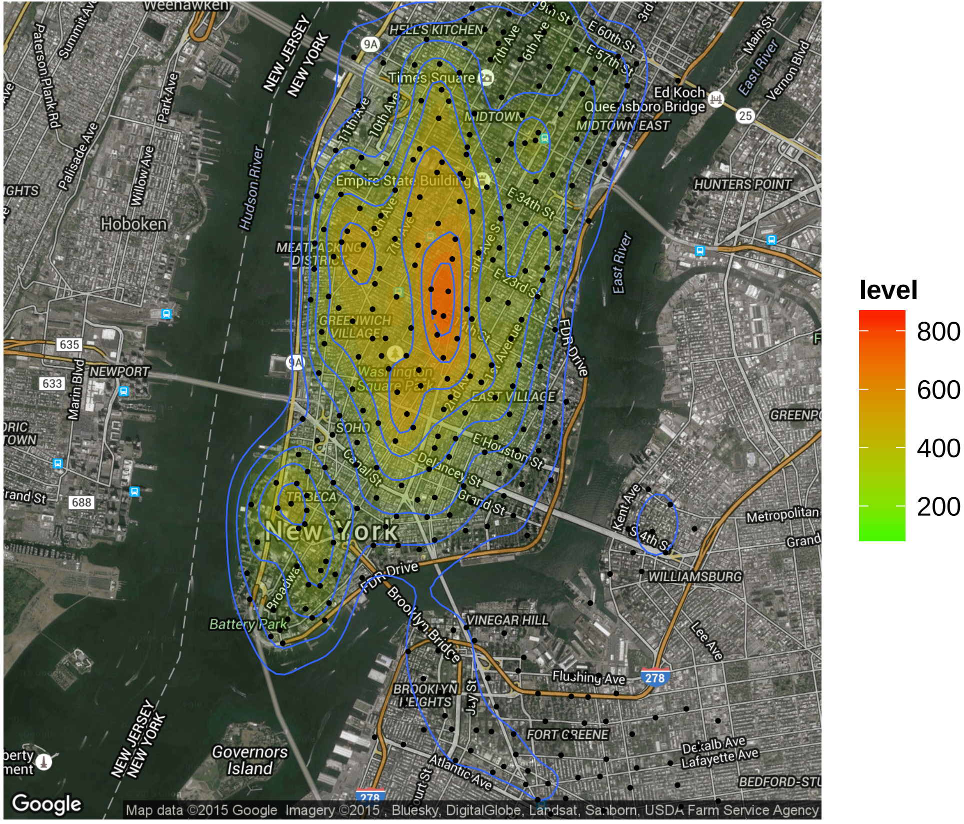


Fig.3 and Fig.4 have almost indistinguishable patterns. It means that the overall start locations balanced the overall end locations quite well and there should have been minimal need to ship the bikes between stations by trucks. However we want to figure out from the data exactly whether bikes were shipped around, and if they were, when and where. We will talk about this in the next section.

Now going back to Fig.3 and Fig.4, we notice that despite the even spread of stations, their popularity was highly concentrated in just a few districts. For example the cross section at Park Avenue and 14th St was extremely popular. A question naturally arising from here is: were 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. We will continue with more details in Part 3 of this project.

3. Manual relocation of bikes

As mentioned earlier, when borrowing and returning bikes in certain stations were not in equilibrium, the company should have shipped 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 = dat0$bike,
                   ti = as.integer(as.POSIXct(master$ti)),
                   date = as.integer(strftime(master$ti, format = "%d")),
                   wday = as.integer(strftime(master$ti, format = "%u")),
                   dur = dat0$dur / 60, # in minutes
                   si = dat0$si,
                   sf = dat0$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 was returned to a certain station before it was borrowed from the same station. But there were 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)
```

```
## Source: local data frame [6 x 4]
##
##    bike nrides nreloc freq
## 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 0.10
```

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

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

In average, a bike were shipped 12 times for every 100 rides. Such manual relocations costed the company extra money and were only necessary if certain stations had insufficient bikes or insufficient parking places. We will find in Part 3 where and when this problem occurred.

4. Usage of bikes

When a bunch of bikes were shipped from one station to another, it suggested that these bikes were idle at the first station for some time where nobody cared about. In think about such “waste” of bikes, it is interesting to know how many “working hours” each bike had per day, and whether the usage changed along a week. The average working time per bike per day is simply the total time of rides 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 number of bikes in the system in July 2014 was the number of bikes ever being riden in the same month. In other words, two possiblities are ignored: 1. there were bikes never used over the whole month, and 2. there were bikes that joined the fleet or retired from the fleet in the middle of the month. We assume that they were rare enough to be ignored.

Now we look deeper into the daily statistics of how people used 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)
```

```
## Source: local data frame [10 x 6]
##
##    date day nrides nbikes  dur_tot  dur_ave
## 1     1   2  34854   4375 74.32674 13.23013
## 2     2   3  26582   4212 54.02608 12.60920
## 3     3   4  27587   4281 59.08458 13.28744
## 4     4   5  13612   3415 33.97751 15.48607
## 5     5   6  22913   3927 67.49374 18.27483
## 6     6   7  23822   3867 77.63931 20.21972
## 7     7   1  31863   4557 72.03498 14.02583
## 8     8   2  32713   4475 70.18901 13.31130
## 9     9   3  34426   4498 73.97201 13.33069
## 10   10   4  36288   4550 80.19551 13.71067
```

Note that in this table, dur_tot was the total working time per bike on that day, which was averaged over all bikes in the system instead of all bikes on the road in that day, since we want to take into account those idle bikes by assigning 0 working time on them. It seems that people used the system quite differently on weekdays and on 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)
```

```
## Source: local data frame [7 x 5]
##
##   day   nrides  nbikes  dur_tot  dur_ave
## 1   1 31636.75 4479.25 68.79516 13.44424
## 2   2 32901.60 4469.00 70.37751 13.19128
## 3   3 33589.80 4504.80 71.84065 13.23206
## 4   4 34929.00 4549.00 76.80990 13.62463
## 5   5 30275.25 4331.25 70.07436 14.53294
## 6   6 27025.25 3904.00 73.83748 17.01226
## 7   7 26497.75 3842.00 73.15060 17.18331
```

While the average daily use of each bike was not too different between weekdays and weekends, there were more bikes borrowed and more daily rides, but shorter trip each ride on weekdays. It seems reasonable that a lot of people rode for short trips to commute on weekdays, and a relatively smaller amount of people rode on 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 date, 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 were the same. It is not enough to say that Sundays had 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 were “cursed” stations where once a bike was parked there, it spent much longer time before being borrowed again. If there were such cases, what were 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 in an extended study, but it is not as urgent as the main issues we focus on in this project.

5. Summary of Part 1

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 was shipped 12 times in every 100 times it served. Section 4 shows that a bike worked roughly 1.2 hours per day, and the average ride durations on weekends were definitely longer than those on weekdays.