Leila Erbay

R – Hubway Lab – Part I

**A. Explore (and clean) the station data**

0. Let's start with looking at the hubway\_stations file. Import it as a data frame.

**stations <- read.csv("hubway\_stations.csv")**

1. What are the column names? Find out using R.

**id, terminal, station, municipal, lat, lng, status**

2. How many stations are there? Find out using R.

**137 unique stations**

Here is the dictionary for this file:

* id: station id assigned by MAPC; corresponds to strt\_statn and end\_statn in trips table
* terminal: Hubway-assigned station identifier
* station: station name
* municipal: Municipality
* lat: station latitude
* lng: station longitude
* status: Existing station locations and ones that have been removed or relocated

3. Look at the summary of this data frame with Hubway station information. Some of these bike stations don't exist anymore; create a data frame with only those that still exist. How many are left?

**130 unique existing stations**

**CHECK IN WITH NEIGHBOR TO CONFIRM YOUR ANSWER HERE**

**B. Explore the trip data**

0. Read in hubway\_trips.csv to R

**trips <- read.csv("hubway\_trips.csv")**

1. This file is quite large. (And we've already removed move than 75% of the data!) How many rows and columns does it have?

**13 Col**

**350615 Row**

Below is the data dictionary:

* seq\_id: unique record ID
* hubway\_id: trip id
* status: trip status; "closed" indicates a trip has terminated
* duration: time of trip in seconds
* start\_date: start date of trip with date and time, in EST
* strt\_statn: id of start station
* end\_date: end date of trip with date and time, in EST
* end\_statn: station id of end station
* bike\_nr: id of bicycle used
* subsc\_type: subscription type - "Registered" is user with membership; "Casual" is user without membership
* zip\_code: ZIP code of user (only available for registered users)

**\*\*data includes an apostrophe(') prefix\*\***

* birth\_date: birth year of user
* gender: gender of user

2. How many unique user zip codes are in this dataset?

**375 Unique Zip Codes**

3. How many unique bicycles are in this dataset?

**882 unique bikes**

4. Calculate the count of rides for each unique bicycle.

**numRides <- count(trips,"bike\_nr")**

**CHECK IN WITH THE CLASS HERE, CONFIRM THE ABOVE ANSWERS**

5. Which bicycle is ridden most frequently?

**B00401**

6. Which bicycle is ridden least frequently?

**T01093**

7. Calculate the total duration of all rides for each bicycle. Hint: tapply() or aggregate()

**duration <- count(trips, "bike\_nr", "duration")**

**##OTHER WAYS TO FIND DURATION**

**duration <- aggregate(duration~bike\_nr, data= trips, sum)**

**duration <- summaryBy(duration~bike\_nr,data = trips,FUN= sum )**

**duration <- data.frame(tapply(trips$duration, trips$bike\_nr, sum))**

**CHECK IN WITH THE CLASS HERE, CONFIRM THE ABOVE ANSWERS AND DISCUSS TAPPLY VS. AGGREGATE, MAKE SURE YOU CAN IMPLEMENT BOTH SOLUTIONS.**

8. Which bicycle has been ridden for the longest total duration in this dataset? Shortest total duration?

**Bike w/ Longest Duration: B00585**

**BIke w/ Shortest Duration: T01380**

9. Consider only trips on the bicycle that has been ridden for the longest duration, which station is its most frequent end station? Which station is its most frequent start station? Return the station ids.

**Most Frequent End Station: 22**

**Most Frequent Start Station: 22**

10. Look up the name of the above station ids in the stations data frame. What are the names of the most frequent start and end stations for this bicycle?

**South Station - 700 Atlantic Ave.**

**CHECK IN WITH THE CLASS HERE, CONFIRM THE ABOVE ANSWERS**