Research Analyst Tableau Focus Candidate Test.

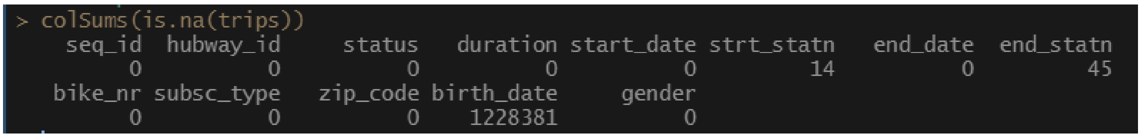
Jerónimo Houlin.

16/10/2021

Wheelie Wonka Co.

To explain bicycle availability in Boston Massachusetts, using the city's bike sharing system for bike stations and trips, the first step is to clean up the database.

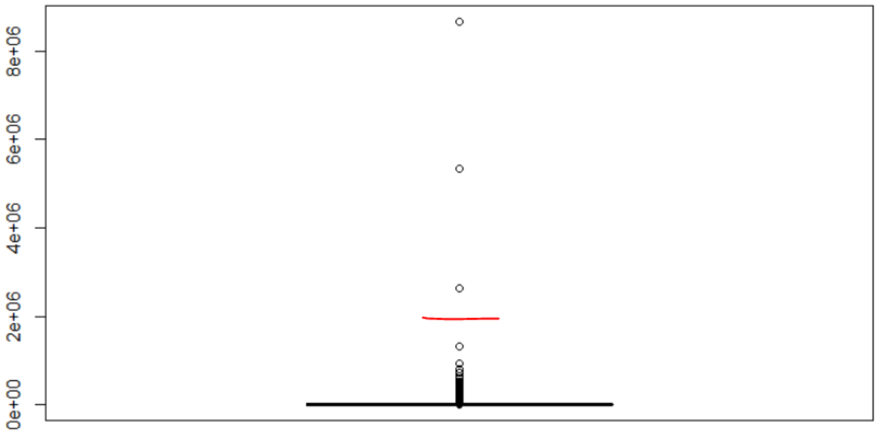
For this we check where NA values are located.



```
> colSums(is.na(trips))
    seq_id  hubway_id     status   duration start_date strt_statn   end_date  end_statn
         0          0          0          0          0         14          0         45
   bike_nr subsc_type   zip_code birth_date     gender
         0          0          0    1228381          0
```

Given that the sources of NA's are variables that cannot be replaced, for example distorting user's birth dates would end up giving us a false customer target, we can assign them NULL values.

There are some negative values in duration, which are replaced for the average of the field assuming these where due to a mistake of how the information was input. Also, any outliers where cut off;



* Trip Duration Box Plot.

Assuming 2,000,000 seconds or 23 days might be too much (or a stolen bike).

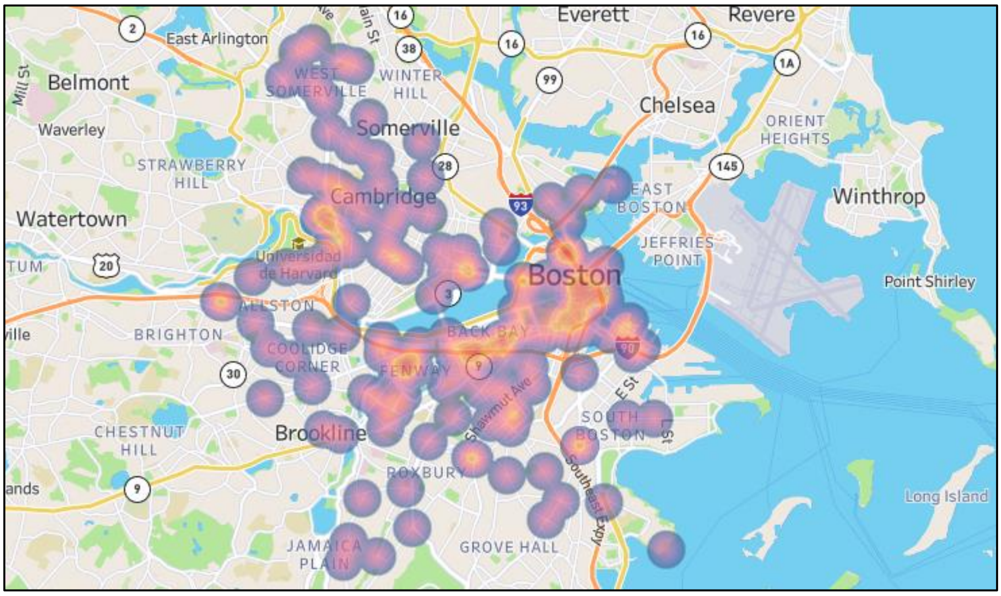
The other manipulations on our data frames (that are in the attached R script) are string transformation for the start and end date to a m/d/y and H/M/S type, and the junction of both the stations id's and the trips beginning and end id's. With this manipulation we get a new CSV output named "trips_sync" which has each trip origin "lat - lon" coordinates and destination information so that we can map both starting points and arrival points in tableau.

Given the entire time frame, we can appreciate the following information on the type of costumer for Boston's bike trips:

GENERATION	AMOUNT	%
Traditionalists	1,213	0.35
Baby Boomers	59,926	17.08
Gen X	74,437	21.22
Gen Y	214,917	61.26
Gen Z	151	0.04

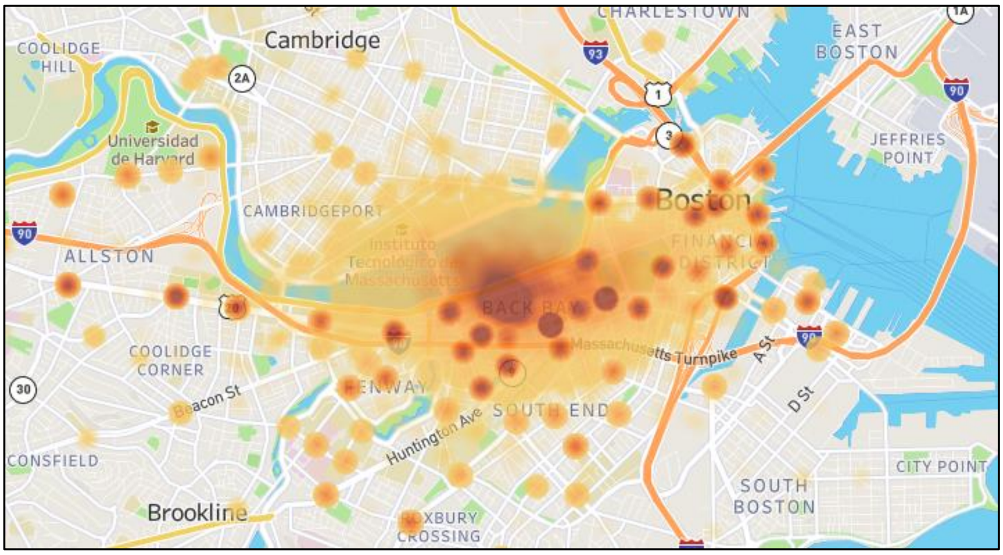
Where 75.44% of bike users are men and average trip duration is 13:23 minutes.

The municipalities where the most trips where originated are Boston (77.38%) and Cambridge (18.79%), but the "number one" hotspot where people pick up the most bikes is at Back Bay, Boston:



*Origin sheet in Tableau (Brookline = 1.30% and Somerville = 2.51% of all trips)

When we look the starting point for the trips, sorted by the trips duration (most dense are trips of longer duration) we notice that mainly Boston and in the outskirts of the city become the hotspots for longer trips.

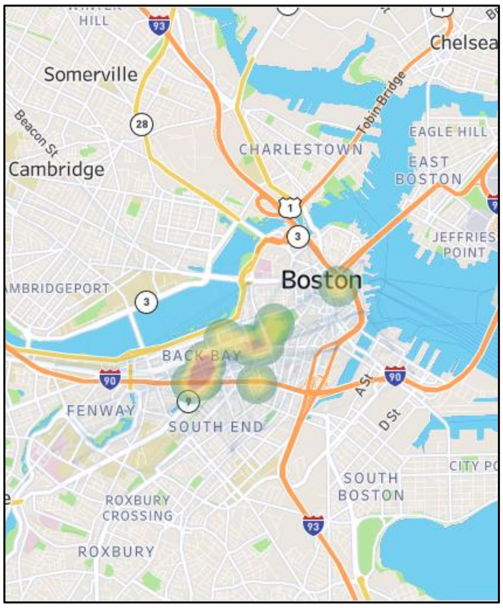


*Duration sheet in Tableau

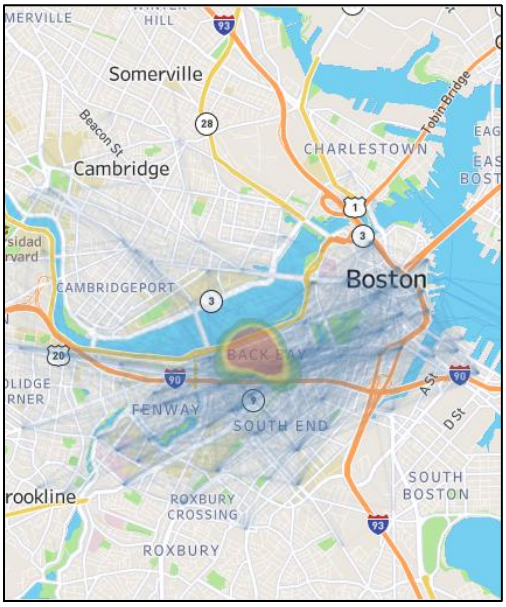
To figure out the mode for the duration of trips, we created a "getmode" function (line 80 of the R script) which gives us a meaningful 6:51 minutes, therefore shorter trips are most common.

If we look at the paths and arrival heat maps for every trips in the time frame, filtered by each generation, we get:

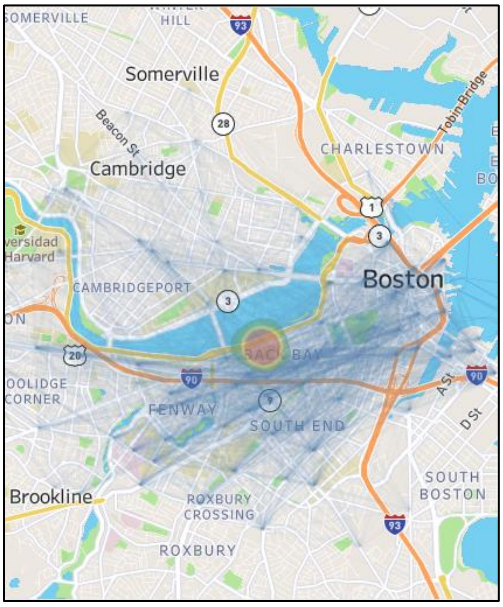
Traditionalists



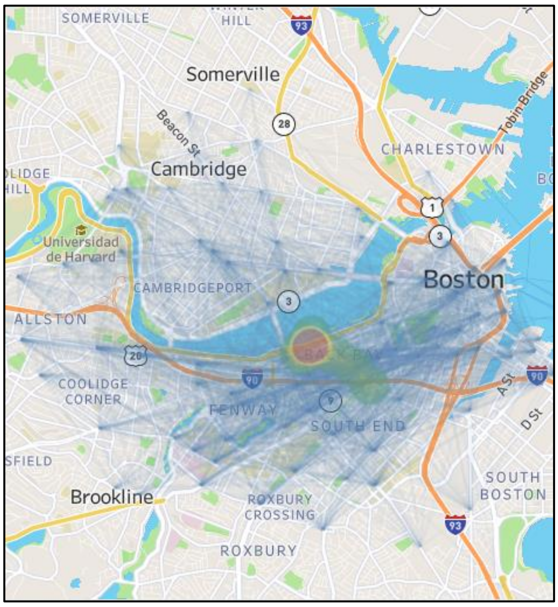
Baby Boomers



Gen X



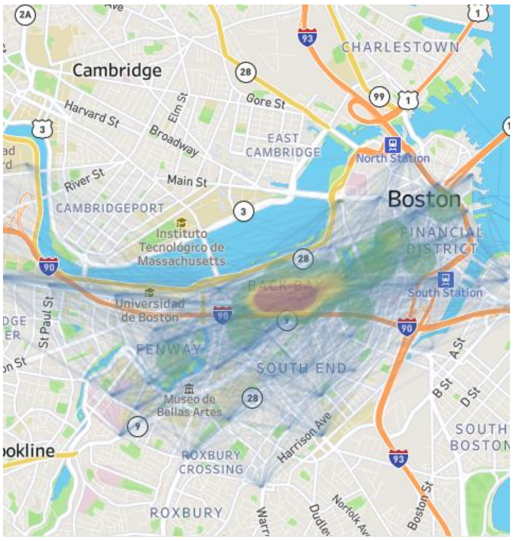
Gen Y



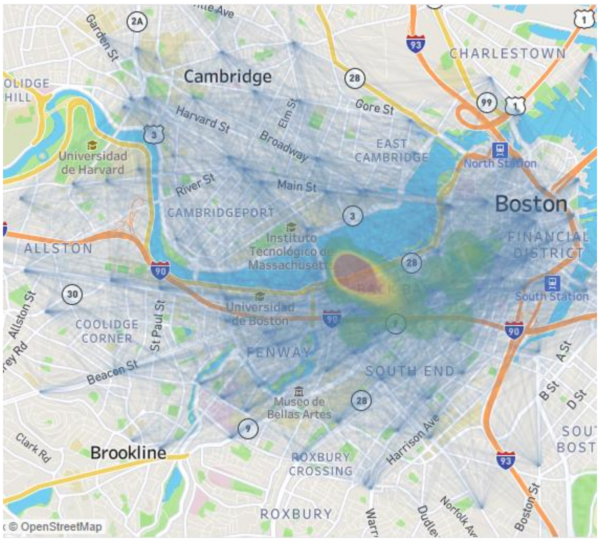
So even though Back Bay is also the spot for most arrivals, Gen Y users (the majority) have most reach.

Lastly, we can compare trips for the Gen Y users over the years:

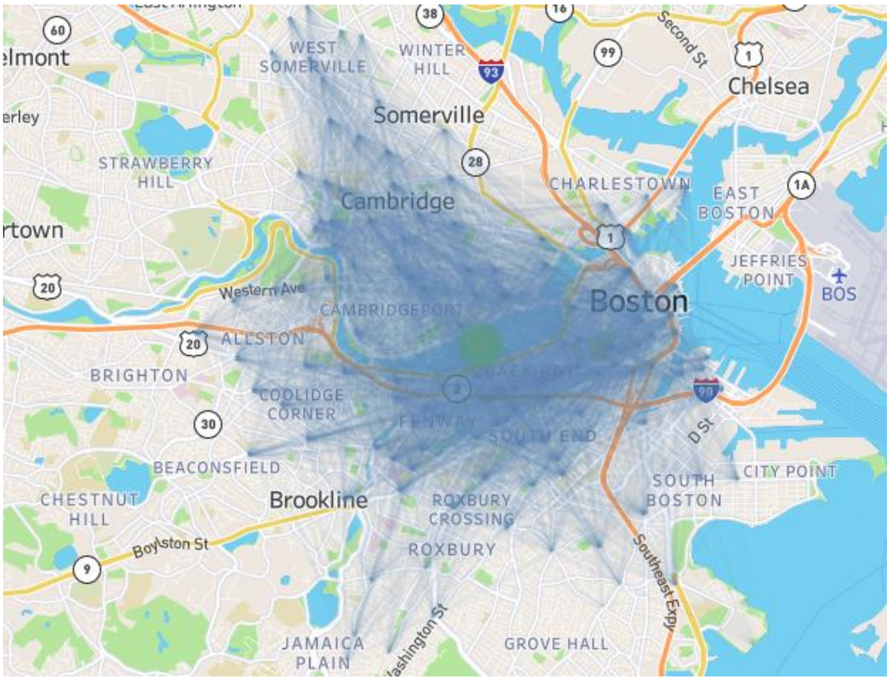
2011



2012



2013



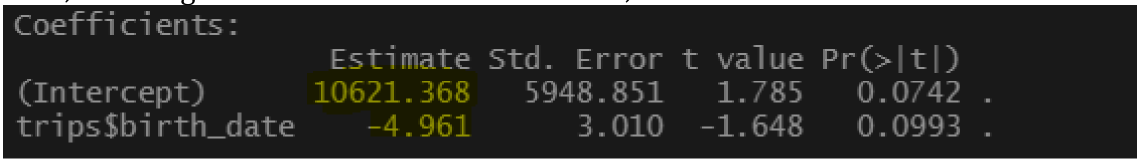
Where, over the years we can tell bike rides for the Gen Y users have spread over to locations northbound like West Somerville and south to Jamaica Plain.

If one would like to see the evolution of how these paths evolve over every hour of the day, you could reference the attached tableau file, sheet named “hourly”. There we can see that the bikes are most often than not used from 8am, until midday.

To explain trip duration I also created a linear model for every variable wich I found most significant, given:

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

First, using a user's birth date, the linear model shows:

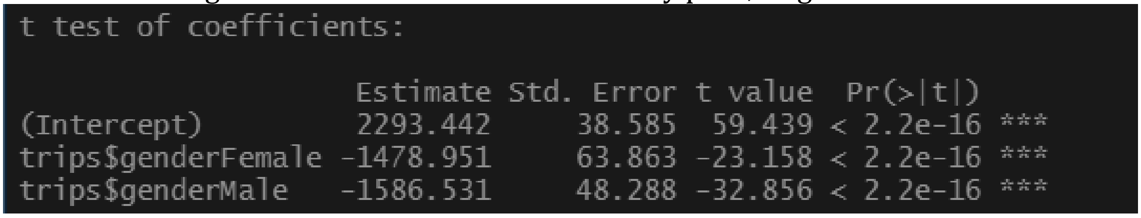


```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       10621.368   5948.851   1.785   0.0742 .
trips$birth_date     -4.961      3.010  -1.648   0.0993 .
```

Where the intercept and slope values are shown, so, for example, if we were to have an individual born in 1999, then the linear model would predict that his trip duration might be:

$10621.368 + (-4.961 * 1999) = 704.33$ seconds or 11:44 minutes

Of course the significance levels here would be very poor, so gender tells us:

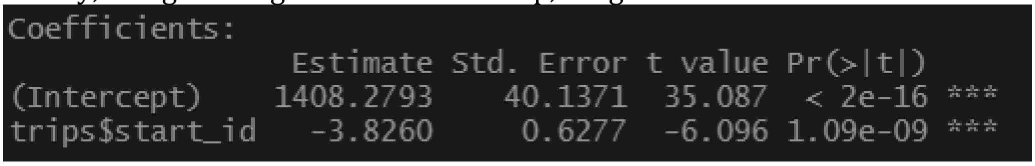


```
t test of coefficients:

                    Estimate Std. Error t value  Pr(>|t|)
(Intercept)         2293.442     38.585  59.439 < 2.2e-16 ***
trips$genderFemale -1478.951     63.863 -23.158 < 2.2e-16 ***
trips$genderMale   -1586.531     48.288 -32.856 < 2.2e-16 ***
```

Here, the Boolean variable "being a male", says that the trip might last 11:47 minutes. Whilst "being a female", predicts a trip duration of 13:34 minutes, therefore, even though male users are dominant, trip duration is actually more inclined if the user is a woman.

Finally, using the origin station id of a trip, we get:



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1408.2793    40.1371  35.087  < 2e-16 ***
trips$start_id   -3.8260     0.6277  -6.096 1.09e-09 ***
```

Meaning that a trip originated in Cambridge st. and Joy st. (id number 8) would last around 22:57 minutes, just like gender, with a significance close to 100%.