# New York City Taxi Data: FOILED for the Public

*Ben Arancibia*

*September 29, 2014*

***Introduction*** This dataset was obtained by a Civic Technology expert located in New York City, who works for the open data firm Socrata. The work he has done to visuzale the Taxi Data can be found here: https://github.com/chriswhong/taxitracker. This dataset tracks all taxis during 2013 including number of passengers, dollar amounts of individual fares, tips, and times the taxi was active. There is a visuzalition that follows one taxi for 24 hours then switches to another to look at the different patterns and choices of each taxi. The visualizatio can be found here: http://nyctaxi.herokuapp.com/

Source: New York Taxi and Limousine Commission obtained by Freedom of Information Law (FOIL).

The dataset is huge when you look at the total processed data set. The total dataset is about 2.5GB and can be downloaded fully from here: http://www.andresmh.com/nyctaxitrips/. For this exercise we will only look at 100 taxi records. This data can be found on the github account referenced earlier. I have downloaded it to my working directory.

***Data Profile***

```
require(ggplot2)
```

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

```
require(DMwR)
```

```
## Loading required package: DMwR
## Loading required package: lattice
## Loading required package: grid
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```
require(dplyr)
```

```
## Loading required package: dplyr
##
## 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
```

```
require(psych)
```

```
## Loading required package: psych
##
```

```
## Attaching package: 'psych'
##
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
data_csv <- read.table("/Users/bcarancibia/CUNY_IS_607/100RandomTrips.csv", header=TRUE, sep=",")
```

The attribute information of the table is:

```
names(data_csv)
```

```
##  [1] "medallion"   "pickuptime"  "dropofftime" "passengers"  "pickupx"
##  [6] "pickupy"     "dropoffx"    "dropoffy"    "fare"        "paymenttype"
## [11] "surcharge"   "mtatax"      "tip"         "tolls"       "total"
```

medallion = The ID of the Taxi

pickuptime = Time passenger was picked up

dropofftime = Time passenger was dropped off

passengers=number of passengers

pickupx = x coordinate of location (for mapping)

pickupy = y coordinate of location (for mapping)

dropoffx = x coordinate of drop off location

dropoffy = y location of drop off location

fare = amount to drive

paymenttype = how the fare was paid

surcharge = tax

mtatax = special transit tax

tip = tip for driver

tolls = toll roads if used

total = total amount aggreagted

Quick summary of the data:

```
summary(data_csv)
```

```
##                              medallion          pickuptime
##  3BCD8C31FAA3F956FCF1E94F56D04A60:227   9/28/13 0:01:   4
##  0FE34002F6E240EBAE51520DEF0D2259:123   9/28/13 0:13:   4
##  E9ECAA3852ABB734244A2DEE0F9204E8:107   9/28/13 0:27:   4
##  D0E11AB0F51BFD9FF4053F8A585D1A89: 99   9/28/13 0:37:   4
##  8139A6C9596767B37F84DACB7E200BDD: 98   9/28/13 0:57:   4
##  B872D33BF89C6C9520CEDBDDDD421AF9: 98   9/28/13 1:12:   4
##  (Other)                         :601   (Other)     :1329
##       dropofftime     passengers      pickupx          pickupy
##  9/28/13 0:11:   4   Min.   :1.00   Min.   :-74.0   Min.   :40.6
##  9/28/13 0:25:   4   1st Qu.:1.00   1st Qu.:-74.0   1st Qu.:40.7
##  9/28/13 0:37:   4   Median :1.00   Median :-74.0   Median :40.8
##  9/28/13 0:44:   4   Mean   :1.61   Mean   :-74.0   Mean   :40.8
##  9/28/13 1:05:   4   3rd Qu.:2.00   3rd Qu.:-74.0   3rd Qu.:40.8
##  9/28/13 1:36:   4   Max.   :6.00   Max.   :-73.8   Max.   :40.8
##  (Other)     :1329
##     dropoffx        dropoffy         fare      paymenttype   surcharge
##  Min.   :-74.2   Min.   :40.6   Min.   : 3.0   CRD:781     Min.   :0.000
##  1st Qu.:-74.0   1st Qu.:40.7   1st Qu.: 6.5   CSH:572     1st Qu.:0.000
##  Median :-74.0   Median :40.8   Median : 9.0               Median :0.000
##  Mean   :-74.0   Mean   :40.8   Mean   :11.1               Mean   :0.315
##  3rd Qu.:-74.0   3rd Qu.:40.8   3rd Qu.:13.0               3rd Qu.:0.500
##  Max.   :-73.8   Max.   :40.9   Max.   :63.5               Max.   :1.000
##
##      mtatax            tip            tolls            total
##  Min.   :0.000   Min.   : 0.00   Min.   : 0.000   Min.   : 4.0
##  1st Qu.:0.500   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.: 8.0
##  Median :0.500   Median : 1.00   Median : 0.000   Median :10.7
##  Mean   :0.498   Mean   : 1.21   Mean   : 0.142   Mean   :13.3
##  3rd Qu.:0.500   3rd Qu.: 2.00   3rd Qu.: 0.000   3rd Qu.:15.6
##  Max.   :0.500   Max.   :12.06   Max.   :10.250   Max.   :81.0
##
```

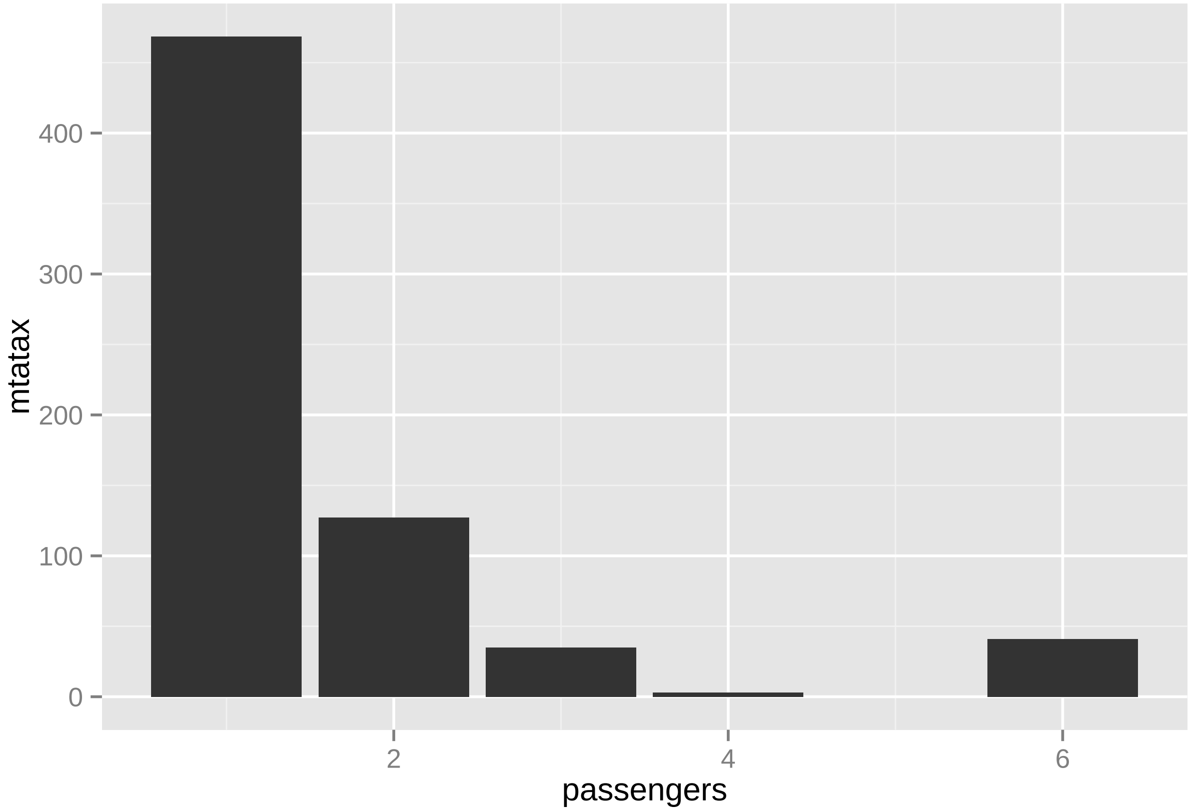
Looking at the summary you cannot really discern any information, one of the main reasons this is because there are four out of the fifteen feilds that are geographic related.

Lets turn it into a data frame to do very very surface level data analysis.

```
taxi <- data.frame(data_csv)
```

I was curious about the relation between number of passengers and mtatax. Which group of passengers paid the most mtatax. Theoretically, the tax should be higher when the number of passengers increased. This was not the case at all, look at the amount of tax collected in relation to the number of passengers in taxis.

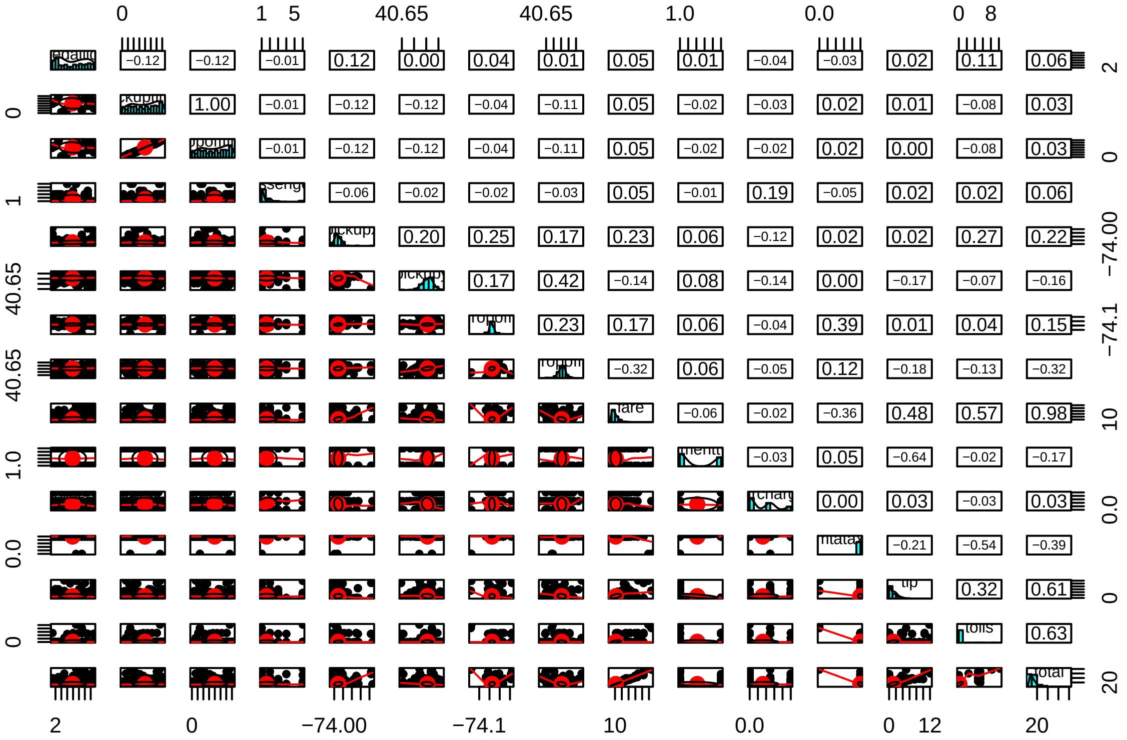
```
p <- ggplot(data=taxi, aes(x=passengers, y=mtatax))
p+geom_histogram(stat="identity")
```



As you can see from the visualization the highest amount of MTA Tax is paid by 1 person passengers followed by two person passengers. I don't think that the actual tax percentage is higher, just that there are more one and two person passenger taxi rides compared to the other amount of taxi cab rides.

***Correlations*** Using the psych library, I took a look at all the variables to see if there are any correlations. It is a little difficult to see all the correlations.

```r
cor <- taxi[sample(nrow(taxi), 1000),]
pairs.panels(cor)
```



When looking at the different correlations, there are some that you can automatically ignore. For example all the x and y coordinates will be coorelated because all the taxi trips were done in New York City. It would be useful to look in those cells though to see if there are any outliers. One thing that does look like a correlation is the fare with the total. The higher the fares the higher the total amount paid to the taxi. This makes complete sense. Other than that there do not seem to be any correlations.

### Conclusion

Based off the different findings and diving deeper into the dataset. There do not appear to be any correlations within the data besides the more fares, tolls, and other amounts increase, the higher the total fare. One thing that I am curious about is why the longer amount of time between pick up and drop off is there an increase in total amount. Also, with the same logic the father from the pickup the higher the fair should be theoretically. There is a lot of analysis that can be done with this dataset, but the real strength of the data is the ability for it to be analyzed with the geographic lense to the data. Plotting this data on map, like the original user did, would make this a more powerful data set.