TDI_r_markdown

Nima Dolatnia

July 23, 2018

Introduction

The idea of this project came from a converstation that I had with a Lyft driver a couple of weeks ago. During my travel, the driver mentioned that she also works for Uber and for some reason users of one of the apps pay way more tip compare to the others! This stroke me as a surprise and wondered if the amount of tip that a person pays can be predicted. This information can be useful to both the company (and perhaps the driver) as well as the user of the service. To analyze the amount of tip tha people pay, I have selected the NYC yellow taxi dataset from Jan 2017. It has more than 9 million instances and has enough features to be able to obtain meaningful patterns from it.

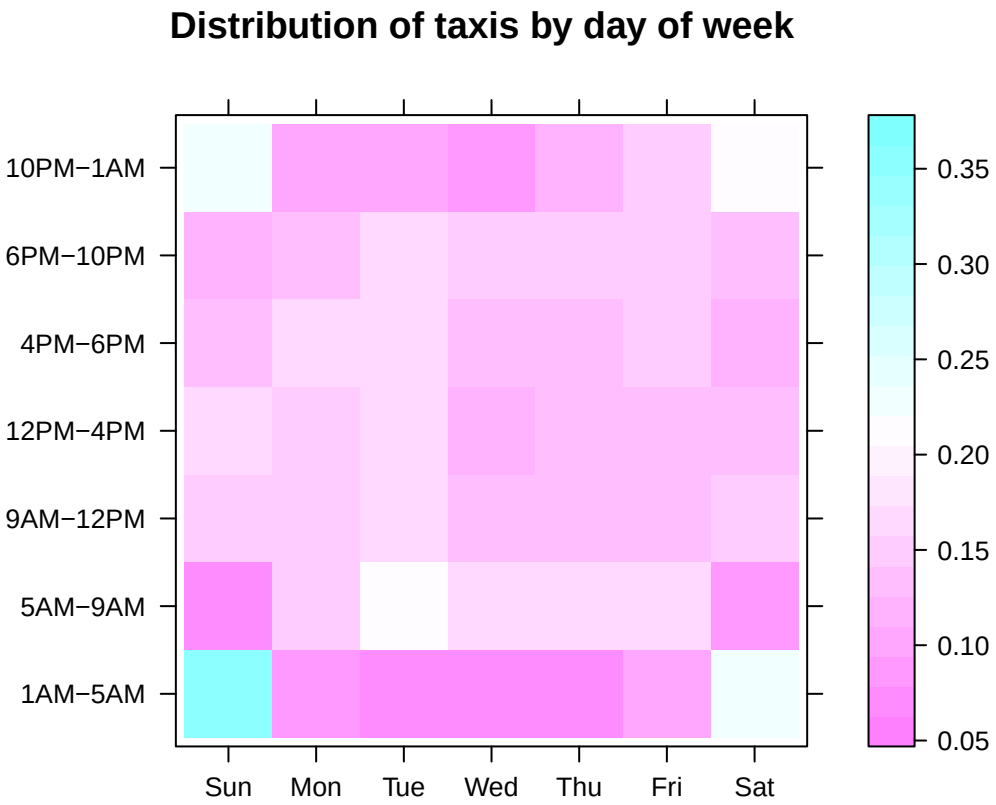
Exploring the data

After reading the data in, we extract the days of the week, hours of the day and each trip's duration in seconds.

```
##  VendorID    tpep_pickup_datetime tpep_dropoff_datetime passenger_count
##  1:4397921   Length:9710124       Length:9710124        Min.   :0.000
##  2:5312203   Class :character     Class :character      1st Qu.:1.000
##              Mode  :character     Mode  :character      Median :1.000
##                                                       Mean   :1.629
##                                                       3rd Qu.:2.000
##  trip_distance     RatecodeID   store_and_fwd_flag  PULocationID
##  Min.   :  0.000   1 :9459658   N:9664833          Min.   :  1.0
##  1st Qu.:  0.950   2 : 198715   Y:  45291          1st Qu.:114.0
##  Median :  1.600   3 :  16820                      Median :162.0
##  Mean   :  2.814   4 :   4569                      Mean   :164.1
##  3rd Qu.:  2.900   5 :  30086                      3rd Qu.:233.0
##   DOLocationID   payment_type  fare_amount           extra
##  Min.   :  1.0   1:6505711    Min.   :  -350.0   Min.   :-55.2000
##  1st Qu.:107.0   2:3144709    1st Qu.:     6.5   1st Qu.:  0.0000
##  Median :162.0   3:  46256    Median :     9.0   Median :  0.0000
##  Mean   :161.8   4:  13447    Mean   :    12.4   Mean   :  0.3235
##  3rd Qu.:234.0   5:      1    3rd Qu.:    13.5   3rd Qu.:  0.5000
##     mta_tax          tip_amount        tolls_amount
##  Min.   :-0.5000   Min.   : -41.000   Min.   :-15.0000
##  1st Qu.: 0.5000   1st Qu.:   0.000   1st Qu.:  0.0000
##  Median : 0.5000   Median :   1.300   Median :  0.0000
##  Mean   : 0.4975   Mean   :   1.751   Mean   :  0.2764
##  3rd Qu.: 0.5000   3rd Qu.:   2.260   3rd Qu.:  0.0000
##  improvement_surcharge  total_amount       tip_percent
##  Min.   :-0.3000       Min.   :  -350.3   Min.   :  0.00
##  1st Qu.: 0.3000       1st Qu.:     8.3   1st Qu.:  0.00
##  Median : 0.3000       Median :    11.3   Median : 19.00
##  Mean   : 0.2997       Mean   :    15.5   Mean   : 14.31
##  3rd Qu.: 0.3000       3rd Qu.:    16.8   3rd Qu.: 23.00
##    pickup_hour      pickup_day      dropoff_hour     dropoff_day
```

```
##  1AM-5AM : 535763   Sun:1477410   1AM-5AM : 555384   Sun:1485641
##  5AM-9AM :1496536   Mon:1369976   5AM-9AM :1421727   Mon:1371239
##  9AM-12PM:1384927   Tue:1542560   9AM-12PM:1375944   Tue:1540597
##  12PM-4PM:2033496   Wed:1257135   12PM-4PM:2026060   Wed:1258075
##  4PM-6PM :1156910   Thu:1326441   4PM-6PM :1130920   Thu:1323344
##  trip_duration_sec
##  Min.   :-3624015
##  1st Qu.:     374
##  Median :     618
##  Mean   :     891
##  3rd Qu.:    1001
```

Here's the plot of the proportion of pick-ups for each time interval across days of the week.

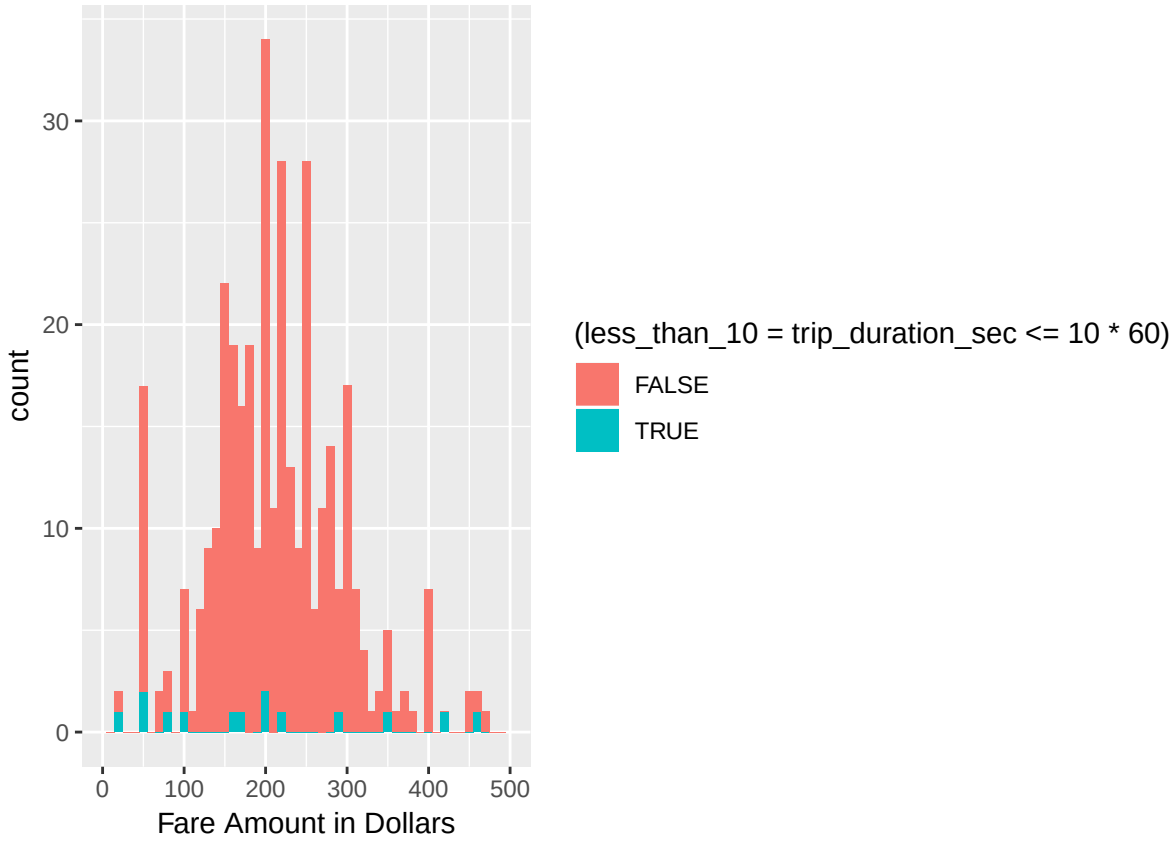


Besides the obvious that Saturday nights i.e. Sunday 1-5AM has the highest number of pick-ups as expected, Tuesday's are particularly busier compare to the rest of week espcially in the morning from 5 to 9 AM. This might be because people have already put the Monday blues behind and want to push through the week motivated.

Outliers

Identifying outliers can sometimes be subjective and require extra information from the source of data. However in this case, things such as negative fare amount or more than 5 passengers are very suspicious. Here we showed records that have at least one of the followings: negative or more than 50 miles trip distance, zero or more than five passengers or lastly negative fare amount. The plot demonstrates these dubious records (350947 instances) that has more than a 50 mile trip and whether or not this trip happened in less than 10

minutes!

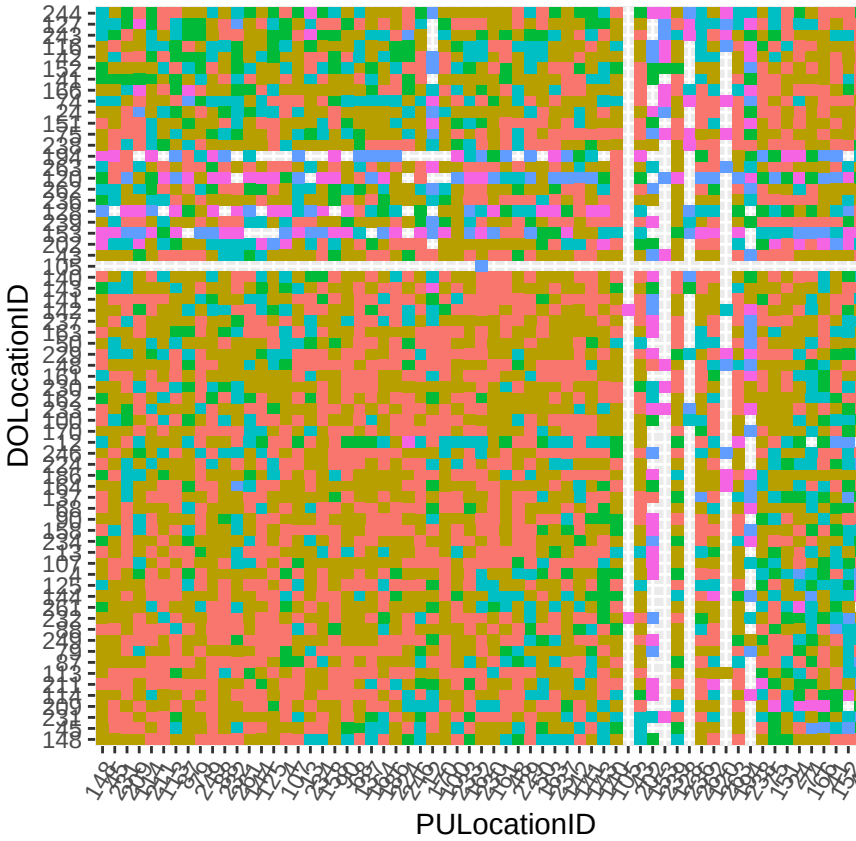


Adding Boroughs

In the latest available dataset (2017), pick up and drop off locations are identified by an ID. A separate csv file, howevere, is provided that specifies the zones and boroughs for each id. We merge the two datasets. The following table shows the number of trips between boroughs.

	Bronx	Brooklyn	EWR	Manhattan	Queens	Staten Island	Unknown
Bronx	5154	76	7	2222	197	0	132
Brooklyn	281	89364	82	35845	7316	62	435
EWR	0	4	645	12	1	1	33
Manhattan	42227	281301	14389	8208546	291278	1407	17655
Queens	8766	75518	449	297942	148154	563	9636
Staten Island	0	6	1	6	3	431	7
Unknown	245	1196	246	20018	2685	19	145561

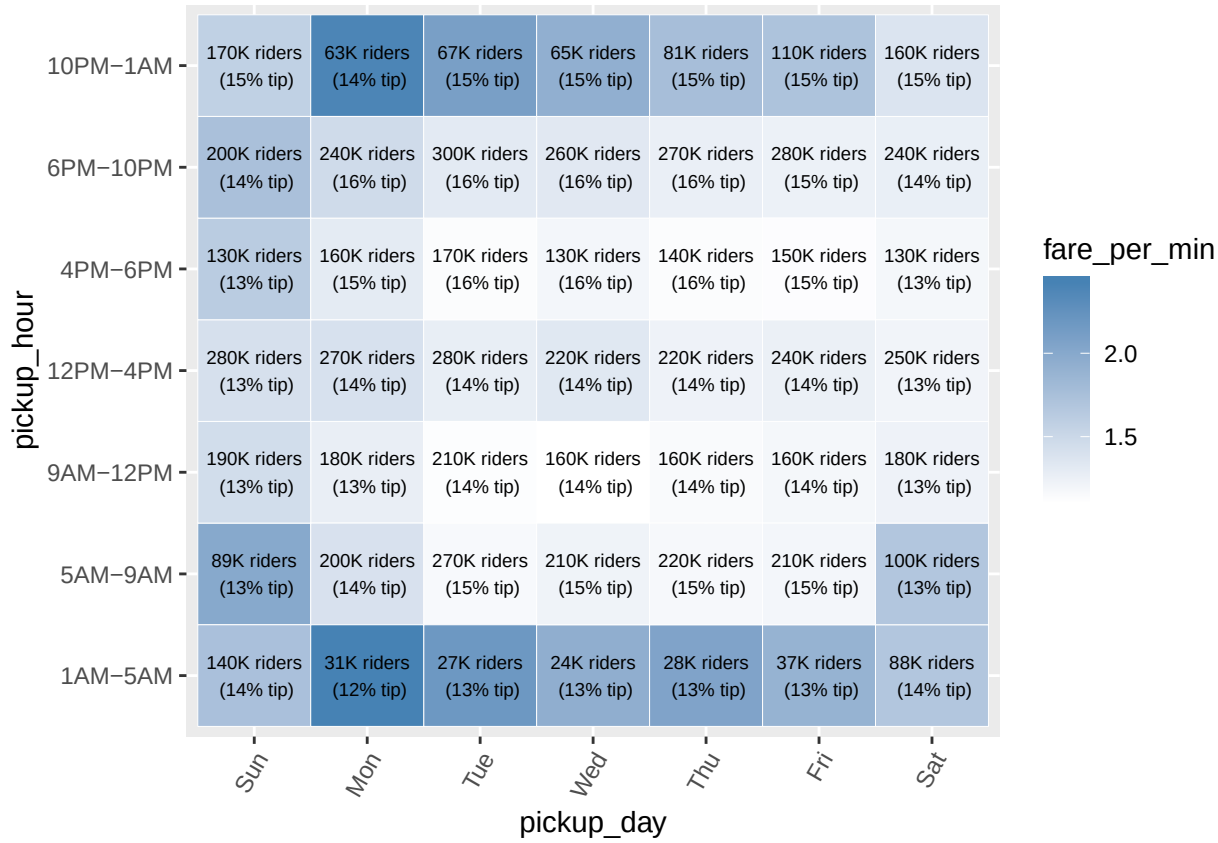
Since the trips from Manhattan to Manhattan takes the bulk of the data, we focus on Manhattan for now.



The following the tip brackets for all locations.

One interesting thing for instance that can be seen from this plot is that people who go from location 243 (i.e. Washington Heights North Zone) to location 148 (i.e. Lower East Side) on average pay 0 tips but for the reverse trip pay between 15 to 20 percents!

The following plot shows the fare dollar amount per minute for each day and hour as well as tip percentage that poeple pay on average for each time slot.



We can see that from 1 to 5 am during the week people consistantly pay less tip. Also from 4 to 10 pm i.e. after work, people seems to be more generous compare to the other times of the week days. During the week, later at night i.e. from 10 pm to 5 am, fare amount per minute is clearly higher. This can perhaps be due to light traffic, however, this different is less obvious over the weekends.

Fitting an actual model

For this section, I considered using random forest (RF) to predict the tip percentage.

After splitting the data into two subsets namely trianing and test, we need to reduce the number of categories for location ids. This is because RF in r cannot handle more than 53 levels for categorical variable. For this reason, we sorted the data based on the number of records for each location and then merge the smallest counts into one group called 'Rest'. Next, I'm planning to tune the model and in future consider other boroughs and perhaps different months to see if there's consistant trends across months.