NYC Taxi & FHV Project - Exploratory

Code

Jie Li, Xiaofan Zhang

8/11/2019

I. Introduction

This is a comprehensive Exploratory Data Analysis for 300 millions of for-hire vehicle (Uber, Lyft, Via) trips originating in New York City from **2018-01-01** to **2018-12-31**, and we focuse on trip counts and duration in New York competition with tidy R, ggplot2, and plotly.

The goal of this challenge is to process large data sets and to understand the duration of FHV in NYC based on features: trip location, pick-up, drop-off time, and weather effect. Also, we are interested in the difference betweeen three companies such as market shares, targeted customers, and business strategy. First of all, we analysis and visualize the original data, engineer new features, aggregate time-series variables to understand the data and pattern. Second, we compare three companies (Uber, Lyft, Via) over various time frame on trip amount and duration to analyze the market share and business strategy. Lastly, we add external NYC weather data to study how the weather impact on the trip duration and order requests in order to understand users behavior.

II. Description of the data source

The data were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page)).

The For-Hire Vehicle (“FHV”) trip records since 2009 until present including fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID. We are focusing on the time period from **2018-01-01** to **2018-12-31**, so the data comes in the shape of 200+ million observations, and each row contains one trip infomation.

The base license number is matching with different vehicle companies, so that we will join the `base-number` file to define the vehicle types, and we only focus on Uber, Lyft, Via at this point.

The NYC Taxi Zones map provided by TLC and published to NYC Open Data(https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc (https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc)). This map shows the NYC taxi zones corresponding to the pick up zomes and drop off zones, or location IDs, included in the FHV trip records. The taxi zones are roughly based on NYC Department of City Planning’s Neighborhood Tabulation Areas (NTAs) and are meant to approximate neighborhoods.

The NYC Weather data is provided by National Centers For Environmental Information (https://www.ncdc.noaa.gov/data-access (https://www.ncdc.noaa.gov/data-access)). NCEI is the world’s largest provider of weather and climate data. Land-based, marine, model, radar, weather balloon, satellite, and paleoclimatic are just a few of the types of datasets available. The weather data we are using is collected from NY Central Park Station (USW00094728) from **2018-01-01** to **2018-12-31**, which contains daily weather records such as wind, precipitation, snow and snow depth.

Statistics through December 31, 2018:

- 17.2 GB of raw data
- 300+ million for-hire vehicle total trips
- 365 daily weather records

Existing problem:

- R reads entire data set into RAM all at once. Total 17.2 GB of raw data would not fit in local memory at once.

- R Objects live in memory entirely, which cause slowness for data analysis.
- The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness.

III. Description of data import / cleaning / transformation

3.1 Libraries and Dependencies

Code

3.2 Data collection

First of all, we write a `shell` script to download original data from public websites

Code

3.3 Data Import

Due to local memeory issue and efficiency issue in R, the solution of processing large data is using high-performance version of base R's data.frame `data.table` and randomly sampling 3% of total data.

we use `fread` function to boost data process, select the vehicle company (Uber, Lyft, Via) based on the license number, and store into `data.table` format to perform our structured data. Each row contains trip information such as travel time, pick-up, drop-off date, time, location ID, and vehicle type.

Weather data can be read by local `csv` file, and inner join on trip data by key `date` .

Code

3.4 Data Processing

Next, we use `lubridate:ymd_hms` and `lubridate:mdy_hms` transformat string to standard time stamp variables, and calucate the travel time in **second** by sbustracting drop-off time and pick-up time.

Code

3.5 Data Structure Overview

Let's have an overview of the first 10 rows of data.

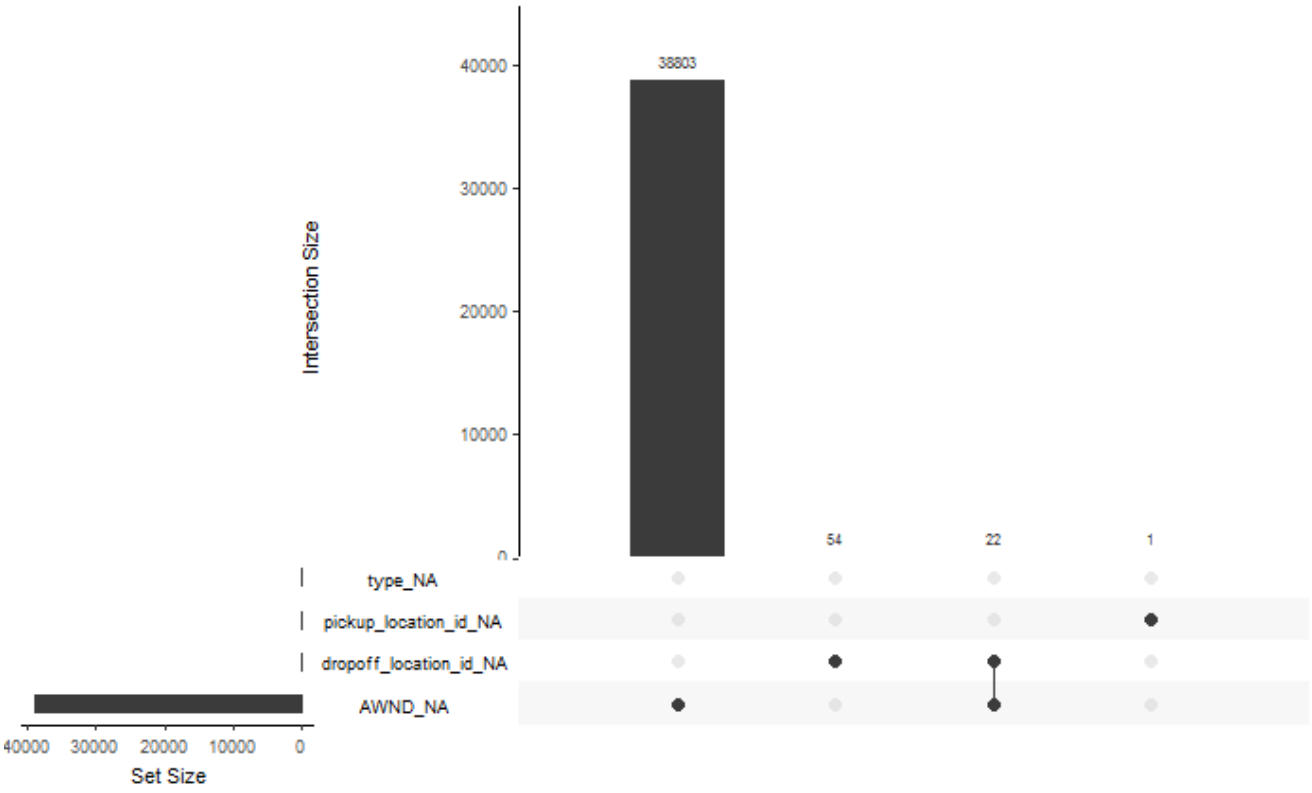
travel_time <int>	pickup_datetime <S3: POSIXct>	dropoff_datetime <S3: POSIXct>	date <date>	mo… <int>	… <int>	wkday <fctr>	pick_hc <ir
1633	2018-01-03 16:33:37	2018-01-03 17:00:50	2018-01-03	1	3	Wednesday	
365	2018-01-03 16:55:25	2018-01-03 17:01:30	2018-01-03	1	3	Wednesday	
2249	2018-01-03 16:54:46	2018-01-03 17:32:15	2018-01-03	1	3	Wednesday	
129	2018-01-03 17:09:33	2018-01-03 17:11:42	2018-01-03	1	3	Wednesday	
900	2018-01-03 16:29:48	2018-01-03 16:44:48	2018-01-03	1	3	Wednesday	
2852	2018-01-03 17:01:27	2018-01-03 17:48:59	2018-01-03	1	3	Wednesday	
2650	2018-01-03 16:22:59	2018-01-03 17:07:09	2018-01-03	1	3	Wednesday	

travel_time <int>	pickup_datetime <S3: POSIXct>	dropoff_datetime <S3: POSIXct>	date <date>	mo… <int>	… <int>	wkday <fctr>	pick_hc <ir
2994	2018-01-03 16:39:17	2018-01-03 17:29:11	2018-01-03	1	3	Wednesday	
1550	2018-01-03 16:08:57	2018-01-03 16:34:47	2018-01-03	1	3	Wednesday	
1052	2018-01-03 16:27:42	2018-01-03 16:45:14	2018-01-03	1	3	Wednesday	

1-10 of 155,088 rows | 1-8 of 17 columns Previous **1** 2 3 4 5 6 ... 100 Next

3.6 Data Missing & Outliers

Let’s visualize the pattern of missing value in the dataset.

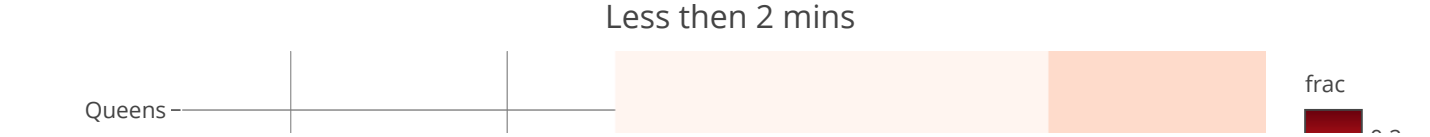


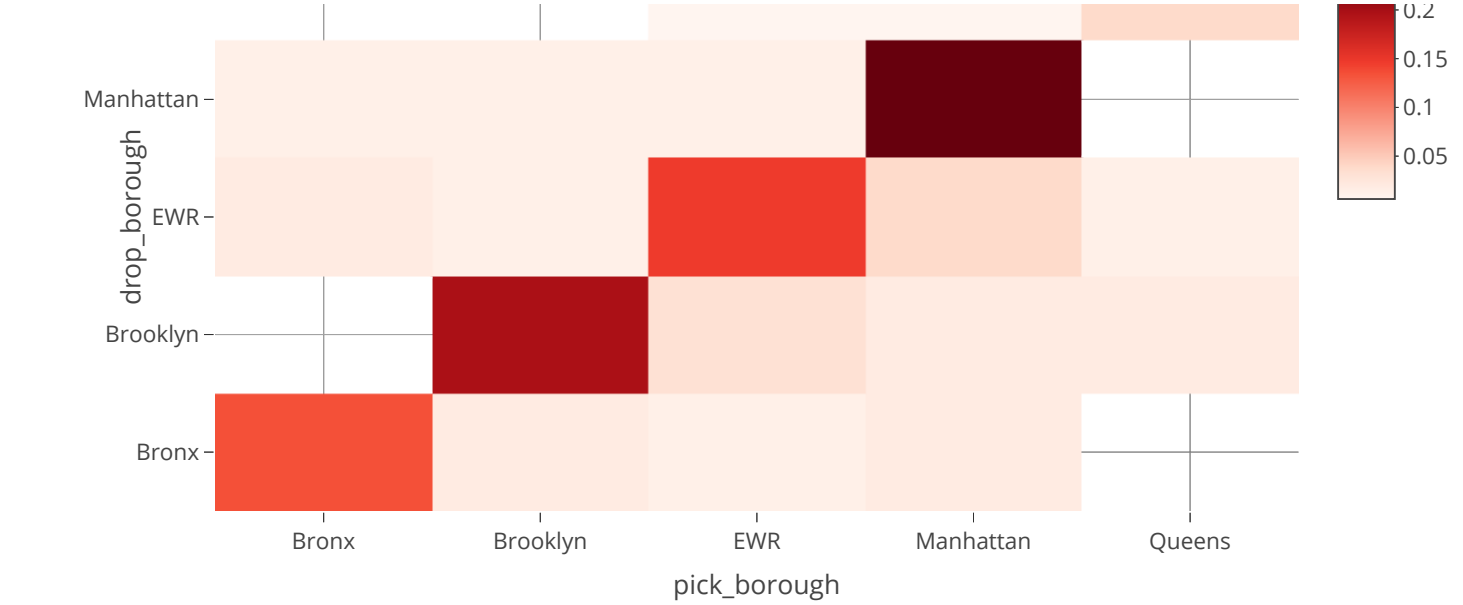
travel_time <int>	pickup_datetime <S3: POSIXct>	dropoff_datetime <S3: POSIXct>	date <date>	mo… <int>	… <int>	wkday <fctr>	pick_hc <ir
905	2018-03-01 06:30:50	2018-03-01 06:45:55	2018-03-01	3	1	Thursday	
1484	2018-03-12 22:12:41	2018-03-12 22:37:25	2018-03-12	3	12	Monday	
2162	2018-03-13 22:10:06	2018-03-13 22:46:08	2018-03-13	3	13	Tuesday	
2959	2018-03-16 20:04:40	2018-03-16 20:53:59	2018-03-16	3	16	Friday	
2286	2018-03-21 01:14:49	2018-03-21 01:52:55	2018-03-21	3	21	Wednesday	
1762	2018-03-24 02:17:15	2018-03-24 02:46:37	2018-03-24	3	24	Saturday	
1280	2018-03-25 03:00:08	2018-03-25 03:21:28	2018-03-25	3	25	Sunday	
2062	2018-03-26 21:10:52	2018-03-26 21:45:14	2018-03-26	3	26	Monday	
4077	2018-03-27 07:10:12	2018-03-27 08:18:09	2018-03-27	3	27	Tuesday	
1892	2018-04-09 20:14:54	2018-04-09 20:46:26	2018-04-09	4	9	Monday	

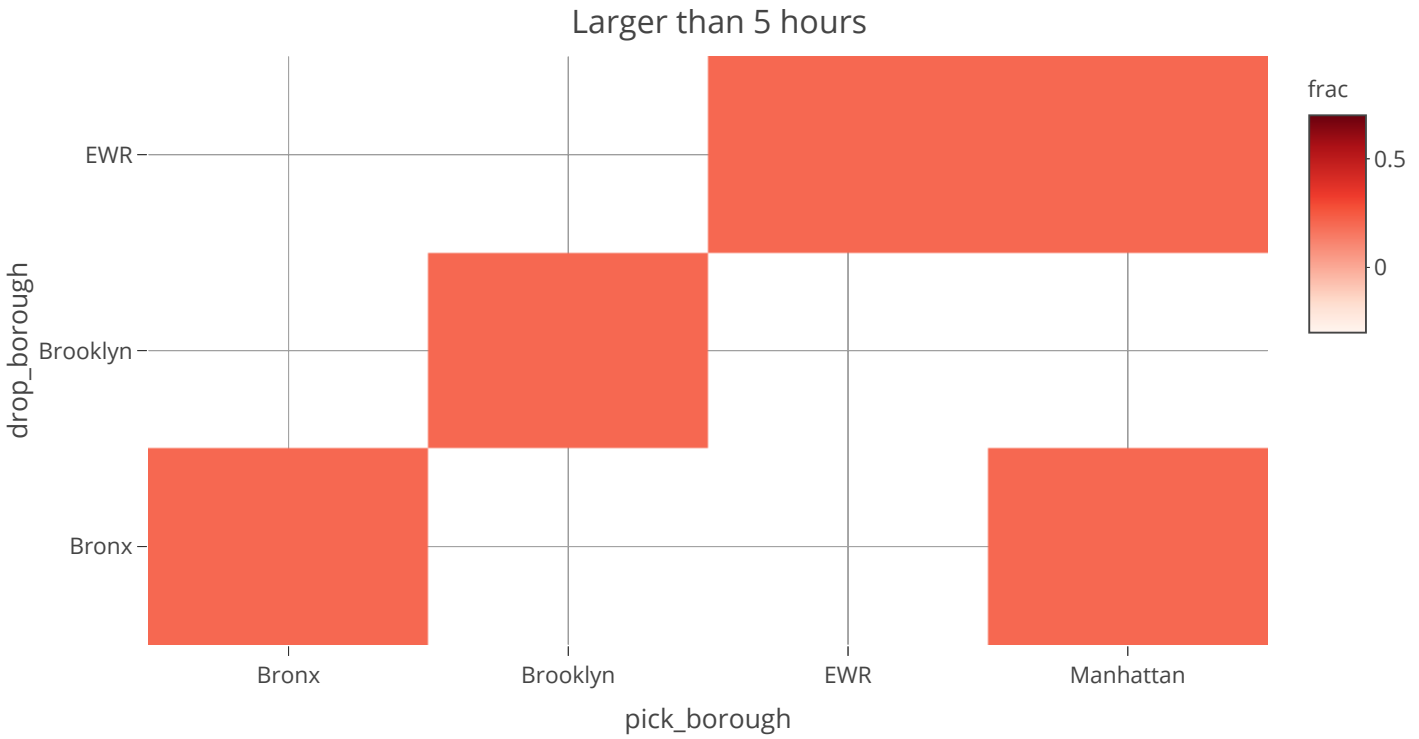
1-10 of 38,880 rows | 1-8 of 17 columns Previous **1** 2 3 4 5 6 … 100 Next

Due to FHV companies' policy, the trip is allowed to be cancelled in 2 minutes. In New York City, it's unlikely the travel time is longer than 5 hours.

Following **heat map** shows the percentage of trips are less than 2 minutes and larger than 5 hours for each pick-up and drop-off borough.



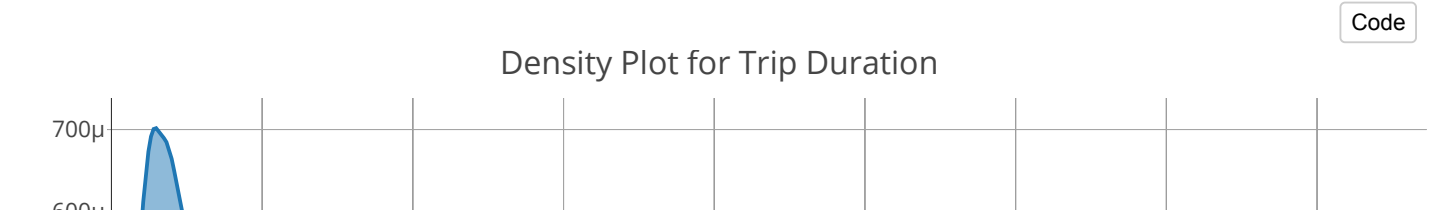


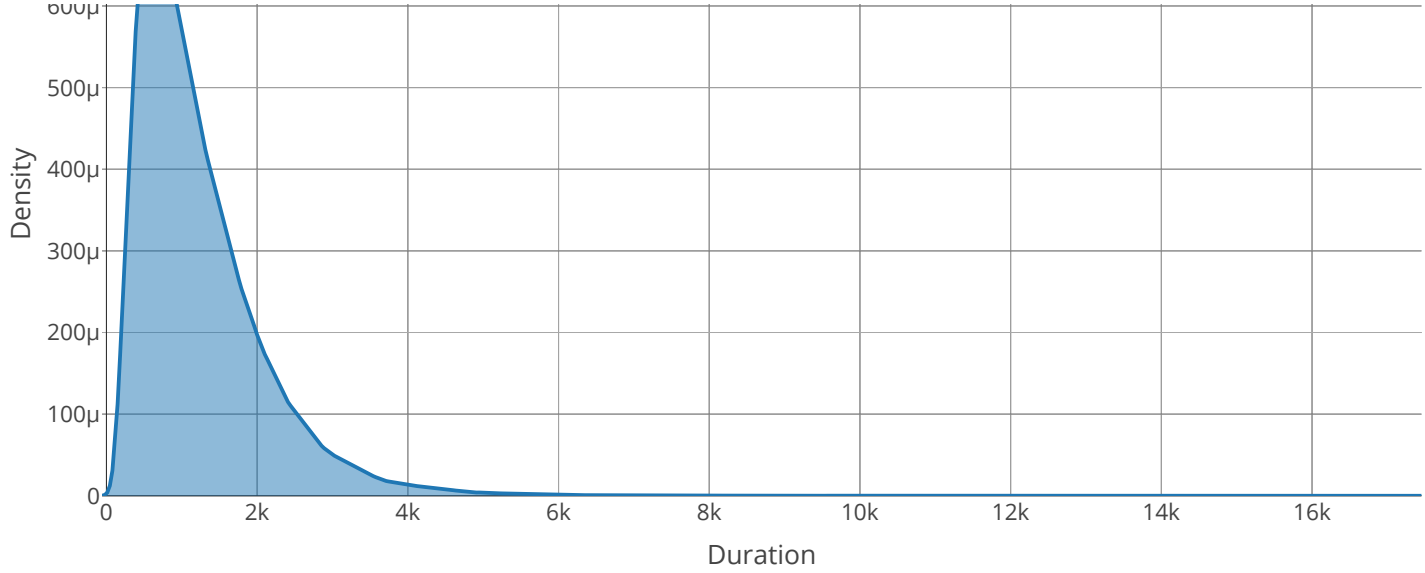


We find: * We found there are 25% missing value in the `AWND` and few in the `pickup_location_id` and `dropoff_location_id`. To conclude accurate analysis, we are going to fill in median of `AWND`, and remove data if `locationID` is missing * A lot of trip orders in the same borough are less than 2 minutes duration, which might be correct * However, it is impossible the trip is crossing two borough taking less than 2 minutes, so we are considered as cancalled orders.

3.7 Data Transformation

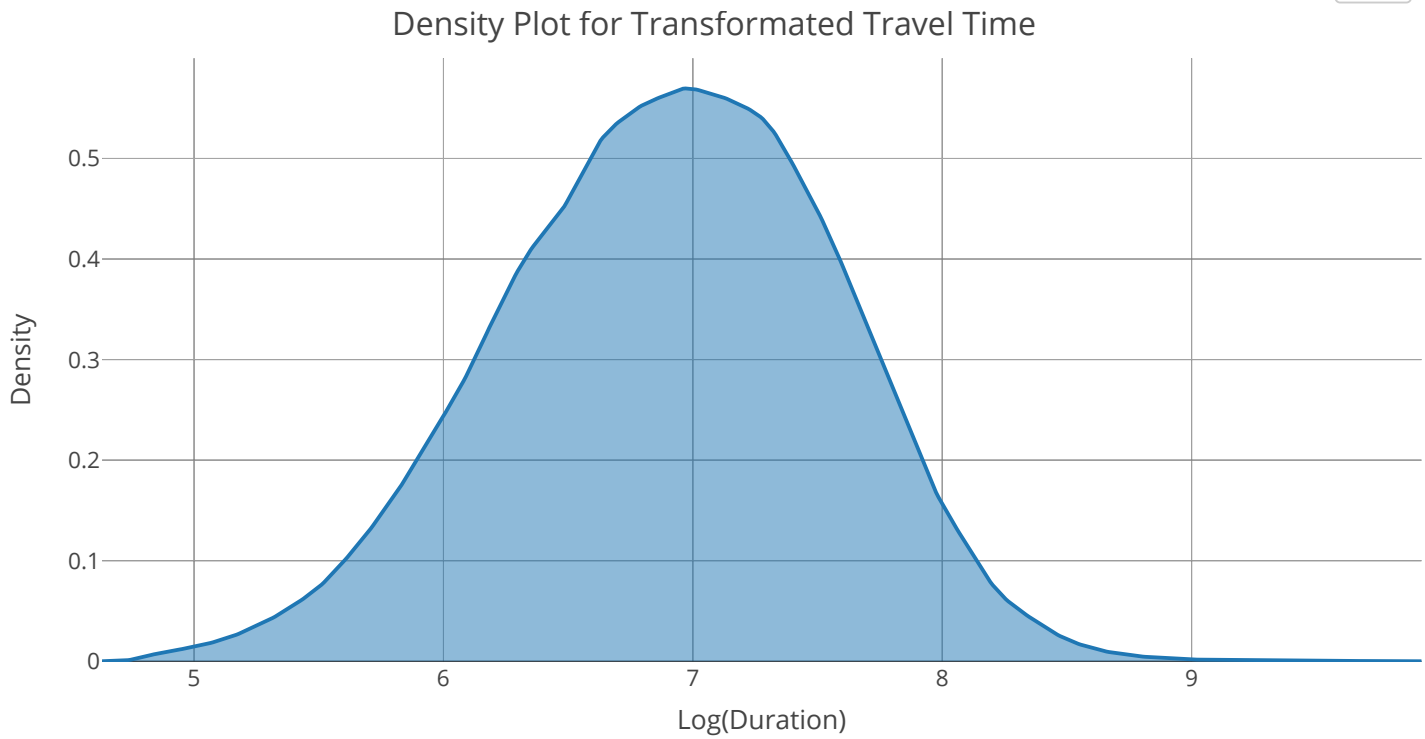
The density plot shows the duration distribution has a significantly right skew.





We might take **Log** transformation on the duration to solve the skewness issue for furture modeling.

Code



3.8 Data Aggregation

we can aggregate data by different levels in order to calcuate daily hour pick-ups, and median travel time in weekday.

Code

month.abb <chr>	wkday <fctr>	pick_hour <int>	N <int>
Jan	Wednesday	16	15
Jan	Wednesday	17	20
Jan	Wednesday	18	20

month.abb <chr>	wkday <fctr>	pick_hour <int>	N <int>
Jan	Wednesday	19	20
Jan	Wednesday	20	20
Jan	Wednesday	21	20
Jan	Wednesday	22	20
Jan	Wednesday	23	20
Jan	Thursday	0	20
Jan	Thursday	6	18

1-10 of 1,936 rows Previous **1** 2 3 4 5 6 ... 100 Next

Code

wkday <fctr>	pick_hour <int>	d.med <dbl>
Wednesday	16	1214.0
Wednesday	17	1210.0
Wednesday	18	1067.0
Wednesday	19	962.0
Wednesday	20	938.0
Wednesday	21	943.0
Wednesday	22	949.0
Wednesday	23	983.0
Thursday	0	936.5
Thursday	6	1133.0

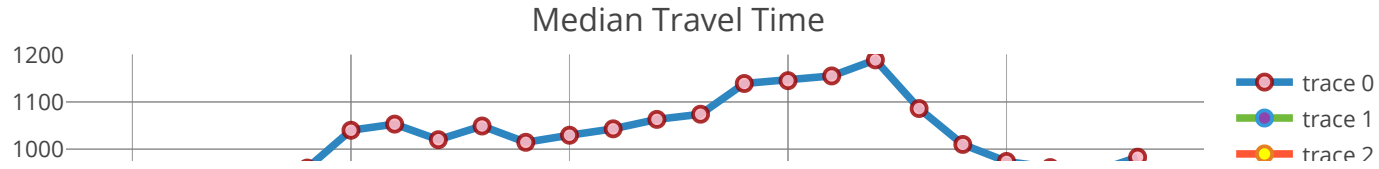
1-10 of 168 rows Previous **1** 2 3 4 5 6 ... 17 Next

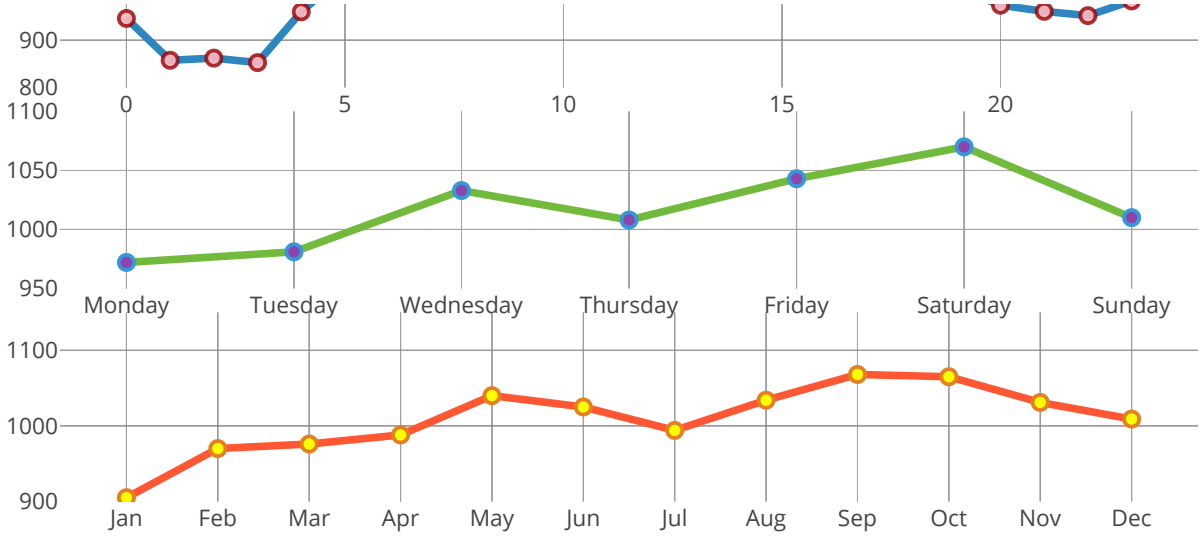
V. Results

5.1 Overall

First of all, we would likd to view overall median travel time based on hour, weekday and month bases. * The median is more robust measurement because it has less effect on outliers. * Hourly base is showing the peak hour effects in a typical day. * The weekly base tells the difference between work day and weekends. * The monthly base has a good explaination on seasonality.

Code





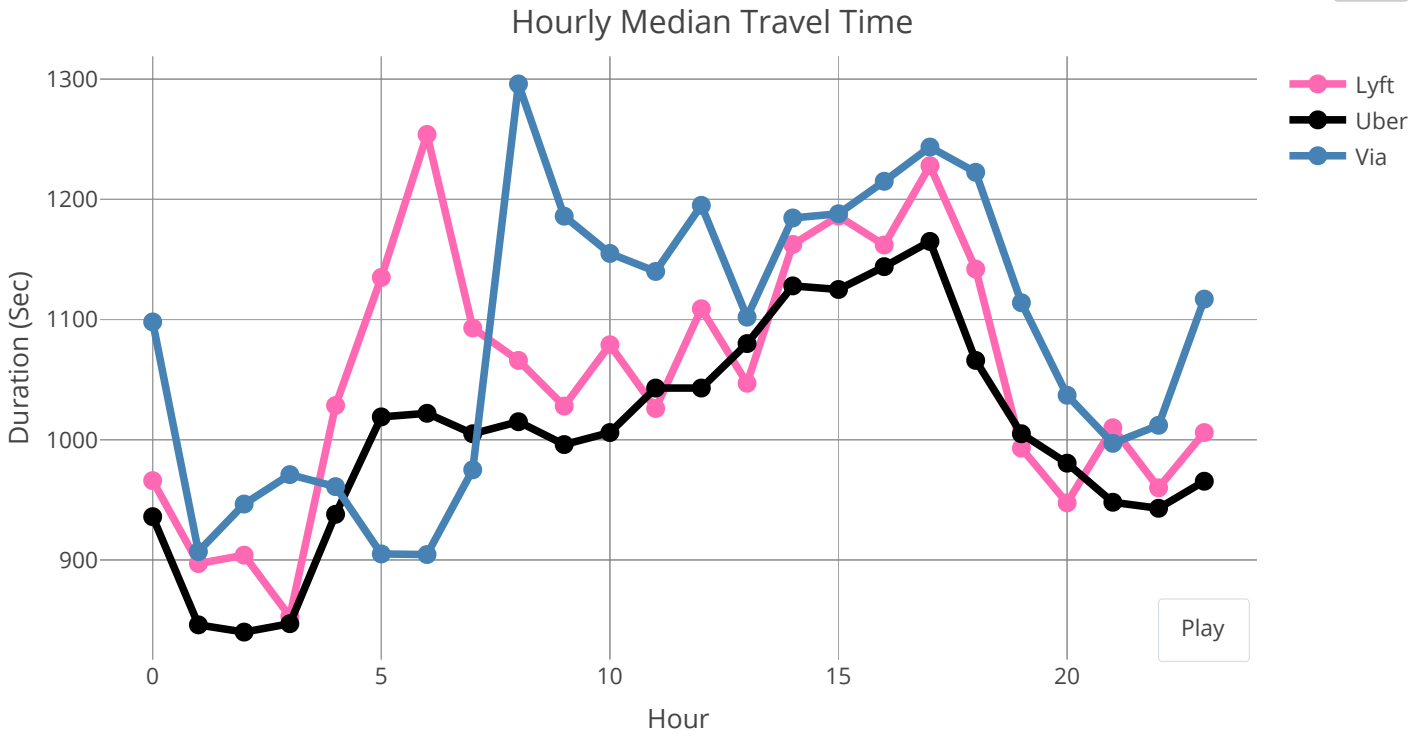
We find:

- The rush hour occurs from 8 AM - 7 PM about 18 minutes, othervise less traffic is after 8 PM
- The highest travel time is 20 minutes at 4 PM, the lowest travel time is 14 minutest at 2 AM (intersting!)
- During weekends, the travel time is higher than weekdays
- For monthly, it shows very strong seasonality that spring and fall have higher travel time; summer and winter are much lower.

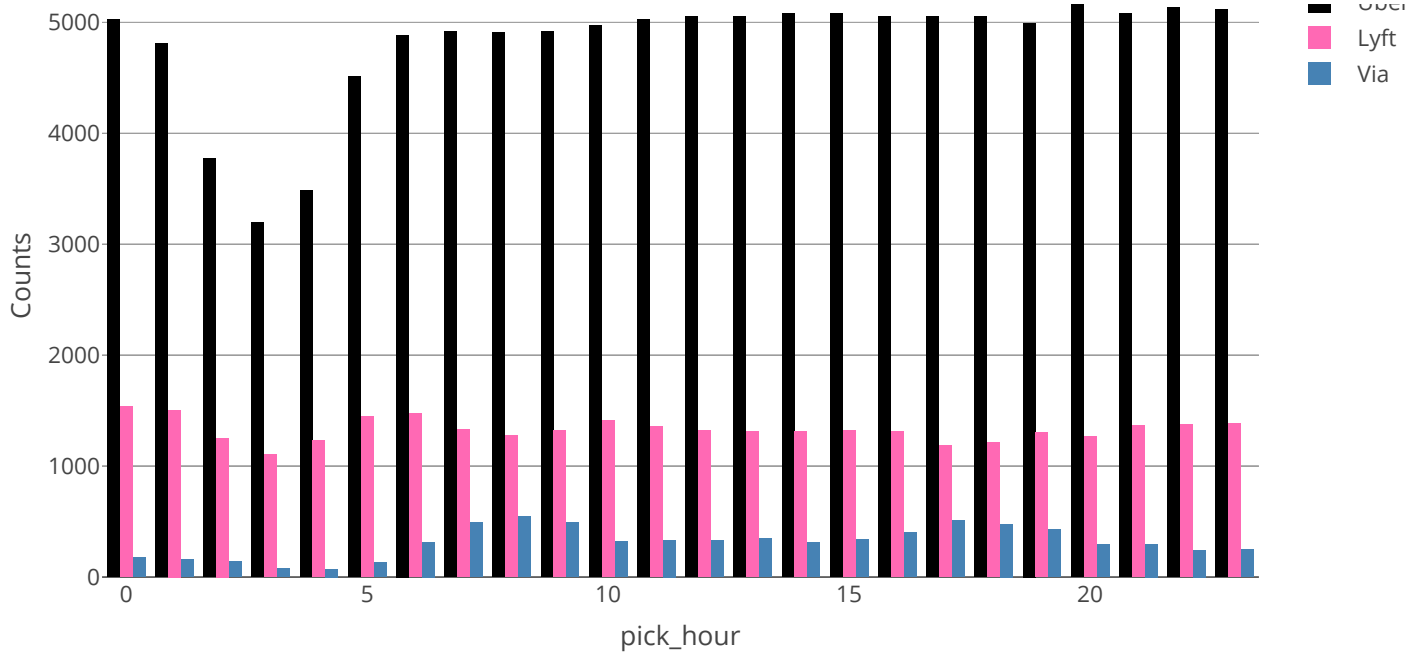
5.2 Types

We investigate on how the median travel time dependes on the different companies in hourly, weekly and monthly bases.

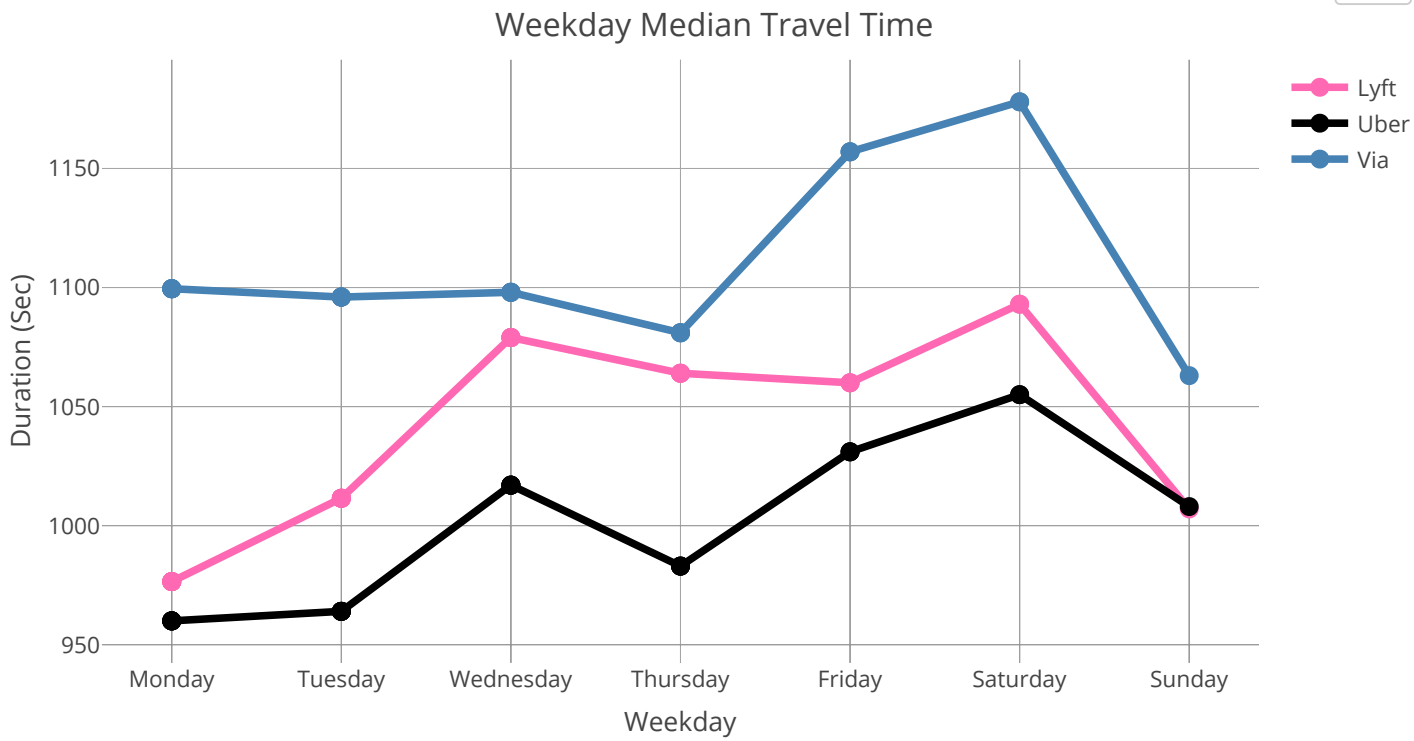
Code



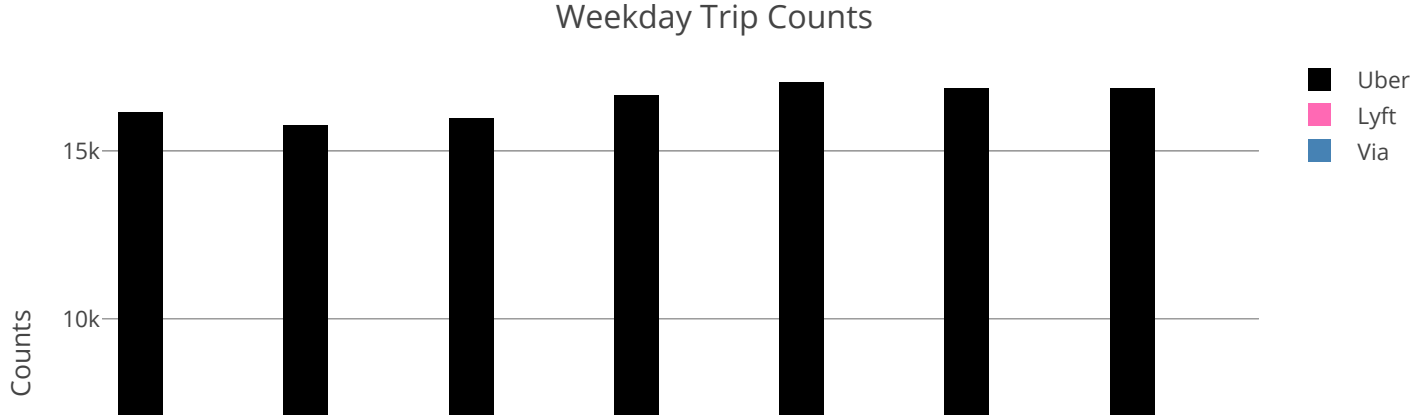
Code

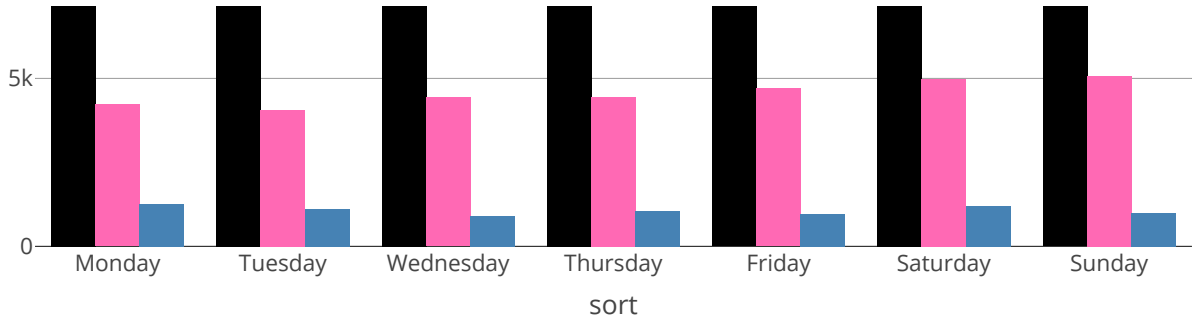

Uber
Lyft
Via
Counts
5000
4000
3000
2000
1000
0
0
5
10
15
20
pick_hour

Code

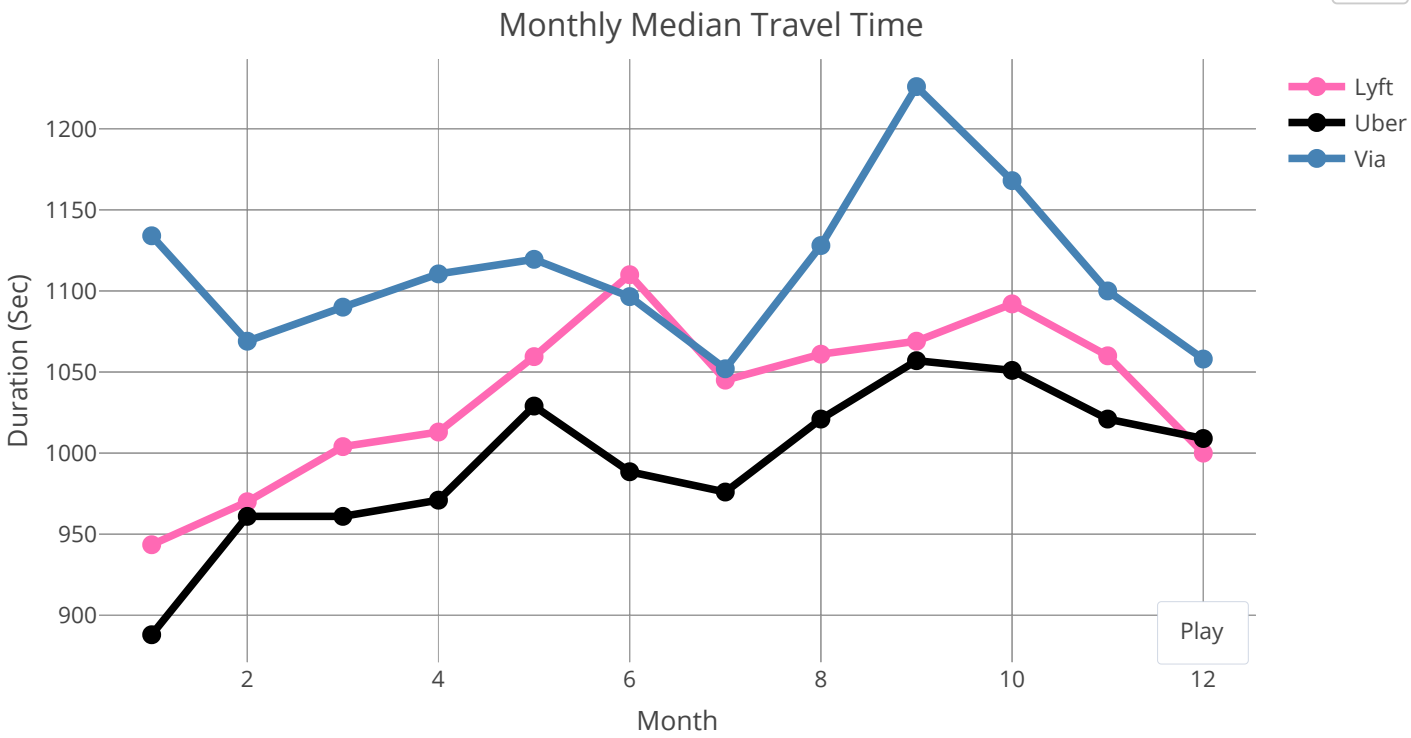

Weekday Median Travel Time
Lyft
Uber
Via
Duration (Sec)
1150
1100
1050
1000
950
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday

Code

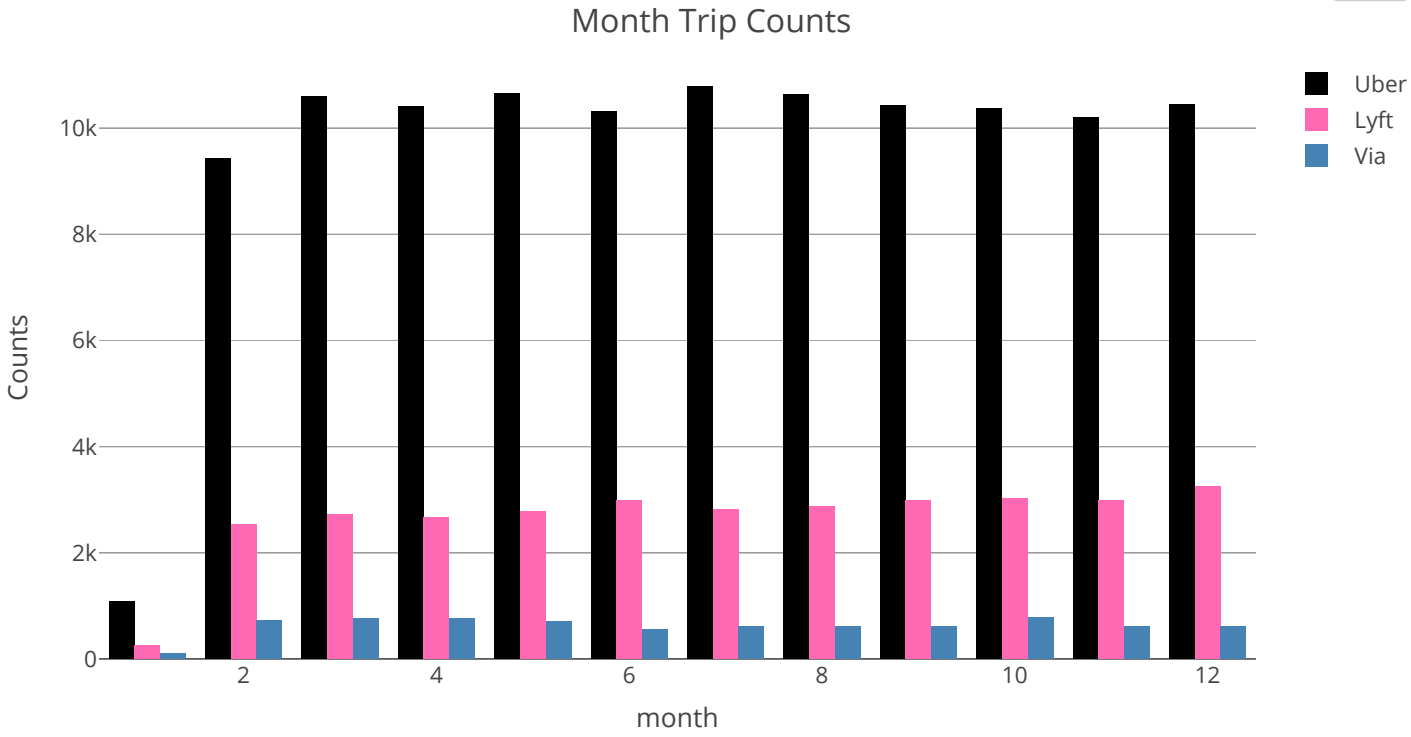

Weekday Trip Counts
Uber
Lyft
Via
Counts
15k
10k


5k
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
sort

Code


Monthly Median Travel Time
Lyft
Uber
Via
1200
1150
1100
1050
1000
950
900
Duration (Sec)
2
4
6
8
10
12
Month
Play

Code

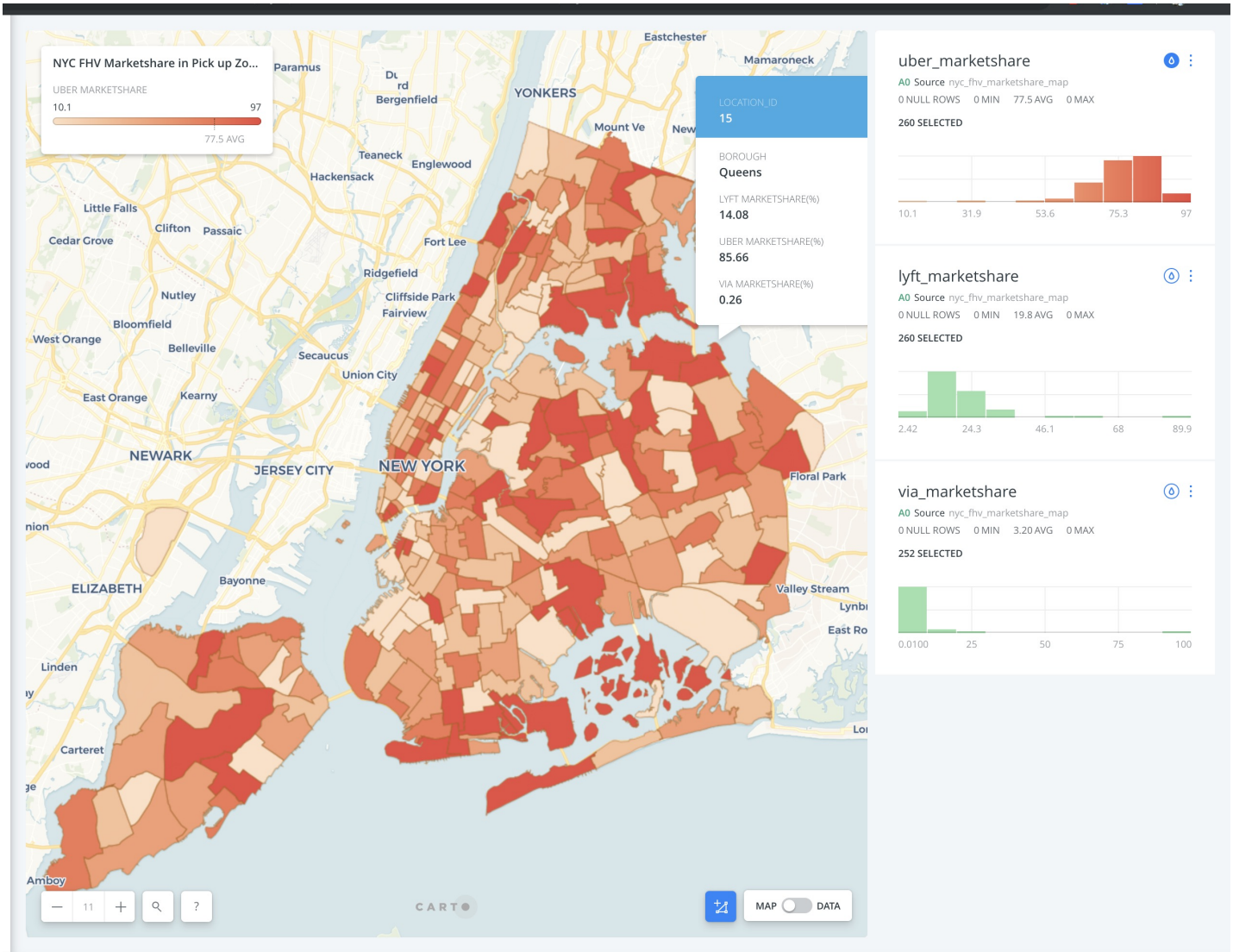

Month Trip Counts
Uber
Lyft
Via
10k
8k
6k
4k
2k
0
Counts
2
4
6
8
10
12
month

We find:

- For tipical day, Uber, Lyft and Via have simmilar trip duration in each hour
- For weekly base, Lyft has a litter higher trip duration than others, especially on Monday (interesting!)
- For monthly base, Via is the highest because most trips are share riders, which takes longer time
- Overally, Uber has lowest trip duration comparing other two!

5.3 Market Share

We also study the market shares on the both space and time line, so create an interactive map **NYC FHV Marketshare map** to indictate percentage of marketshare for Uber, Lyft and Via at different pick up zone. By simply clicking the map, you can see marketshare data in each zone. The legend lies in the right hand side, where you can also alter different views for each types of taxi by clicking three Teardrop-shaped buttons of applying auto style.
(https://zxf71699.carto.com/builder/62d8c815-2839-41fe-95e0-84ac6e4eccb6/embed
(https://zxf71699.carto.com/builder/62d8c815-2839-41fe-95e0-84ac6e4eccb6/embed))



NYC FHV Marketshare map (https://zxf71699.carto.com/builder/62d8c815-2839-41fe-95e0-84ac6e4eccb6/embed
(https://zxf71699.carto.com/builder/62d8c815-2839-41fe-95e0-84ac6e4eccb6/embed))

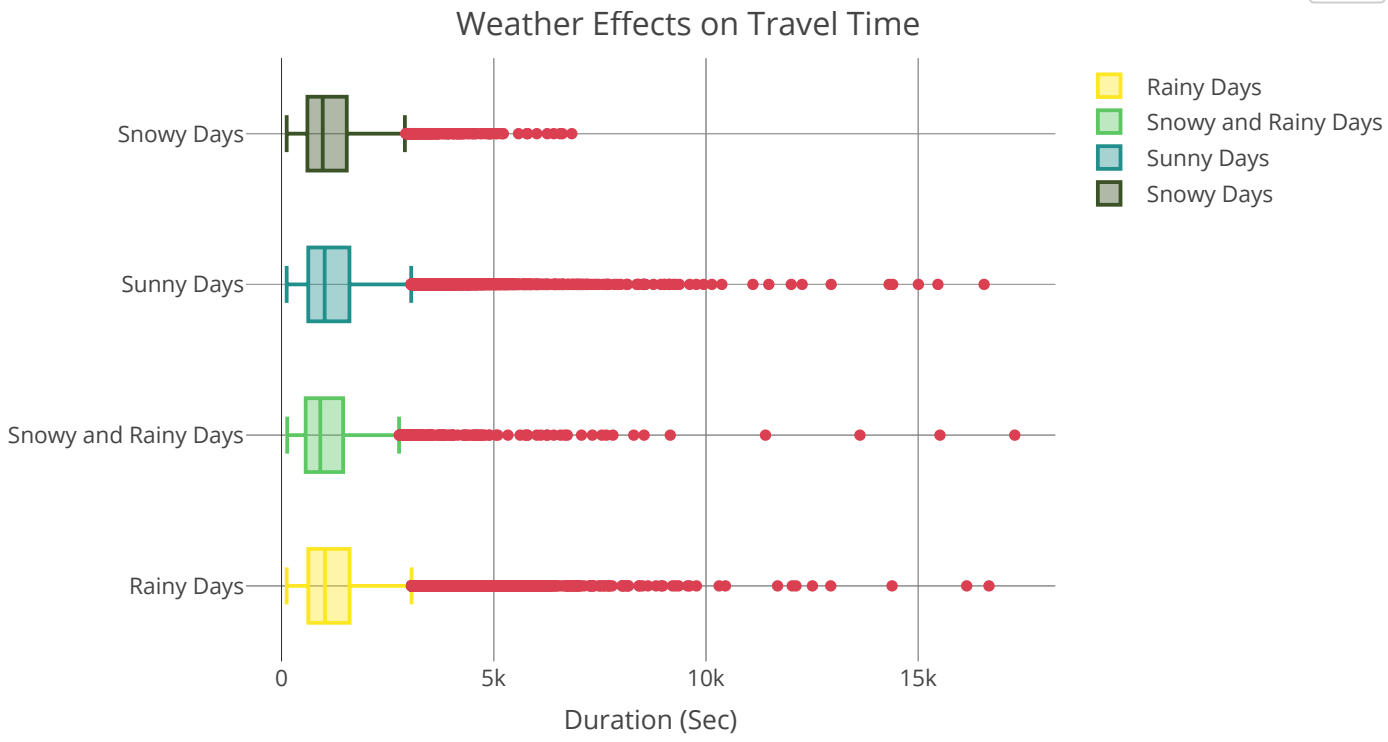
We find:

- Uber has dominant on the FHV market, reaching 75% entire market
- Lyft is second dominant on the FHV market, and weekends have higher numbers of trips.
- Via is a growing company, so it takes a small proportion of market, but it focuses on the peaking hour.

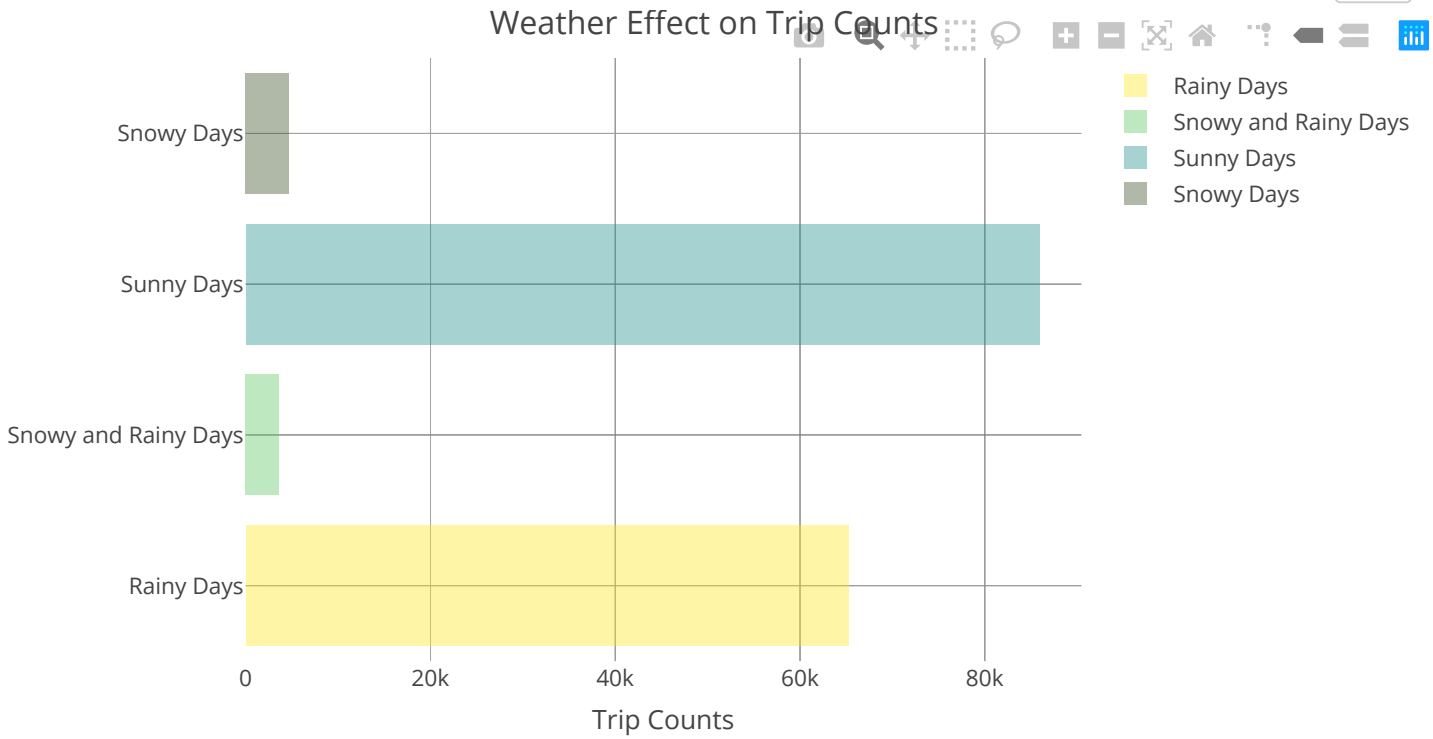
5.4 Weather Effect

We have encouraged to supplement our analysis with combining the extural NYC weather data to study how weather is impacted on the trip duration. Of particular interest here will be the rain, snow fall, and sun statistics.

Code



Code



We find:

- For sunny days, there are the largest amount of trip requests. It tells most people prefer to hand out.
- Rain causes the larger amount order requests and longer trip duration
- For snowy days, there are a few outliers, so it might tells more likely occurs extremely cases in the bad weather.

- It seems more like snow would lead to shorter trips, so it could simply mean passengers were more likely to travel shorter distances, or stay at home.