Code

Group 9 Final Project: NYC Taxi-FHV Analysis

Jie Li, Xiaofan Zhang

April 25, 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**. We are focusing on the trip counts and duration in the different time windows to find out the data insights and user behavior. All the data analysis process has been uploaded to GitHub (https://github.com/Jay4869/NYC-Taxi-FHV-Project).

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

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 (https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc) provided by TLC and published to NYC Open Data. 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 (https://www.ncdc.noaa.gov/data-access) is provided by National Centers For Environmental. 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 January 1 to December 31, 2018:

- 17.2 GB of raw data
- 200+ 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

We list libraries for our data process, manipulation and visualization.

Code

3.2 Data collection

We write a `shell` script to download raw data from public NYC TLC websites

Code

3.3 Data Import & Cleaning

We use `data.table:fread` function to speed up loading data for each month, and select the vehicle company (Uber, Lyft, Via) based on the license number. At the time, we also export subset monthly data into `csv` file as back up. we combine all monthly data into `tibble` format to perform our strucuted data. Each row contains trip information such as pick-up, drop-off date, time, location ID.

Due to local memeory issue in R, we process half-year data at one time, and use **aggregation** technique to compute results and perform entire year analysis. We will explain more detail below.

Code

3.4 File structure and content

Let's have an overview of the first 5000 `Jan` and `Dec` data. We find the time format is different, so we would like to work on variable convertion and transformation such as standard time stamp.

	V1	Pickup_DateTime	DropOff_datetime	PUlocationID	DOlocationID	type
	<int>	<fctr>	<fctr>	<int>	<int>	<fctr>
1	1	2018-01-30 21:06:50	2018-01-30 21:15:34	56	129	Uber
2	2	2018-01-30 21:20:36	2018-01-30 21:35:29	129	112	Uber
3	3	2018-01-30 21:04:45	2018-01-30 21:16:34	47	42	Uber
4	4	2018-01-30 21:11:51	2018-01-30 21:40:35	49	131	Uber
5	5	2018-01-30 21:43:39	2018-01-30 21:49:59	98	121	Uber
6	6	2018-01-30 21:36:53	2018-01-30 21:44:55	235	235	Uber
7	7	2018-01-30 21:48:30	2018-01-30 21:51:30	169	235	Uber
8	8	2018-01-30 21:55:18	2018-01-30 22:15:28	235	208	Uber
9	9	2018-01-30 21:21:25	2018-01-30 21:46:49	231	265	Uber
10	10	2018-01-30 21:24:37	2018-01-30 21:35:23	123	29	Uber

1-10 of 5,000 rows Previous **1** 2 3 4 5 6 … 100 Next

	V1	Pickup_DateTime	DropOff_datetime	PUlocationID	DOlocationID	type
	<int>	<fctr>	<fctr>	<int>	<int>	<fctr>
1	1	12/2/2018 09:53	12/2/2018 10:15	179	181	Lyft
2	2	12/2/2018 18:21	12/2/2018 18:42	161	79	Lyft

V1 <int>	Pickup_DateTime <fctr>	DropOff_datetime <fctr>	PUlocationID <int>	DOlocationID <int>	type <fctr>
3	12/2/2018 23:01	12/2/2018 23:17	256	225	Lyft
4	12/8/2018 22:30	12/8/2018 22:54	233	9	Lyft
5	12/4/2018 18:31	12/4/2018 19:00	65	50	Lyft
6	12/4/2018 18:31	12/4/2018 18:58	95	132	Lyft
7	12/4/2018 18:31	12/4/2018 19:05	165	63	Lyft
8	12/4/2018 18:31	12/4/2018 19:17	37	76	Lyft
9	12/4/2018 18:31	12/4/2018 18:58	255	225	Lyft
10	12/4/2018 18:31	12/4/2018 18:53	225	76	Lyft

1-10 of 5,000 rows Previous **1** 2 3 4 5 6 ... 100 Next

3.5 Data Transformation

Next, we use `lubridate:ymd_hms` and `lubridate:mdy_hms` transformat string to standard time stamp variables, and calucate the trip duration in **minute** by sbustracting drop-off time and pick-up time. Also, we factorize the company types to save memory usage and furture visualization.

Code

PUlocationID <int>	DOlocationID <int>	type <fctr>	pick <S3: POSIXct>	drop <S3: POSIXct>	duration <dbl>
56	129	Uber	2018-01-30 21:06:50	2018-01-30 21:15:34	8.733333
129	112	Uber	2018-01-30 21:20:36	2018-01-30 21:35:29	14.883333
47	42	Uber	2018-01-30 21:04:45	2018-01-30 21:16:34	11.816667
49	131	Uber	2018-01-30 21:11:51	2018-01-30 21:40:35	28.733333
98	121	Uber	2018-01-30 21:43:39	2018-01-30 21:49:59	6.333333
235	235	Uber	2018-01-30 21:36:53	2018-01-30 21:44:55	8.033333
169	235	Uber	2018-01-30 21:48:30	2018-01-30 21:51:30	3.000000
235	208	Uber	2018-01-30 21:55:18	2018-01-30 22:15:28	20.166667
231	265	Uber	2018-01-30 21:21:25	2018-01-30 21:46:49	25.400000
123	29	Uber	2018-01-30 21:24:37	2018-01-30 21:35:23	10.766667

1-10 of 5,000 rows Previous **1** 2 3 4 5 6 ... 100 Next

Code

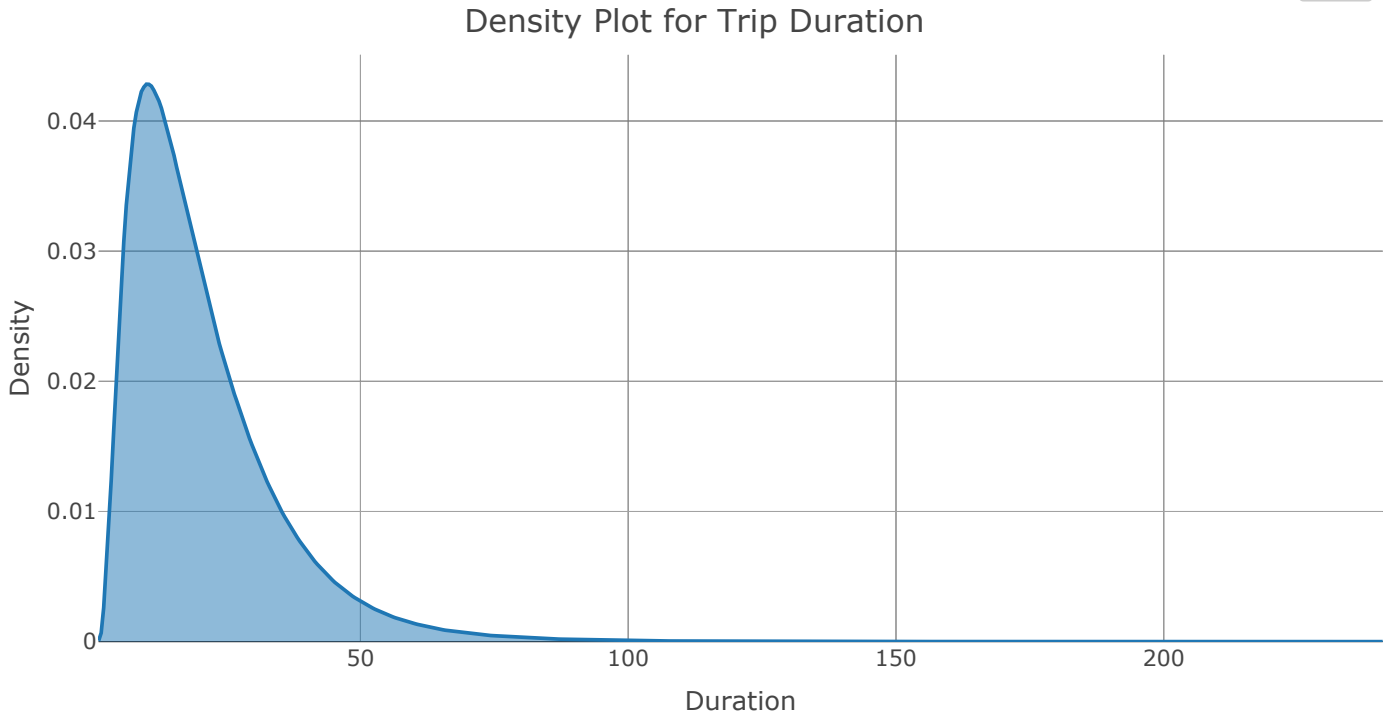
PUlocationID <int>	DOlocationID <int>	type <fctr>	pick <S3: POSIXct>	drop <S3: POSIXct>	duration <dbl>
179	181	Lyft	2018-12-02 09:53:00	2018-12-02 10:15:00	22
161	79	Lyft	2018-12-02 18:21:00	2018-12-02 18:42:00	21
256	225	Lyft	2018-12-02 23:01:00	2018-12-02 23:17:00	16
233	9	Lyft	2018-12-08 22:30:00	2018-12-08 22:54:00	24
65	50	Lyft	2018-12-04 18:31:00	2018-12-04 19:00:00	29

PUlocationID <int>	DOlocationID <int>	type <fctr>	pick <S3: POSIXct>	drop <S3: POSIXct>	duration <dbl>
95	132	Lyft	2018-12-04 18:31:00	2018-12-04 18:58:00	27
165	63	Lyft	2018-12-04 18:31:00	2018-12-04 19:05:00	34
37	76	Lyft	2018-12-04 18:31:00	2018-12-04 19:17:00	46
255	225	Lyft	2018-12-04 18:31:00	2018-12-04 18:58:00	27
225	76	Lyft	2018-12-04 18:31:00	2018-12-04 18:53:00	22

1-10 of 5,000 rows Previous **1** 2 3 4 5 6 ... 100 Next

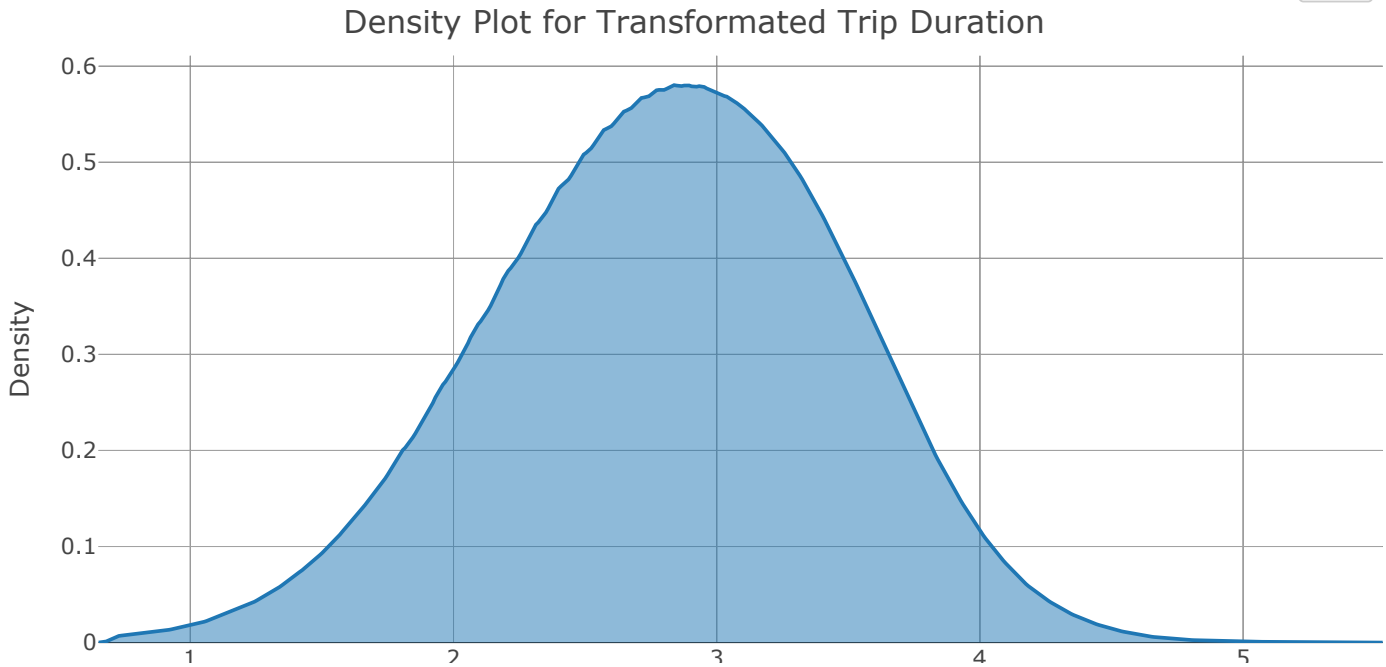
The density plot shows the duration distribution has a significantly right skew.

Code



We require to take **log** transformation on the duration to solve the skewness issue to normal distribution for furture modeling and statistical inference.

Code

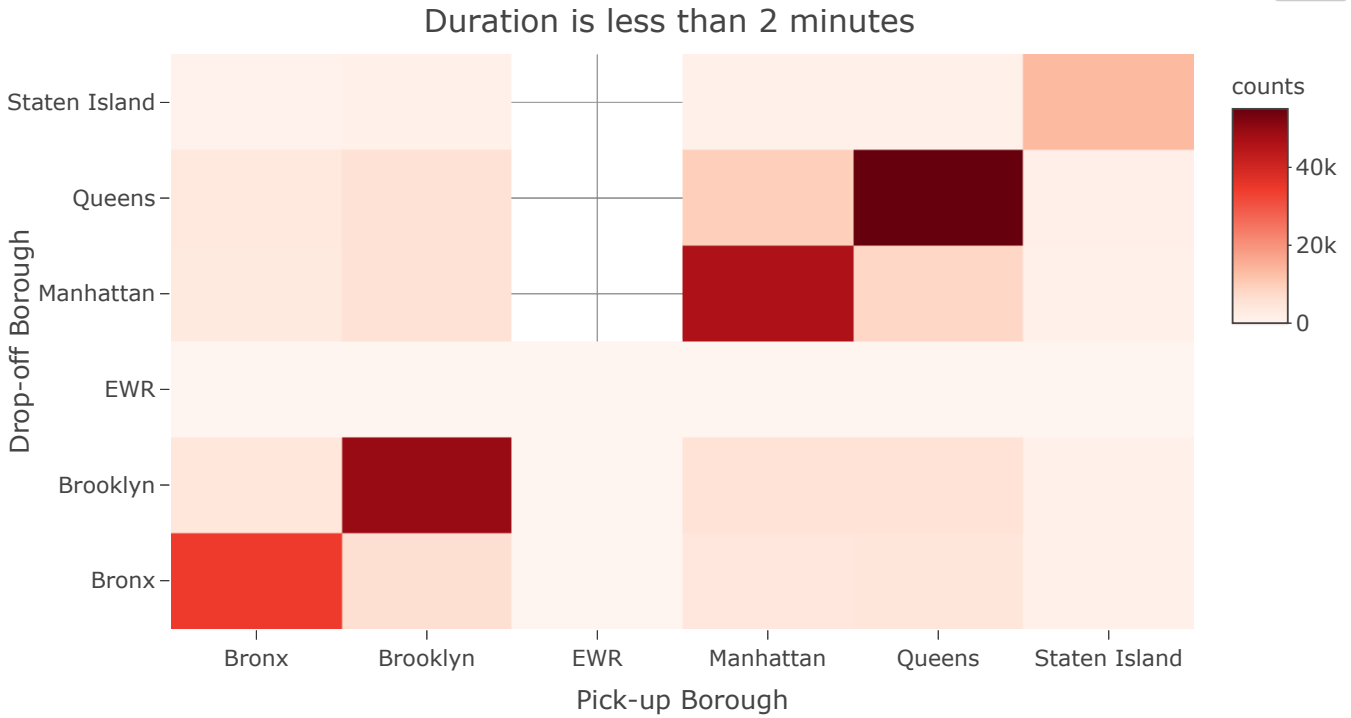


Log(Duration)

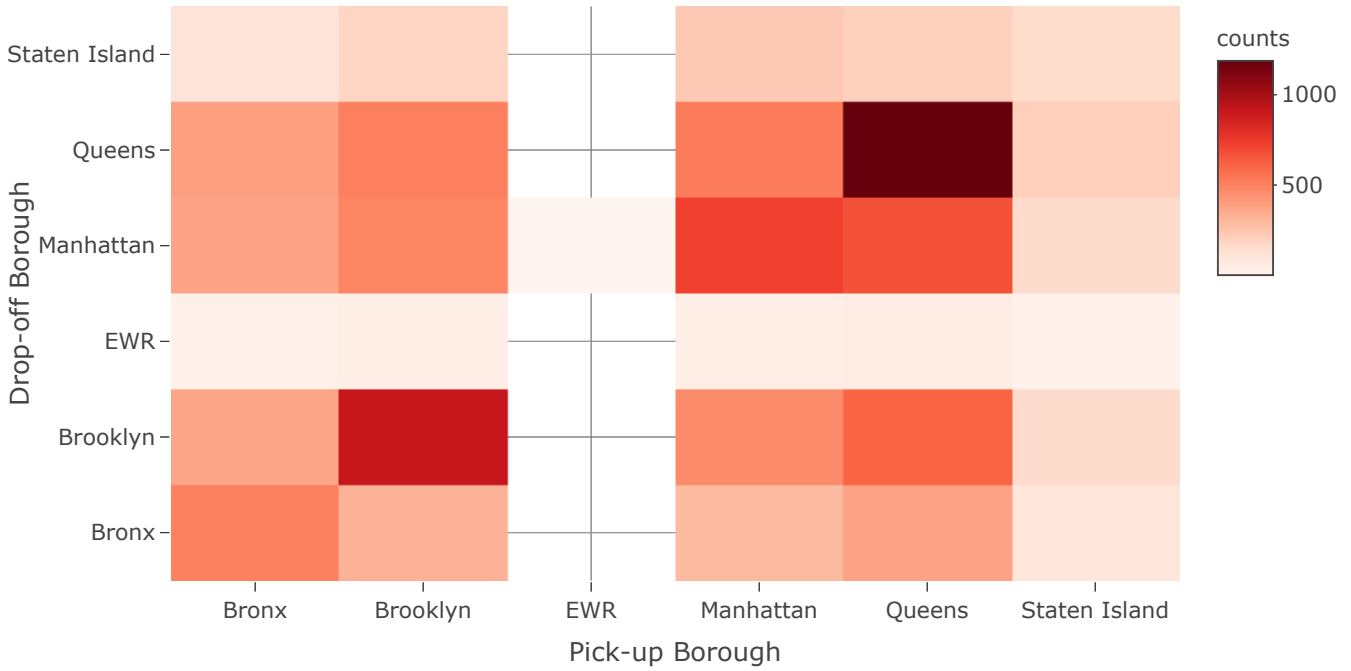
3.6 Data Missing & Outliers

Due to most companies allow to cancel the order in 2 minutes, and drives might miss the passengers in order to cancel the order. We use **heat map** to visualize the trips counts based on the pick-up and drop-off location in order to identify the possiblity.

Code



Code



Code

PUlocationID	DOlocationID	type	pick	drop	duration
<lgl>	<int>	<fctr>	<S3: POSIXct>	<S3: POSIXct>	<dbl>
NA	37	Lyft	2018-04-02 21:43:00	2018-04-02 22:00:00	16.816667

PUlocationID	DOlocationID	type	pick	drop	duration
<lgl>	<int>	<fctr>	<S3: POSIXct>	<S3: POSIXct>	<dbl>
NA	164	Lyft	2018-04-01 17:53:00	2018-04-01 18:18:00	25.083333
NA	17	Lyft	2018-04-04 01:23:00	2018-04-04 01:43:00	19.350000
NA	170	Lyft	2018-04-14 22:18:00	2018-04-14 22:34:00	16.100000
NA	144	Lyft	2018-04-02 09:39:00	2018-04-02 10:09:00	29.750000
NA	107	Lyft	2018-04-01 12:42:00	2018-04-01 12:52:00	10.133333
NA	40	Lyft	2018-04-04 20:09:00	2018-04-04 20:23:00	14.100000
NA	107	Lyft	2018-04-06 12:21:00	2018-04-06 12:39:00	17.966667
NA	100	Lyft	2018-04-08 17:15:00	2018-04-08 17:28:00	12.333333
NA	256	Lyft	2018-04-14 17:21:00	2018-04-14 17:41:00	20.683333

1-10 of 91,932 rows Previous **1** 2 3 4 5 6 ... 100 Next

We find:

- There are 279,693 trip records are less than 2 minutes duration, which might be incorrect or cancelled request.
- Also, there are 7900+ trips have duration longer than 4 hours, which does not make sense for close borough.
- There are missing values when pick-up location is EWR in our dataset. More specific, there are 91,932 records missing pick-up location. To conclude accurate analysis, we are going to remove all `NA` records.

3.7 Data Aggregation

By Solving local memory issue in R, since we are interested in the number of trips, and trip duraction, we don't have to store all data into R. The idea is that we can process half-by-half year data and aggregate into different levels such as hour, weekday, day, and month. Then, we combine aggregated results to make visualization plots, which are much smaller.

Code

hour	wday	type	n
<int>	<fctr>	<fctr>	<int>
0	Friday	Lyft	243525
0	Friday	Uber	809884
0	Friday	Via	28960
0	Monday	Lyft	198294
0	Monday	Uber	610848
0	Monday	Via	17374
0	Saturday	Lyft	397000
0	Saturday	Uber	1271771
0	Saturday	Via	49804
0	Sunday	Lyft	456128

1-10 of 504 rows Previous **1** 2 3 4 5 6 ... 51 Next

Code

month	type	d.med
<fctr>	<fctr>	<dbl>

month <fctr>	type <fctr>	d.med <dbl>
Apr	Lyft	17.10000
Apr	Uber	16.43333
Apr	Via	18.66667
Feb	Lyft	16.50000
Feb	Uber	15.81667
Feb	Via	18.03333
Jan	Lyft	16.21667
Jan	Uber	15.65000
Jan	Via	17.65000
Jun	Lyft	17.96667

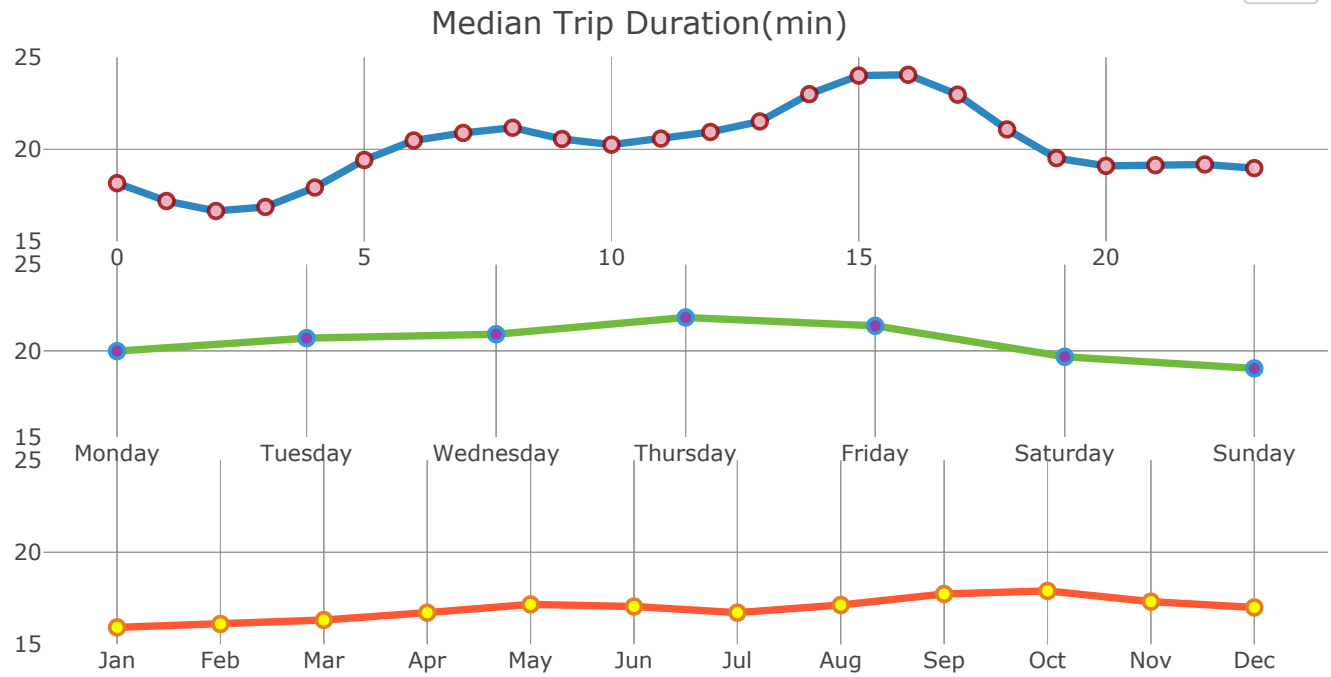
1-10 of 35 rows Previous **1** 2 3 4 Next

V. Results

5.1 Overall Median Duration

We look at overall median trip duration 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:

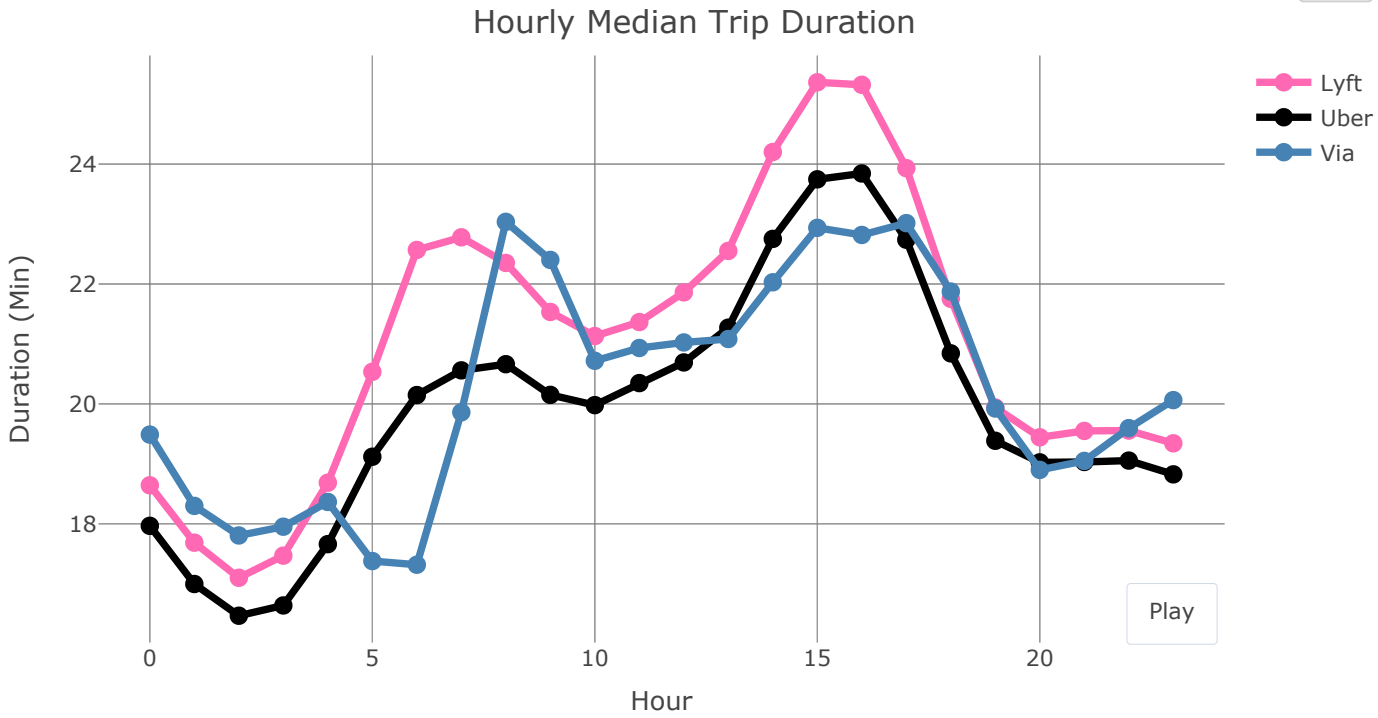
- After mid-night, the shortest trip duration is falling down and reaching 16 minutes; the peak trip duration occurs during 6 AM - 6 PM about 20+ minutes in a typical day, and the longest trip duration is 24 minutest at 4 PM (interesting!)

- For the weekday, the longest trip duration occurs on Thursday, and weekends have the lowest trip duration. We guess maybe more people prefer to stay at home on the weekends.
- The higher trip duration occurs in May and Oct, and the lower happens in Jan. The spring and fall are the best weather for traveling and visitors in NYC, so it might cause traffic.

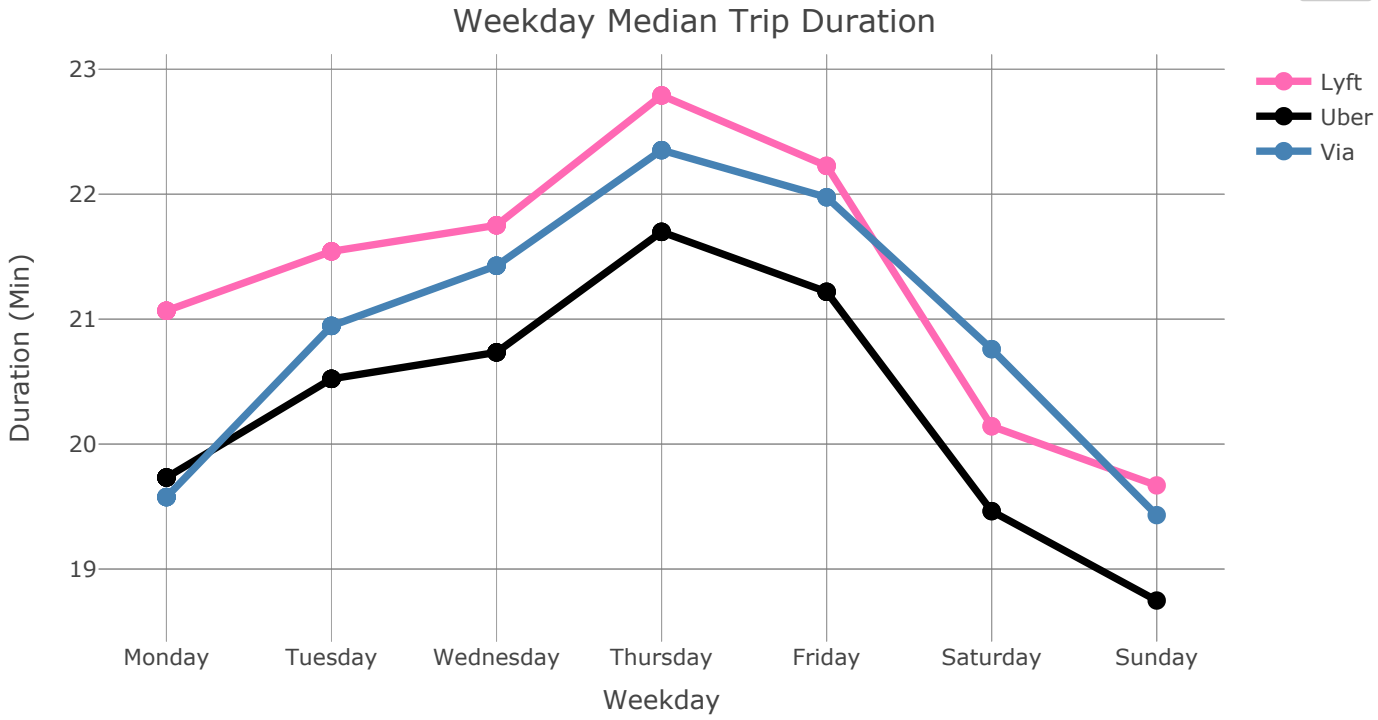
5.2 Durations of FHV Types

We investigate on how the median trip duration depends on the different for hire vehicle types such as Uber, Lyft and Via in hourly, weekly and monthly bases. Note: click `Play` you will find some interesting trends!

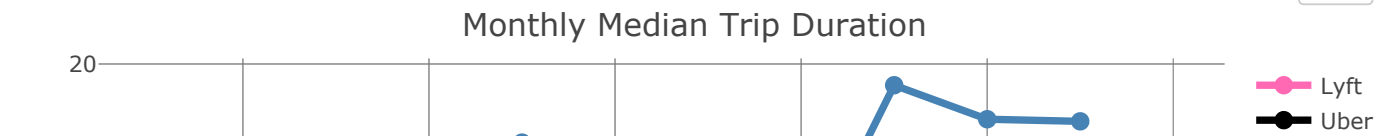
Code

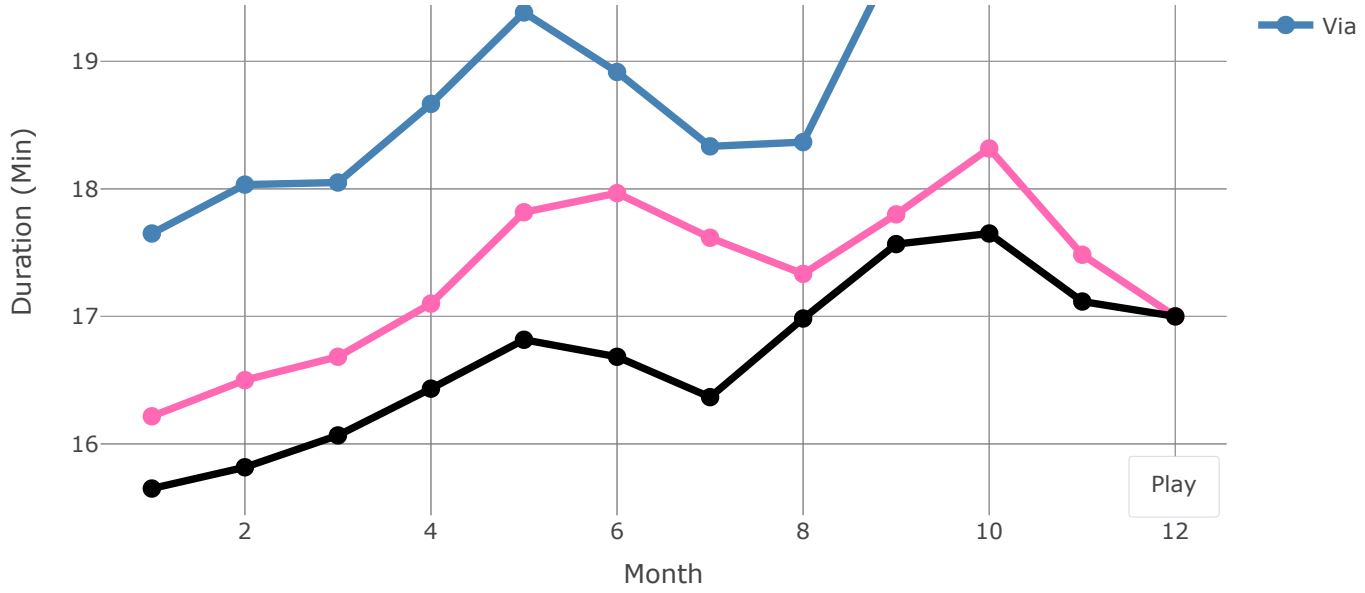


Code



Code



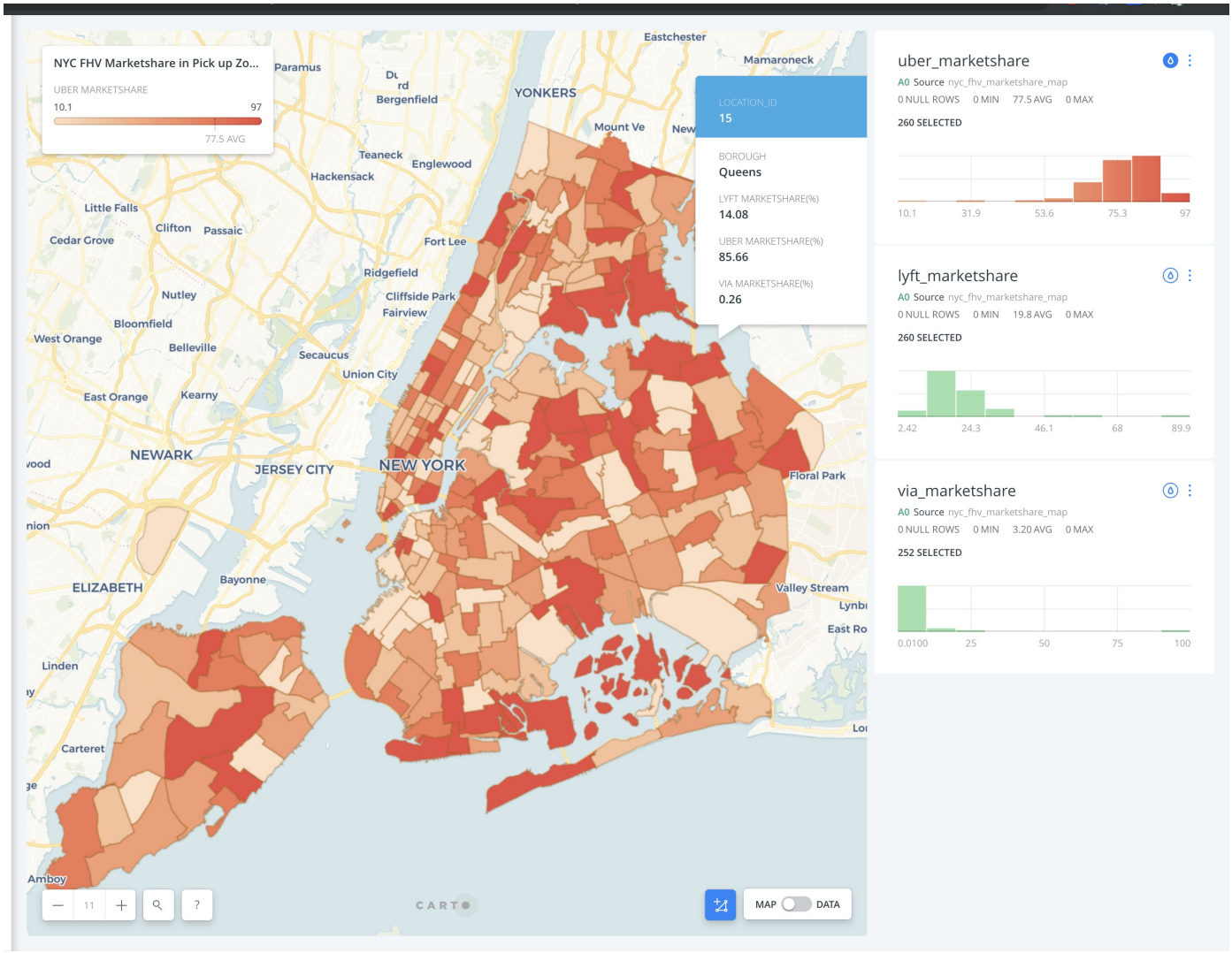


We find:

- For typical 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's median duration is the highest because most trips are share riders, which takes longer time.
- Overall, Uber has lowest trip duration comparing other two!

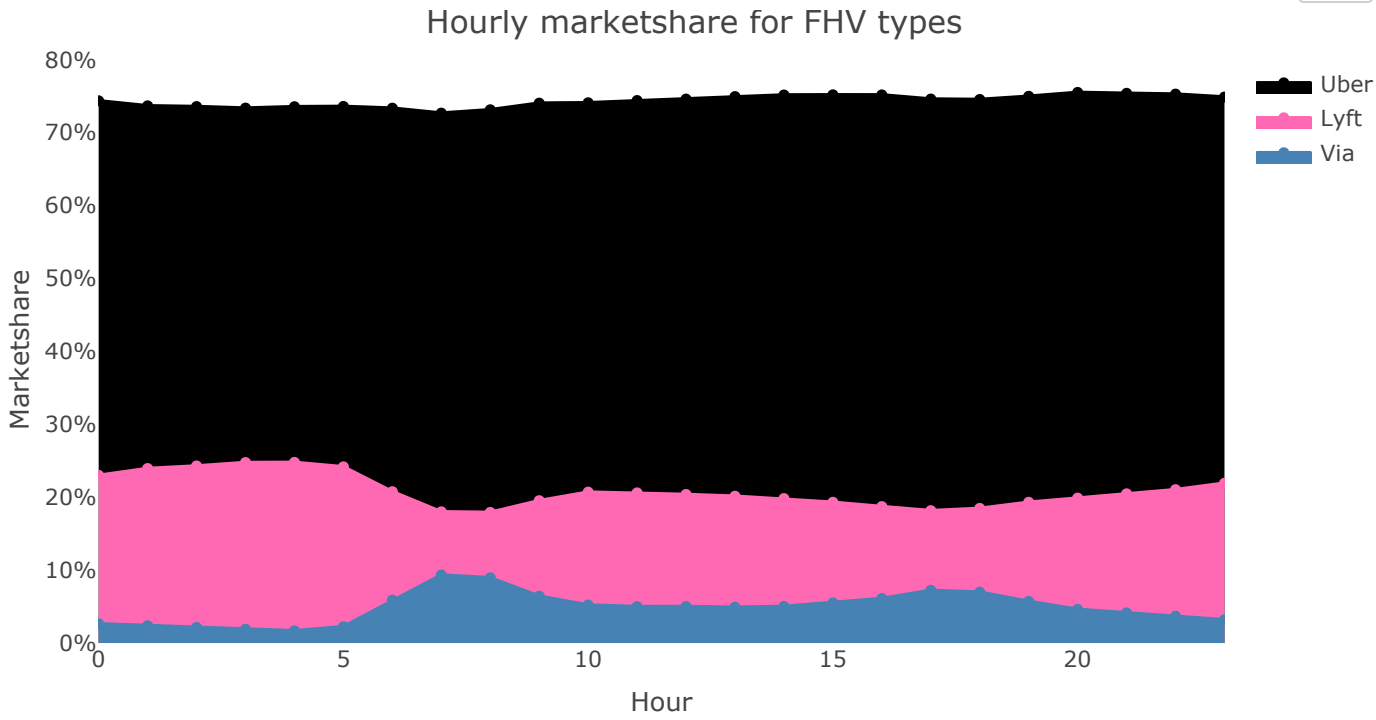
5.3 Market Share (Interactive component)

We also study the market shares on the both space and time line. We create an interactive NYC FHV Marketshare map (https://zxf71699.carto.com/builder/62d8c815-2839-41fe-95e0-84ac6e4eccb6/embed) 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.



NYC FHV Marketshare map

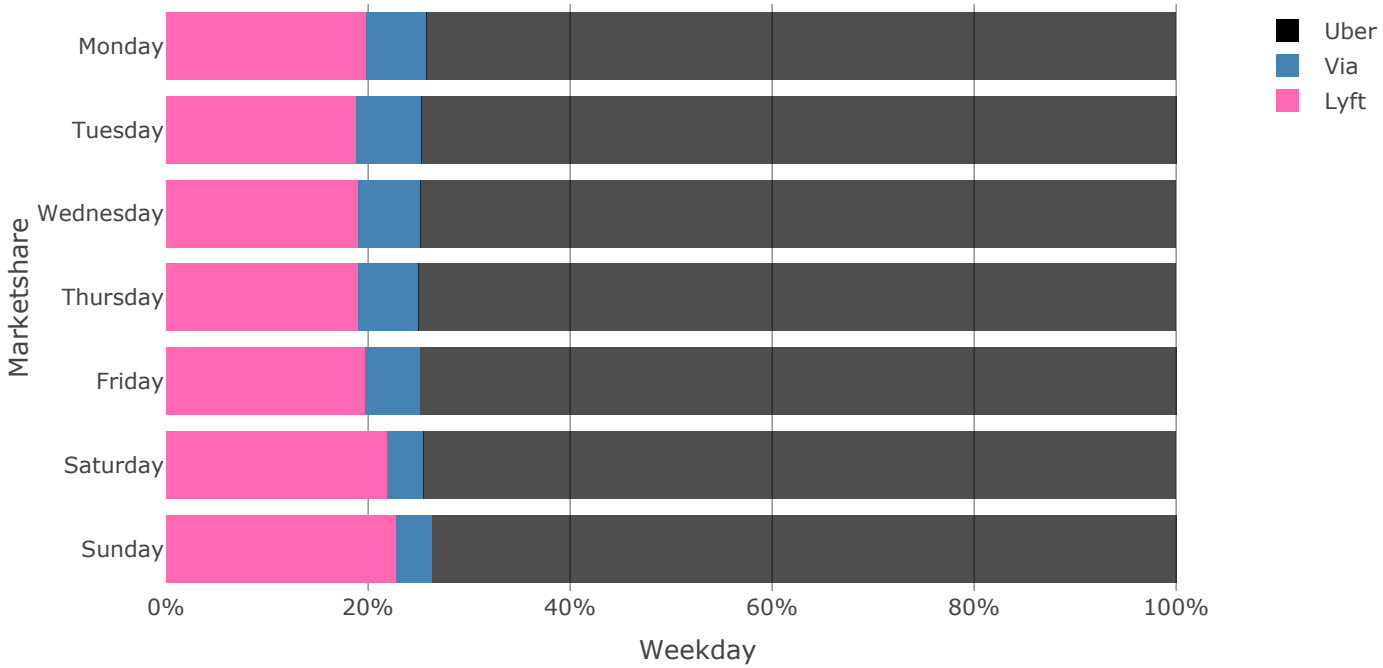
Code



We then research on hourly, weekly marketshare of Uber, Lyft and Via. We plot the following filled line plot and bar plot to observe whether there are some patterns or not.

Code

Weekly marketshare for FHV types



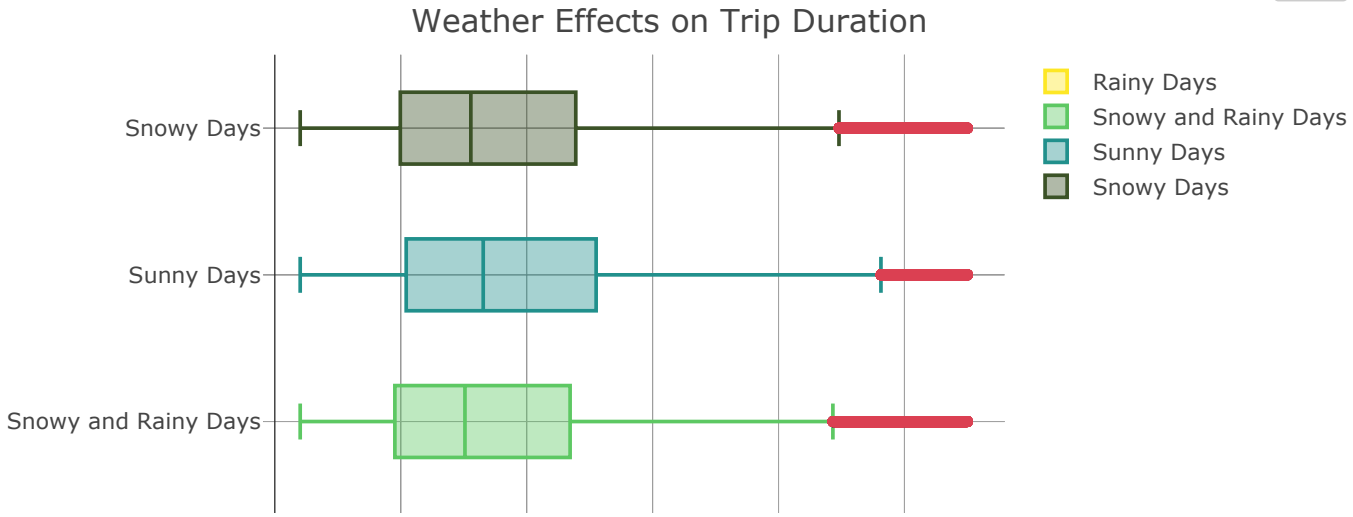
We find:

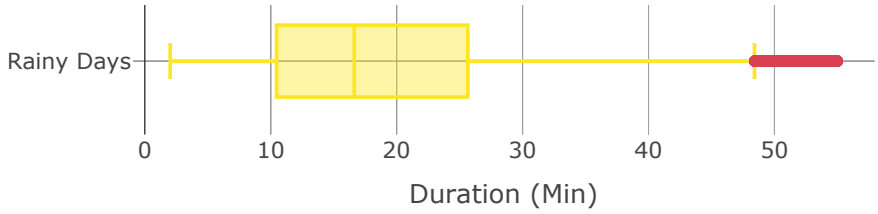
- Space Aspect
  - Service provided by Uber covers almost the whole area and it dominates the market.
  - Lyft makes up no more than 40% market among these four types in each pick-up zone. Service operates well mainly in midtown and downtown and some area of Brooklyn.
  - Via serve a smaller area in NYC and most zone only have no more than 10% Via trips. The interesting point is that Via makes up 100% marketshare in Newark Airport, but we think it is not realistic and maybe data of Uber and Lyft did not cover Newark Airport.
- Time Aspect
  - Uber is dominant on the FHV market, reaching 75% in entire area.
  - Lyft is second dominant on the FHV market, and weekends have higher numbers of trips.
  - Via is a growing company and it takes a small proportion of market, but it focuses on the peaking hour and weekday. It make sense because Via offers packages at weekdays for workers and students.

5.4 Weather Effect

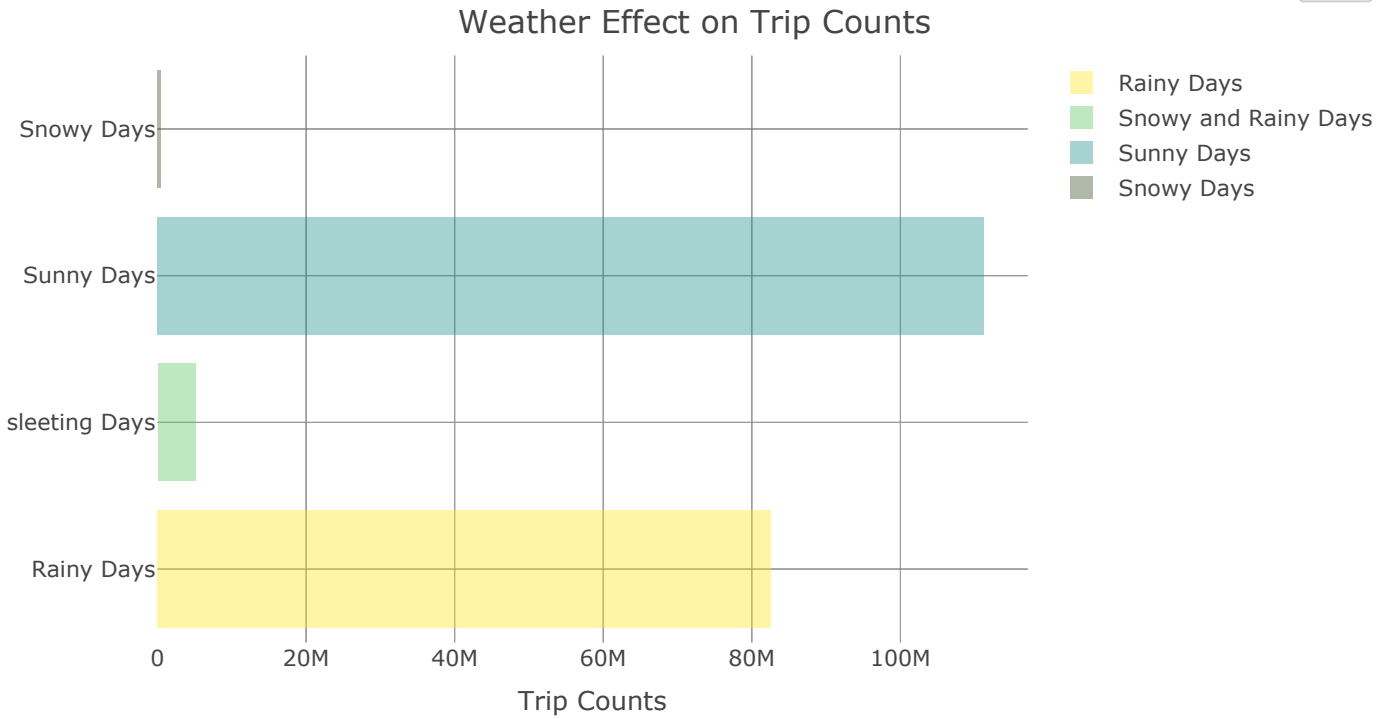
We have encouraged to supplement our analysis with combining the external NYC weather data to study how weather impacted on the trip duration. The particular interest here will be the rainny, snowny, sleeting(mixture of snow and rain) and sunny weather. We ploted box plot to observe the distribution of trip durations in these four weather conditions and we generated a bar plot for the number of trip requests for different weather respectively. Note that there are about 25k+ trips with durations larger than 60 miniutes which only made up 0.012% of the whole dataset, so we remove those records for better visualization purpose.

Code





Code



We find:

- For sunny days, there are the largest amount of trip requests. It tells most people prefer to hang 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 is more likely that snow would lead to shorter trips, so it could simply mean passengers were more likely to travel shorter distances, or stay at home.

VI. Conclusion

We conclude:

- FHV data has about 90k missing value. Since we have very large dataset, removing missing value will be better idea.
- The trip exists a lot of outliers (duration <= 2 mins and >= 4 hours), so we need to investigate on the specific direction and time consumption to identify the possiblity.
- The duration is heavy skew, so we require **log** transformation to normal distribution for future modeling.
- In general, trip durations fluctuate all day and long trips happen at rush hour which makes sense. Interesting point is that longer trips happen in Thursday, it may be worth further investigation though. We also found trips with longer duration occurs in May and Oct when a larger number of vistors come to the city.
- Overall, Via ususally have longer duration because they offer a large number of shared ride and Uber seems has lowest trip durations. But in weekly base, Lyft generally takes more time for trips except weekend.
- It is obvious that Uber still dominate the FHV market. Via offer services in limitted areas compared to others. But it is interesting that Lyft shares less market during rush hour and weekdays. We guess that lots of Lyft drivers may be part-time and they may be out of service during weekdays or rush hour.
- Although distribution of trip durations is quite similar for these four weathers. The interesting point is that median of trip durations in sunny days are slightly larger than others.

Future Thoughts:

- We can expend our data time frame such that collect and combine the latest three years data to study the pattern.
- It will be a good idea to estimate the distance betweeen each location block, so we can analyze the medican speed, which might be more intuition.
- The yellow and green cabs also have a big proportion of market share, and directly collected by goverment, so the data is more accurate and reflect to reality.
- Another concern is that rainy days are much less than sunny days, so we should normalize the trips by weather days, which might be more make sense.
- I found the paper said: the minimum and maximum temperatures show a strong correlation with each other because it also reflects to the the quality of weather and human behavior. We should include the temeratures to visualize the impact on trip duration.