NYC Taxi & FHV Project

@Author: Jie Li(jl5246), Xiaofan Zhang(xz2735), Hao Wu(hw2664)

Summary

This is a comprehensive statistical inference analysis for 300 millions of for-hire vehicles (Uber, Lyft, Via) trips originating in New York City from **2018-01-01** to **2018-12-31**. The goal of this challenge is to determine the major effect on the trip duration from **Upper West** to **Airports** (JFK, LGA) by time zone and weather factors. The results of the analysis show that rush hour (7-10AM & 4-7PM) during weekdays is the major effect on trip duration, taking 36 minutes to LGA, 56 minutes to JFK, which are **34%** longer trip time than traveling in the weekends.

Introduction

How long does it take to get to an NYC Airport?
What is the best time to leave?
How can I choose JFK and LGA airport?
How the traffic is impacted by the weather?

For answering those questions above, we apply exploratory analysis (EDA) and transformation on the raw data to determine the missing value, skewness, and time zone. Then, we construct the hypothesis test to study time zone, weather and interaction effects on the trip duration using one-way ANOVA, two-way ANOVA, and block design techniques. Lastly, we train the regression model to interpret the results.

Data Source

The raw data were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery. The For-Hire Vehicles (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 300+ million observations, and each row contains one trip information.

The NYC Taxi Zones map provided by TLC and published to NYC Open Data. This map shows the NYC taxi zones corresponding to the pickup zomes and drop off zones, or location IDs, included in the FHV trip records. The taxi zones are roughly based on the NYC Department of City Planning’s Neighborhood Tabulation Areas (NTAs) and are meant to approximate neighborhoods.

The NYC Weather data is provided by the 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.

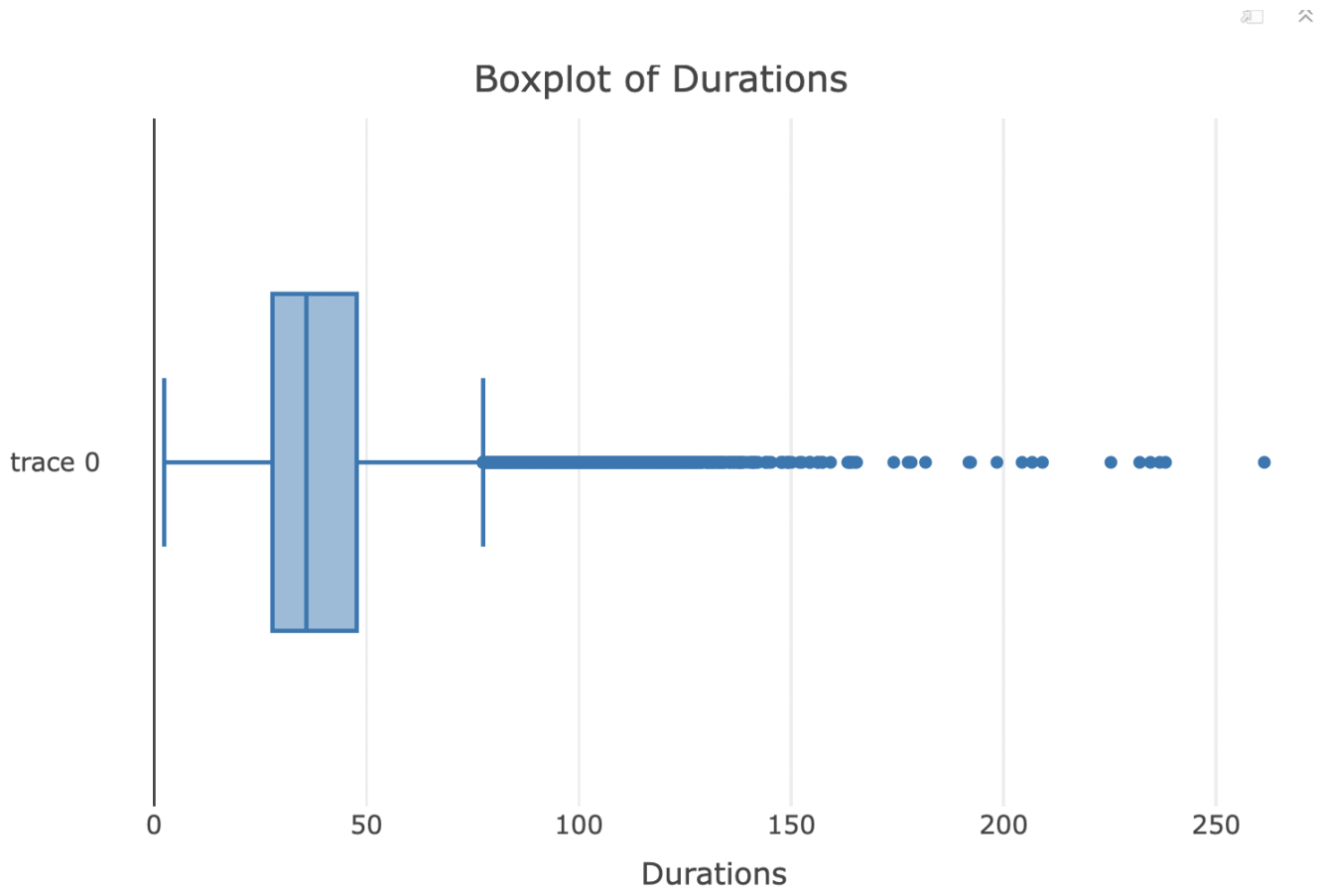
Raw data tatistics through January 1 to December 31, 2018:

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

Exploratory Data Analysis & Transformation

Data Process/Cleaning/Sampling

We are interested in how the trip duration to airports effects on Columbia Students, so we assume that most pick-up location will occur near the school because most students would like to live closely. The Drop-off location will be two airports (JFK, LGA).



Based on the google trip plan, we find the minimum duration requires 18 minutes to LGA, and 25 minutes to JFK. We create the boxplot of trip duration and find 2503 outliers that the trip duration large than 90 minutes, and 53 trip duration less than 15 minutes are

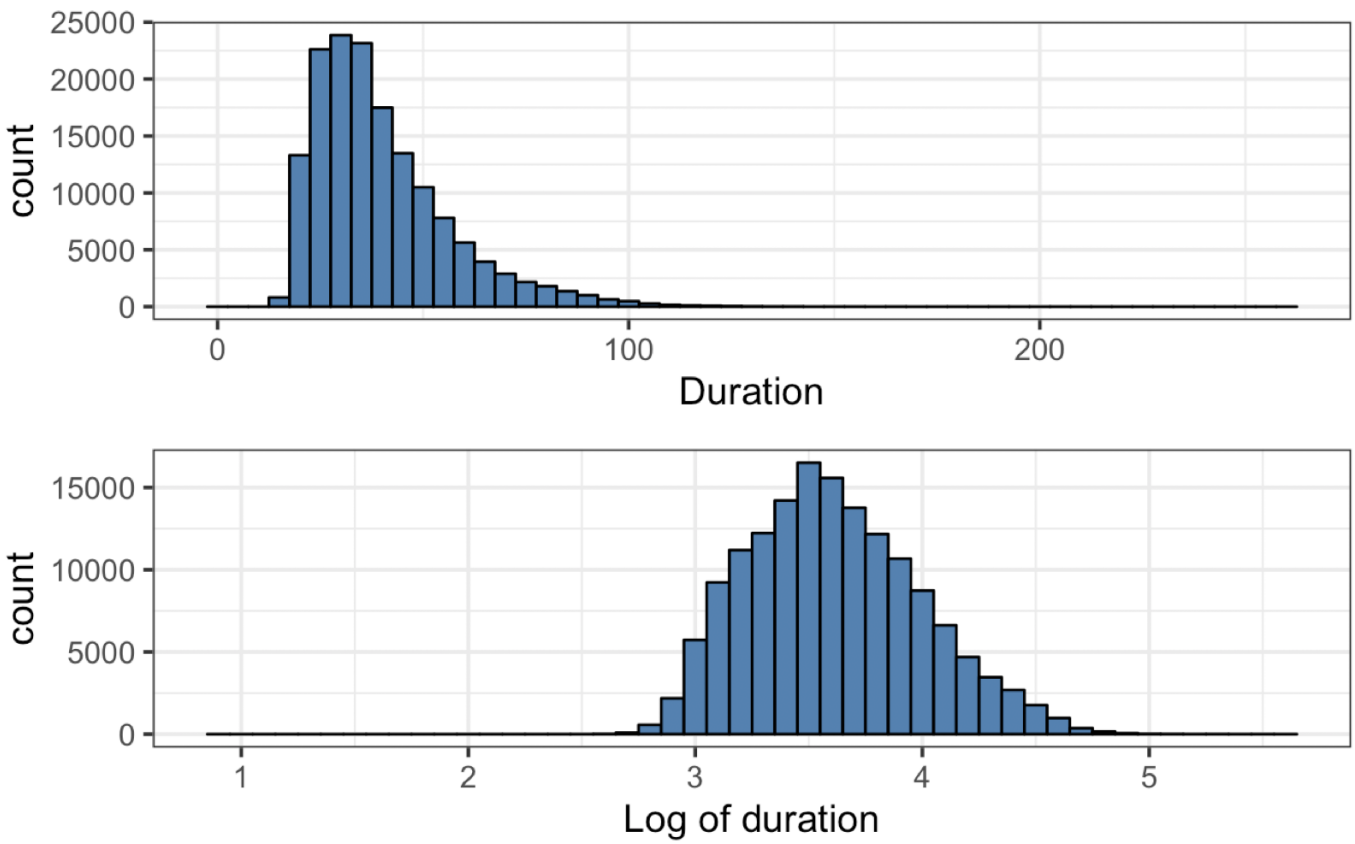
abnormal cases.

Due to large data size, we would like to remove those outliers, and also randomly sample 20 trips per day to reduce our data dimension, and perform a balanced test for the future.

Data Transformation

We plot a histogram of trip durations and find the distribution of trip duration is significantly right-skewed.

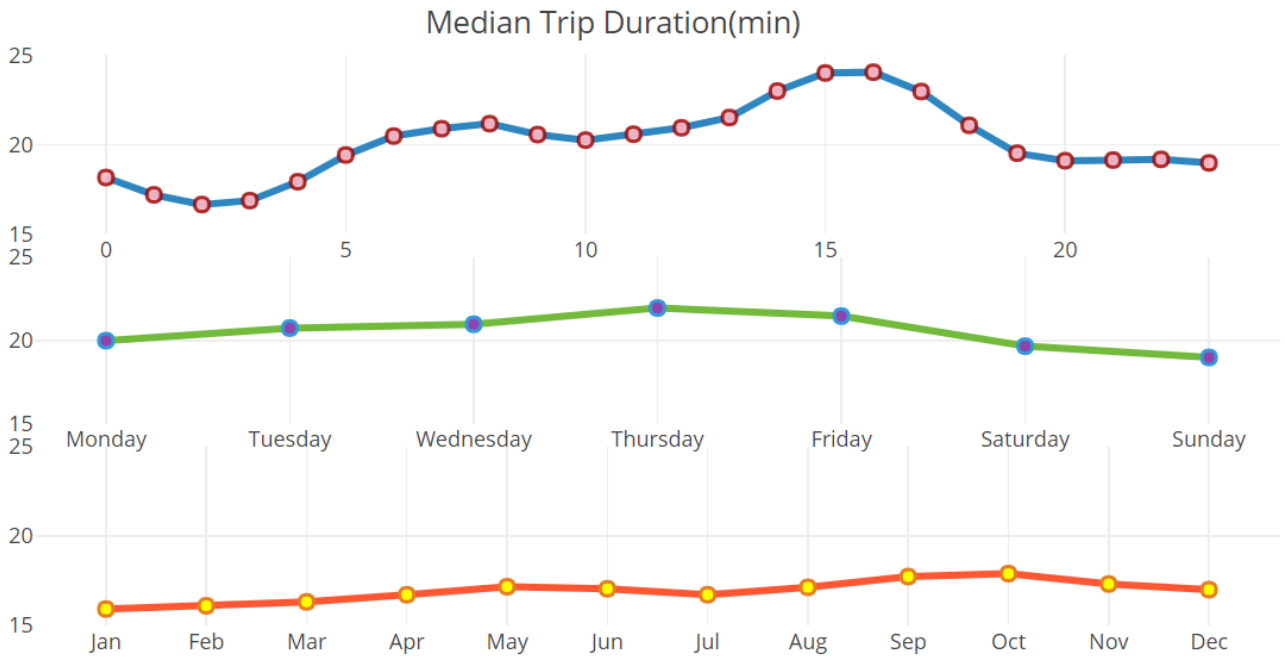
In order to solve the skewness issue to a normal distribution for future modeling and statistical inference, so we apply **Log** transformation on trip duration. Following histogram, it is showing a transformed duration is very close to normal distribution.



Duration over Time

We look at overall average trip duration based on hour, week and month bases. Since we have removed the outliers, the average is more intuitive for understanding. The hourly base is showing the rush hour effects in a typical day. The weekly base tells the difference

between workday and weekends. The monthly base has a good explanation of seasonality and weather trend.

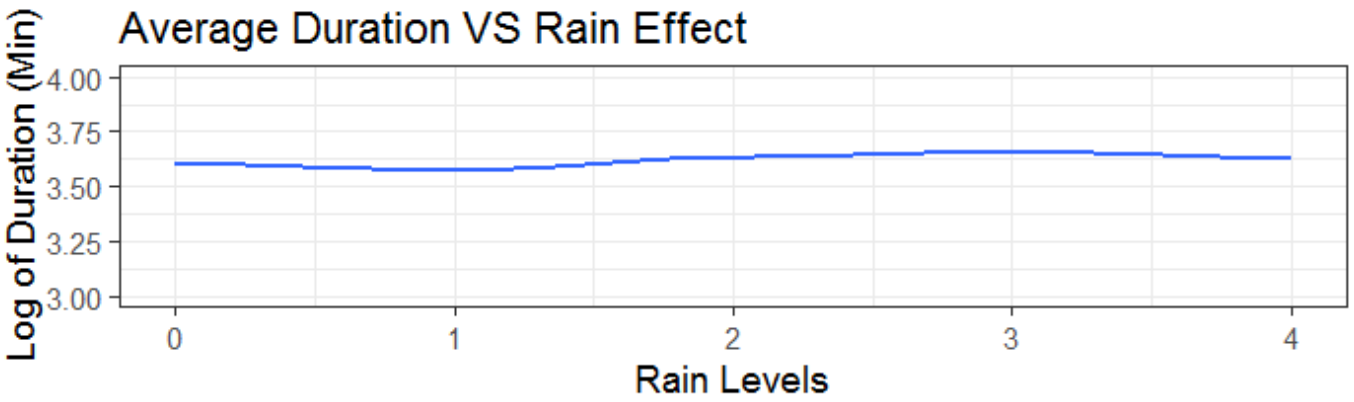


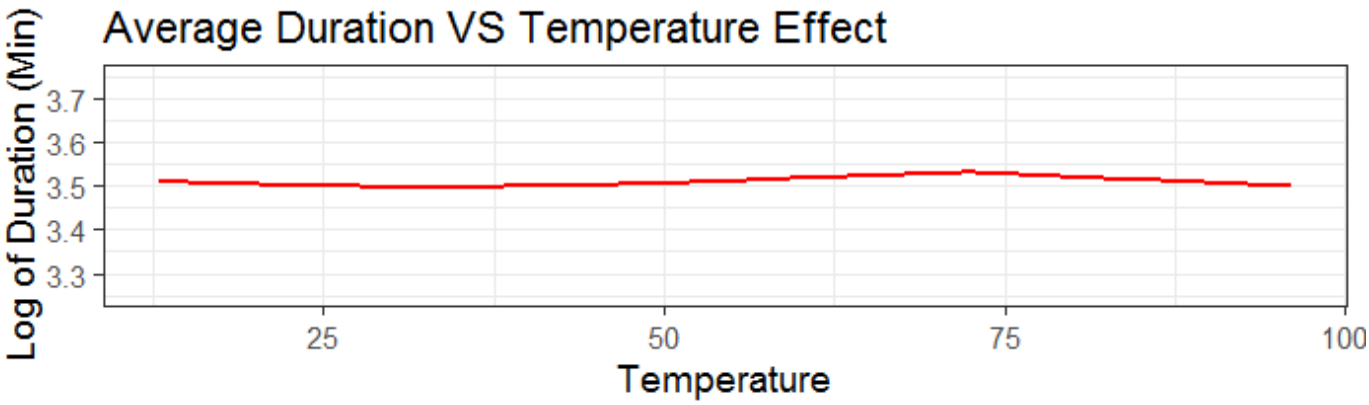
Weather Trend

We have encouraged to supplement our analysis by combining the external NYC weather data to study how the weather is impacted on the trip duration. Of particular interest here will be the rain and temperature statistics.

Rain factor has been decoded to four levels based on the amount of precipitation. For example, 0 means no rain, and 4 means violent. We find the rain levels might cause a longer trip duration.

The temperature might change human behavior. For example, most people prefer to stay at home when the temperature is very low as winter. On the other hand, more people would like to go outside with friends on warm days, which cause longer duration.





Analysis of Result

LGA and JFK Airport

First of all, we are interested in whether overall trip duration have significant difference to two airport or not. We set the null hypothesis: no difference of trip durations to the two airports, and use one-way ANOVA to test.

The ANOVA table is:

	Df	Sum Sq	Mean Sq	F value	P-value
Drop-off location	1	409.4	409.4	5887	< 2e-16
Residuals	6858	477.0	0.1		

The p-value of drop-off location is less than 0.05, so we have strong evidence to reject H_0. In conclusion, the time spending from Upper West to airport is significantly different. It is trivial to understand because the distance to LGA from Upper West is much shorter than JFK. Therefore, we would like to split the dataset by airports and analyze factors separately.

Time Effect

Secondly, we generally spilt 24 hours into several groups by hourly average durations.

For JFK and LGA, we have:

- Group 1: hour 23-6
- Group 2: hour 7-10, 16-19
- Group 3: hour 11-15, 20-22

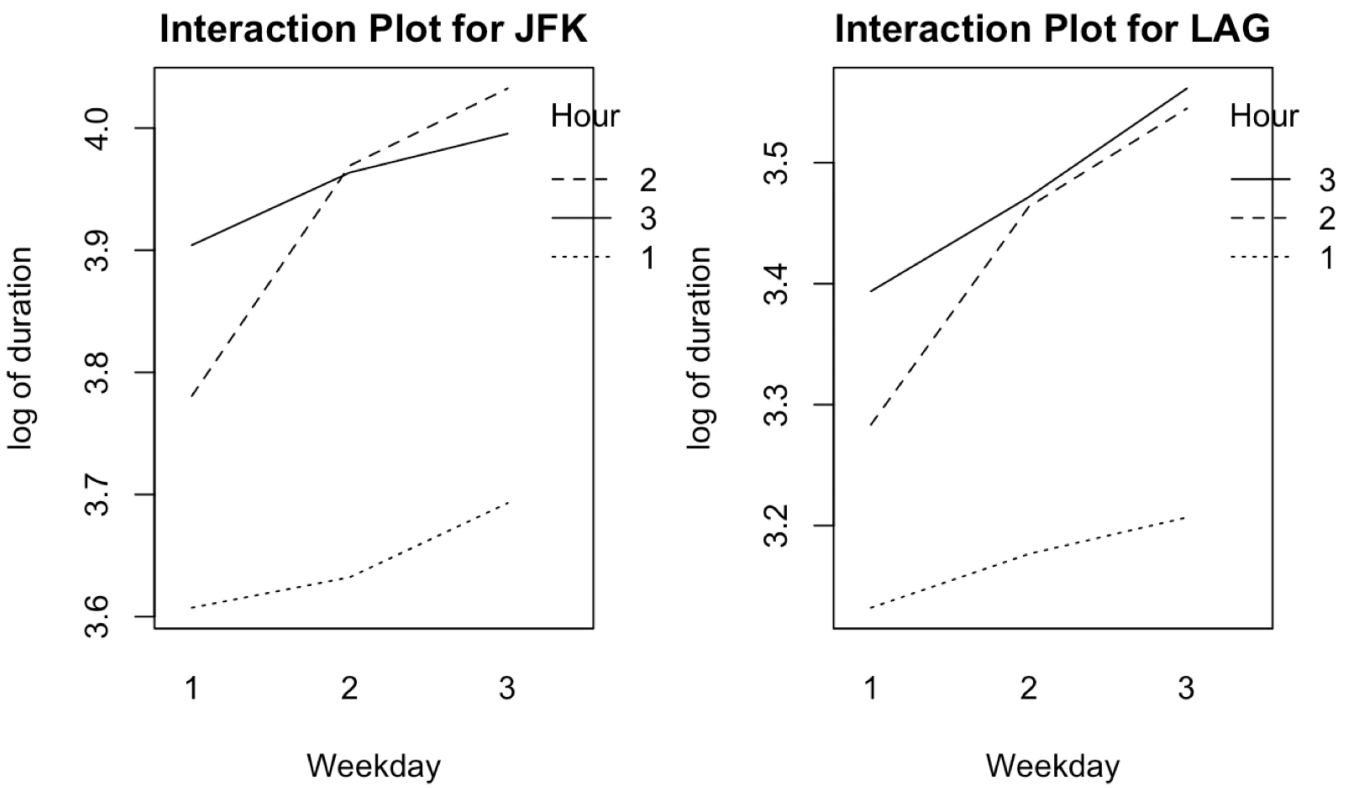
Then, we also divide windows for weeks by weekly average duration.

For JFK and LGA, we have:

- Group 1: Sat, Sun
- Group 2: Mon, Tue, Wed
- Group 3: Thu, Fri

For intuition, we realize that people may take longer time to commute by taxi at rush hour during weekdays, as there is heavy traffic in New York City. On the other hand, in the weekends, the trips would be longer at night when traffic speed is slow as many people go out for party or other entertainment activities. Thus, we are curious about whether there exists any interaction effect for weekdays and hours, so we plot interaction plot to observe that.

We plot interaction plot between weekdays and hours as following:



We observed that weekday and hour do have some interaction effect for both airports. In order to further study, we use F test and compute type III Anova.

The type III Anova table is:

- JFK

	Sum Sq	Df	F value	P-value
(Intercept)	43280	1	917596.233	< 2.2e-16
time	52	2	549.239	< 2.2e-16
weekday	9	2	98.712	< 2.2e-16
time:weekday	4	4	20.781	< 2.2e-16
Residuals	148	3134		

- LGA

	Sum Sq	Df	F value	P-value
(Intercept)	28056	1	795529.73	< 2.2e-16
time	57	2	593.92	< 2.2e-16
weekday	14	2	144.82	< 2.2e-16
time:weekday	3	4	18.16	8.77e-15
Residuals	177	3708		

The p-value of `time:weekday` for both airports are all less than significant level, 0.05, so we will reject H_0 and conclude that there exists interaction effect for both JFK and LGA.

Let's move to JFK airport. As we have detected interaction effect, so we do not need to test single effect of time(hour) and weekday. Now we will work on multiple comparisons for different combinations of weekdays and hours. We use Tukey test here. We may not be able to utilize LSD or Bonferroni test to make pairwise comparisons between variables since those tests are no longer robust for large combinations. The following table shows the comparison results.

Note: H_i stands for $group_i$ of the time variables and W_i stands for $groups_i$ of the weekday variable, for $i = 1, 2, 3$

combinations	Average log of duration	group index
$H_2 * W_3$	4.032461	1
$H_3 * W_3$	3.995427	2
$H_2 * W_2$	3.969308	3
$H_3 * W_2$	3.963610	3
$H_3 * W_1$	3.904146	4
$H_2 * W_1$	3.780664	5
$H_1 * W_3$	3.692944	6
$H_1 * W_2$	3.632236	7
$H_1 * W_1$	3.607276	7

Next, since there is also an interaction effect beween time(hour) and weekday for the LGA airports, then we will do multiple comprisons for each combination of weekdays and hours. We also use Tukey test here. The following table shows the comparison results.

combinations	Average log of duration	group index
$H_3 * W_3$	3.561377	1
$H_2 * W_3$	3.544810	1
$H_3 * W_2$	3.471657	2
$H_2 * W_2$	3.463655	2
$H_3 * W_1$	3.393732	3
$H_2 * W_1$	3.283494	4
$H_1 * W_3$	3.206756	5
$H_1 * W_2$	3.176610	6
$H_1 * W_1$	3.132125	7

From the two tables above, the average duration for rush hour on workdays is highly likely to be longer than the average duration after 11PM. In addition, it may take a longer time to airports during the workdays than weekends.

Weather effect

Third, in this part, we aim to detect if the weather have a significant effect on trip duration to the airport. Here, we will consider both weather types and temperature. We find there are only nine snowy days for 2018, so we decide to ignore the snow effect. We classify rain intensity as, light rain, moderate rain, heavy rain and violent rain by precipitation according to wikipedia (https://en.wikipedia.org/wiki/Rain).

We test the rain effect on overall trip duration, using block design. We treat drop-off airports as a block and different rain intensity as treatment.

The ANOVA table show as following:

	Df	Sum Sq	Mean Sq	F value	P-value
droplocation	1	0.4838	0.4838	121.504	0.000385
Rain	4	0.0075	0.0019	0.473	0.756973
Residuals	4	0.0159	0.0040		

The p-value of Rain is 0.757 and bigger than 0.05, so we have no strong evidence to reject H_0, and it concludes that the rain does not have a significant effect on trip duration, which is really surprising! We may require more rainy data to do deep analysis.

Train model

Lastly, based on the factor analysis above, we train the linear regression model as following, which use airport location, hour, weekday, interaction between hours and weekdays to predict the trip duration.

$$Duration = e^{3.13+0.48J+0.16T_2+0.28T_3+0.04W_2+0.08W_3+0.14C_1+0.03C_2+0.18C_3+0.05C_4}$$

Where

- J represents dummy variable for JFK airport
- C_1, C_2, C_3, and C_4 represent dummy variables for the combinations of two groups $H_2 * W_2$, $H_3 * W_2$, $H_2 * W_3$, and $H_3 * W_3$.

Variable	Intercept	J	T_2	T_3	W_2
Coefficient	3.13	0.48	0.16	0.28	0.04
p-value	< 2e-16	< 2e-16	< 2e-16	< 2e-16	0.00239

Variable	W_3	C_1	C_2	C_3	C_4
Coefficient	0.08	0.14	0.03	0.18	0.05
p-value	4.78e-09	< 2e-16	0.08051	< 2e-16	0.00728

From the table above, at $\alpha = 0.05$, all the p-values of the coefficients are close to 0. Therefore we can conclude all the coefficients are statistically significant.

Conclusion

We visualize and analyze 6,860 trip durations of for-hiring-vehicles (FHV) from Upper West to LGA and JFK airports to study the major effects. We applied ANOVA table, block design and multiple comparisons to test the significance of features. We found in general, it takes less time to commute to LGA airport than JFK. Additionally, we detect that time effect and interaction term do exist in trip durations. Surprisingly, both rainfall intensity and temperature do not have a significant effect on trip durations. By the findings above, we build a linear regression model to interpret the result.

In most cases, the worst hour traveling to airports are rush hour such as 7-10AM and 4-7PM in the weekdays. For example, **the expected trip time from Upper West heading JFK Airport takes 56 minutes, for LGA Airport takes 36 minutes**. 5% of trips during rush hour takes over 86 minutes to JFK, 54 minutes to LGA — good luck making your flights in that case!

If you left Upper West heading airports after 11PM, you would face an expected trip time of 37 minutes to JFK, 23 minutes to LGA, with a 90% chance of getting there in less than 50 minutes and 30 minutes. Furthermore, **traveling to airports after 11PM at the weekends will be 34% shorter trip time than weekdays**. Plan your schedule, save your time!