Geospatial Analysis of Yellow Taxi and High Volume For-hire Services And The Impacts On The Environment

Yuxin Yan
Student ID: 1067815

August 16, 2021

1 Introduction and Data selection

Air pollution is a serious environmental issue in NYC (New York city) which gas emissions from taxi and HVFHS (high volume for-hire services) could not be ignored. This report aims to explore potential factors leading to more PM2.5 emissions, as well as how COVID-19 pandemic influenced PM2.5 levels and trips. The purpose is to give advice to the city government on how to improve subway system and regulate limousine companies in order to reduce PM2.5 levels.
All the taxi datasets were obtained from the Taxi & Limousine Commission City of New York government website [1]. The time period chosen were March to April, 2019 which involves the peak of outbreak in NYC . For the comparison purpose, data from the same months in 2020 was also selected.

1.1 Yellow Taxi Data

Yellow Taxi datasets were chosen because they include information of trip distance which potential influencing factor average speed could be calculated. Also, yellow taxi could go into the most air polluted area Manhattan. There were 18 million instances and 18 attributes before preprocessing.

1.2 High Volume For-Hire Services Data

Ride-sharing companies have dominated the licensed vehicles market since 2017. However, taxi discharges less gas compared to ridesharing vehicles [2]. According to studies [3], New York's laws to mandate hybrids joining taxi fleet has helped to reduce air pollution in Manhattan. It is suggestive to explore HVFHS. The dataset had 63 million instances with 7 features before preprocessing.

1.3 PM2.5 Level Data

A research [4] indicates that the COVID-19 virus can be carried by particulate matter, so it is vital to study PM2.5 on this occasion. PM2.5 levels data was sourced from the United States Environmental Protection Agency website [5].
The specified data used were daily averages of PM2.5 density (in ug/m^3) from March to April 2019, 2020, from two monitoring stations: CCNY (The City College of New York) and IS143 (J.H.S. 143 Eleanor Roosevelt) located in high-traffic borough Manhattan. There were 119 instances for each site with 20 features in total.

1.4 Subway Entrances and Usage Data

One method to reduce PM2.5 emissions is building integrated public transport. According to a study [6], new subway openings could effectively lower PM2. 5 concentrations.
To give sound recommendations, 1868 locations of subway entrances were visualized. Data was sourced from New York Government website [7]. Daily subway usage frequency was crawled from MTA Turnstile Data websites [8] in the time period March and April, 2019. Frequency was obtained by counting the number of turnstile records which assumed to be the true ridership numbers.

2 Data Preprocessing

2.1 Cleaning

2.1.1 Yellow Taxi data cleaning

Remove noisy data:

- Filtered pick-up and drop-off time to be within each corresponding month before merging.
- Filtered $VendorID$ to 1 and 2 as clarified in the data dictionary, there were some instances with $VendorID$ 4.
- Filtered $Passenger\_count$ in the range $[1, 6]$. According to the liveries [9], the maximum number of passengers allowed in a yellow taxi is six. There were some instances with more than 6 passengers which were considered as faulty data as an assumption.
- Filtered $Trip\_distance$ to be greater than 0, there were some instances with 0 trip distance. These were considered as error made by taximeter.
- Removed trips which absolute values of fares were less than \$2.5 which were invalid according to TCL standard fares [10]. Negative fares were considered as prepaid fees rather than faulty data from the payment type.
- Filtered $PULocationID$ and $DOLocationID$ to the range [1, 263] based on the data user guide. There were some instances with out-of-range location ID.
- Filtered mta_tax to be 0.5 per trip according to NYC government [11]. There were some instances with tax not equal to 0.5.
- Removed instances where pick-up and drop-off times were the same. These instances were associated with fare amount which makes little sense. The assumption was made that the drivers forgot to open the meter.
- Converted pick-up and drop-off datetime to *datetime* objects.

Remove outliers:

- Removed outliers with trip distance larger than median + 3*IQR by calculating IQR as 3^{rd} quartile minus 1^{st} quartile. There were some instances with extreme trip distances from the boxplot [Figure 20]. This filtered long trips more than 3.8km which gas emissions could not be monitored from the pick-up location.

Remove irrelevant data:

- Filtered *paymenttype* to be 1, 2, 3 and $RateCodeID$ to be 1, 2, 5 as interested types.
- Removed irrelevant attributes.

2.1.2 HVFHV data cleaning

$PULocationID$ and $DOLocationID$ were filtered to the range [1, 263]. There was no missing data besides SR_Flag with Nan standing for non-shared rides. Irrelevant attributes were dropped. Pick-up and drop-off time were converted to *datetime* objects.

2.1.3 Extra data cleaning

There was little cleaning to be done for subway and PM2.5 datasets. Irrelevant attributes were removed. Date of measurement was converted to *datetime* objects. For the subway entrances data, Yes/No were replaced with boolean values.

2.2 Feature engineering

2.2.1 Taxi dataset feature engineering

For the yellow taxi dataset these were included:

- Added a trip duration column measures how long the trip took in minutes. Outliers larger than median+3*IQR were removed. 1.5*IQR was not chosen because there might exist some trips driven in a low speed with short distance but long trip time. In this case, its gas emissions could still influence the pick-up location's PM2.5 concentrations.
- Added a weekday/weekend indicator: true for weekdays, false for weekends.
- Added a peak time indicator. Rush hour were roughly considered to be 6:00-10:00 and 16:00-20:00.
- Added a speed column which measures of average speed (km/min) of a vehicle and outliers larger than median+10*IQR were removed. These trips had an average speed more than 0.97 km/min which were considered to be unusual.

For the HVFHS data, a trip duration measured in minutes column was calculated and outliers were removed using the same criteria as taxi data. For both datasets, a date column was added to aggregate daily data.

2.2.2 Aggregating data

Many aggregations of data were created for daily analysis. For both car types, daily trips count and an indicator of weekday were created. For yellow taxi, sum of trip duration ($hour$), sum of trip distances (km) and average speed (km/min) were created.

3 Analysis and Geospatial visualization

While beginning throughout exploration, several questions came up that might help to steer the direction of investigation. These questions were divided into two themes: "What factors related to taxi and limousine would influence the density of PM2.5 in NYC?" and "How did COVID-19 impact taxi industry and PM2.5 levels?".
To investigate potential factors influence PM2.5 emissions, yellow taxi, HVFHS and CCNY PM2.5 daily data were chosen to make correlation heat maps. The pick-up locations of taxi and HVFHS data were filtered to zones 152,116 and 42 where these zones are located near CCNY. The heat maps reveals that there exist positive correlations between trip distances, number of pick-ups, trip duration and PM2.5 concentrations, while average speed is negatively correlated with PM2.5 levels especially for 2020 [Figure 2].

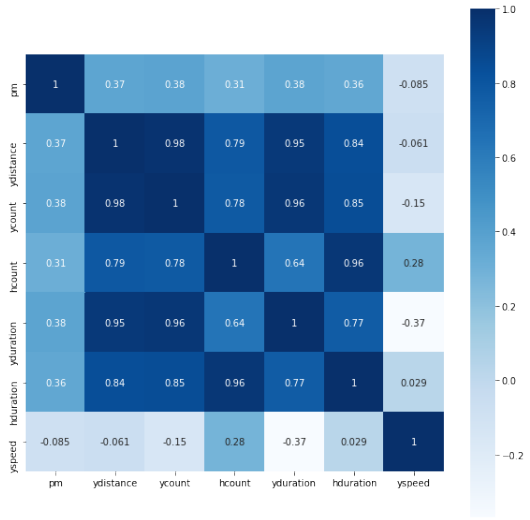


Figure 1: 2019 Taxi features and CCNY PM2.5 concentrations correlation heat map

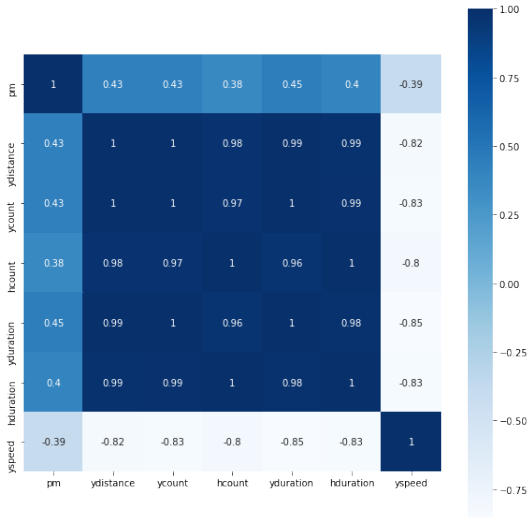


Figure 2: 2020 Taxi features and CCNY PM2.5 concentrations correlation heat map

3.1 Preliminary analysis

From the density distribution plots, PM2.5 concentrations from 2019 and 2020 follow approximate normal distributions with nearly identical standard deviations while 2020's data is more right skewed with a lower mean. It is inferred that lockdown has effectively reduced PM2.5 levels. This would be tested in the modelling part.

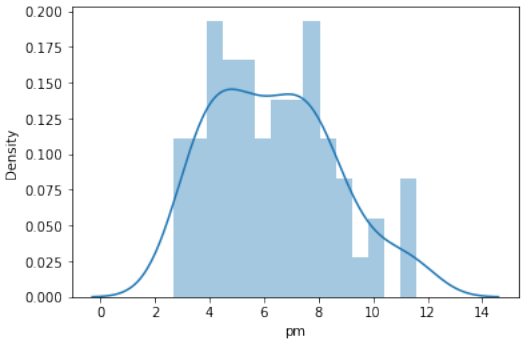


Figure 3: 2019 PM2.5 concentrations distribution

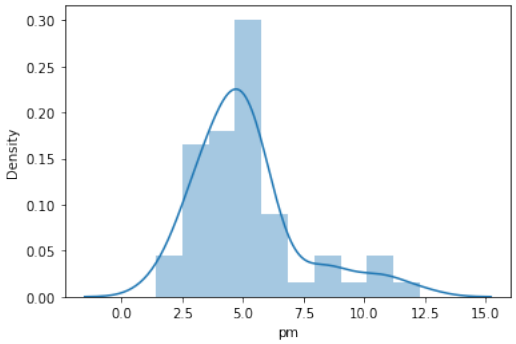


Figure 4: 2020 PM2.5 concentrations distribution

Other factors of trip frequency and speed were selected for geospatial analysis.

3.2 Geospatial Limousine demand analysis

In order to investigate zones where people have higher traffic demand, trips frequency was visualized. NYC subway entrances (grey dots) were added to analysis which zones might need more subway services.

3.2.1 Yellow taxi demand

Manhattan makes up most of the yellow taxi market [Figure 5], but it also enjoys a relative convenient subway system. Therefore, Manhattan were filtered in order to investigate other regions. It is clear

from the map [Figure 6] that there is a great travelling demand but no subway stations around LaGuardia Airport (LGA) and John F. Kennedy International Airport (JFK).

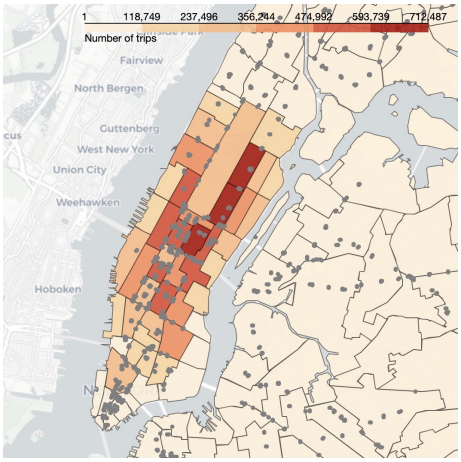


Figure 5: Yellow taxi trip count in Manhattan

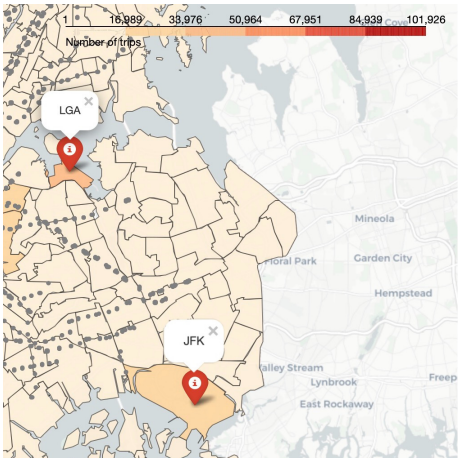


Figure 6: Yellow taxi trip count in Airports

3.2.2 HVFHS demand

HVFHS' market is more spread than that of yellow taxi. Zone 79 has the highest pick-up frequency but limited subway entrances around [Figure 7]. After some investigation [12] [Figure 8], there is a big club, some hotel villages and some cafes located in this zone where many social activities happen. Consequently, it has high traffic demand.

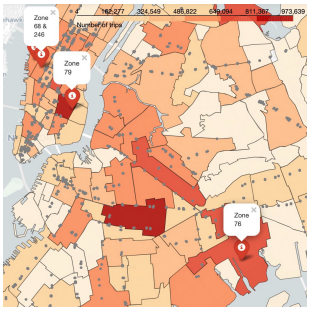


Figure 7: HVFHS demand map

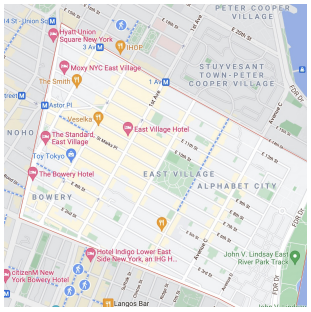


Figure 8: Zone79 Google Map

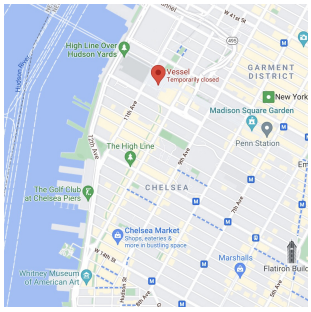


Figure 9: Zone68 Google Map

Zone 246 and zone 68 have relative high demand but very limited subway entrances [Figure 7]. Chelsea Market, Golf Club, Whitney Museum of American Art and Vessel are situated in these zones which would be attractive for both visitors and locals [Figure 9].
Apart from Manhattan, there are some zones with high transportation demand. Besides two airports mentioned, in Brooklyn, zone 76 is a living quarter where local people might have daily journey need [Figure 21].

3.3 Geospatial yellow taxi speed analysis

Although the correlation indicates a moderate negative relationship between speed and PM2.5 levels, it is still considered as an influential factor. According to a study, lower speed would result more PM2.5 emissions [13]. The maps [Figure 10, 11] suggest that Manhattan, Brooklyn and Bronx have

lower average ground speed. This might generate more PM2.5. Meanwhile, lockdown had increased average speed especially for high traffic regions.

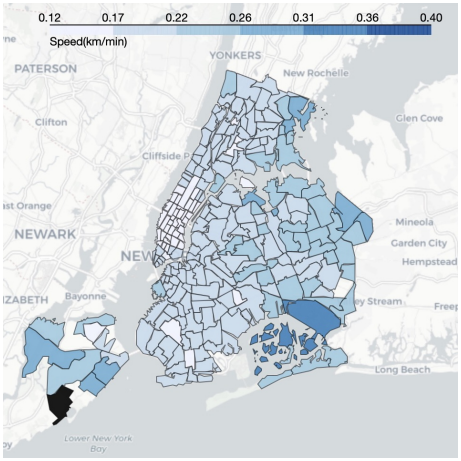


Figure 10: Median of daily average speed in 2019

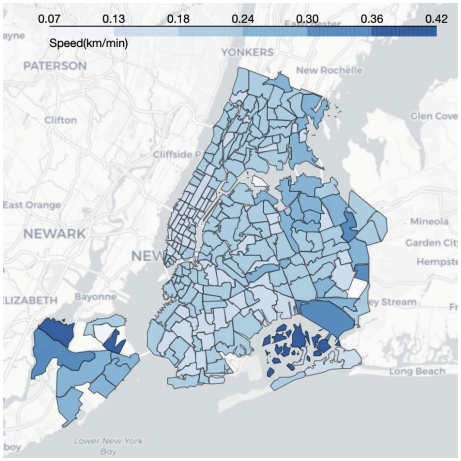


Figure 11: Median of daily average speed in 2020

3.4 Weekday subway and HVFHS demand analysis

Compared to HVFHS, Subway usage frequency remained stable throughout a week. HVFHS tended to experience demand peak around Saturday. It was inferred that commuters prefer subway as a cheap vehicle.

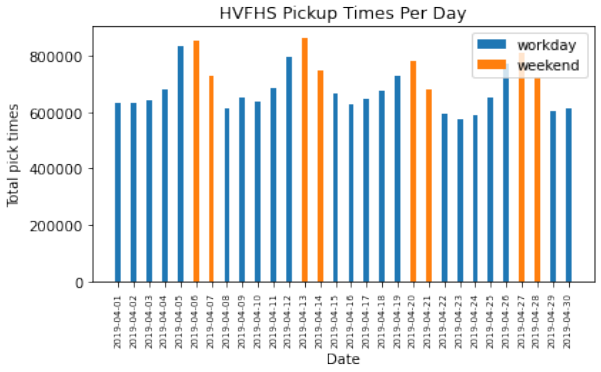


Figure 12: HVFHS daily pickup times in April 2019

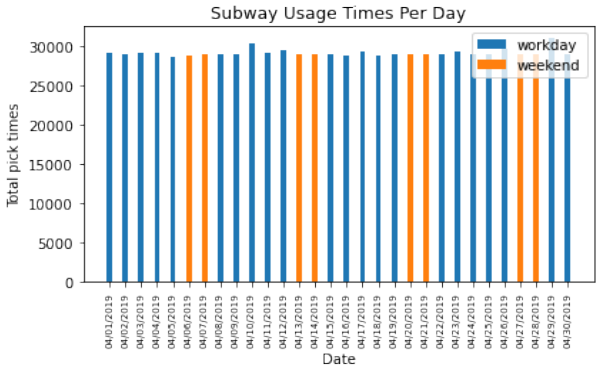


Figure 13: Subway daily usage times in April 2019

4 Statistical Modelling

To address hypothesis testing, linear regression (1) was chosen and assumptions were checked.

$$Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I) \tag{1}$$

The first hypothesis COVID-19 pandemic significantly reduced PM2.5 Levels in NYC was tested using ANCOVA. To achieve this, interaction model (2) and additive model (3) with continuous response variable of daily mean PM2.5 concentrations from monitoring sites CCNY, IS143 were analysed. The

predictors are categorical factor year and continuous variable time calculated as the number of days from the 1^{st} January. This could be interpreted as fitting two regression lines to each population.
The model assumptions were checked using eyeballing, F-test and diagnostic plots. PM2.5 density shows a downward trend along with time and data is independent. The p-value of F-test indicates two population having equal variances. QQ-plot [Figure 17] shows a slight right-skewed distribution for the response variable. Therefore, log transformation was applied to PM2.5 levels to make it follow normal distribution.
The graph [Figure 14] shows a moderate negative correlation between time and PM2.5 levels, and R^2 value indicates how well the regression lines fit the data. From the ANCOVA outputs [Figure 15, 16], there is no significant interaction under the significance level of 0.05, i.e., there is not much difference between the rate of decreases of PM2.5 concentrations for 2019 and 2020. This suggests during the outbreak, PM2.5 density decreased along with time in a similar rate as the same season in 2019 near CCNY. However, it is not rigorous to deduce that lockdown did not mitigate pollution in NYC.

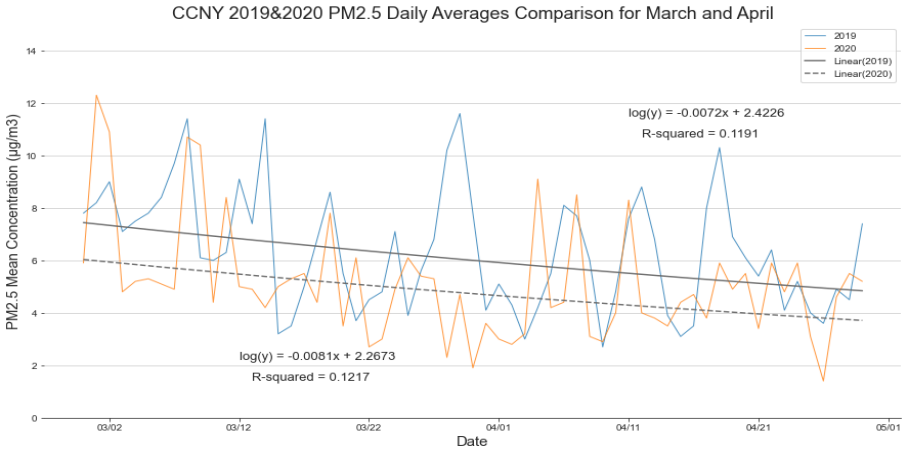


Figure 14: CCNY PM2.5 regression model after log transformation

The second finding was there are significant interactions between peak time and year, weekdays and year. Two factor models with (3) and without (4) interaction using response trip duration were compared. Data used was sampled yellow taxi records. The ANOVA results [Figure 18, 19] show that COVID-19 pandemic made differences of trip duration between weekdays and weekends, peak and non-peak hours longer compared to 2019. Thus trip duration was shorter in rush hours and weekdays in 2020.

5 Recommendations and discussion

To control PM2.5 emissions, actions could be taken from three perspectives for the government. While it was tested lockdown did not change the rate of the reduction of PM2.5 levels in CCNY, heavily polluted areas may not be the case. Controlling the trip frequency could have certainly reduced PM2.5 levels, less trip duration and faster speed also count. Therefore, public transport system should be improved to ease traffic congestion. Government could build subway stations near LGA and JFK. Zone 79, 68 , 246 and 76 should open more subway entrances. Also, weekend discount could be considered to attract more people to take subway in order to decrease for-hire services demand on weekends.
The second action to consider is to introduce more clean-energy vehicles. From previous analysis, high traffic and low speed are potential factors which result more PM2.5. Low average driving speed areas

Manhattan and Brooklyn have high taxi and HVFHS demand. Gasoline yellow taxi could be replaced by electric cars. Also, government can make regulation on the percentage of clean-energy vehicles for HVFHS companies.
Last but not least, enough air quality data is crucial in identifying air quality improvements. It is worth noting that some of NYC's most polluted areas like Midtown Manhattan don't have any monitoring stations. Without reliable data, it is hard to investigate whether air pollution mitigating actions actually work.

6 Conclusion and Limitations

Taxi trip frequency, trip duration and distance had positive relationships with PM2.5 emissions while driving speed was negatively correlated with PM2.5. To reduce pollution in a sustainable way, sound subway system could be established. It has been showed that LGA and JFK airports, zone 79, 246, 68 and 76 could be added more subway services.
Although PM2.5 levels declined over the outbreak at a substantial rate, it is not rigorous to attribute the change totally to COVID-19 pandemic, it could also ascribable to seasonal and weather change.
There were some limitations of the project. The available information for HVFHS is limited which trip distance and speed can not be analysed. Also, to be more rigorous, trip attributes like speed and frequency should be averaged inside each zone instead of selecting records where pick up location is close to monitoring site.

7 Appendix

7.1 Model formula

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \xi x_{ij} + \sigma_{ij} \tag{2}$$

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \sigma_{ij} \tag{3}$$

$$y_{ijk} = \mu + \tau_i + \beta_j + \xi_{ij} + \sigma_{ijk} \tag{4}$$

$$y_{ijk} = \mu + \tau_i + \beta_j + \sigma_{ijk} \tag{5}$$

7.2 ANOVA/ANCOVA results

```
Analysis of Variance Table

Model 1: log(pm) ~ time + year
Model 2: log(pm) ~ (time + year)^2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    119 16.156
2    118 16.148  1 0.0082988 0.0606 0.8059
```

Figure 15: ANCOVA for CCNY PM2.5

```
Analysis of Variance Table

Model 1: log(pm) ~ time + year
Model 2: log(pm) ~ (time + year)^2
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    113 21.546
2    112 20.833  1   0.71234 3.8296 0.05285 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16: ANCOVA for IS143 PM2.5

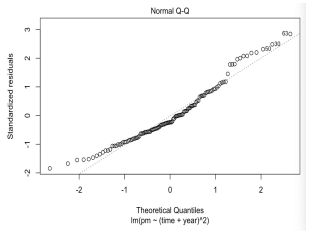


Figure 17: QQ plot for CCNY PM2.5 levels

```
Analysis of Variance Table

Model 1: trip_duration ~ PEAK + year
Model 2: trip_duration ~ PEAK * year
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1  5e+05 24974310
2  5e+05 24973592  1       718 14.375 0.0001498 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 18: ANOVA for interaction model between year and PEAK

```
Analysis of Variance Table

Model 1: trip_duration ~ WD + year
Model 2: trip_duration ~ WD * year
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1  5e+05 24890682
2  5e+05 24889326  1    1355.6 27.232 1.805e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 19: ANOVA for interaction model between year and WD

7.3 Related graphs

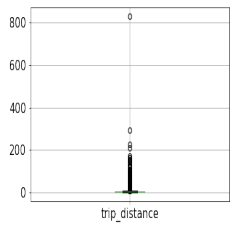


Figure 20: Trip distance outliers

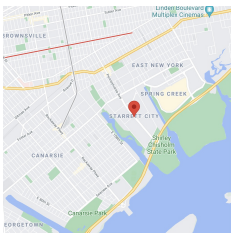
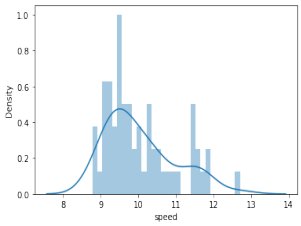


Figure 21: Google Map



Zone76 Figure 22: 2019 Average speed distribution

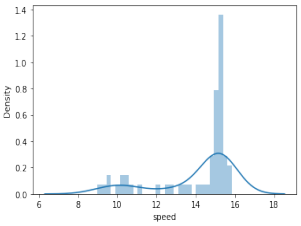


Figure 23: 2020 Average speed distribution

References

[1] Www1.nyc.gov. 2021. *TLC Trip Record Data - TLC.* [online]
Available at: <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>.

[2] Funes, Y., 2019. *Loophole Allows NYC Uber and Lyft Cars to Pollute More Than Yellow Cabs.* [online] Gizmodo.
Available at: <https://gizmodo.com/loophole-allows-nyc-uber-and-lyft-cars-to-pollute-more-1835123271>.

[3] Fry, D., Kioumourtzoglou, M., Treat, C., Burke, K., Evans, D., Tabb, L., Carrion, D., Perera, F. and Lovasi, G., 2019.
Development and validation of a method to quantify benefits of clean-air taxi legislation. *Journal of Exposure Science & Environmental Epidemiology,* 30(4), pp.629-640.

[4] Martelletti, L. and Martelletti, P., 2020. Air Pollution and the Novel Covid-19 Disease: a Putative Disease Risk Factor. *SN Comprehensive Clinical Medicine,* 2(4), pp.383-387.

[5] US EPA. 2021. *Download Daily Data | US EPA.* [online]
Available at: <https://www.epa.gov/outdoor-air-quality-data/download-daily-data>[Accessed 3 August 2021].

[6] Lu, H., Zhu, Y., Qi, Y. and Yu, J., 2018. Do Urban Subway Openings Reduce PM2.5 Concentrations? Evidence from China. *Sustainability,* 10(11), p.4147.

[7] Data.ny.gov. 2021. [online]
Available at: <https://data.ny.gov/Transportation/NYC-Transit-Subway-Entrance-And-Exit-Data/i9wp-a4ja>
[Accessed 3 August 2021].

[8] Web.mta.info. 2021. *mta.info | Turnstile Data.* [online]
Available at: <http://web.mta.info/developers/turnstile.html> [Accessed 7 August 2021].

[9] 2016. *Drivers of Taxicabs and Street Hail Liveries.* [ebook]
Available at: <https://www1.nyc.gov/assets/tlc/downloads/pdf/rule_book_current_chapter_54.pdf>
[Accessed 1 August 2021].

[10] Www1.nyc.gov. 2021. *Taxi Fare − TLC.* [online]
Available at: <https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page> [Accessed 1 August 2021].

[11] Tax.ny.gov. 2021. *Information on the taxicab and hail vehicle trip tax.* [online]
Available at: <https://www.tax.ny.gov/bus/taxi/> [Accessed 1 August 2021].

[12] Google.com. 2021. *Before you continue to Google Maps.* [online]
Available at: <https://www.google.com/maps/22.38131,114.168639,11z> [Accessed 6 August 2021].

[13] William M, H. and William R, B., 2021. *Evaluating the Contribution of PM2.5 Precursor Gases and Re-entrained Road Emissions to Mobile Source PM2.5 Particulate Matter Emissions.* [online]
Available at: <https://www3.epa.gov/ttn/chief/conference/ei13/mobile/hodan.pdf> [Accessed 4 August 2021].