Uber Fares Dataset Analysis

Daniel Wang, Minzhe Guo, & Yuntao Shang

MUgSS Datathon, 3/11/23

Goal

What is the best strategy for uber drivers to take orders?

– At what time in a given week the fare rate/miles is the highest?

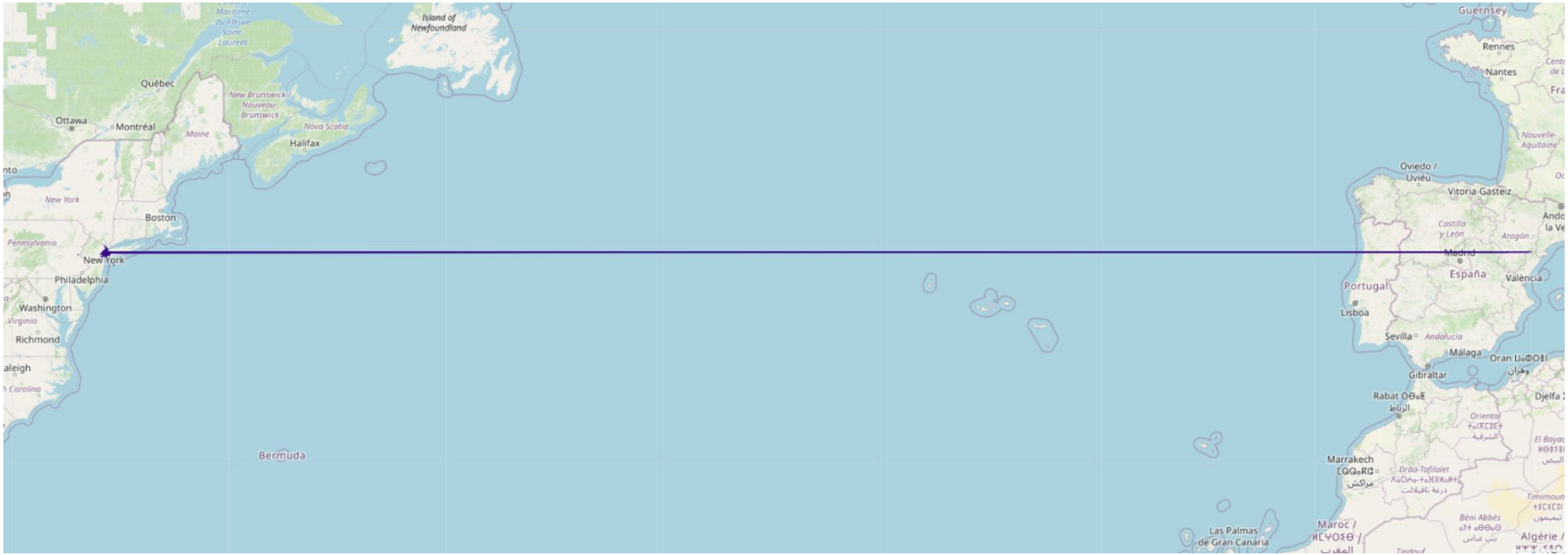
– At what locations they should pick-up the passengers?

Data Cleaning

- Remove geographical location that doesn't make sense.
  (Pickup & Drop-off latitude and longitude = 0)

- Translate & parse pickup-datetime into a more friendly format for later data manipulation
  (2015-05-07 19:52:06 UTC => Year: 2015, Month: 05, Weekday: 3, Hour: 19)

- Add "distance" column to the dataset based on each records' Pickup & Drop-off location
- Remove "distance" that doesn't make sense(with standard deviation)


Island of Newfoundland
Québec
Ottawa
Montréal
Maine
New Brunswick / Nouveau-Brunswick
Nova Scotia
Halifax
New York
Boston
Pennsylvania
New York
Philadelphia
Washington
Virginia
Richmond
Bermuda
Guernsey
Rennes
Nantes
Oviedo / Uviéu
Vitoria-Gasteiz
Castilla y León
Aragón
Madrid
España
Valencia
Portugal
Lisboa
Sevilla
Andalucía
Málaga
Gibraltar
Rabat
Marrakech
Las Palmas de Gran Canaria
Maroc /
Algérie /

Methodology

- EDA
- Map Projection
- Regression analysis

Regression Analysis

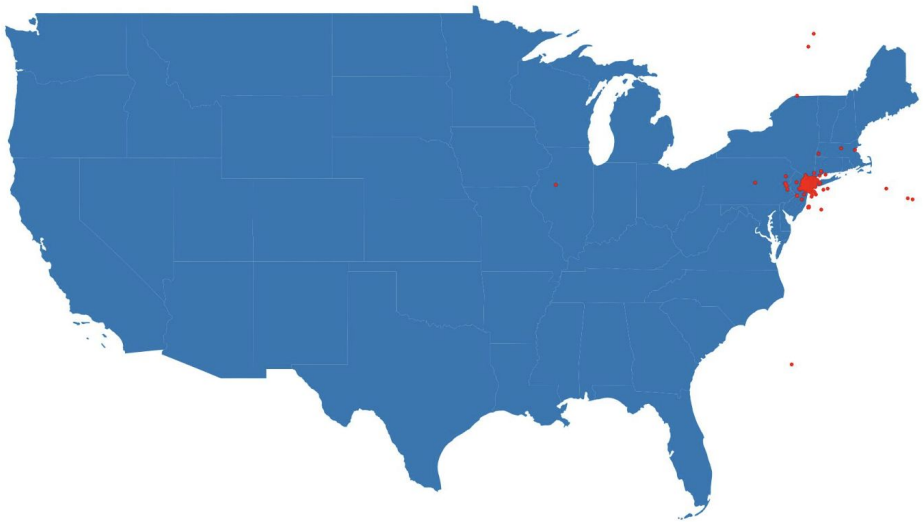
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.286e+03  2.395e+01 -53.690  < 2e-16 ***
pickup_longitude   1.297e-01  2.593e-02   5.000 5.73e-07 ***
pickup_latitude    1.164e-01  6.209e-02   1.874   0.0609 .
dropoff_longitude  3.534e-02  2.653e-02   1.332   0.1828
dropoff_latitude   2.694e-02  6.145e-02   0.438   0.6611
passenger_count    6.820e-02  1.583e-02   4.309 1.64e-05 ***
year               6.479e-01  1.189e-02  54.485  < 2e-16 ***
month              1.090e-01  6.426e-03  16.969  < 2e-16 ***
weekday            2.065e-02  1.133e-02   1.823   0.0683 .
hour              -3.124e-02  3.383e-03  -9.233  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.72 on 196029 degrees of freedom
Multiple R-squared:  0.01626,   Adjusted R-squared:  0.01621
F-statistic:   360 on 9 and 196029 DF,  p-value: < 2.2e-16
```

Response Variable: fare
Predictor Variables: pickup location, drop off location, number of passengers, year, month, weekday, and hour

Estimate slopes are small
R-squared: 0.01626

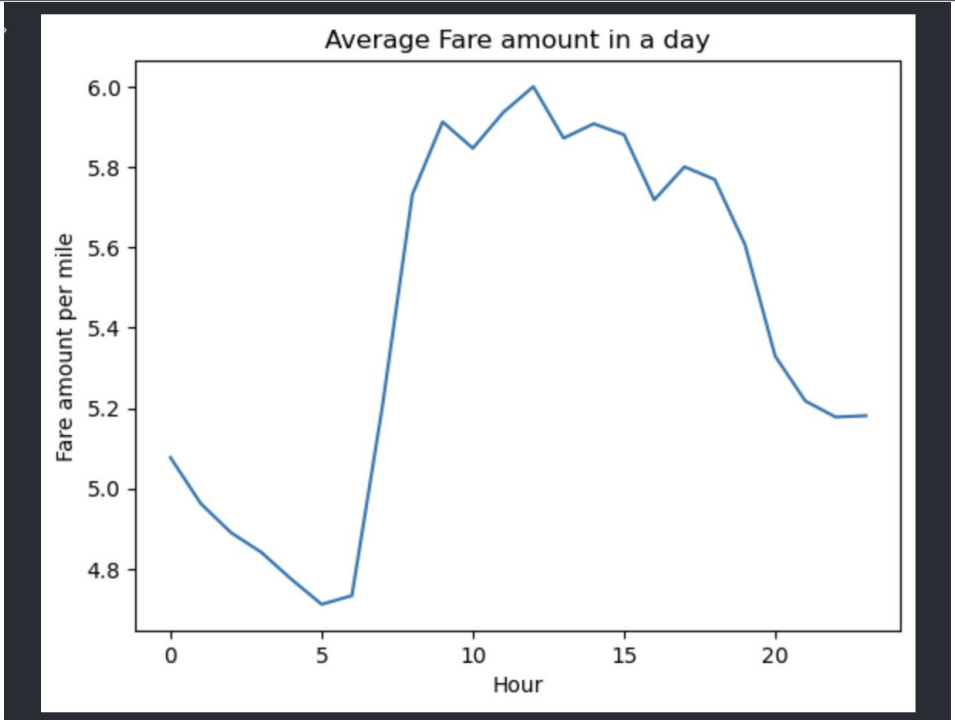
Regression Analysis

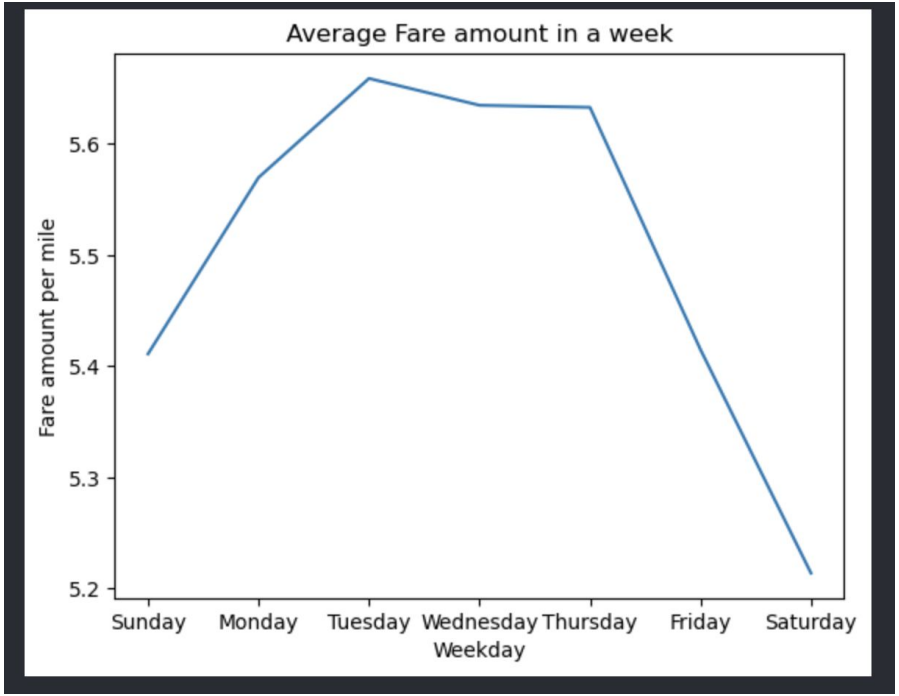


Understanding the Dataset:
1. Location in large scale (based on longitude and latitude) and the day of a week do not seem to influence the amount of fare.
2. There are evidence that hour of a day and number of passenger influence fare rate.
3. By plotting the data on a map, we find out that most orders in the data set located in NYC.

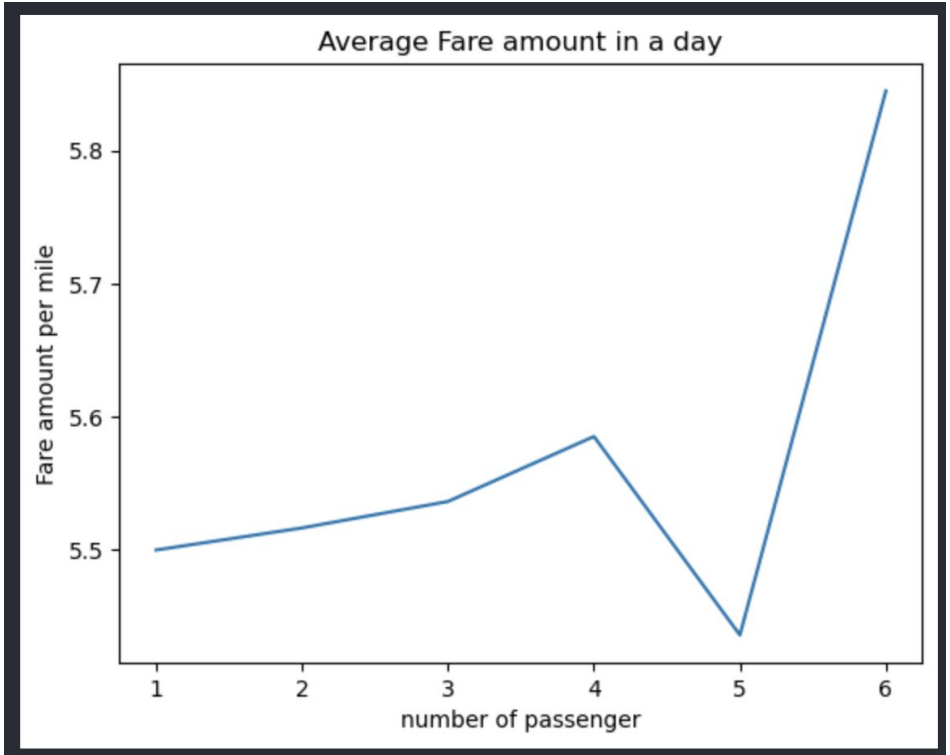
Based on these findings, we continued the following research.

Results and Findings

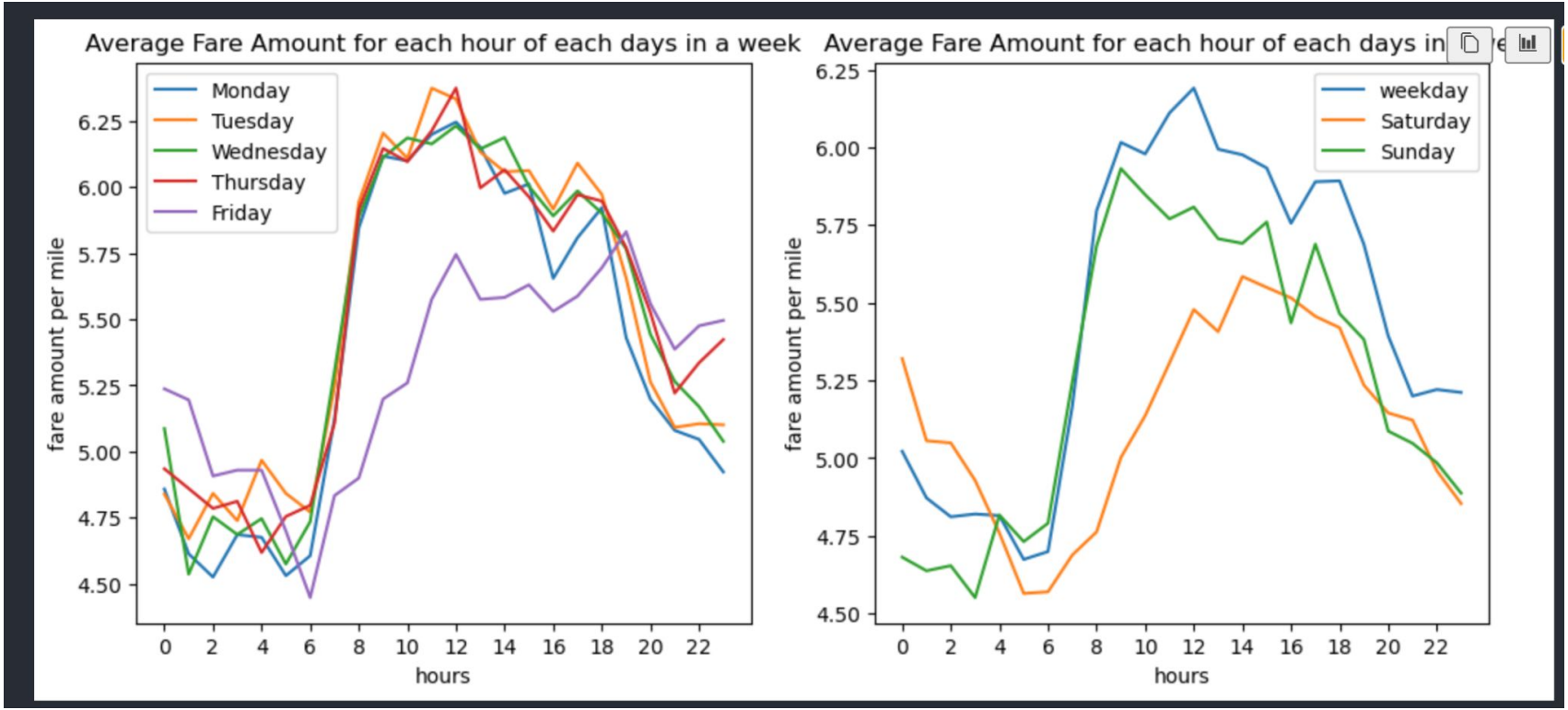




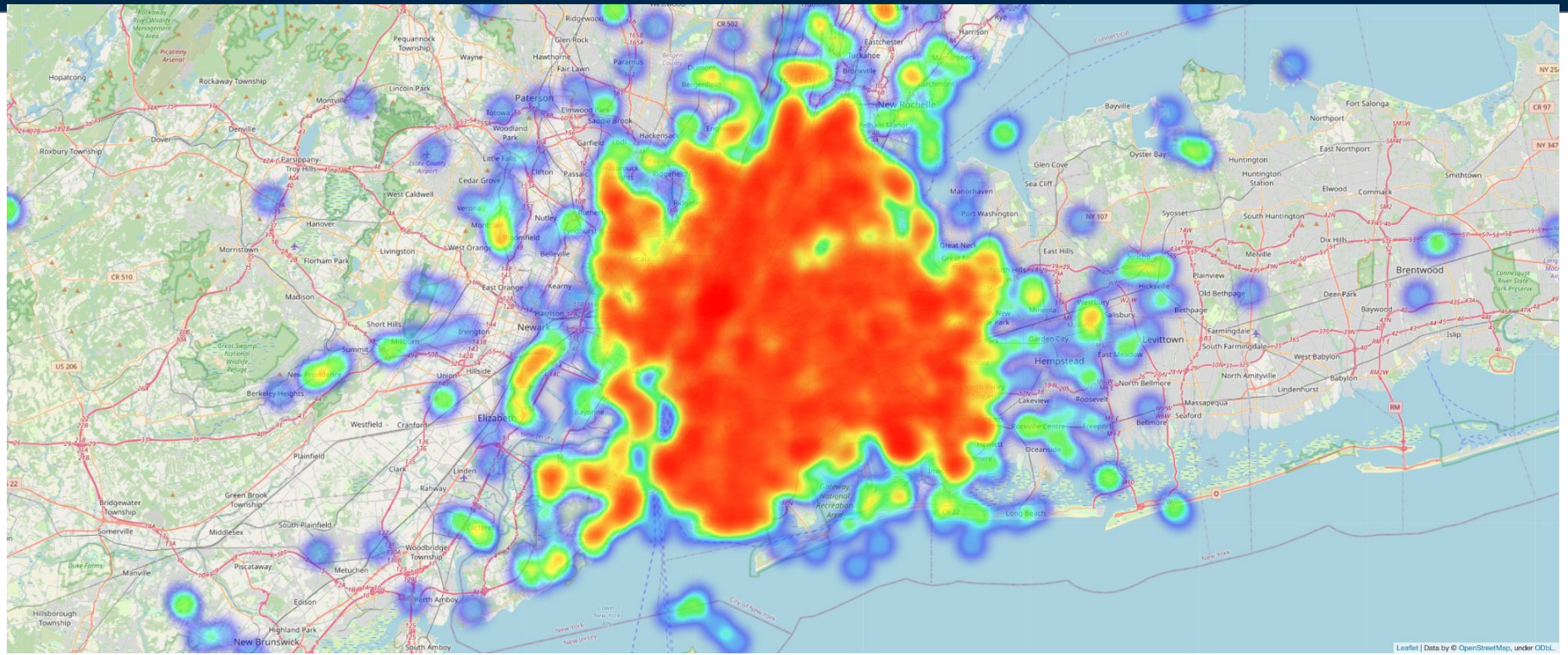
Results and Findings



Results and Findings

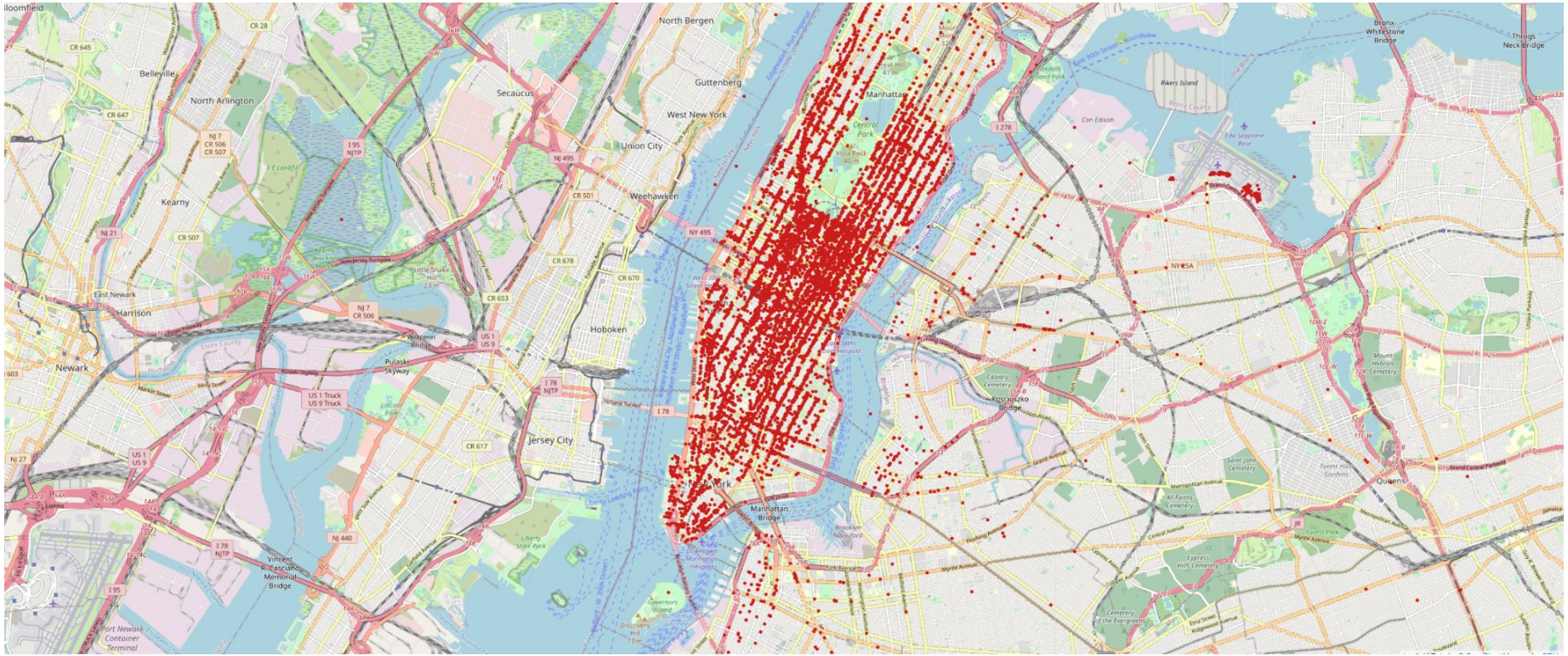


Results and Findings



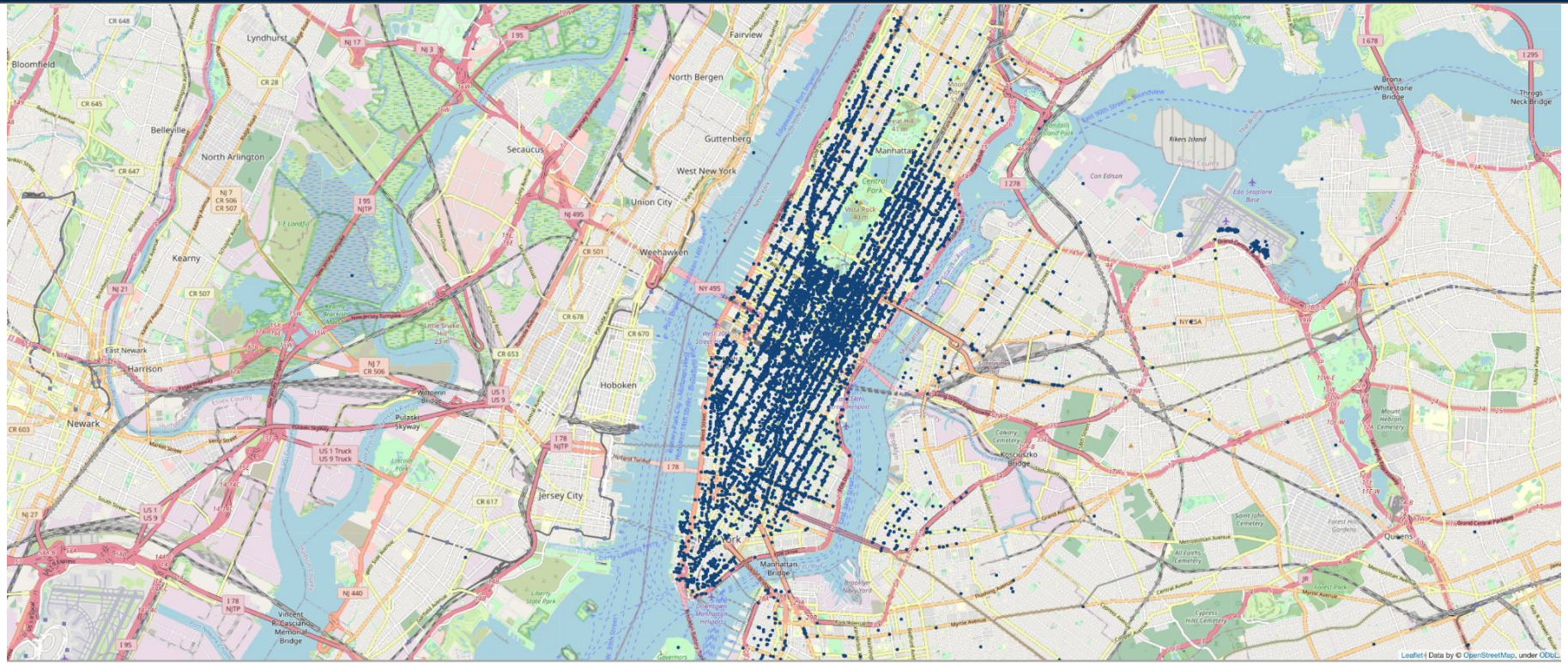
Results and Findings

Pickup Location



Results and Findings

Drop-off Location



Results and Findings

Conclusion

- Fares per mile in NYC is affected by day, time, and number of passengers
- Manhattan downtown, LaGuardia Airport, and John F. Kennedy International Airport are the three most rewarding places for the uber drivers
- Uber drivers in NYC should have different strategies between weekdays and weekends
    - For Weekdays, uber drivers should take as many orders as possible in 12 pm. The best time during weekday,from the data, is around Thursday 12 am.
    - For Saturday, they should take as many orders as possible between 12:00 am to 4 am