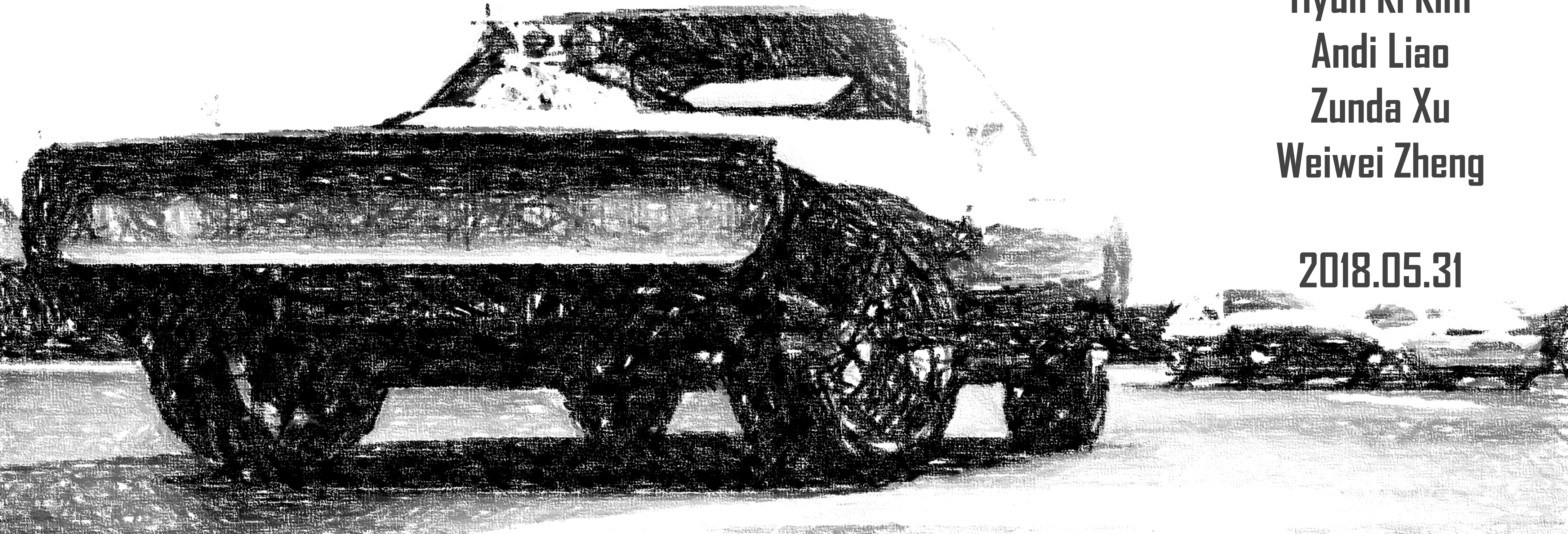
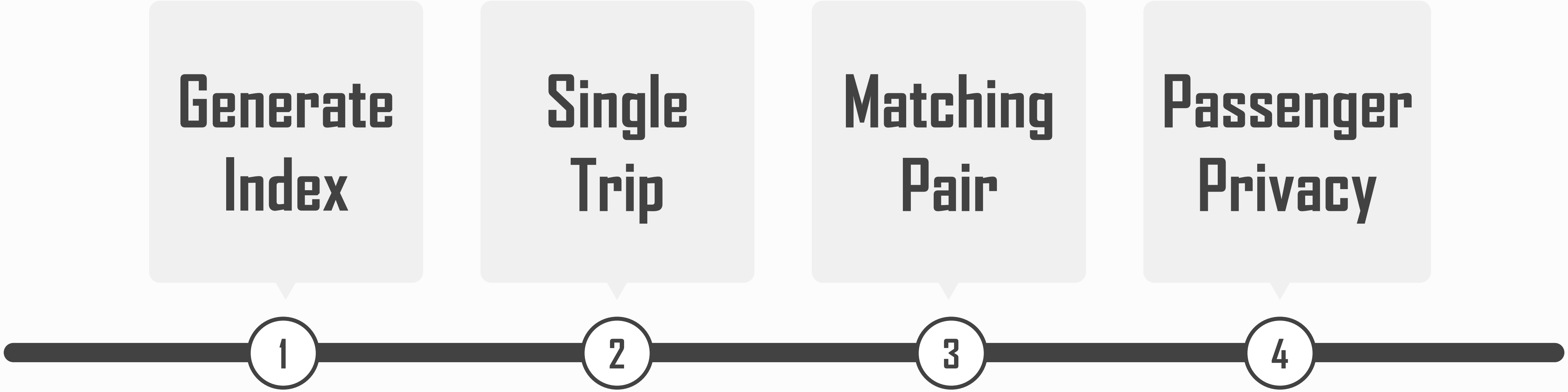
FAST & FURIOUS



Hyun Ki Kim
Andi Liao
Zunda Xu
Weiwei Zheng

2018.05.31

CONTENTS



1 Generate Index

2 Single Trip

3 Matching Pair

4 Passenger Privacy

Data Summary

Data size of one month

- 1.2 GB
- Over 10 million rows

Duration

- Single trip/Matching pair:
  - Focus on 2015/07 – 2016/06
- Passenger privacy:
  - 2009/01 – 2016/06, 217 GB in total

Generate Index

1

Index for travelling time / tip

- Why use index?
- Components:
  - Weather:
    - use wunderweather API + mapping
  - Location:
    - k-means clustering(k = 50 / 100)
  - Pick up Time:
    - month, weekday, hour

Single Trip

Predict travelling time / tip

- Multiple linear regression(MapReduce)
  - Find parameters of regression through Cholesky decomposition

Sample Results

- Location index and hour index statistically significant

3

Matching Pair – Prediction

Most Similar Trip

- Sum of squared difference (MPI)
- Single difference (MapReduce)

Results

- Mean squared error (MapReduce)
- Compare with dummy regressors (MapReduce)

Matching Pair – Causality

Most Similar Trip

- Fixed effect (MapReduce)
    - Control five indices
    - Sum of absolute difference

Results

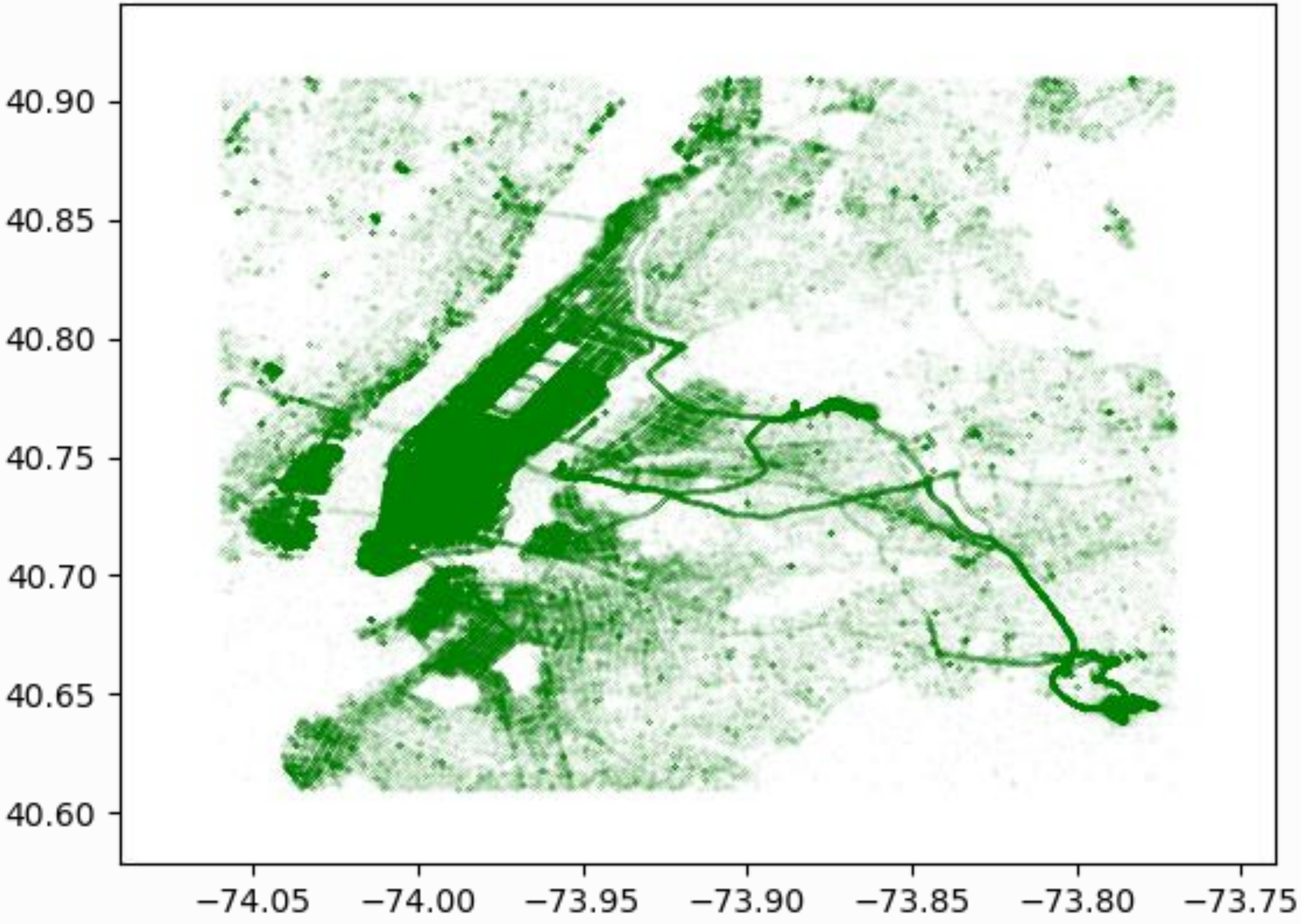
- Simple linear regressions (MapReduce)

Passenger Privacy

Motivation

- 2009/01 ~ 2016/06: latitude and longitude
- 2nd Annual NYC TLC Hackathon (2016/10)
- 2016/07 ~ 2017/12: area code

Passenger Privacy



Passenger Privacy

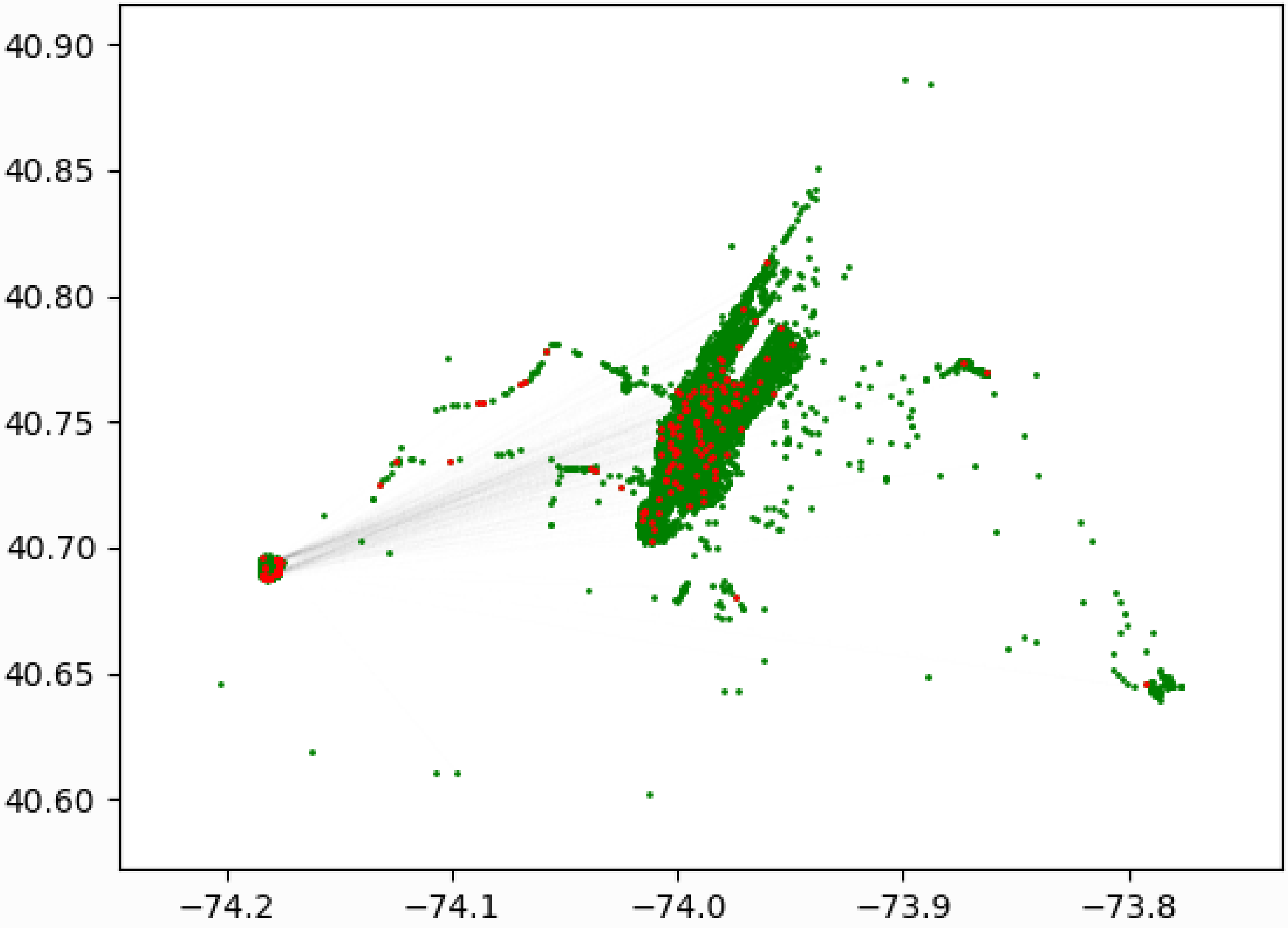
Tip_rate >= 50% and Tip_amount >= $ 15

Top Pickup	Top Dropoff
JFK	Newark
LGA	JFK
Penn Station	LGA
Grand Central Terminal	Penn Station

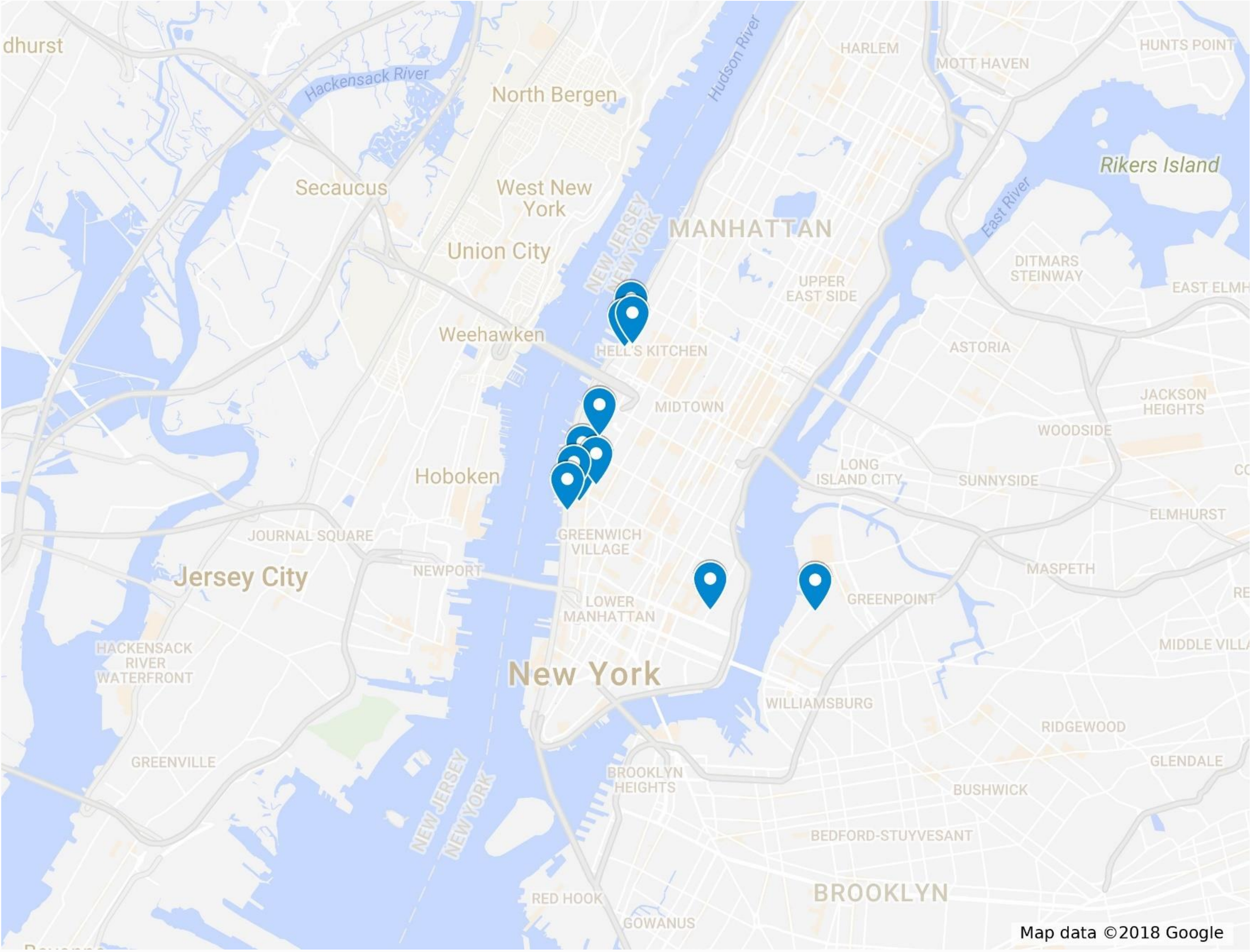
etc: Le Bain, Morgan Stanley, Google, Trump Tower, Plaza hotel…

4

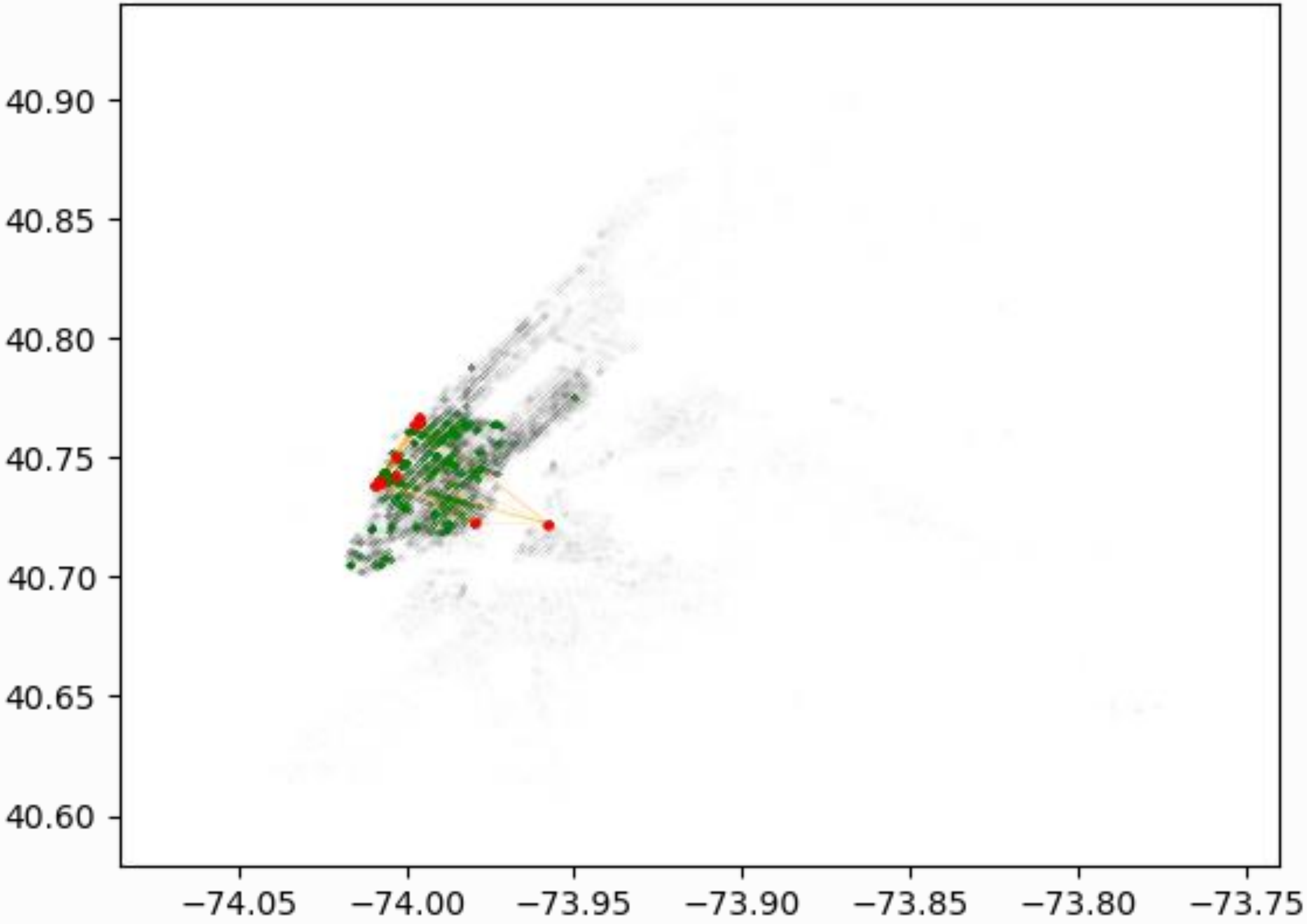
Passenger Privacy



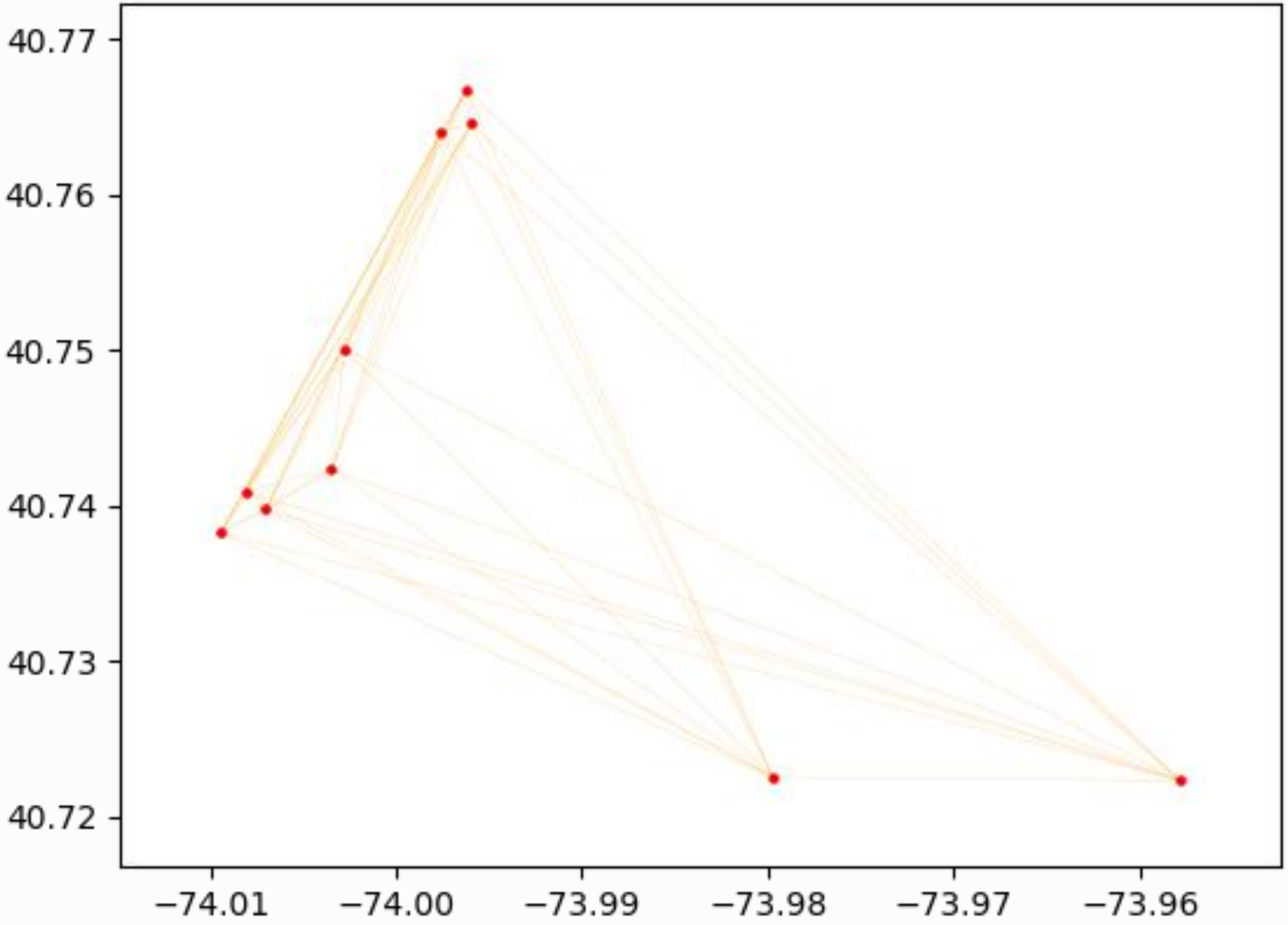
Passenger Privacy

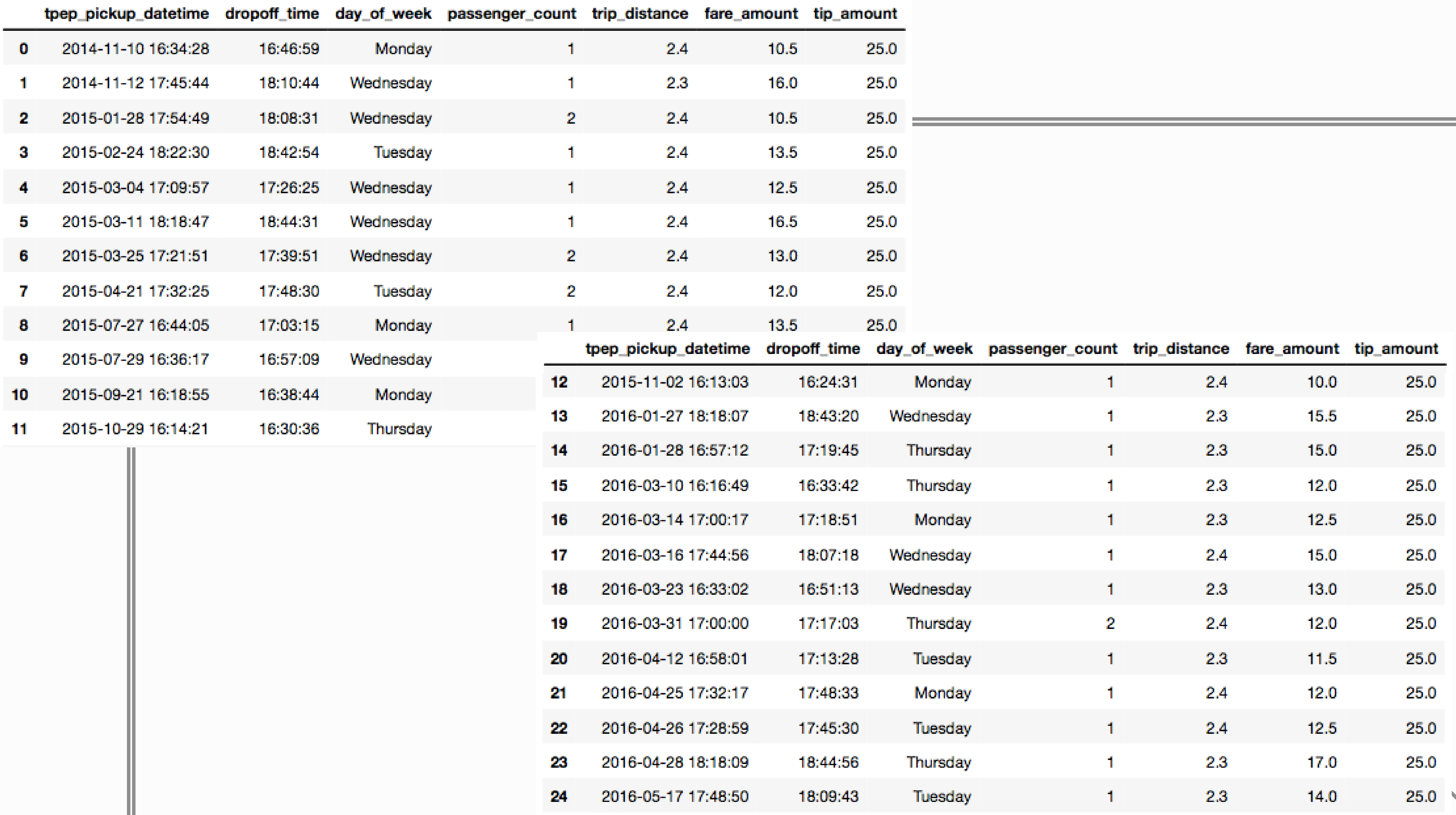


Passenger Privacy



Passenger Privacy





	tpep_pickup_datetime	dropoff_time	day_of_week	passenger_count	trip_distance	fare_amount	tip_amount
0	2014-11-10 16:34:28	16:46:59	Monday	1	2.4	10.5	25.0
1	2014-11-12 17:45:44	18:10:44	Wednesday	1	2.3	16.0	25.0
2	2015-01-28 17:54:49	18:08:31	Wednesday	2	2.4	10.5	25.0
3	2015-02-24 18:22:30	18:42:54	Tuesday	1	2.4	13.5	25.0
4	2015-03-04 17:09:57	17:26:25	Wednesday	1	2.4	12.5	25.0
5	2015-03-11 18:18:47	18:44:31	Wednesday	1	2.4	16.5	25.0
6	2015-03-25 17:21:51	17:39:51	Wednesday	2	2.4	13.0	25.0
7	2015-04-21 17:32:25	17:48:30	Tuesday	2	2.4	12.0	25.0
8	2015-07-27 16:44:05	17:03:15	Monday	1	2.4	13.5	25.0
9	2015-07-29 16:36:17	16:57:09	Wednesday				
10	2015-09-21 16:18:55	16:38:44	Monday				
11	2015-10-29 16:14:21	16:30:36	Thursday				

	tpep_pickup_datetime	dropoff_time	day_of_week	passenger_count	trip_distance	fare_amount	tip_amount
12	2015-11-02 16:13:03	16:24:31	Monday	1	2.4	10.0	25.0
13	2016-01-27 18:18:07	18:43:20	Wednesday	1	2.3	15.5	25.0
14	2016-01-28 16:57:12	17:19:45	Thursday	1	2.3	15.0	25.0
15	2016-03-10 16:16:49	16:33:42	Thursday	1	2.3	12.0	25.0
16	2016-03-14 17:00:17	17:18:51	Monday	1	2.3	12.5	25.0
17	2016-03-16 17:44:56	18:07:18	Wednesday	1	2.4	15.0	25.0
18	2016-03-23 16:33:02	16:51:13	Wednesday	1	2.3	13.0	25.0
19	2016-03-31 17:00:00	17:17:03	Thursday	2	2.4	12.0	25.0
20	2016-04-12 16:58:01	17:13:28	Tuesday	1	2.3	11.5	25.0
21	2016-04-25 17:32:17	17:48:33	Monday	1	2.4	12.0	25.0
22	2016-04-26 17:28:59	17:45:30	Tuesday	1	2.4	12.5	25.0
23	2016-04-28 18:18:09	18:44:56	Thursday	1	2.3	17.0	25.0
24	2016-05-17 17:48:50	18:09:43	Tuesday	1	2.3	14.0	25.0

Thank you !