An analysis of 2015 NYC Green Taxi Trip Data

Analyzing large datasets with Python

Data Source NYC OpenData

24th February 2017

Carlo Motta

A look at the dataset

- 19 233 766 (> 19 millions) observations
- ~3Gb
- Features:

0	vendorid	10	Trip_distance
1	pickup_datetime	11	Fare_amount
2	dropoff_datetime	12	Extra
3	Store_and_fwd_flag	13	MTA_tax
4	rate_code	14	Tip_amount
5	Pickup_longitude	15	Tolls_amount
6	Pickup_latitude	16	Ehail_fee
7	Dropoff_longitude	17	Improvement_surcharge
8	Dropoff_latitude	18	Total_amount
9	Passenger_count	19	Payment_type
10	Trip_distance	20	Trip_type

- To make the problem tractable with my modest resources, I reduced the number of observations via random sampling (~5%)

Some stats

	vendorid	rate_code	Pickup_longitude	Pickup_latitude
mean	1.782501	1.099008	-73.826362	40.689645
std	0.412545	0.637620	2.836832	1.563842
min	1.000000	1.000000	-115.174675	0.000000
25%	2.000000	1.000000	-73.959190	40.699361
50%	2.000000	1.000000	-73.945099	40.746895
75%	2.000000	1.000000	-73.916901	40.803665
max	2.000000	99.000000	0.000000	41.292233

	Dropoff_longitude	Dropoff_latitude	Passenger_count	Trip_distance
mean	-73.828218	40.689380	1.371483	2.882150
std	2.797645	1.542149	1.044929	2.947592
min	-115.174850	0.000000	0.000000	0.000000
25%	-73.966963	40.700444	1.000000	1.070000
50%	-73.944321	40.748108	1.000000	1.900000
75%	-73.909386	40.792949	1.000000	3.640000
max	0.000000	41.637573	8.000000	105.740000

Some stats

	Fare_amount	Extra	MTA_tax	Tip_amount
mean	12.331019	0.353044	0.486347	1.221430
std	10.256481	0.365783	0.086079	3.952988
min	-200.000000	-1.000000	-0.500000	-100.000000
25%	6.500000	0.000000	0.500000	0.000000
50%	9.500000	0.500000	0.500000	0.000000
75%	15.000000	0.500000	0.500000	2.000000
max	499.000000	1.000000	0.500000	480.000000

	Tolls_amount	Ehail_fee	Improvement_surcharge	Total_amount
mean	0.112948	NaN	0.290724	14.795530
std	1.033035	NaN	0.053963	12.065141
min	0.000000	NaN	-0.300000	-200.000000
25%	0.000000	NaN	0.300000	7.800000
50%	0.000000	NaN	0.300000	11.300000
75%	0.000000	NaN	0.300000	17.800000
max	235.000000	NaN	0.300000	499.300000

Cleaning data

- I parsed the data and applied **filters** to some features

- e. g.

  $|x - x_{AVG}| < 3\sigma$ x = latitude, fare, ...

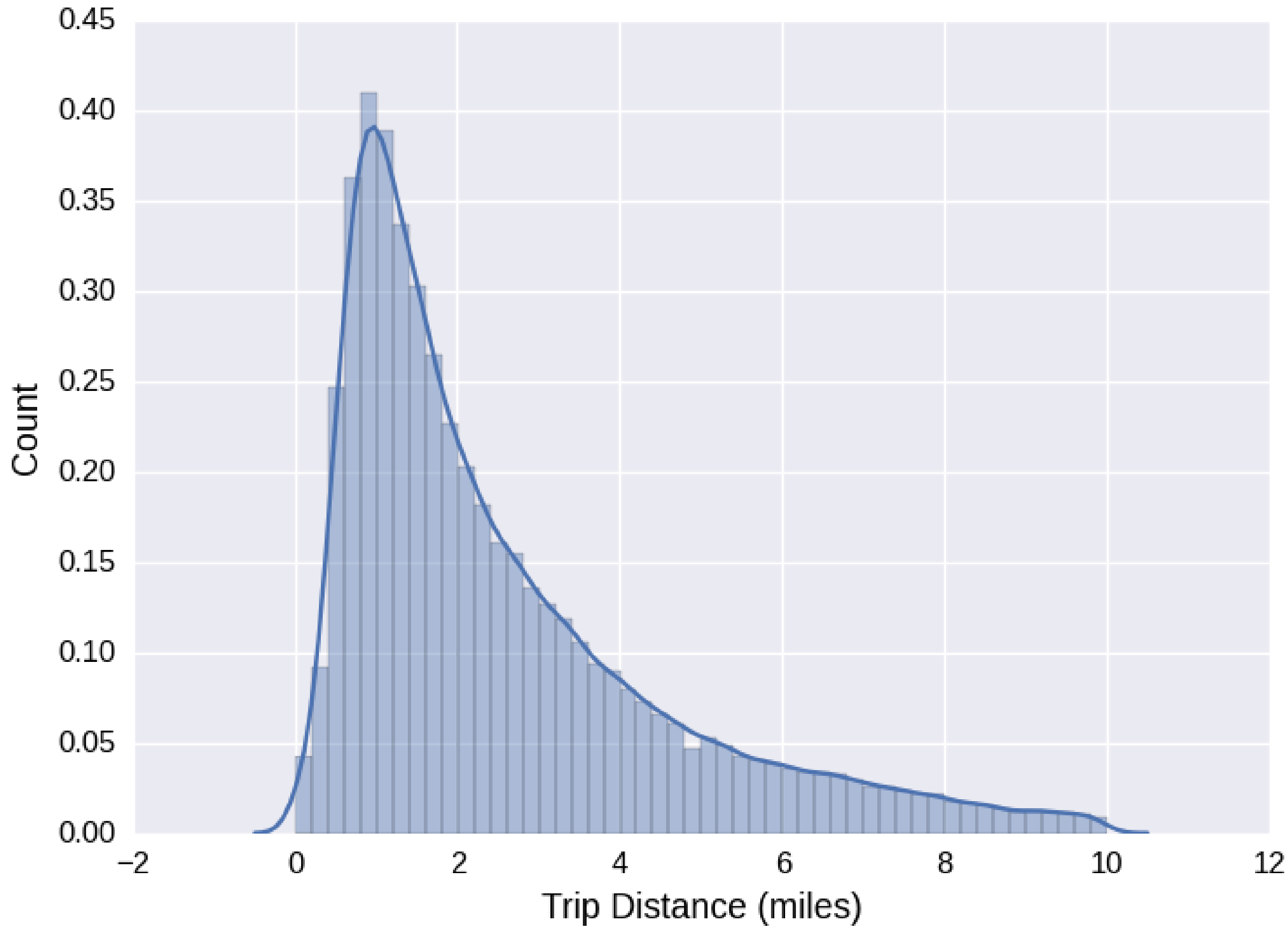
  $x > 0$ x = num. Passengers

  (...)

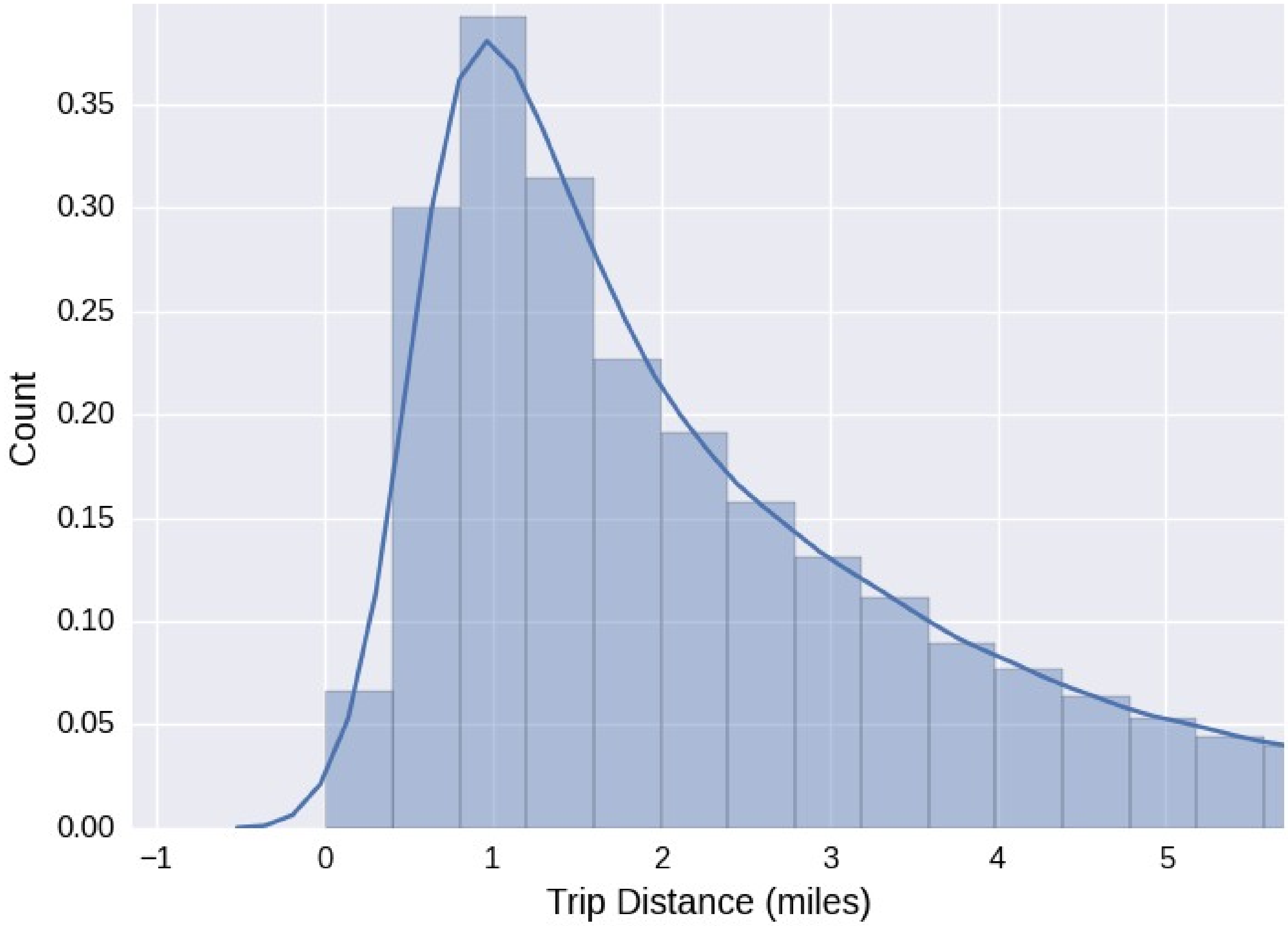
- Created derived variables:

Week_day	
Month_day	
Hour	1-to-24
Shift type	Morning/afternoon/night
Speed	
Tip_percentage	
With_tip	
Origin	(Manhattan or not)

Trip Distance

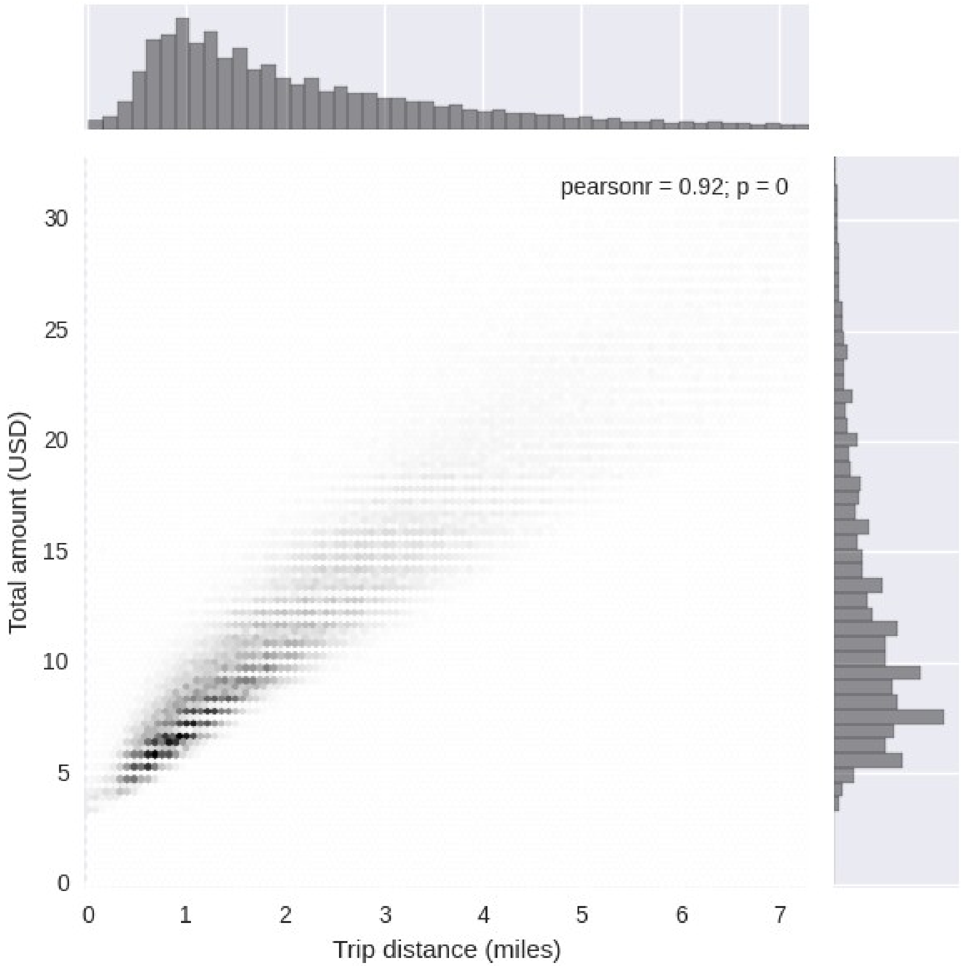


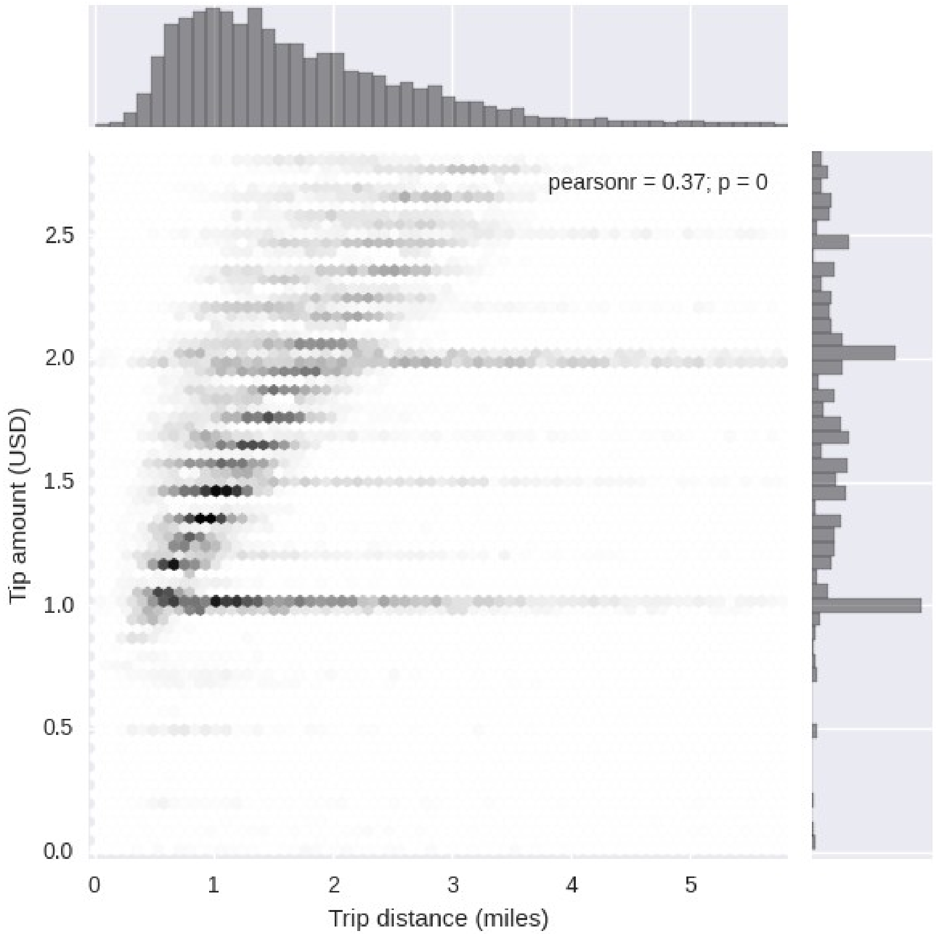
Trip Distance



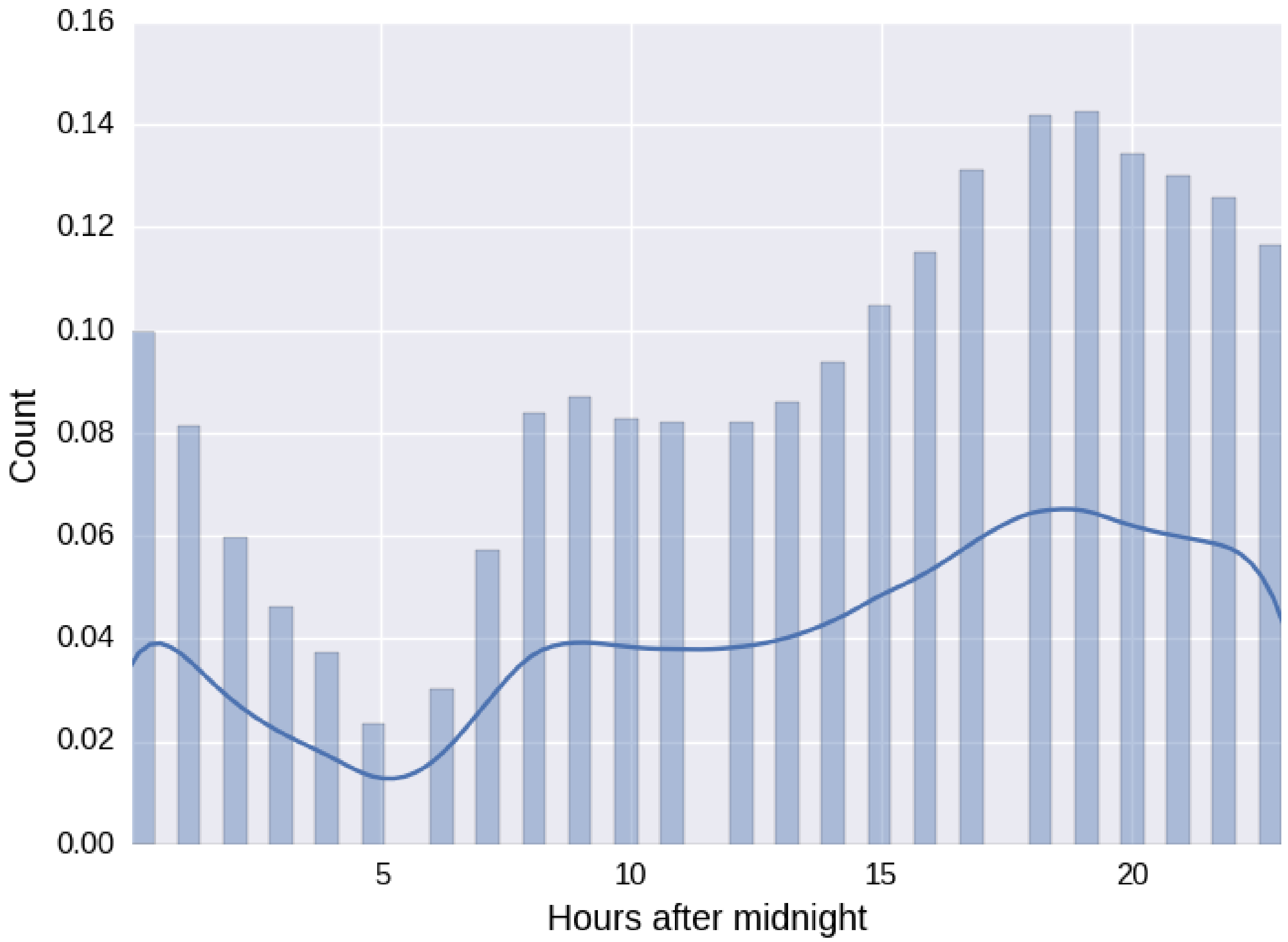
Right-skewed distribution due to lower bound (0)

Correlations

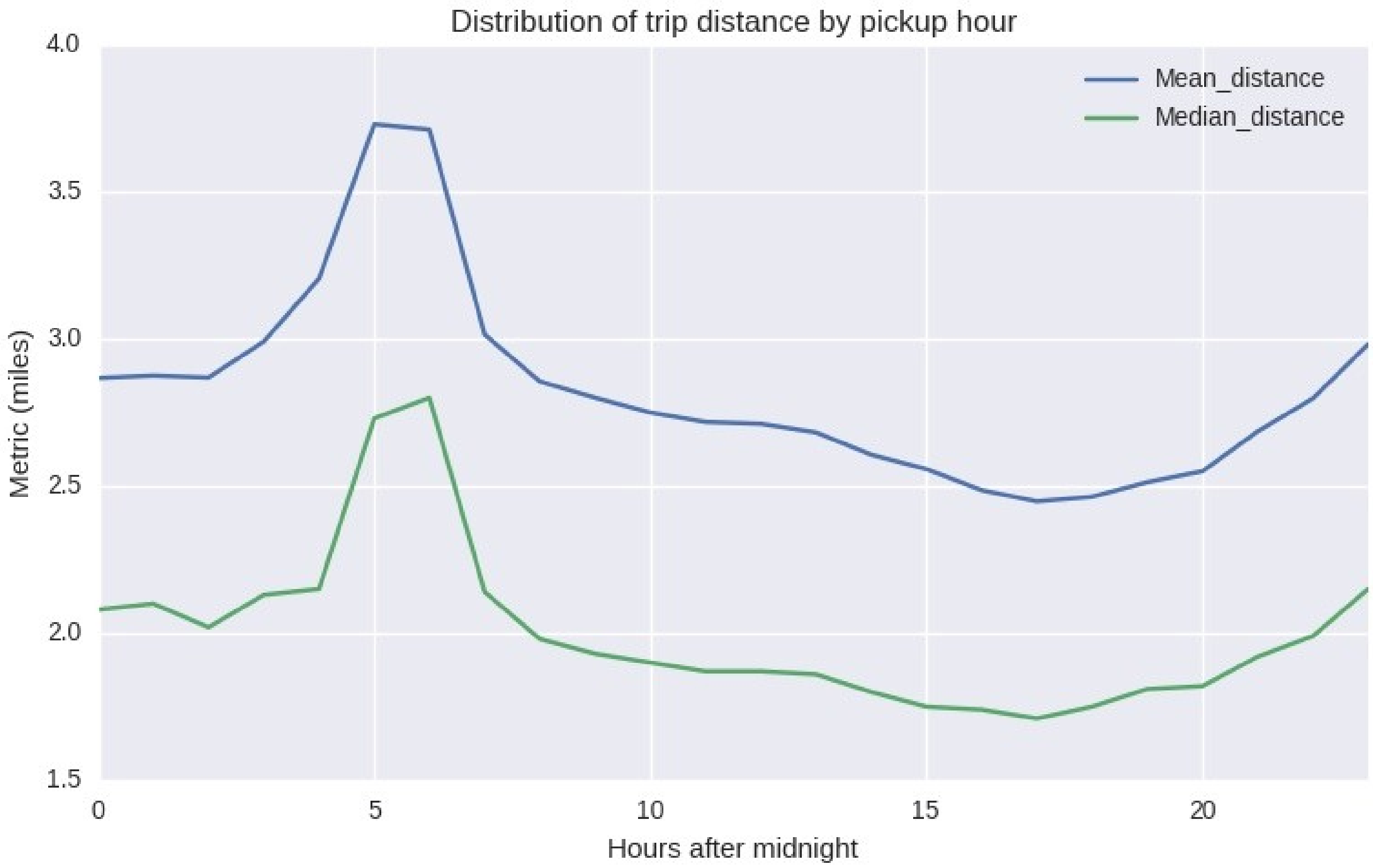




Pickup hour

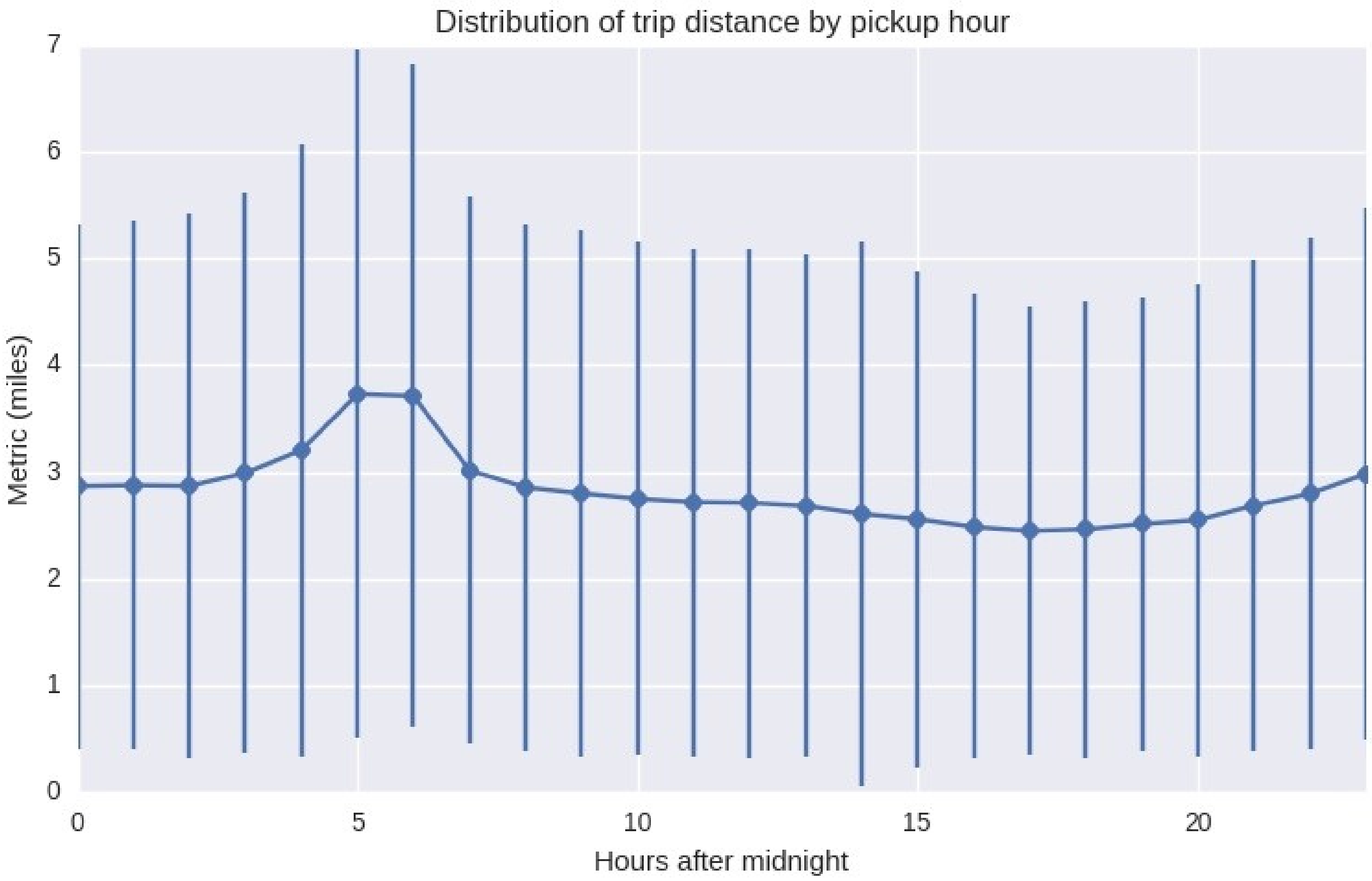

0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Count
5
10
15
20
Hours after midnight

Trip Distance by hour

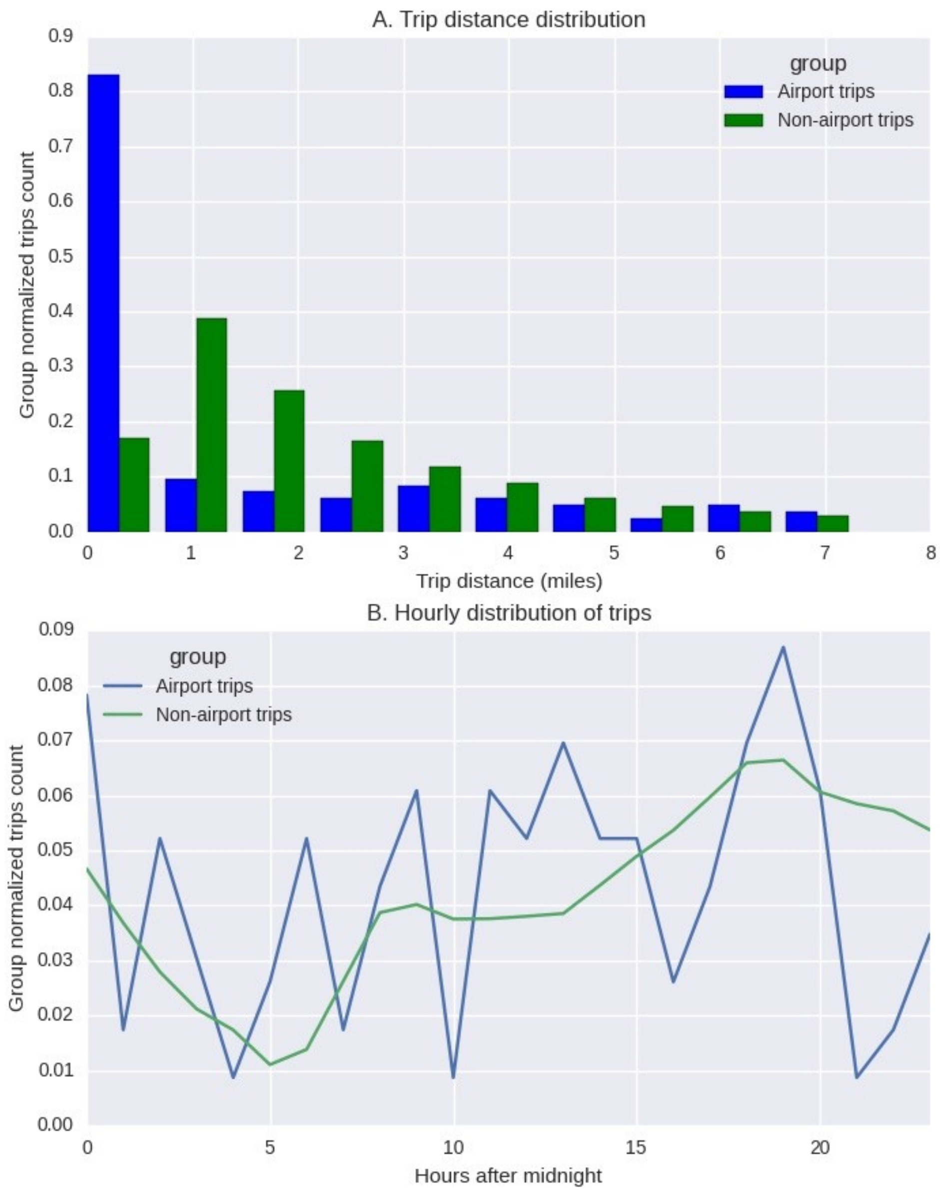


Longer travels in the morning 5am-7am

Trip Distance by hour

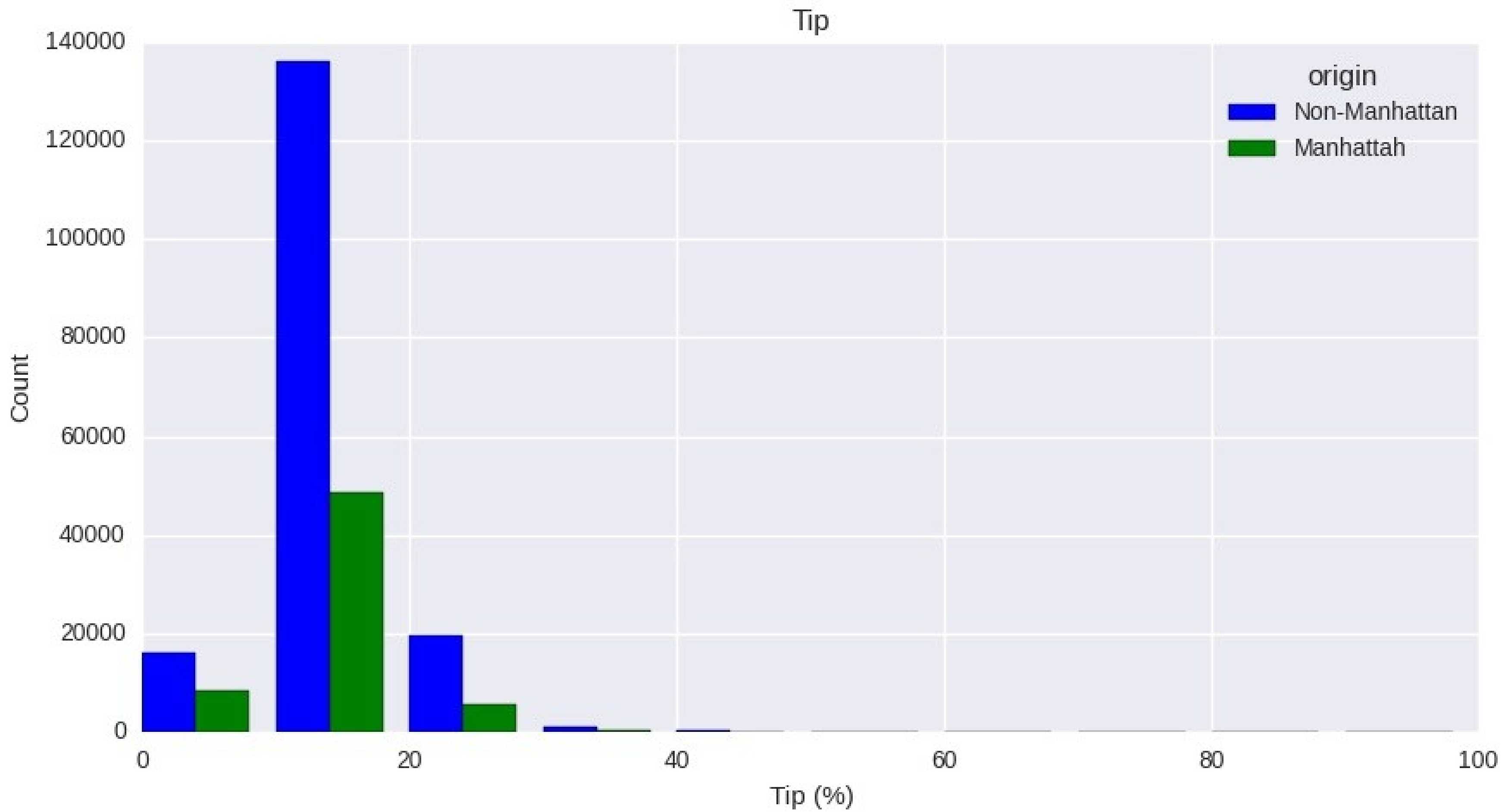


Trip Distance by hour



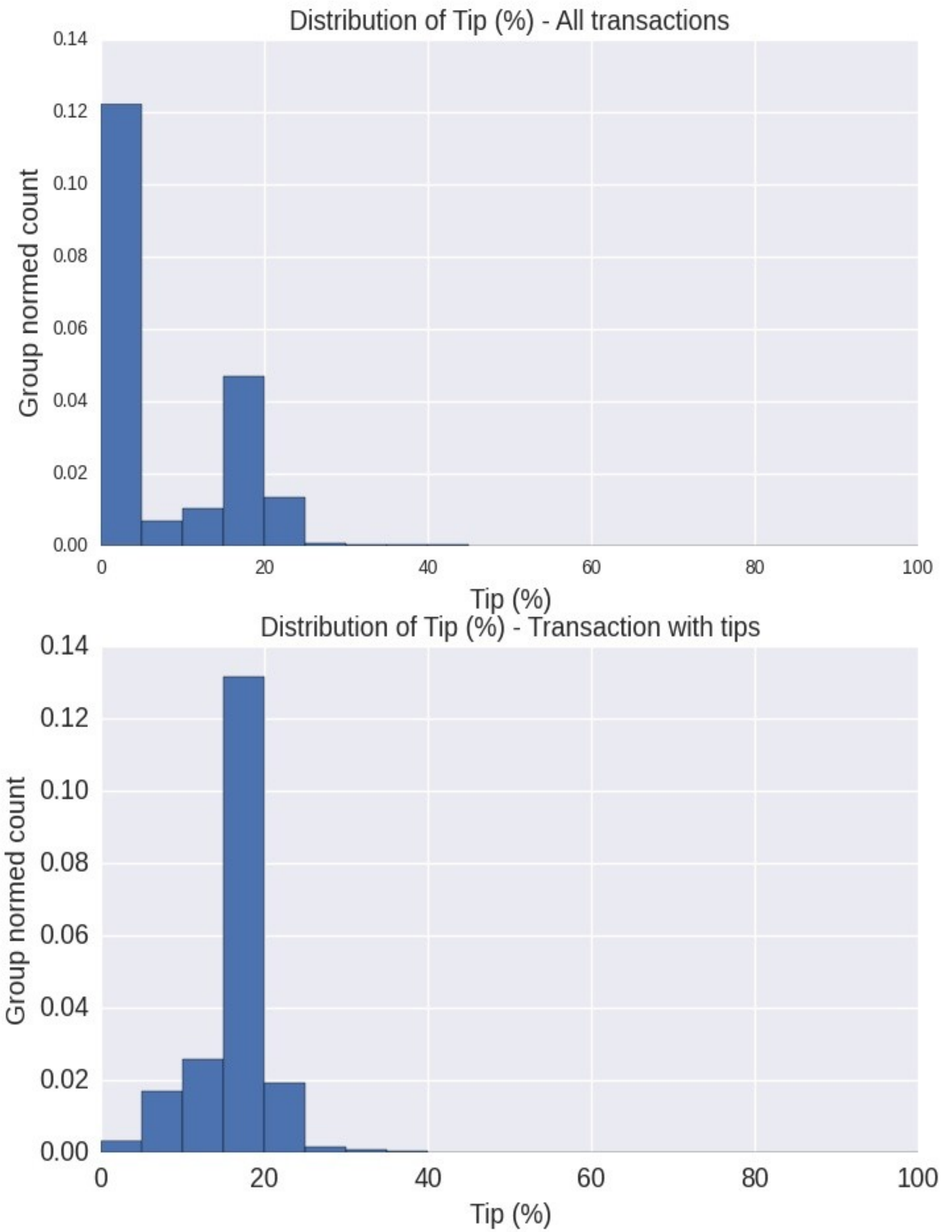
- Number of trips to/from NYC airports: 115
- Average total charged amount (before tip) of trips to/from NYC airports: $ 23.9934782609 per trip

Tip and Trip origin

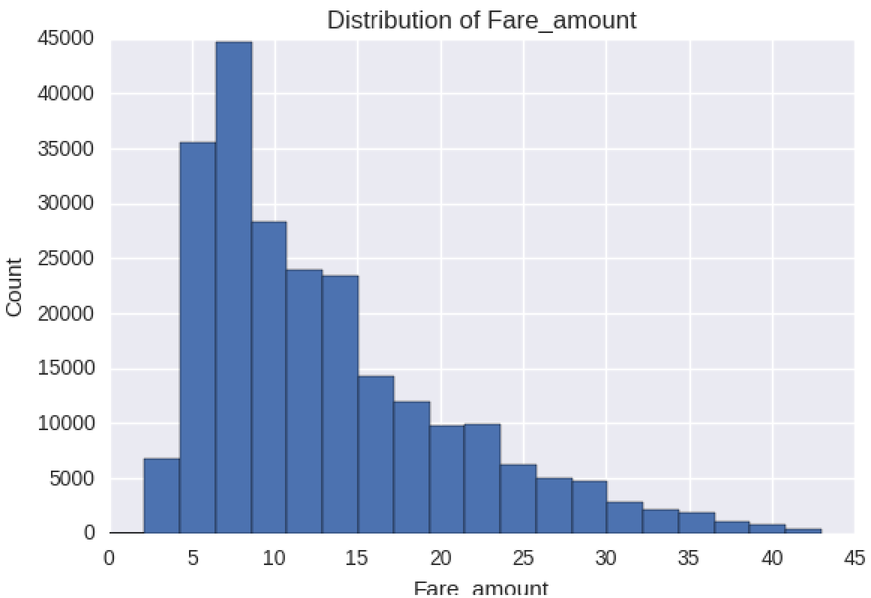


t-test results:
statistics = 37.11635088218631,
p-value = 1.0641940231098476e-299
→ the two distributions are different at 95% level of confidnce

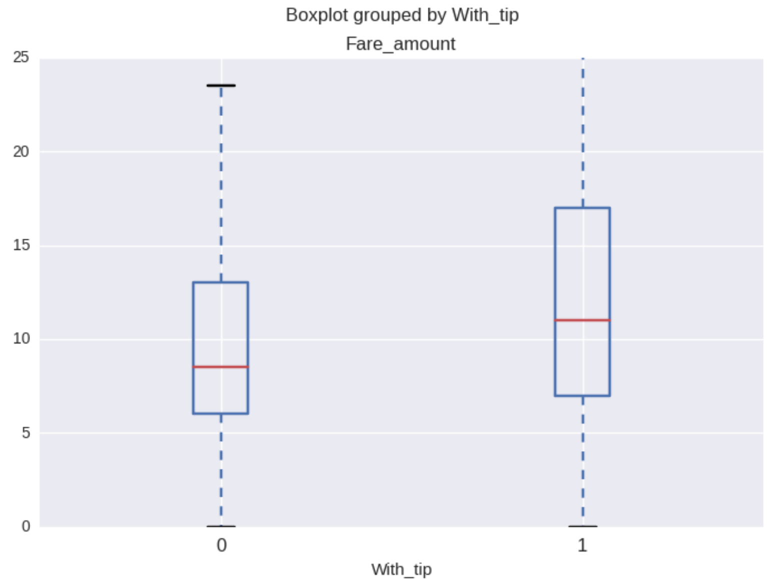
Distribution of tips



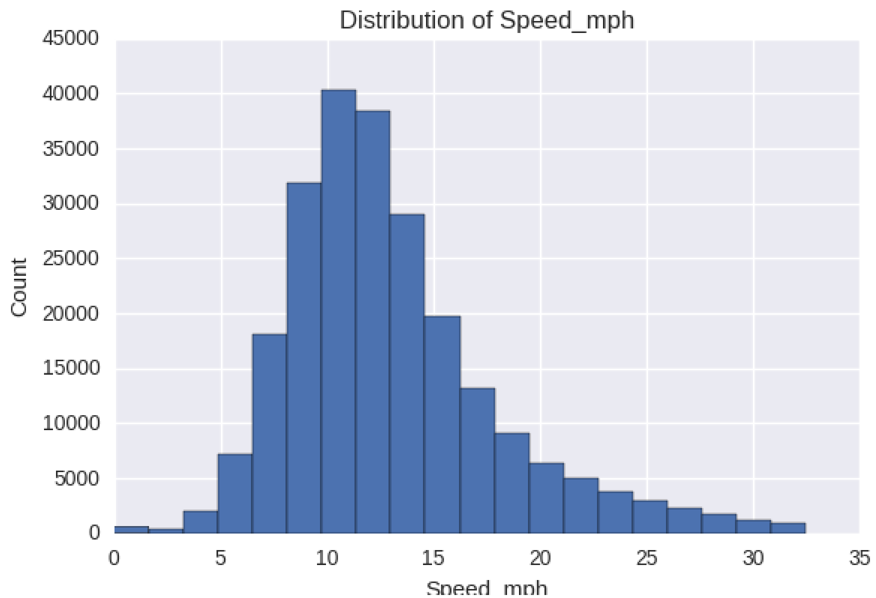
Distribution of tips

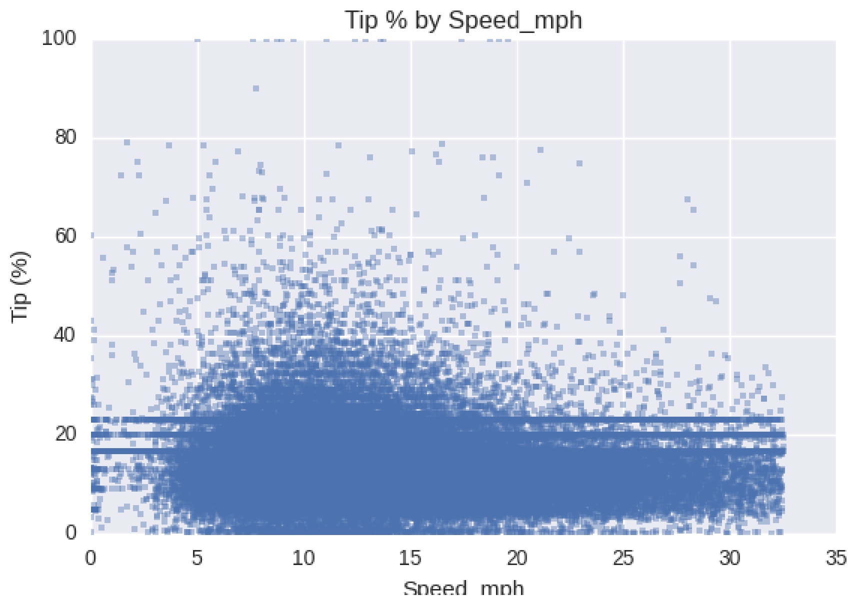


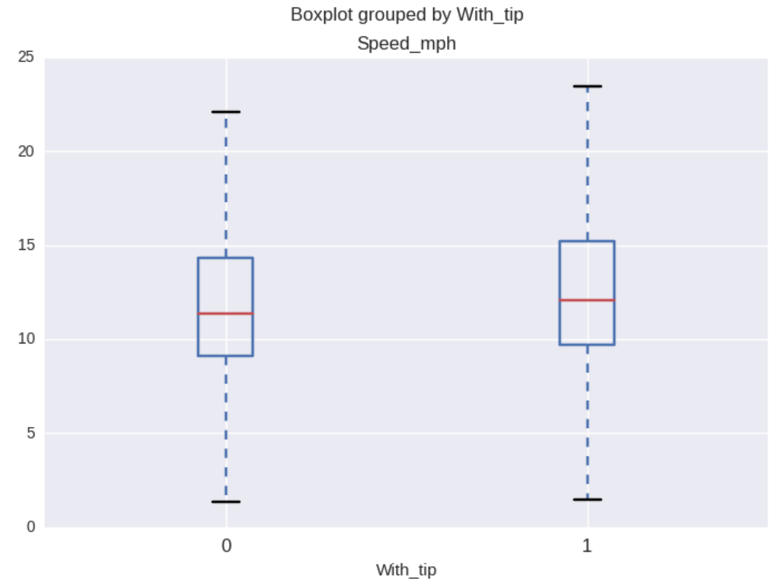
t-test results: (-65.553446118735991, 0.0)



Distribution of tips

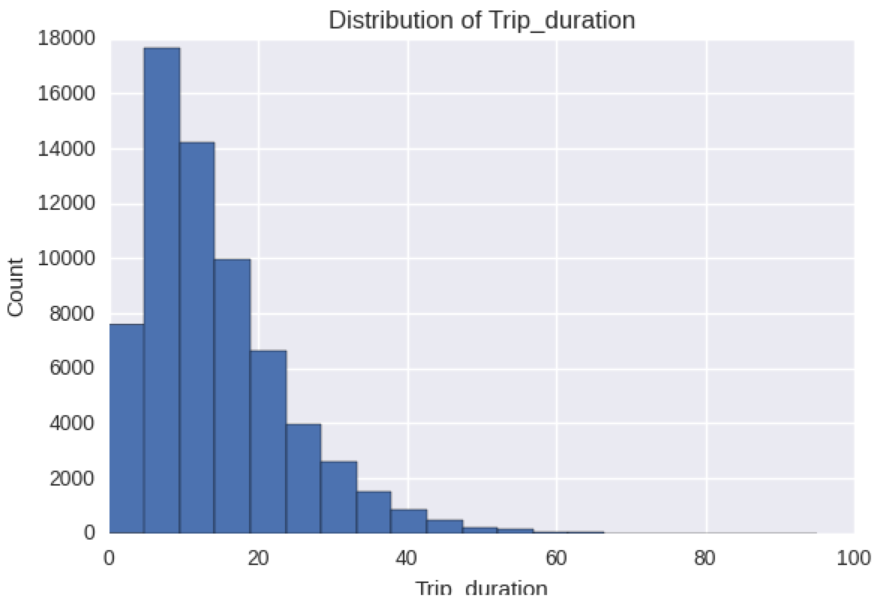


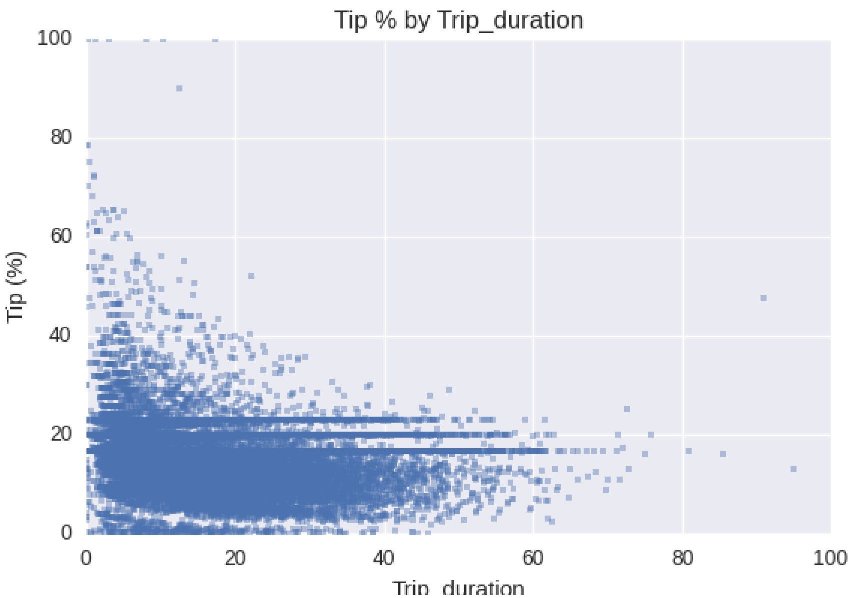




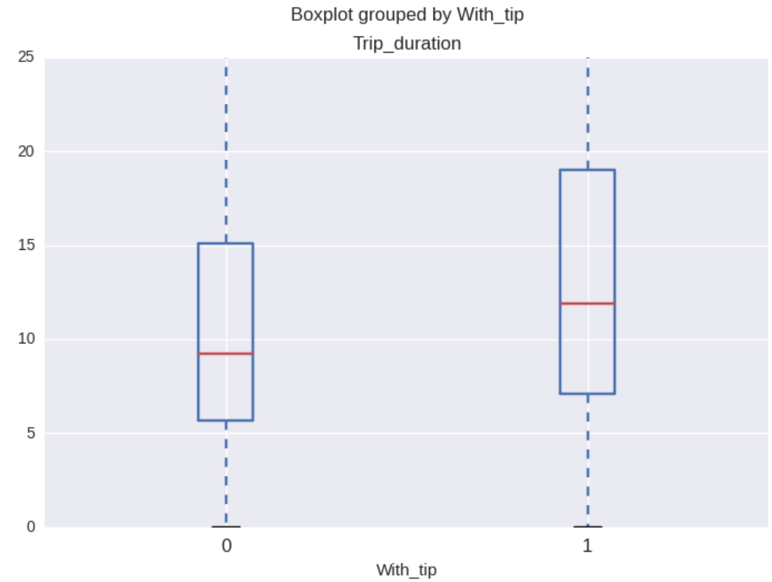
t-test results:
(-23.83456220, 2.47366441e-125)

Distribution of tips



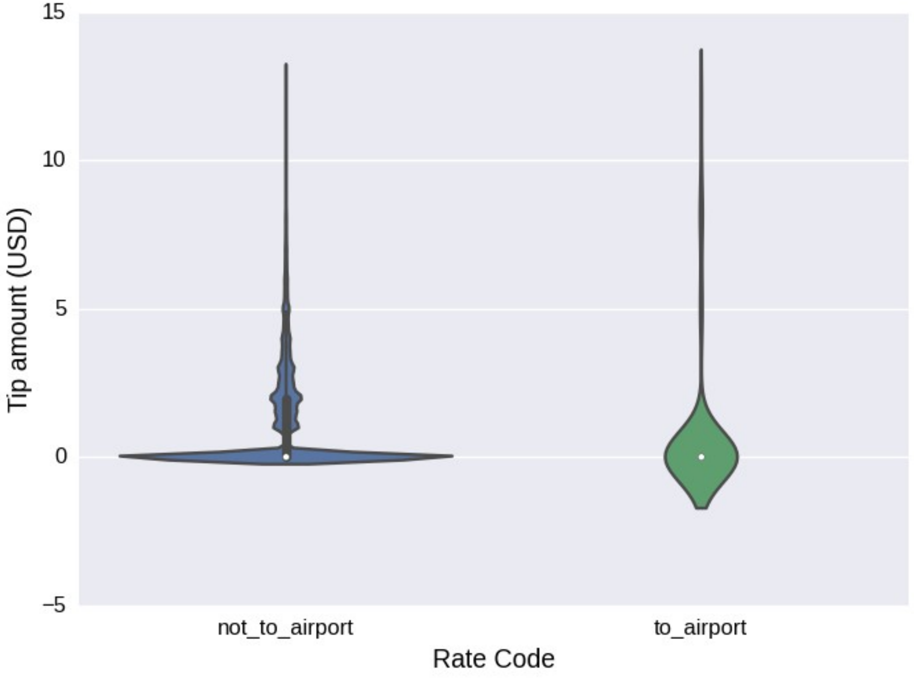


t-test results:
(-2.1104725, 0.03481)

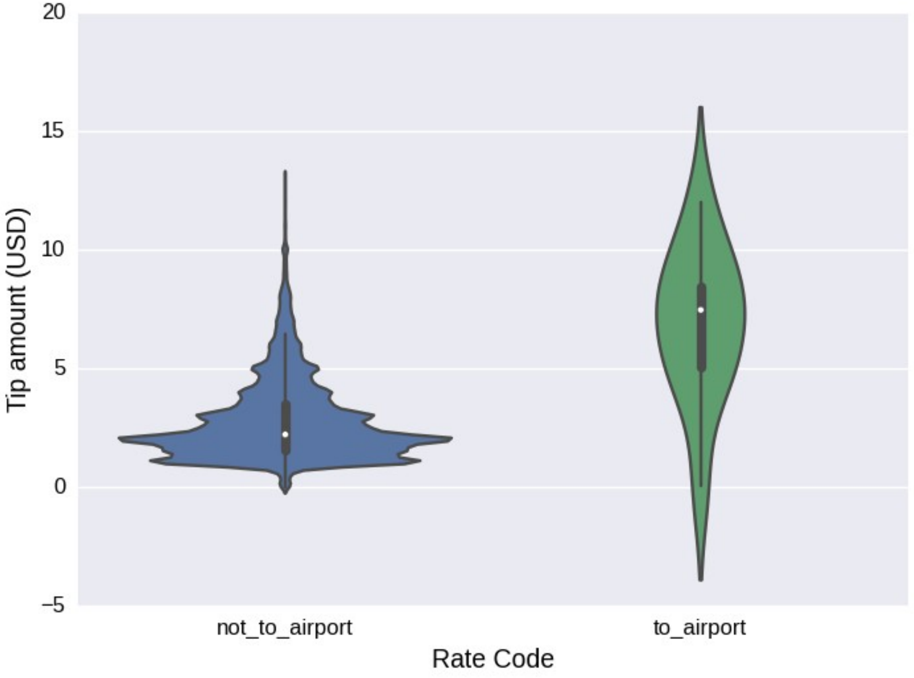


Tip and Trip type

Including **NO** tip

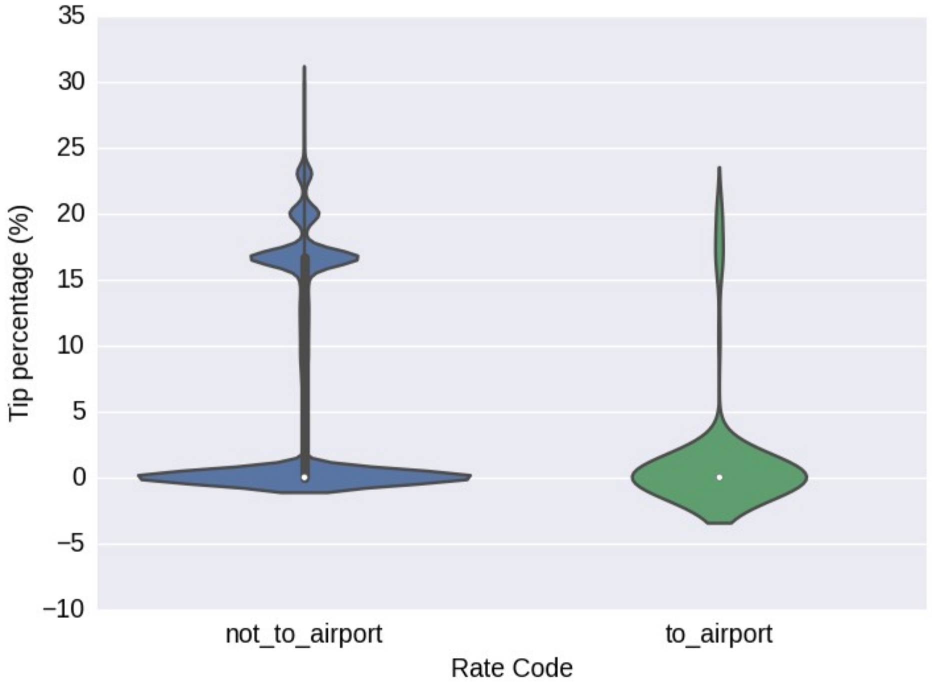


Excluding **NO** tip

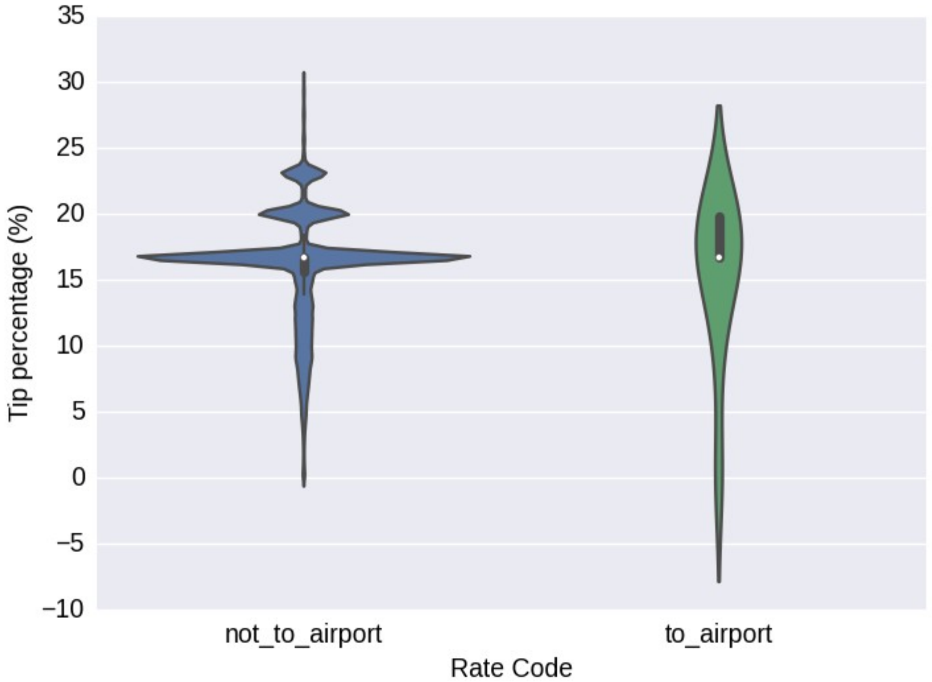


Tip % and Trip type

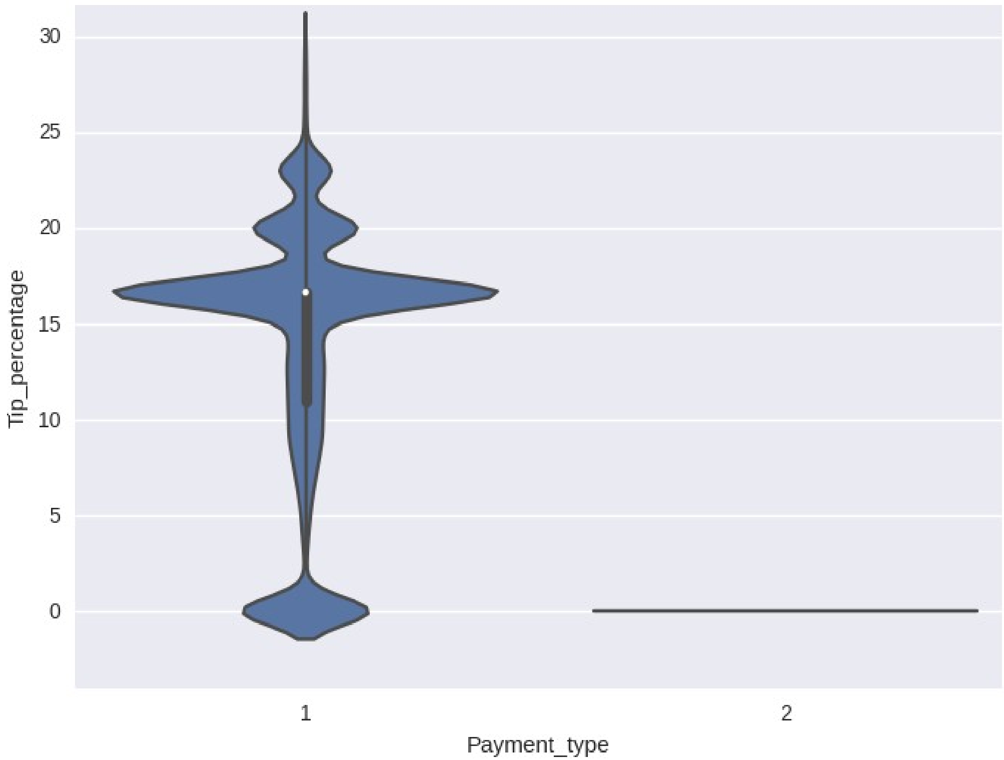
Including **NO** tip



Excluding **NO** tip



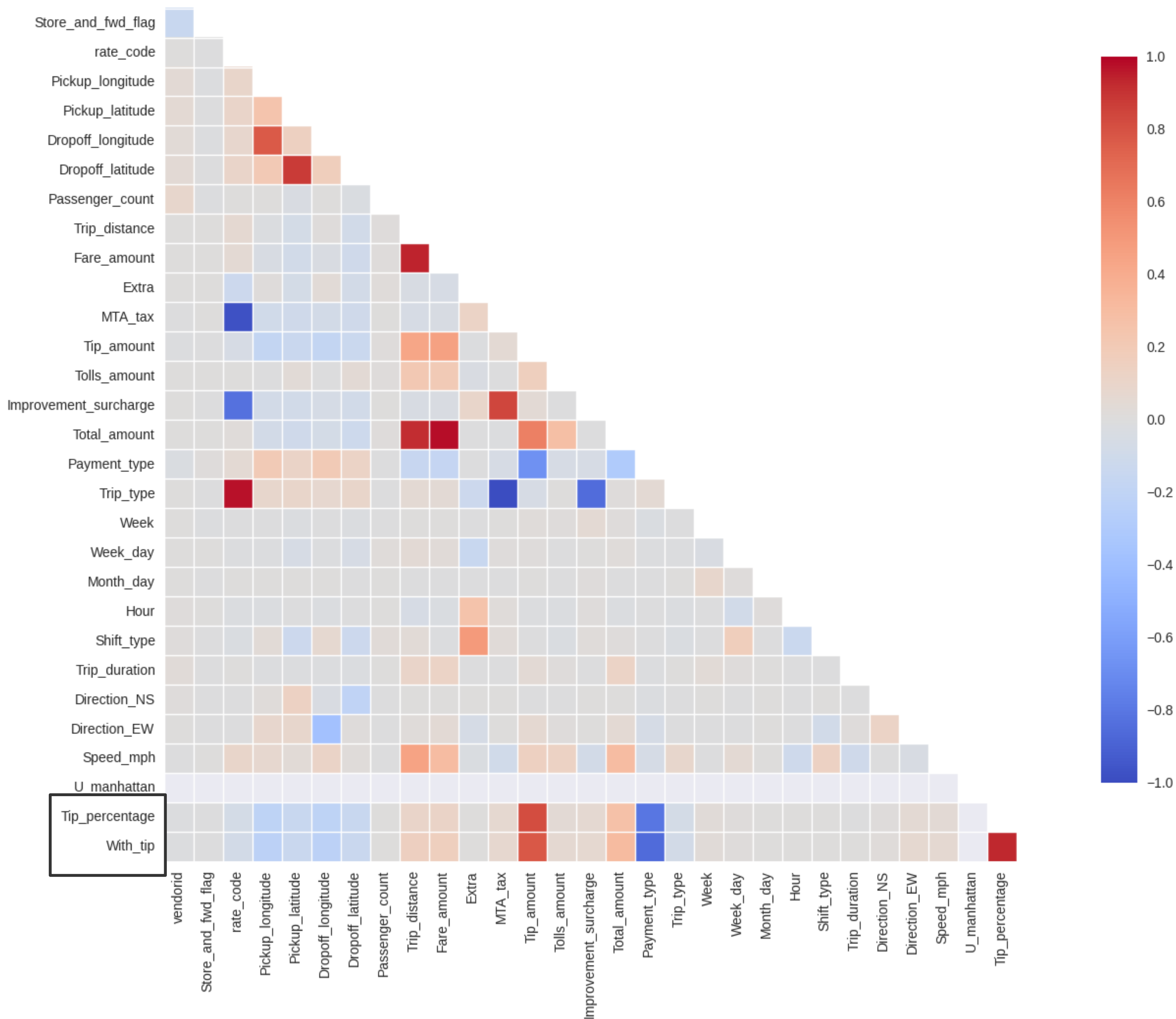
Tip % and payment method



Credit Card Cash

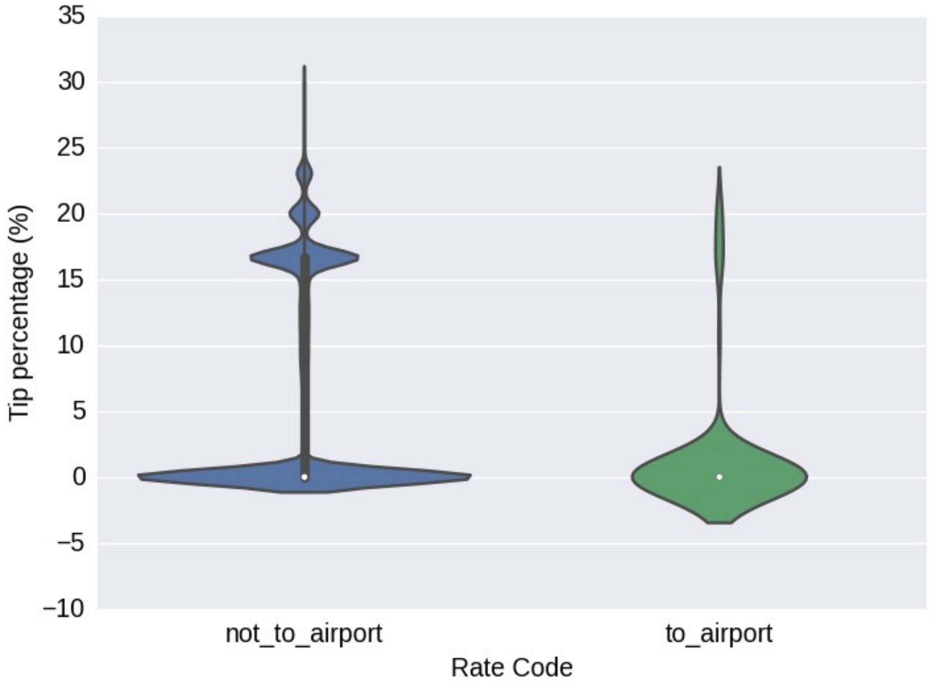
- >>> len(df3[(df3.Payment_type == 2) & (df3.Tip_amount>0)]) → 0
- CURIOUS: no transaction with cash has a tip
- Checking back on the full dataset: only 19 cash transactions have a tip >0, i.e. ~0.0001%

Correlation plot

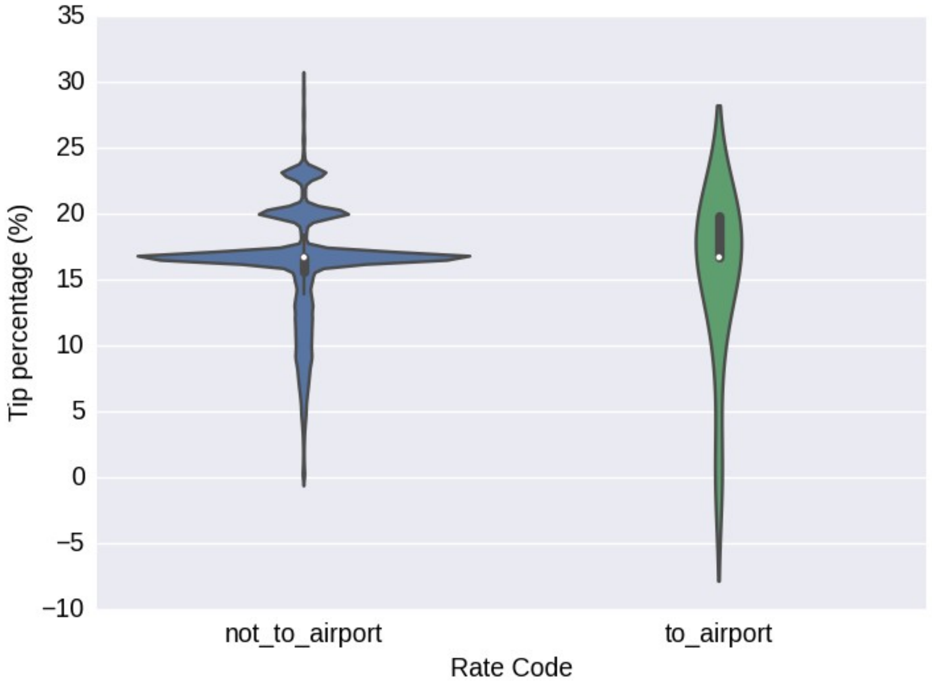


Tip and Trip type

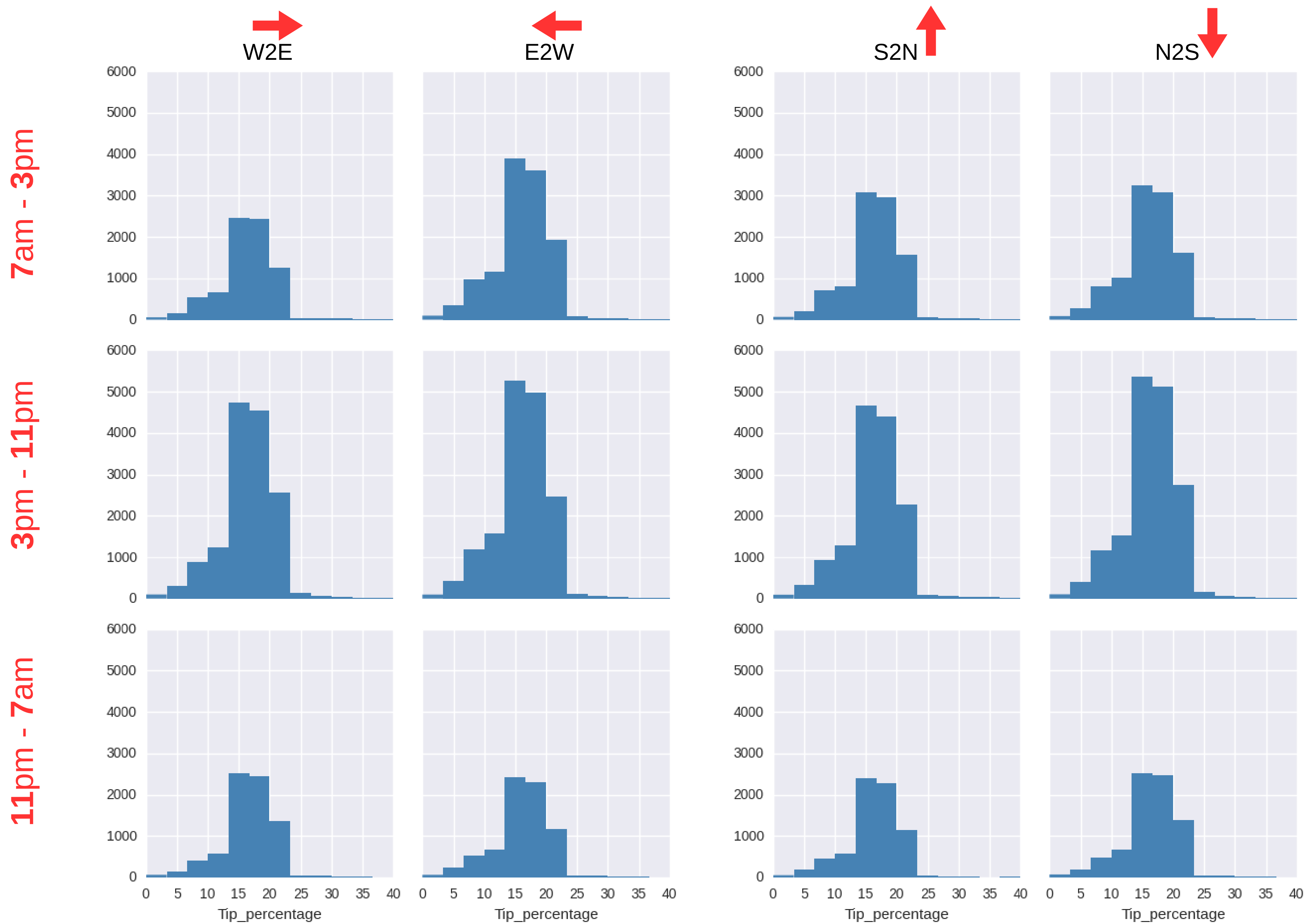
Including **NO** tip



Excluding **NO** tip



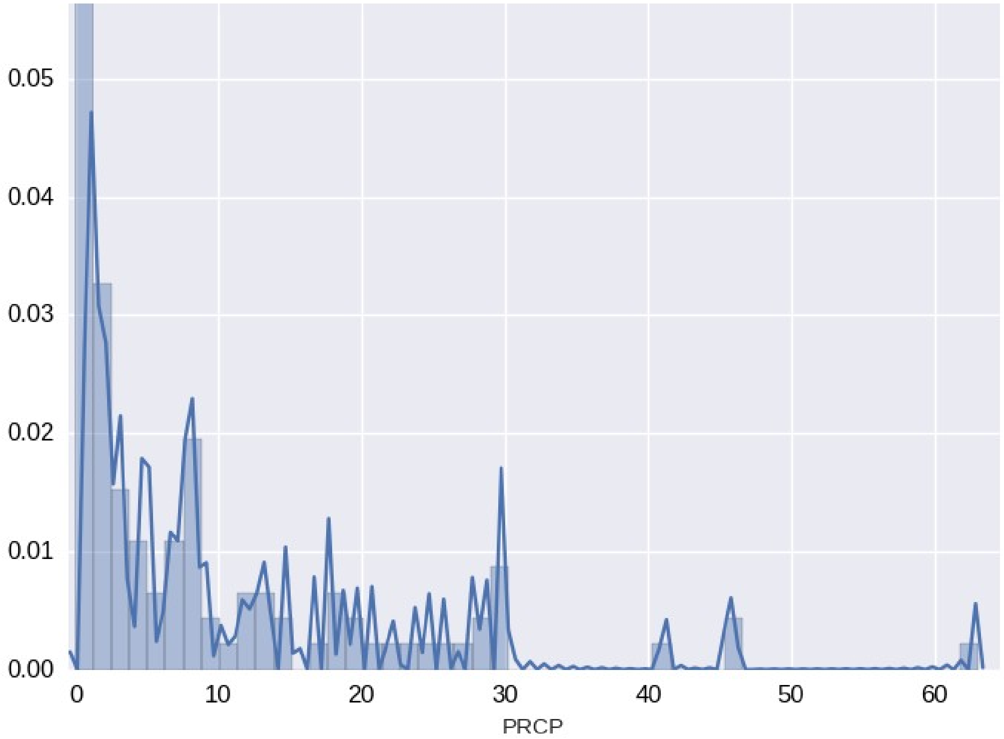
Tip and Shift type



Precipitations

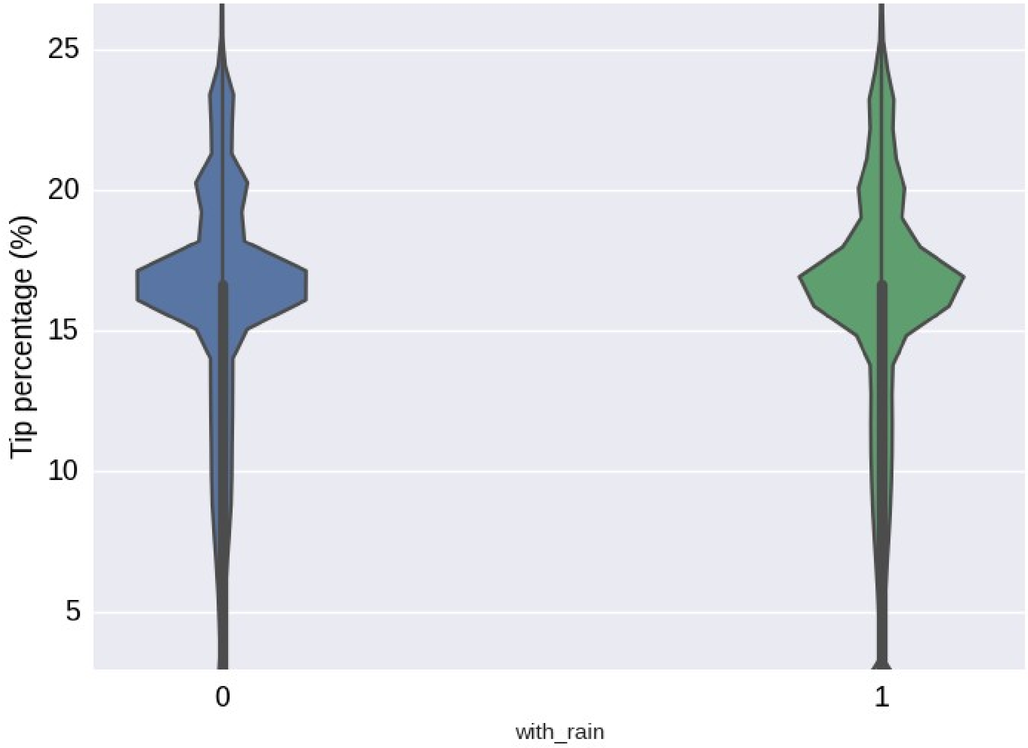
- Climate data obtained from:

National Oceanic and Atmospheric Administration
U.S. Department of Commerce



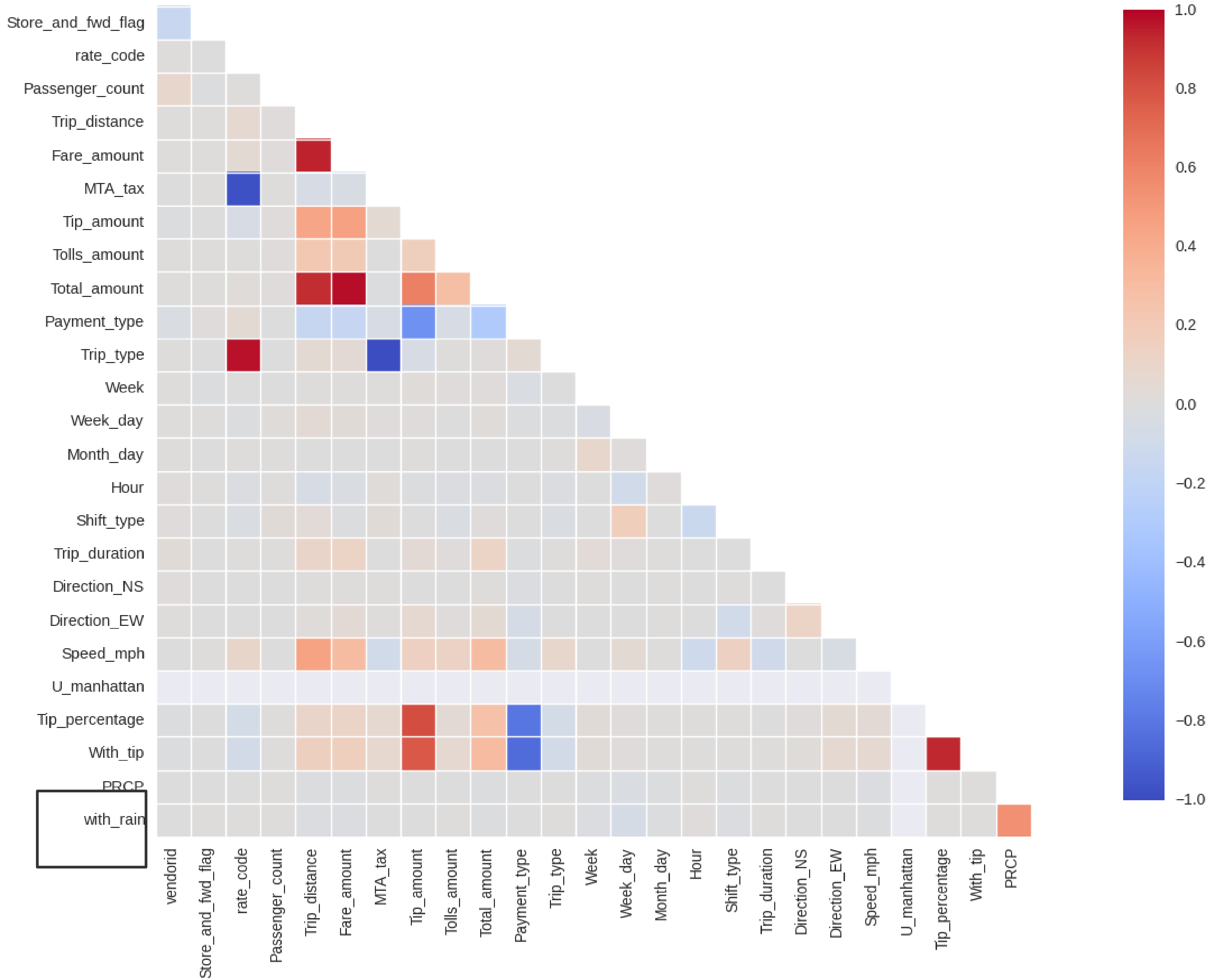
- Rain on 30% of days in a year

Precipitations



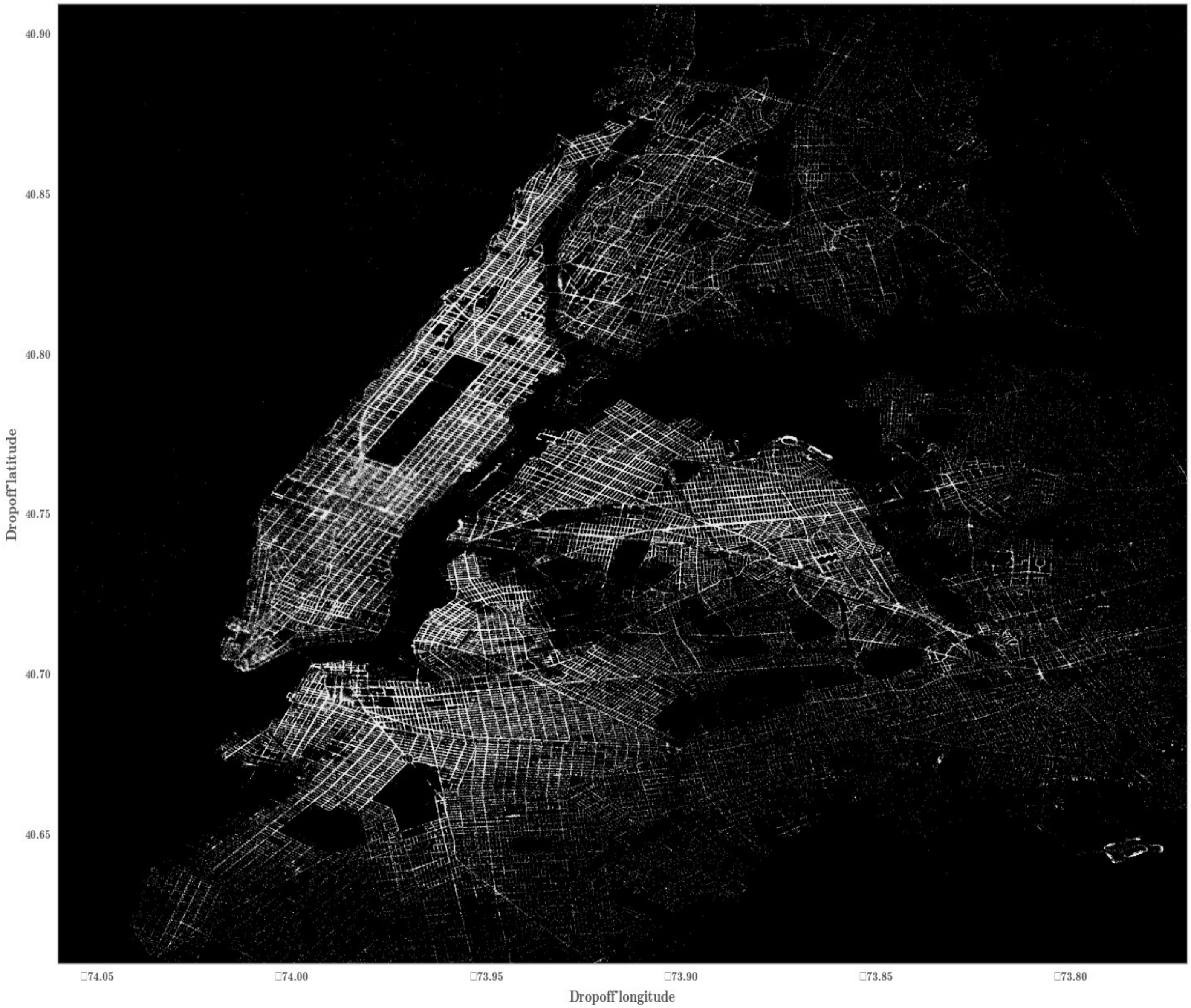
- Rain has no effect on tip

Correlation plot



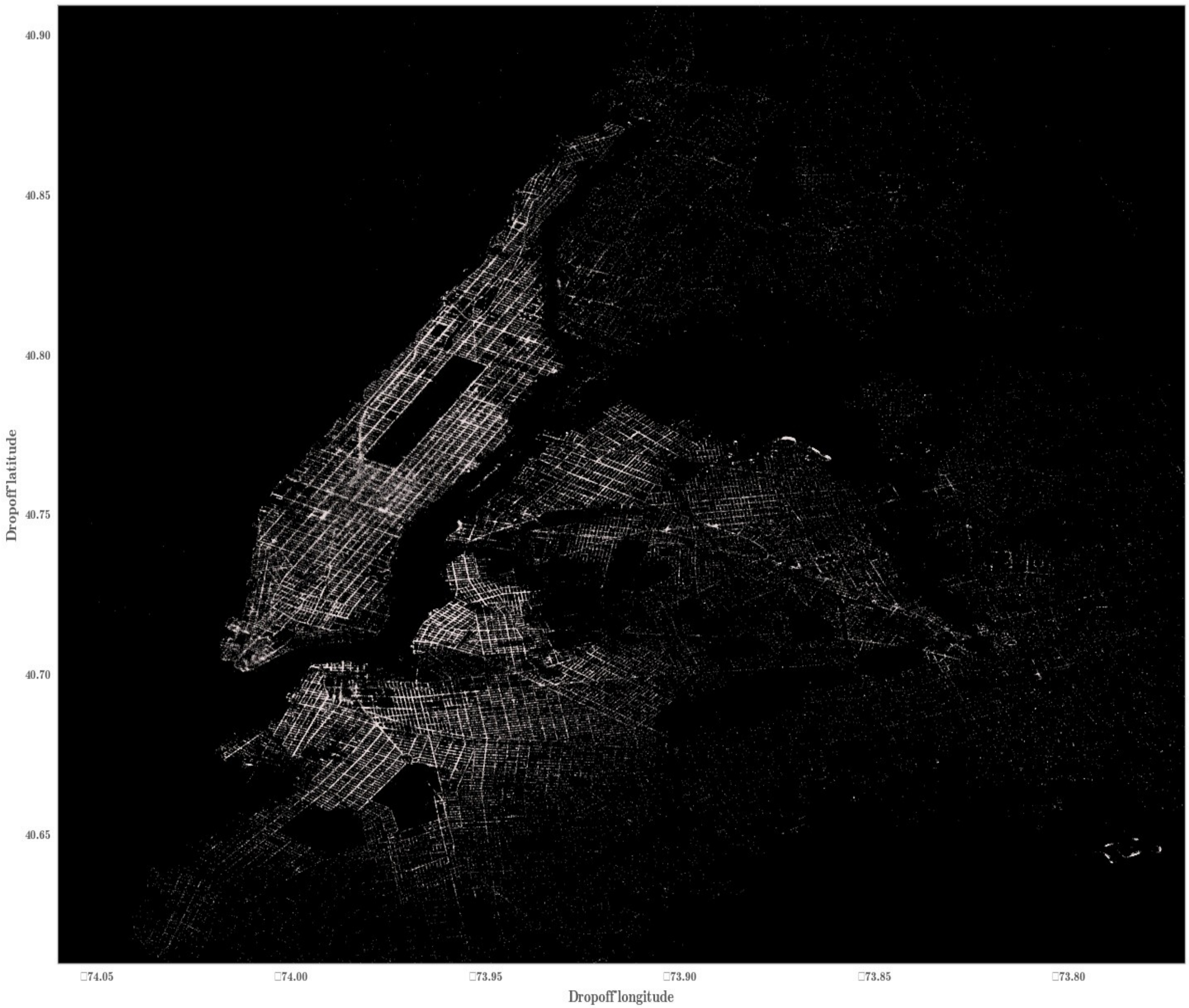
Mapping the data

Dropoff



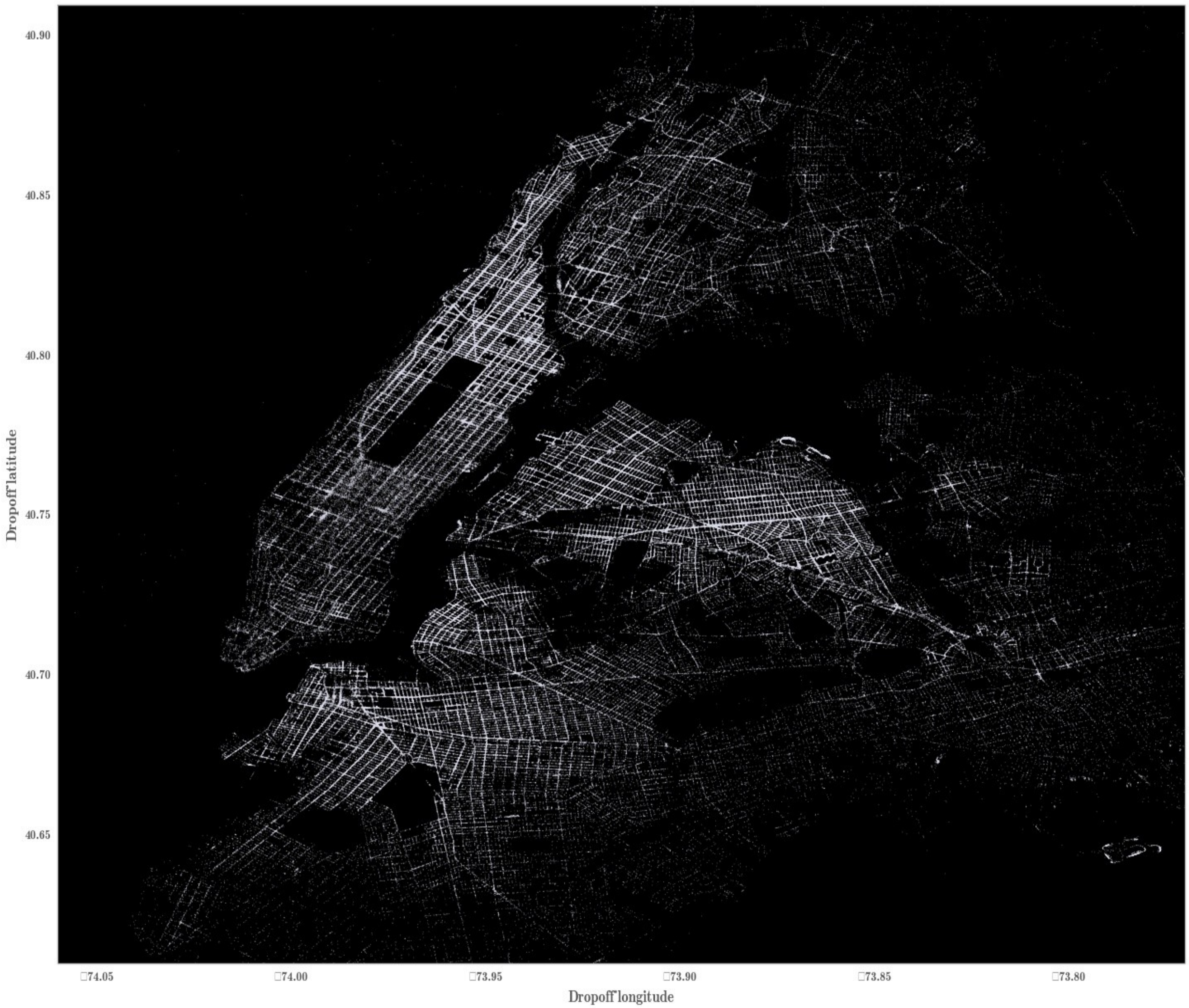
Mapping the data: tip>mean

Dropoff



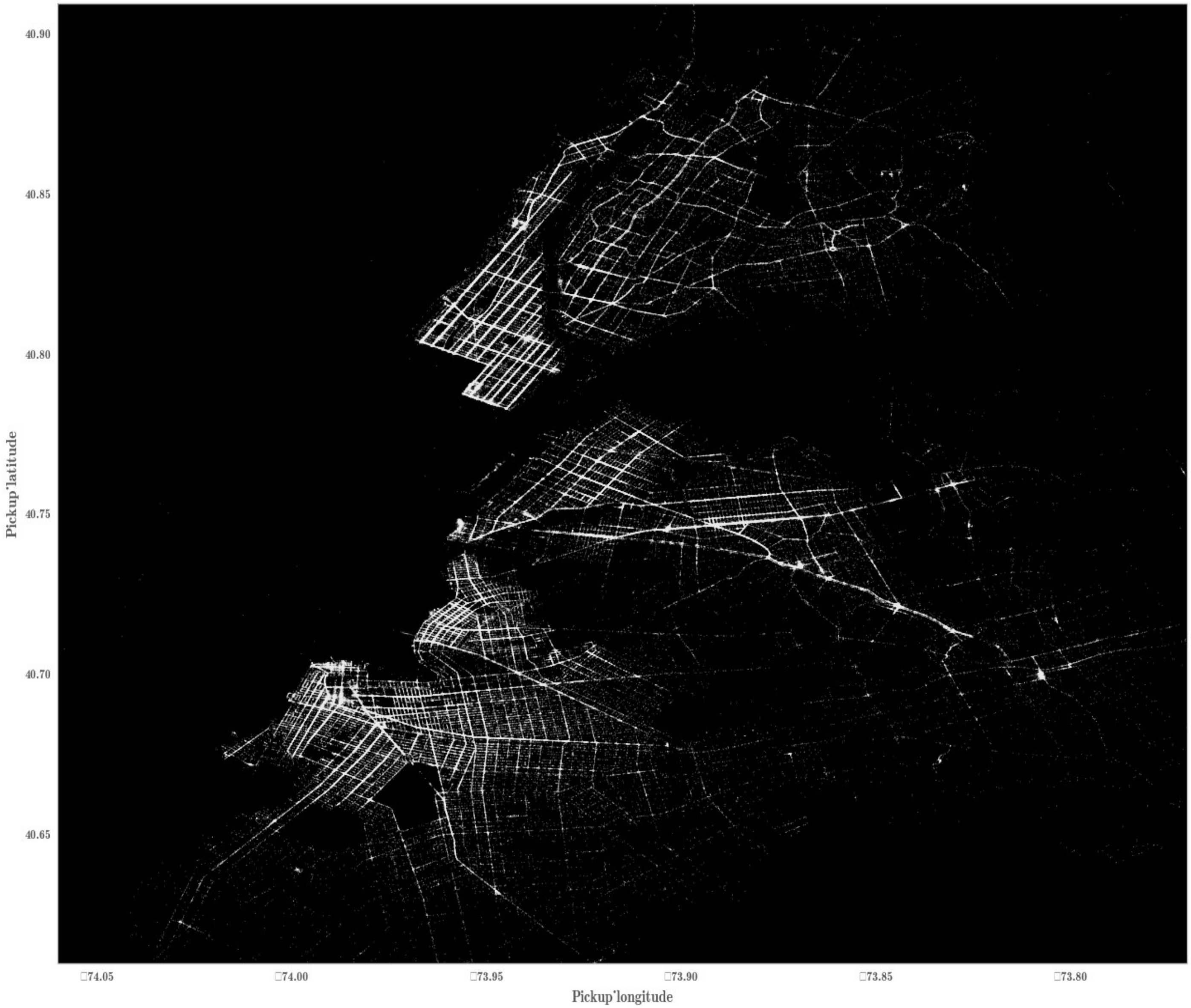
Mapping the data: tip<mean

Dropoff



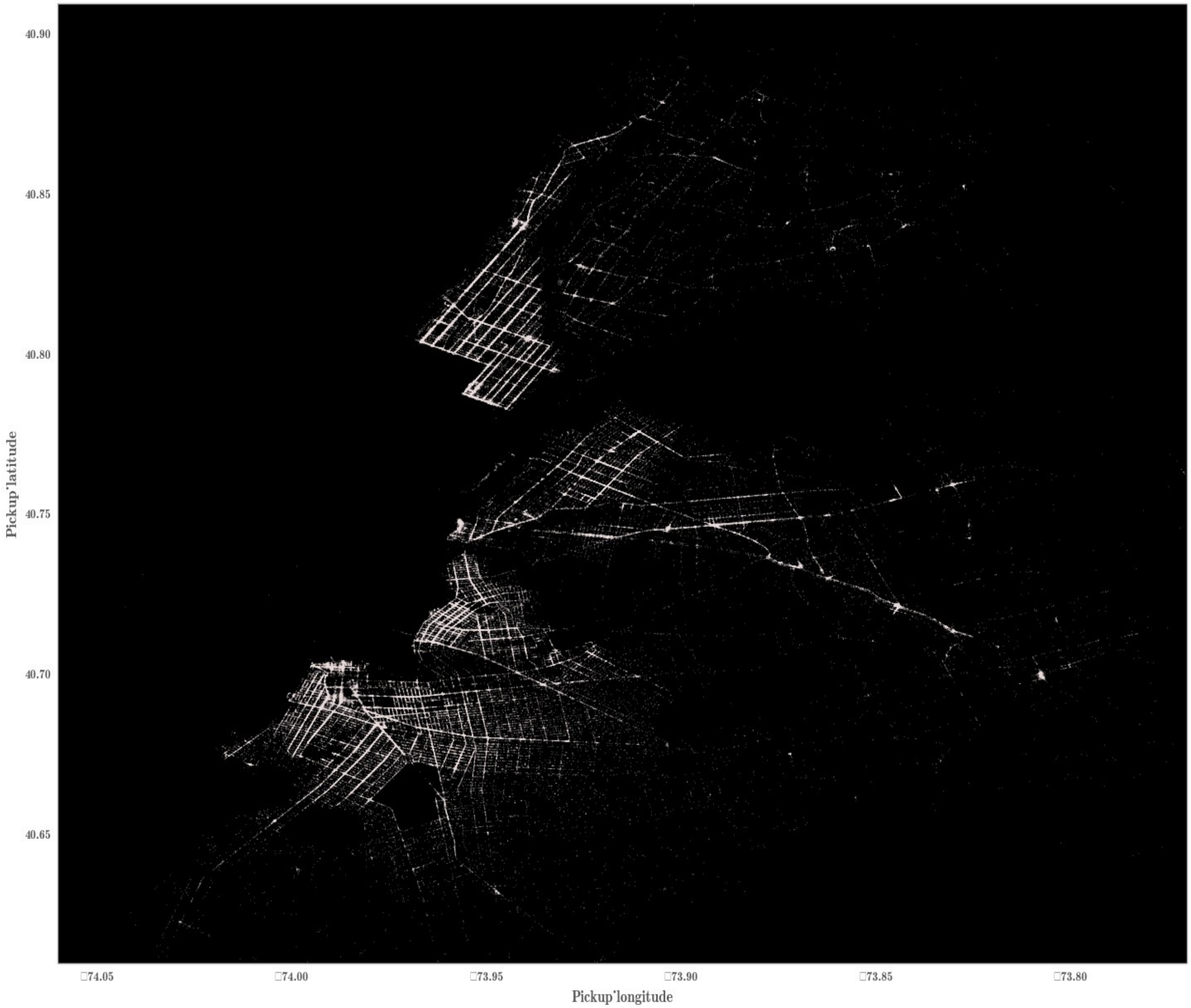
Mapping the data

Pickup



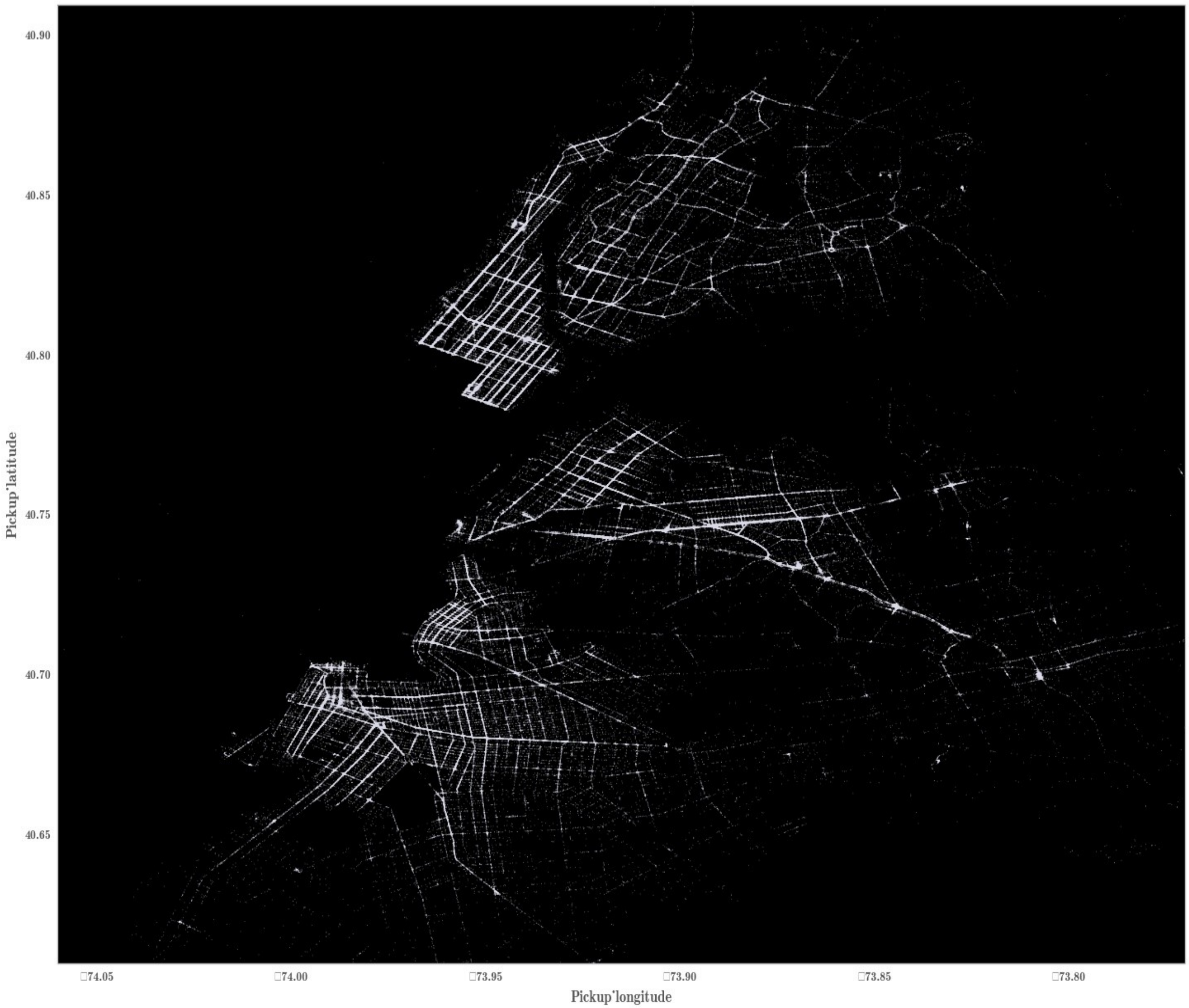
Mapping the data: tip>mean

Pickup



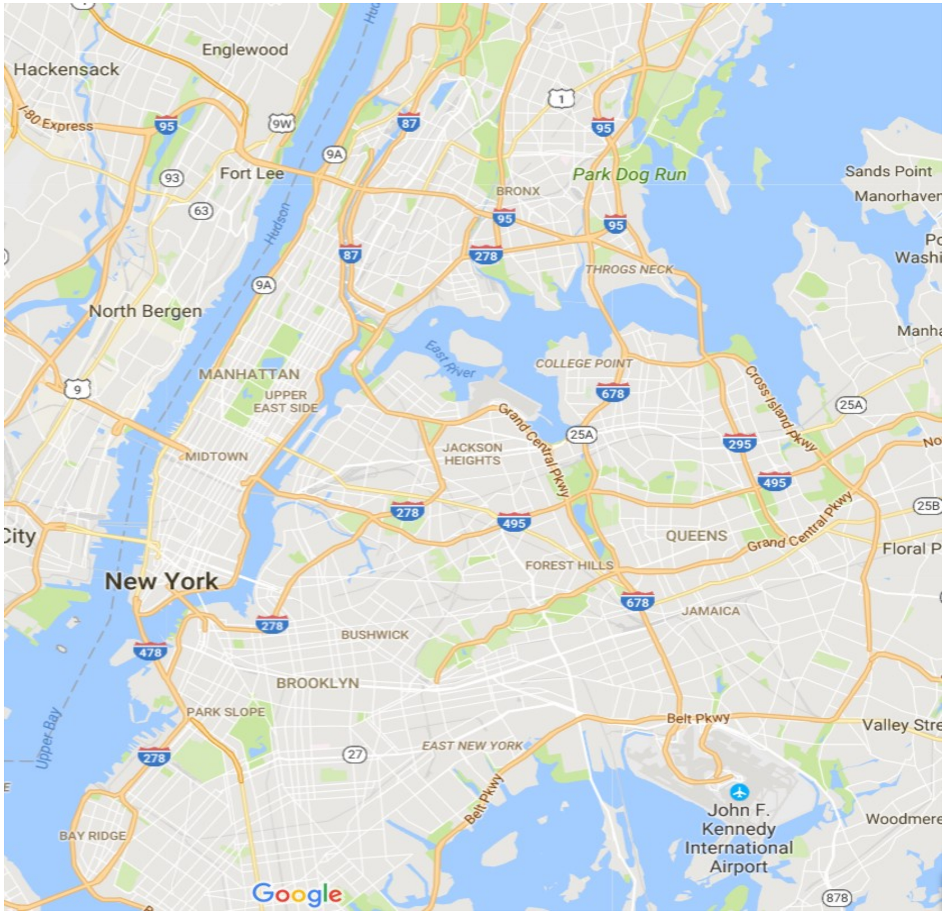
Mapping the data: tip<mean

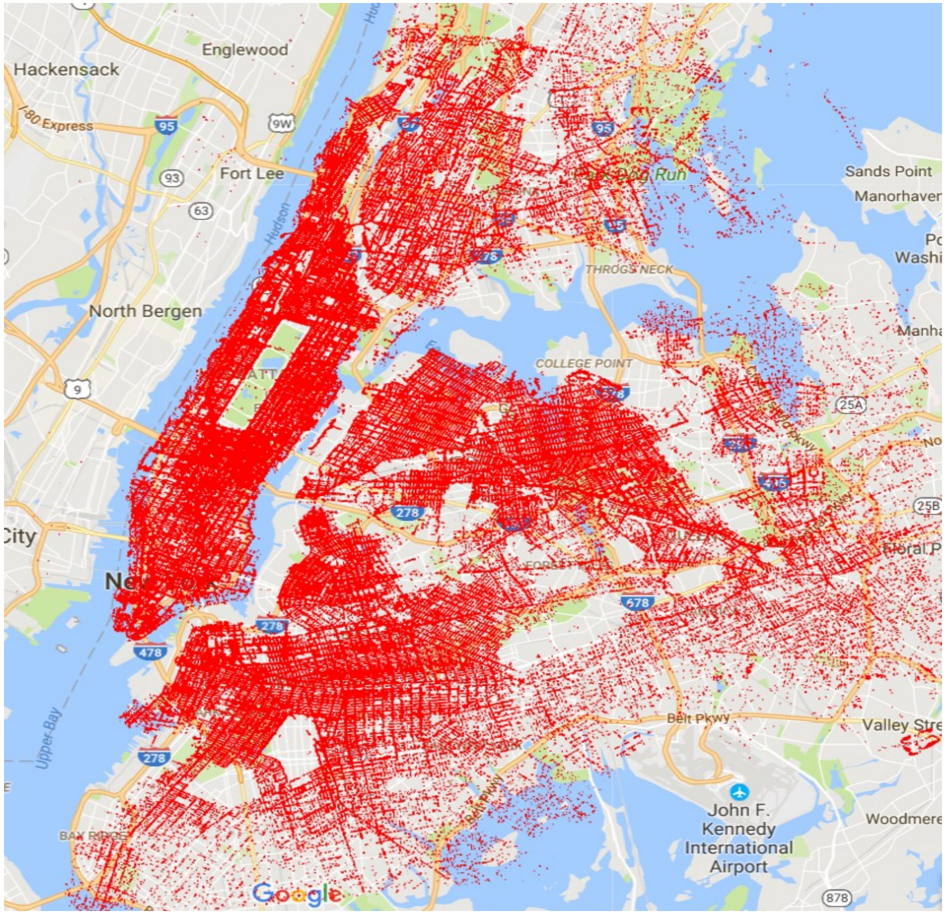
Pickup



Mapping the data

Dropoff





Classification model

- Classify a transaction with and without tip
- Target = with_tip
- Predictors = Payment_type, Total_amount, Trip_duration, Speed, MTA_tax, Extra, Hour, Direction_NS, Direction_EW, with_rain
- Training sample size: 10000
- 5-fold cross validation
- Optimized number of trees: 150
-

Classification model

Model report:
Accuracy: 0.9937
AUC Score (Train): 0.999914898879
CV Score - Mean : 0.9959447 | Std : 0.001198673 | Min : 0.9941585 | Max : 0.997252