Stream Processing

Project presentation

2020/2021

Table of Contents

Data - Taxi trips from New York City

Data

- Provided data consists of reports of taxi trips including starting point, drop-off point, corresponding timestamps, and information related to the payment.
- Data are reported at the end of the trip, i.e., upon arrive in the order of the drop-off timestamps.
- Events with the same dropoff_datetime are in random order.
- Quality of the data is not perfect.
  - Some events might miss information such as drop off and pickup coordinates or fare information.
  - Moreover, some information, such as, e.g., the fare price might have been entered incorrectly by the taxi drivers thus introducing additional skew.

Taxi trips from New York City

Attributes	Description
medallion	an md5sum of the identifier of the taxi - vehicle bound
hack_license	an md5sum of the identifier for the taxi license
pickup_datetime	time when the passenger(s) were picked up
dropoff_datetime	time when the passenger(s) were dropped off
trip_time_in_secs	duration of the trip
trip_distance	trip distance in miles

Taxi trips from New York City

Attributes	Description
pickup_longitude	longitude coordinate of the pickup location
pickup_latitude	latitude coordinate of the pickup location
dropoff_longitude	longitude coordinate of the drop-off location
dropoff_latitude	latitude coordinate of the drop-off location

Taxi trips from New York City

Attributes	Description
payment_type	the payment method - credit card or cash
fare_amount	fare amount in dollars
surcharge	surcharge in dollars
mta_tax	tax in dollars
tip_amount	tip in dollars
tolls_amount	bridge and tunnel tolls in dollars
total_amount	total paid amount in dollars

Taxi trips from New York City

medallion	an md5sum of the identifier of the taxi - vehicle bound
hack_license	an md5sum of the identifier for the taxi license
pickup_datetime	time when the passenger(s) were picked up
dropoff_datetime	time when the passenger(s) were dropped off
trip_time_in_secs	duration of the trip
trip_distance	trip distance in miles
pickup_longitude	longitude coordinate of the pickup location
pickup_latitude	latitude coordinate of the pickup location
dropoff_longitude	longitude coordinate of the drop-off location
dropoff_latitude	latitude coordinate of the drop-off location
payment_type	the payment method - credit card or cash
fare_amount	fare amount in dollars
surcharge	surcharge in dollars
mta_tax	tax in dollars
tip_amount	tip in dollars
tolls_amount	bridge and tunnel tolls in dollars
total_amount	total paid amount in dollars

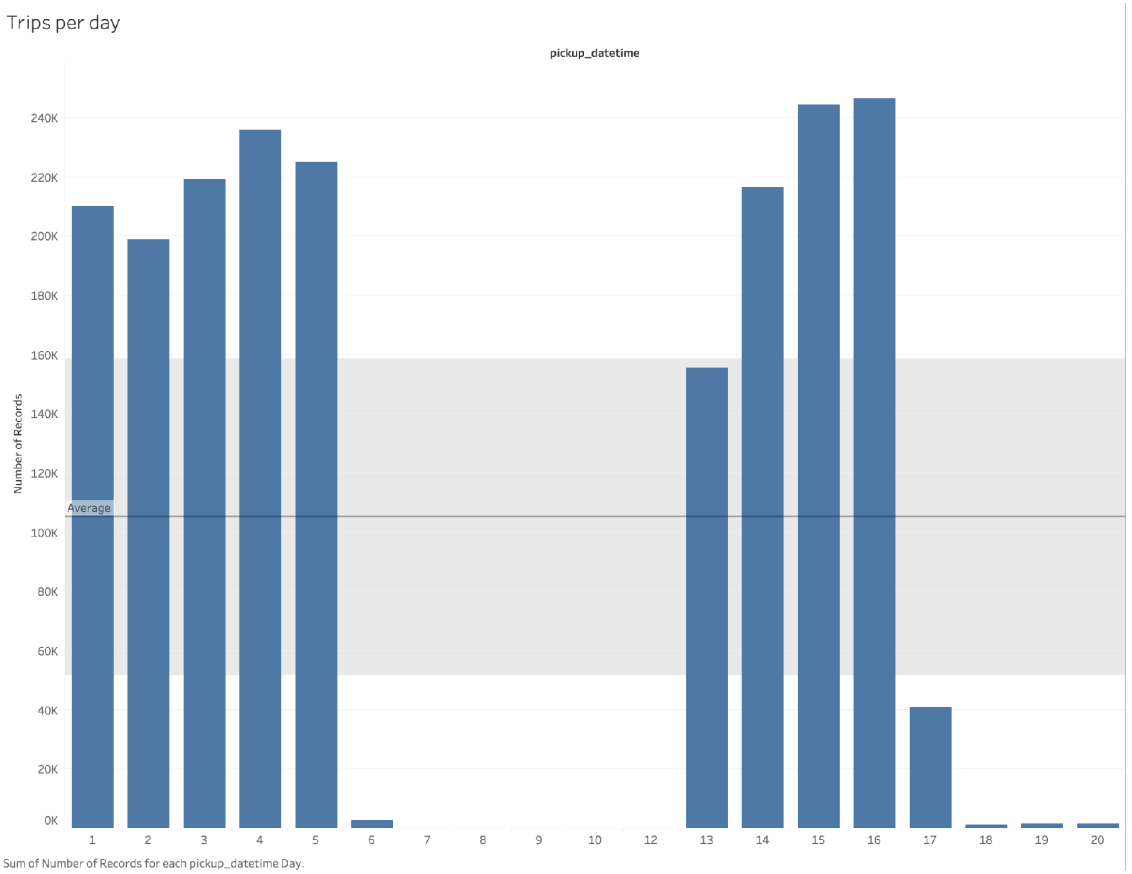
Where to get the data?

- ACM DEBS 2015 Grand Challenge
  - http://www.debs2015.org/call-grand-challenge.html

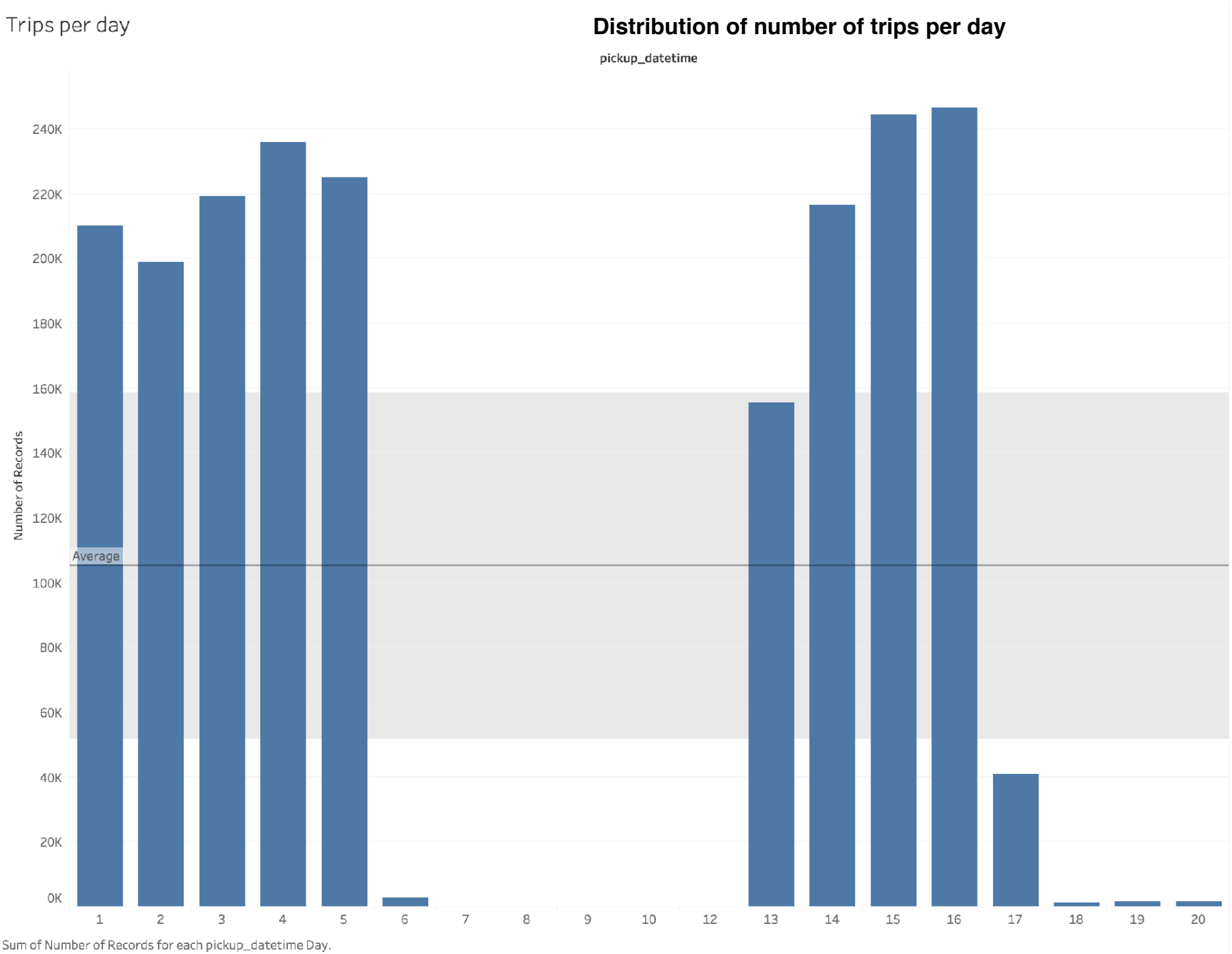
- 20 days (roughly 2 million events) of data (~130 MB)
- Data for the whole year 2013 (~173 million events) (~12 G) (~33 G expanded)

Preliminary information based on sample data

- ~10 800 Taxis
- ~20 300 Drivers
- 20 days
- ~2 million records (trips)

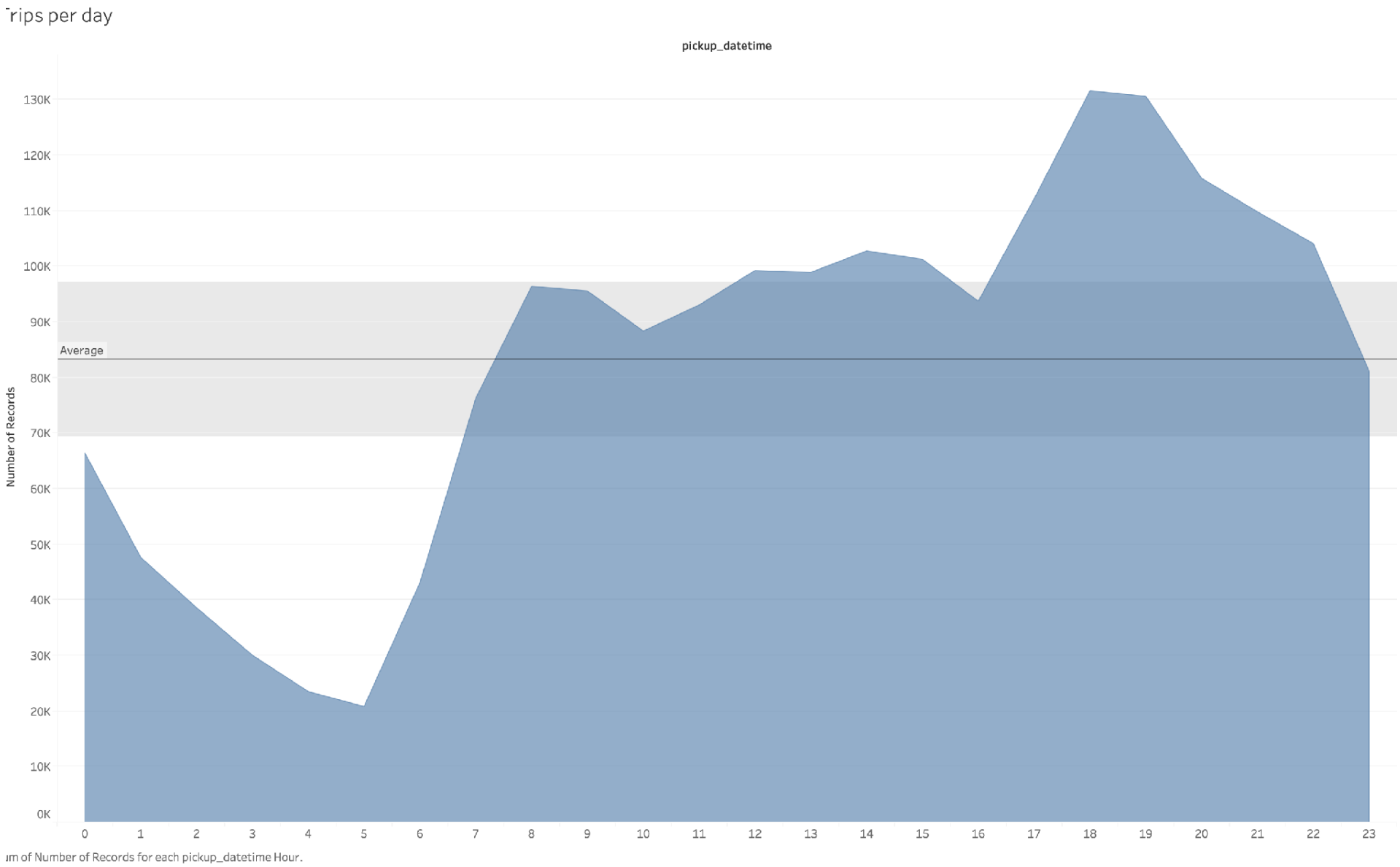


Trips per day

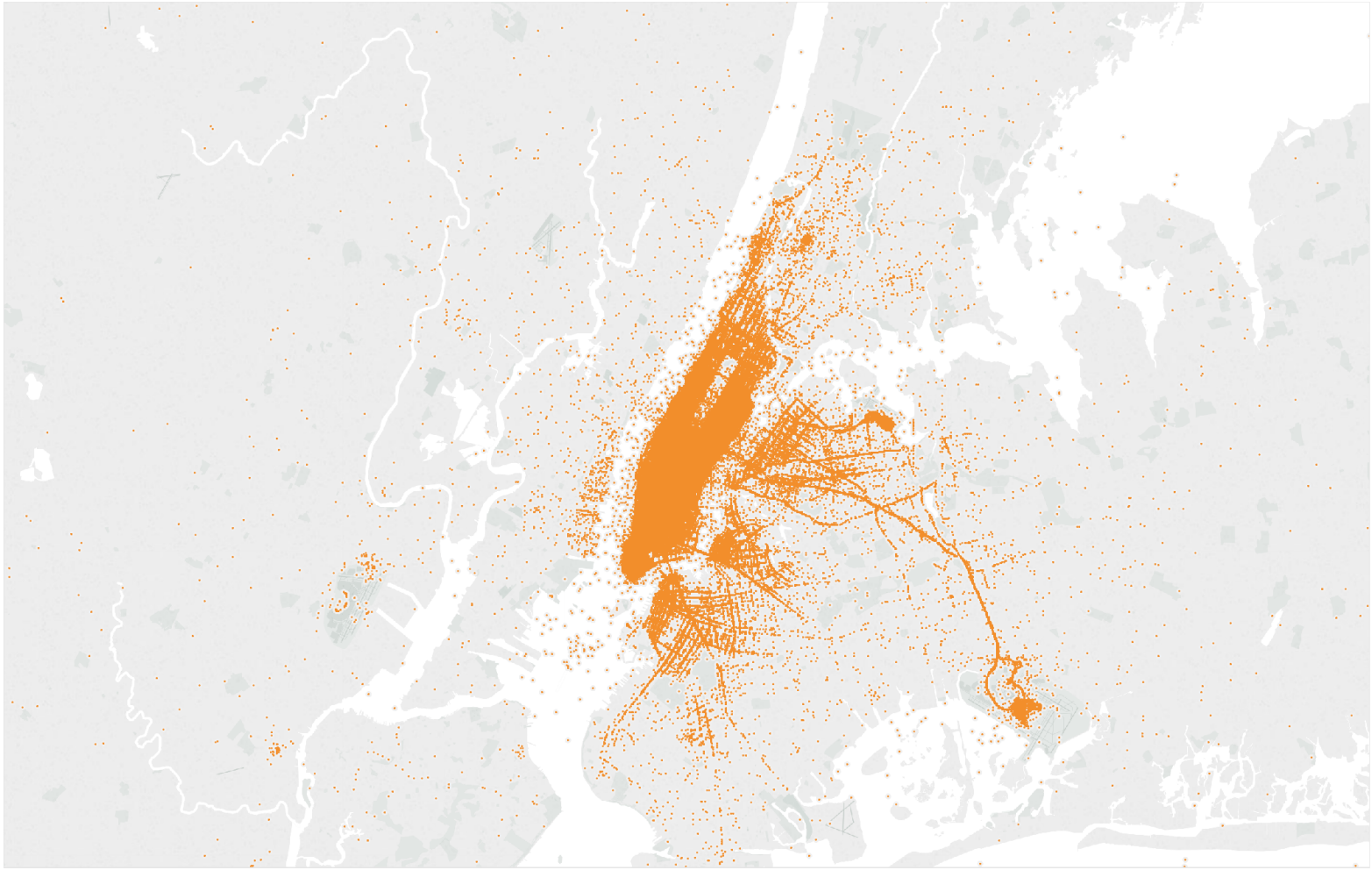


Sum of Number of Records for each pickup_datetime Day.

Distribution of number of trips per hour

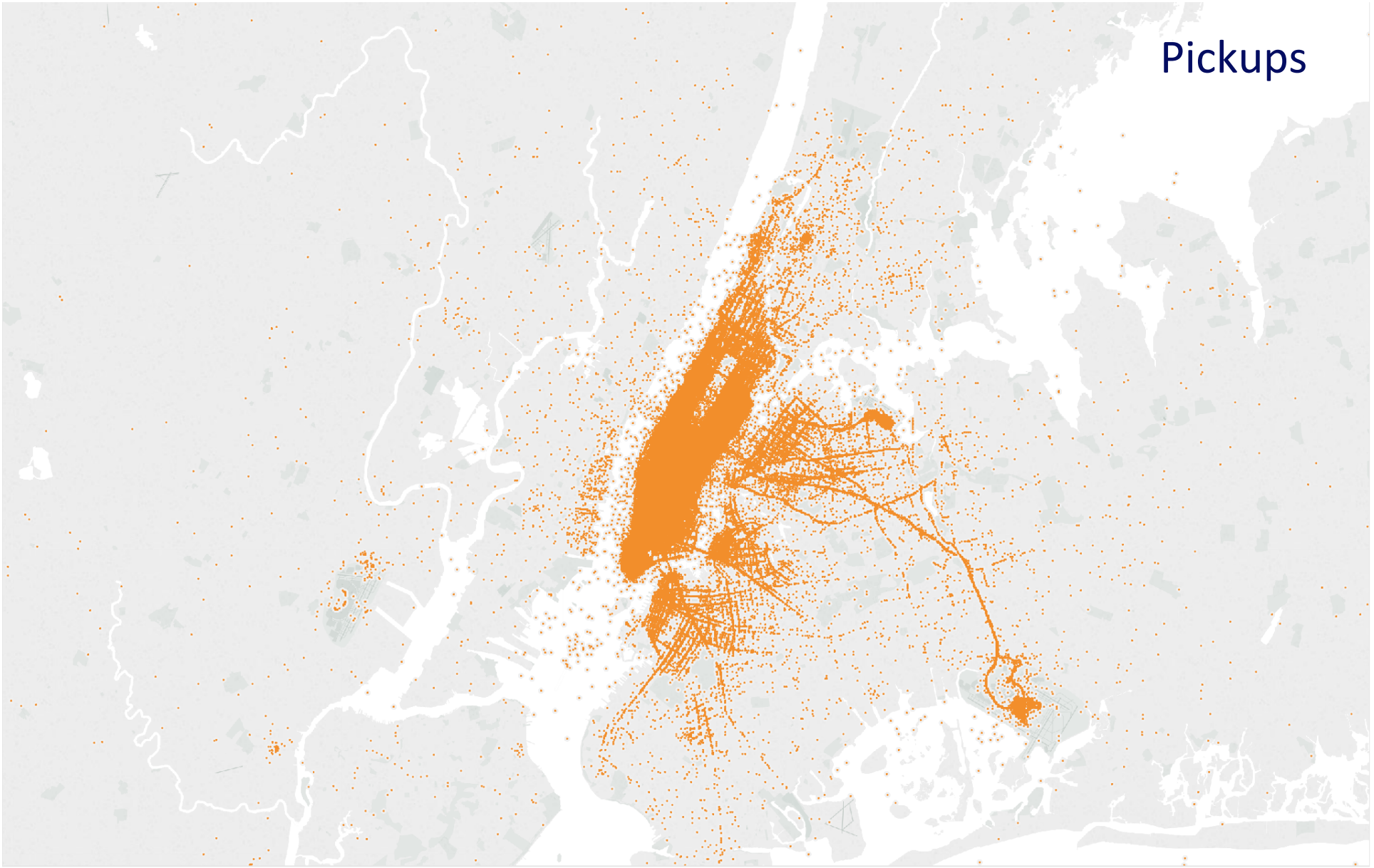


Data for 20 days: Pickups



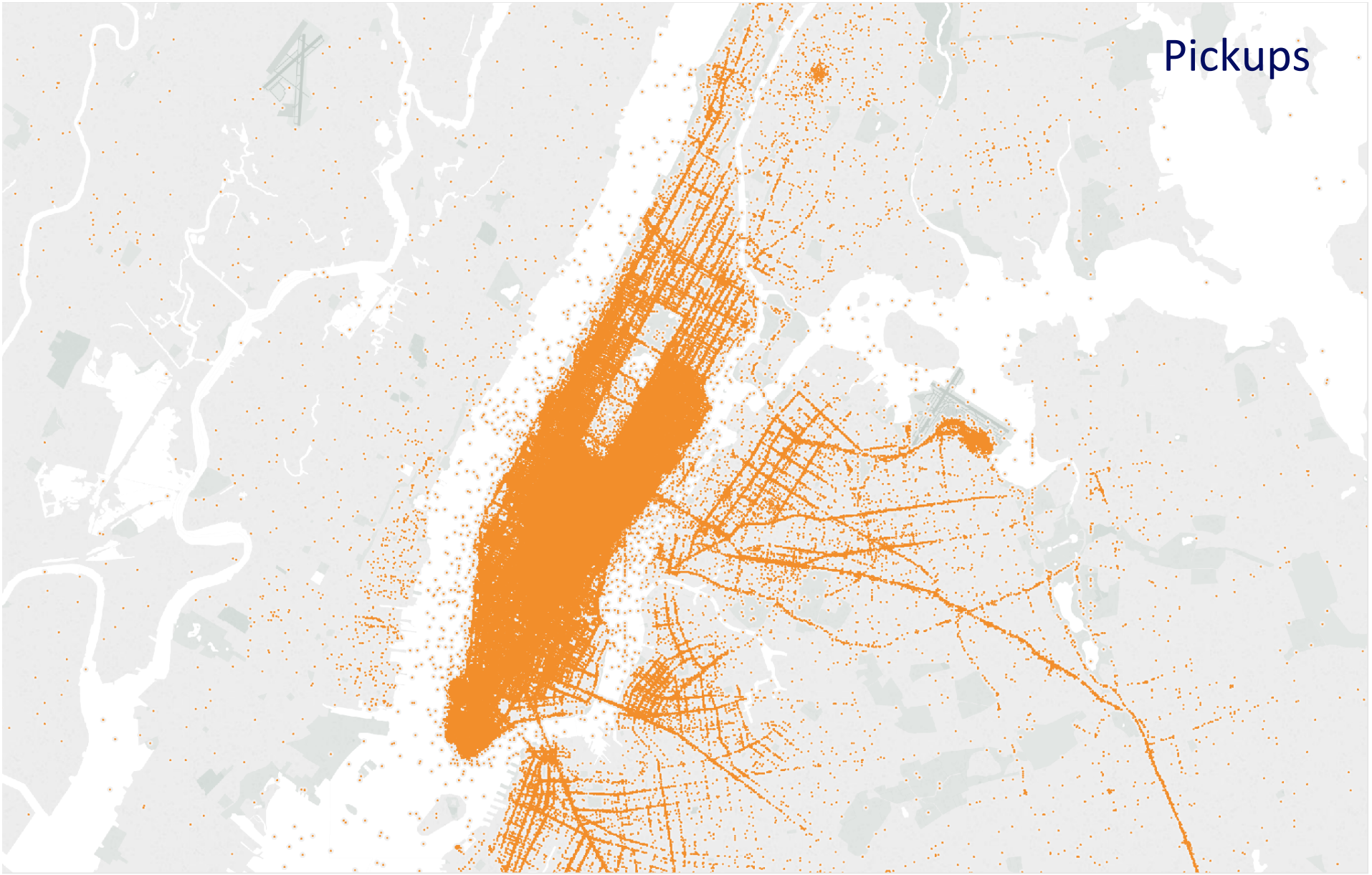
Map based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 01/01/2013 00:00:00 to 20/01/2013 23:59:27.

Pickups



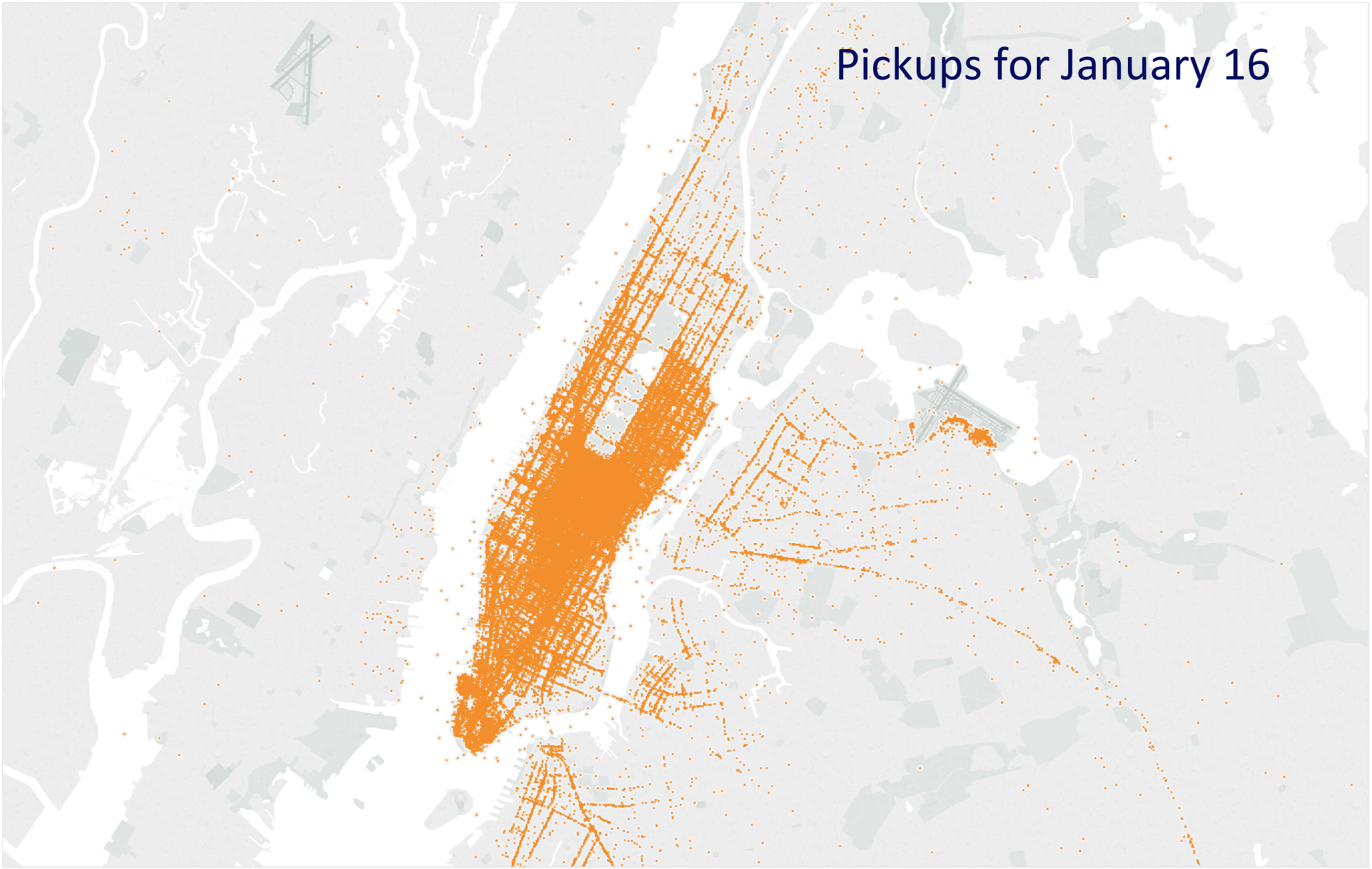
Map based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 01/01/2013 00:00:00 to 20/01/2013 23:59:27.

Pickups



ap based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 01/01/2013 00:00:00 to 20/01/2013 23:59:27.

Pickups for January 16



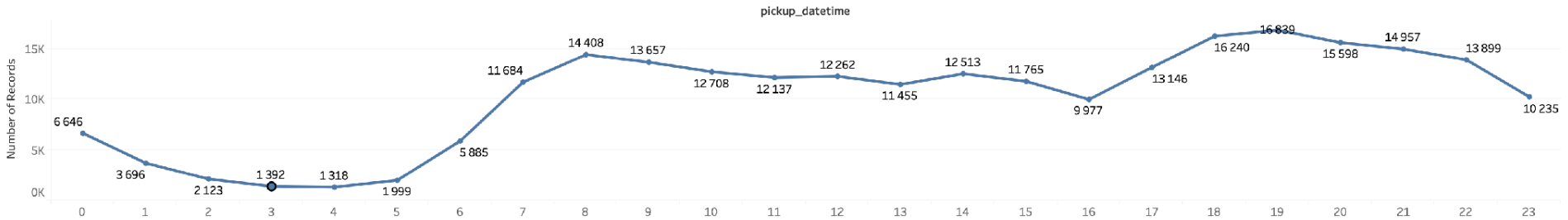
Map based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 16/01/2013 00:00:00 to 16/01/2013 23:59:27.

Pickups hour evolution during the hour 3 of day

pickup_datetime
16/01/2013 00:00:00 to ..

Hour of pickup_datetime
3
Show history

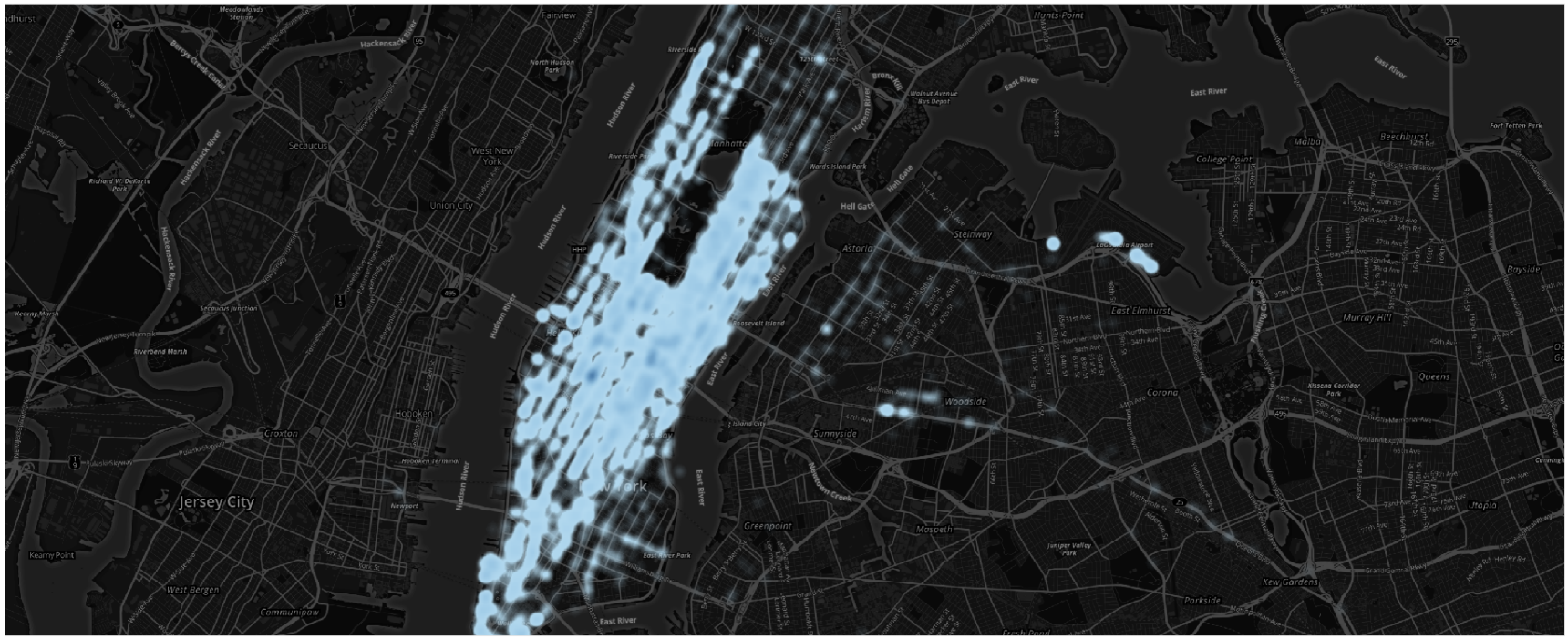
Hour Evolution a day



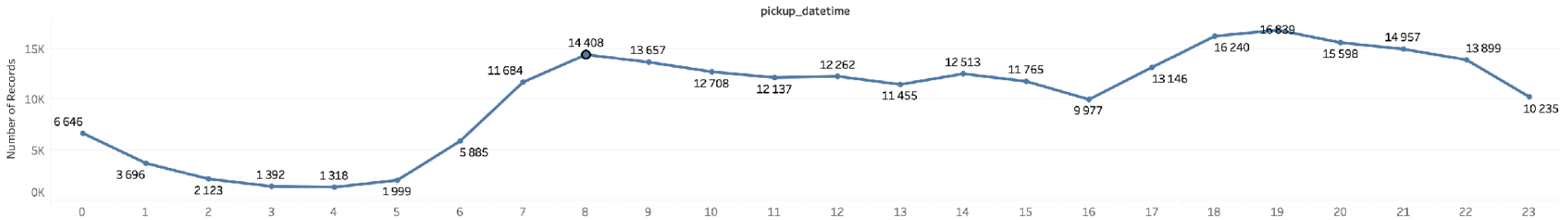
Pickups hour evolution during the hour 8 of day

pickup_datetime
16/01/2013 00:00:00 to ..

Hour of pickup_datetime
8
Show history



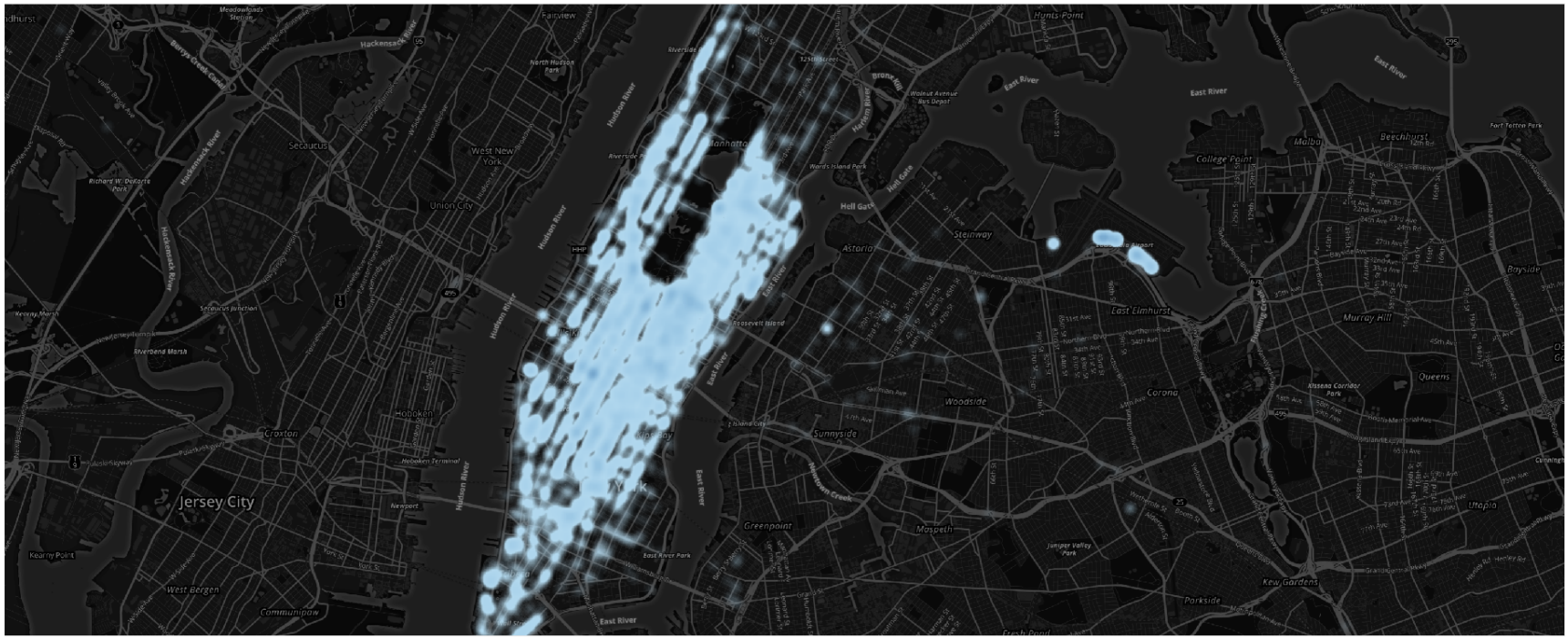
Hour Evolution a day



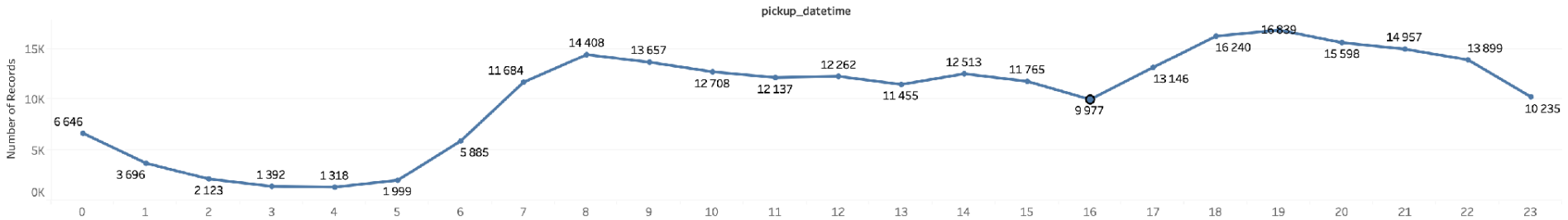
Pickups hour evolution during the hour **16** of day

pickup_datetime
16/01/2013 00:00:00 to ..

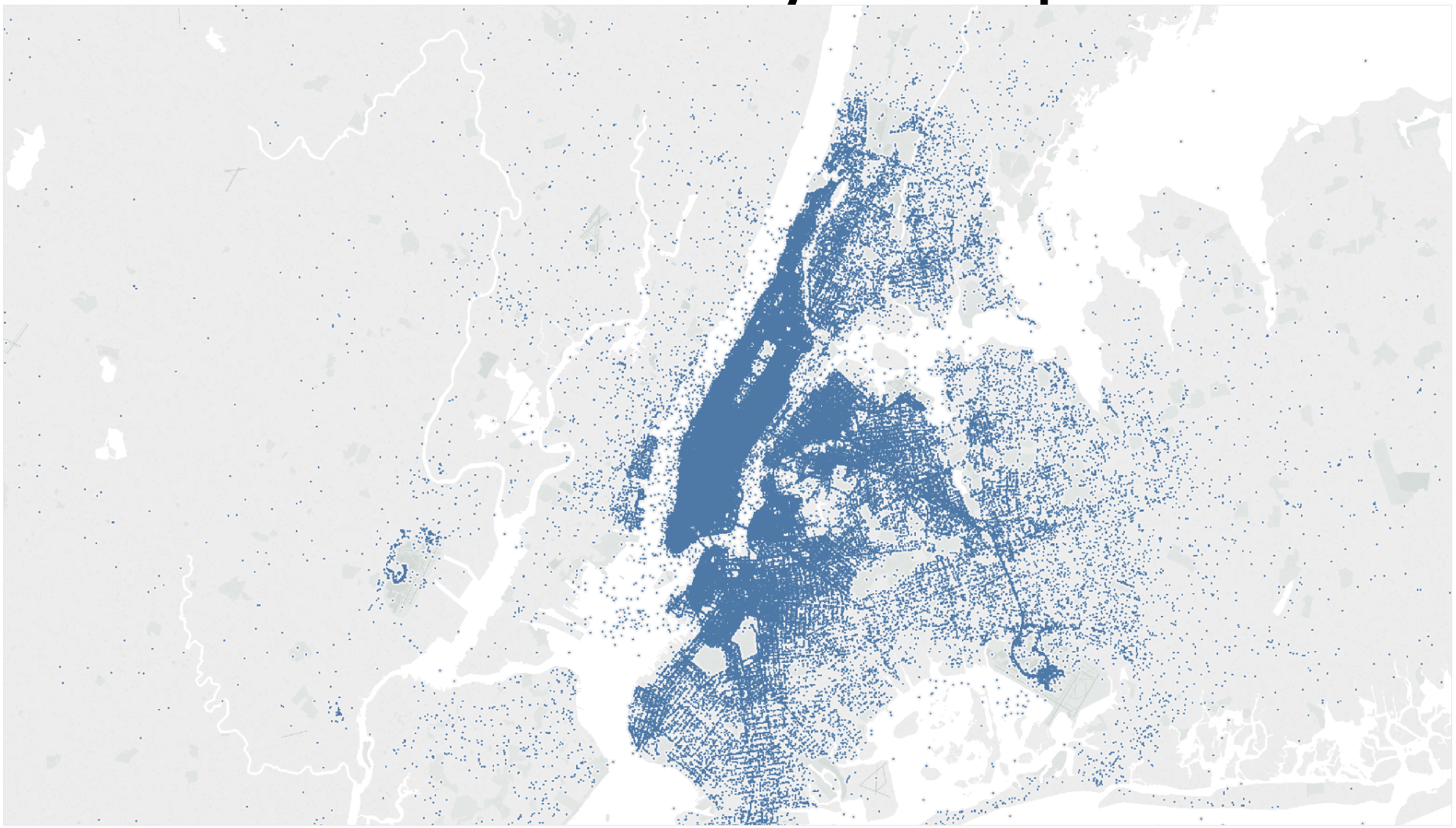
Hour of pickup_datetime
16
Show history



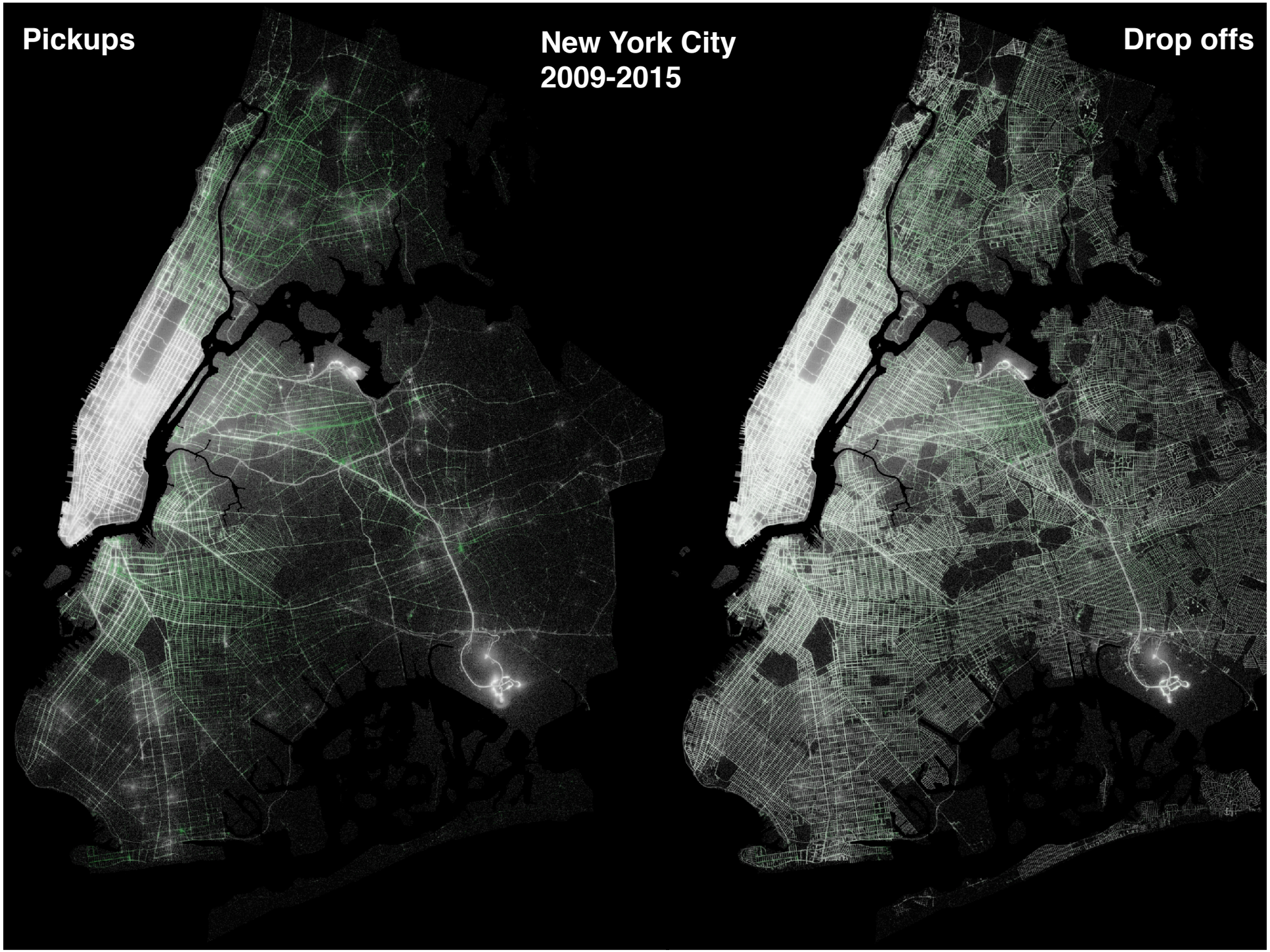
Hour Evolution a day



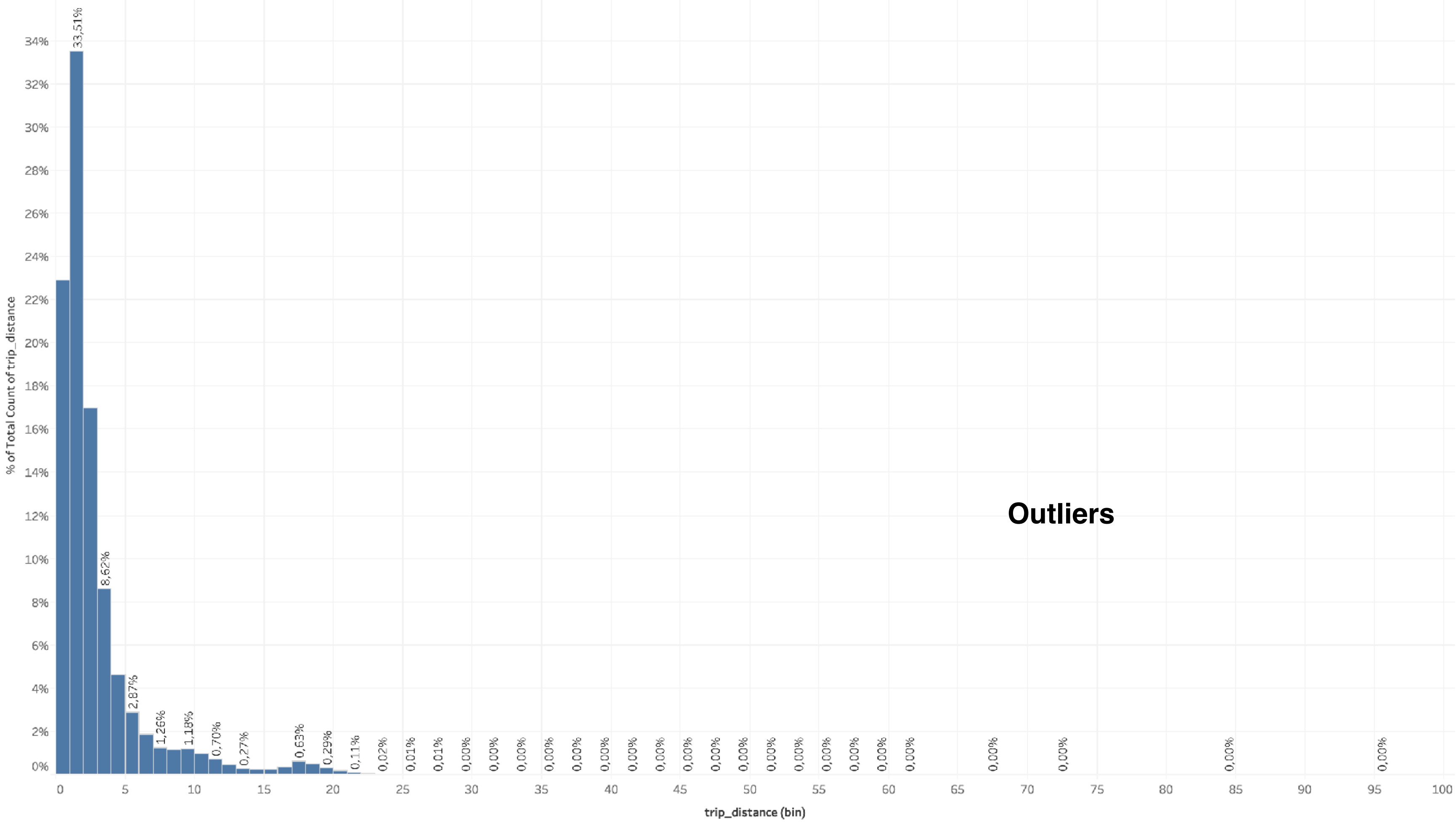
Data for 20 days: Drop-off



Map based on dropoff_longitude and dropoff_latitude.

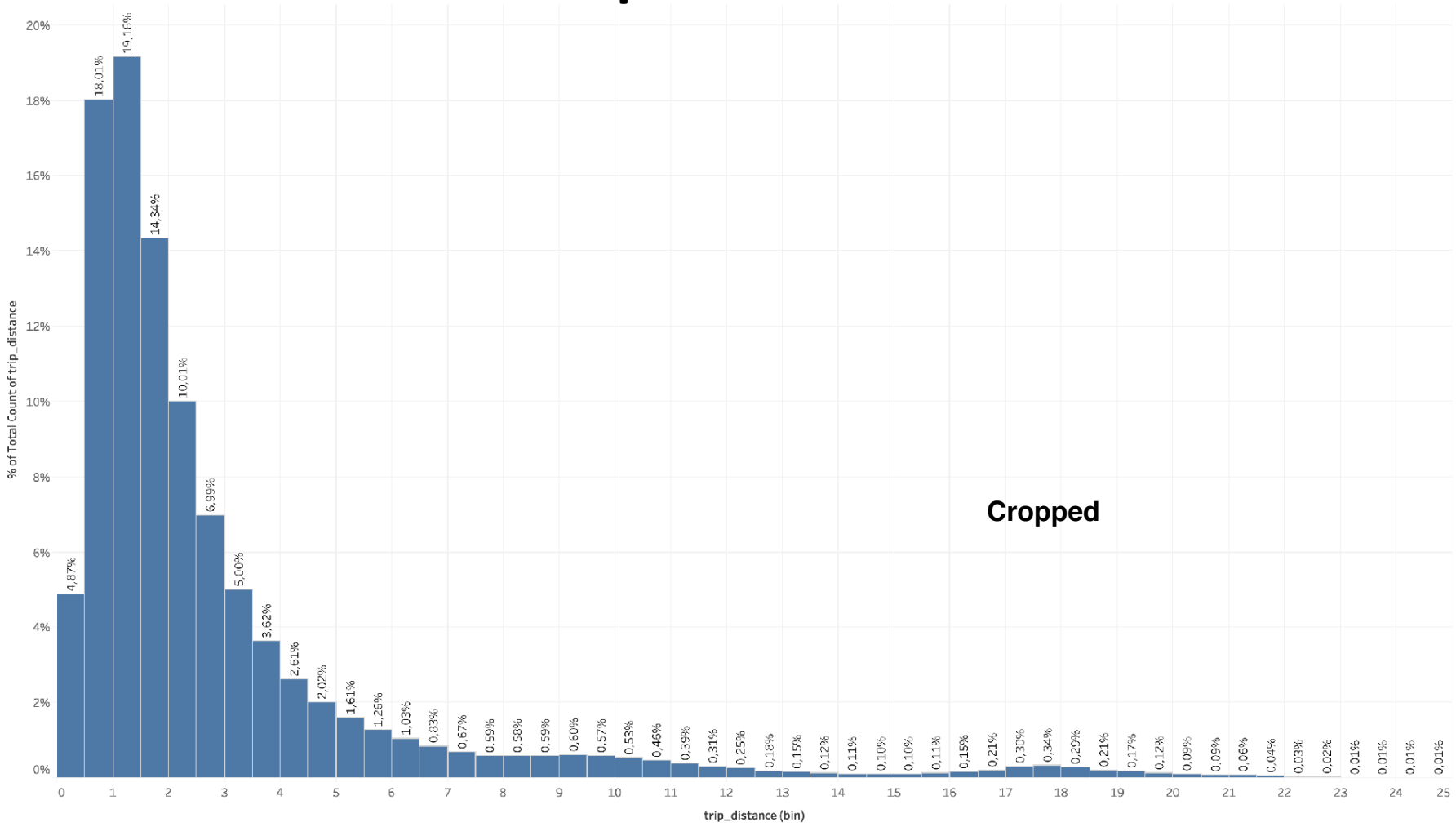

Pickups
New York City
2009-2015
Drop offs

Trip Distance



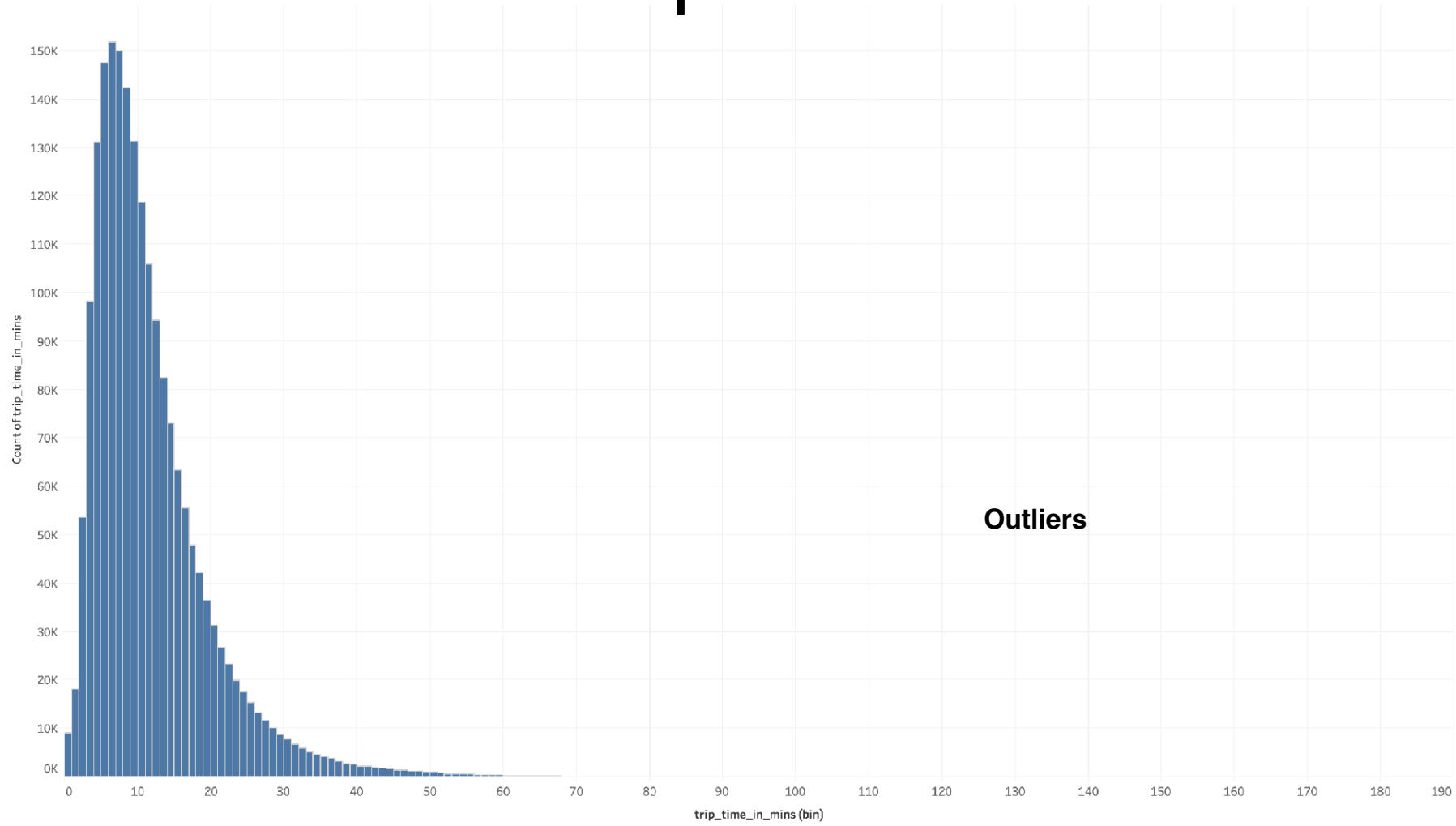
The trend of % of Total Count of trip_distance for trip_distance (bin).

Trip Distance



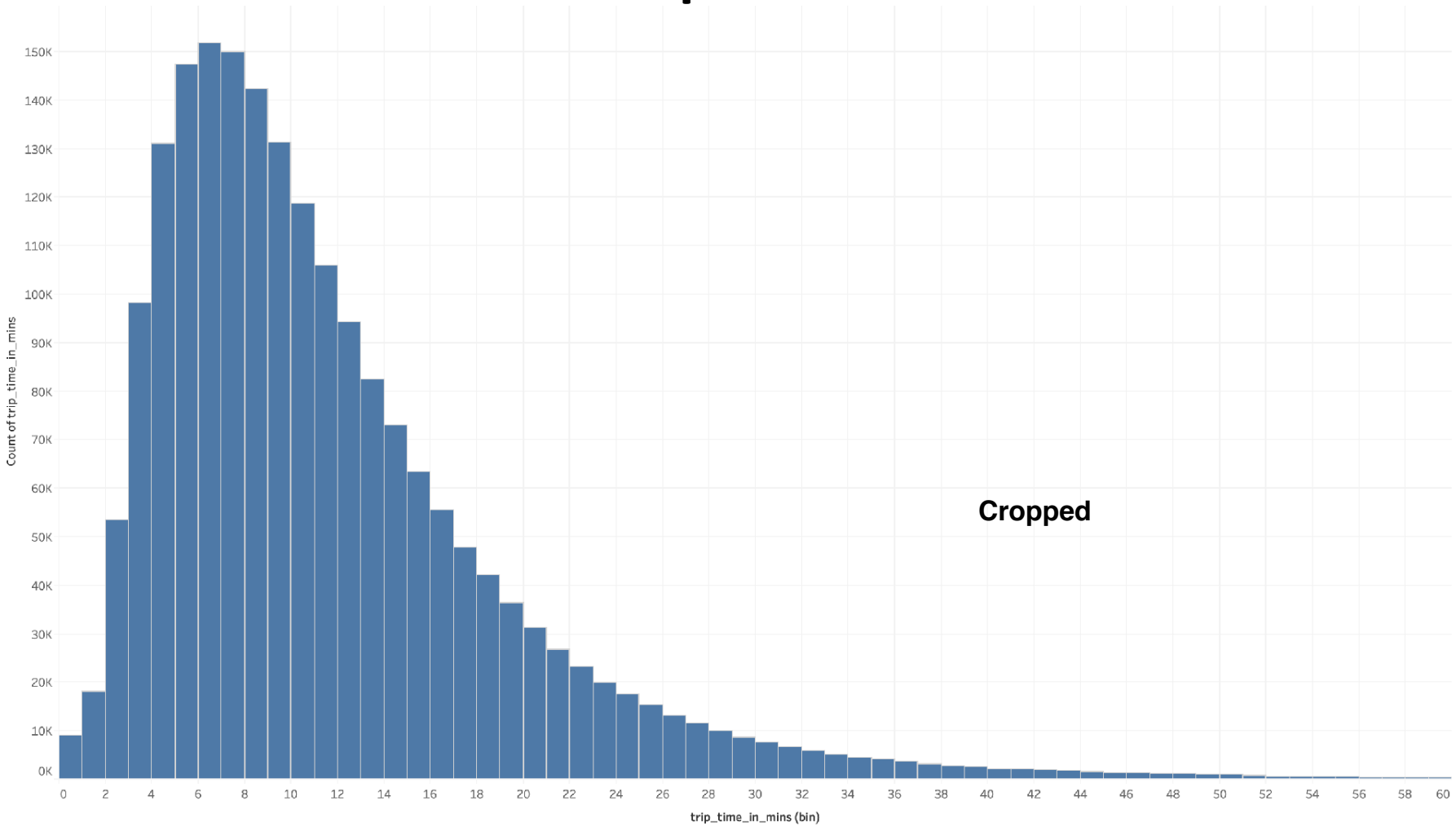
The trend of % of Total Count of trip_distance for trip_distance (bin).

Trip Time



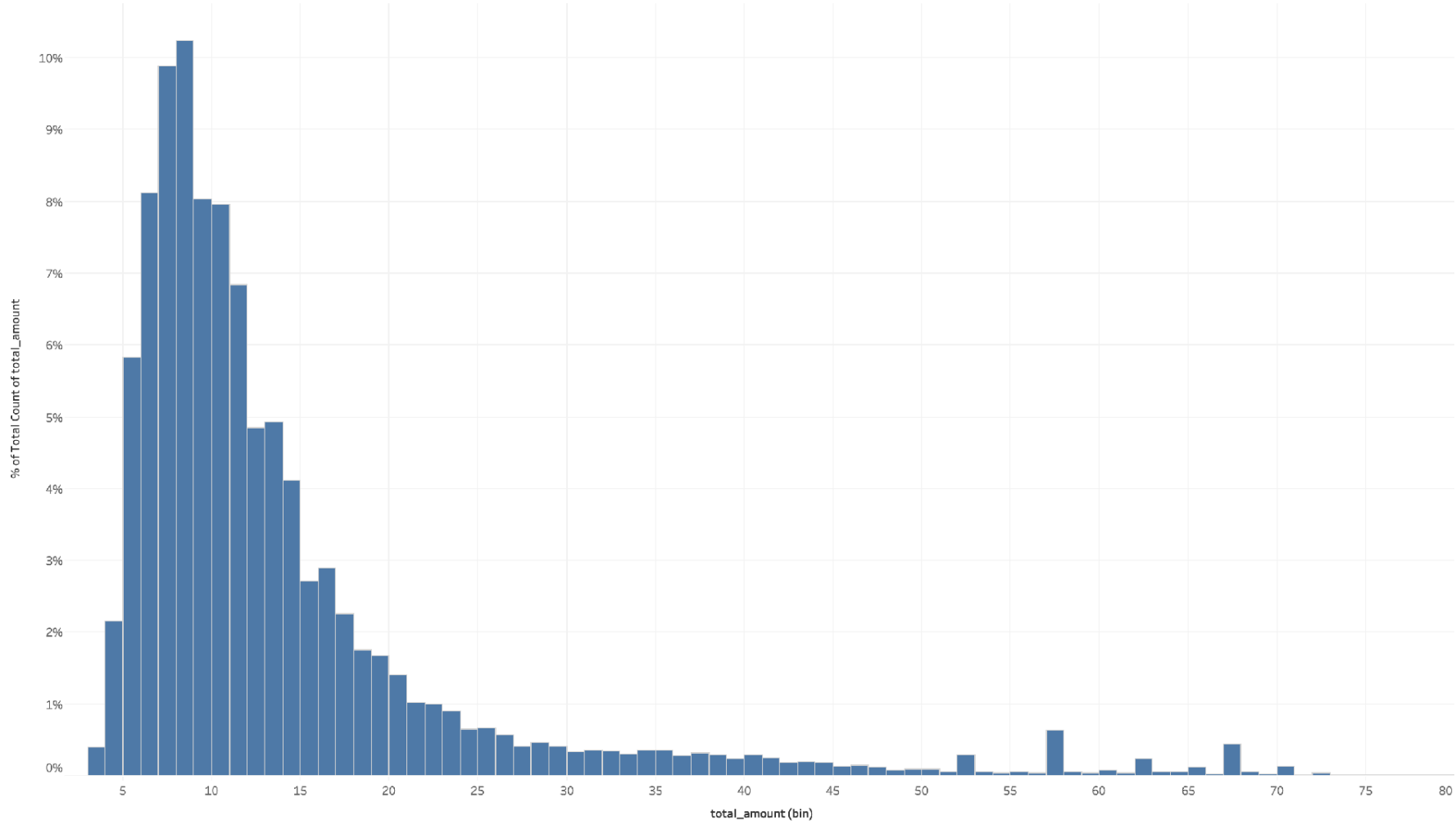
The trend of count of trip_time_in_mins for trip_time_in_mins (bin).

Trip Time



The trend of count of trip_time_in_mins for trip_time_in_mins (bin).

Trip Total Amount



The trend of % of Total Count of total_amount for total_amount (bin).

Full Assigment

- Full assignment text and supplemental code and scripts:

https://github.com/smduarte/ps2021