Project Presentation

Table of Contents

- **Data - Taxi trips from New York City**
- **Questions**
- **Recommendations**

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

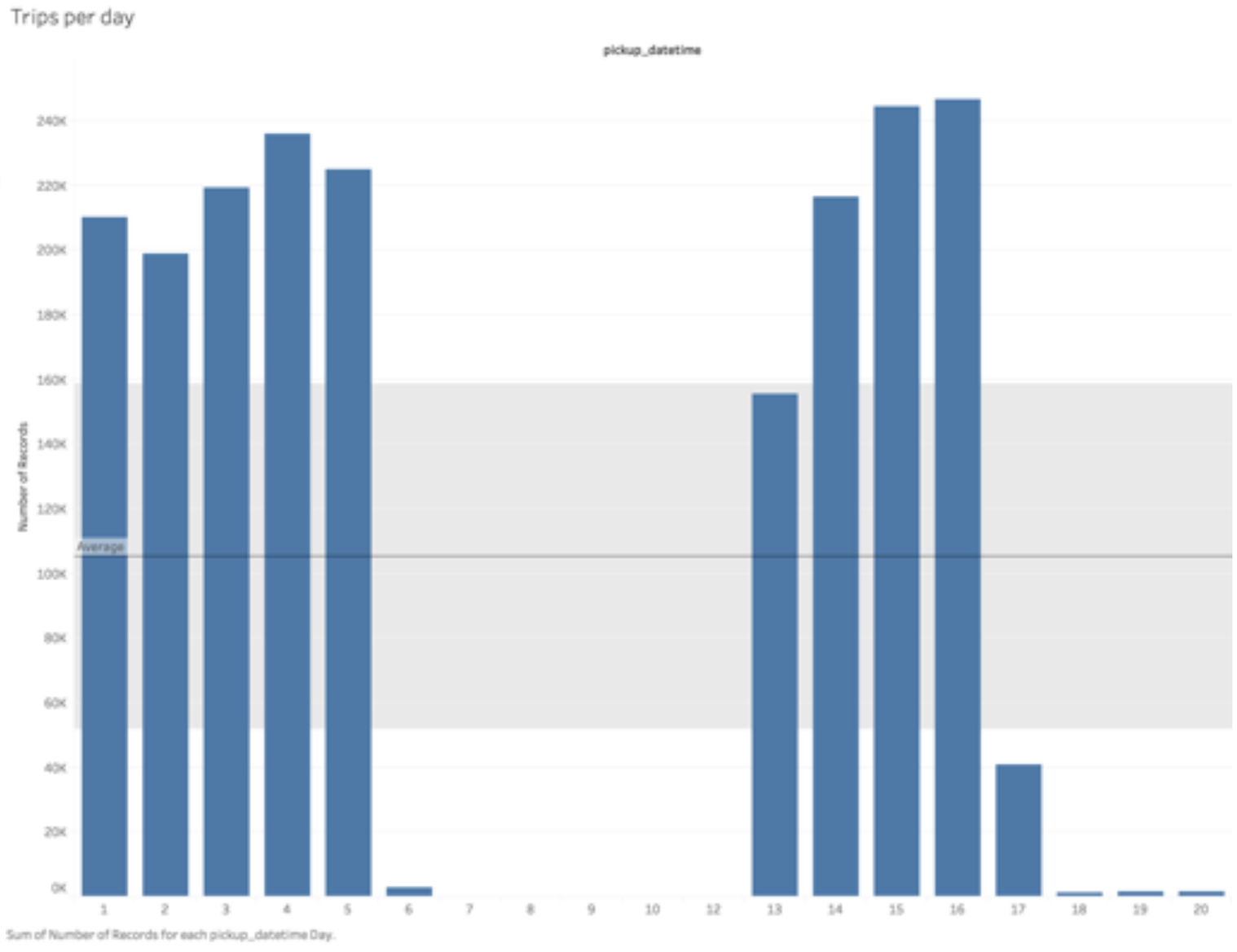
Where to get the data?

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

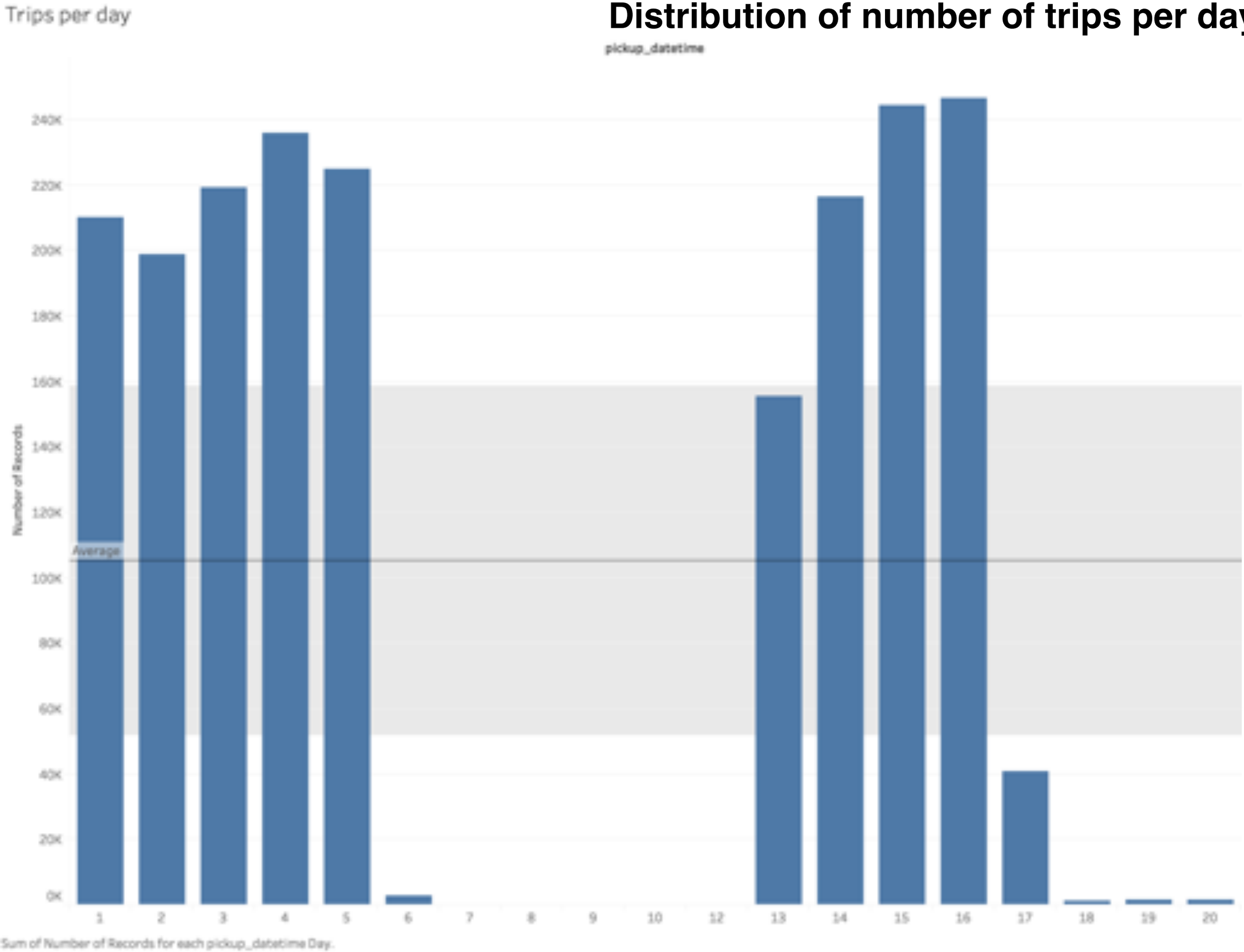
- **20 days (roughly 2 million events) of data (~130 MB)**
- **Data for the whole year 2013 (~173 million events) (~12 G) (~33 G expanded)**

Preliminary information based on sample data

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

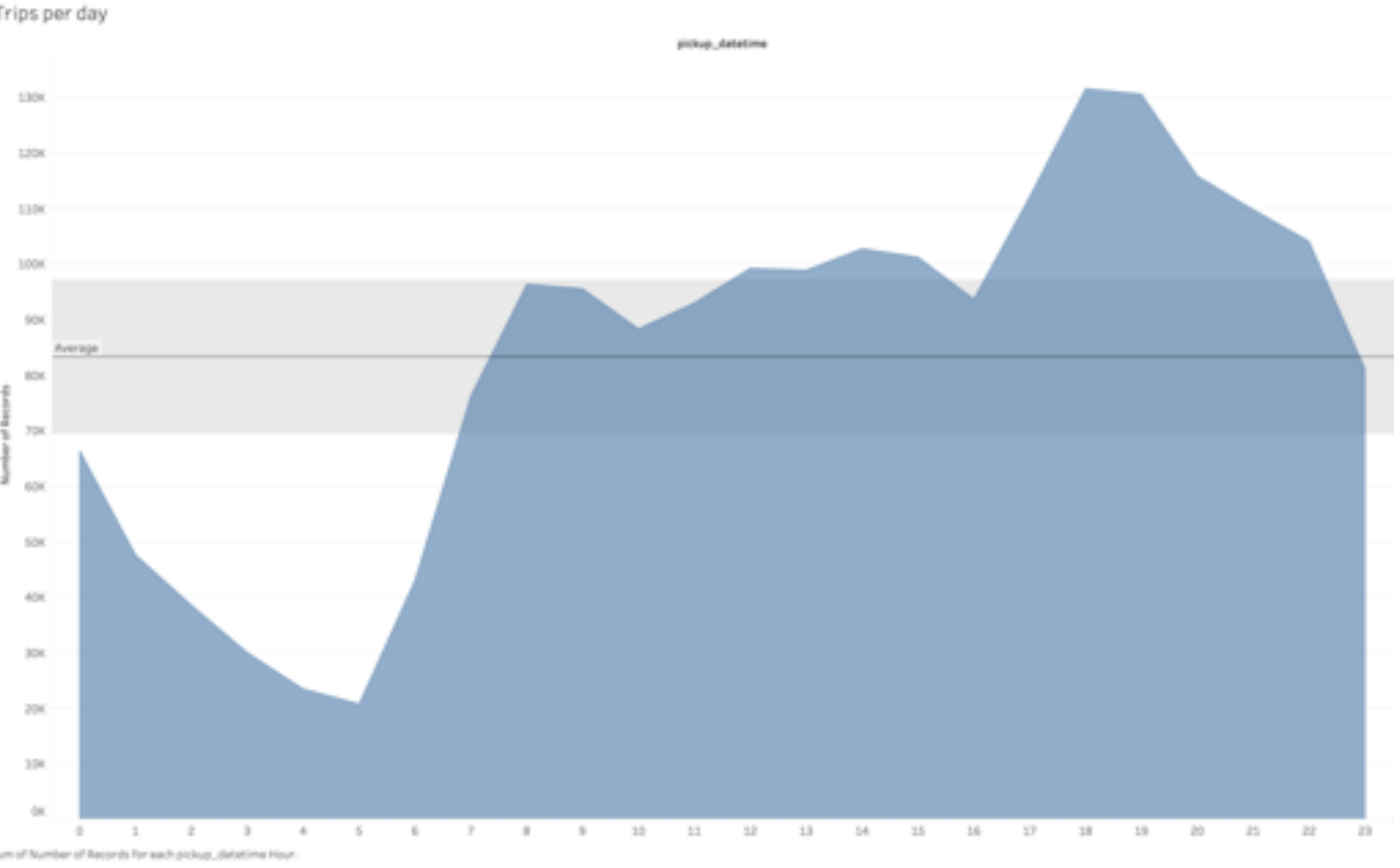


Distribution of number of trips per day

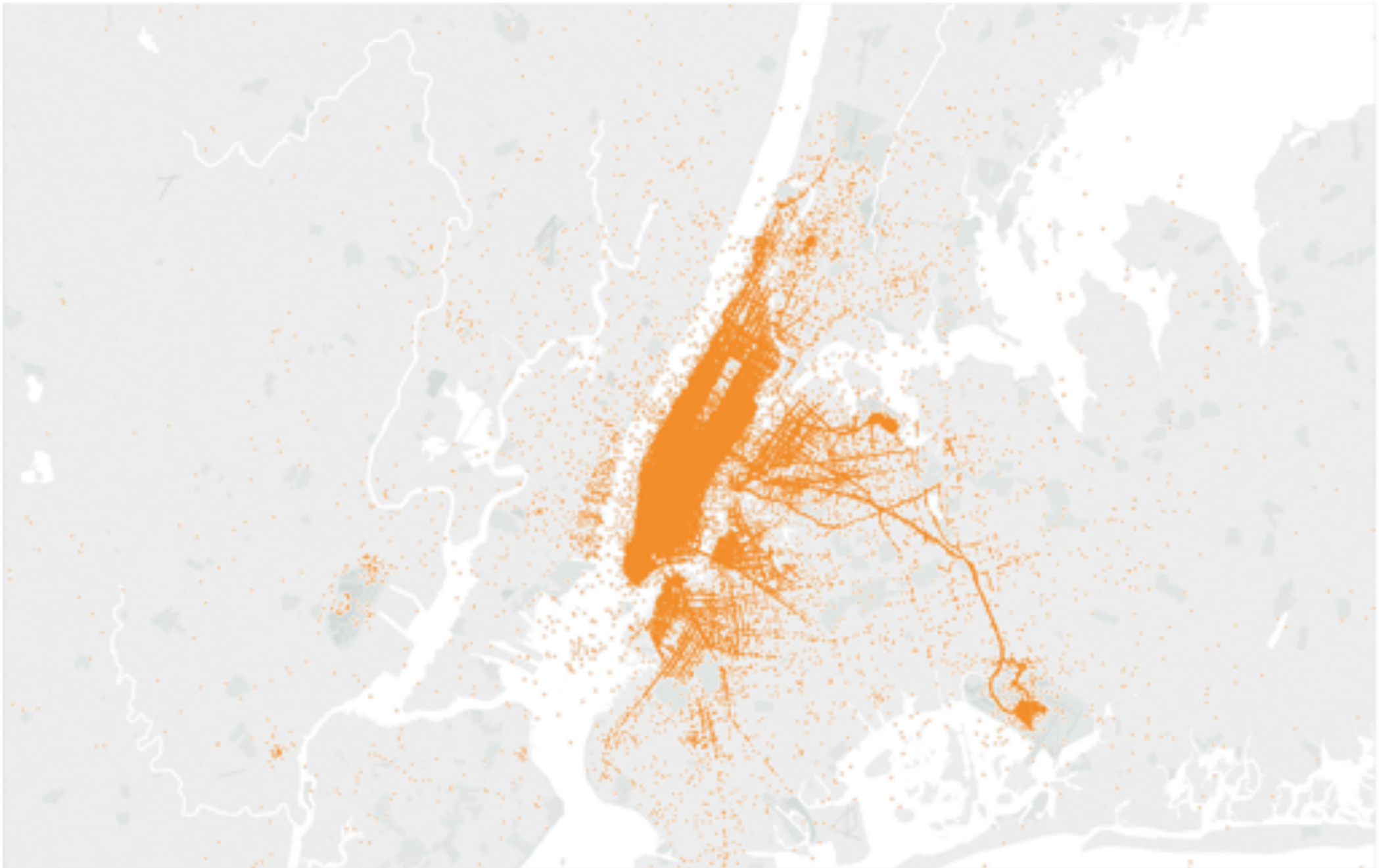


Sum of Number of Records for each pickup_datetime Day.

Distribution of number of trips per hour

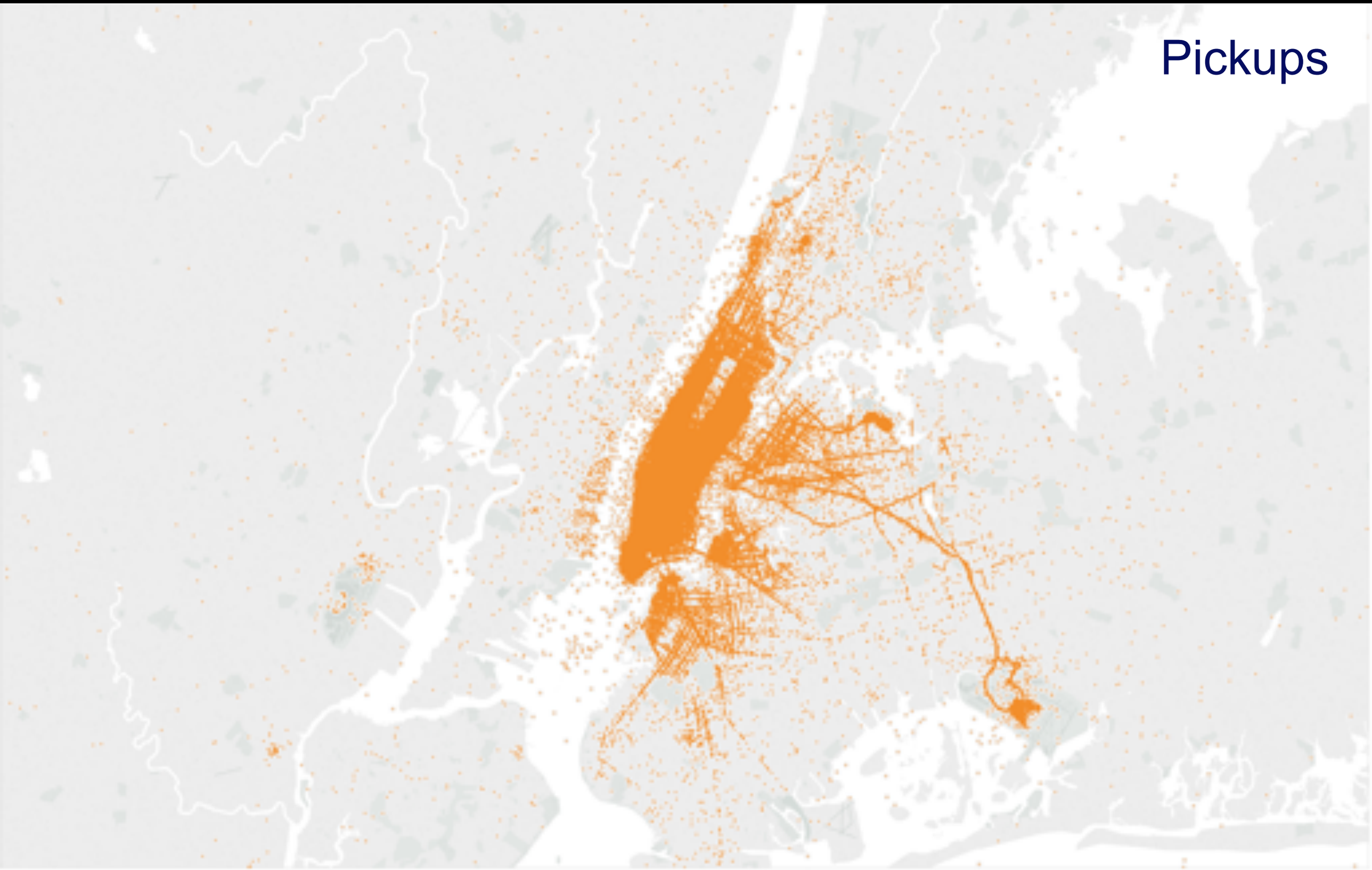


Data for 20 days: Pickups



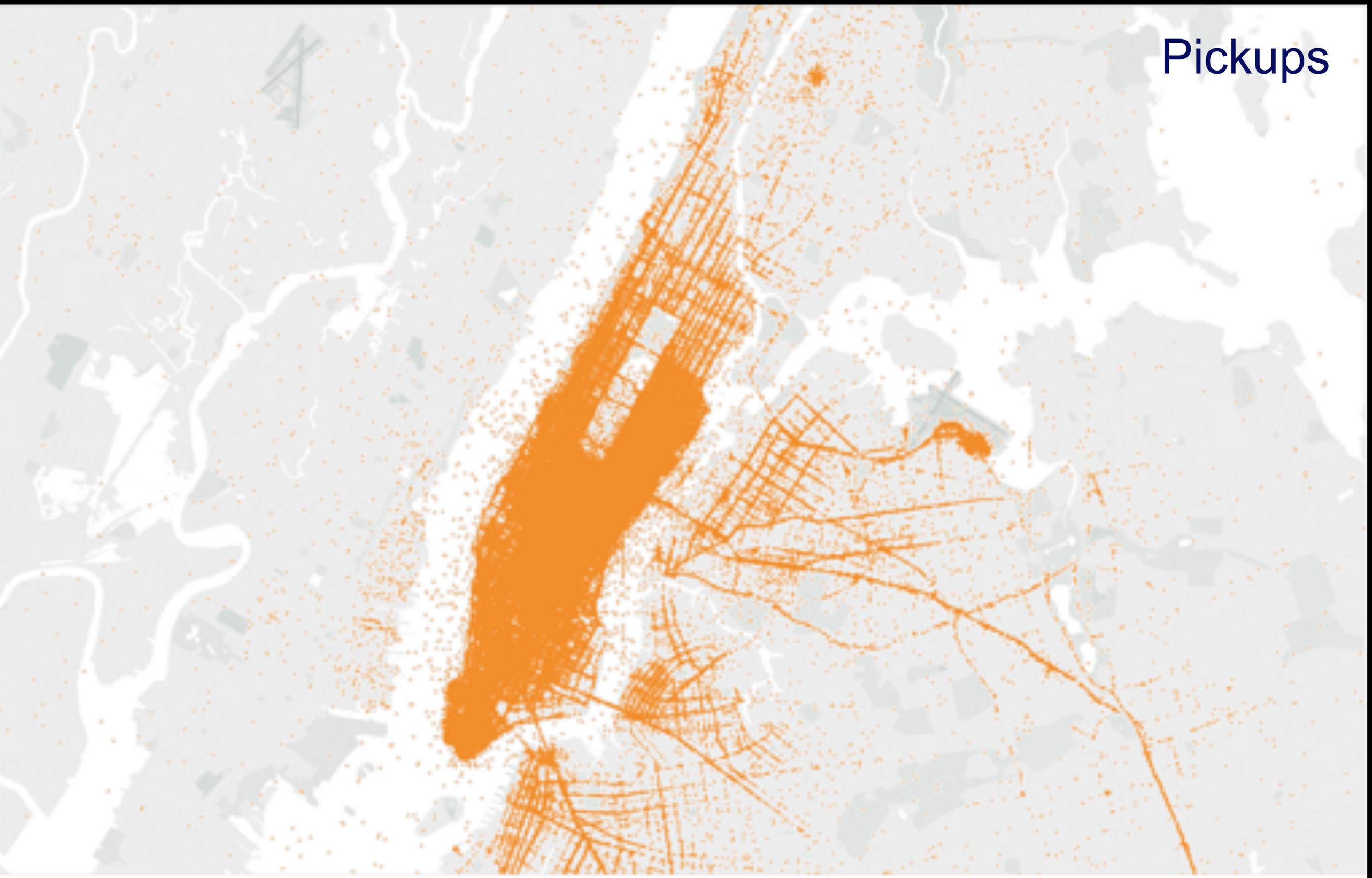
Map based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 01/01/2013 00:00:00 to 20/01/2013 23:59:27.

Pickups



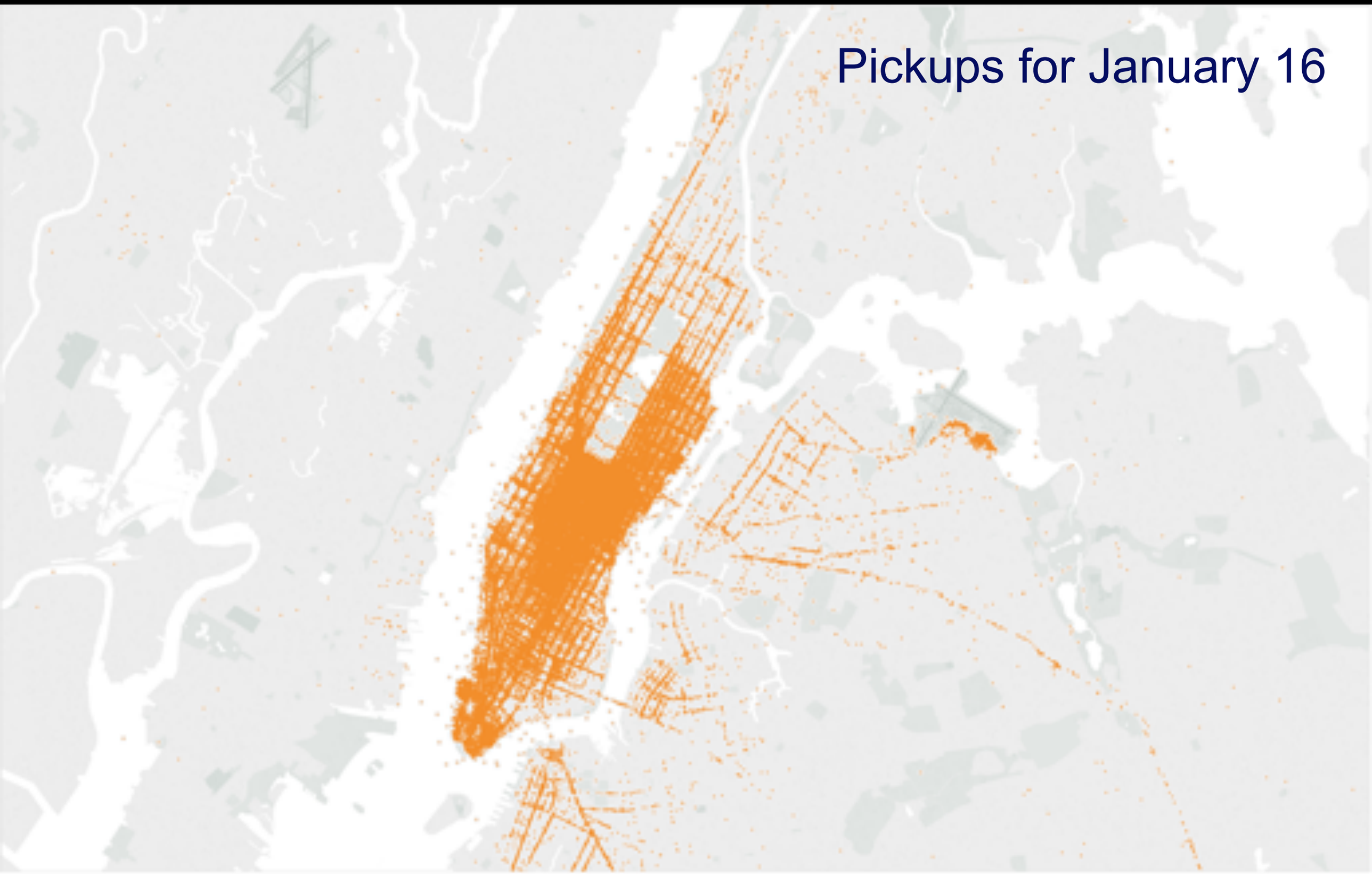
Map based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 01/01/2013 00:00:00 to 20/01/2013 23:59:27.

Pickups



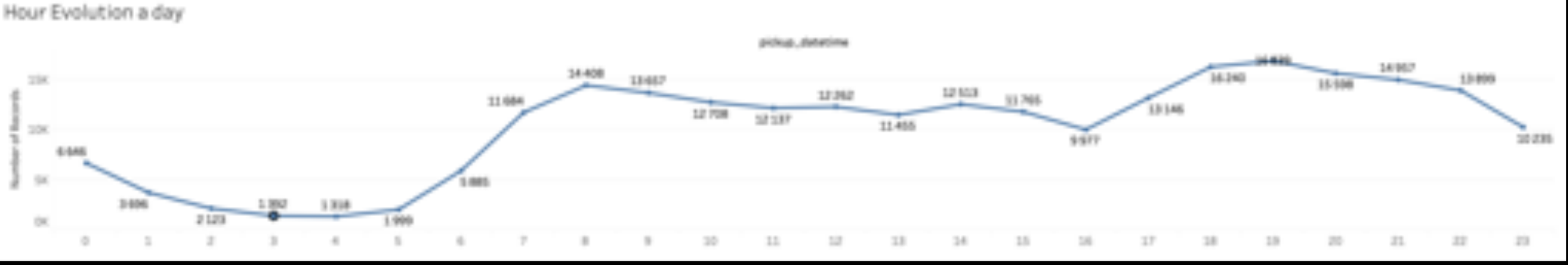
lap based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 01/01/2013 00:00:00 to 20/01/2013 23:59:27.

Pickups for January 16

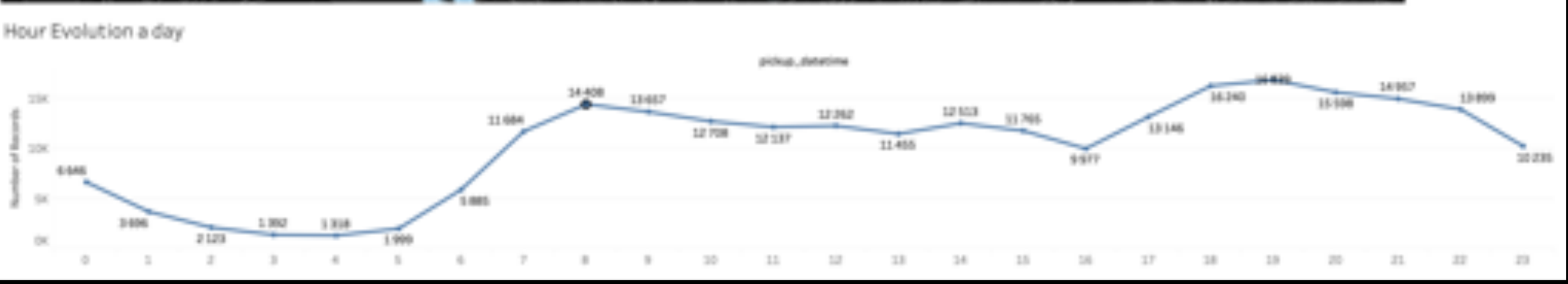


Map based on pickup_longitude and pickup_latitude. The data is filtered on pickup_datetime, which ranges from 16/01/2013 00:00:00 to 16/01/2013 23:59:27.

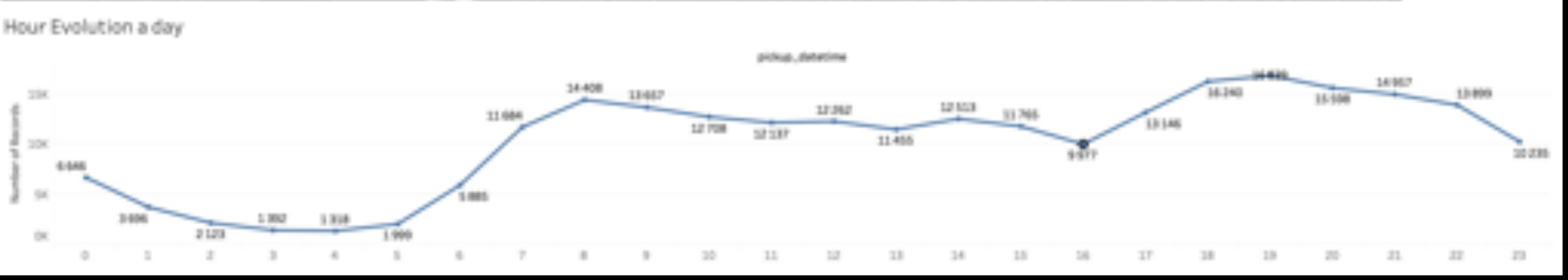
Pickups for January 16 at 3 AM



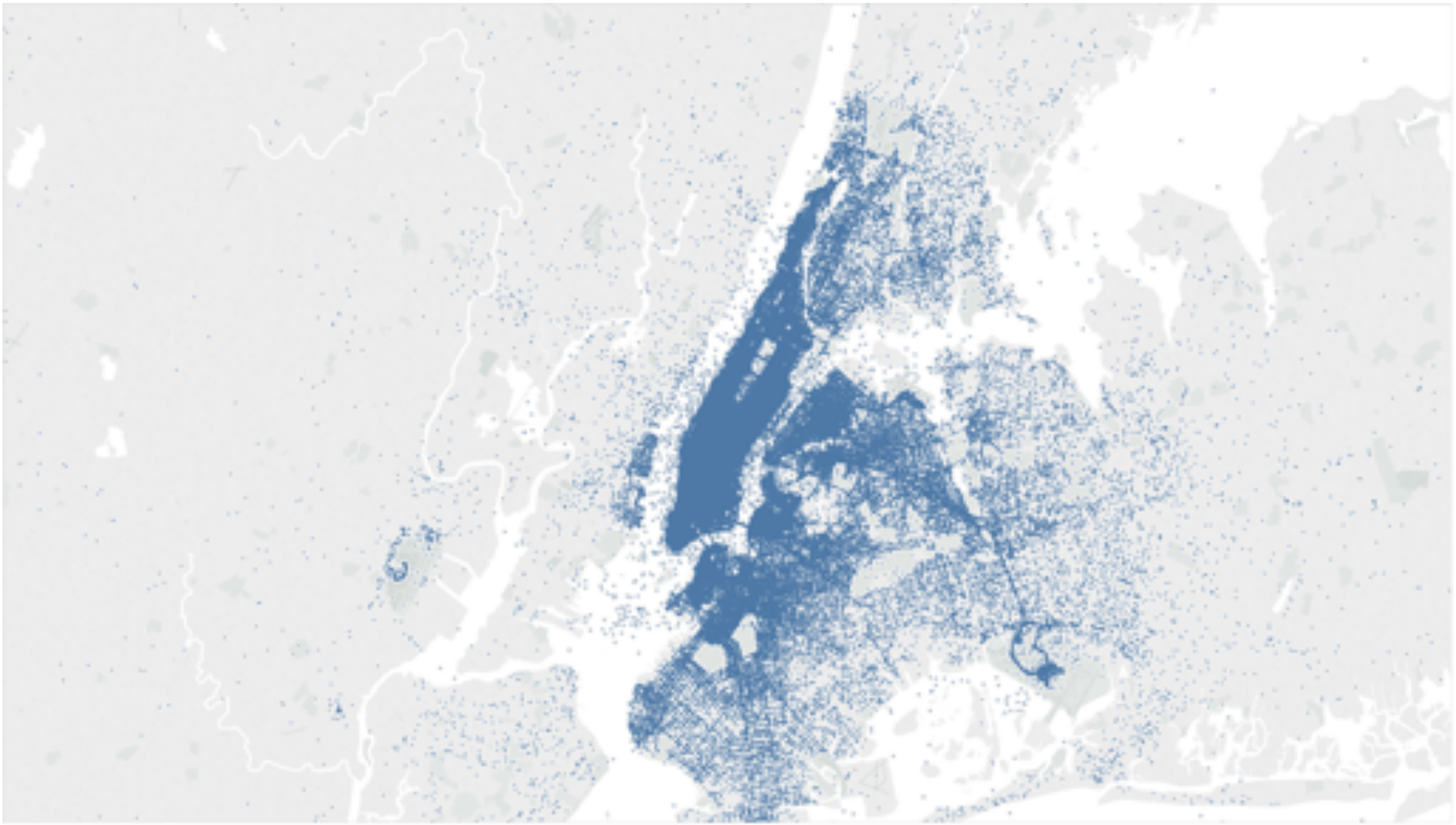
Pickups for January 16 at 8 AM



Pickups for January 16 at 4 PM



Data for 20 days: Drop-off

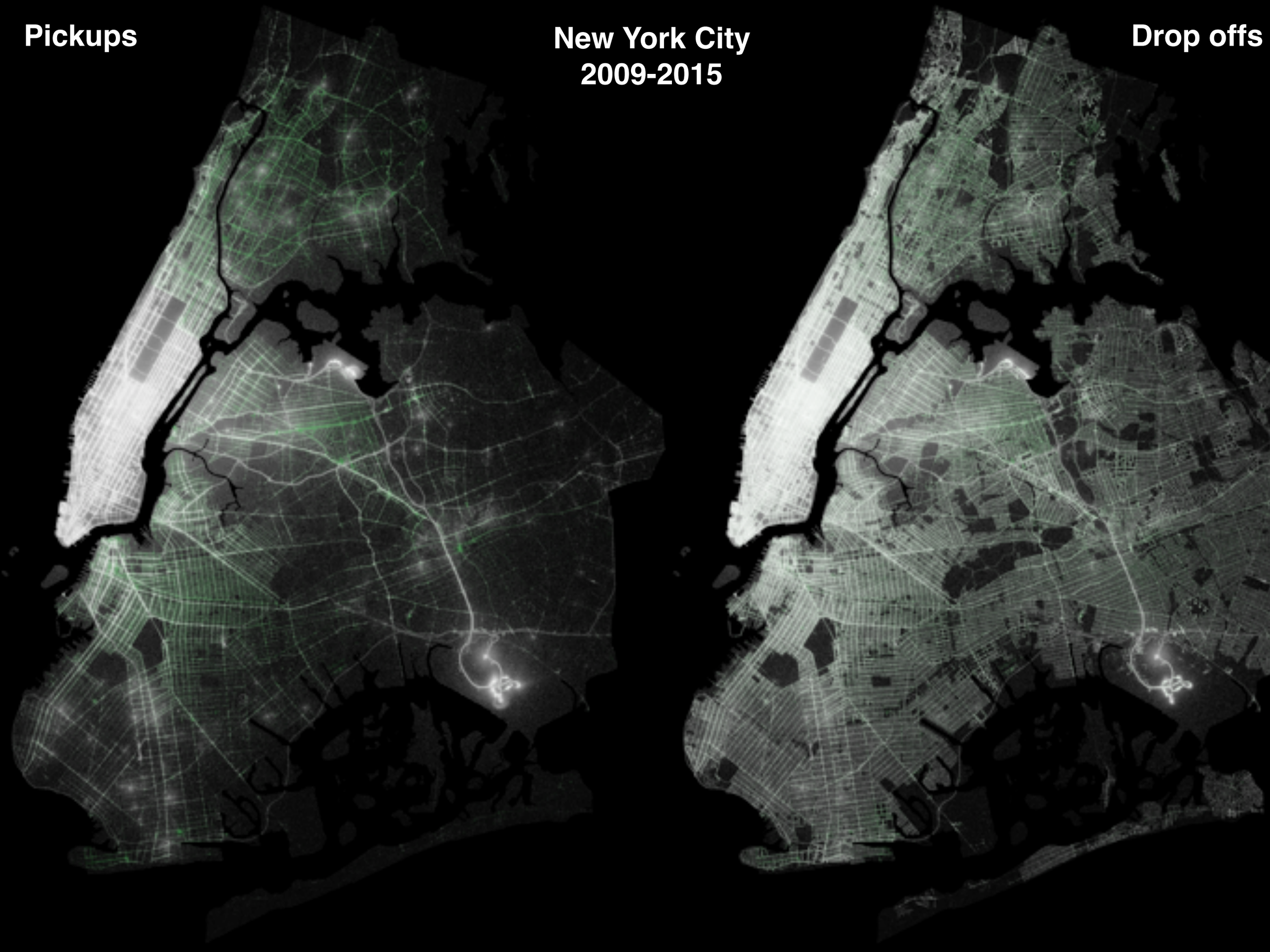


Map based on dropoff_longitude and dropoff_latitude.

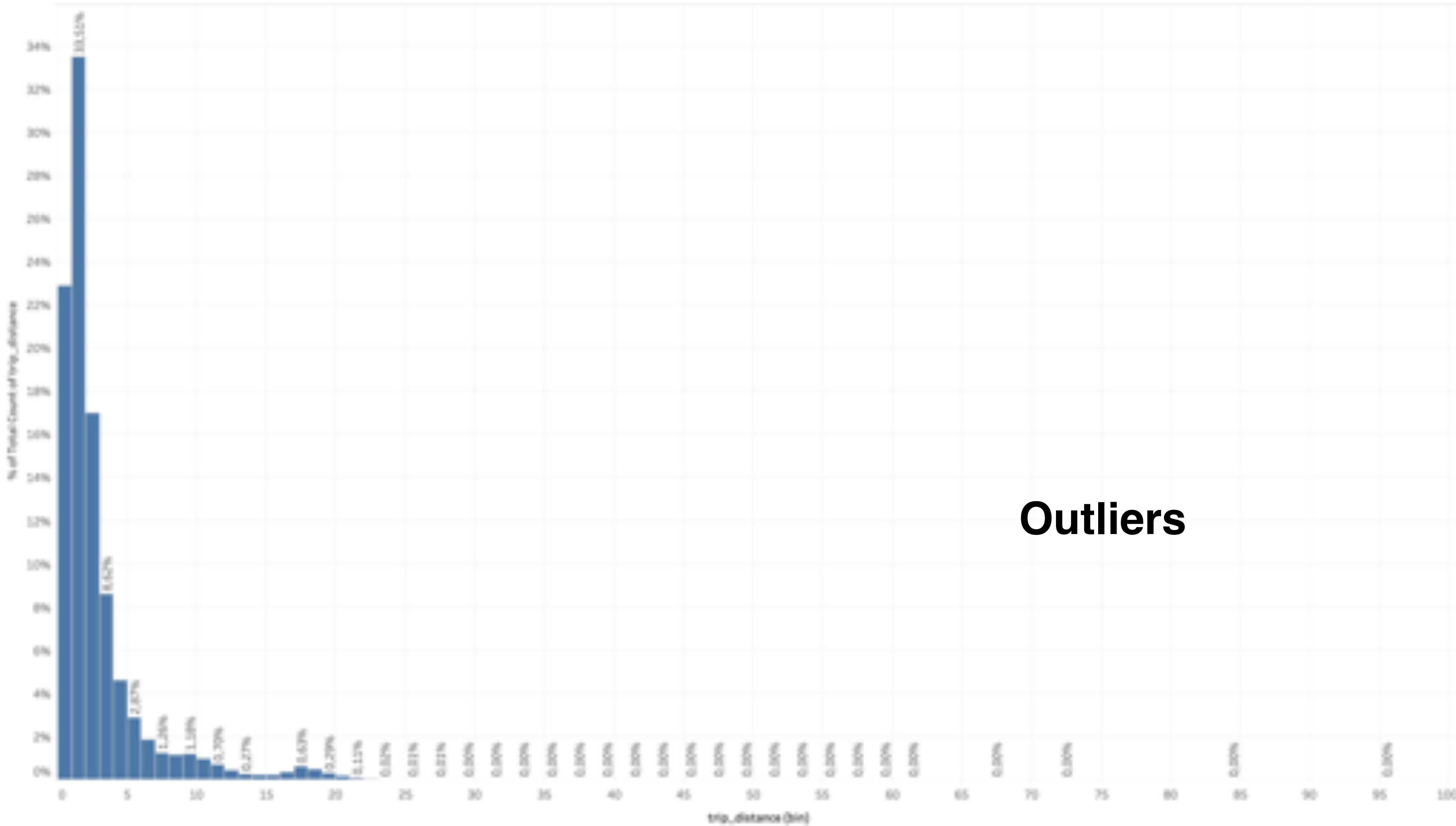
Pickups

New York City
2009-2015

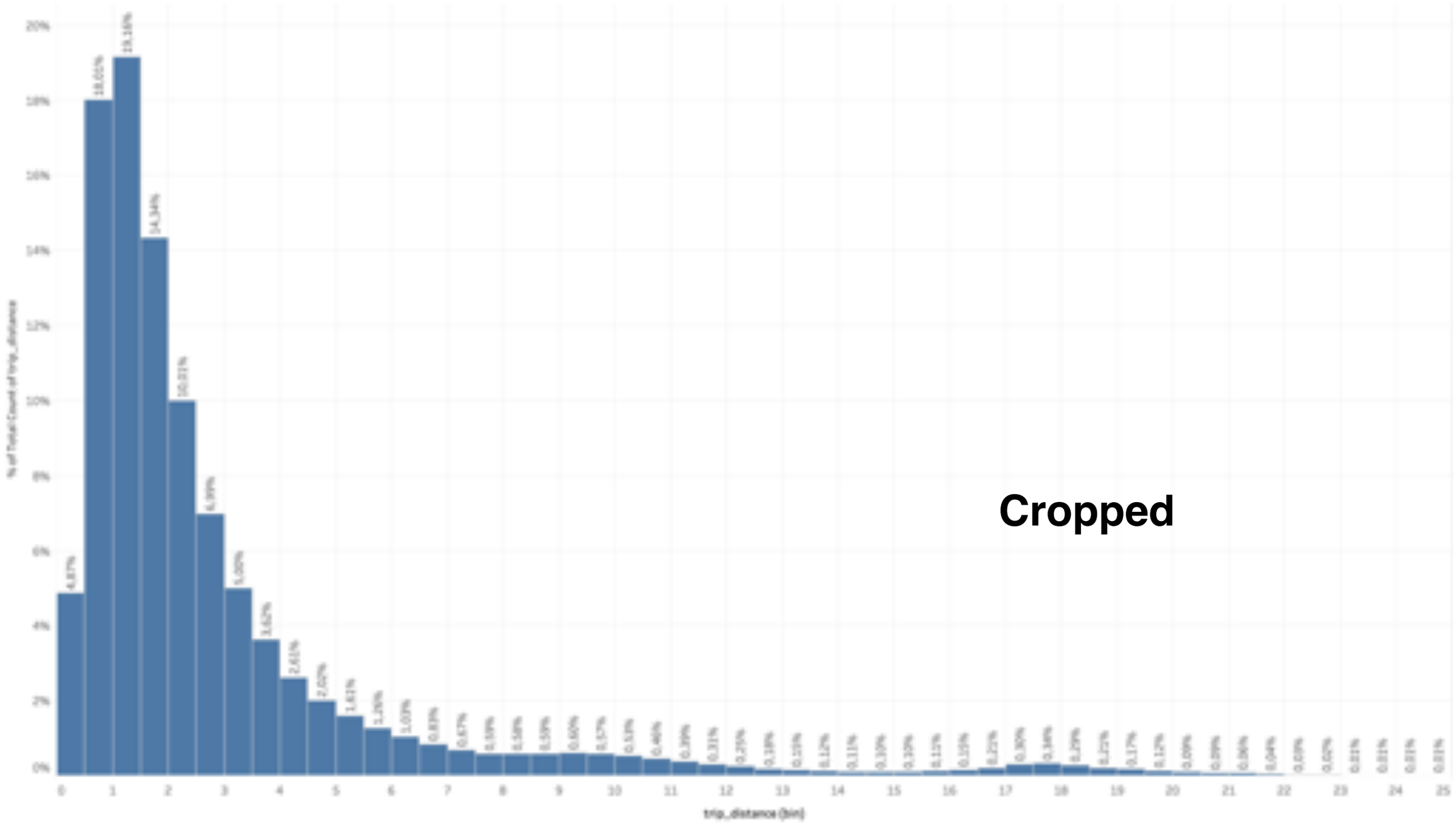
Drop offs



Trip Distance



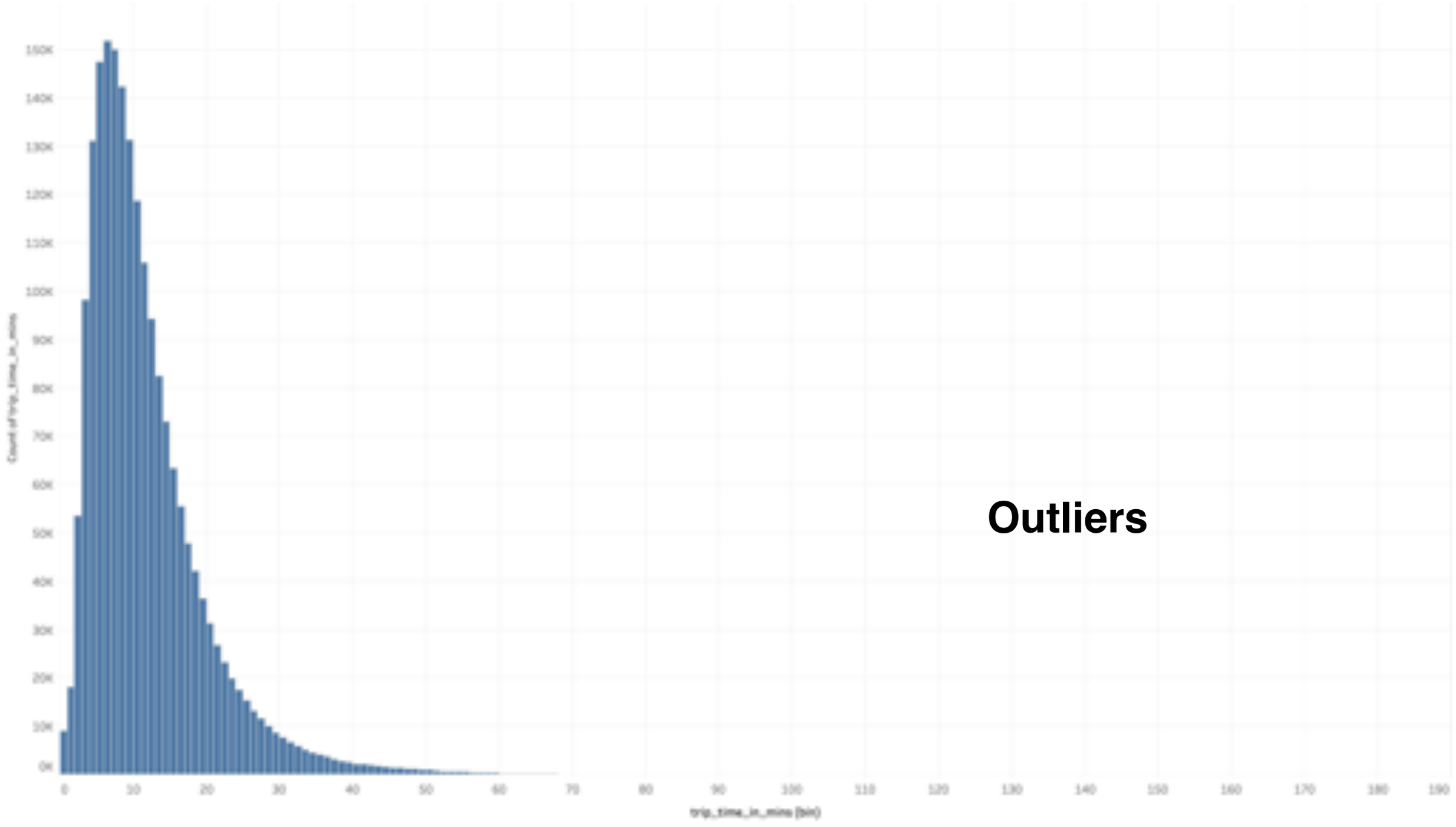
Trip Distance



Cropped

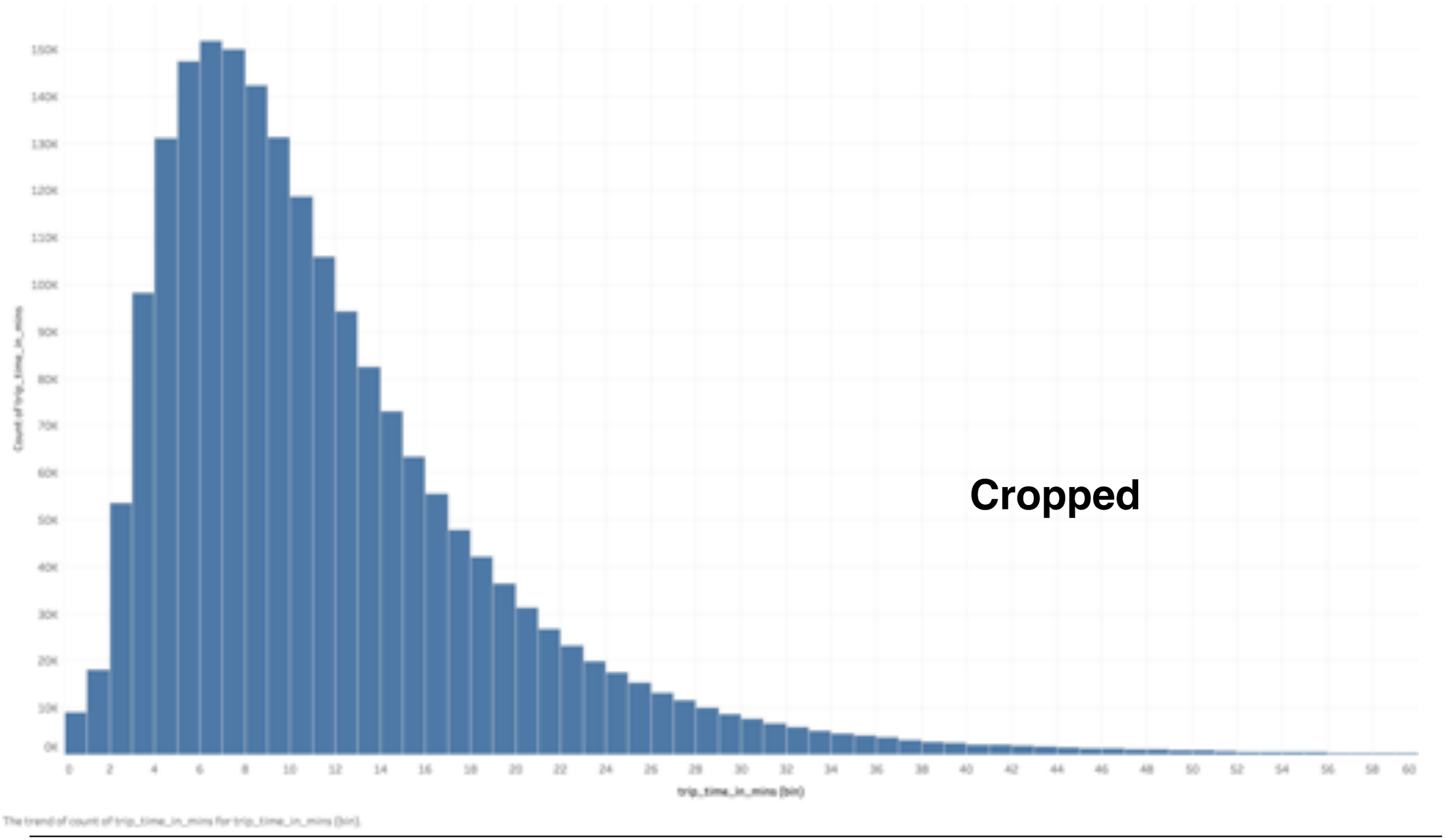
The trend of % of Total Count of trip_distance for trip_distance (bin).

Trip Time



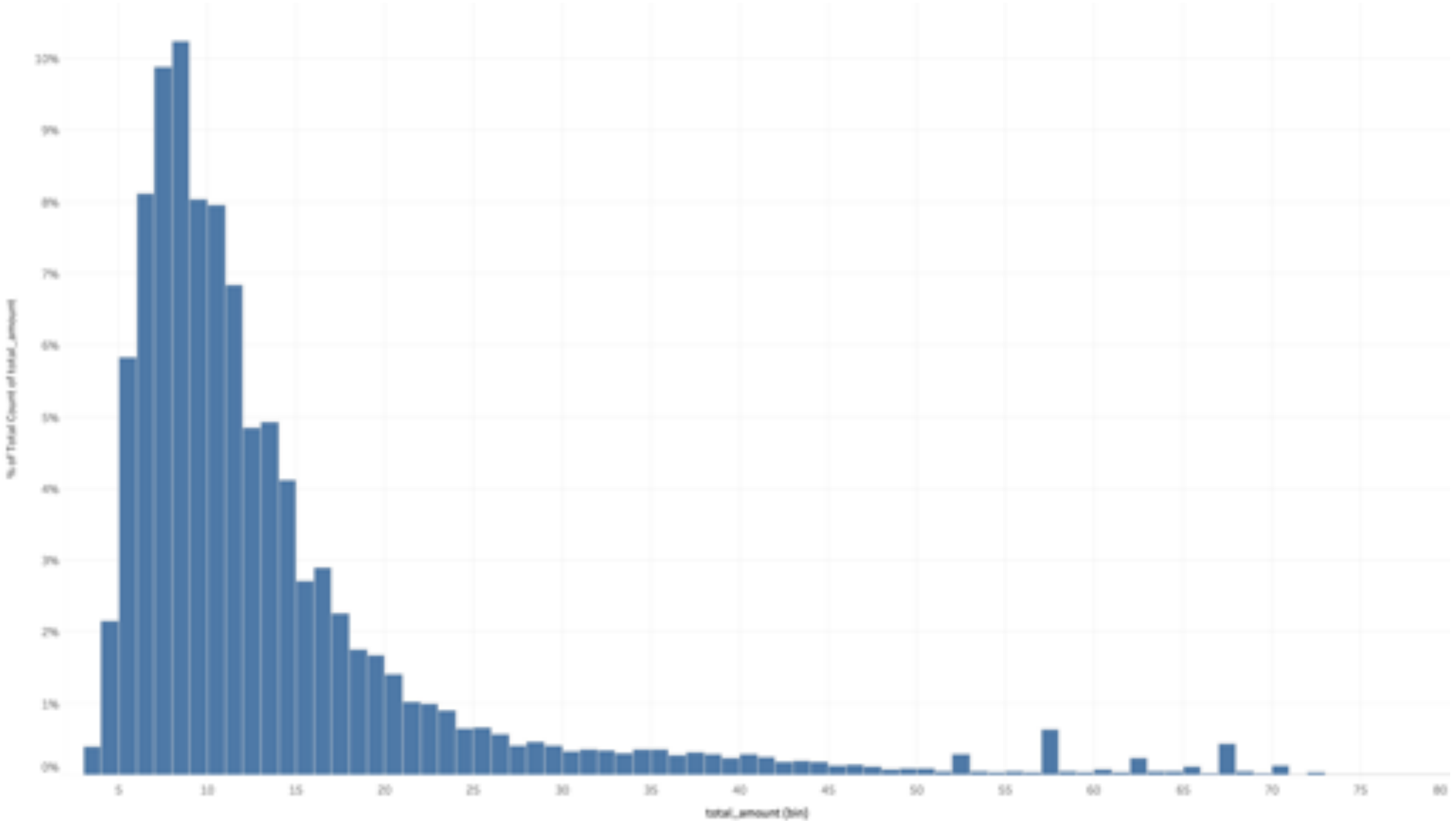
Outliers

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).

Questions

Questions

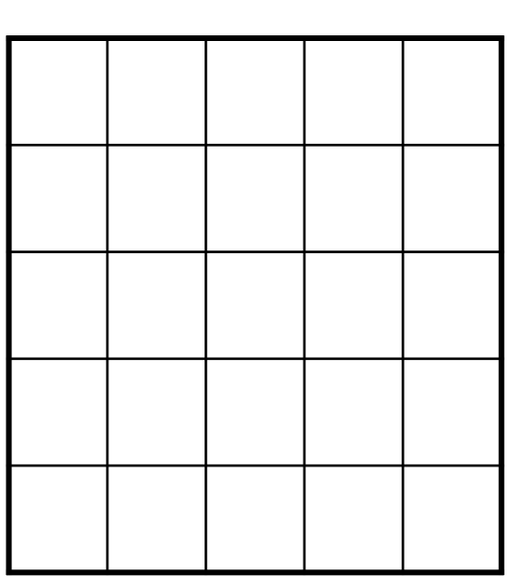
- **Q1: Find the top 10 most frequent routes during the last 30 minutes.**
- **Q2: Identify areas that are currently most profitable for taxi drivers.**
- **Q3: Alert whenever the average idle time of taxis is greater than a given amount of time (say 10 minutes).**
- **Q4: Detect congested areas.**
- **Q5: Select the most pleasant taxi drivers.**

.....

- **You may propose two additional and optional queries**

Questions

- **Q1: Find the top 10 most frequent routes during the last 30 minutes.**
  - A route is represented by a starting grid cell and an ending grid cell.
  - All routes completed within the last 30 minutes are considered for the query.
  - The output query results must be updated whenever any of the 10 most frequent routes changes.



  - Use a grid of 300 x 300 cells
  - Each cell is a square of 500 x 500 m
  - Cell 1.1, located at 41.474937, -74.913585 (in Barryville)
  - The coordinate 41.474937, -74.913585 marks the center of the first cell
  - All trips starting or ending outside this area are treated as outliers (not be considered)

Questions

- **Q2: Identify areas that are currently most profitable for taxi drivers.**
  - The profitability of an area is determined by dividing the area profit by the number of empty taxis in that area within the last 15 minutes.
  - The profit that originates from an area is computed by calculating the average fare + tip for trips that started in the area and ended within the last 15 minutes.
  - The number of empty taxis in an area is the sum of taxis that had a drop-off location in that area less than 30 minutes ago and had no following pickup yet.
  - For this problem use a cell size of 250m X 250m, i.e., a 600 x 600 grid

Questions

- **Q3: The city wants to be alerted whenever the average idle time of taxis is greater than a given amount of time (say 10 minutes)**
  - The idle time of a taxi is the time mediating between the drop off of a ride, and the pickup time of the following ride.
  - It is assumed that a taxi is available if it had at least one ride in the last hour.

Questions

- **Q4: Detect congested areas**
  - Areas where, when the taxis enter there, the rides increase in their duration.
  - For that, there should be alerts when a taxi has a peak in the duration of the ride that is followed by at least 3 rides all increasing in their duration.
  - The alert should contain the location where the taxi started the ride which had the peak duration.

Questions

- **Q5: Select the most pleasant taxi drivers**
  - To distinguish the most pleasant taxi drivers, it should be nice to have an event, emitted once a day, signaling the taxi driver with the highest total amount of tips in that day.

Recommendations

Recommendations

- **Read all the available information**
  - http://www.debs2015.org/call-grand-challenge.html
- **Get familiar with the sample data**
- **Prepare filters to exclude non used data**
  - Out of area
  - Extreme values and Null values that affect computation
- **Compute Streams with converted coordinates to cell grids**
  - Simplified flat earth assumption for mapping coordinates to cells in the queries. You can assume that a distance of 500 meter south corresponds to a change of 0.004491556 degrees in the coordinate system. For moving 500 meter east you can assume a change of 0.005986 degrees in the coordinate system.