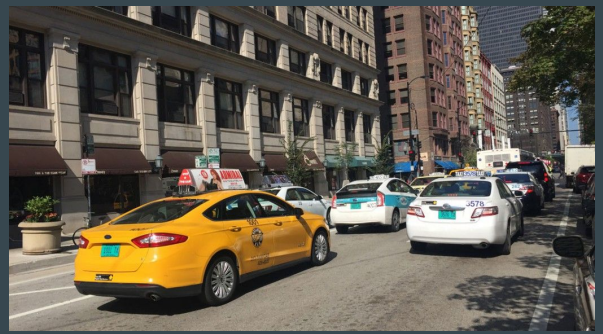
Analyzing Chicago Taxi Trip Data

Arif-Chuang-Scivittaro



Chicago Taxi Trips Dataset from Chicago Data Portal

- 109M rows, 23 columns, 40GB
- 27M trips in 2015 alone
- Includes data on over 100M taxi rides dating back to 2013
- Each row contains:
  - Trip ID
  - Taxi ID
  - Times/lengths of trip
  - Start/ending dates and coordinates
  - Fare amount and type of payment
- Columns used:
  - Taxi ID
  - Start/end coordinates, time
  - Community area

Statistics Obtained using MapReduce

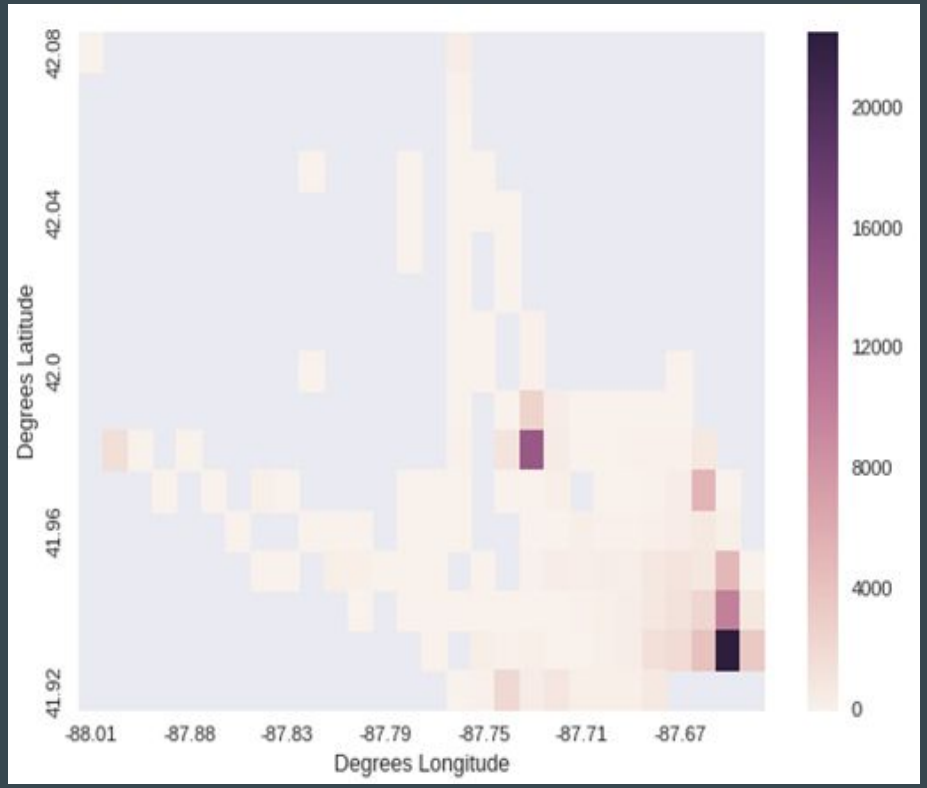
- Average Ride Length (Miles): 2.62
- Average Time: 12 min
- Average tip as a % of fare: 22.3%
- Average tip: $13.50
- Most popular community area pickup: Near North Side
- Most popular community area drop-off: Loop

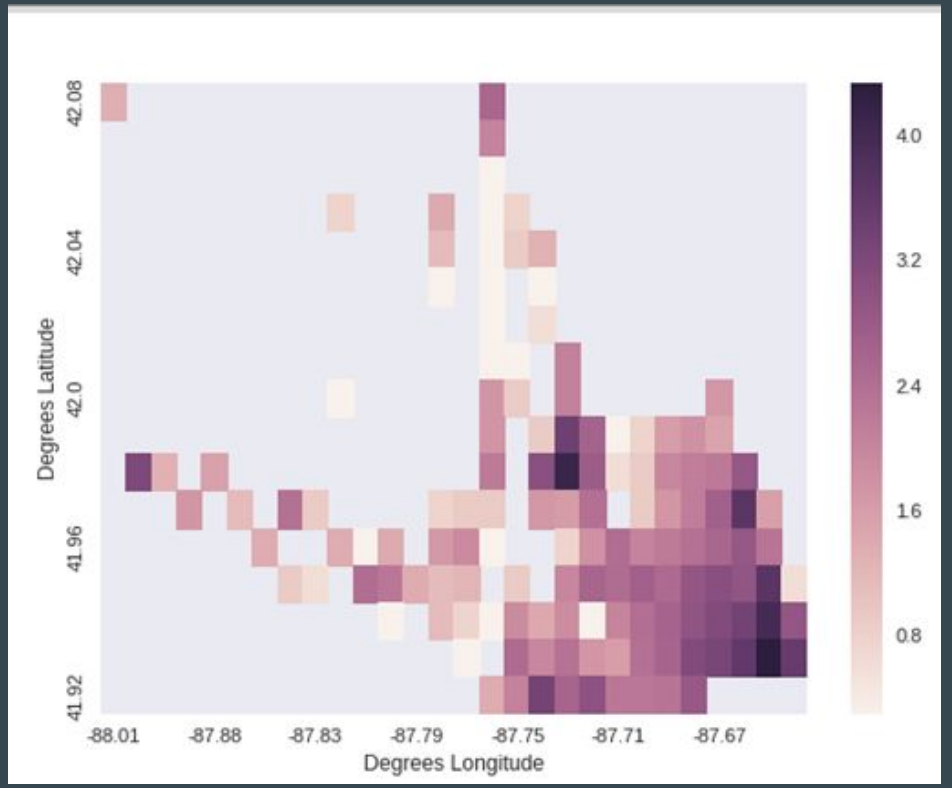
RIDE INTERSECTION HEATMAP

Taxi Trip Comparison Heatmap

- Compared every pair of trips in the dataset
- Line segment drawn from locations of start and end of trips approximated a route
- Routes that intersected were of special interest
- Narrow angles of intersection indicated two trips could take similar routes
- Used MRJob increment the counts of spots on a grid of Chicago close to each trip intersection

Results


Degrees Latitude
42.08
42.04
42.0
41.96
41.92
-88.01
-87.88
-87.83
-87.79
-87.75
-87.71
-87.67
Degrees Longitude
20000
16000
12000
8000
4000
0


Degrees Latitude
42.08
42.04
42.0
41.96
41.92
-88.01
-87.88
-87.83
-87.79
-87.75
-87.71
-87.67
Degrees Longitude
4.0
3.2
2.4
1.6
0.8

Conclusions

- Areas with most hits were around downtown and towards O'Hare
- Intersections of straight segments tend to occur around popular pickup and dropoff points - more precisely approximating a route could help us learn more about congestion in the middle of a trip

Next Steps

- More closely approximate the route taken by the taxi, with more precision than a straight line segment
- Compare trips whose projected routes do not cross but are very close
- Can results be extrapolated to non-taxi traffic flow to estimate total traffic and congestion in Chicago?
- Analyze why the southern portions of Chicago do not appear in the data

DRIVER BEHAVIOR

Drivers ‘in the hole’

- In the hole: when a driver is unlikely to find their next passenger near where they will drop off their current passenger
- If I drop someone off in Woodlawn, what is the likelihood I will find a new passenger within 30 minutes?
- Used MRJob to construct the route for every (driver, date) pair
- Examine each ride in each route - yield the dropoff neighborhood, and if they found their next ride within 30 minutes
    - Multi-step MRJob

Results

- Initial results on a 5M subsample are not surprising - highest likelihoods near the loop and O'Hare

Next Step: Supply and Demand

- Can we analyze how drivers respond to changes in demand?
- Following a drop-off in the south side, do drivers tend to head back up to the loop despite there being demand in the south side?
- Do fares tend to change with demand shifts?
- How does the ‘mass’ of taxis shift throughout a day?

CHALLENGES

Assumptions

- One-to-one relationship between Taxi IDs and Drivers
  - Companies generally own taxi medallions which allows them to lease out cars to drivers in shifts
- Once a driver starts their day, they take no breaks
- Straight line between start point and end point approximates a route well
  - Highways do not follow the grid pattern

Challenges

- Location information and time are not exact, for privacy
- Locations: by ‘census tract,’ generally sufficiently small

- Outputting results from MRJob into CSV
  - Inside ‘if __name__ == ‘__main__”, redirect STDOUT to out.tsv
  - Open out.tsv and write its rows to out.csv

- Heatmap: packages require all points to be loaded into memory
  - Packages used a list of points to generate heatmaps
  - Used Seaborn package for presentation