Big Data Web Services for Uber and Taxi Riders and Drivers

Huiyu Sun and Siyuan Hu

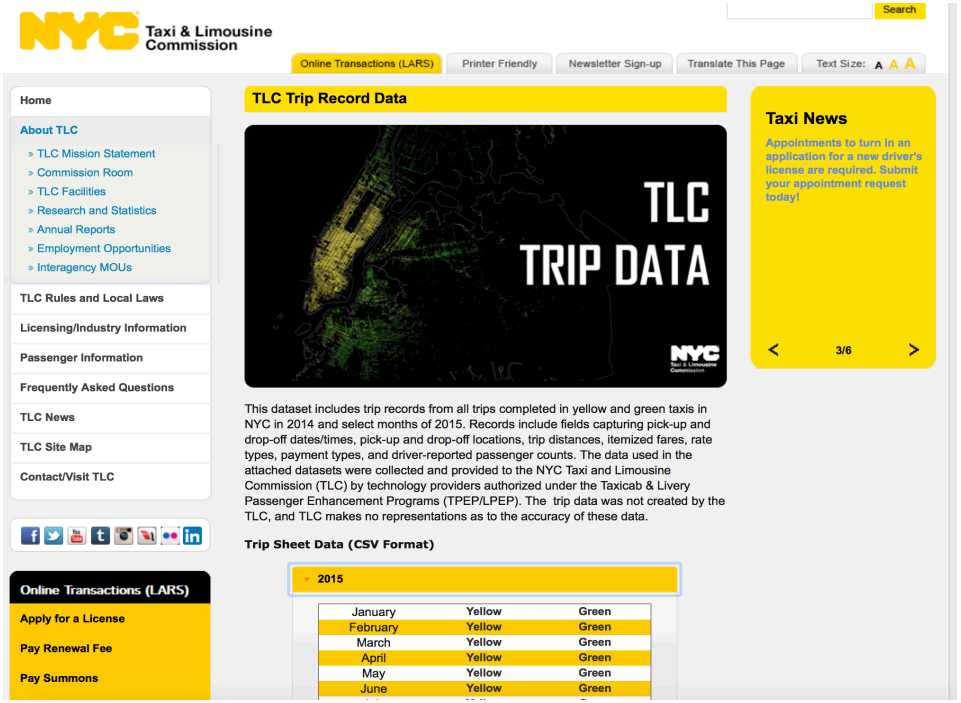
Department of Computer Science

New York University

NEW YORK UNIVERSITY

Background

- Taxi and Uber in NYC are equipped with GPS and fare collection systems.
- Trip data uploaded and made available to the public.
- NYC Taxi & Limousine Commission (TLC) website: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml



Data Source

- Taxi dataset: includes trip records from all trips completed in yellow and green taxis in NYC between 2009 and 2015.
- Total size: 50GB (2011-2015).
- Uber dataset: trip records between 2014-2015.
- Total size: 1GB (2014-2015).

VendorID	lpep_pickup_d	Lpep_dropoff_d	Store_and_f	RateCodeID	Pickup_longi	Pickup_latitu	Dropoff_lon	Dropoff_latit	Passenger_c	Trip_distanc	Fare_amoun	Extra	MTA_tax	Tip_amount	Tolls_amoun	Ehail_fee	improvemen	Total_
2	12/1/15 00:12	12/1/15 00:18	N	1	-73.844681	40.721508	-73.836334	40.7088776	1	1.27	6.5	0.5	0.5	1.56	0		0.3	
2	12/1/15 00:48	12/1/15 00:59	N	1	-73.80703	40.6996574	-73.86367	40.691143	1	3.57	12.5	0.5	0.5	2	0		0.3	
2	12/1/15 00:06	12/1/15 00:20	N	1	-73.961815	40.8056412	-73.92598	40.8241234	2	3.51	13.5	0.5	0.5	4.44	0		0.3	
2	12/1/15 00:43	12/1/15 00:59	N	1	-73.945221	40.8083839	-73.959587	40.8013573	1	2.43	12.5	0.5	0.5	2.76	0		0.3	
2	12/1/15 00:04	12/1/15 00:09	N	1	-73.939018	40.805542	-73.943977	40.8137398	5	0.89	5.5	0.5	0.5	1	0		0.3	
2	12/1/15 00:38	12/1/15 00:43	N	1	-73.941574	40.8061485	-73.953438	40.8094292	5	0.74	5.5	0.5	0.5	2.04	0		0.3	
2	12/1/15 00:56	12/1/15 01:09	N	1	-73.949654	40.8022308	-73.978699	40.7459984	5	4.56	15	0.5	0.5	0	0		0.3	
2	12/1/15 00:07	12/1/15 00:15	N	1	-73.903549	40.7454185	-73.893211	40.7348633	5	1.24	7	0.5	0.5	0	0		0.3	
2	12/1/15 00:08	12/1/15 00:11	N	1	-73.939011	40.8437843	-73.93335	40.8507767	1	0.68	4.5	0.5	0.5	1.16	0		0.3	
2	12/1/15 00:03	12/1/15 00:05	N	1	-73.939781	40.7942314	-73.945175	40.7844963	1	0.8	4.5	0.5	0.5	0	0		0.3	
2	12/1/15 00:07	12/1/15 00:15	N	1	-73.945137	40.7844849	-73.935616	40.8319588	1	3.81	12.5	0.5	0.5	0	0		0.3	
2	12/1/15 00:02	12/1/15 00:08	N	1	-73.917572	40.7700005	-73.916908	40.7839088	1	1.13	6.5	0.5	0.5	1	0		0.3	
2	12/1/15 00:15	12/1/15 00:18	N	1	-73.917091	40.7712097	-73.922966	40.7766342	1	0.45	4	0.5	0.5	1.59	0		0.3	
2	12/1/15 00:02	12/1/15 00:16	N	1	-73.829025	40.7134132	-73.885078	40.7497063	1	8.13	24	0.5	0.5	5.06	0		0.3	
2	12/1/15 00:20	12/1/15 00:32	N	1	-73.884254	40.7476082	-73.864532	40.7610817	1	1.79	10	0.5	0.5	0	0		0.3	
2	12/1/15 00:44	12/1/15 00:50	N	1	-73.844521	40.7205963	-73.81601	40.7081146	1	2.06	8	0.5	0.5	1.86	0		0.3	
2	12/1/15 00:02	12/1/15 00:27	N	1	-73.940392	40.8406372	-74.013962	40.7024574	1	11.08	32.5	0.5	0.5	0	0		0.3	
2	12/1/15 00:33	12/1/15 00:38	N	1	-73.844353	40.7208786	-73.859543	40.729351	1	1.03	6	0.5	0.5	1.82	0		0.3	
2	12/1/15 00:45	12/1/15 00:54	N	1	-73.84462	40.7201309	-73.878273	40.7245827	1	2.49	10	0.5	0.5	2.26	0		0.3	
2	12/1/15 00:31	12/1/15 00:40	N	1	-73.925591	40.7457619	-73.885178	40.7438622	5	2.2	8.5	0.5	0.5	0	0		0.3	
2	12/1/15 00:24	12/1/15 00:32	N	1	-73.829788	40.7597351	-73.863602	40.7577019	1	2.13	9	0.5	0.5	0	0		0.3	
2	12/1/15 00:03	12/1/15 00:15	N	1	-73.925385	40.7617722	-73.825378	40.742836	2	7.36	21.5	0.5	0.5	0	0		0.3	
2	12/1/15 00:31	12/1/15 00:40	N	1	-73.881218	40.7560158	-73.861694	40.7449112	2	1.94	9	0.5	0.5	0	0		0.3	
2	12/1/15 00:03	12/1/15 00:09	N	1	-73.929657	40.7563591	-73.917061	40.7707291	1	1.21	6.5	0.5	0.5	0	0		0.3	
2	12/1/15 00:10	12/1/15 00:14	N	1	-73.917061	40.7707253	-73.901886	40.7763481	1	1.18	6	0.5	0.5	1.46	0		0.3	
2	12/1/15 00:16	12/1/15 00:21	N	1	-73.941658	40.8179092	-73.953606	40.8178101	1	1.02	6	0.5	0.5	0	0		0.3	
2	12/1/15 00:39	12/1/15 00:42	N	1	-73.925308	40.7619019	-73.911583	40.7586098	5	0.92	5	0.5	0.5	0	0		0.3	
2	12/1/15 00:28	12/1/15 00:33	N	1	-73.954605	40.8053093	-73.944359	40.8172989	1	1.25	6.5	0.5	0.5	2.34	0		0.3	
2	12/1/15 00:41	12/1/15 00:49	N	1	-73.955322	40.8045578	-73.93206	40.8005753	1	1.55	8	0.5	0.5	0	0		0.3	
2	12/1/15 00:17	12/1/15 00:22	N	1	-73.955467	40.7140388	-73.939011	40.7261238	1	1.37	6.5	0.5	0.5	2.34	0		0.3	
2	12/1/15 00:19	12/1/15 00:25	N	1	-73.909866	40.7754822	-73.910255	40.7656555	1	1.16	6	0.5	0.5	0	0		0.3	
2	12/1/15 00:46	12/1/15 00:54	N	1	-73.952393	40.8108482	-73.942093	40.8261185	1	1.73	8	0.5	0.5	0	0		0.3	
2	12/1/15 00:00	12/1/15 00:05	N	1	-73.911766	40.767807	-73.914078	40.7569237	1	1.01	5.5	0.5	0.5	1.36	0		0.3	
2	12/1/15 00:51	12/1/15 00:59	N	1	-73.913689	40.7655945	-73.910912	40.7762642	1	1.27	7.5	0.5	0.5	2	0		0.3	
2	12/1/15 00:34	12/1/15 00:45	N	1	-73.874741	40.7351265	-73.907501	40.7167625	1	2.67	11	0.5	0.5	0	0		0.3	
2	12/1/15 00:34	12/1/15 00:39	N	1	-73.91893	40.7589378	-73.911865	40.767704	1	0.75	5	0.5	0.5	0	0		0.3	
2	12/1/15 00:30	12/1/15 00:33	N	1	-73.957672	40.7178688	-73.956001	40.7259216	1	0.7	4	0.5	0.5	0	0		0.3	

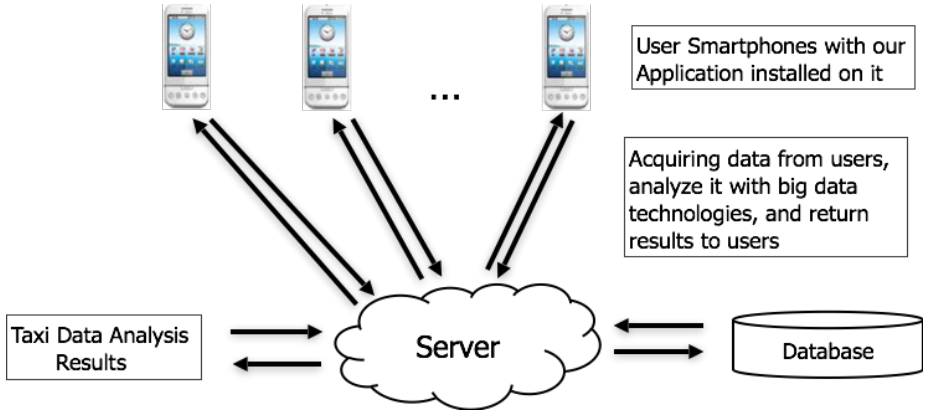
Date/Time	Lat	Lon	Base
4/1/14 00:11	40.769	-73.9549	B02512
4/1/14 00:17	40.7267	-74.0345	B02512
4/1/14 00:21	40.7316	-73.9873	B02512
4/1/14 00:28	40.7588	-73.9776	B02512
4/1/14 00:33	40.7594	-73.9722	B02512
4/1/14 00:33	40.7383	-74.0403	B02512
4/1/14 00:39	40.7223	-73.9887	B02512
4/1/14 00:45	40.762	-73.979	B02512
4/1/14 00:55	40.7524	-73.996	B02512
4/1/14 01:01	40.7575	-73.9846	B02512
4/1/14 01:19	40.7256	-73.9869	B02512
4/1/14 01:48	40.7591	-73.9684	B02512
4/1/14 01:49	40.7271	-73.9803	B02512
4/1/14 02:11	40.6463	-73.7896	B02512
4/1/14 02:25	40.7564	-73.9167	B02512
4/1/14 02:31	40.7666	-73.9531	B02512
4/1/14 02:43	40.758	-73.9761	B02512
4/1/14 03:22	40.7238	-73.9821	B02512
4/1/14 03:35	40.7531	-74.0039	B02512
4/1/14 03:35	40.7389	-74.0393	B02512
4/1/14 03:41	40.7619	-73.9715	B02512
4/1/14 04:11	40.753	-74.0042	B02512
4/1/14 04:15	40.6561	-73.9531	B02512
4/1/14 04:19	40.725	-73.9844	B02512
4/1/14 04:20	40.695	-74.1783	B02512
4/1/14 04:26	40.9859	-74.1578	B02512
4/1/14 04:27	40.6879	-74.1814	B02512
4/1/14 04:38	40.6878	-74.1816	B02512
4/1/14 04:47	40.7234	-73.9974	B02512
4/1/14 04:49	40.7336	-73.99	B02512
4/1/14 05:08	40.7141	-74.0094	B02512
4/1/14 05:12	40.7893	-73.9709	B02512
4/1/14 05:18	40.7747	-73.991	B02512
4/1/14 05:19	40.7689	-73.9876	B02512
4/1/14 05:23	40.7744	-74.0149	B02512
4/1/14 05:24	40.7393	-73.9974	B02512
4/1/14 05:24	40.7776	-73.9752	B02512
4/1/14 05:27	40.6483	-73.7829	B02512
4/1/14 05:34	40.6907	-74.1782	B02512
4/1/14 05:36	40.7217	-73.9875	B02512

Motivation

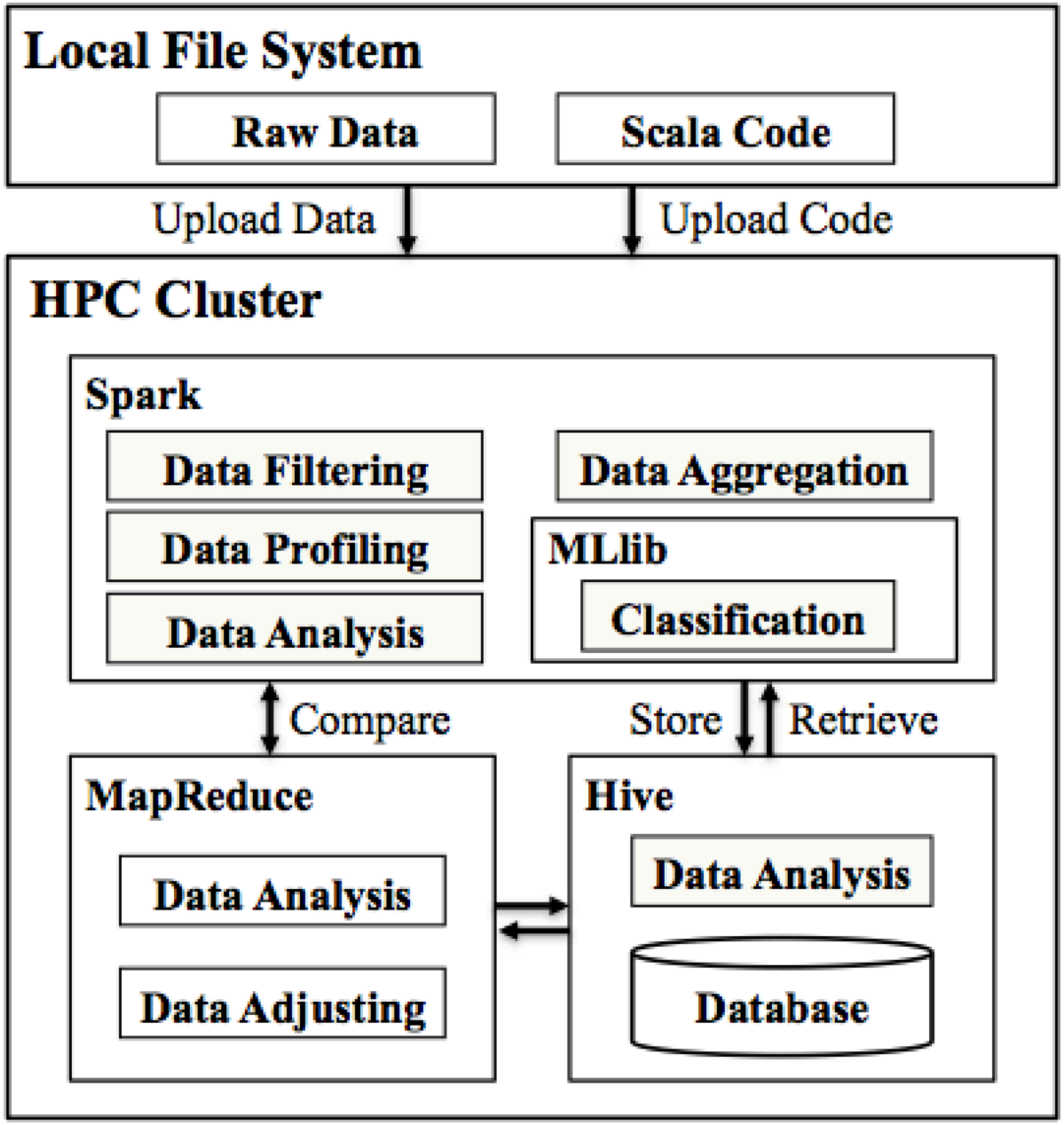
- Results can benefit: taxi companies, urban planning companies, taxi drivers, etc.
- Comparing taxi with Uber gives insight into their respective business strategies.
- Leverage historical data to understand current and future taxi and Uber rides and their transportation patterns.

Goals

- Understand traffic and travel patterns, discover relationships, and make predictions on the taxi and Uber network
- Understand to what extent Uber has affected taxi
- Propose a service architecture to integrate the back-end system that analyzes taxi data with application front-end



Data Processing Architecture



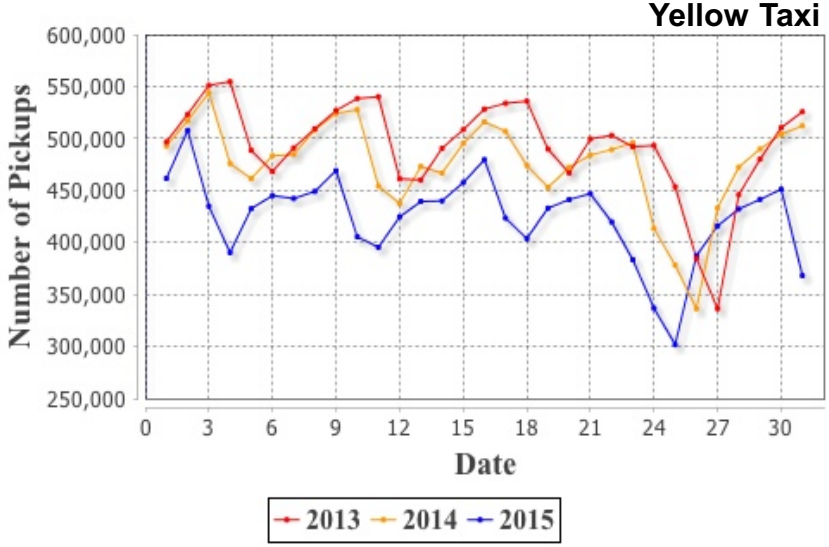
- Data Analytics and Visualization

- Application Specifics

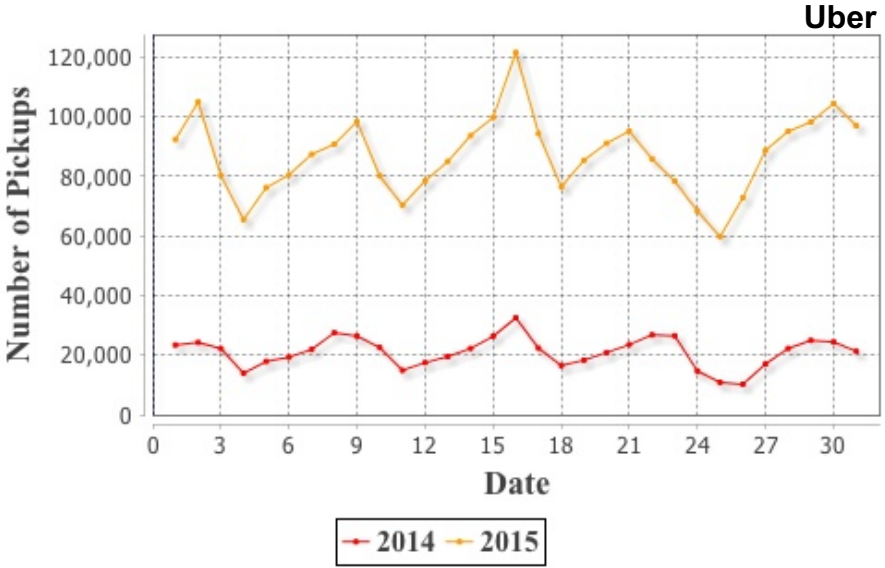
Pick Ups by Day

- Plots daily pickup numbers against day of month.
- Insight: Daily pickups differ from Green Taxi to Yellow Taxi to Uber, but all showing periodic patterns
- Put more focus on certain days.

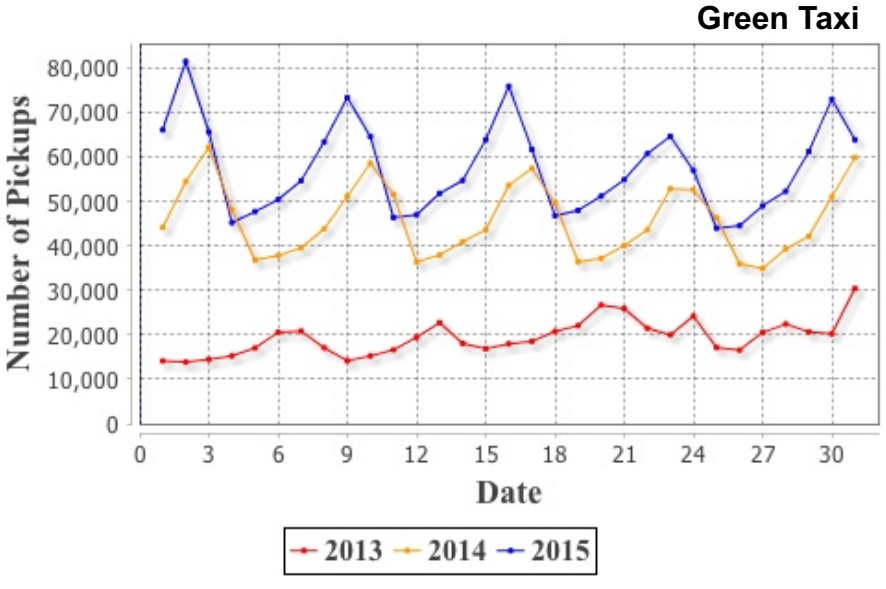
Yellow Taxi



Uber

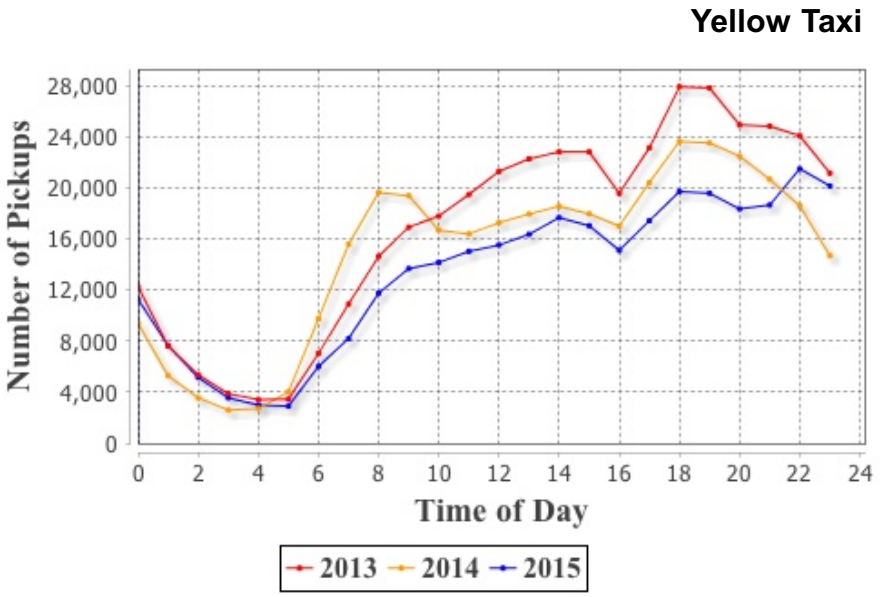


Green Taxi



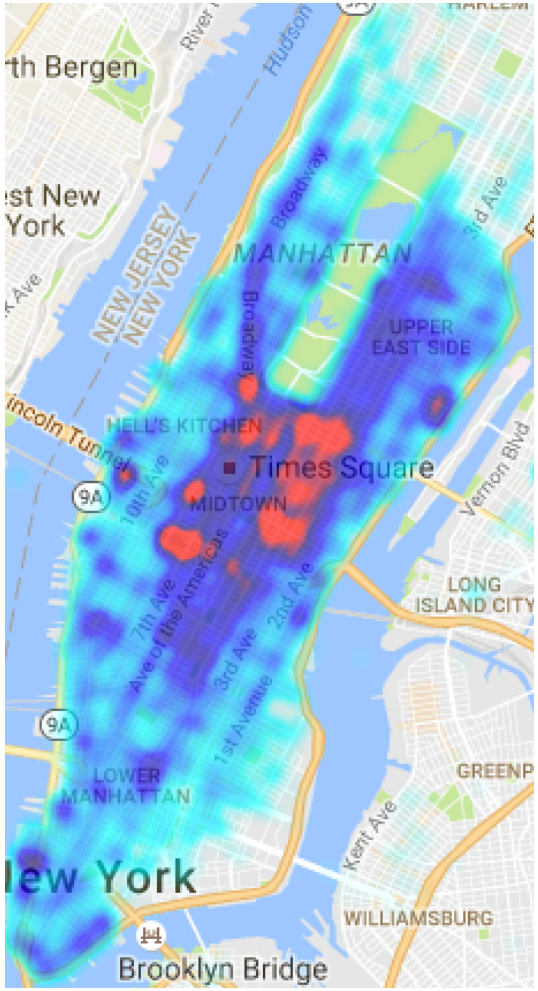
Pickups by Hours – Yellow Taxi

- The diagram shows hourly pickups of yellow taxi on the same date in different years
- Hourly pickups peak around 6-7pm, hits bottom around 4-5am, showing an average drop of 700% from peak to bottom
- Insight: Periodic patterns for the same date in different years reveal that we can use historic pickups data to predict future pickups

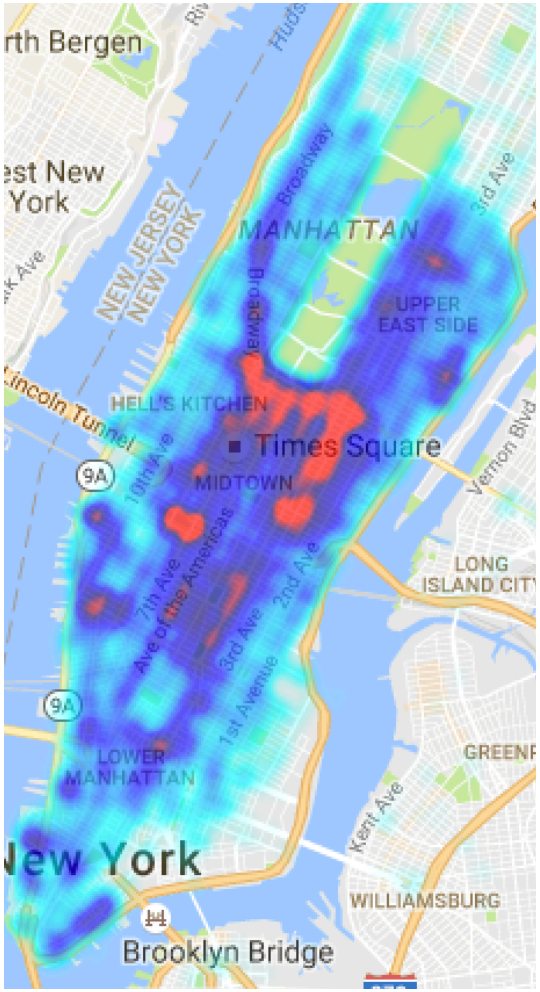


Heatmaps By Hours – Yellow Taxi

- Plots pickup heatmaps by hour of day (Red – most activity)
- Comparing 2014 with 2015, shows a strikingly similar trend
- Insight: can now confidently use historic data to predict where a large number of pick-ups will occur in the future



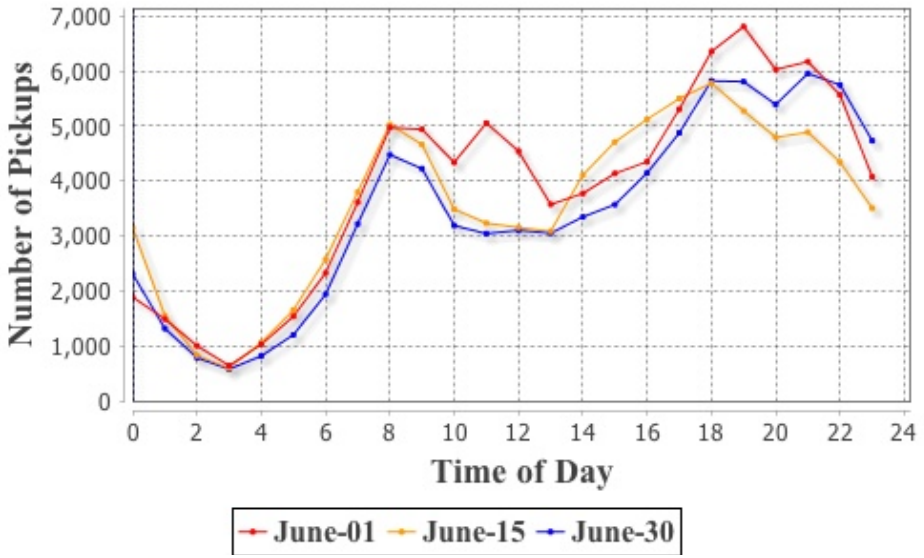
2014

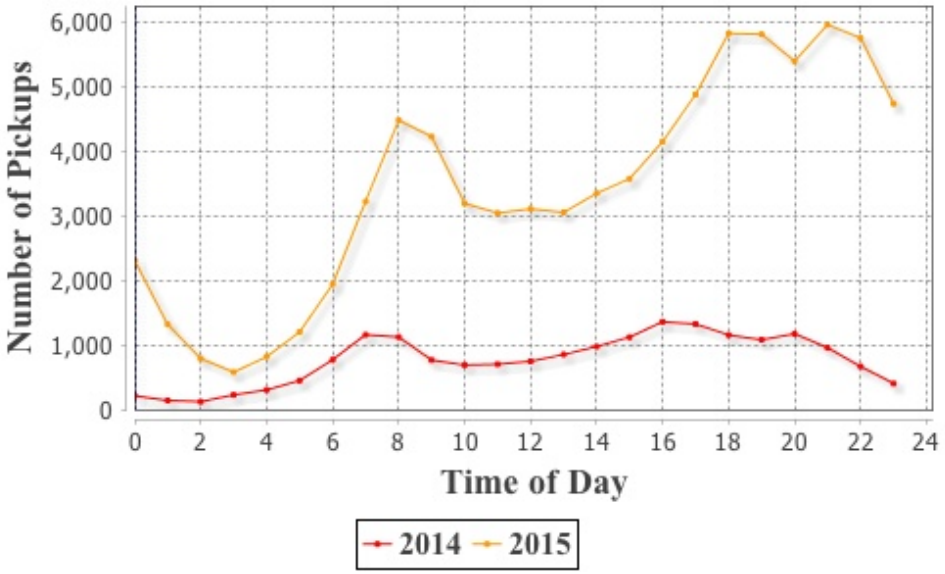


2015

Pickups by Hours - Uber

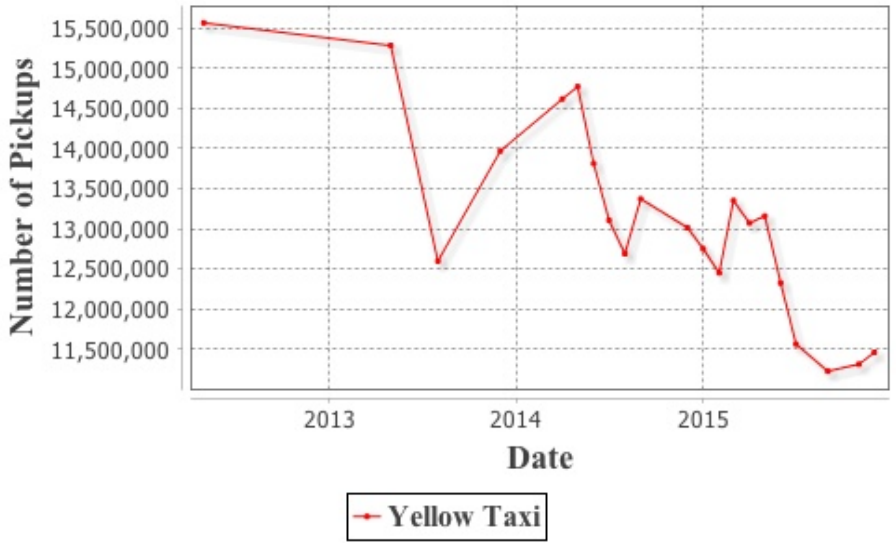
- Similar periodic patterns apply to Uber as well
- Insight: Uber drivers can utilize pickups visualization to know when is a good time to catch a passenger

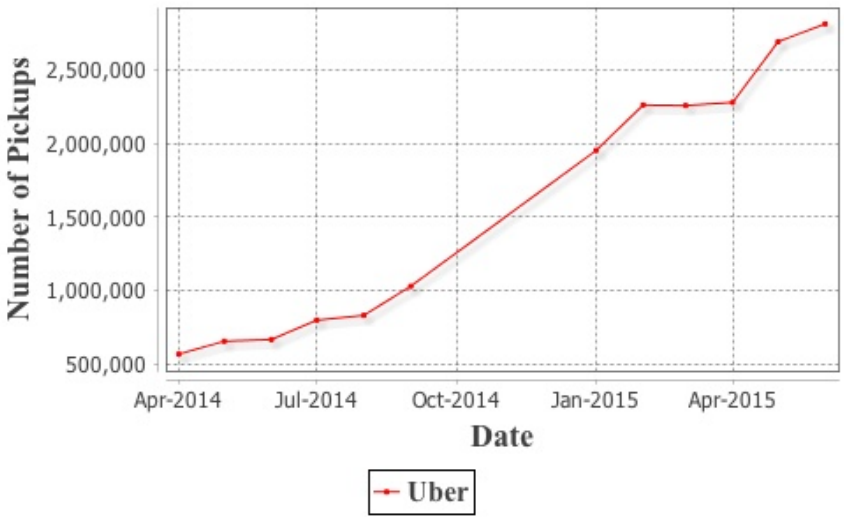


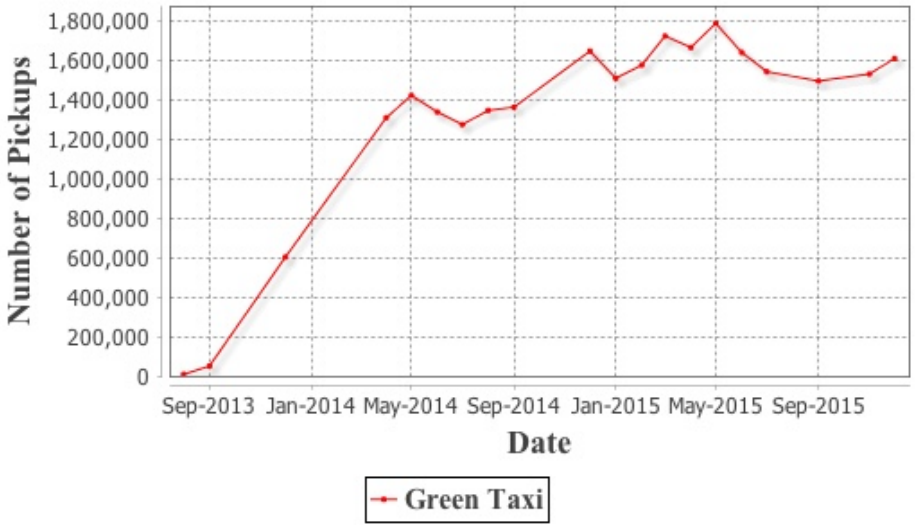


Pickups By Month

- Plot monthly pick up numbers for 2011-2015.
- Yellow taxi numbers declining. Green taxi increasing initially, then plateaus. Uber increasing.

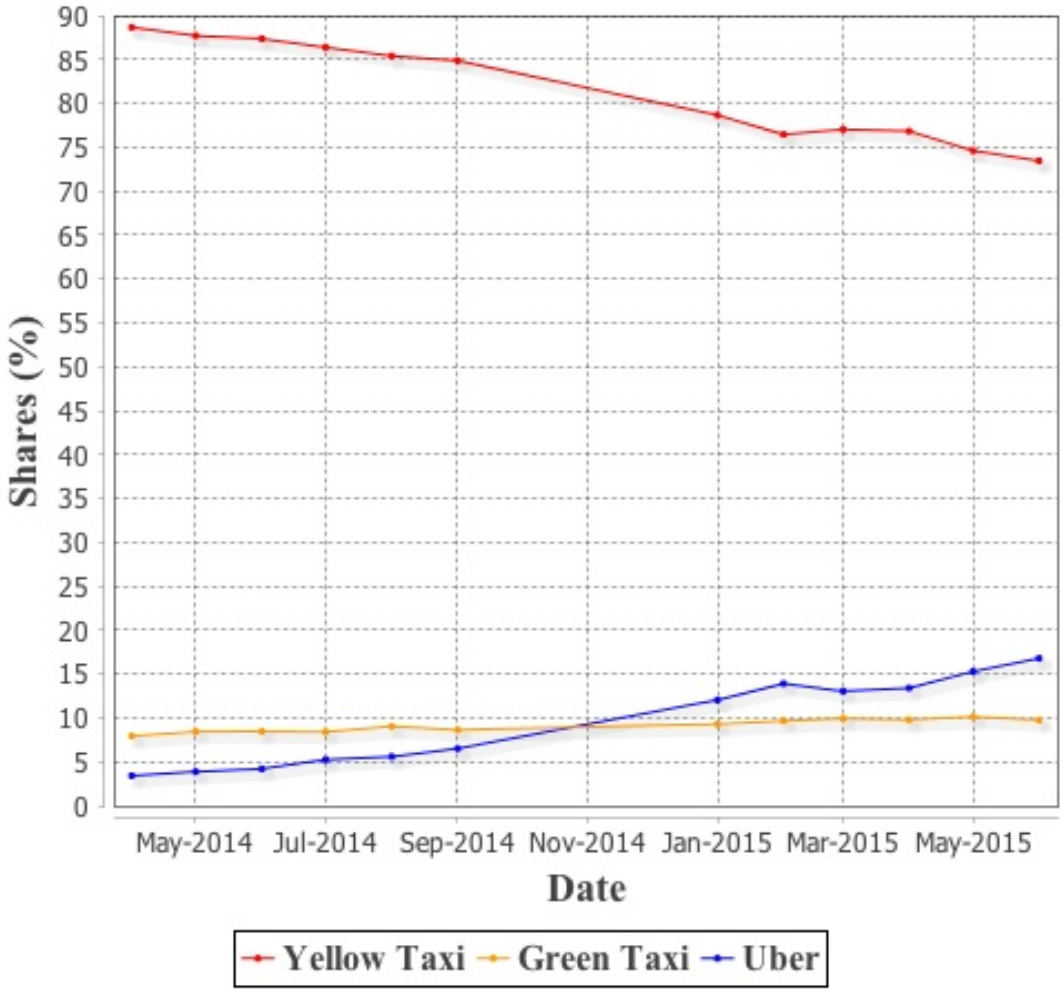






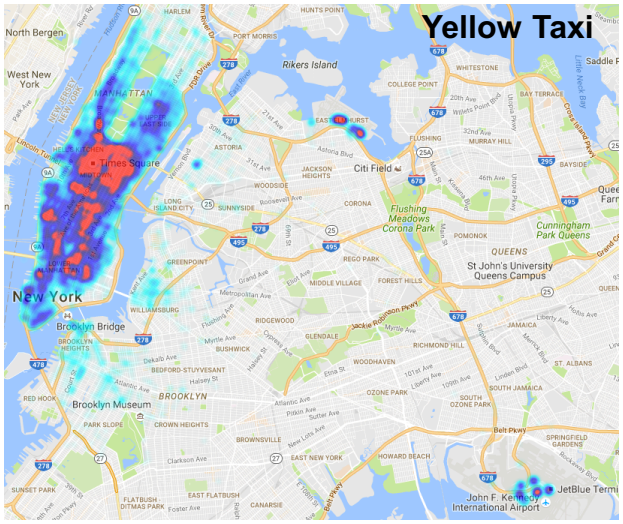
Pickups By Month - Combined

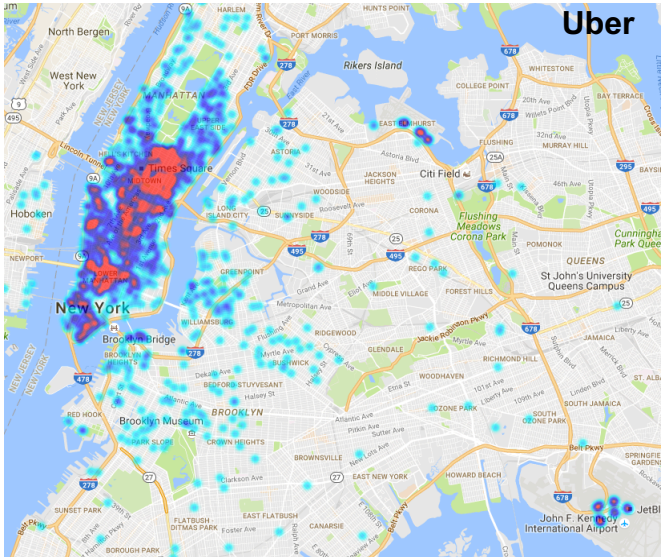
- Compare percentage of shares reveals growth pattern:
  Uber increasing
  Yellow decreasing
  Green staying flat
- Insight: we can make an assumption from the diagram that the rise in Uber causes the decline of yellow taxi.

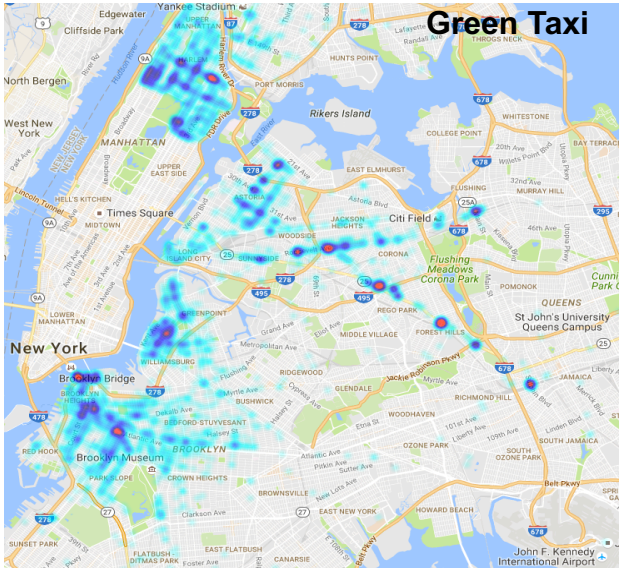


Location Heatmaps

- Yellow taxi: most pickups in Manhattan below Harlem, and at LaGuardia and JFK airports
- Green taxi: Harlem, Brooklyn, Queens, and Bronx
- Uber: combining locations from yellow and green taxis

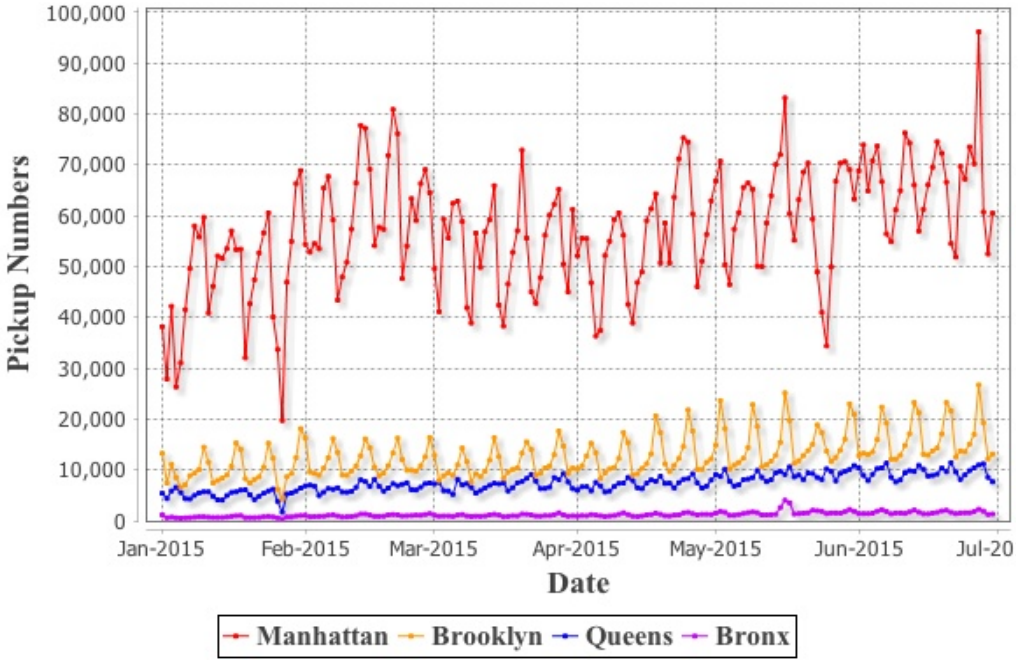


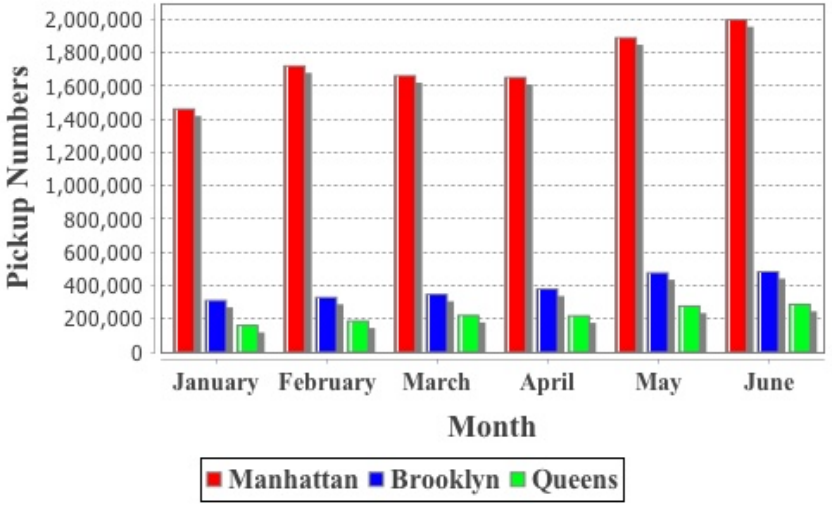




Location by Boroughs - Uber

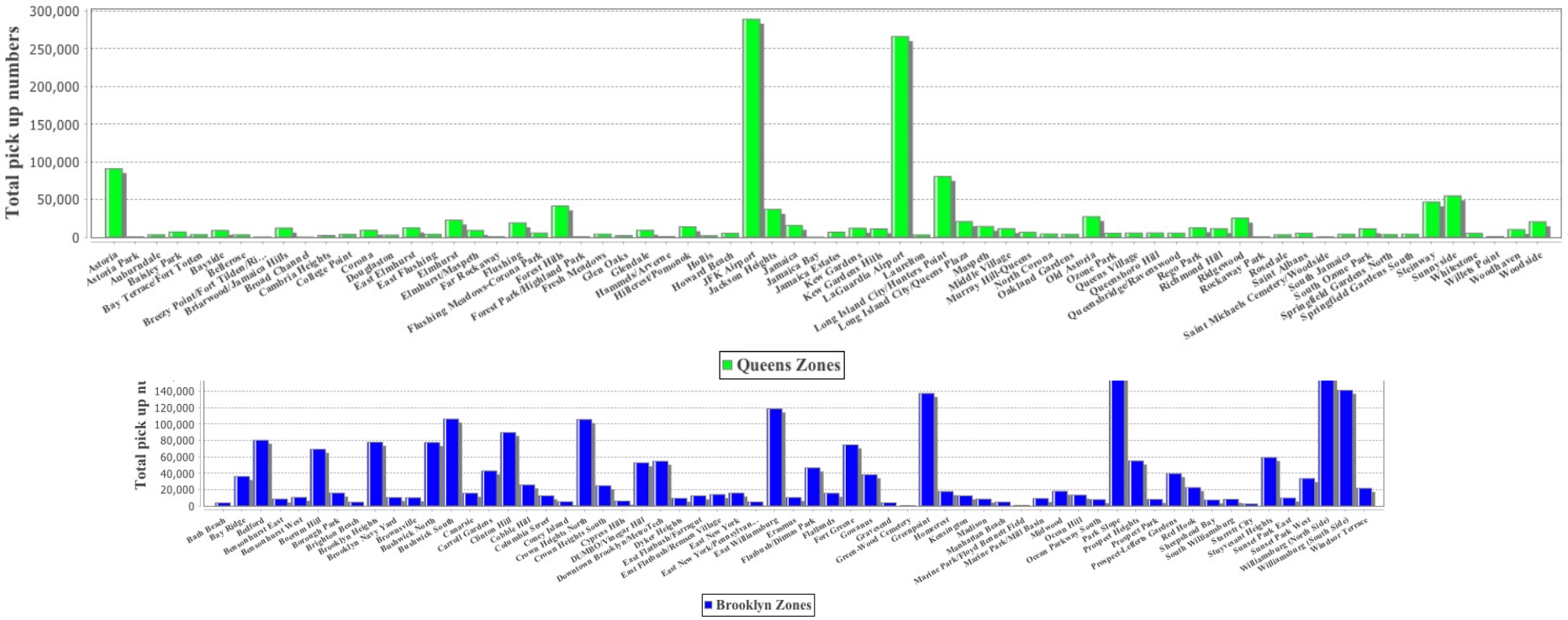
- Plots daily and monthly Uber pick ups over 6 months in 2015 by each borough
- Insight: pick up numbers gradually increasing for all boroughs, but increase rates of Brooklyn and Queens are higher than Manhattan



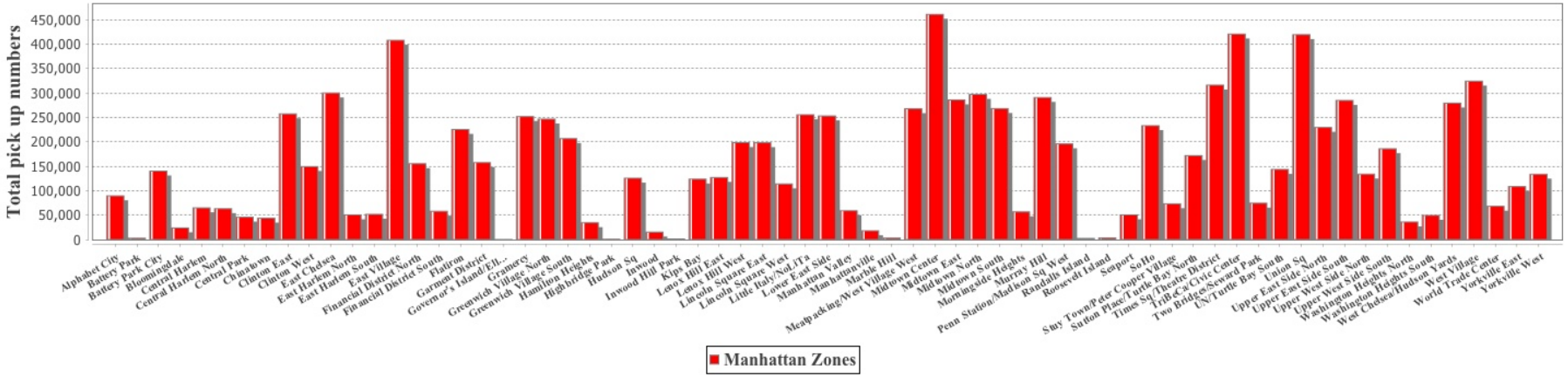


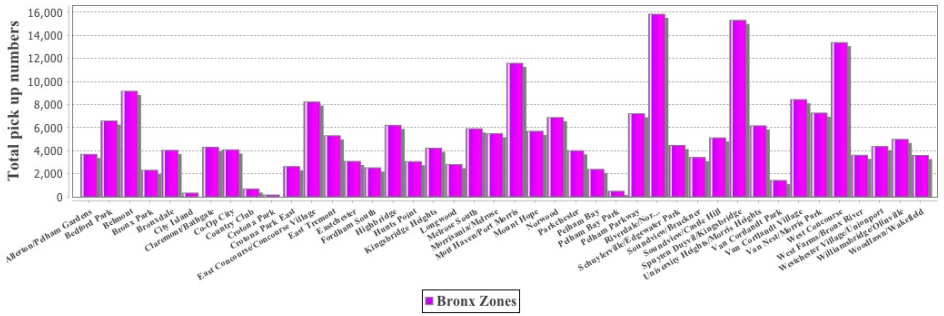
Queens & Brooklyn Zones - Uber

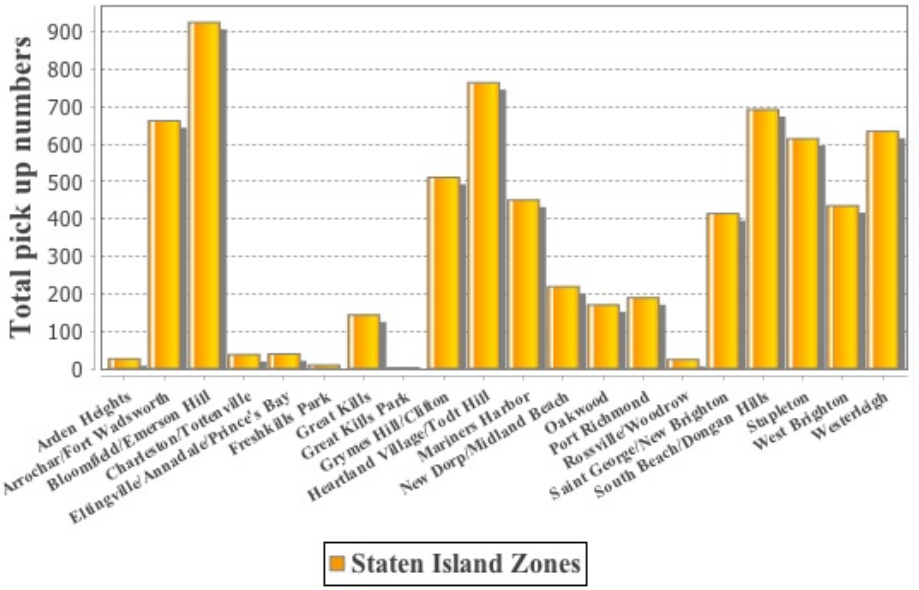
- Plots pickup numbers by different zones (districts inside a borough)
- Zones in Queens and Brooklyn vary a lot in terms of pickups and concentrate in only a few zones. This information can be used by drivers
- Insight: chance of getting a ride increases for zones with a high number of pick ups



Other Borough Zones - Uber

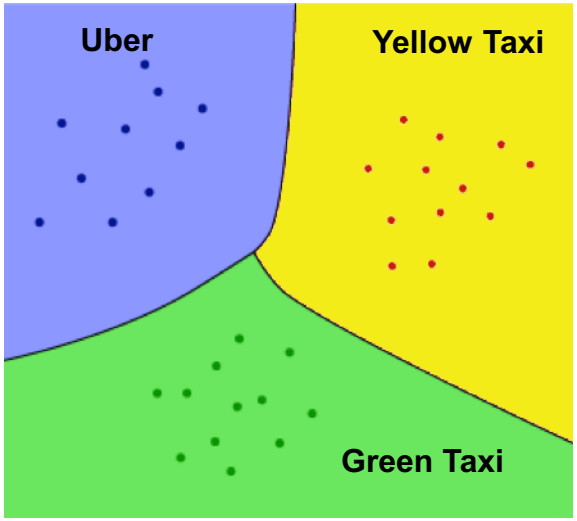






Classification – Spark MLlib

- Classify a new point on the map to be one of three categories: yellow, green or Uber.
- Train classifiers to find the dominating category for each region.



Classifier – Logistic Regression

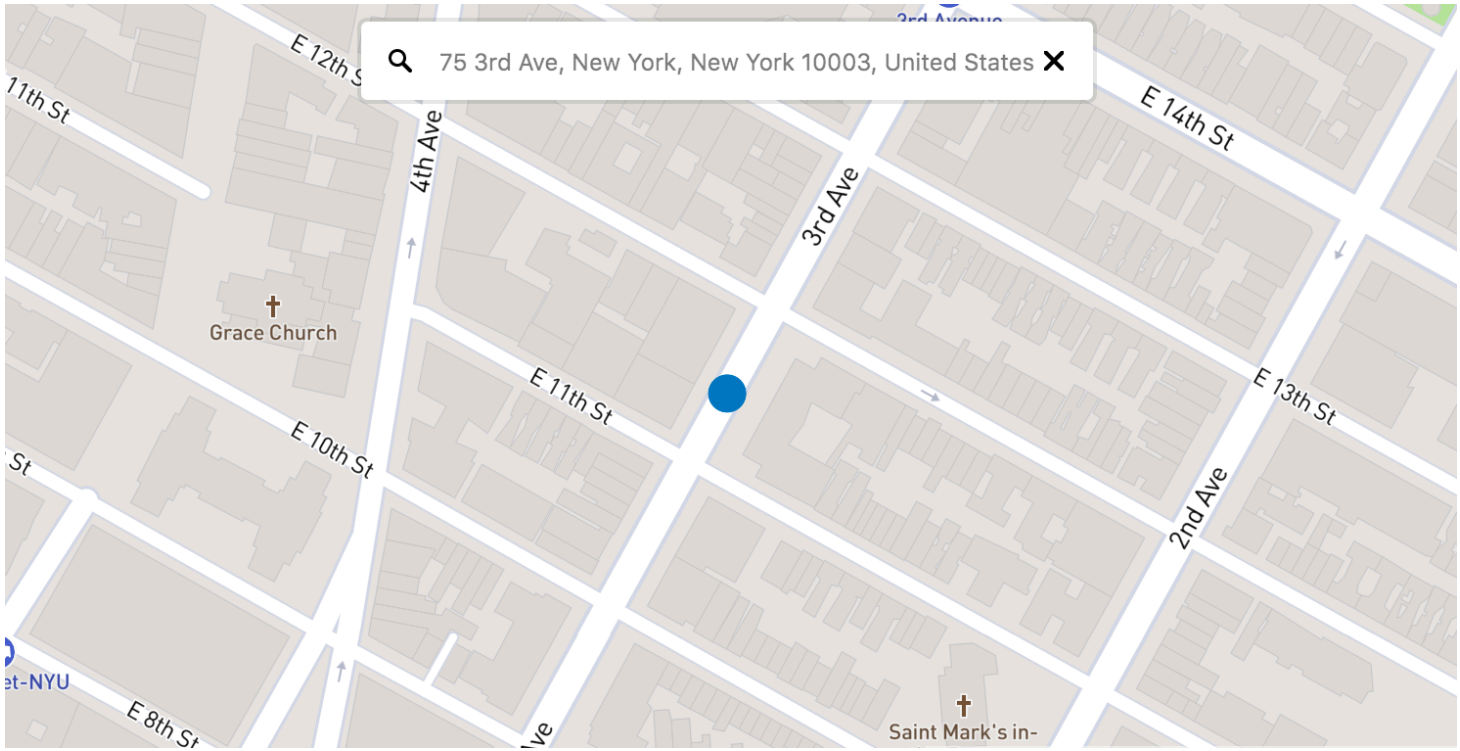
- Classifier Kernel: Multi-label Logistic Regression with Limited-Memory BFGS
- Datasets: Pickup coordinates of Yellow/Green/Uber
- Input: given any location in NYC
- Output: suggestion to users on whether to choose Yellow Taxi, Green Taxi or Uber in terms of pickups frequency in that location
- Training/Testing data split: 80%/20%
- Accuracy: 60%

Model Inputs and Outputs

Given a point from user (-74.01,40.76)
➔ Take adequate random points around the given point (-74.01,40.76)
➔ Spark-Submit
➔ Output prediction based on the top result, which in this case is Uber

Combine the model with UI

Address given by user ➔ Google Geocoding API
➔ Fetch coordinate (X_0, Y_0) as the input
➔ Output prediction based on the given point among Yellow/Green/Uber



Recommendation System

We propose a recommendation system as follows:

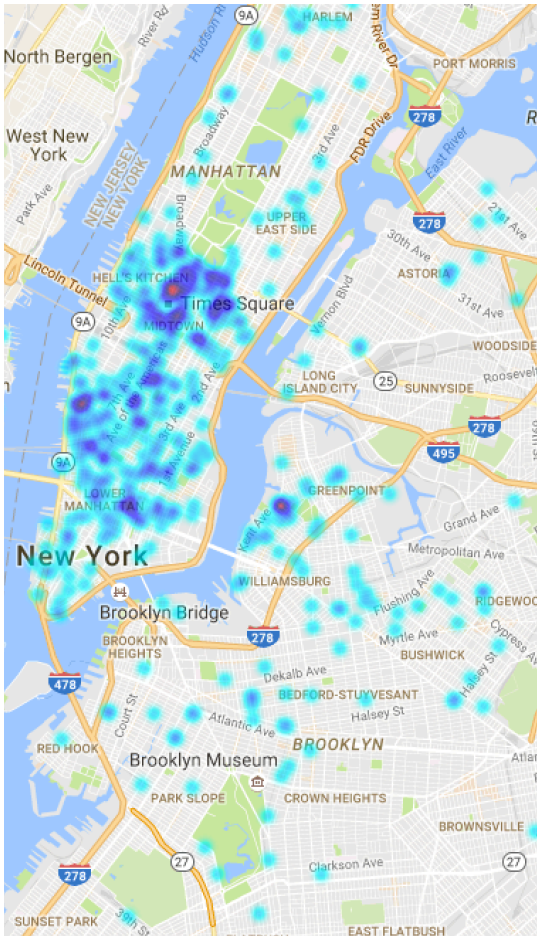
- **Feature: Given a new point, obtain the points region, then classify the region into yellow, green or Uber**
- **For passengers: given any location in NYC, can suggest to passengers whether yellow taxi, green taxi or Uber is an optimal choice**
- **For Taxi drivers: spatially avoid areas predicted to be dominated by Uber and temporally focus on the dates with higher historical pickups**
- **For Uber drivers: spatially focus on the zones with higher historical pickups**

Applying Model – Test assumption

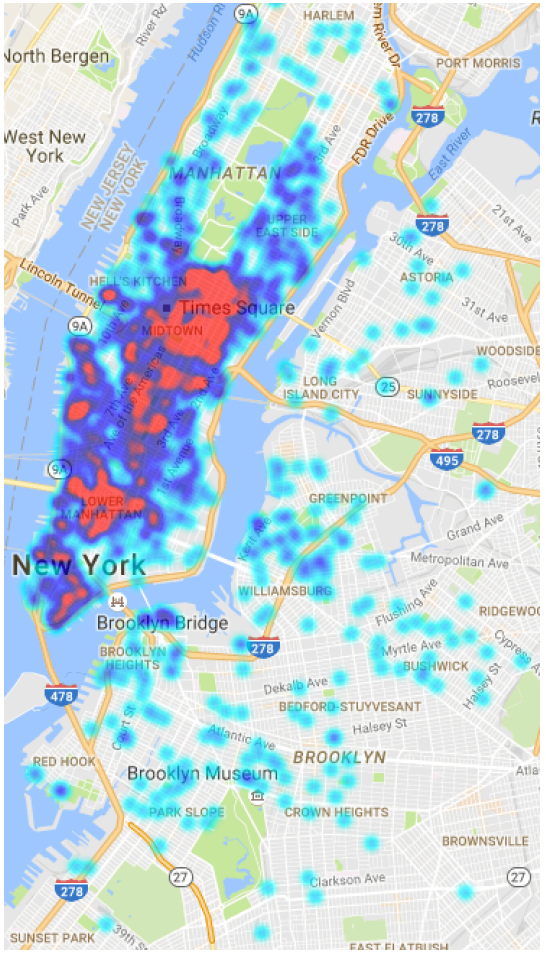
- Sample: 3 million random location points for testing
- Predicting on 2014-04 model: 599895 points were classified as Uber
- Predicting on 2014-08 model: 798479 points were classified as Uber, predicted 200K more points as Uber
- Conclusion: Uber is directly affecting the number of taxi pickups, previous taxi heavy regions are gradually taken by Uber
- Insight: NYC Taxi Commission needs to change strategy on which locations to send taxi

Location Heatmaps - Uber

- Heatmaps also reveal that Uber is expanding and taking over taxi, which shows Uber's business strategy has directly influenced taxi pickups



2014-04



2014-08

Future Works

- Improve classification accuracy
- Try different classifiers and compare results
- Incorporate more features into the service architecture

Thank You!