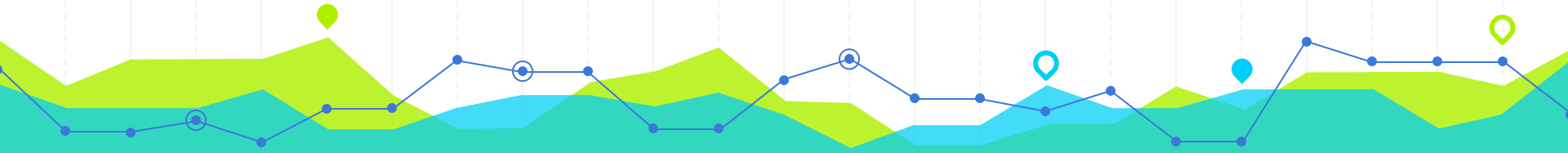
Exploring NYC Transit Data

Using open source tools



Midterm Project Presentation
DATASCI W209
March 10th, 2016

James | Drew | Greg

Introduction

NYC Taxi & Limousine Commission

The NYC Taxi & Limousine Commision (TLC) has made publicly available all green and yellow taxi trips between 2009 and 2015. This is a massive data source at > 1.1 Billion records!

+

Uber, by way of an information request submitted by 538 under FOIL, has made available NYC pickup data between April & September 2014 and January - July 2015, respectively .

UBER

+

citi bike

System Data is made readily available on Citi Bike Share trips

NYC OpenData

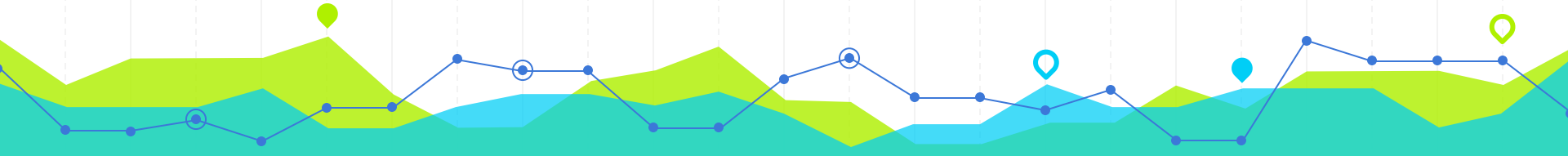
Relevant Open and accessible datasets include:
1. Weather
2. Felony Incidents
3. MTA information - turnstile usage

=

Visual Research Potential

A unique opportunity to:

- Build on some of the unique and amazing work done by:
  - 538,
  - Chris Whong
  - Todd Schneider
- Leverage techniques and frameworks taught in the course: downsampling, context, form and D3 respectively.
- Really tell a story through data - explore questions about NYC *in a way not traditionally* possible



Goal & Motivations

Acquire, integrate and analyze NYC open transit data to uncover macro transit trends that can be further explored in a narrow context to tell a story about a specific group of people

Hypothesis

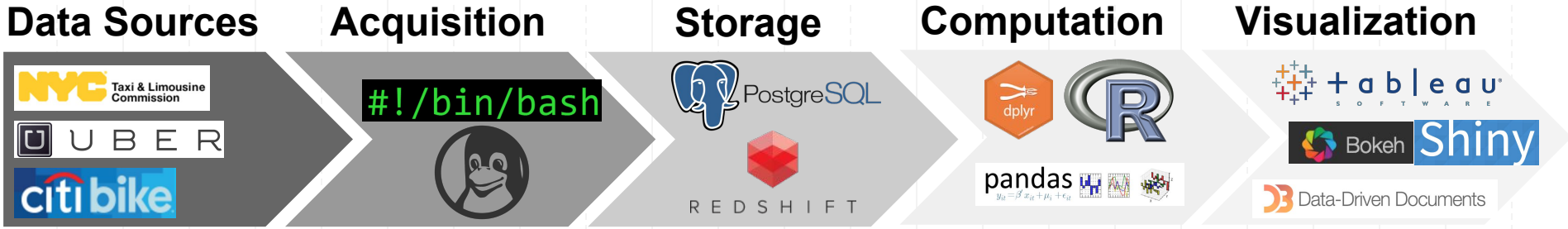
- We hypothesize that robust well integrated transit datasets can answer real questions about how New Yorkers (r*ead:* humans) live, work and play
- Any increase or decrease in demand for NYC transit modalities has far reaching implications given transportation's importance and proxy for economic activity (eg. 175 million taxi rides in 2015)
- NYC is the perfect petri dish for analysis with its sky high population and transport density

Applicability

- We seek to visually explore both broad and narrow conceptual questions
- **Broadly**
  - Can we pinpoint the efficiency of the transit system at various points in time?
  - Can we visually infer how exogenous changes influence modality choice / driver behavior?
- **Narrowly**
  - Can we tell a story about NYU students and their economic behavior given transit choice and options?*

*to be further developed / evolved

Architecture / Process Overview



- Download entirety of NYC Taxi, Uber, Citibike and other selected open data via custom bash scripts
- Parse & clean, standardize information and load into a `Trips` schema based on dataset conformity and relevancy

- Preprocess into PostgreSQL & PostGIS to obtain geometry data
- Join main transit data together and on attributes from other open data (Weather, Felony Incidents and MTA information)
- Load processed data into Redshift for high performance columnar analysis ability
- Leverage R / Python to investigate particular questions

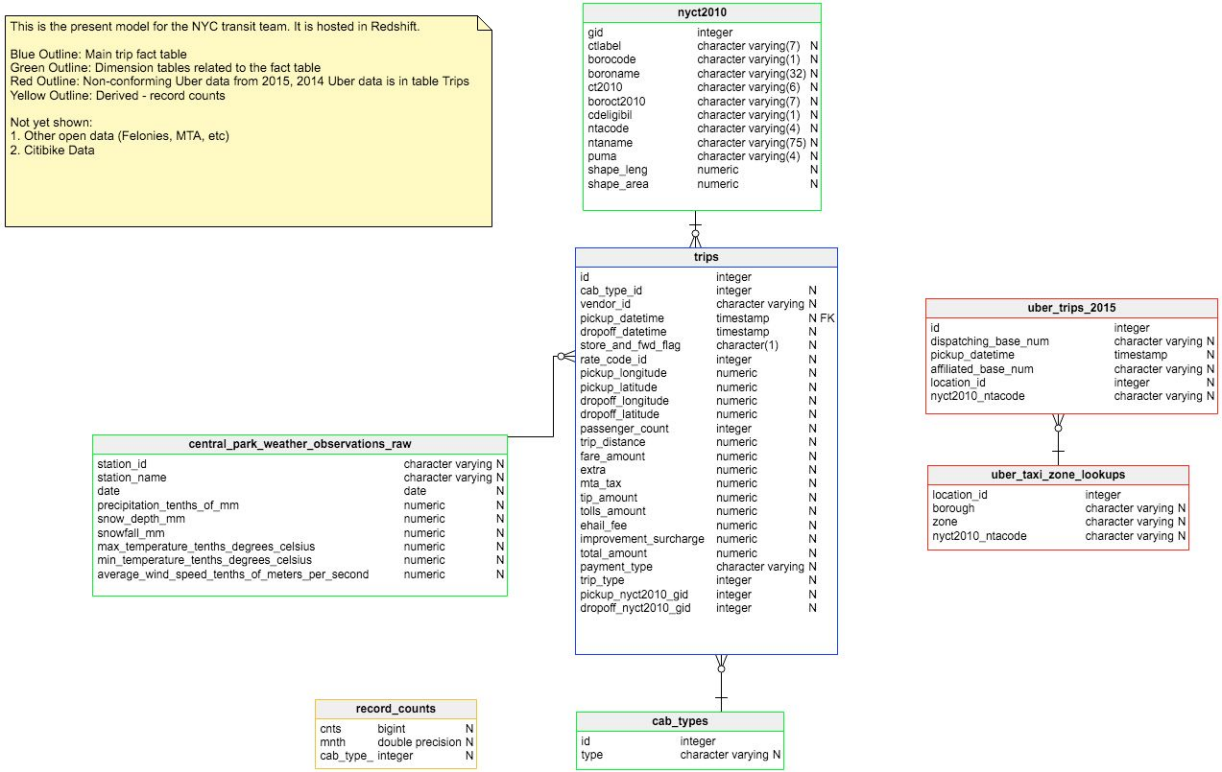
- Create initial prototype in Tableau
- Visualize final results **as a web-app** via Bokeh or Shiny & leveraging D3 for map visuals

ER Diagram

This is the present model for the NYC transit team. It is hosted in Redshift.

Blue Outline: Main trip fact table
Green Outline: Dimension tables related to the fact table
Red Outline: Non-conforming Uber data from 2015, 2014 Uber data is in table Trips
Yellow Outline: Derived - record counts

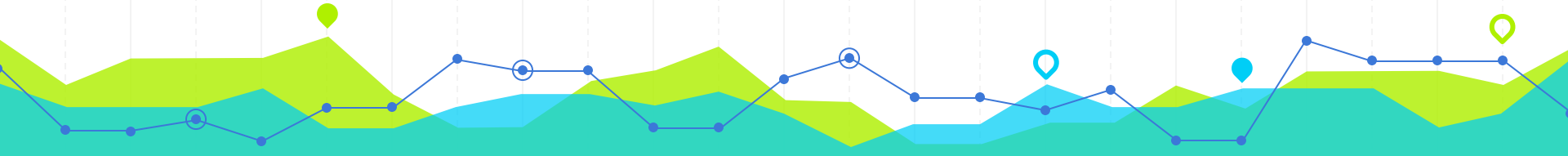
Not yet shown:
1. Other open data (Felonies, MTA, etc)
2. Citibike Data



Target Audience

- Urban Planners
  - “Do trends in traffic flow predict a need for new / different zoning or better transportation infrastructure?”
- Traffic Engineers
  - “Does indicated traffic flow conflict with existing traffic constraints such as large impending construction projects?”
- Sociologists / policy-makers / economists
  - “Do trends in taxi vs. Uber / Lyft usage serve the public in the best way?”
- Data scientists / visualization enthusiasts
  - “What cool examples can I see for visualizing data?”

Data Sources



Taxi Data

NYC Taxi & Limousine Commission

Synopsis
This is the granddaddy dataset and is extremely comprehensive in both size and breadth including: VendorId,

Summary Information

- URL: Taxi TLC Data / DD
- Date Range
  - Yellow: 2009 - 2015
  - Green: Aug 2013 - 2015
- Total Files
  - 116
- File Size (Total // Average):
  - 267G // 2.3G
- Average Record Count:
  - 14M

Prior Work done by and self published by Todd Schneider [Image attribution]

Uber Data

UBER

Synopsis
This dataset is fairly limited and only includes: date and time of pickup, lat/long, and a base code that corresponds to a TLC station name

Summary Information

- URL: Uber Data
- Date Range
  - July 2014 - September 2014
  - Jan 2015 - June 2015
- Total Files
  - 7
- File Size (Total // Average):
  - 727M // 103M
- Average Record Count:
  - 1M

Manhattan Dominates Both Ubers And Cabs

Residential pickup rates by **borough**

	PICKUP INDEX (100 = AVG.)	
BOROUGH	UBER	CABS
Bronx	8	9
Brooklyn	79	29
Manhattan	357	431
Queens	19	22
Staten Island	1	0

Pickup data from April through Sept. 2014

SOURCE: TAXI & LIMOUSINE COMMISSION

Lower Income Means Fewer Pickups

Residential pickup rates by **median income** of census tract

	PICKUP INDEX (100 = AVG.)	
MEDIAN INCOME	UBER	CABS
$0-25K	21	26
25-50	32	39
50-75	75	67
75-100	160	167
100-125	419	462
125-150	649	725
150+	539	564

Pickup data from April through Sept. 2014

TAXI & LIMOUSINE COMMISSION, CENSUS BUREAU

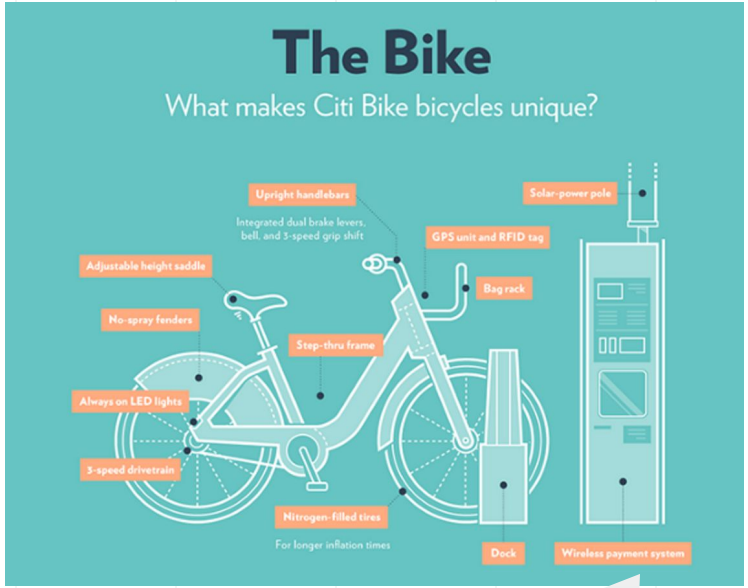
Some prior work done by and published on http://fivethirtyeight.com/ [Image attribution]

Bikeshare Data citi bike

Synopsis
This dataset provides trip history information including: Duration, Start and End Time, Station Names & Lat / Long, Bike ID, User Type, Gender & Birthday

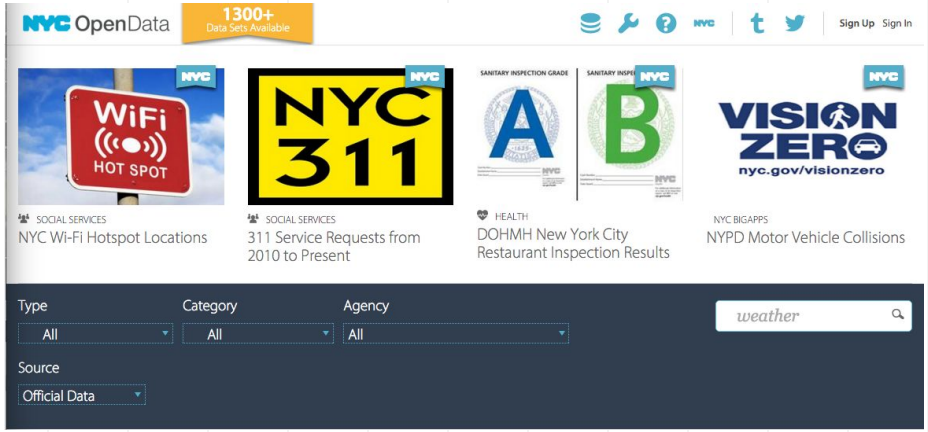
Summary Information

- URL: System Data
- Date Range
  - July 2013 - December 2015
- Total Files
  - 30
- File Size (Total // Average):
  - 4G // 130M
- Average Record Count:
  - 1M



For those not familiar w/ bike share data, the units when docked roughly look like this. [Image attribution]

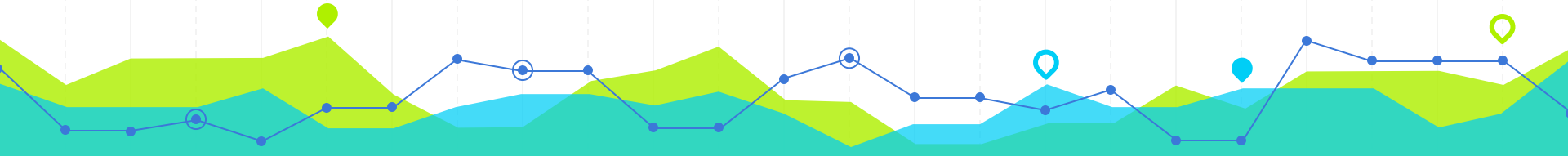
NYC Open Data



*to be further developed / evolved

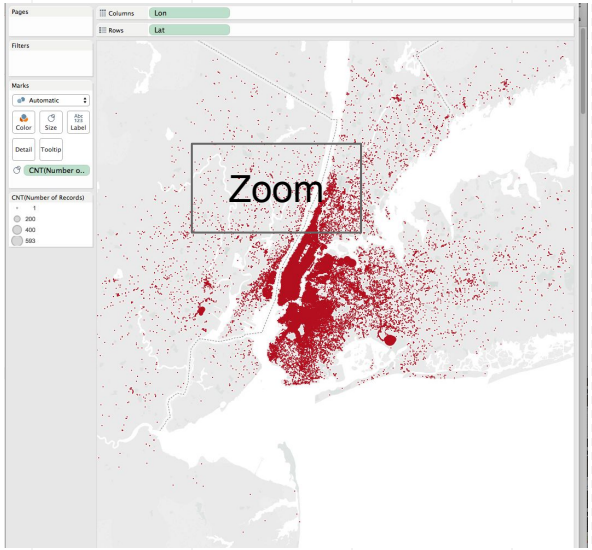
- **Crime**: using long/lat location to evaluate impact on transit
- **Weather**: how does climate conditions impact transit
- **Subway/Bus/Rail**: volume, location implications on other forms of transit

Current Mockup

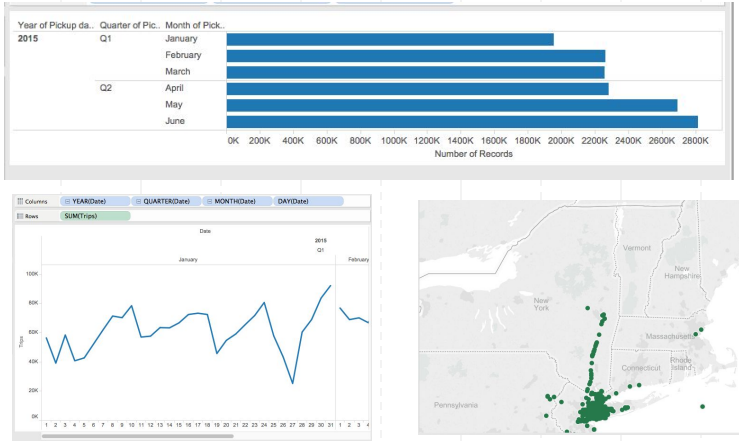


Anticipated Interactions

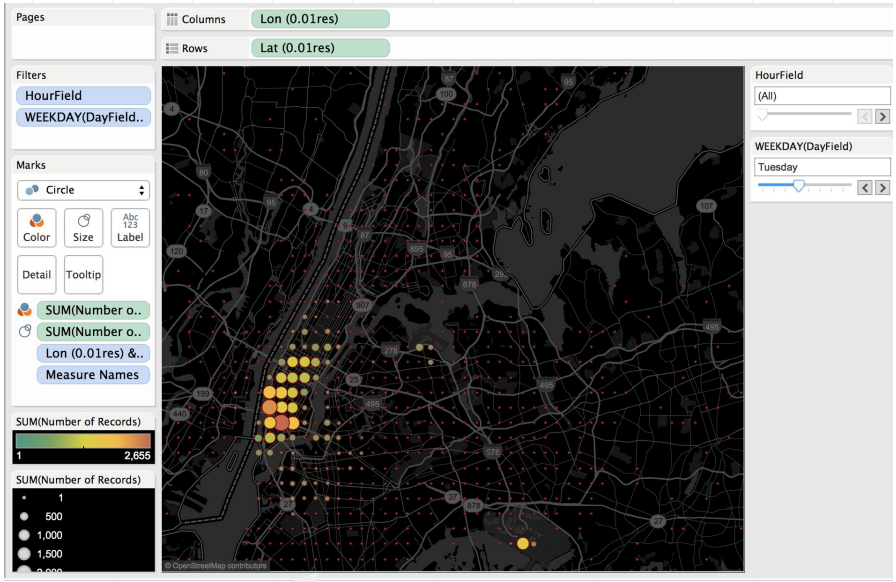
Faceted Filter



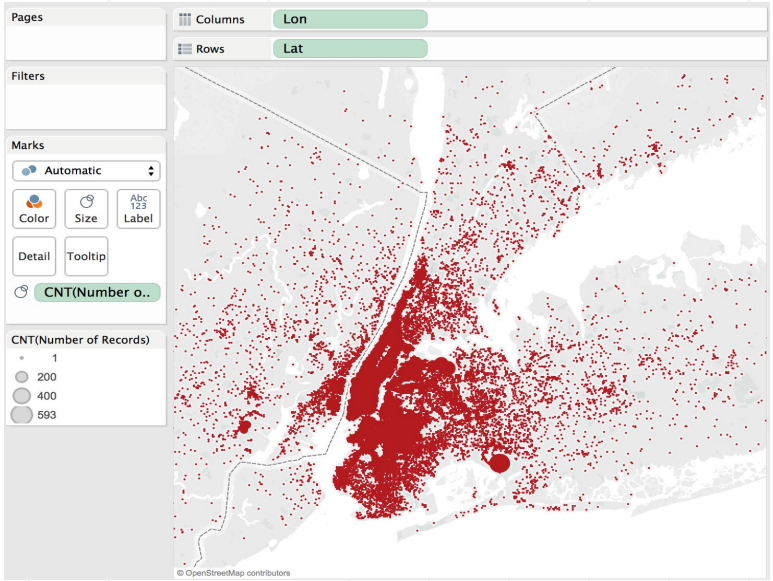
- Overview, zoom, filter, details-on-demand
- Navigate from broad to specific analyses
- Potential visualizations optimized for personas/tasks
- Interactivity: brushing and linking, overview + detail, zooming and panning



Mockup - Iteration #1

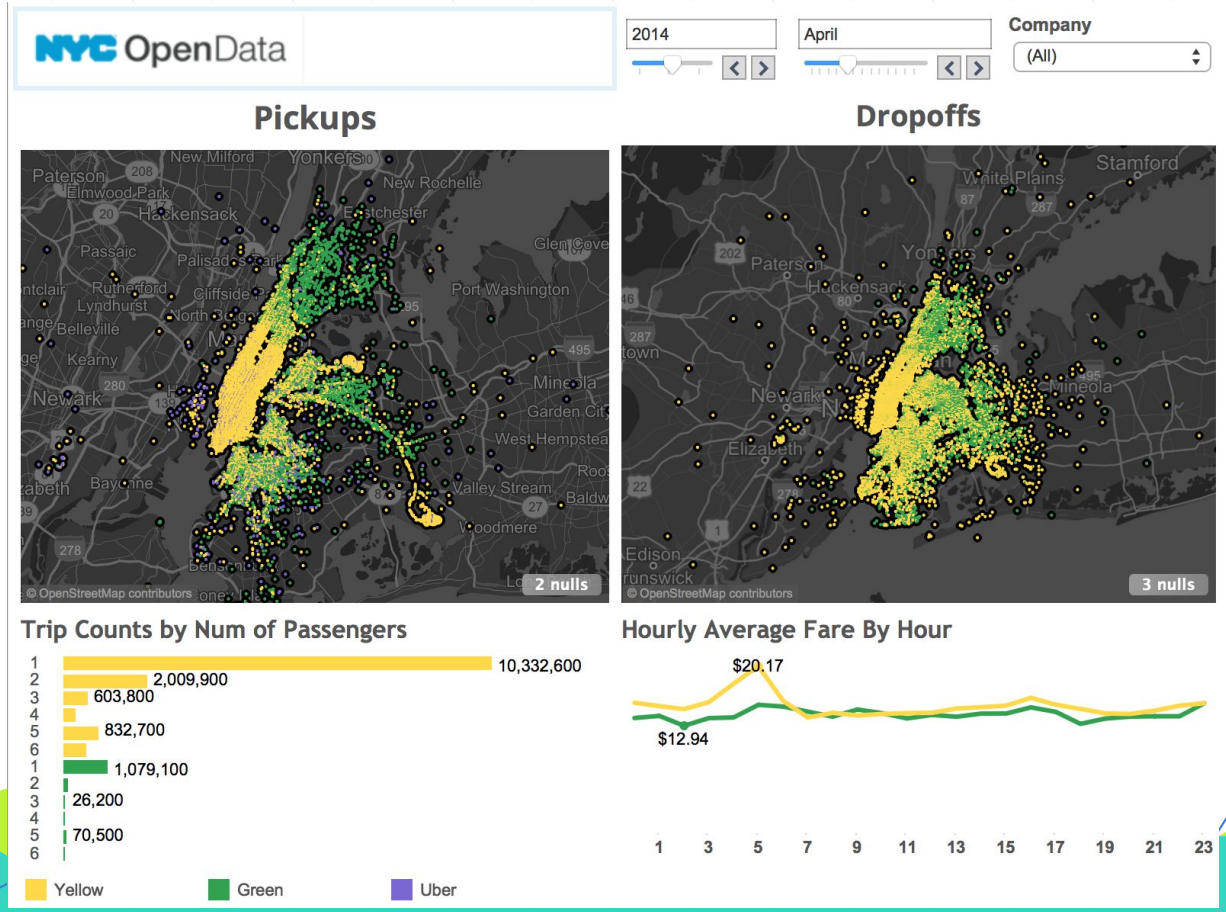


Some work done to explore and aggregate counts by lat/long pairs over time

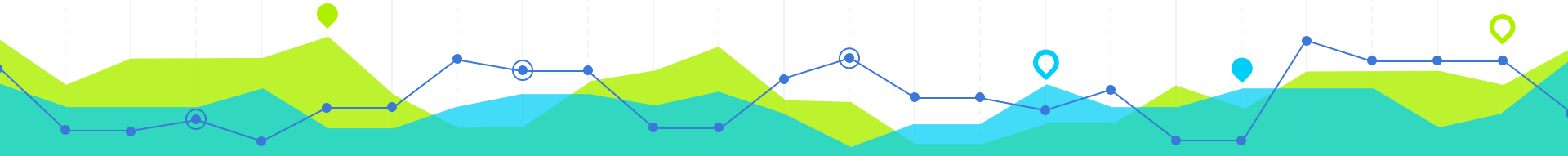


An additional attempt to map and show how disbursed all Uber pickups are

Mockup - Iteration #2



Testing Plan



Expected Feedback Capture

Initial Thoughts (Verbatims):

1. Without clicking on anything, what do you expect each drop-down select to do?
2. Examine the titles for each of the visualizations. What do you think each title means?
3. What do you expect will happen if you choose a [ ]? What about after choosing more than one?

Interactive Exercises

1. Depending on final visualization form, intend on capturing user feedback that requires interactive exploration and discovery
2. Likely include 5 - 10 “challenges” for the user to complete

Usability Questions

1. Test questions along the lines of: the following will be administered:
    a. User Satisfaction
    b. Usage Simplicity
    c. Productivity
    d. Useful Error Messaging
    e. Ease of Navigation
    f. Information Organization
    g. Meeting expectations

Final Analyses & Feedback

1. Was the layout intuitive? If not, how would you adjust the layout?
2. Beyond the questions listed above, what improvements could be made to the tool?
3. What do you wish the tool could do for you?

Sending Us Feedback

jamesgray@ischool.berkeley.edu

gregce@gmail.com

drewplant@ischool.berkeley.edu

Questions?