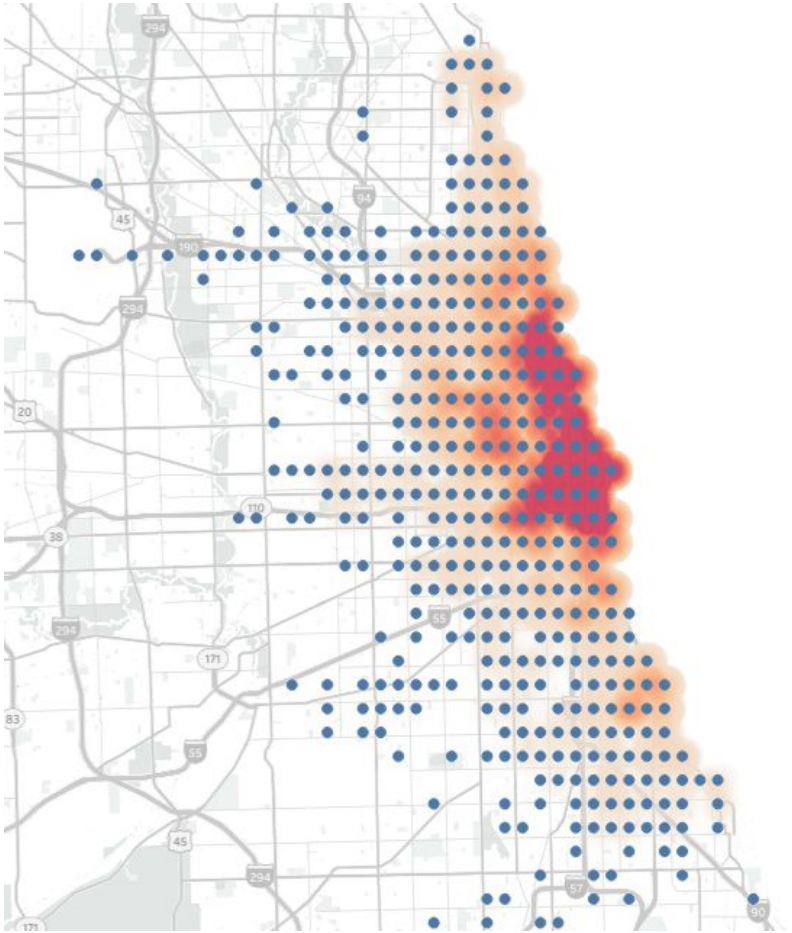
Business Opportunities for Last Mile Transportation

Neeraj Tadur, Scott Shepard, Dan Dobrzynski, Kevin Stutenberg

Contents

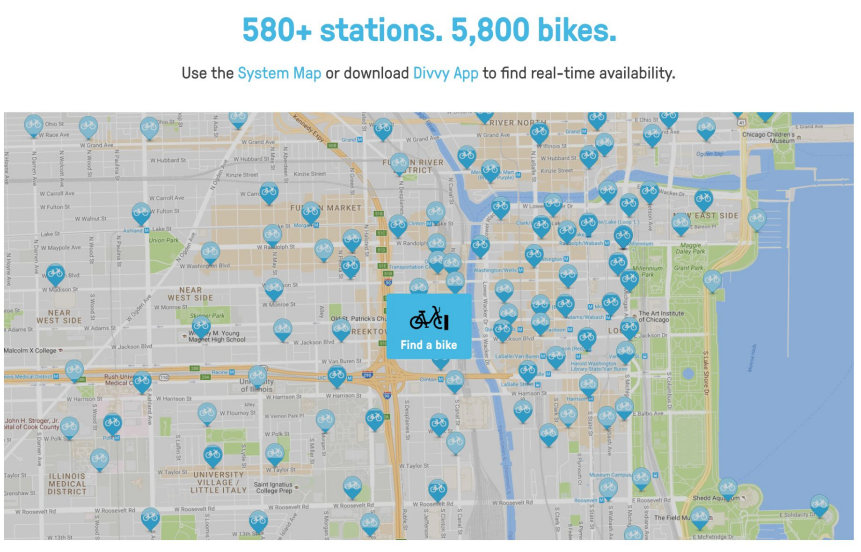


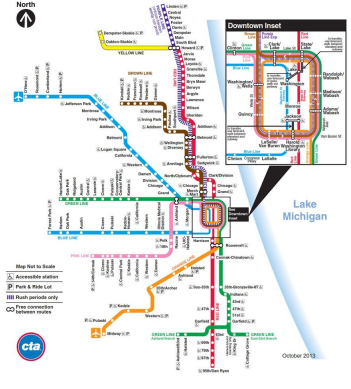
Executive Summary

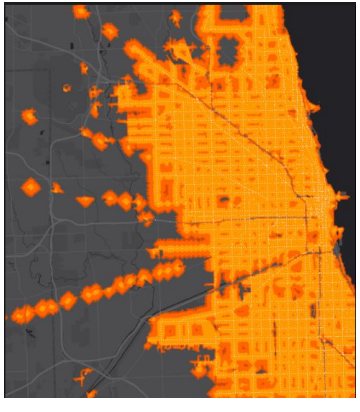
- YesSQL is a data consultancy that works with startups and small companies in need of full-service data warehouse development and strategic planning.
- YesSQL has been tasked by Citrus, a start-up **electric scooter company**, who is looking to **expand services in the Chicago** Metropolitan area.
- Citrus is looking to help **solve the "last-mile" problem**, connecting parts of the city that are close to but **not within walking distance** of major transportation connector hubs.

Business Use Case

- Citrus understands that there are **other ‘last-mile’ commuter options** such as **taxi services** and **Divvy** bike rental.
- YesSQL will **target, load, warehouse, report, and visualize** relevant metropolitan transportation data from CTA, taxi services, and Divvy.
- Citrus executive leadership will be able to target deployment locations where their **scooter solutions can supplement or introduce last-mile solutions.**
- YesSQL solution should be one that will ensure a successful deployment and be extensible such that Citrus executives can use it **beyond initial deployment and into future expansion projects.**







Data Pipeline

Data Sources

Data Name	Table	Records
Divvy	Divvy Stations	602
	Divvy Trips	14,496,257
CTA	CTA Stations	300
	CTA Daily Ridership	910136
Taxi	Taxi Trips	10,565,534

Data Overview - Divvy

Data comes from Divvy JSON feed and from ZIP archives

- Stations
  - location information
  - capacity and available bikes
- Trips
  - time
  - from and to stations

```
8
9 •  select * from trips limit 5;
```

100% 13:3

Result Grid | Filter Rows: Search | Export:

index	trip_id	starttime	stoptime	bike_id	tripduration	from_station_id
0	4118	2013-06-27 12:11:00	2013-06-27 12:16:00	480	316	85
1	4275	2013-06-27 14:44:00	2013-06-27 14:45:00	77	64	32
2	4291	2013-06-27 14:58:00	2013-06-27 15:05:00	77	433	32
3	4316	2013-06-27 15:06:00	2013-06-27 15:09:00	77	123	19
4	4342	2013-06-27 15:13:00	2013-06-27 15:27:00	77	852	19

Data Overview - CTA Rail

CTA rail ridership data from Chicago data portal

- CTA Station data set (300 x 18 factors)
  - https://data.cityofchicago.org/Transportation/CTA-L-Rail-Stations-kml/4qtv-9w43
  - CTA rail station specific data
  - Location (Lat/Long),
  - Station ID - Key for ridership data
  - Binary indicators for Rail Line
  -
- CTA Daily Ridership data sets (881,184 x 5col)
  - https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Daily-Totals/5neh-572f
  - Number of rides per day / station
  - Independent of Rail direction
  - Daytype- (Weekend/ Weekday/ Holiday)

Image courtesy of: https://www.transitchicago.com/holidayfleet/

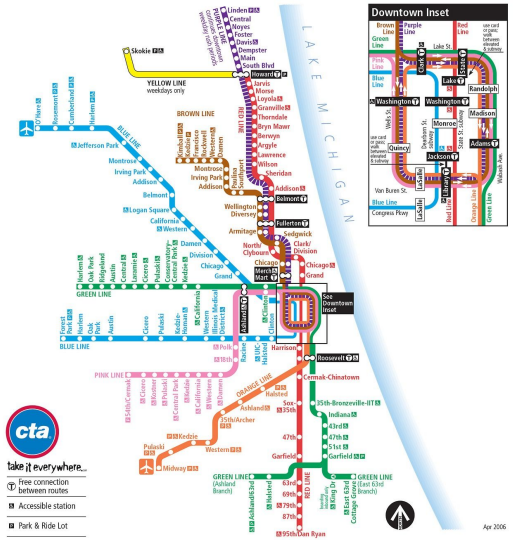


Image Courtesy: https://chicagotransitguide.com/maps/cta-map/

Data Preview:

Trip ID (character)	Taxi ID (character)	Trip Start Timestamp (character)	Trip End Timestamp (character)	Trip Seconds (integer)	Trip Miles (double)	Pickup Census Tract (double)	Dropoff Census Tract (double)	Pickup Community Area (integer)
01deea4e27a483f1fe86aa5c8168ff066103f921	bd6dfef80649b77a404ad83b02e7a159af9119641d5d3883a4e7af473e11bde4305fba8cbf94d25660bfe776d2d9a83257d47c97c7686c46df529d0eee678355	01/27/2016 01:30:00 PM	01/01/1900 12:00:00 AM	NA	0.0	NA	NA	NA
5b554a40e15bd31e4bc23f5dc9463ea17495c489	bd6dfef80649b77a404ad83b02e7a159af9119641d5d3883a4e7af473e11bde4305fba8cbf94d25660bfe776d2d9a83257d47c97c7686c46df529d0eee678355	01/27/2016 03:30:00 PM	01/01/1900 12:00:00 AM	NA	0.0	NA	NA	NA
a4c71ba5d2c54a0655c3fa5c3f19ba5b1e1e3027	bd6dfef80649b77a404ad83b02e7a159af9119641d5d3883a4e7af473e11bde4305fba8cbf94d25660bfe776d2d9a83257d47c97c7686c46df529d0eee678355	01/28/2016 03:45:00 PM	01/01/1900 12:00:00 AM	NA	0.0	NA	NA	NA
0380f8478ff2d0c4fc304f7c6e911550a895a4f0	bd6dfef80649b77a404ad83b02e7a159af9119641d5d3883a4e7af473e11bde4305fba8cbf94d25660bfe776d2d9a83257d47c97c7686c46df529d0eee678355	01/28/2016 03:45:00 PM	01/01/1900 12:00:00 AM	NA	0.0	NA	NA	NA
401f4434967b5075446f1a3902e36536987a94d9	bd6dfef80649b77a404ad83b02e7a159af9119641d5d3883a4e7af473e11bde4305fba8cbf94d25660bfe776d2d9a83257d47c97c7686c46df529d0eee678355	02/09/2016 09:30:00 AM	01/01/1900 12:00:00 AM	NA	0.0	NA	NA	NA

Data Overview - Chicago Taxi Data

"Taxi trips reported to the City of Chicago in its role as a regulatory agency. To protect privacy but allow for aggregate analyses, the Taxi ID is consistent for any given taxi medallion number but does not show the number, Census Tracts are suppressed in some cases, and times are rounded to the nearest 15 minutes. Due to the data reporting process, not all trips are reported but the City believes that most are."

- Taxi Trip Data- 2016, 2017
  - https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/data
  - Available Data
    - **Trip Start/End Date and Time**
    - **Taxi ID, company/ payment type**
    - **Geography**
      - Census Tract & Community Area
      - **Lat & Long- Pickup & Dropoff**
    - **Costs**
      - **Fare/ Tips / Tolls /Total Cost**
  - Very large dataset- Initial filtration through data portal

https://en.wikipedia.org/wiki/Yellow_Cab_Company

DESIGN - ETL

Fetch Data	Production Database	Summary & Linking Tables
Variety of methods: - Python - R - Manual loading Preprocessing - Compression - Column Reduction - Remove nulls	Data sources are in separate tables: - Divvy Stations - Divvy Trips - CTA Stations - CTA Daily Trips - Taxi Trips Unlinked to one another	Python & SQL Area Gridding - Link sources through grid_id Summary tables: - Divvy Daily Ridership - Daily Grid Activity

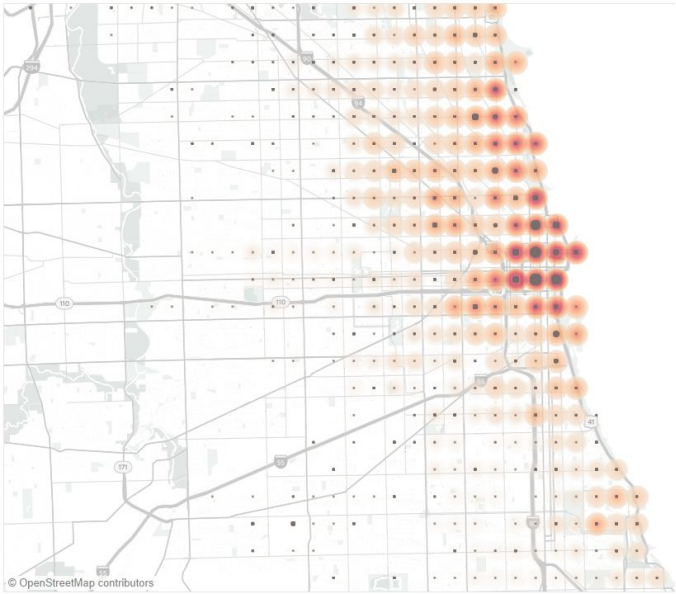
```
r = requests.get(divvy_url)
stations = r.json()
for station in stations:
    cursor.execute(add_station, station)
```

Github Link

ETL- Grid Key Development

- Primary and foreign keys
- Factors of interest- Timing and Location
  - Timing
    - Transformation of tabular data to DateTime format
  - Location- Development of unique grid key
    - ~1 km resolution (.01 deg) desired
    - Grid value- 8 character key created off of Latitude and Longitude
    - Grid used as key for all tables with latitude and longitude
      - DIVVY station, CTA station, and Taxi pickup/dropoff
    - Grid SQL Coding:

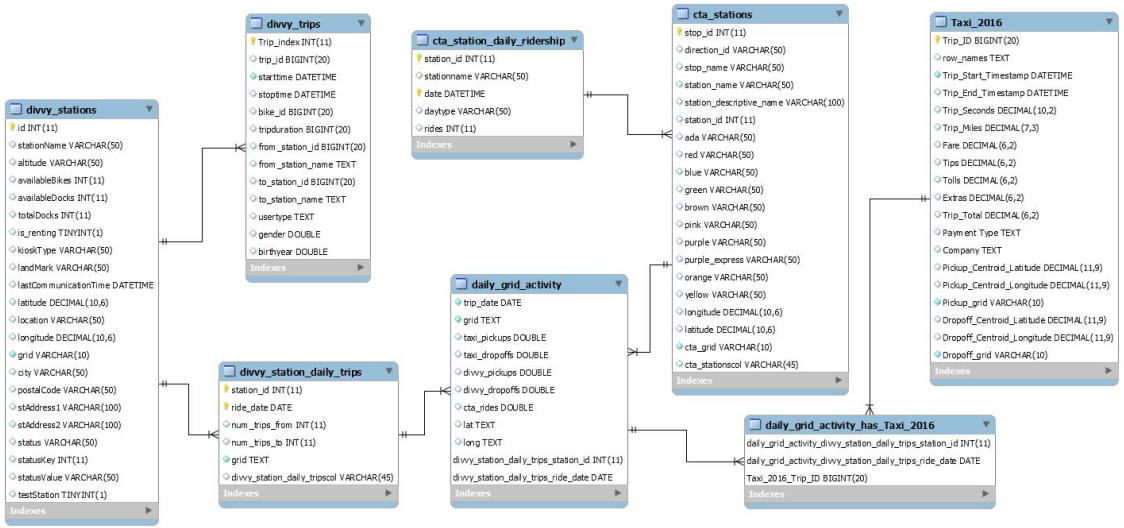
Taxi Intensity vs Divvy Intensity



Grid Key Example:
- **Lat: 41.90907**
- **Long: -87.90304**
- **Grid ID: 41908790**

```
GRID = REPLACE(REPLACE(concat(ROUND(Pickup_Centroid_Latitude,2),ROUND(Pickup_Centroid_Longitude,2)),'.',''),'-','');
```

Production Database

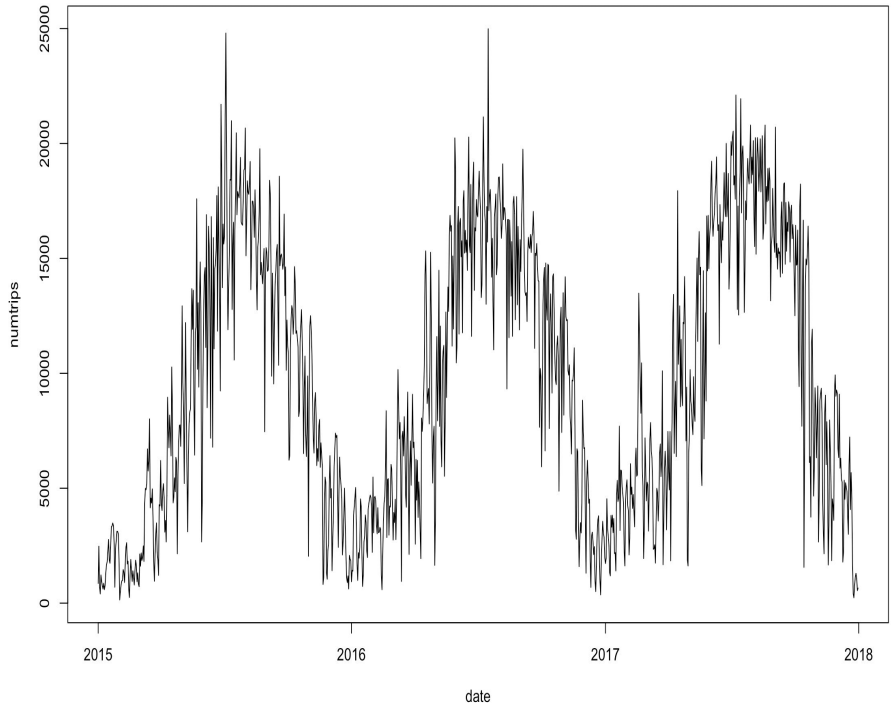


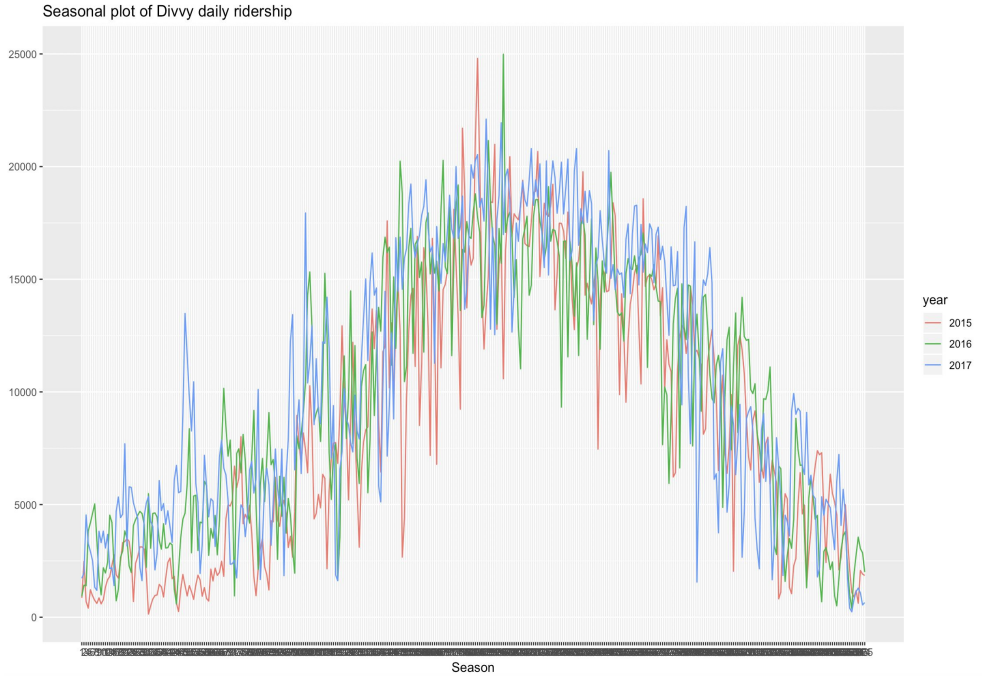
Normalized Data

- Primary normalized relationships are gridID, dateID (CTA, DIVVY), Trip_Start_Timestamp (Taxi)
- Unique IDs from source data include Trip_IDs, station_ID, stop_ID
- Database keys held to not null include: Date and count variables for trip data / Grid location for all location data

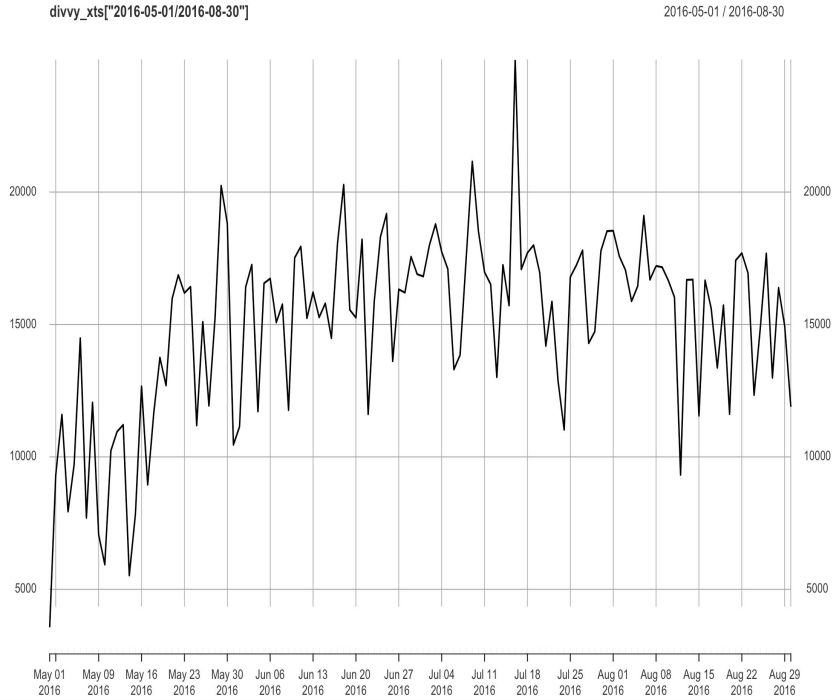
Results

Time Series Analysis





Seasonality Time series plot



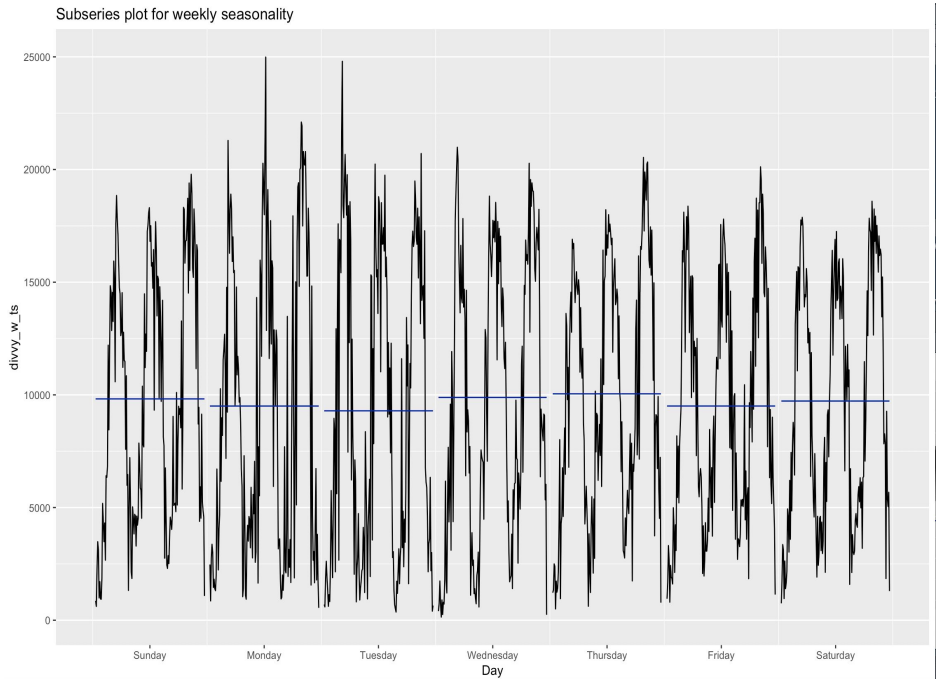
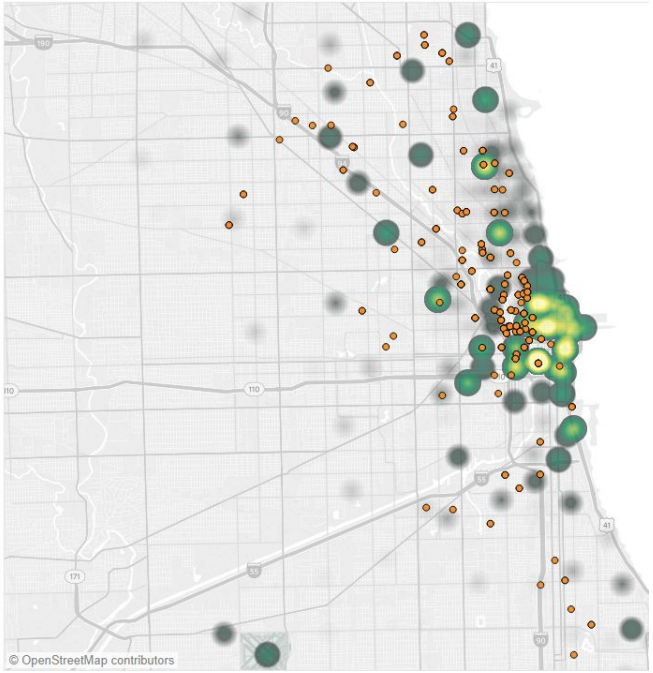
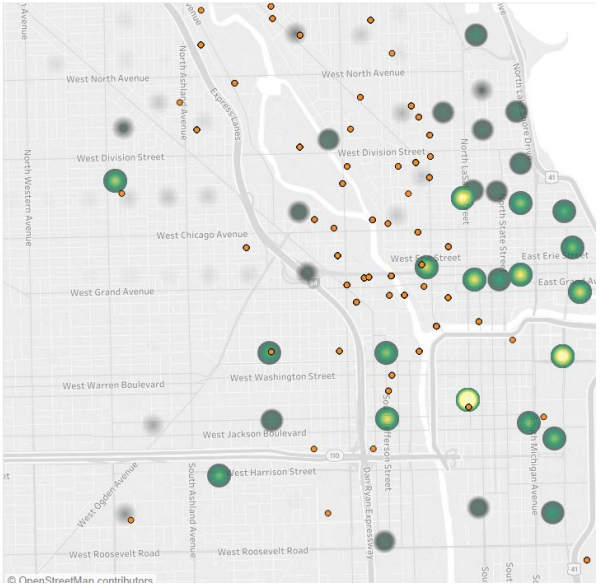


Tableau - Mapping Analytics

Dual Axis Taxi Pickups(Density) and Divvy Stations



Dual Axis Taxi Pickups(Density) and Divvy Stations



CTA Stations /Nearby Divvy Stations

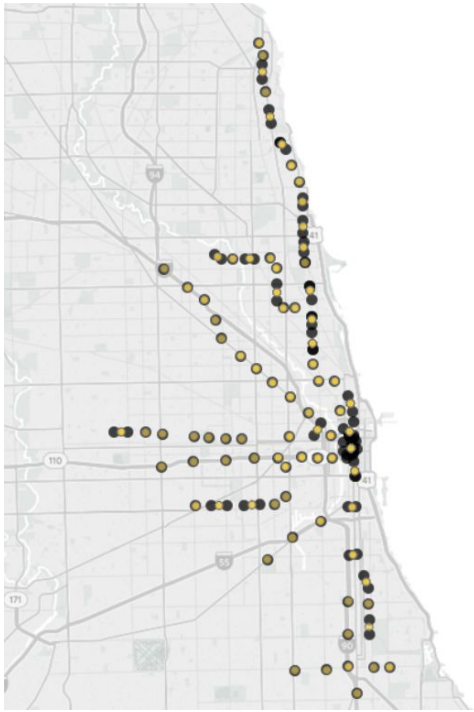
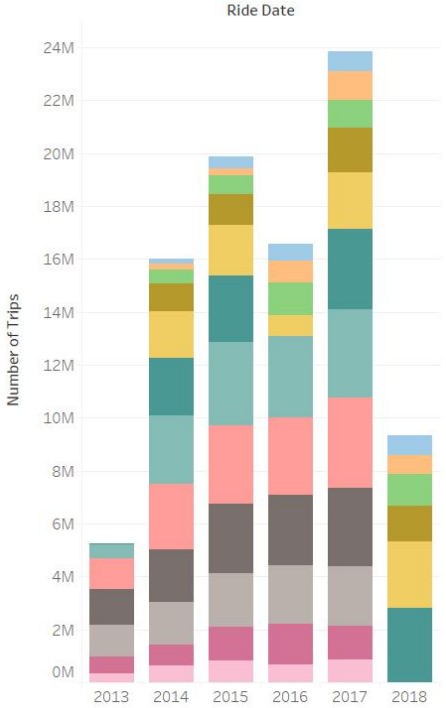


Tableau - Numerical Analytics

Total Divvy Trips By Year and Month



CTA Station to Divvy Distance

Station Name	
54th/Cermak	1.498
95th/Dan Ryan	0.989
95th/Ran Ryan	0.989
Cumberland	4.997
Dempster-Skokie	2.597
Forest Park	2.250
Harlem	1.741
Harlem/Lake	1.483
Jefferson Park	1.057
Kostner	0.922
Linden	0.654
Midway	3.106
O'Hare	8.124
Oak Park	0.999
Oakton-Skokie	2.458
Rosemont	6.000

Divvy Station (Largest Ridership)

Station Name	
Canal St & Adams St	108.4
Canal St & Madison St	85.4
Clinton St & Madison St	89.0
Clinton St & Washington Blvd	117.8
Columbus Dr & Randolph St	78.5
Daley Center Plaza	64.0
Franklin St & Monroe St	73.5
Kingsbury St & Kinzie St	85.8
Lake Shore Dr & Monroe St	126.2
Lake Shore Dr & North Blvd	99.0
LaSalle St & Jackson Blvd	60.8
McClurg Ct & Illinois St	62.0
Michigan Ave & Lake St	60.4
Michigan Ave & Oak St	92.0
Michigan Ave & Washington St	74.0
Millennium Park	90.8
Shedd Aquarium	76.4
Streeter Dr & Grand Ave	191.7
Theater on the Lake	109.1

Recommendations / Future Work

- Recommendations:
  a. Downtown activity is order of magnitudes greater than neighborhoods
  b. Focus locations on pockets of dead zone activity just outside the loop
  c. Investigate locations near CTA stations that do not have closeby Divvy options

Next Steps:

- Create separate reporting Database
  a. Flesh out schema for reporting
  b. Seperate from
- Additional Data
  a. Weather
  b. Added data from taxi trips
  c. Uber / Lyft Data addition
- Additional Analysis
  a. User level analysis - Trip clustering by user, age, distance, time of day.

Lessons Learned:

- Data processing
  a. OpenRefine memory allocation- RAM dependent
- Data Storage
  a. AWS- Requires paid server connections, can be SLOW!
  b. mySQL import can be SLOW - R and Python good alt.
  c. Table/Key development best during table creation
  d. Local storage improves speed in this instance
- Tableau Issues
  a. Lacked functionality with moderate server connection

Appendix

Data References

Divvy public data sources:

- https://www.divvybikes.com/system-data
- https://feeds.divvybikes.com/stations/stations.json

We plan to use the Divvy public datasets and create a database that will help us visualize and explore popular routes, which stations they connect to, when they are used, and what for.

CTA “L” data:

Daily Ridership: https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Daily-Totals/5neh-572f

Station Locations: https://data.cityofchicago.org/Transportation/CTA-L-Rail-Stations-kml/4qtv-9w43

Taxi data set: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/data