Green Cab Data Challenge

Miya Wang

nwang26@fordham.edu

Research Preview

Initiatives:

- Identify business risks and opportunities for driver-to-be
- Deliver industry overview to newbie driver
- Optimize revenue potential for veteran driver

Timeframe:

- August 1, 2013 – June 30, 2016

Report Includes:

- Green cab program landscape
- Time-wise business opportunities and strategies
- Place-wise business opportunities and strategies

DATA & Methodology

- Ridership
  - Crawled from NYC open data (over 45million records)
  - Timeframe: Aug 1, 2013 – Jun 30, 2016 (1064 days, 152 weeks)
- Weather
  - Gathered through Dark Sky API
  - Timeframe: Jan 1, 2013 – Dec 31, 2016 (29,999 hours)
- Festival
  - Collected through US Federal Holiday Calendar API
  - Timeframe: Jan 1, 2013 – Dec 31, 2016
- Zip code
  - Crawled through Geolocation Service API
  - Timeframe: Jan 1, 2016 – Jun 30, 2016

relational database

SQLite

- Clean and transform data (anomaly detection)
- Filter or subset or aggregate

APACHE Spark

small csv files

- Trend analysis
- Exploratory analysis
- Social network analysis

Key Findings and Recommendations

Opportunity:

- Green cab business has grown **stable** with regular earning per ride ($12.4) and regular ridership (1.5 million per month)
- Most of the time, business opportunities are **predictable** with accumulating passengers at regular **time and place**.
- Most passengers tip with a **generous** amount (22% on average).
- Overall, weather has **no** effect **most of the time**(90%), **boosts** ridership **occasionally** (6%) and **rarely** (<1%) **harms** the business.

Risk:

- A suspected periodical **stagnation** (from June to October) is found in ridership.
- Earnings per ride.is **shrinking** silently by losing several cents.
- Bad weather (e.g. snowy) on **weekend** has the ability to **largely** jeopardize business opportunities.

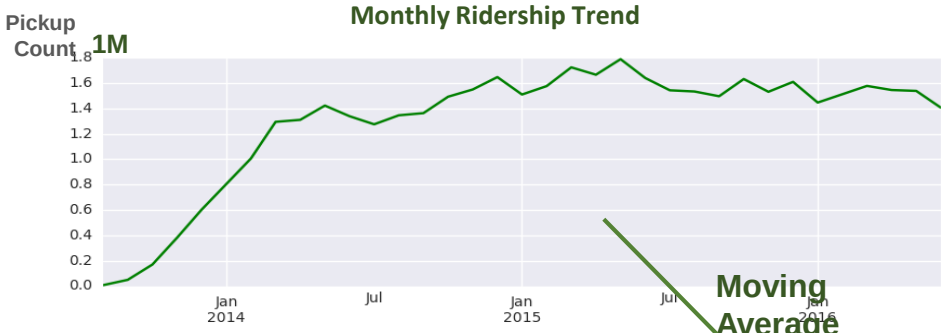
Strategies:

- Drive at 6pm, 4pm-Midnight every Friday and Saturday, 0-3am Sunday and 8am-9am Monday through Friday.
- Drive around Columbia University neighborhood, East Harlem, Astoria, Washington Heights, Long Island City and East Elmhurst.

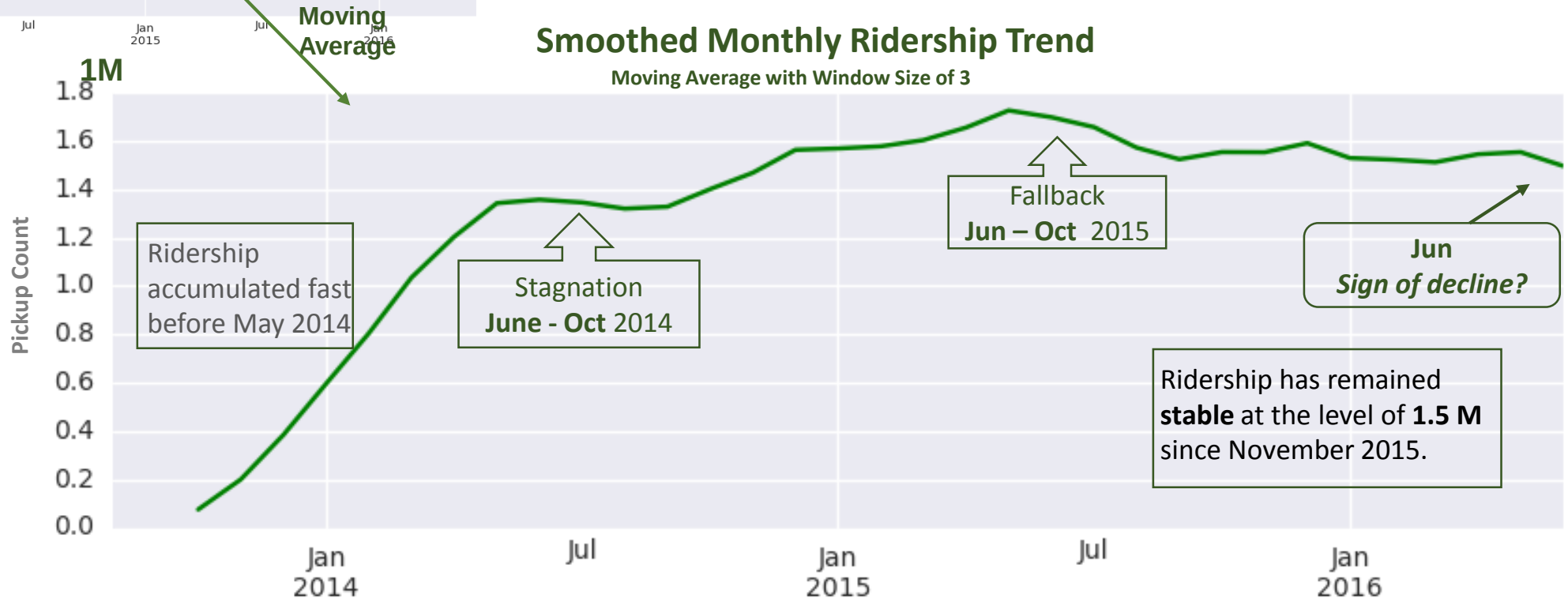
Landscape

- ridership
- earning per ride
- tip behavior

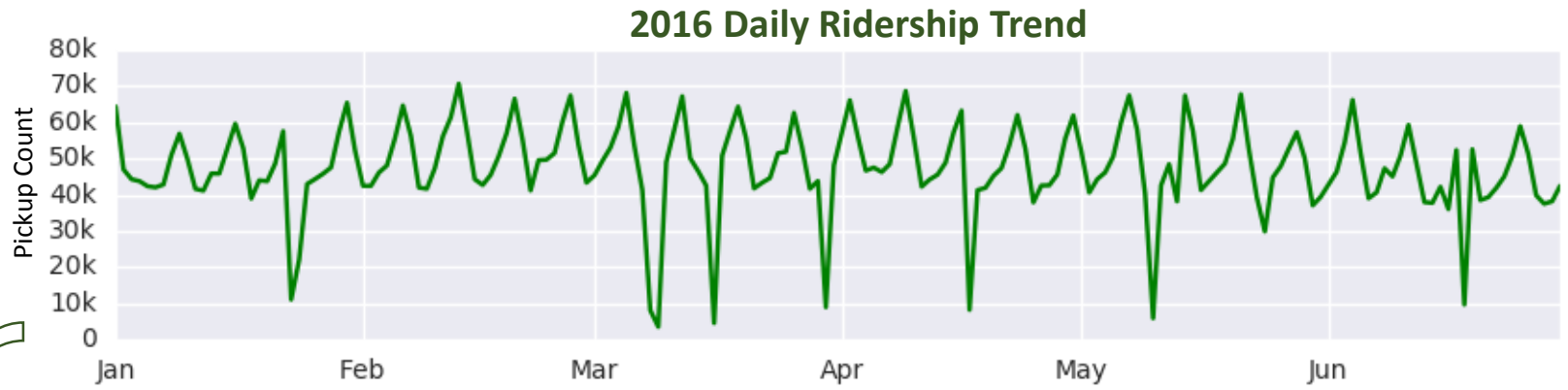
Ridership Has Grown Stable Overall, But With a Suspected Periodical Shrinkage



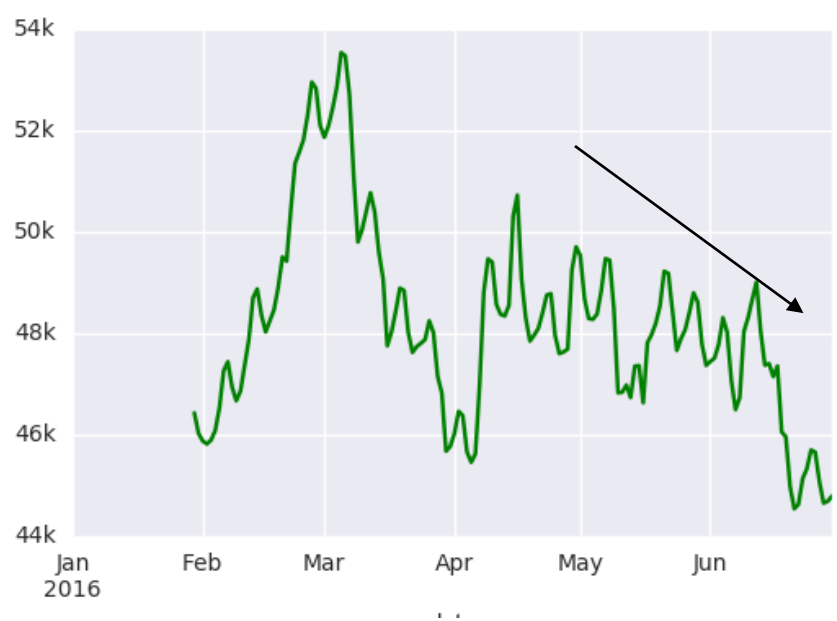
- During the period of **June to Oct** 2014, ridership experienced a stagnation.
- From **June to Oct** 2015, ridership experienced a fallback.
- In **June** 2016, ridership showed a sign of decline.

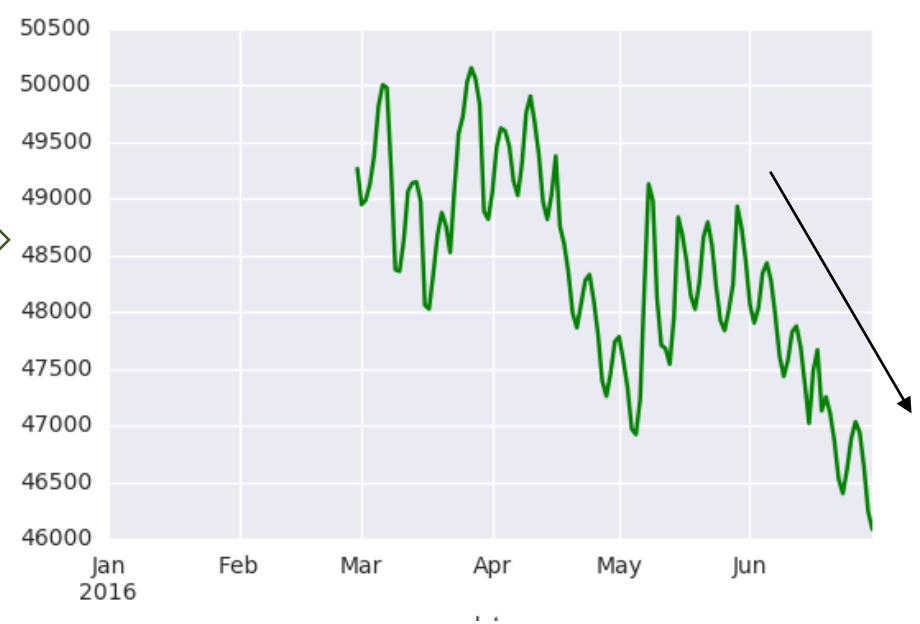


Daily Pickups Dropped Quietly in the Second Quarter of 2016



Moving Average

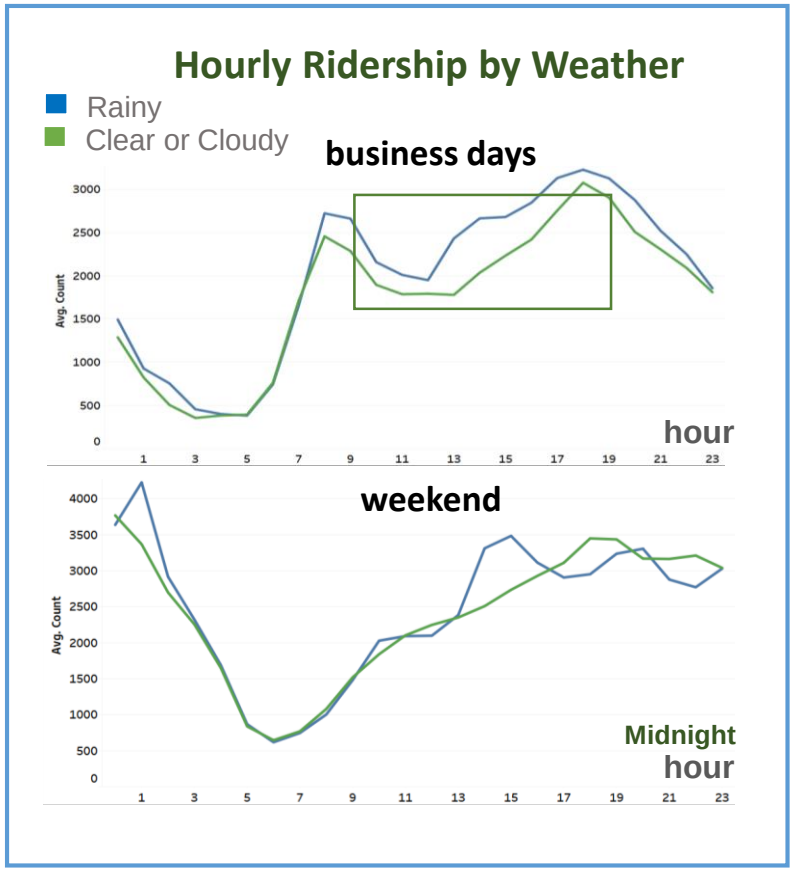


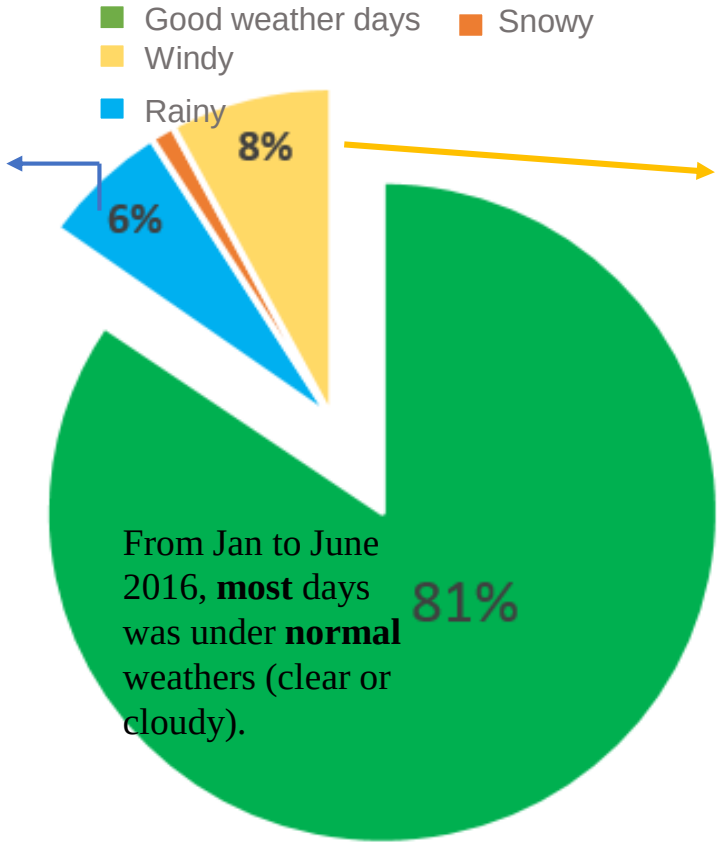


June 2016 probably marks the Start of periodical decline.

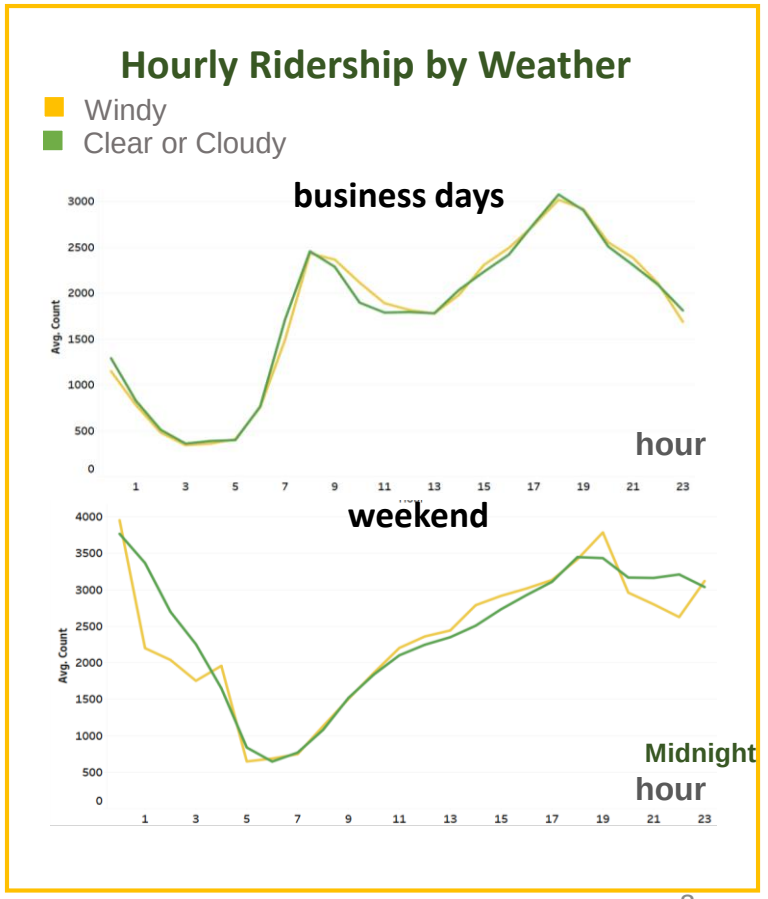
Weather Is Barely A Concern To Ridership Most of The Time

- **Daytime** ridership on **rainy** business days experience some **increase**.

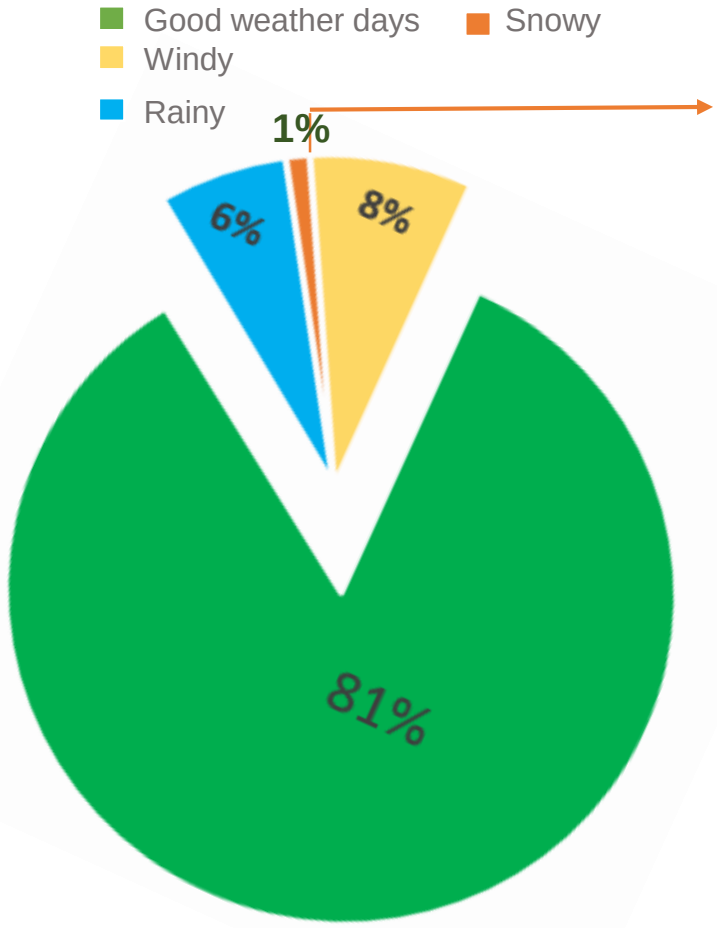


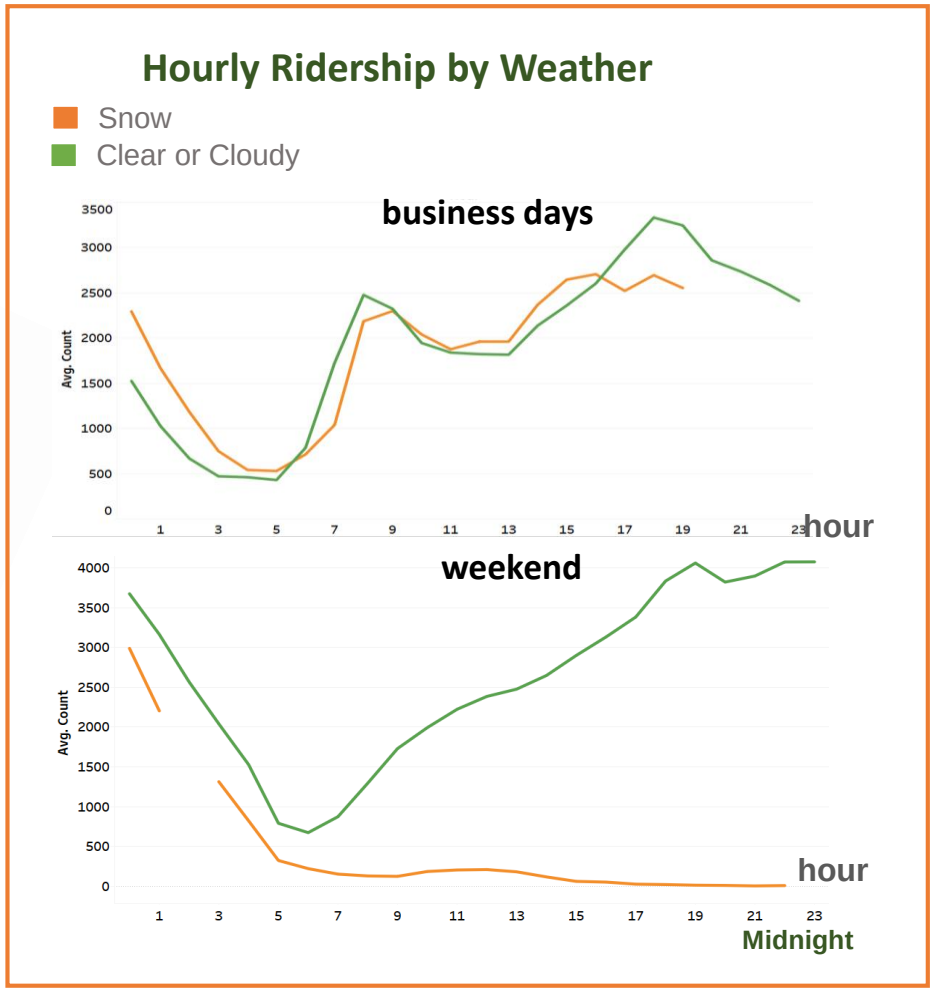


- Ridership on **windy** days **barely** change.



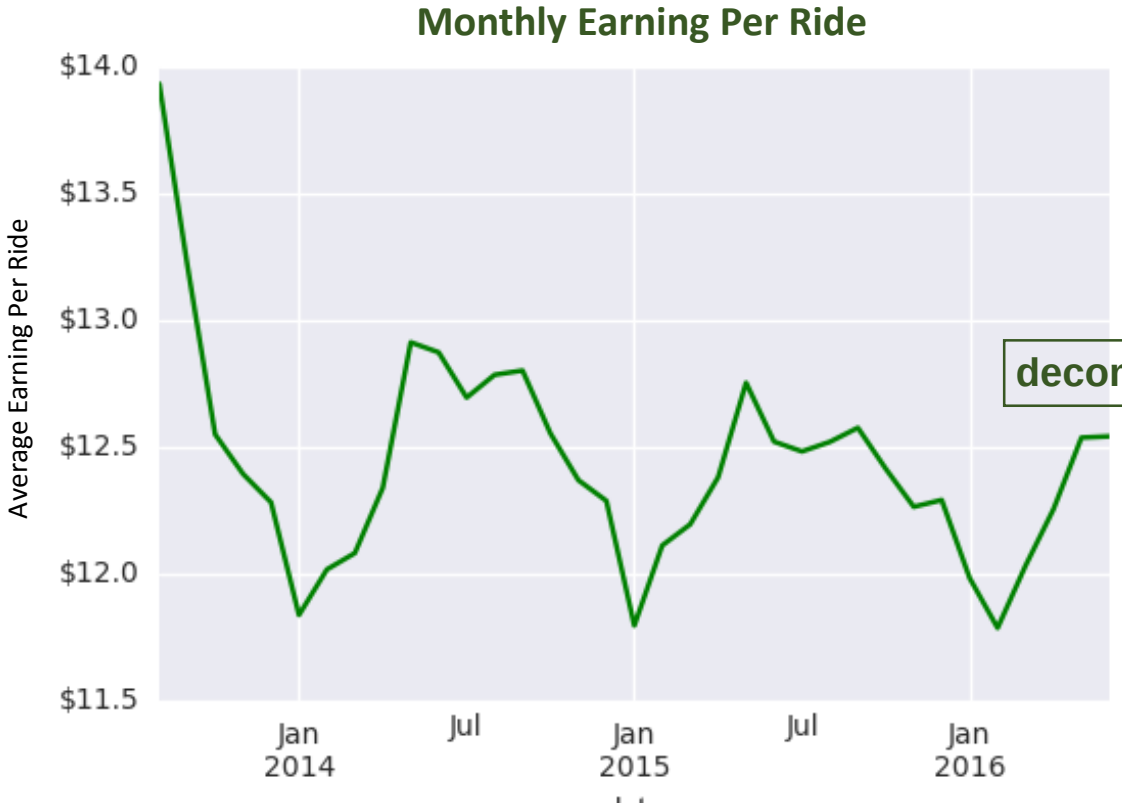
Compared To Those During Business Days, Weekend Ridership Is More Vulnerable To Bad Weathers



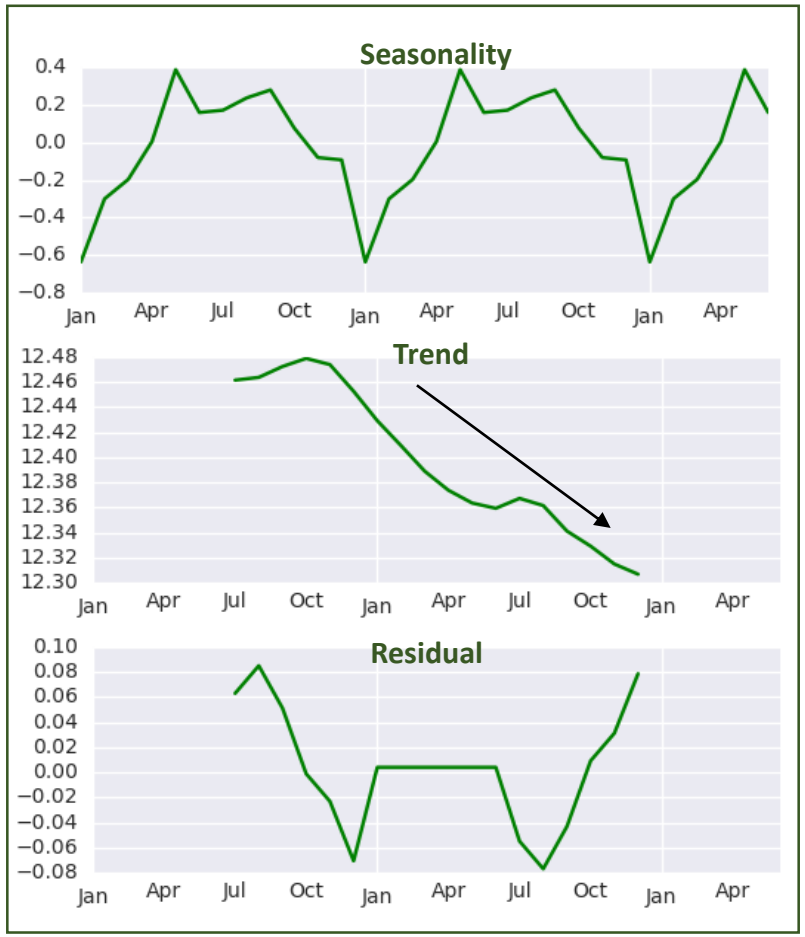


- Ridership on snowy **business** days is **slightly** depressed.
- Ridership on snowy **weekends** is **largely** restrained**,** especially those in the **evenings.**

Seasonally Higher Earning Per Ride Offset Fewer Ridership, But With Limited And Weakened Effect



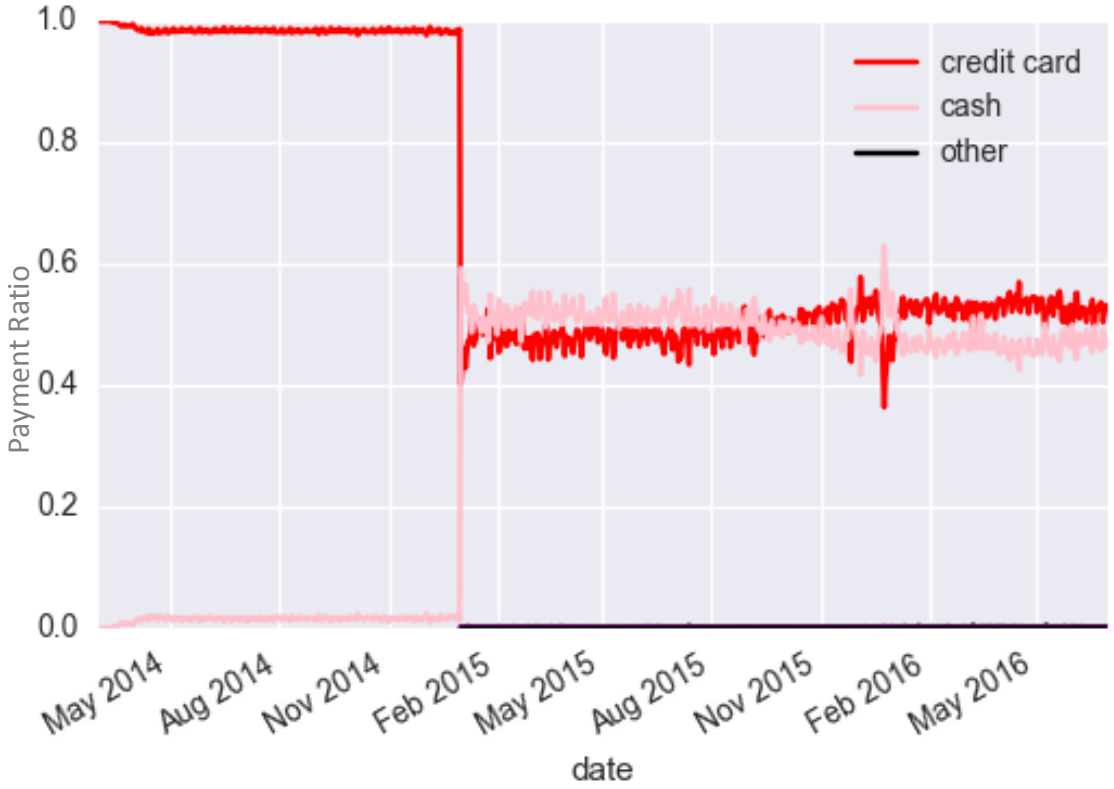
decompose

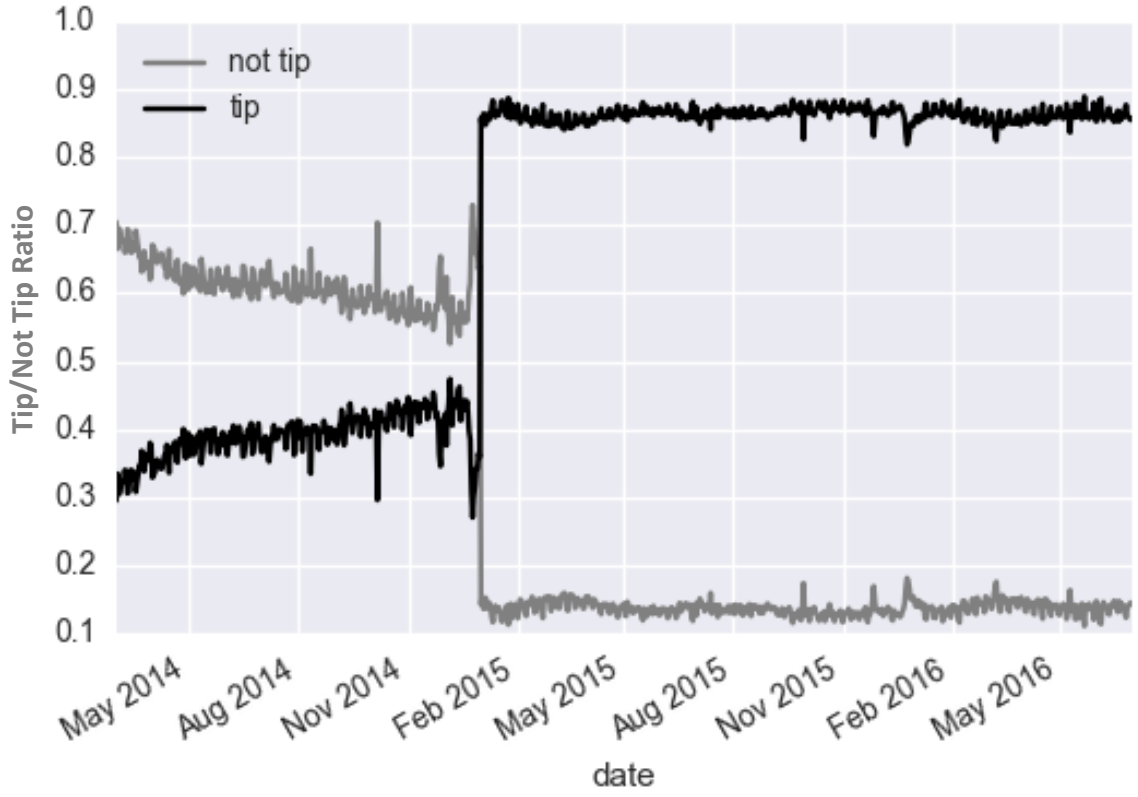


- Fare averages at $12.4 with a small deviation of 5 cents.
- Higher earnings per ride happens during **June to October**, while lows during November to May.
- When ridership was shrinking, earing per ride was on its seasonal highs.

The Reliability of Tip Data Before 2015 is Under Question

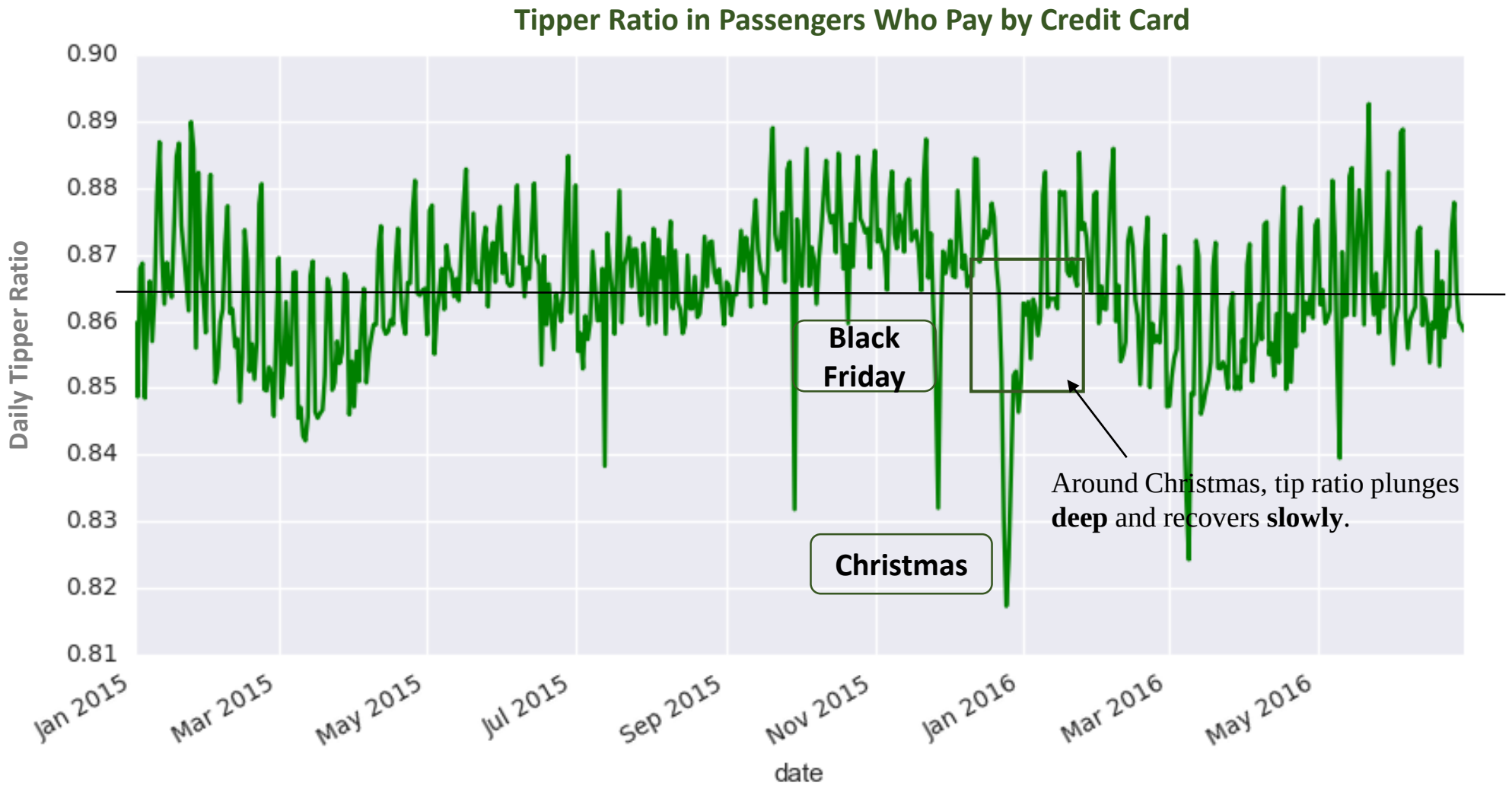
- **Before** 2015, most all passengers paid by card and more than half of them did not tip;
- **After** 2015, around half of passengers pay by card and over 80% of them tip.





Tip

Most of the time, Over 85% Of Passengers Who Pay By Card Tip

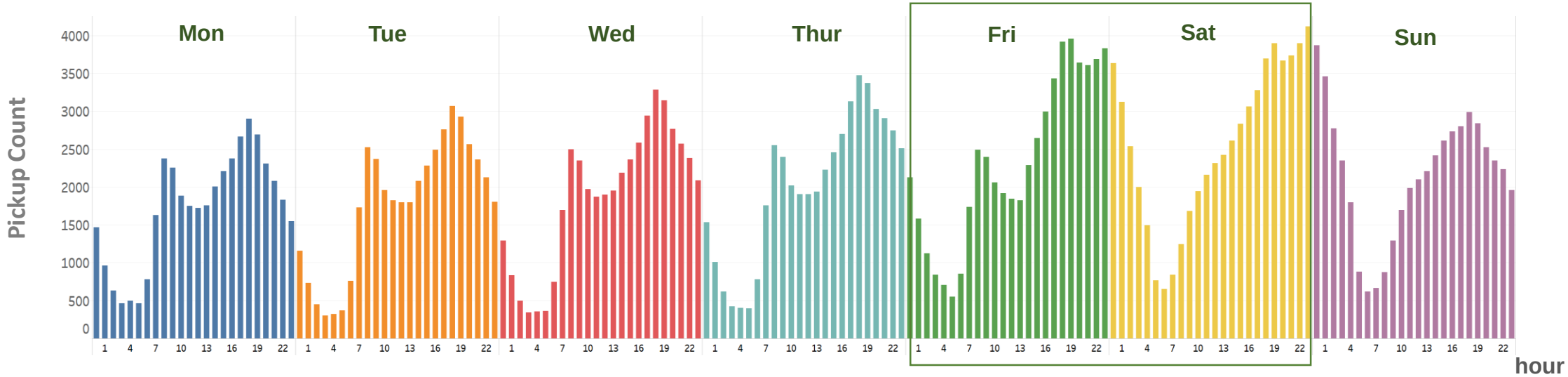


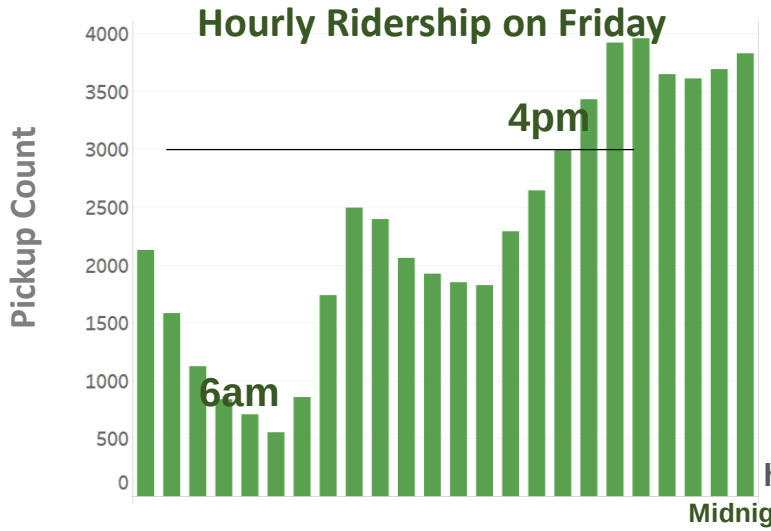
Strategy

- Time
- Place

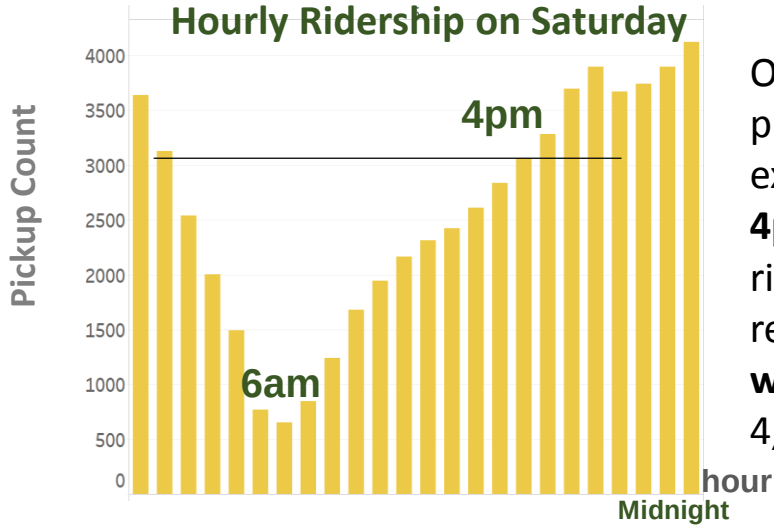
- For higher accuracy and relevancy, only post-2016 data is used
- Time frame: Jan 1, 2016 - Jun 30, 2016. (182 days, 6 months and 26 weeks.)

4 PM - Midnight Every Friday and Saturday, Demands Soar





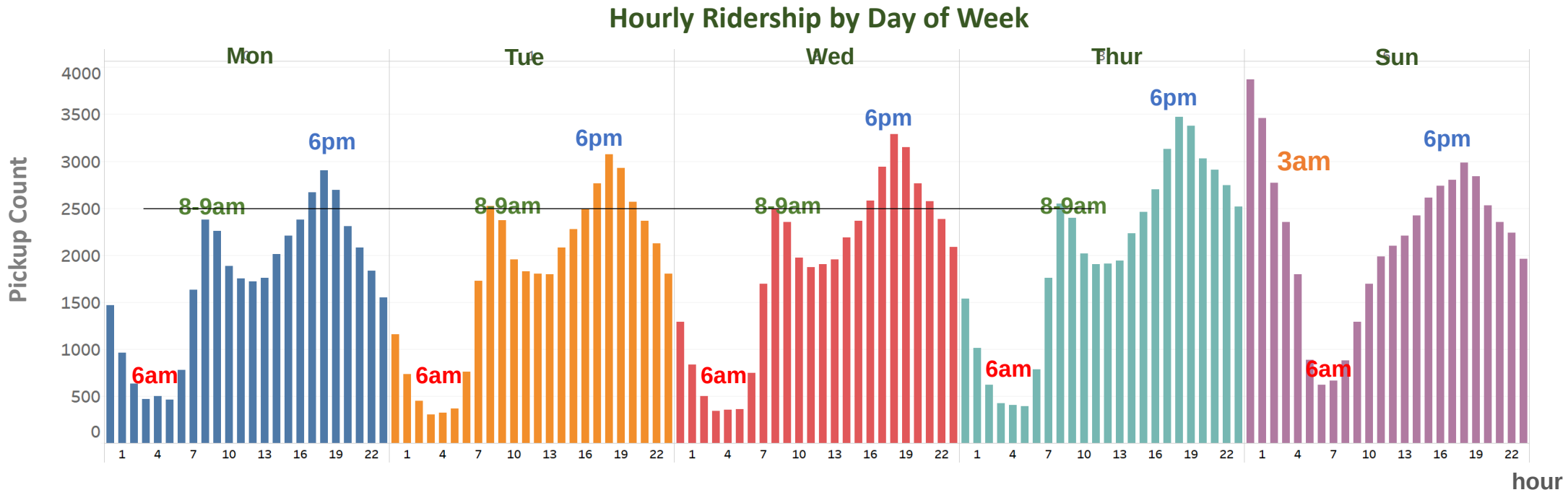
On Friday, Hourly pickup volume exceeds 3,000 from **4pm** and keeps above 3,500 till **midnight**.



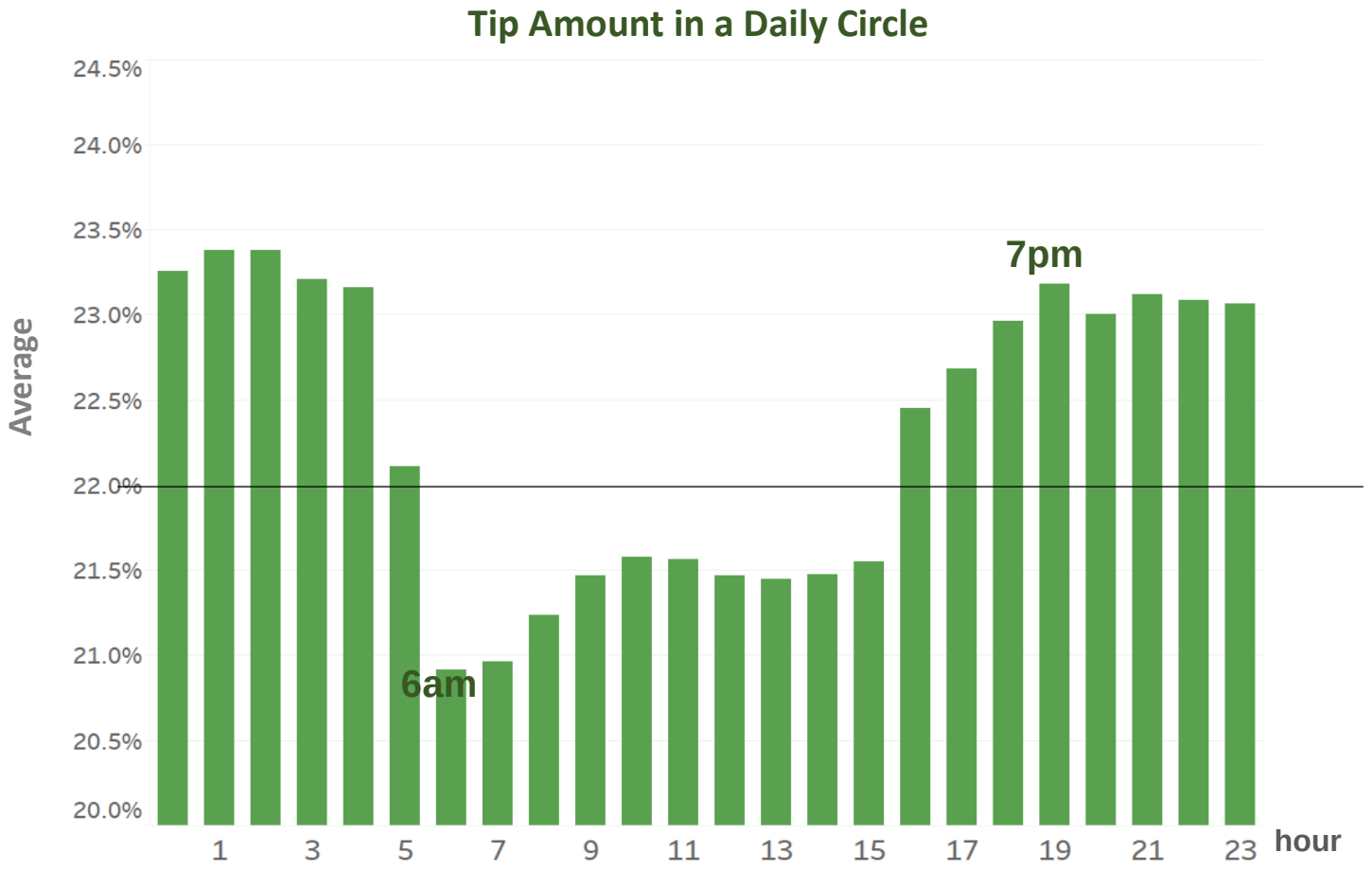
On Saturday, hourly pickup volume exceeds 3,000 from **4pm** and keeps rising till **midnight**, reaching the **weekly peak**, over 4,000 rides.

6 PM Everyday, 0-3 AM Sunday And 8 AM-9 AM Monday through Friday, Ridership Peaks

- Overall, rides at **night** outnumber those during daytime; pickups in the **afternoon** exceed those in the morning.
- Morning rush happens during **8am to 9 am** from Monday to Friday.
- Daily peak happens at **6pm** during business days.
- Daily peak happens around **Midnight** on Friday, Saturday and Sunday.

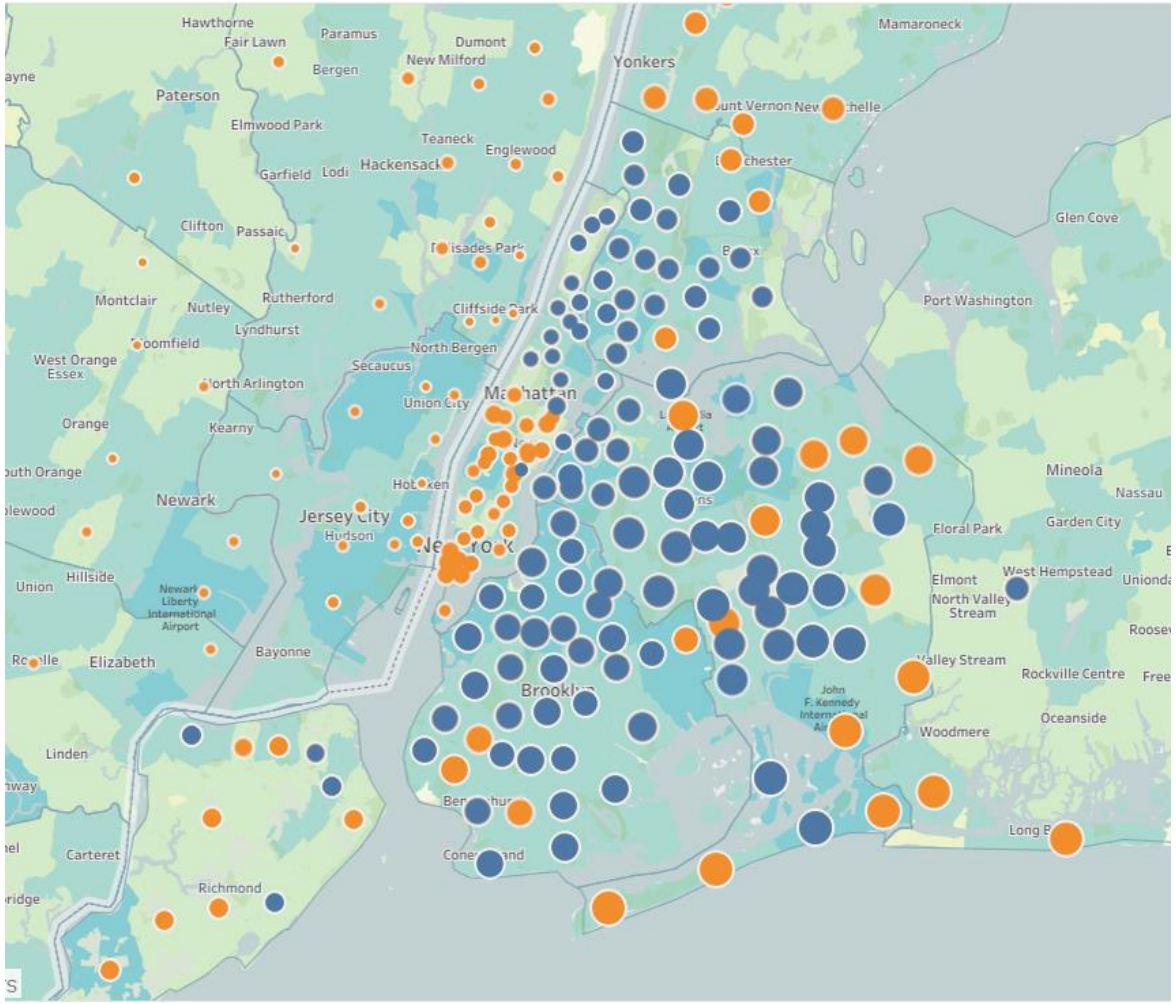


0-4 AM, 6 PM-Midnight Everyday, Tip Amount Stays High



Tip amount averages at around 22% with a deviation of 1.3%.

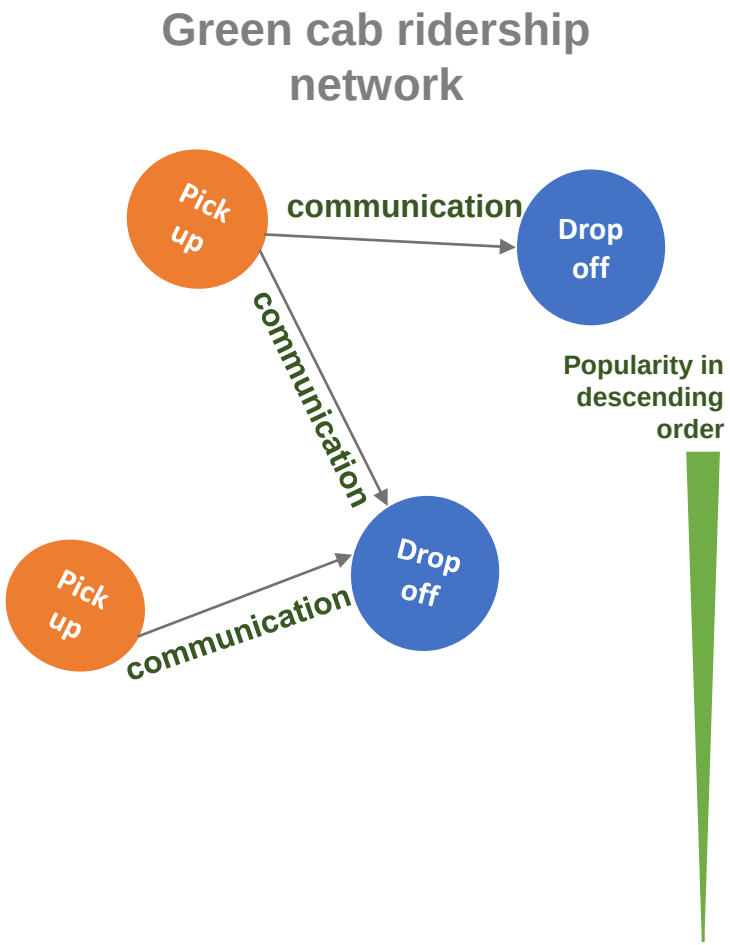
In Densely Populated Areas, Ridership Soar



- **Blue** dots represent **pickup** locations while **orange** dots indicate **drop-off** places.
- Rides **frequency** is captured by the **size** of dots: larger dots indicate higher ridership.
- Area **population density** is captured by the color of layer: darker layer indicates higher population density.

- Places with **higher population density** have **more business opportunity.**
- Compared to pickup places, **drop off** places, except those in Manhattan, are **sparsely distributed** and sometimes **far-away**, for example, Yorkers in upper state , Rosedale in Queens and Newark in Jersey.
- Picking up passengers at Queens or Brooklyn and dropping them off in Manhattan, typically lower Manhattan, is one typical green cab business routine.

Near Columbia University, East Harlem and Long Island City, Business Opportunity Blooms



network properties:
In-degree is a count of the number of edges directed to the node.
Out-degree is the number of ties that the node directs to others.
Degree is the sum of in-degree and out-degree

⇩

popular drop-offs:

11369: East Elmhurst
10016: Madison Ave/Empire State Building
11102/11103/11105/11106: Astoria
10463:Riverdale, Bronx
10001: Chelsea
10035/10029: East Harlem
10010: Baruch College
10019: Columbus Circle
10003: East Village

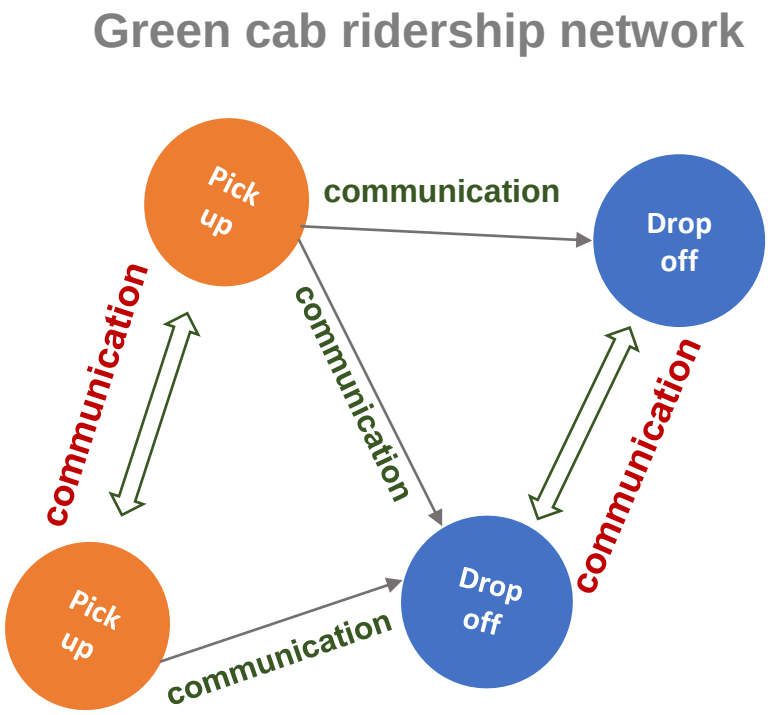
popular pick-ups:

10027: **Columbia University**
10035/11029: **East Harlem**
11101/10026: **Long Island City**
10032: Washington Heights
11222/ 11105/11103 :Astoria
11231: IKEA Brooklyn
11373: **Elmhurst**
11249: East Flatbush, Brooklyn
11375: Forest Hills

popular places:

11369/11373:East Elmhurst
10027: **Columbia University**
10035/10029: **East Harlem**
11102/11103/11105: Astoria
11101: **Long Island City**
10032: Washington Heights
11231: Red hook, Brooklyn
10463: Riverdale, Bronx
11222:Little Poland, Brooklyn

Near Columbia University, East Harlem, Astoria and Washington Heights, Sustainable Business Opportunity Brews



betweenness centrality:
quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.
quantifies the control of **resource** on the communication or interaction.

Higher betweenness centrality indicates that more **efficient** communication happens through the node.

Under the context of cab business, higher betweenness centrality indicates **faster loop of pickup and drop-off and pickup** happens at a place.

sustainable places:
10027: **Columbia University**
10035/10029:: **East Harlem**
11102/11105/11103 : **Astoria**
10032: **Washington Heights**
11377: Woodside
11373: Elmhurst

Efficiency in descending order