Green Cab Data Challenge

Miya

12/27/2016

EXECUTIVE SUMMARY

Help Green taxi driver in NYC understand key trends and optimize their revenue potential through a journey of exploratory data analysis on four years' historical data about green cab ridership.

OBJECTIVE

- Augment passenger VOLUMNE☺
- Optimize financial EFFICIENCY☺
- Increase TIP☺

ROADMAP

- Build data collection pipeline
- Load the data
- Prepare the data:
  - ✓ Data Integration and validation
  - ✓ Data Cleaning and transformation
- Conduct exploratory analysis:
  - ✓ Statistical analysis
  - ✓ Social network analysis
  - ✓ Classification modeling
- Develop insights/recommendations

DATA COLLECTION *(Python)*

- Web crawling pipeline including:
  - ✓Fetch data about ridership, weather, federal holiday
  - ✓Add attributes (e.g. “ride time” from existing “pick up time & drop off time”, “year, month, day of week and hour” from existing “date”, “zip code” from existing “longitude/latitude”, etc.)
  - ✓Insert data into tables in a relational database (SQLite)
  - a raw data of 6 gigabytes

LOAD DATA (Spark)

- Data Integration (weather data with 29,999 rows and ridership data with 45,299,607 rows)
- Data Validation (data type, missing value, anomaly detection etc.)

✓ loud noise:

- Tax, fare, tolls, trip distance etc. being negative
- Tip being hundreds
- Ride time being 0
- Passenger number over 5

✓ hidden noise:

- Cab speed being either over 25mph (speed limit in NYC) or under 10 mph in the city
- Fare per minute less than $0.5 or over $30
- Tip percentage over 100%
- … …

LOAD DATA (Spark) Cont’d

If throwing away all polluted records, 8,314,112 (one fifth) would remain.

- Assume noise is randomly distributed among data.
- Deal with noise only when the noise would directly affect analysis, e.g. negative fare values would affect density analysis of efficiency(fare/ride time), but not trend analysis of ride count. Thus, when we count rides, records containing negative fares wouldn’t be dropped.

METHODOLODY

- Aggregate or filter or subset data with Spark
- Visualize and conduct analysis with Python/Pivot table/Tableau/R
- Develop insights

ANALYSIS

I. VOLUME
II. EFFICIENCY
III. TIP

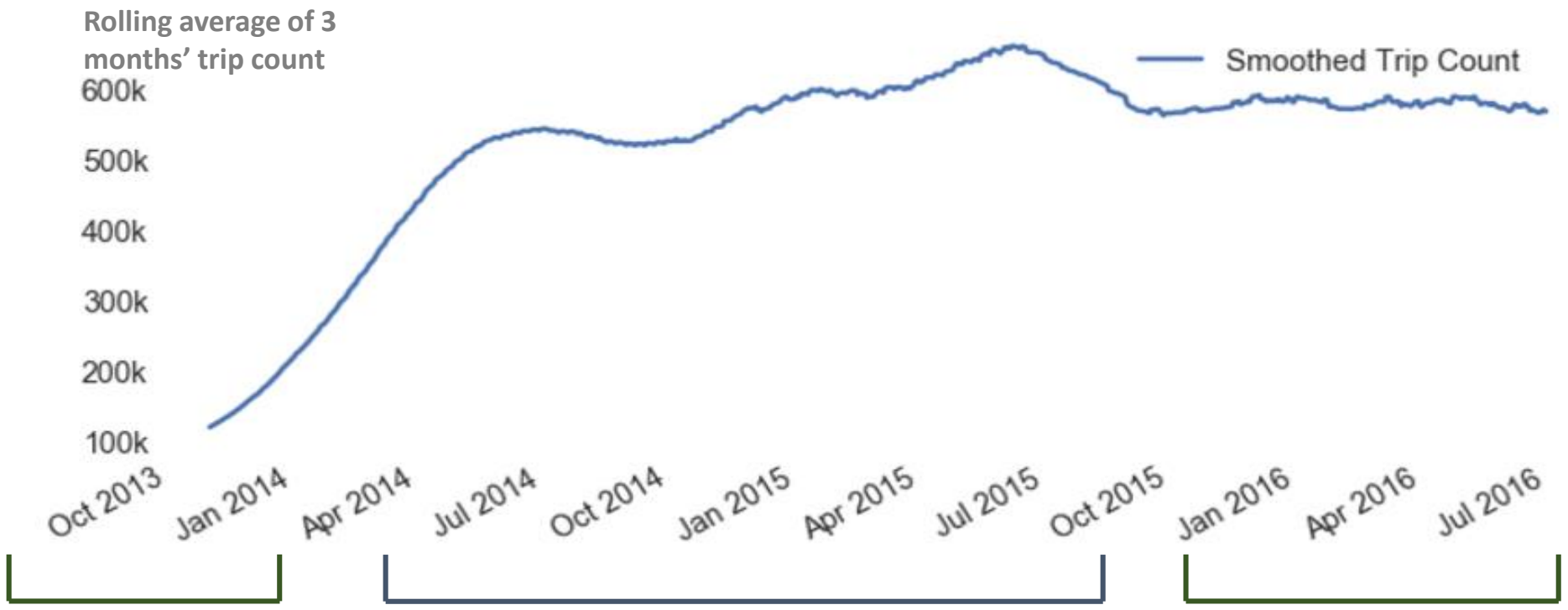
I. Volume

Overview

- Trend analysis
- Segmentation:
  - ✓Day of week
  - ✓Day of week + hour
  - ✓Weather + hour
  - ✓Holiday + hour
- Social network analysis
- ***Raw data used***

Trend

Green cab market has matured



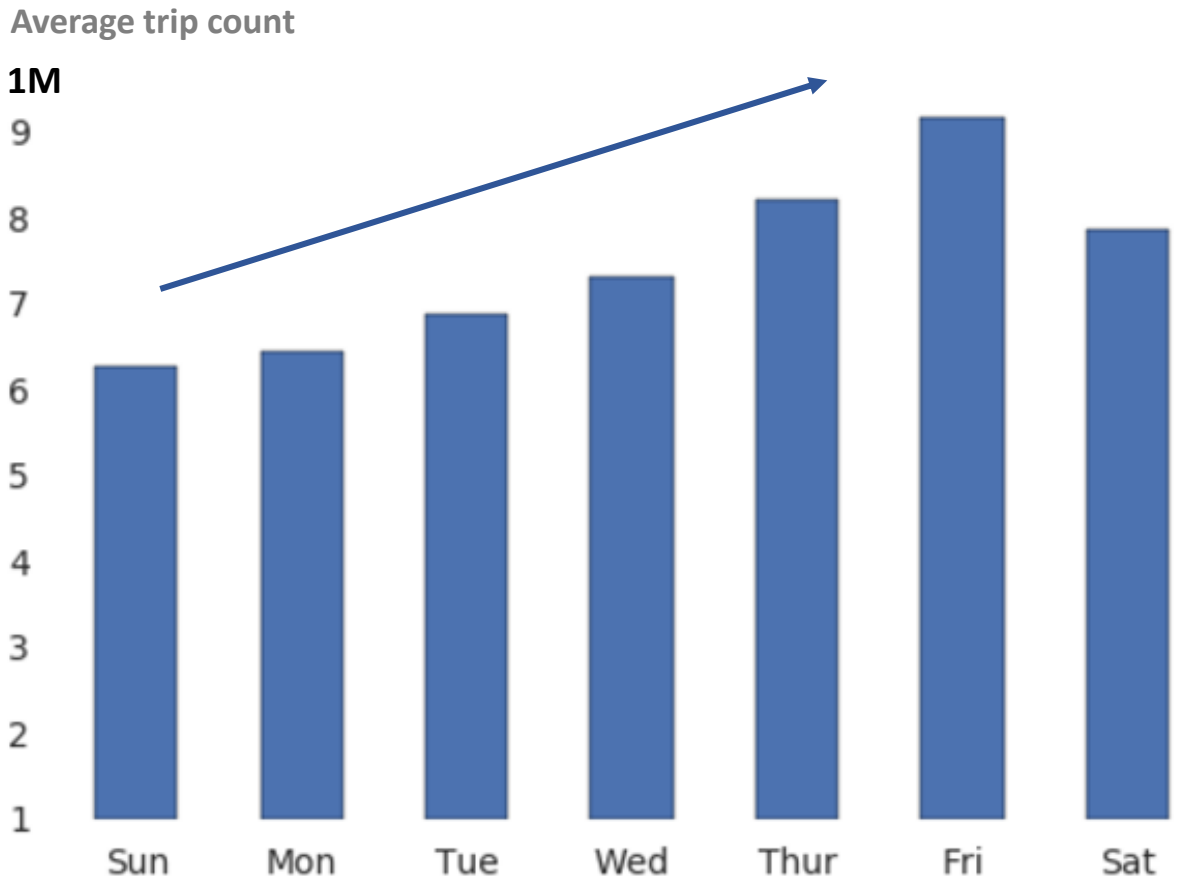
Stage 1: Innovation **Stage 2: Growth** **Stage 3: Maturity**

DATA: TAXI AND LIMOUSINE COMMISSION

Product life-cycle theory by Raymond Vernon

Segmentation —— Day of week

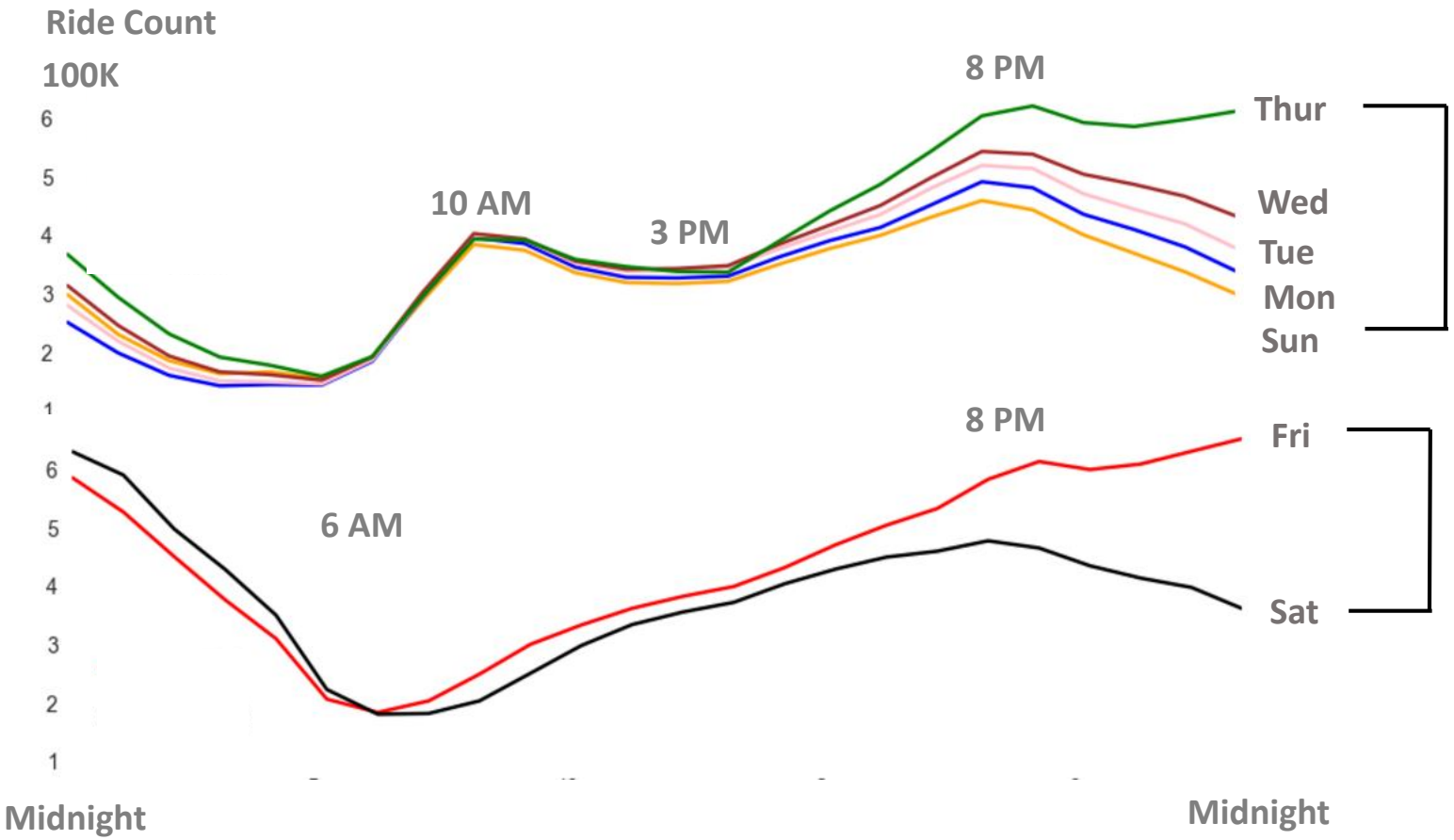
Ridership grows from Sunday and peaks on Friday



DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— Day of week & Hour

Hourly ridership follows two patterns: one starts in Thursday afternoon and another Saturday evening.



CYCLICAL pattern for traffic:

- 3 pm Thursday till 8 pm Saturday
  - peak at Midnight
  - slow growth since 6 am
- 8 pm Saturday to 3 pm next Thursday
  - peak at 8 pm
  - rapid increase from 6 am

6 am to 3 pm from Sunday to Thursday follows a CONSTANT pattern.

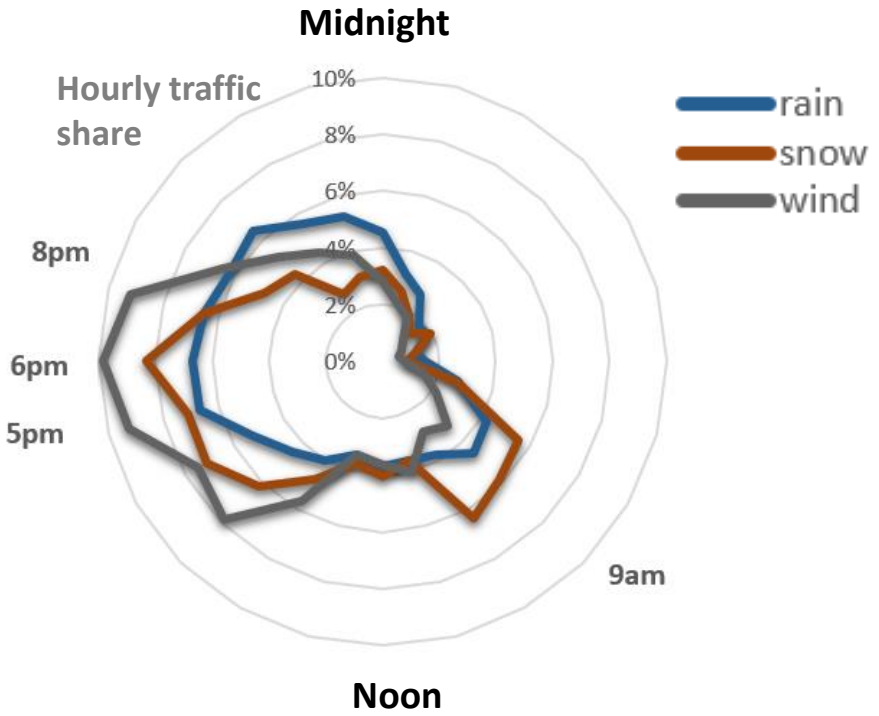
Weekly PEAK: 8 pm till midnight Thursday and Friday.

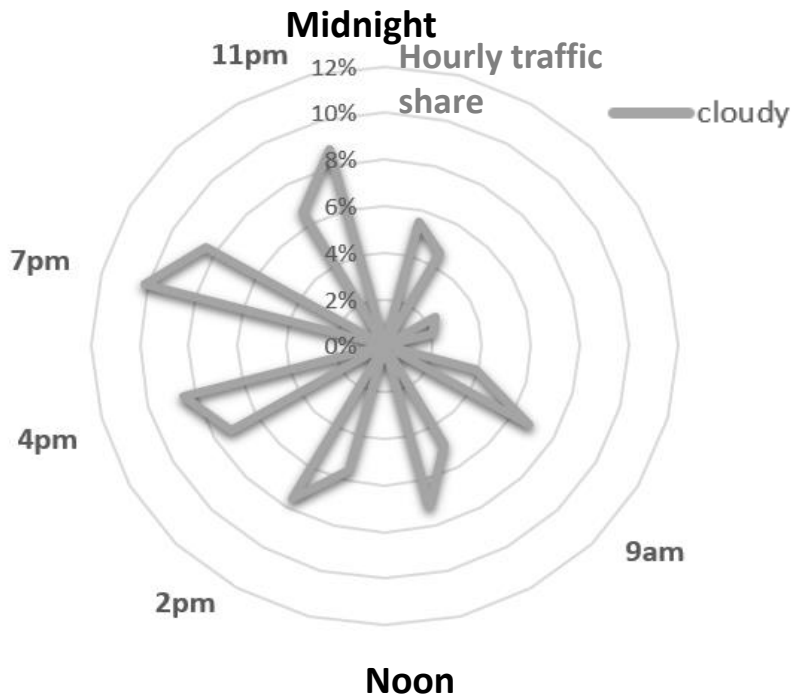
Mini Rush-hour: 7 am till 11 am Sunday to Thursday.

DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— Weather & Hour

Ridership dispersed on rainy/snowy/windy day and centralized on cloudy day





Rainy/Snowy/Windy:

- Traffic evenly-spread during daytime and sparse in the evening.
- Mini rush period usually around 6 pm
- Snow and wind greatly depress evening traffic

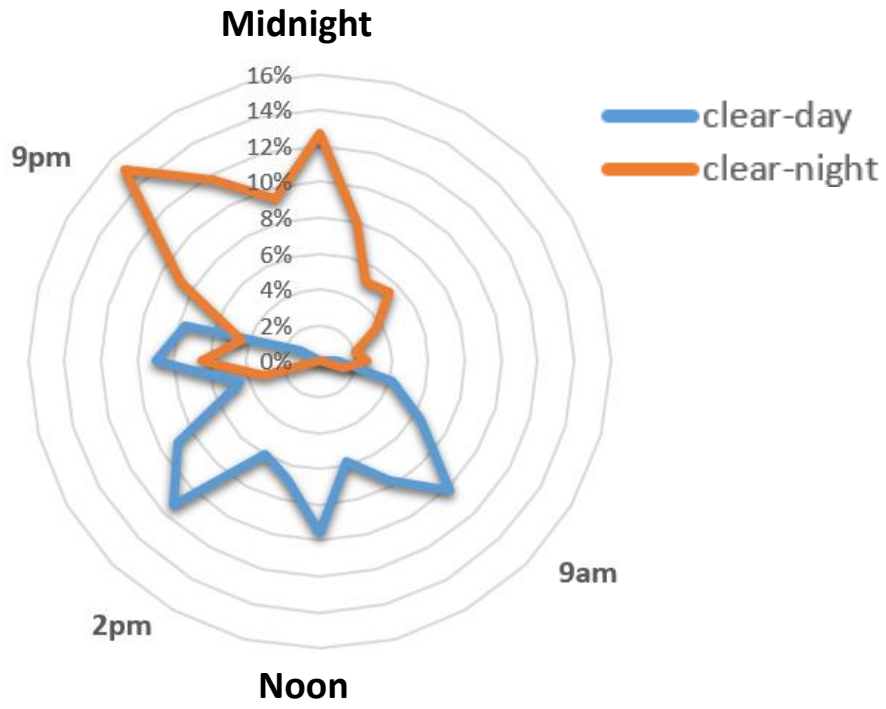
Cloudy:

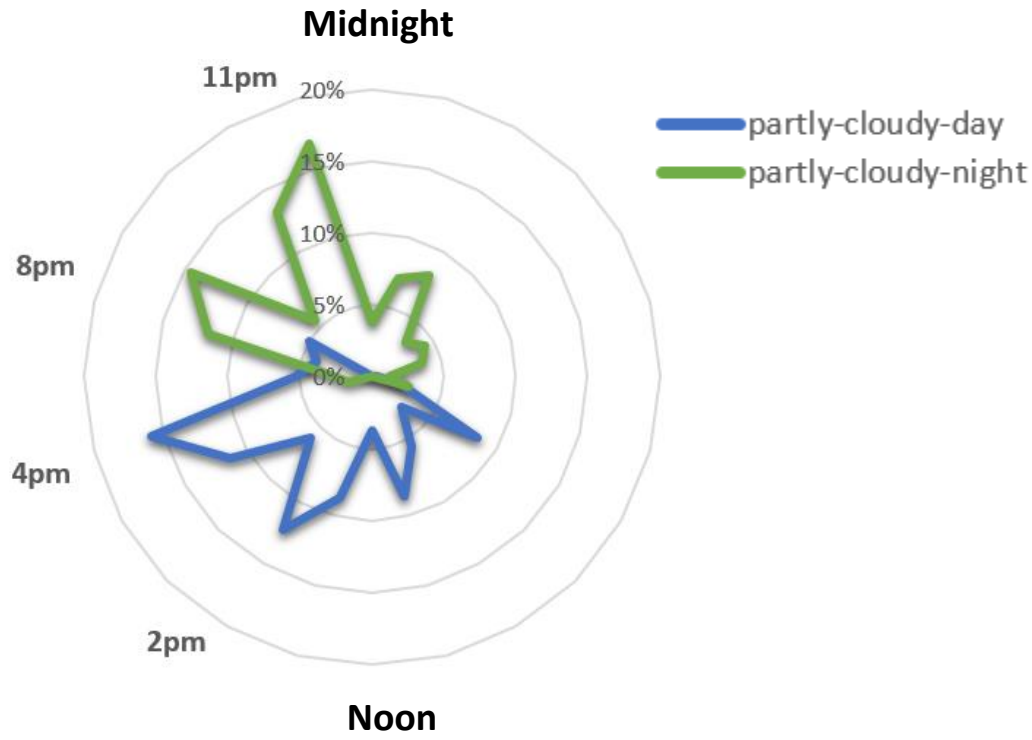
- Traffic concentrated at 9 am, noon, 2 pm, 4pm, 7pm and 11 pm.
- Peak around 7pm.

DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— Weather & Hour Cont'd

Good weather (clear and cloudy) stimulates ridership in a centralized way





- Good weather evening drive ridership till midnight.
- Traffic concentrated at 9 pm/noon/2 pm for clear day, and 9 pm/midnight for clear night.

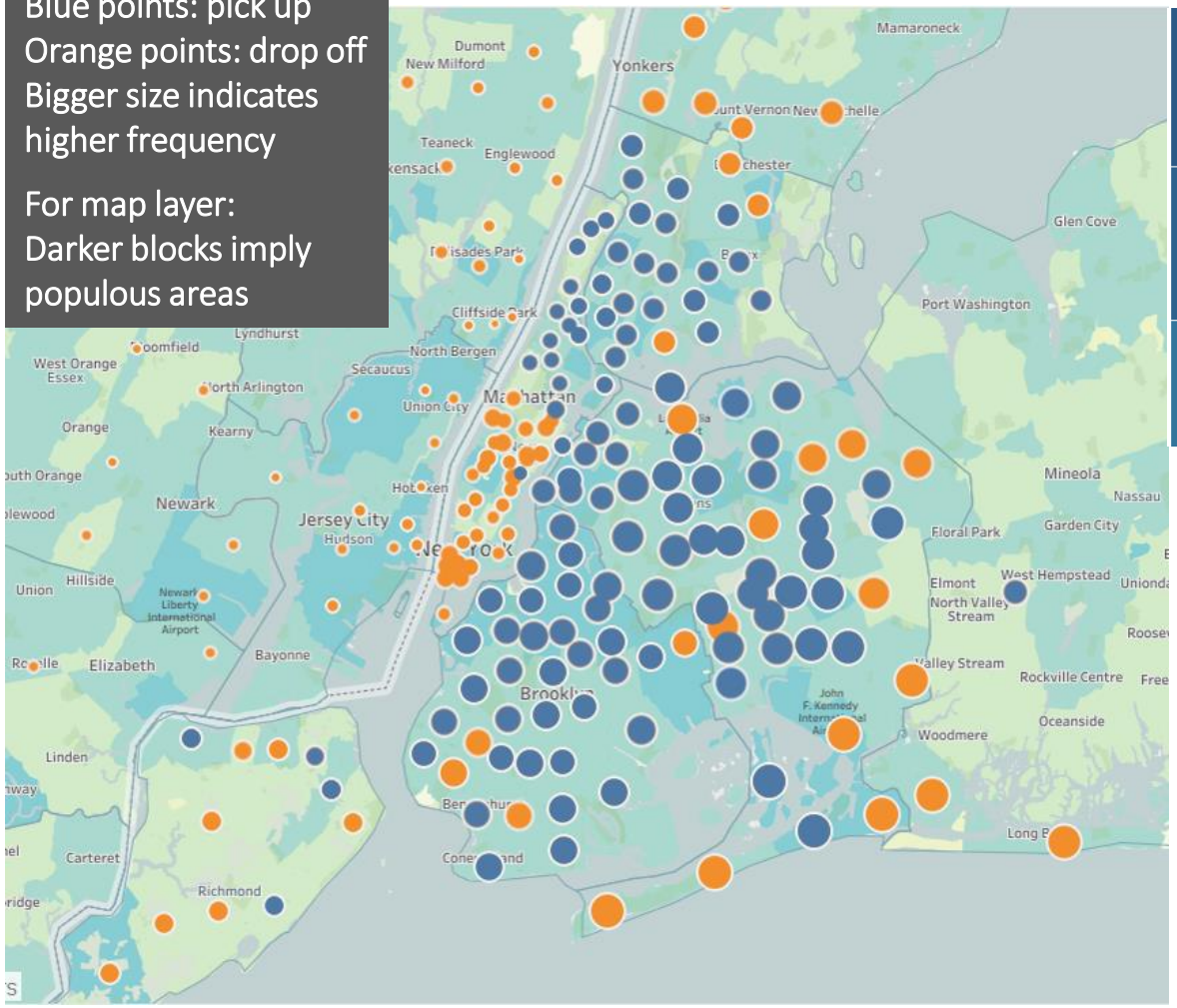
- Traffic concentrated at 4 pm and 11 pm for partly cloudy day and partly cloudy night respectively.

DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— ZIP Code

Blue points: pick up
Orange points: drop off
Bigger size indicates higher frequency

For map layer:
Darker blocks imply populous areas



10027	11101	11105	11369	11103	10031
		11373			10029
	10032			11105	
10035		10026			10017
	11222		10016	10001	10003
		11215			
				11106	10025
10029	11102	11368	11102		
				10035	11010
	11231	11249	10463	10027	10010

popular pickup (Blue) and drop off (Orange) places (zip code)

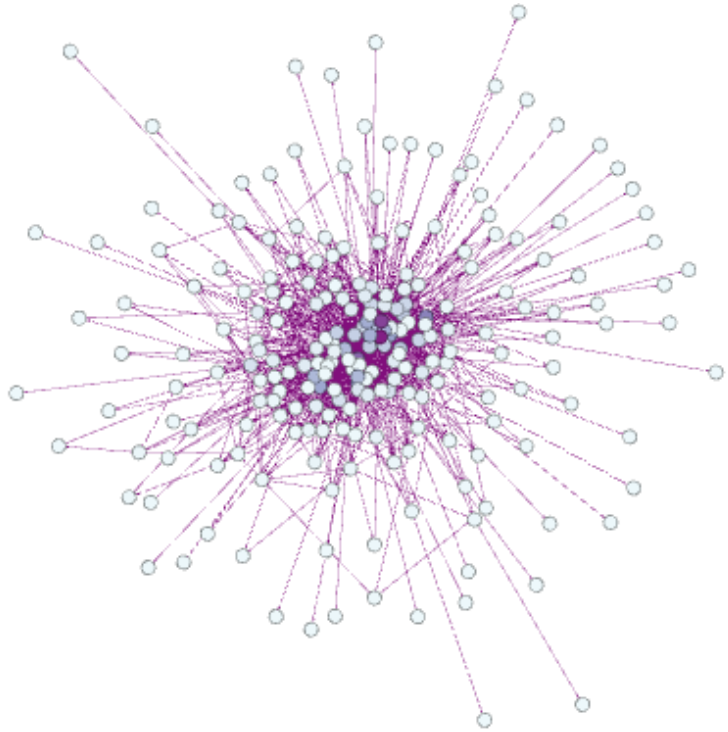
As expected, densely populated areas, such as Williamsburg, Long Island city and Staten Island, witness larger ridership.

Picking up passengers at Queens or Brooklyn and dropping them off in lower Manhattan is one typical green cab business routine.

Green cab ride destinations, except lower Manhattan, are sparsely distributed in far-away upper state, deep east-south Brooklyn and some Jersey areas: Yorkers, Rosedale and Newark.

DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— ZIP Code



Green cab ridership network from February 1 - 14 2016

Darker colored nodes and edges indicate higher degree (ridership)

- "Popular places of departure":
  - nodes with high out-degree (10027, 10035, 10029, 11101, 10032, 11222, 11102, 11231, 11105, 11373)
- "Unpopular places of departure":
  - nodes with low out-degree (11357, 11421, 11692,10471,11365,11366,10303,11427,10306)
- "Popular destination":
  - nodes with high in-degree (11369, 10016, 11102, 10463, 11103, 11105, 10001, 11106,10035, 10027)
- "Unpopular destination":
  - nodes with low in-degree (7057,11362,10155, 10466, 11228, 11561, 7028, 7043, 7650)
- " Traffic hub" :
  - nodes with high degree (11369, 10027, 10035, 10029, 11102, 11105, 11101, 10016, 11103, 10032)
- "Transportation Junction":
  - Nodes with high centrality (10027, 10035, 11102, 10029, 11105, 10025, 10032, 10031, 11103)

Recommendation——Volume

- Get prepared for high volume of ridership at the following time points:
  - 8 PM Everyday
  - Thursday and Friday night
  - 10 AM Every Sunday through Thursday.
- Good weather indicates great but challenging business opportunities: don't miss peak hours on good weather days (9 AM & 2 PM for clear day, 9 PM & Midnight for clear night, 2 PM & 4 PM for cloudy day and 8 PM & 11 PM for cloudy night), as passengers far less likely to show up other times.

Time-wise

- Show up in places with zip codes **10027,** 11222, **11102**, 11231, **11105**, 11373, 11369, **10035, 10029**, **11101**, 10016, 11103, **10032,** especially those in bold, as they are both "popular places of departure" and "traffic hub" (tremendous potential passengers).
- Avoid places with zip codes 10025 and 10031 during peak hours as they are "Transportation Junction": fewer passengers, more traffic jams.

Place-wise

II. Efficiency

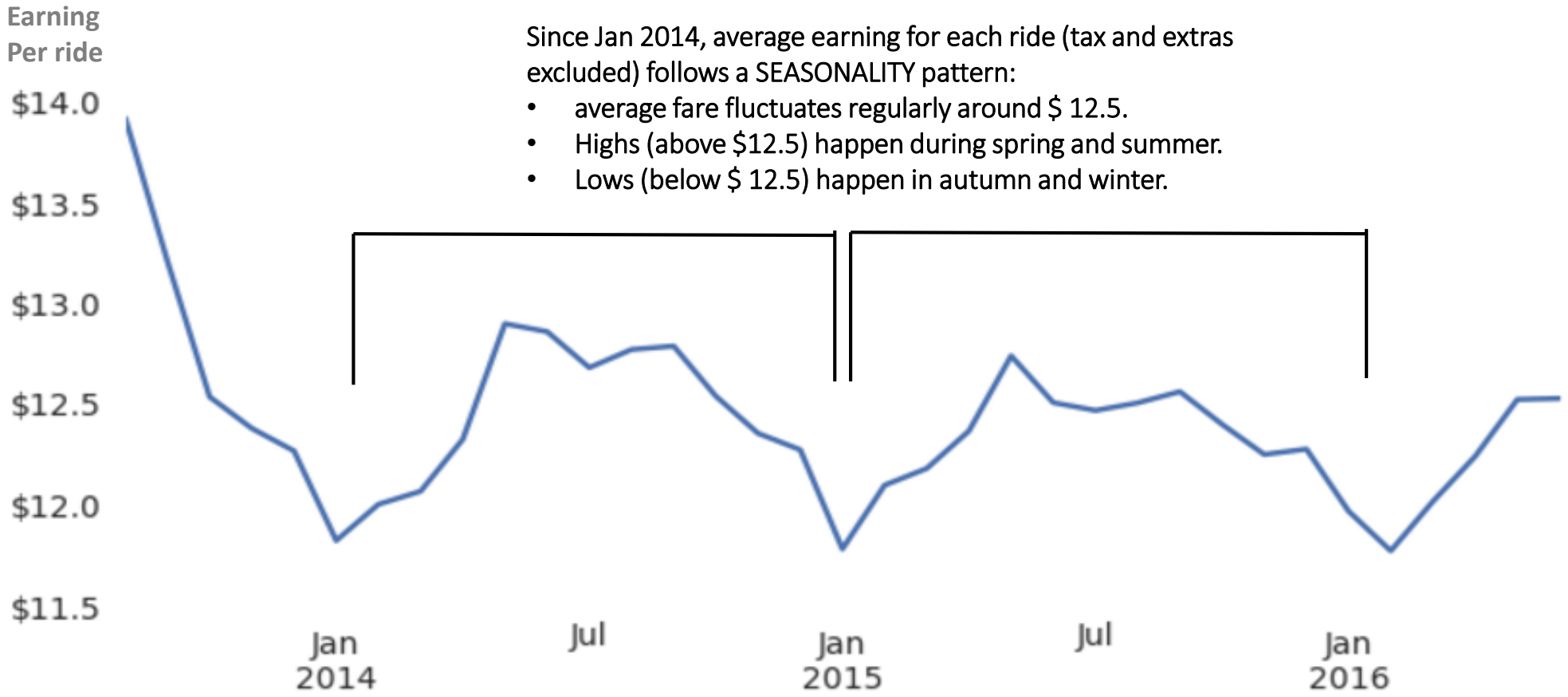
* Money earned per ride per minute

Overview

- Calculate “Efficiency” (Fare amount divided by ride time in minutes)
- Trend analysis
- Density analysis
- Efficiency Prediction (Classify a trip as “efficient” or “inefficient”)
  - ***Cleansed data used***

Trend

Green cab program is a seasonal business



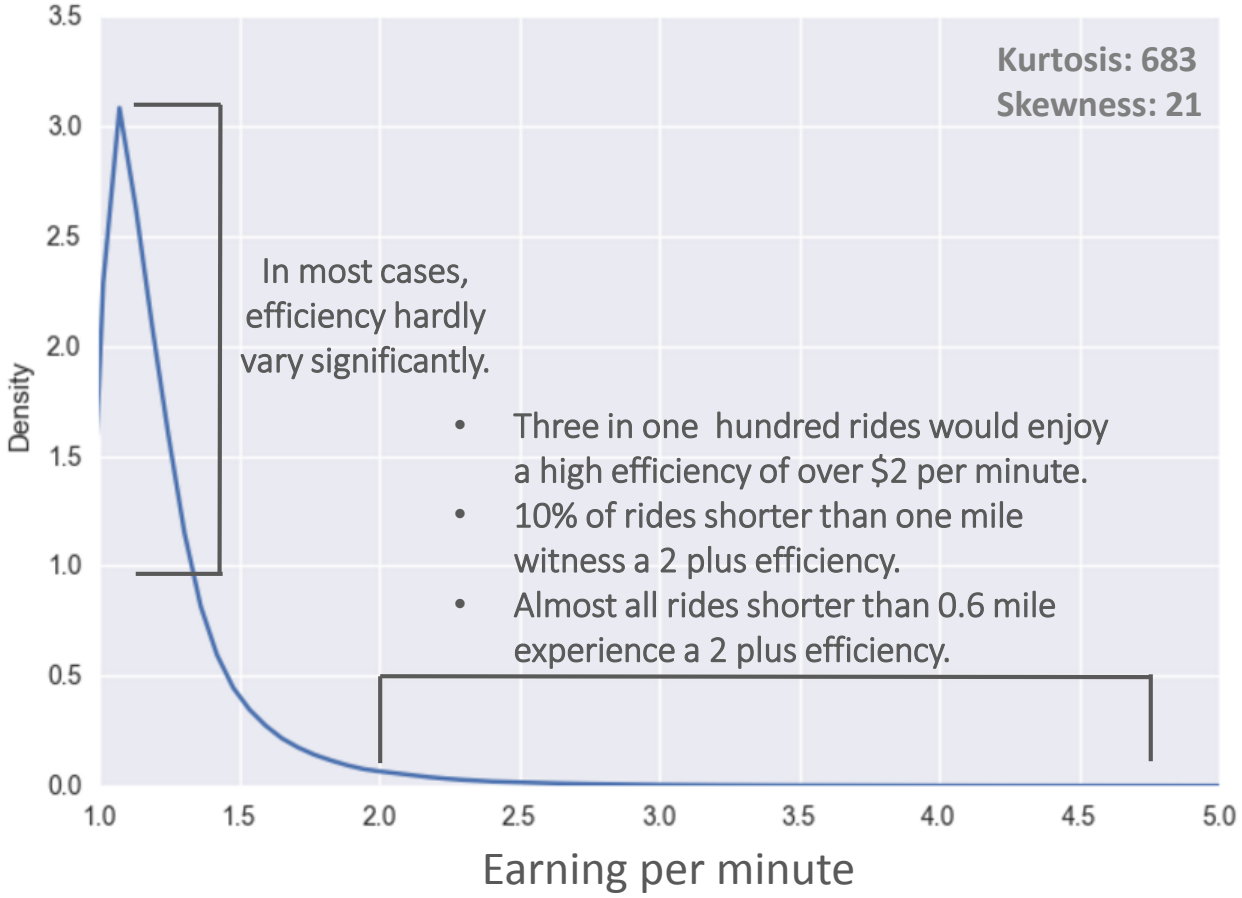
Since Jan 2014, average earning for each ride (tax and extras excluded) follows a SEASONALITY pattern:

- average fare fluctuates regularly around $ 12.5.
- Highs (above $12.5) happen during spring and summer.
- Lows (below $ 12.5) happen in autumn and winter.

DATA: TAXI AND LIMOUSINE COMMISSION

Density

Minutely earning per ride averages at $ 1.27 but could go up to tens occasionally.



Note: x axis is limited to a range of one to five for visualization

Mean	1.266382
Standard Deviation	0.610673
Minimum	1.000255
25%	1.069418
50%	1.159154
75%	1.304348

- Kurtosis being over 683, larger than the statistical normal value of 651, accompanied by a pretty small p-value, the null hypothesis of green cab efficiency following a normal distribution is rejected.
- Skewness value being 21 means there is a lot of weight in the left tail. Efficiency distribution is heavily left skewed, which means super high efficiency happens with a non-negligible probability

DATA: TAXI AND LIMOUSINE COMMISSION

Predictive Analysis—— Overview

- Randomly sample 10% from cleansed dataset
- Label generation:
  - ✓ Efficiency over the average labeled as “Good”, otherwise “Low”
- Features preparation:
  - ✓ Holiday, Passenger number, Rate Code ID, Store and fwd flag, Trip distance, Trip type, Pick up hour, month, day of week, temperature, weather, wind bearing, wind speed and visibility.
- Random Forest VS Gradient Boost
- Model validation through visualization

Predictive Analysis —— Results

- Accuracy (after cross validation):

Random Forest	Gradient Boost
66%	72%

  - Accuracy is good overall.

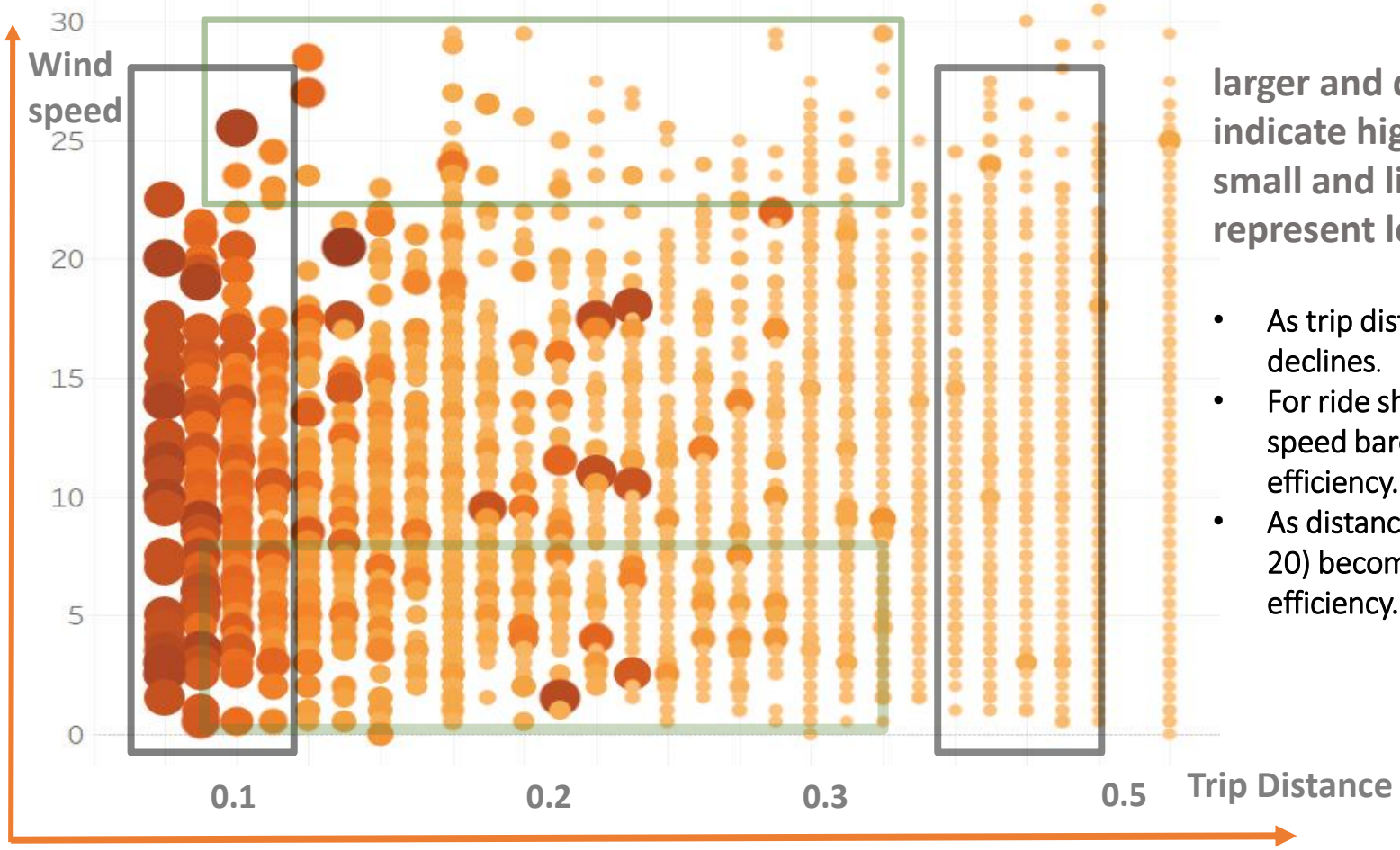
- Top three important features generated by model:

Features	Variance Explained
Trip distance	0.68
Wind Speed	0.06
Temperature	0.06

  - Model may suffer under-fitting and is unstable when trip distance is unknown. More features are needed.

Predictive Analysis —— Validation

For short rides, distance increase and faster wind speed restrains efficiency



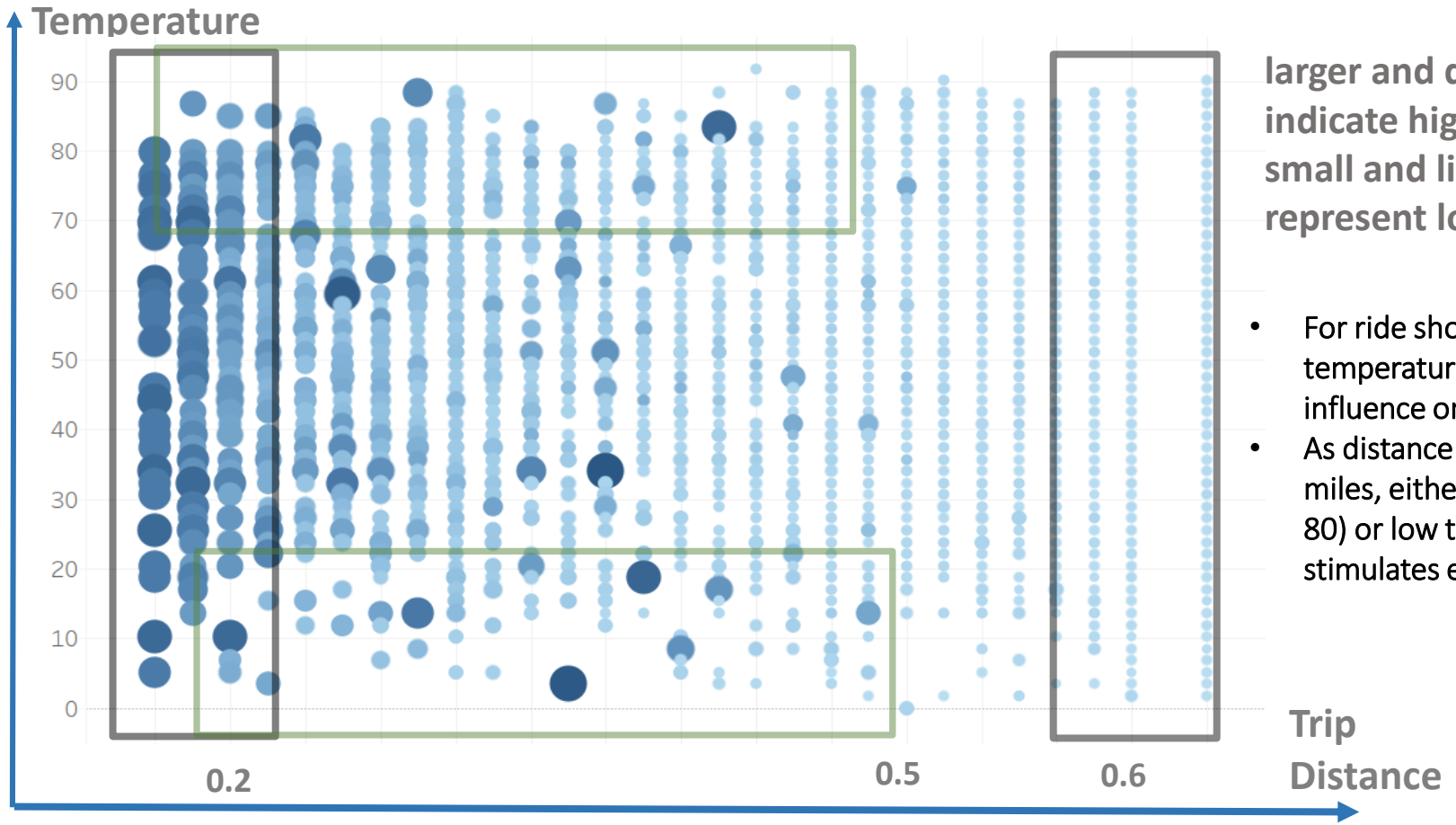
larger and darker circles indicate higher efficiency while small and lighter circles represent lower efficiency

- As trip distance increases, efficiency declines.
- For ride shorter than 0.1 miles, wind speed barely show any influence on efficiency.
- As distance grows, strong wind (over 20) becomes an adverse factor for efficiency.

DATA: TAXI AND LIMOUSINE COMMISSION

Predictive Analysis —— Validation

For short rides, polarized temperatures promote efficiency



larger and darker circles indicate higher efficiency, while small and lighter circles represent lower efficiency

- For ride shorter than 0.2 miles, temperature barely show any influence on efficiency.
- As distance grows its way up to 0.5 miles, either high temperature (over 80) or low temperature (below 20) stimulates efficiency.

DATA: TAXI AND LIMOUSINE COMMISSION

Recommendation——Efficiency

- Expect an average earning ranging from $12 to $13 for each ride.
- Expect an earning of $ 1.26 per minute for an ordinary ride.
- Don’t expect to earn more than $2 per minute per ride. It happens with a small chance (3%).
- Shorten each ride below 1 mile if a high efficiency is needed; Shorten each rides below 0.6 miles if a high efficiency is strongly needed.
- As efficiency is hard to change and rare to vary, focus on ridership (volume) for higher revenue.

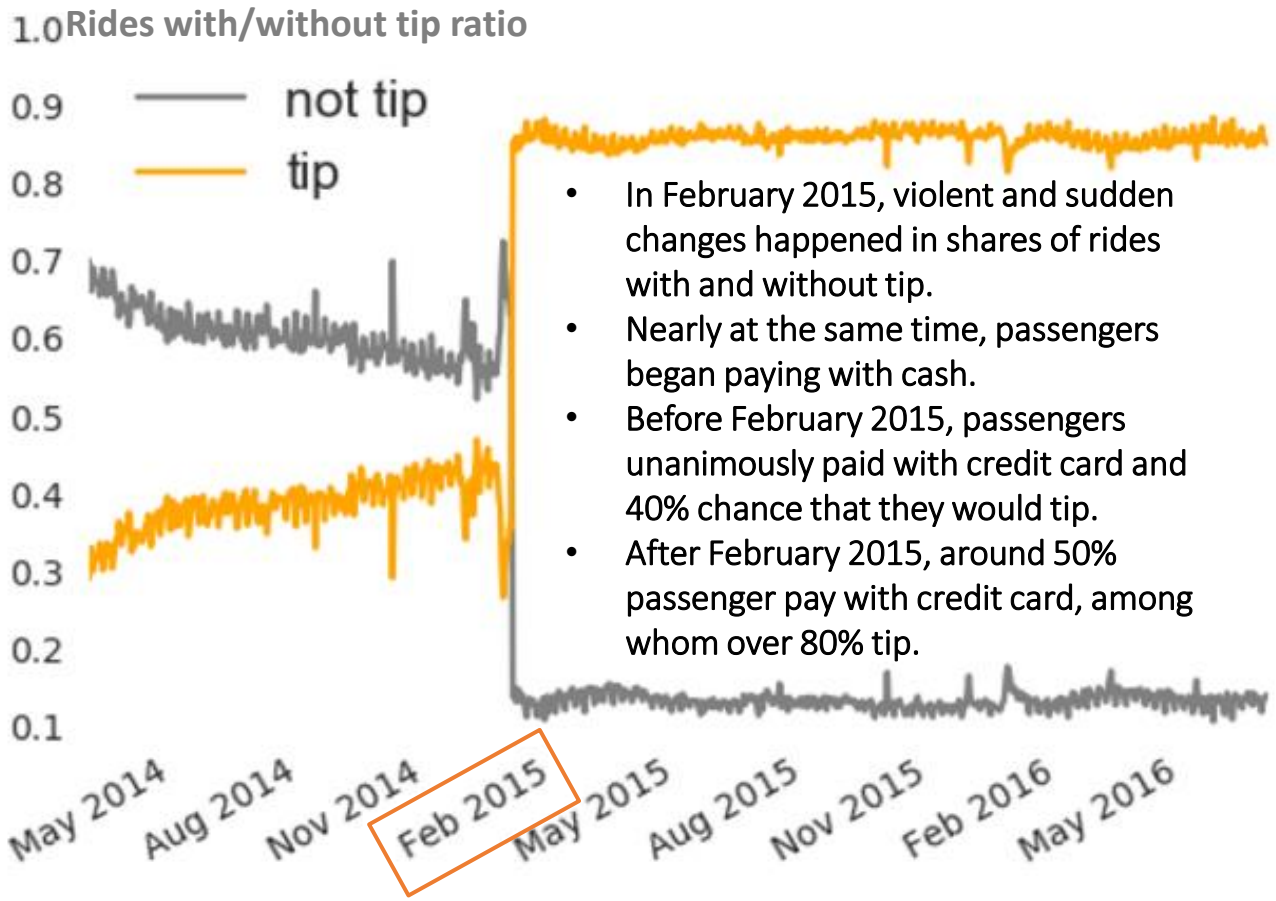
III. Tip

Overview

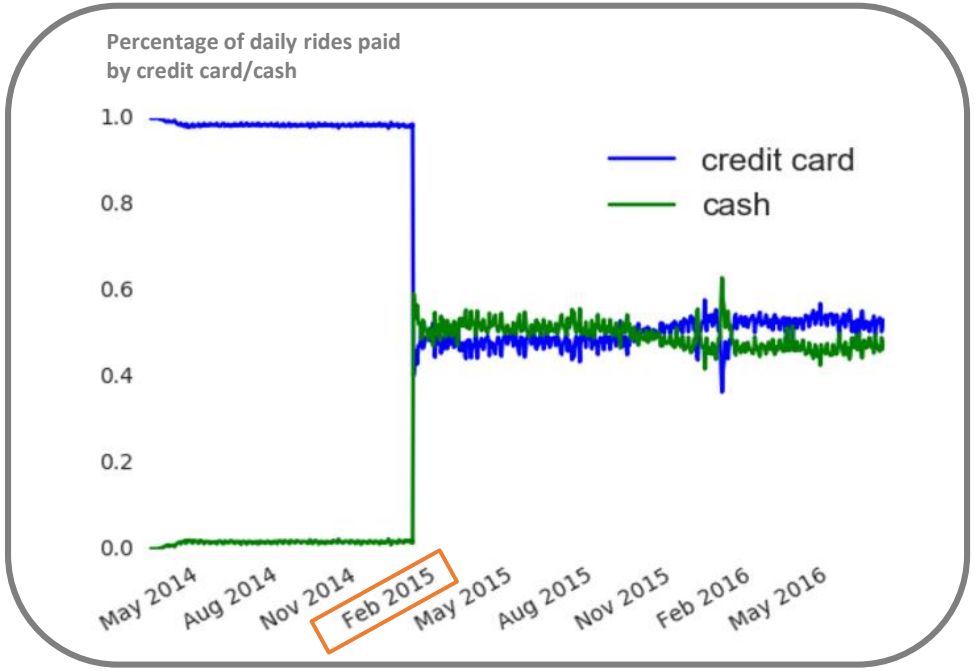
- Trend analysis
- Segmentation:
  - ✓Day of week
  - ✓Hour
  - ✓Weather
  - ✓Rate Code ID
- **Cleansed data used**

Trend

Great changes in payment type happened in February, 2015.
More and more passengers pay with card and over 80% of them tip.



- In February 2015, violent and sudden changes happened in shares of rides with and without tip.
- Nearly at the same time, passengers began paying with cash.
- Before February 2015, passengers unanimously paid with credit card and 40% chance that they would tip.
- After February 2015, around 50% passenger pay with credit card, among whom over 80% tip.

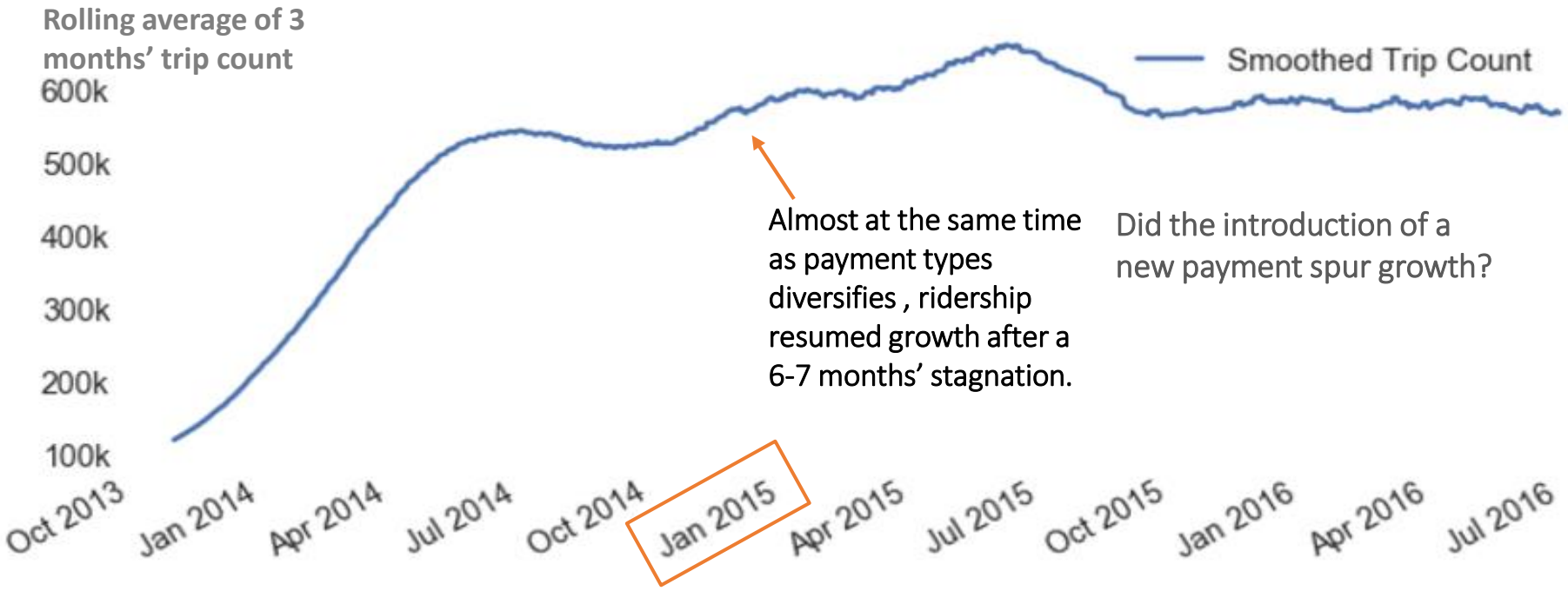


- Despite of occasional decline, credit card payment share has kept gaining popularity since Aug 2015.

DATA: TAXI AND LIMOUSINE COMMISSION

Trend

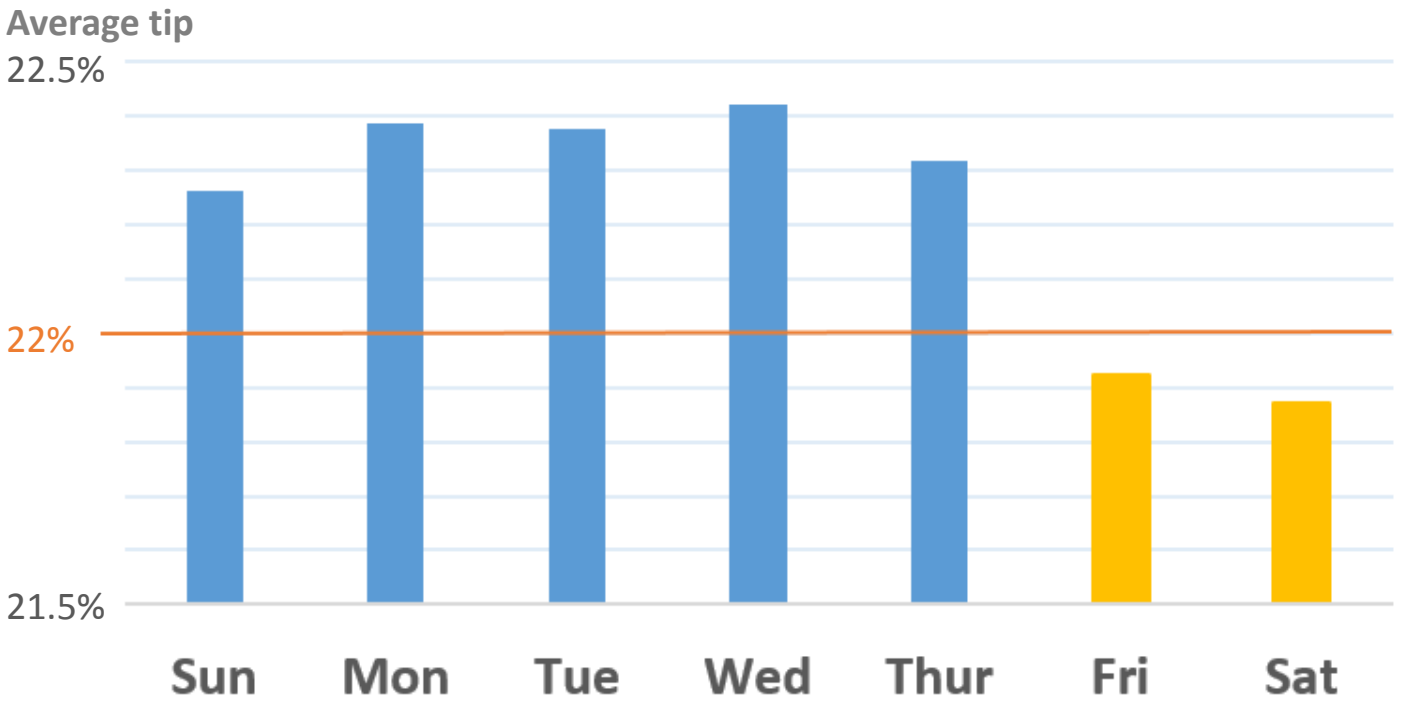
Payment type change in February, 2015 was accompanied with a ridership reboot.



DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— Day of week

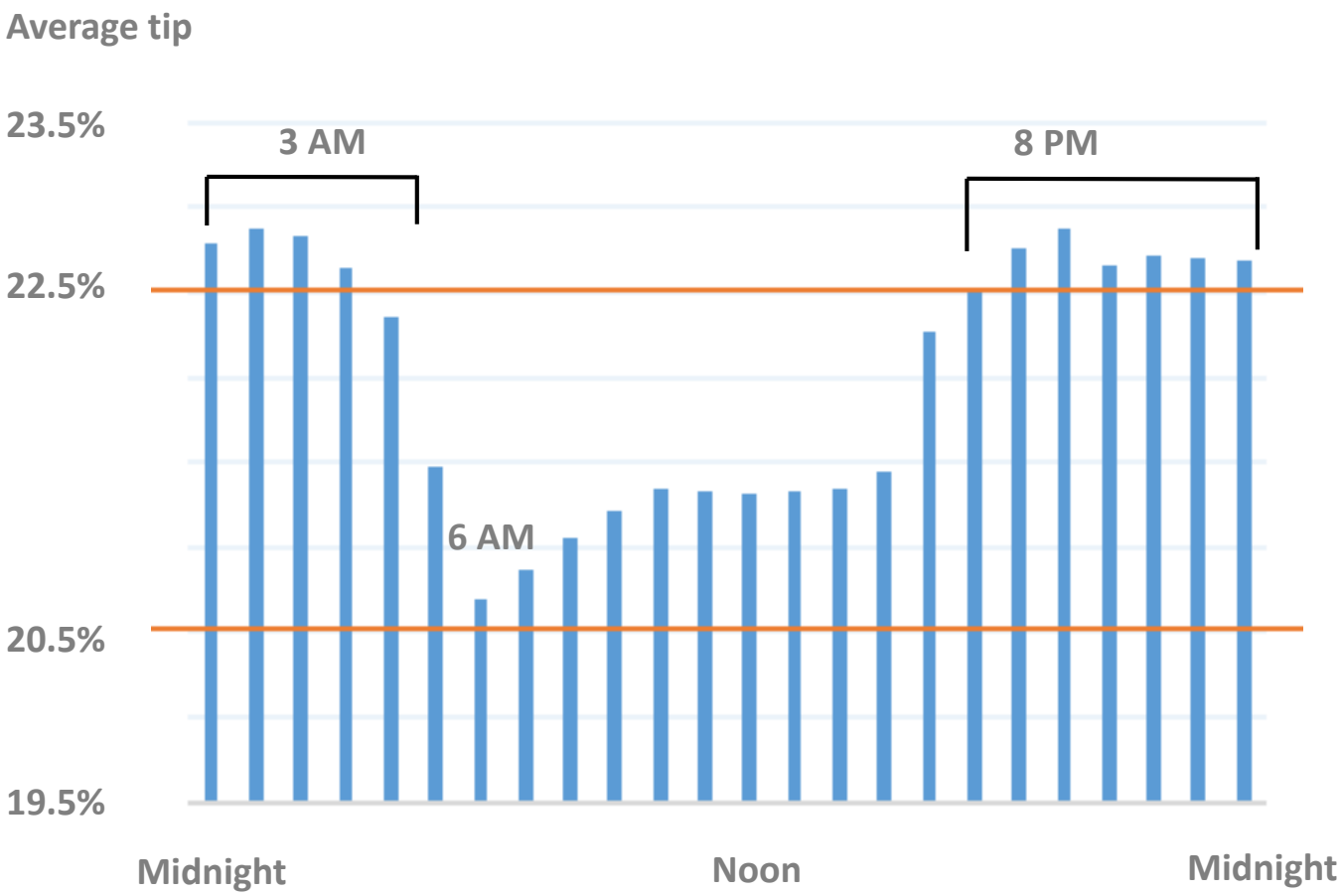
Average tip is better from Sunday through Thursday



DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— Hour

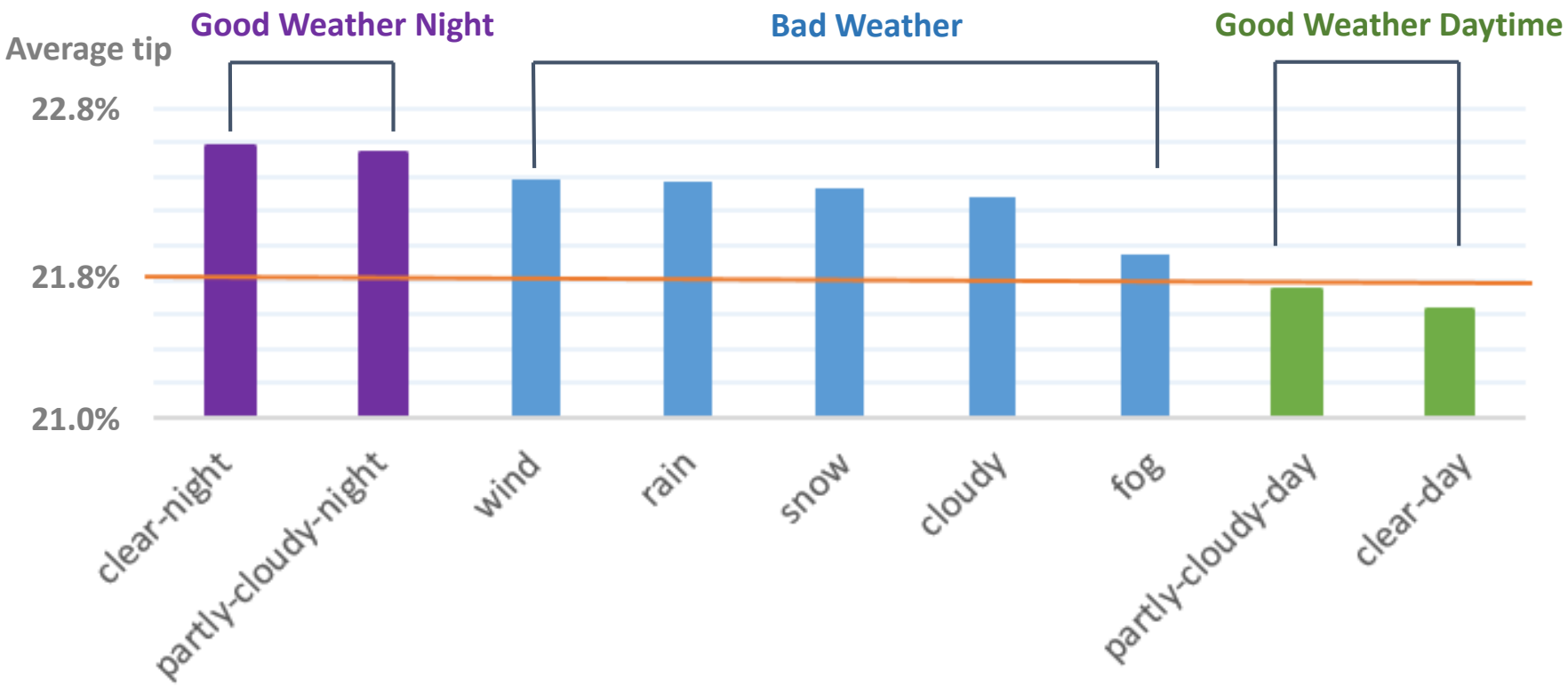
Average tip is better before sunrise or after sunset.



DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— Weather

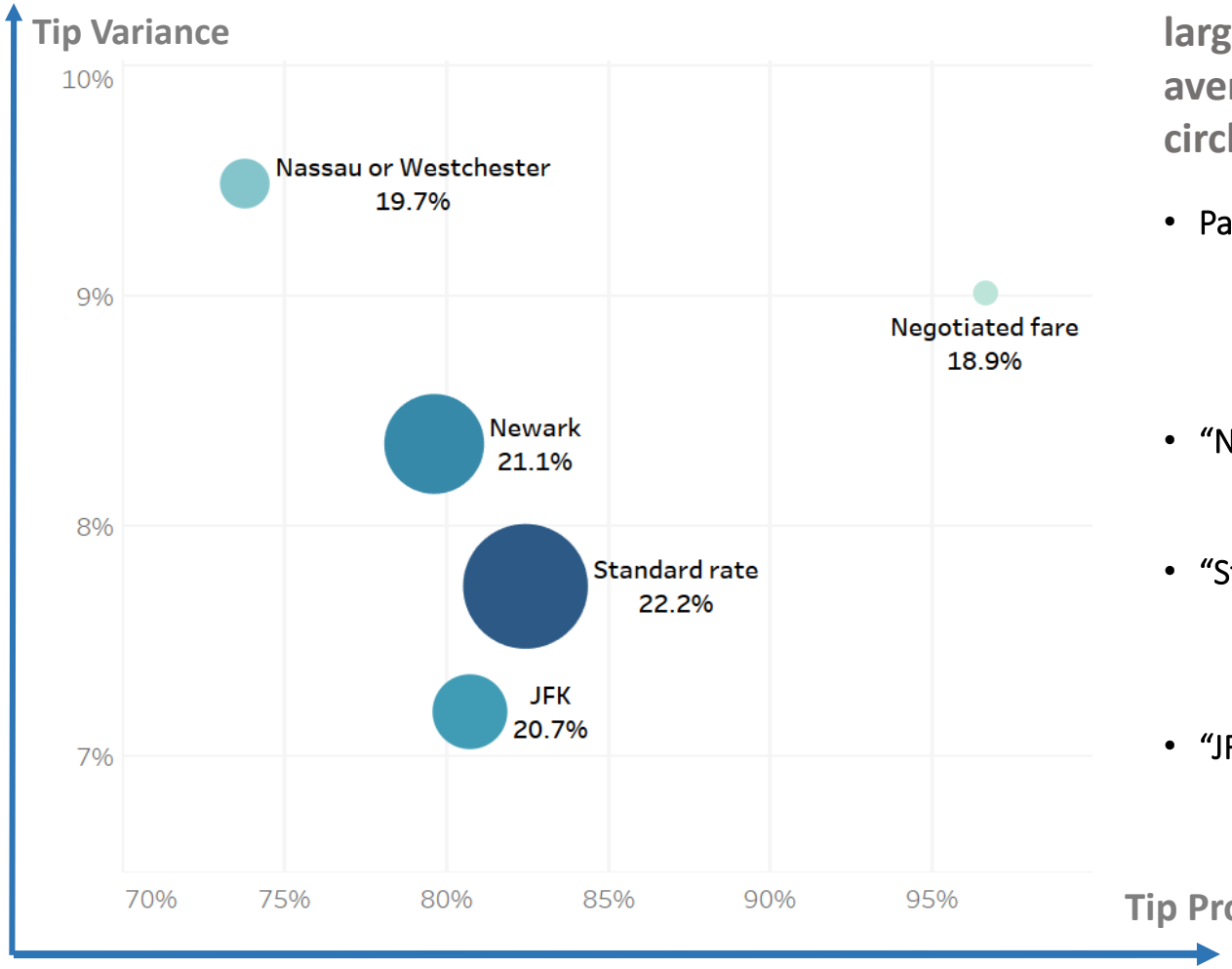
Passengers tip more during bad weather.



DATA: TAXI AND LIMOUSINE COMMISSION

Segmentation —— Rate Code ID

“Standard rate” passengers the most generous while “Nassau or Westchester” the least.



larger and darker circles indicate higher average tip while small and lighter circles represent lower average tip.

- Passengers heading to Nassau or Westchester:
  - not generous tipper
  - unpredictable: great variance in tip behaviors
  - over one quarter of them not tip
  - tip little (less than 20%)
- “Negotiated fare” passengers:
  - almost all of them tip
  - tip less (below 19%).
- “Standard rate” passengers :
  - the most generous
  - over 80% of them tip
  - tip over 22%.
- “JFK” passengers:
  - the most stable tipping behavior
  - tip around 20.7%

DATA: TAXI AND LIMOUSINE COMMISSION

Recommendation —— Tip

- Take credit card payment type as a strong signal (over 80%) that the customer would tip.
- Expect an average tip of 22% from “credit card” payers.
- Expect higher tip:
  - during bad weather hours
  - in the evening
  - on Sunday through Thursday.
- Expect generous tippers among those taking standard rate ride.

Summarized Recommendation

- Don’t miss any good weather Thursday afternoons or nights, as ride demands would soar and average tip would peak.
- 6 AM is always bad time for business opportunity: few passengers, least tip.
- Chase the crowd: densely populated places and traffic hub indicate greater business potentials.
- Show up frequently in the following places (in zip codes): 10027, 10035, 10029, 11101, 10032, 11222, 11102, 11231, 11105, 11373. Passengers depart there.
- Don’t go long away into a destination like Little Neck, NY: earn less per minute and no coming passengers.
    - A bad ride sample: starts from a place near Brooklyn and heads to Nassau or Westchester:
        - ✓ long-distance (low efficiency)
        - ✓ passengers either not tip at all or tip below 20%
        - ✓ no coming passengers around

        Highly likely you drive back for long with a vacant cab. And it is Thursday night with good weather!

Time-wise

Place-wise

Future Work

- With tremendous noise and pollution in raw data detected, it is suggested to improve the data collection pipeline for more accurate analysis.
- A arising question in this project that whether it was the introduction of a new payment that spurred growth around February 2015 is worthy of further exploration.
- Green cab market has grown mature. How to develop it in a sustainable and creative way without any hurt to yellow cab business is an urgent task with significance.