E-scooter ride share data analysis

Analysis of Two Mobility Companies' Operation

Prakash Periasamy

Fremont, CA

2022

My Journey and Technical Background

10+ years

Professional Experience in Industry

Research Scientist	***Graduate Student MS and PhD***	***Hardware (chip) Reliability, Integration Engineer***	***Data Scientist (Product Analytics) - fleet management***
3.5 years	Graduation: 2012	4.5 years	4 years
`Analytics & Experiments in Materials Joining`	`Materials Science`	`Reliability Analytics & Engineering`	`Data Scientist/ Product Engineer`
National Lab in India			
2 peer-reviewed articles; 2 trade secrets 1 presentation award; 1 promotion	*6 peer-reviewed journal articles, 1 Book Chapter(First author); Visiting Scholar (TU-Darmstadt, Germany); 2 poster awards*	*Lead engineer; 8 US Patents; 2 poster awards*	*Lead owner for 2 end-to-end applications; proposal winner ($US 25K); 1 peer-reviewed article (2nd author); 1 promotion*
`C;Visual Basic; Excel; SQL`	`Mathematica; Origin;Igor-Pro`	`Mathematica`	`R; Shiny; R Markdown (notebook); RStudio (IDE); SQL; Docker`

Outline

- Background
- Executive Technical Summary
- Data Description
- EDA
  - Data Cleaning
  - Operational Metric Analysis
  - Temporal & Spatial Analysis
- Scooter Count Estimation
- Concluding Remarks

Background

Extracted e-scooter ride share data analysis from 2 mobility companies (Say Company A and B) operating out of Louisville, KY area

Data Source: https://data.louisvilleky.gov/dataset/dockless-vehicles

Goal:

1) To extract critical operational metrics and compare company A and B

2) Also scooter ID is missing in company B. Can we estimate the number of scooters company “B” operates

Executive Technical Summary

Louisville, KY

Market Where Both Companies Operate

Refer Figure 1

~ 1.5x to 2x

Company A's Operation > Company B

Refer Table 1

12-6pm on Weekends (Sat)

When Demand Peaks (A & B)

Refer Figure 2 & 3

Casual/Recreational

Rider Persona (A & B)

Seems Like Non-Commuter Riders; Figure 2 & 3

About 30 (to 120)

Estimated Scooters of Company B

Logic: Overlapping Trips; Underestimation from True; Refer Figure 4a & 4b

Table 1: Select Operational Metrics (rounded values) for Company A & B

Table 1

SD= Standard Deviation

Operation Metrics	Company A Average *(SD)*	Company B Average *(SD)*
Trips Per Day (counts/day)	**353** *(140)*	**165** *(103)*
Distance Per Trip (miles/trip)	**1.5** *(1.8)*	**1** *(1)*
Duration Per Trip (mins/trip)	**17** *(21)*	**14** *(19)*
Trip Speed (mph)*	~**5.7** *(2.7)*	~**5.6** *(3.1)*
[*Rough estimate. no pause info]		
Average Revenue** Per Day ($/day)	**1261** *(625)*	**500** *(305)*
[**Assumptions: $1 unlockfee +]		

Figure 1

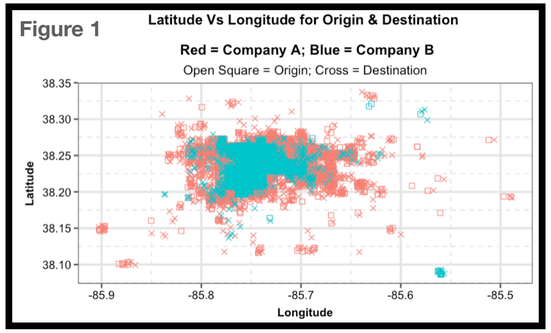


Figure 2

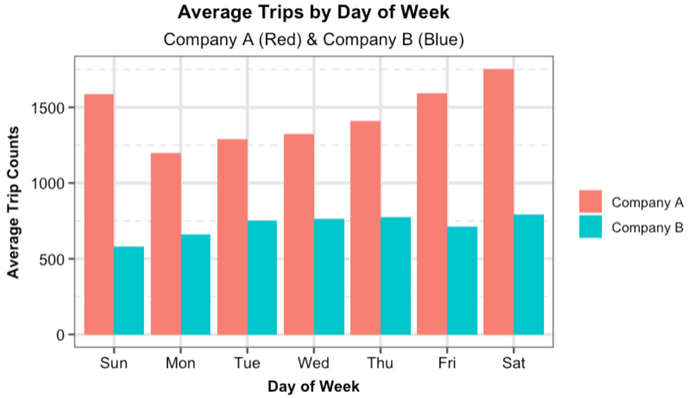


Figure 3

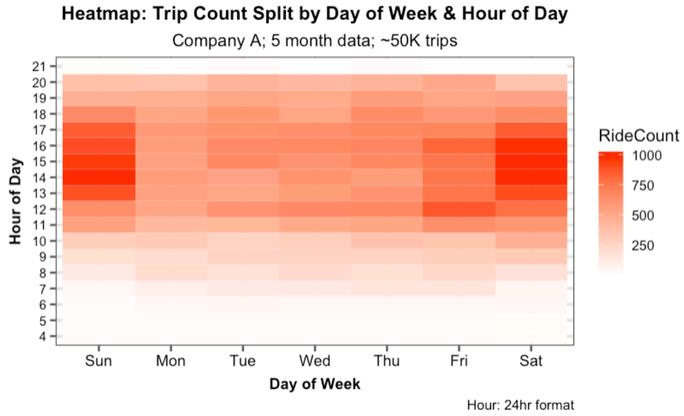


Figure 4a

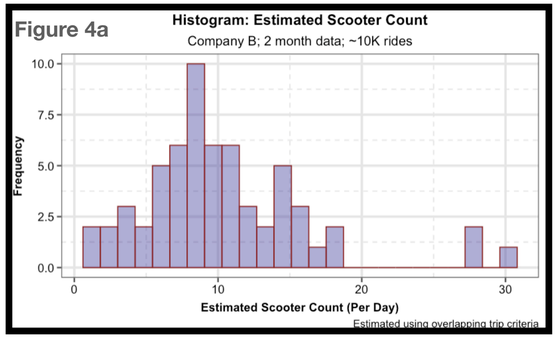
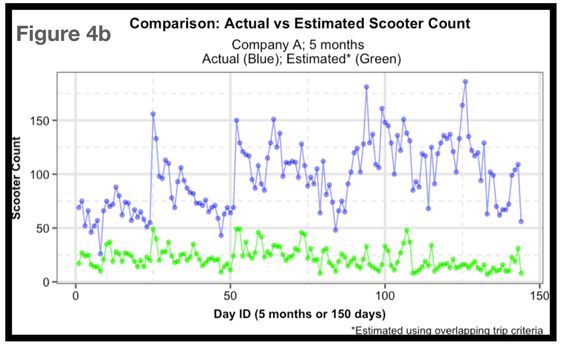


Figure 4b



Data Description

- Data: scooter rental share data from two mobility companies, A & B

TripID	ScooterID	Start & End Time	Start & End Latitude, Longitude	Trip Distance
Unique for each trip	Unique for scooter Given only in A and not in B	Time: Trip start and end	Origin & Destination Location Position	in miles (A), meters (B)

- Company A -> **5 months** data (Aug to Dec'18) & Company B -> **2 months** data (Dec'18 to Jan'19)
- **No duplicates** in both datasets
- **No missing values** in B dataset and A had very negligible % (1 missing out of 50K records)
- Converted B's TripDistance to miles
- Calculated TripDuration from Start & End Time [in mins]
- Calculated TripSpeed = TripDistance/TripDuration [in miles per hour]

EDA: Data Cleaning

Data Cleaning Procedure

I. Latitude (lat) & Longitude (lon) : Origin/Destination

- Fact: 0.01 in latitude (or longitude) ~ 0.7 miles
- Fact/Assumption: e-scooter range ~ 25 miles
- Hence, median lat/lon ± 0.4 = 28 mile radial distance from median location.
  - Based on this, valid boundary region = 28 mile radius
  - Dropped trips that had lat/lon > defined boundary

II. Trip Speed

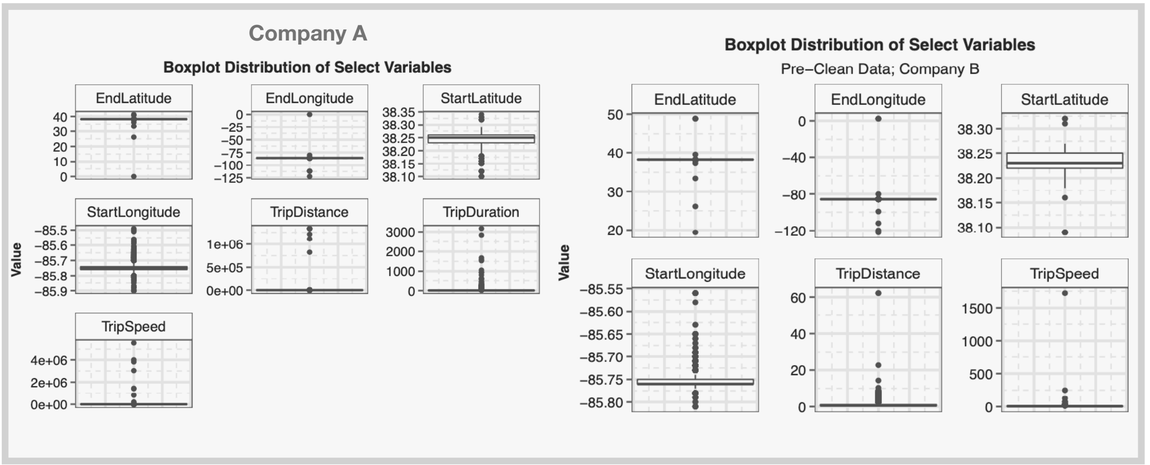
- Fact/Assumption: e-scooter max speed ~25-50 mph
  - Dropped trips with speed > 50 mph
  - This cleaned up “Trip Distance”, “Trip Duration” as well

~ 11% (A) & 5% (B)

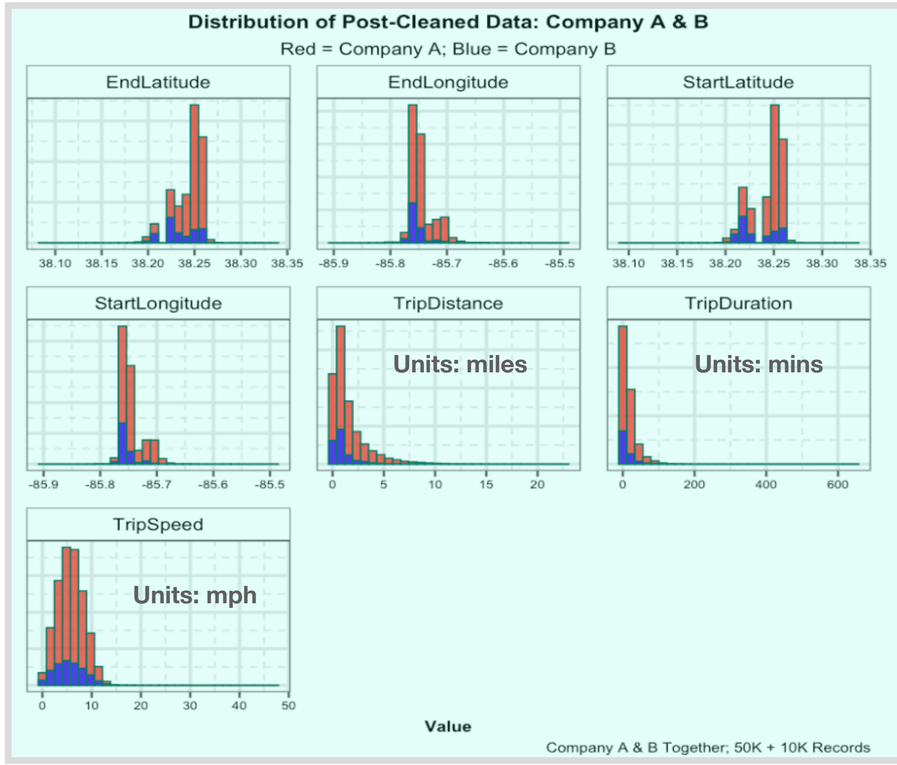
% records dropped post-cleaning

For more details refer backup slide/R Markdown notebook (Section Data Cleaning)

Pre-Cleaned Raw Data (+ Calculated Variables)



~11% (A) & 5% (B) data dropped post cleaning



EDA: Metric Analysis

Company A Operation > Company B

SD= Standard Deviation

Table 1: Select Operational Metrics (rounded values) for Company A & B

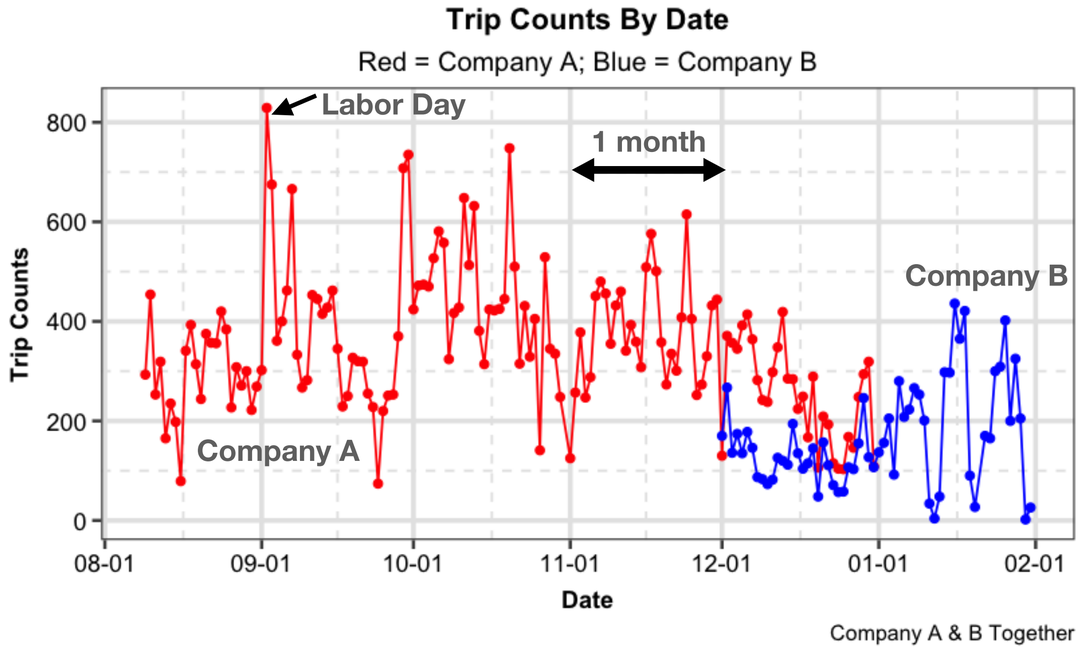
Operation Metrics	Company A Average *(SD)*	Company B Average *(SD)*
Trips Per Day (counts/day)	**353** *(140)*	**165** *(103)*
Distance Per Trip (miles/trip)	**1.5** *(1.8)*	**1** *(1)*
Duration Per Trip (mins/trip)	**17** *(21)*	**14** *(19)*
Trip Speed (mph)* [*Rough estimate. no pause info]	~**5.7** *(2.7)*	~**5.6** *(3.1)*
Average Revenue** Per Day ($/day) [**Assumptions: $1 unlockfee +]	**1261** *(625)*	**500** *(305)*

~ 1.5x to 2x

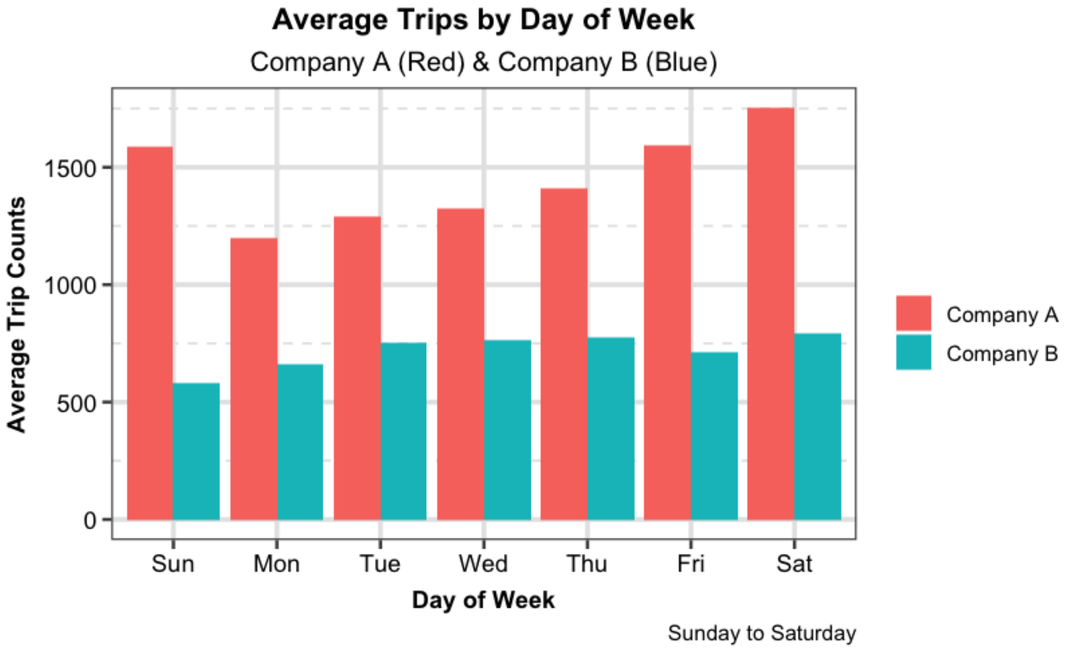
Company A's Operation > Company B

Refer Table 1

Trips By Date & Day of Week
Peak demand on Saturday (A) ; Midweek & Saturday Peak (B)



Company A & B Together



Sunday to Saturday

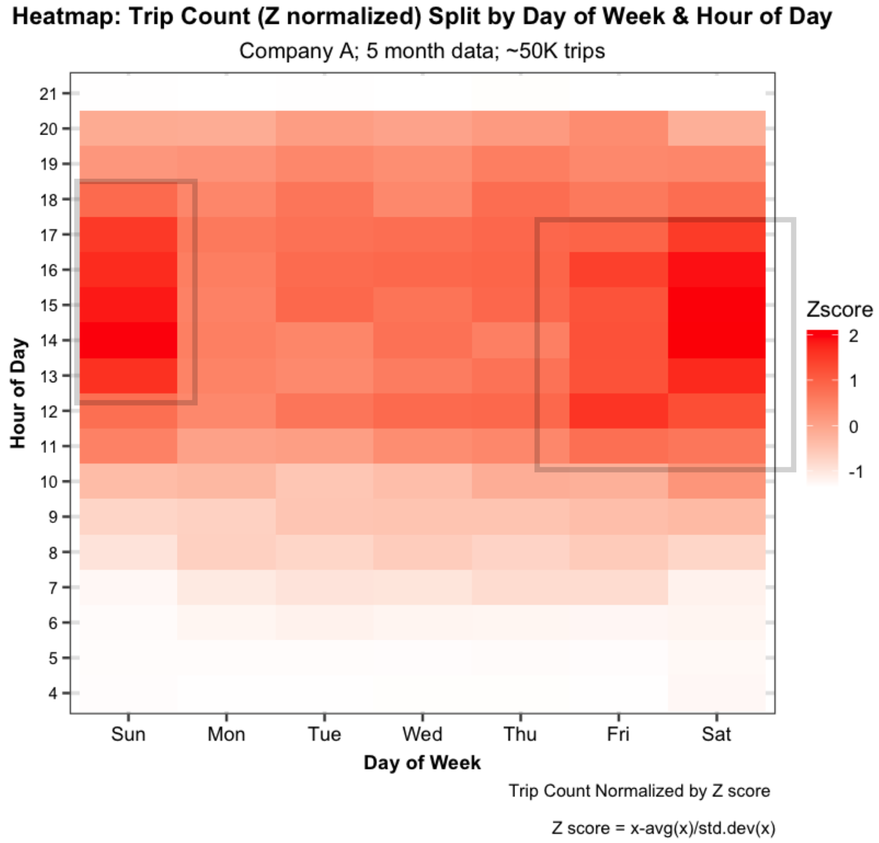
~ 350 trips/day; Peaks on Weekends (~ 1750 trips on sat)

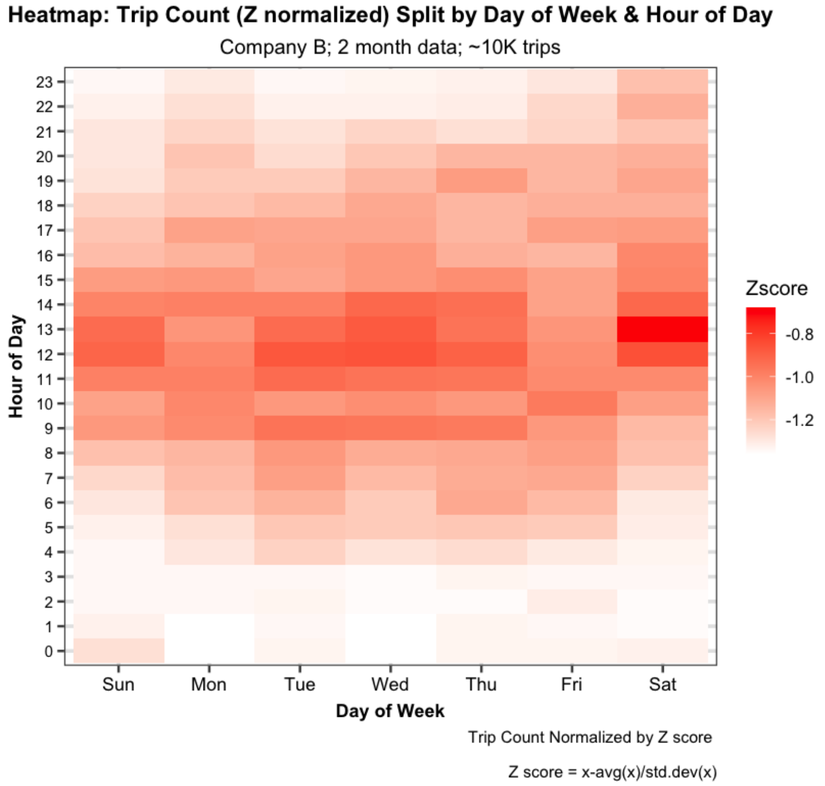
Company A

~ 160 trips/day; Peaks on Midweek & Saturday (~ 750 trips)

Company B

Temporal Analysis: What time is peak demand?
Noon - 6 PM Peak Demand On Weekends (A)



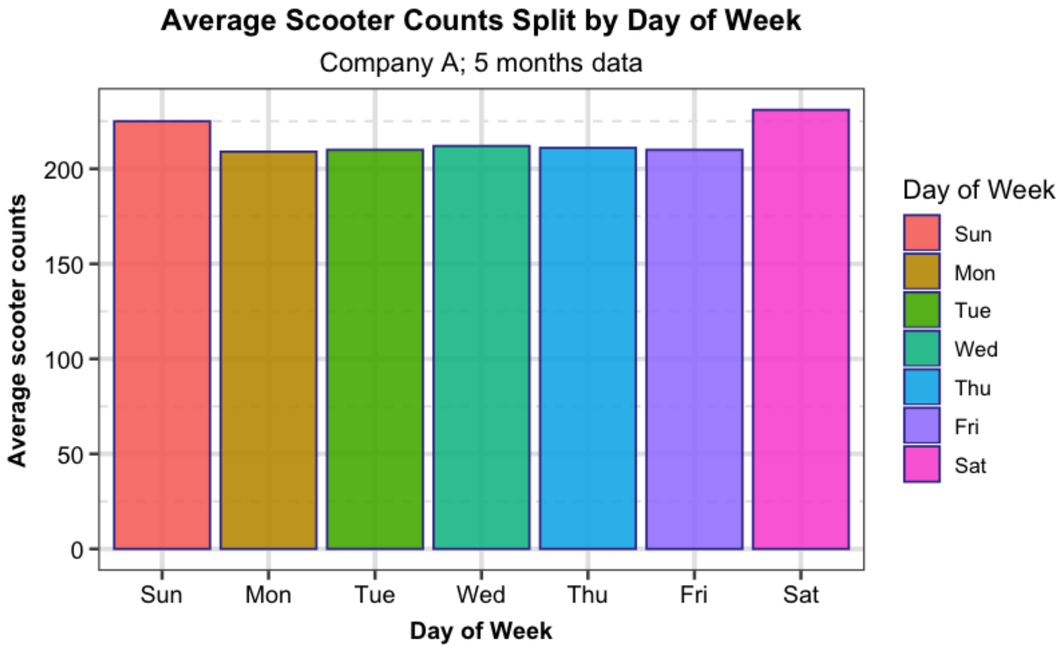


Casual Riders On Weekends (A&B) & Midweek (B) Between Noon to 6 PM

Different peak demand for both A & B

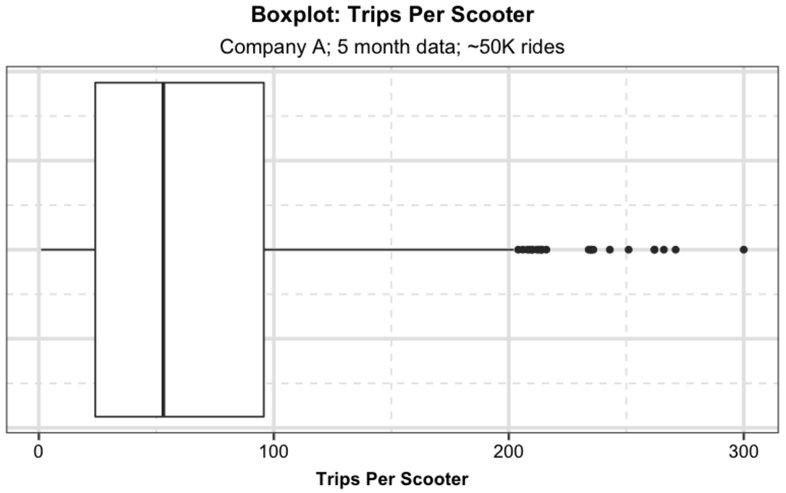
- A & B stations at different centers of attraction? One possible guess (with location positions & peak demand),
  - A near downtown which might explain demands on weekends (midday to evening). Possibly riders going to water front and back
  - B near hospital/univ. campus which might explain demands on midweek (midday). Possibly lunch goers

Scooter Related Metrics (A) : ~ 200 to 250 scooters per day of week



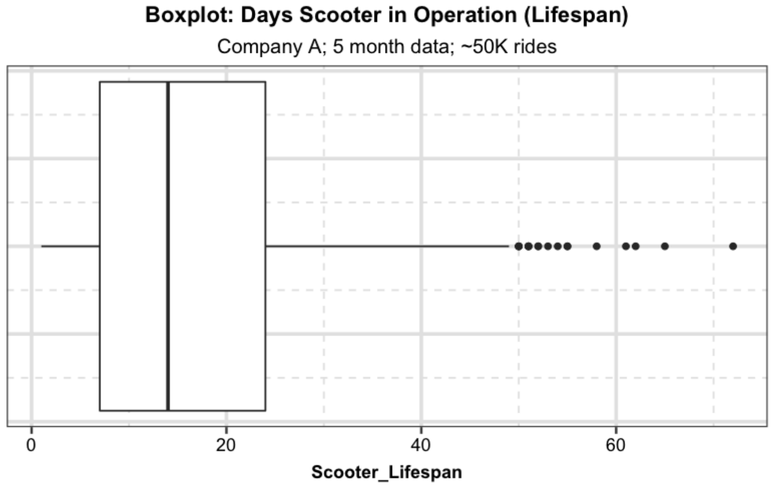
Little Dynamics in Scooter Inventory

Peaks on Weekends



~50 trips per scooter

Max 300 total trips

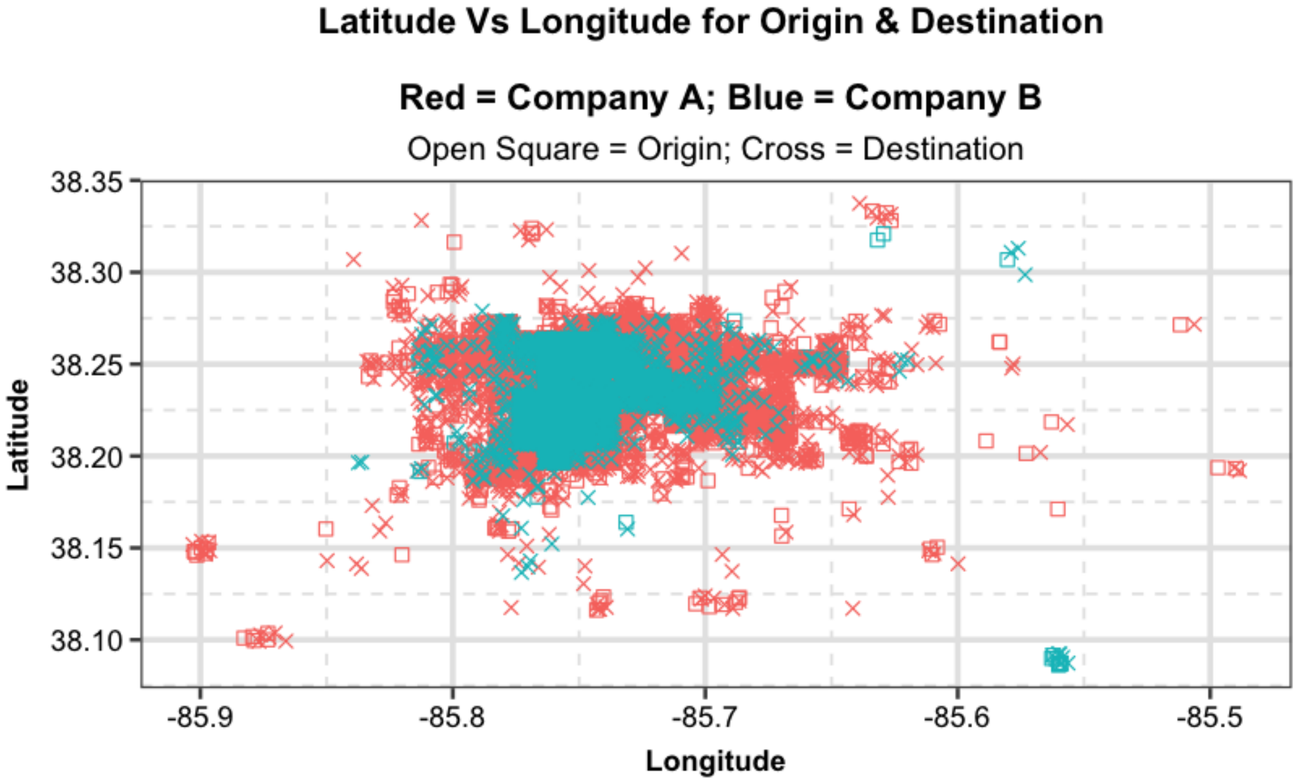


~14 days lifespan

Max ~ 70 days

Spatial Analysis: What is the Area Coverage? Any Common Origin-Destination (O-D) Pairs?

- Both operate out of Louisville, KY
- Company A seems to be centered around 38.25, -85.75 (lat, lon)
  - Near Downtown?
  - Majority of the trips is round trip: O-D location is same
- Company B seems to be centered around 38.23, -85.76 (lat, lon)
  - Hospital/Univ. Campus?
  - About 1.5 miles off from Company A
- Next: Should overlay with Louisville map



Company A's Operation Area Coverage > Company B

Majority are round trips, but good spread in Company A

Estimation: Number of Scooters (Company B)

What is the estimate number of scooters in B?
About 30 scooters and can go up to 120 scooters

Logic

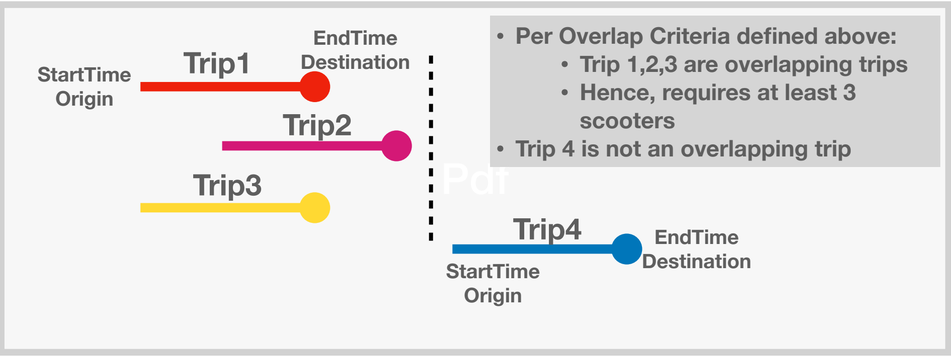
At a minimum, there should be as many number of scooters as there are overlapping trip.

- Because a scooter during a trip, can not be rented for another trip. Hence overlapping trips is a proxy for estimating the minimum number of scooters

Overlap Criteria

For two trips, trip1 and trip2, to overlap, the following condition should satisfy:

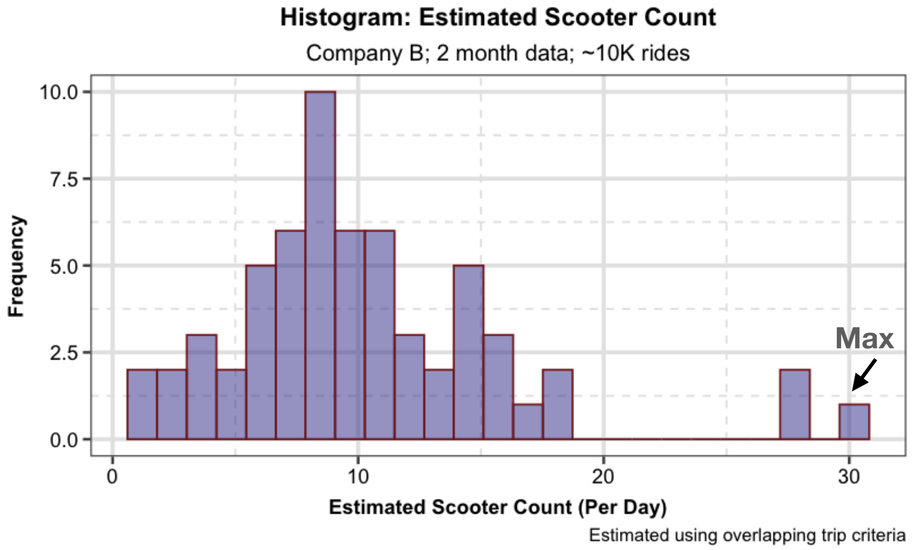
- **StartTime (of Trip1) < EndTime (Trip2) & EndTime(Trip1) > StartTime(Trip2)**

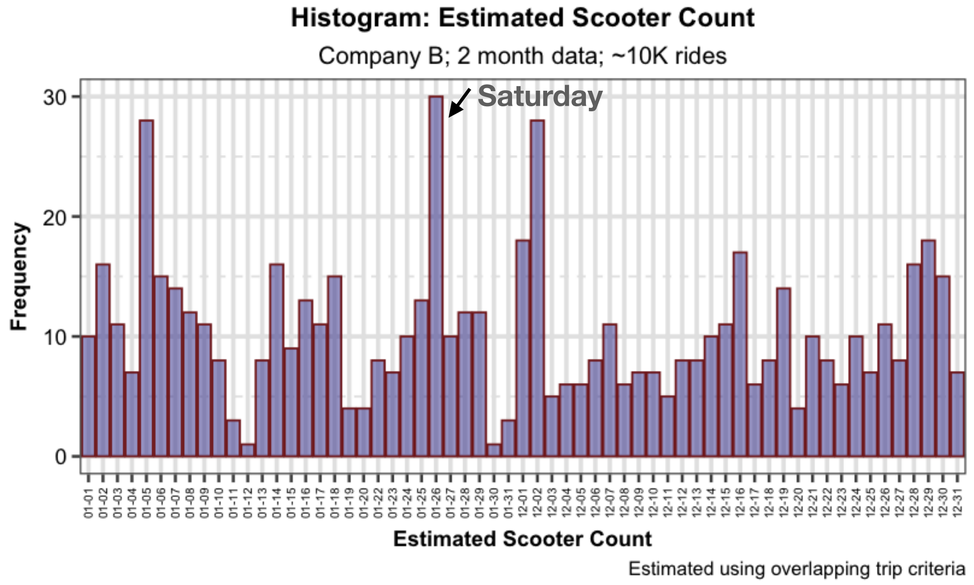


About 30 (to 120)

Estimated Scooters of Company B

See next slide for more info





How does the logic compare with actual scooter numbers in Company A? Why overlapping trip logic underestimates?

Execution Steps

Step 1: Calculate the maximum number of overlapping trips (using the above criteria) within a day. Repeat it for all days and look for the maximum in each case.

- Let's call it "max_overlaptrips_for_each_day"

Step 2: Get the overall maximum value from all of the calculated "max_overlaptrips_for_each_day"

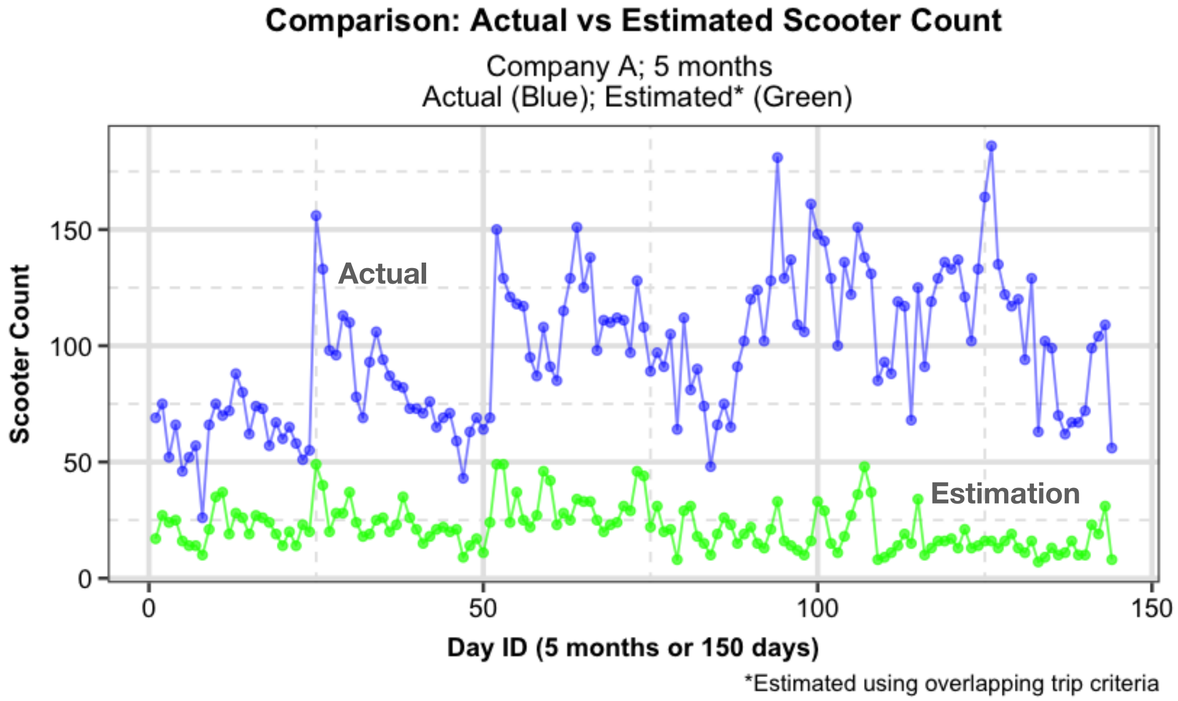
Result from Step 2, gives the estimated minimum of scooters

Limitations/Assumptions in this Strategy:

- **Assumption:** Scooter inventory is assumed to be constant for the entire time period.
  - Company A files indicates a dynamic inventory. We won't be able to capture that for Company B without the ScooterId
- **Limitation:** This strategy would underestimate the true maximum of scooters because:
  - it doesn't take into account start and end locations
  - if they never actually rented out the full capacity (outside the scope of the data)

Strategies to Improve Estimation Accuracy:

- Including location data (origin & destination) might improve the accuracy



4x Delta: Actual Vs Estimated

Scooter Count in Company A

Concluding Remarks/Next Actions

Concluding Remarks

See more detailed info in the attached R Markdown document

Both A & B Operate from Louisville, KY

A & B stations seems to be separated by ~1.5 miles

Company A's Operation > Company B

Daily revenue is ~$1260 ±625 /day (A) vs $500 ± 305/day (B)

50 vs 200* (A) & 30 to 120** (B)**

** Scooter Count Estimation from Overlapping Trip Criteria

* Actual Scooter Count from Data (A)

Executive Technical Summary

Louisville, KY

Market Where Both Companies Operate

Refer Figure 1

~ 1.5x to 2x

Company A's Operation > Company B

Refer Table 1

Weekend (Sat) Midday

When Demand Peaks (A & B)

Refer Figure 2 & 3

Casual/Recreational

Rider Persona (A & B)

Seems Like Non-Commuter Riders; Figure 2 & 3

About 30 (to 120)

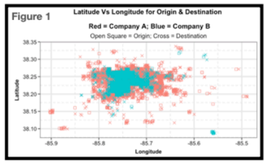
Estimated Scooters of Company B

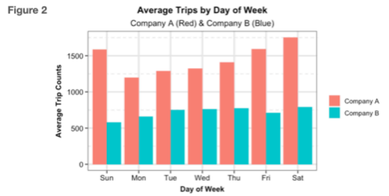
Logic: Overlapping Trips; Underestimation from True; Refer Figure 4a & 4b

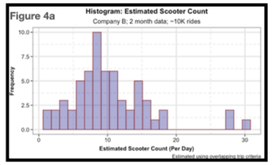
Table 1: Select Operational Metrics (rounded values) for Company A & B

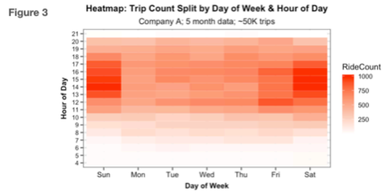
SD= Standard Deviation

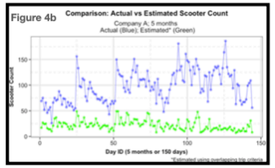
Operation Metrics	Company A Average *(SD)*	Company B Average *(SD)*
Trips Per Day (counts/day)	**353** *(140)*	**165** *(103)*
Distance Per Trip (miles/trip)	**1.5** *(1.8)*	**1** *(1)*
Duration Per Trip (mins/trip)	**17** *(21)*	**14** *(19)*
Trip Speed (mph)*	~**5.7** *(2.7)*	~**5.6** *(3.1)*
[*Rough estimate. no pause info]		
Average Revenue** Per Day ($/day)	**1261** *(625)*	**500** *(305)*
[**Assumptions: $1 unlockfee +]		











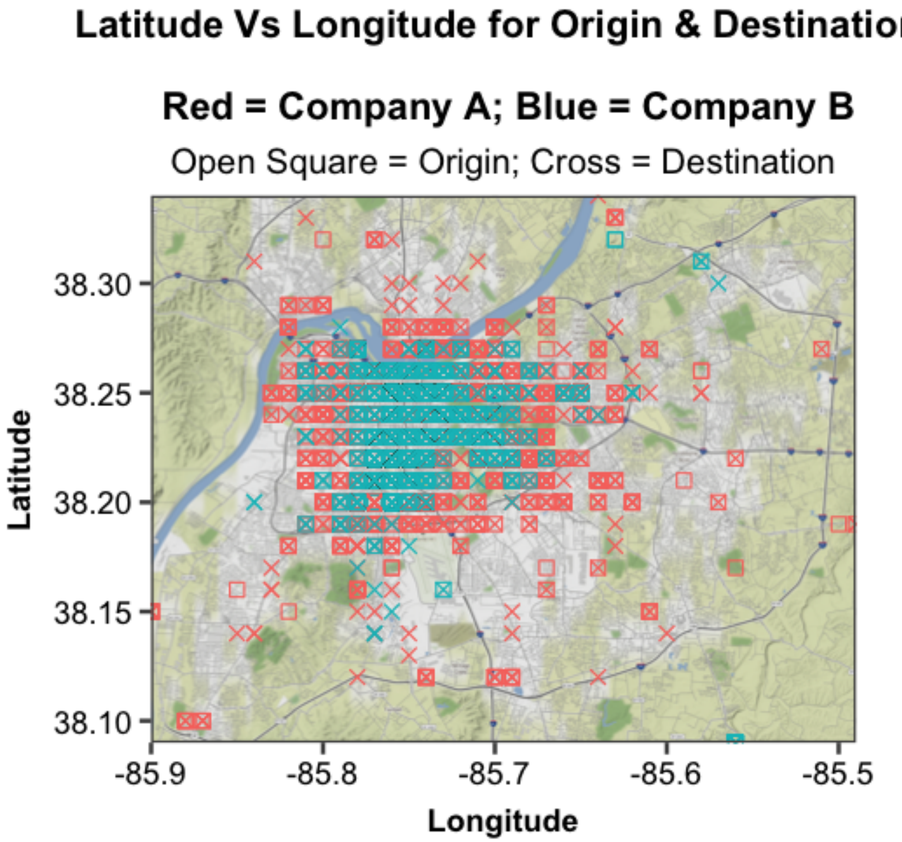
Next Actions

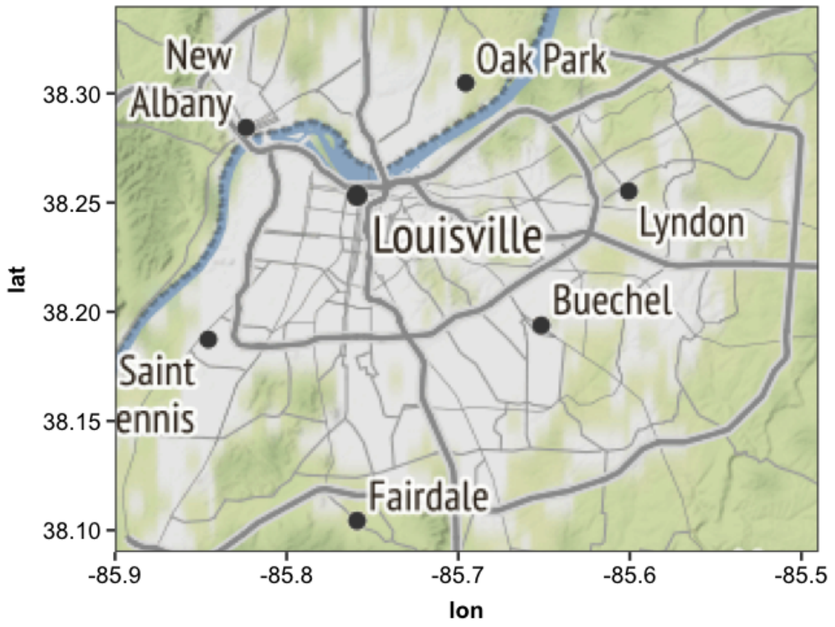
- Do more geo-spatial analysis by overlapping data with Louisville map
- Is there a way to include location information to further improve the scooter count estimation?
- Can you assess the performance of the operation? Bad Starts, Uncharged Scooters
  - Does the 11% vs 5% dropped (erroneous data) can be used as a proxy for performance?
- Is there a way to extract rides due to Juicing (crowd sourcing option to let riders take home the scooters to charge them and drop it off, the next morning)?

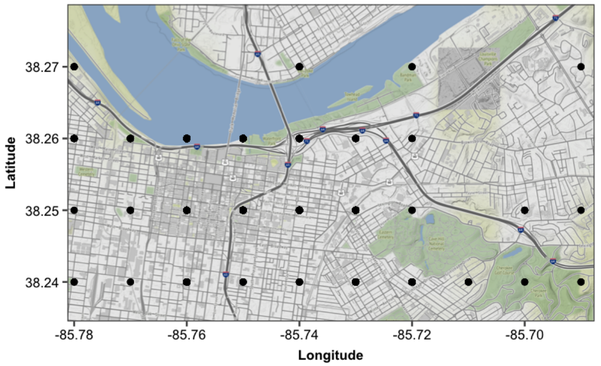
Thank You

Backup Slides

Overlay: Location Data on Real Map







Data Cleaning Procedure

- Latitude and longitude range to divide the entire planet is +/- 90 and +/-180 degrees, respectively. In our data, latitude and longitude are in 2 decimals. Hence it's a coarse precision considering the operation of e-scooters which typically have a range of only about ~25 miles.
- **0.01 in latitude or longitude represents approximately 0.7 miles**. Thus, in our data, location precision is limited to about 0.7 miles
- Louisville, KY's : latitude position is 38.328732, longitude position is -85.764771
- In this data, median StartLatitude and Endlatitude is 38.25
- **So making an approximation to go up to +/-0.4 (+/-28 miles) in latitude with 38.25 as the center, gives the extreme boundaries (37.85 to 38.65) of the latitude position.**
    - This gives a radial range of 28 miles from the center (median) latitude position of 38.25

- Median StartLongitude is -85.75 (as a reference, center)
- So following the same procedure as in EndLatitude, for EndLongitude as well, let's make gross approximation to go up to +/-0.4 (+/-28 miles) with -85.75 as the center, gives the extreme boundaries of longitude to be between approximately -86.15 and -85.35 in longitude.
    - It nicely covers the range covered by all data in StartLongitude as well
- Cleaning up outlier EndLatitude did fix the EndLongitude as well, confirming that those records have both erroneous EndLatitude and EndLongitude

Assumption: e-scooters can travel a max speed of 25-50 mph. Hence dropping records with TripSpeed >50 mph