Transit Data Analysis

Danielle Stacy, Siyue Yang, Jing Guo

Mentors: Candace Brakewood, Cheng Liu, and Kwai Wong
University of Tennessee

Background

Transit is an app used to collect and map real-time public transit data. People may use the app to determine which train or bus route to take, to plan a trip, or to search for the quickest form of transportation among other things. The data collected from the app has been organized into 13 different tables: device, favorites, feed download, installed app, location, nearby view, placemark, session complete, sharing system actions, sharing system purchase, trip, uber request, and user feed session.

Data Tables

- We only used a portion of the tables provided
- Danielle used the session complete, placemark, and favorite tables
  - The favorite table had to be included after the placemark table was found to be incomplete
- Jing used the Uber request data from Transit dataset，as well as the following public datasets from the Internet：Uber raw data, Taxi data in New York, New York Central Park weather data
- Alice used the bikesharing system actions in the Transit dataset together with another data source from a bikeshare operator

Home and Work Inferences of Transit Users

Danielle Stacy
The University of Alabama

Outline

- Research Question
- The Process
- A Clustering Example
- Extracting the Data
  - Which tables were used?
- Clustering and Labeling the Data
  - What clustering algorithm?
  - What labeling process?
- Evaluating the Accuracy
  - What is “accurate”?
- Results and Future Work
  - How were the results?
  - What else could be done?

Question

Can Transit users' home and work locations be inferred from the data collected from the users in the app?

~~In theory, a user would check the app in the morning at home to check the quickest way to work and then check again in the evening from work to search for the quickest way home.~~

~~A user is at home during the night and at work during the day, so their location data with the app should match this.~~

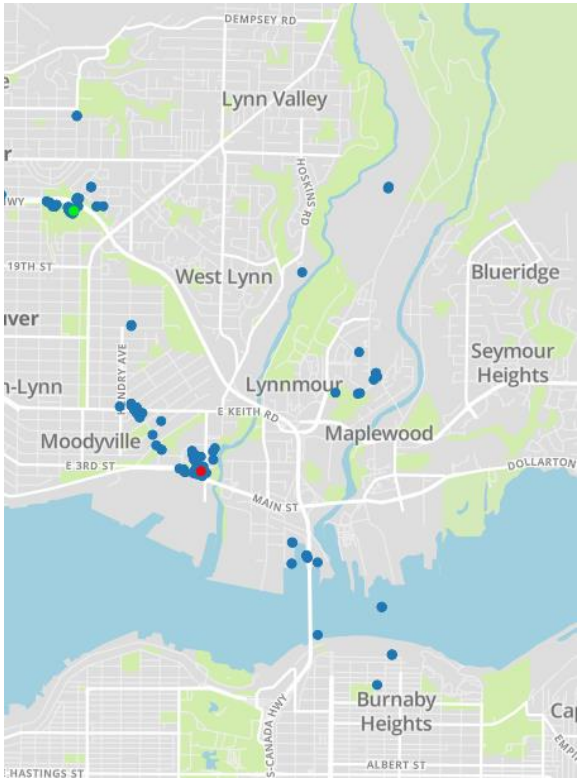
Location data should naturally cluster around two specific locations: home and work. To distinguish the home and work clusters, I will find where the user is more commonly during the weekend. Users work during the week and are at home during the weekend, so during the weekend, their location data should cluster around their home location. The goal is to check this assumption with the data provided from the app and determine if this is a valid process to infer home and work locations.

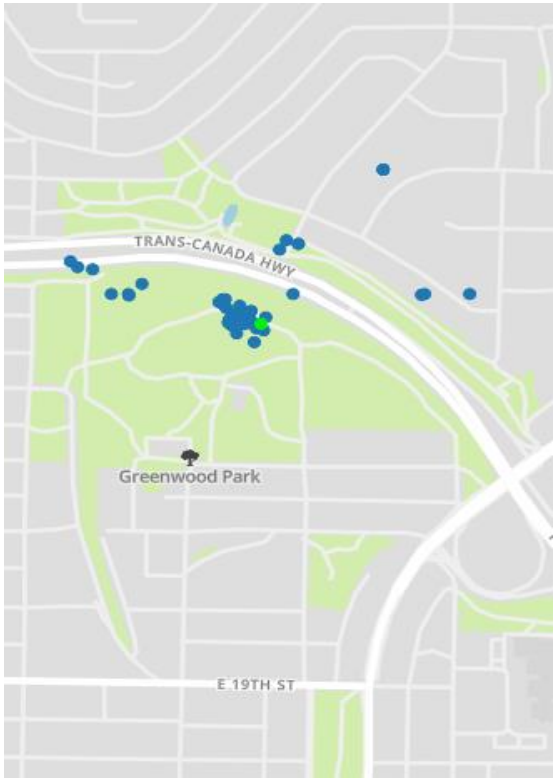
The Process

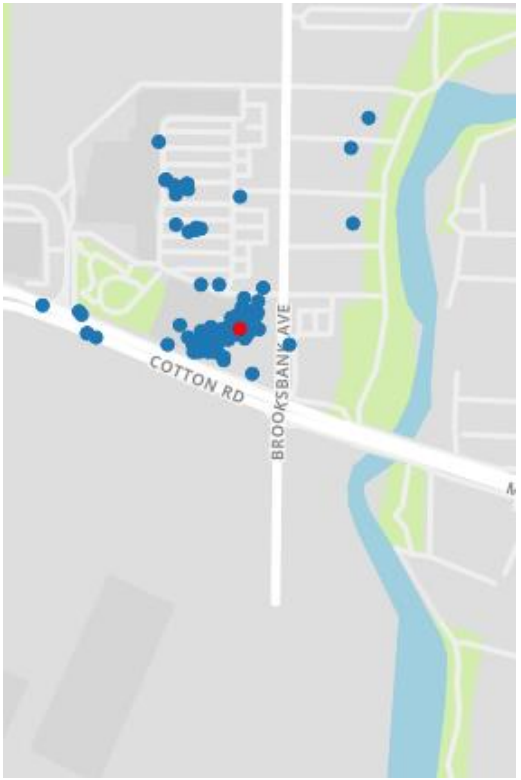
A unique identifier has been assigned to every user, so it is simple to keep track of a specific user across multiple tables. To check a user's location on the weekend, I will use the session complete table that provides a timestamp and location coordinates of the user when they opened the app. If there is clustering at a specific location during the weekend for a user, I would designated that location as the home and the other cluster should represent the work location.

To validate my chosen locations of the user's home and work, I will make use of the placemark and favorite tables. From these table, I can find users' stored home and work locations. I would check the coordinates from this table with the coordinates my algorithm found to establish an accuracy rate. The accuracy rate is determined by how many home and work locations were correctly found within a certain margin of error.

Clustering Example







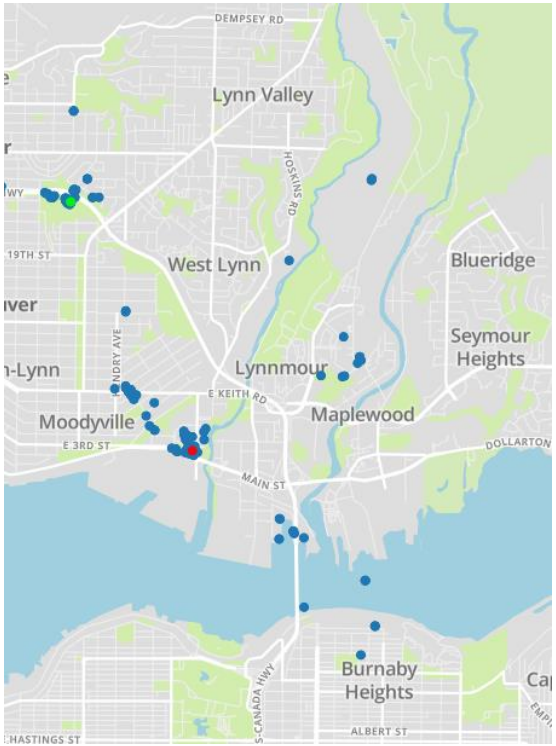
Coordinates have been shifted to protect user anonymity

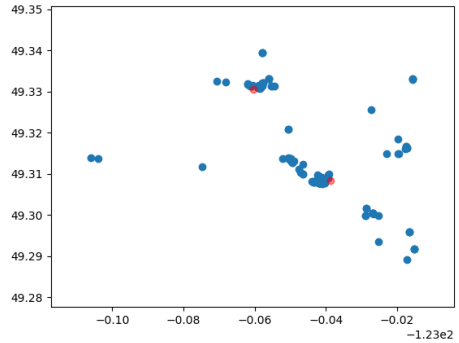
Extracting the Data

- Extract those who have saved home and work locations
- Remove duplicates
- Isolate those who have saved both home and work locations
- Extract location information from those udids
- Use of Python, Bash script, and OpenDIEL

Clustering and Labeling the Data

With the location data extracted, I can find clustering. I chose to use k-means as my clustering algorithm. Once the data is clustered, I am free to label the clusters as home and work. To do this, I find the percentage of data that occurs on the weekend for each cluster. The centroid of the cluster with the higher percentage of weekend data is labeled home and the other is work.






```
Home: [49.3079726, -123.0410949]
Work: [49.3110115, -123.0583576]
Algorithm-found home: [49.30841905263158, -123.0386882631579]
Algorithm-found work: [49.330624179104476, -123.06033597014925]
```

Coordinates have been shifted to protect user anonymity

Evaluating the Accuracy

With the data clustered and labeled, it is time to establish an accuracy rate. Validating the results will be easy since I have only used the data of users who have saved their home and work locations. I have defined a point as “accurate” when its coordinates are within 0.05 of the actual locations.

For example, if the actual coordinates are (38, -75) and my algorithm determined the coordinates are (38.5, -74.5), then the result is deemed accurate.

I also have evaluated the average of how far the algorithm-found coordinates are from the actual coordinates.

Results and Future Work

With the weekend/weekday labeling system and a margin of error or ±0.05, I reached a home accuracy rate of 65.5027% and a work accuracy rate of 56.1108%. The margin of error amounts to over 4 kilometers. The average distance of the algorithm-found coordinates from the actual home coordinates is about 80 kilometers! The average distance of the algorithm-found work from the actual work location is also about 80 kilometers.

```
Home accuracy rate: 0.6550270921131849
Work accuracy rate: 0.5611077664057796
Average distance from home: 0.09856185921476948, -0.7193104142689638
Average distance from work: 0.1440318463053732, -0.7285809828786194
```

Need to improve the accuracy rate and the margin of error. Ideally, the margin of error should only be about 1 kilometer. This may be done with a different clustering algorithm or a different labeling system. Clustering algorithms that are not susceptible to outliers nor require a predetermined number of clusters might be better than k-means. A labeling system that takes into account the patterns of the user's movements throughout the day may be more accurate than simply using the time data.

Uber, Taxi and TRANSIT in New York City

Analysis of characteristics of different travel modes

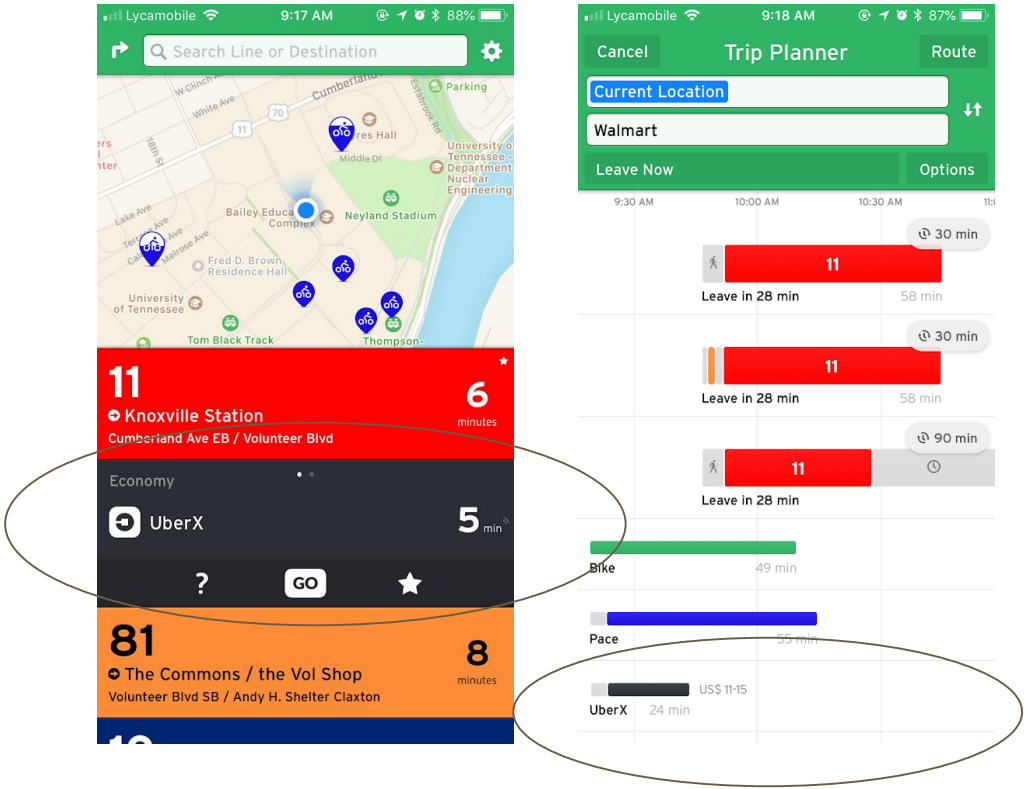
Jing Guo
Changsha University of Science & Technology

Joint Institute for Computational Sciences

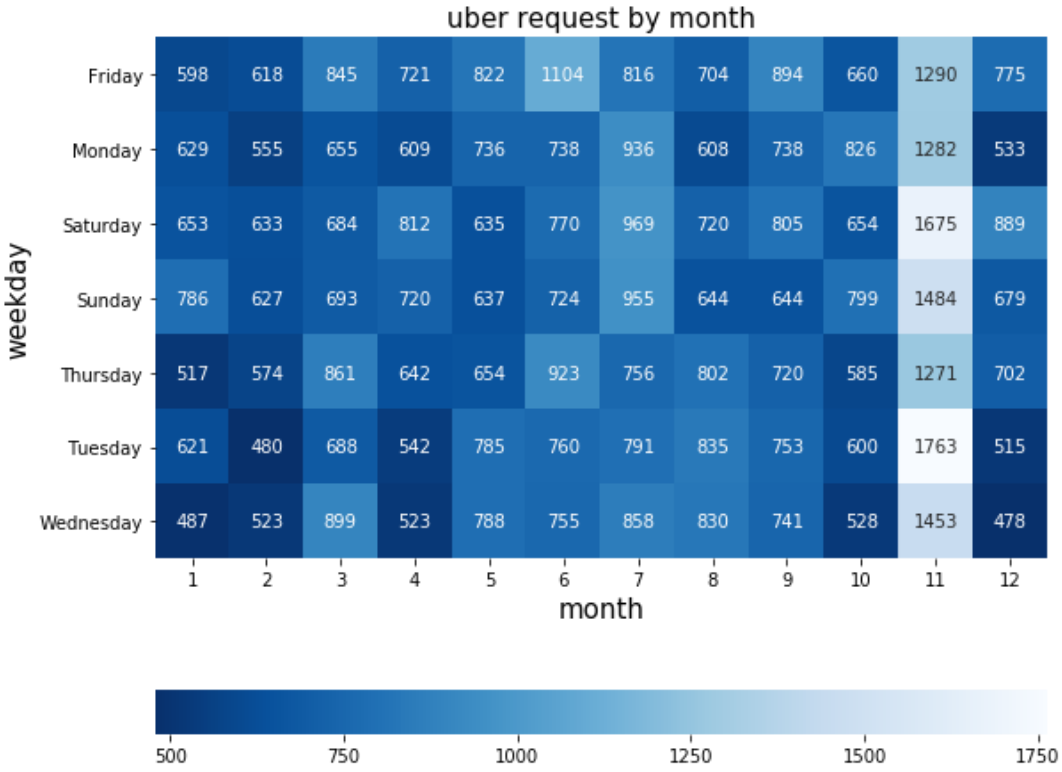
UBER REQUEST ANALYSIS(2016-11~2017-10)

Contents

- User request time trend
    - By weekday
    - By hour
- Uber types selected by user
- Analysis of the frequency of uber request
- Uber request & weather

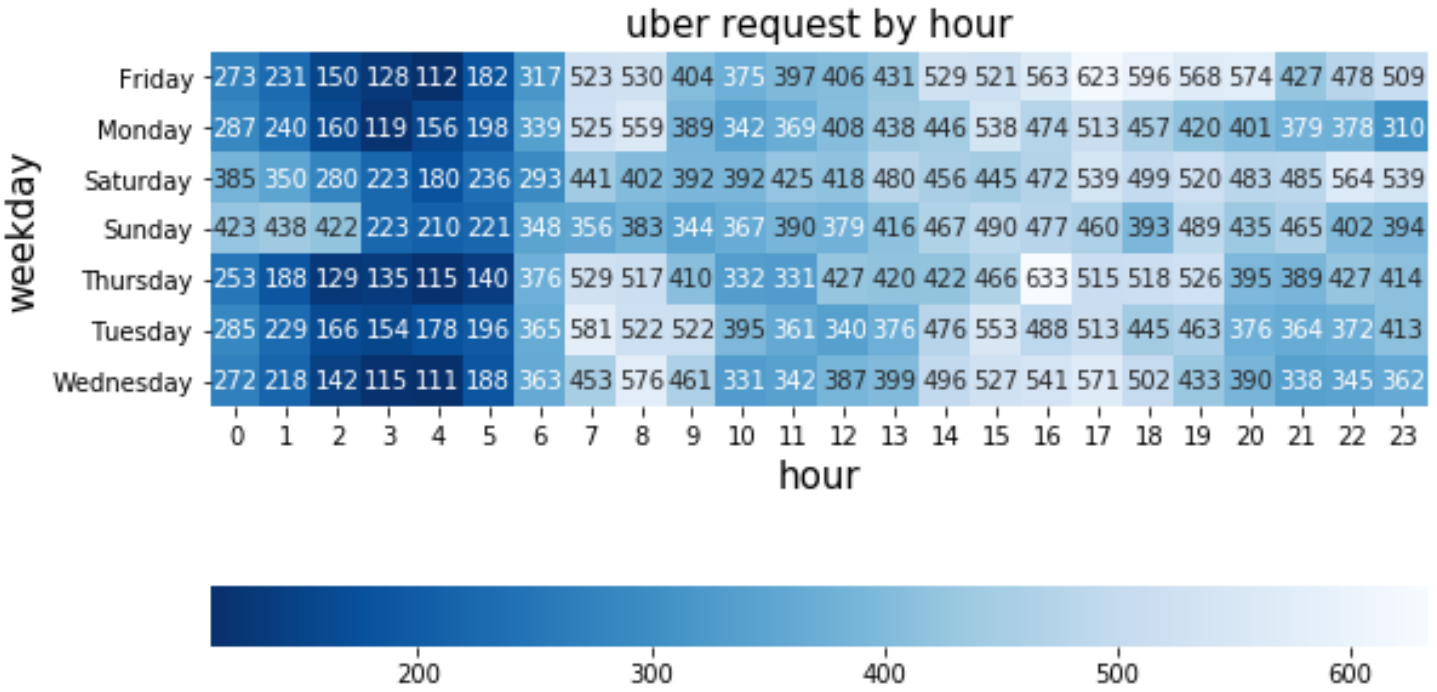


Pickup Time Trend By Month

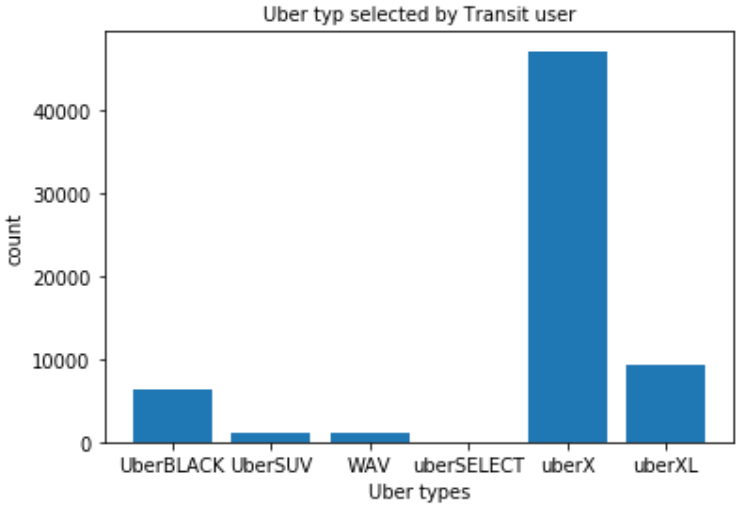


More users in November than in other months.

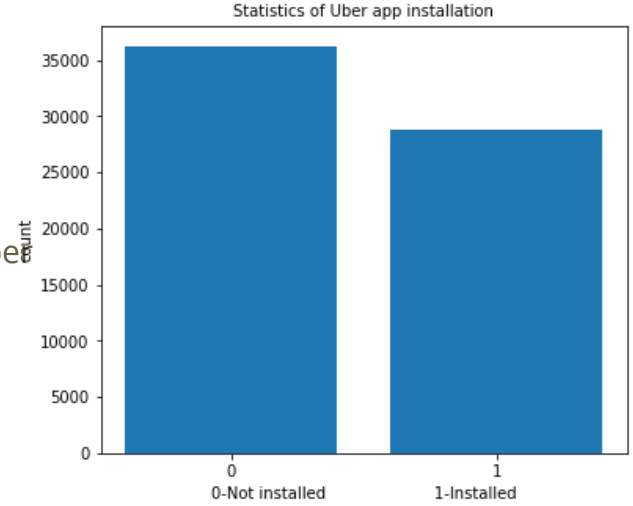
Pickup Time Trend By Hour



Uber types selectes by user



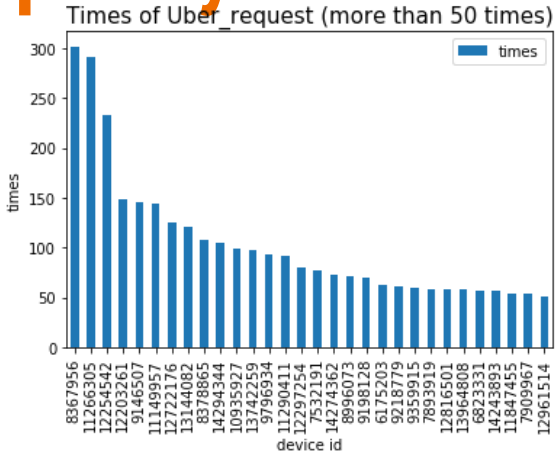
- UberX is the most popular type.
- More than half of the users did not install Uber app

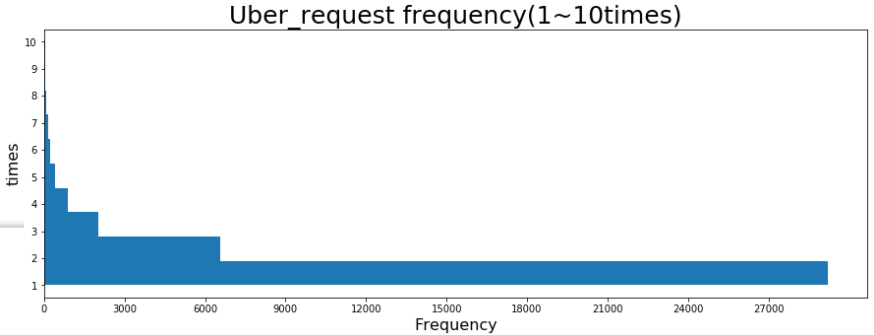


Analysis of the frequency of uber request

```
In [121]: device_id = request['device_id']
          fre_id = device_id.value_counts()
          fre_id
```

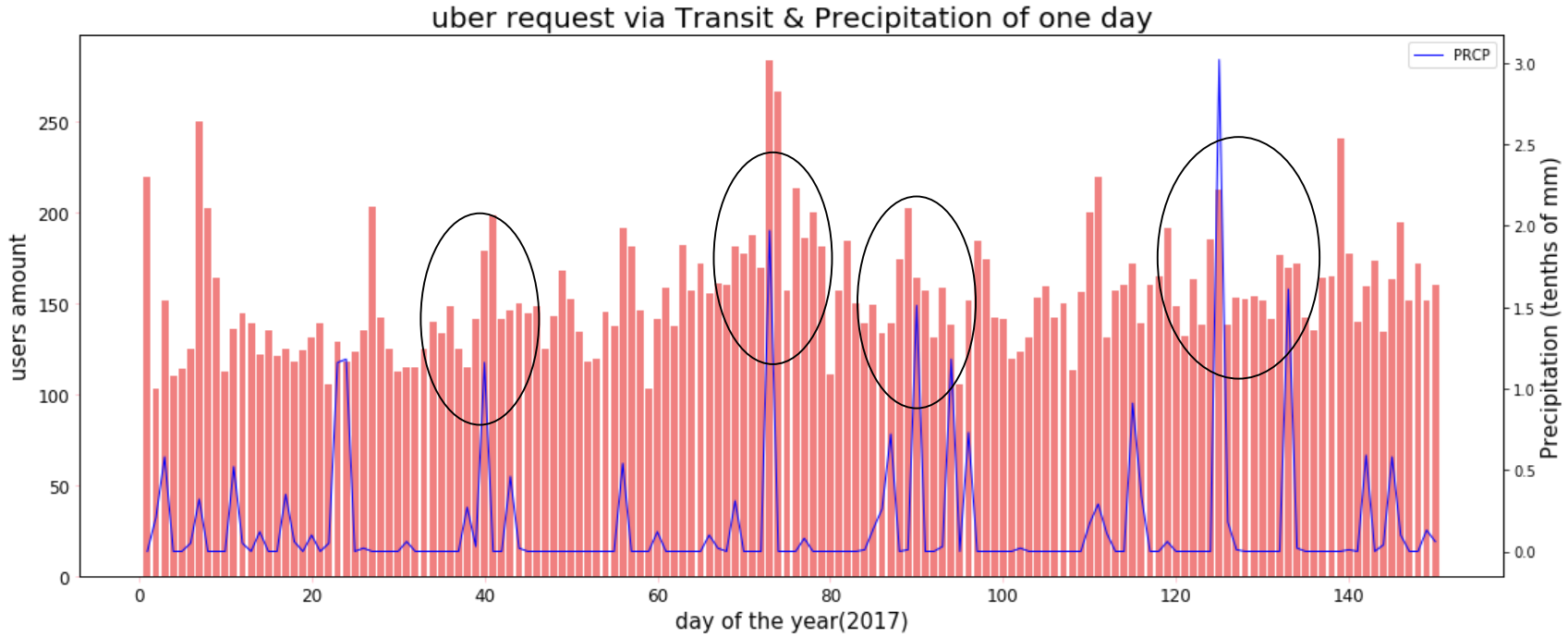
```
Out[121]: 8367956     302
          11266305    291
          12254542    233
          12203261    148
          9146507     146
          11149957    144
          12722176    125
          13144082    121
          8378865     108
          14294344    105
          10935927     99
          13742259     98
          9796934      93
          11290411     91
          12297254     80
          7532191      77
          14274362     73
          8996073      71
          9198128      69
          6175203      62
```





Although some users use Transit to find or book Uber over 50 times in a year, it can be seen from the figure that the proportion of users who have only used it once is very large. This means that most of the Uber request senders will only choose to use Transit to book Uber once.

Uber Request & Weather



The number of Uber requests on the day of the rainy weather is **increasing** compared to the adjacent days.

Uber & Taxi & Uber request via Transit Analysis and Comparison

Dataset

- Uber raw data
  (fromTLC:http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)
- Taxi in New York data
  (fromTLC:http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)
- Uber request data from Transit
- New York Central Park weather data
  (from:National Climatic Data Center https://www.ncdc.noaa.gov/)

Contents

- Proportion analysis of different methods
- Trip time trend (heat map)
- Trip & Weather

Proportion analysis of different methods

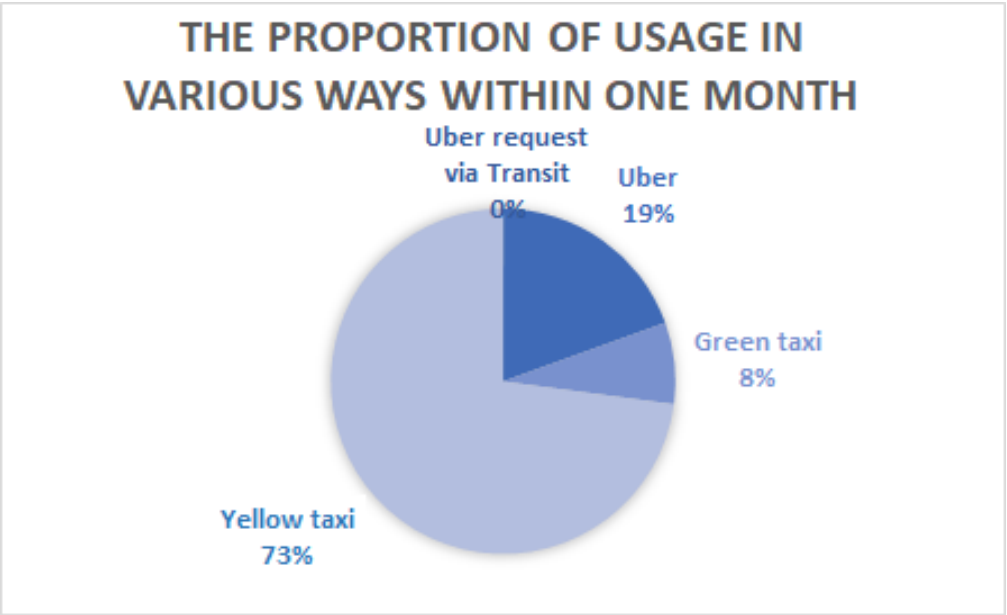
This is a pie chart of the proportion of users who choose to travel by different modes of transportation within one month.

From this we can get: Users who use Transit to book Uber are very few in comparison of all the data.

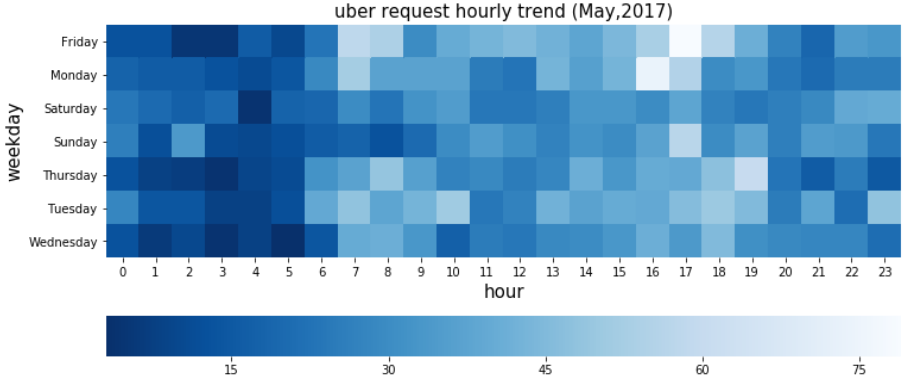
It is also possible to verify previous assumptions: the most primitive purpose of most users of Transit is to query public transit real-time information or use public transportation.

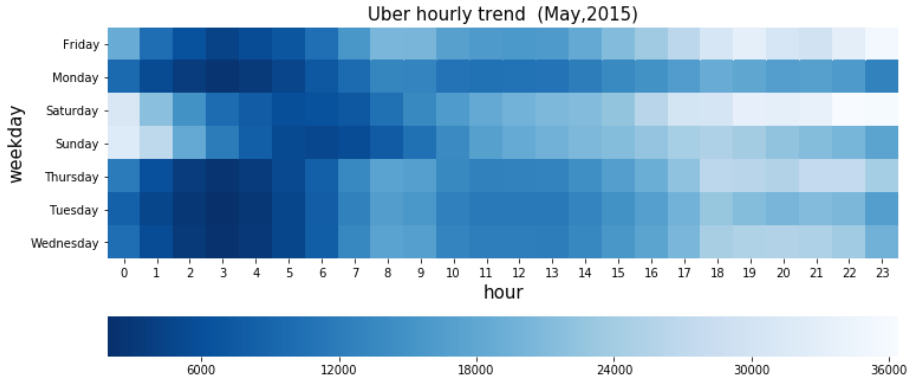
Yellow taxi users account for the largest proportion. They are more widely distributed than green taxis.

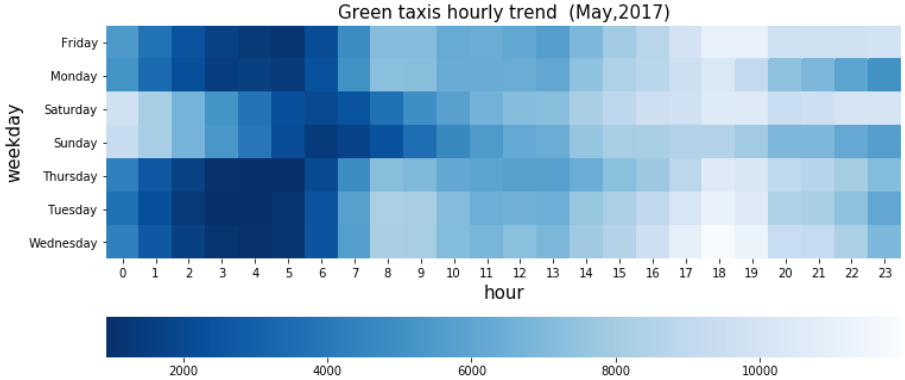
Yellow taxis are still a very important part of user travel choices

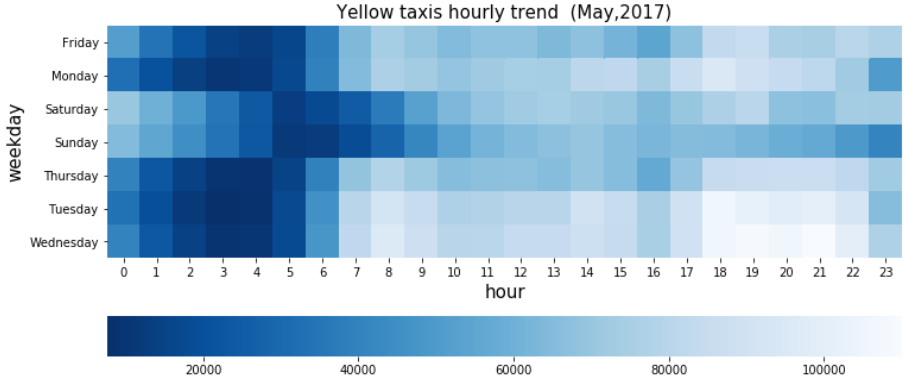


Trip time trend (heat map)

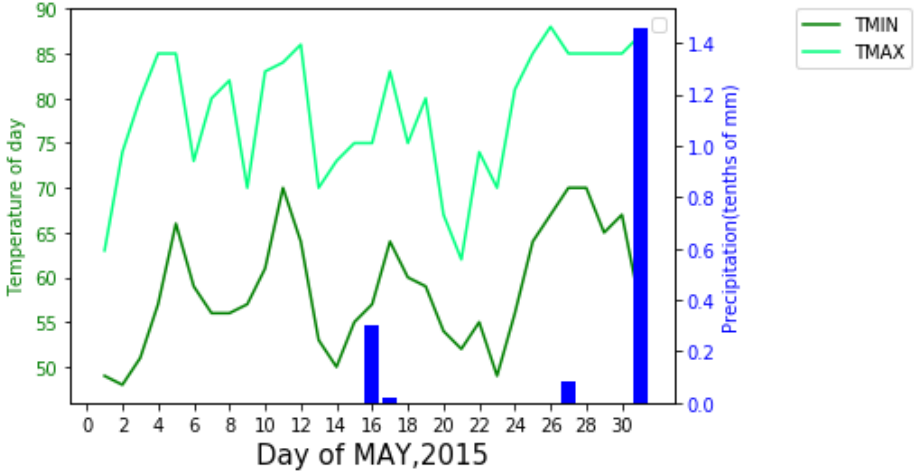


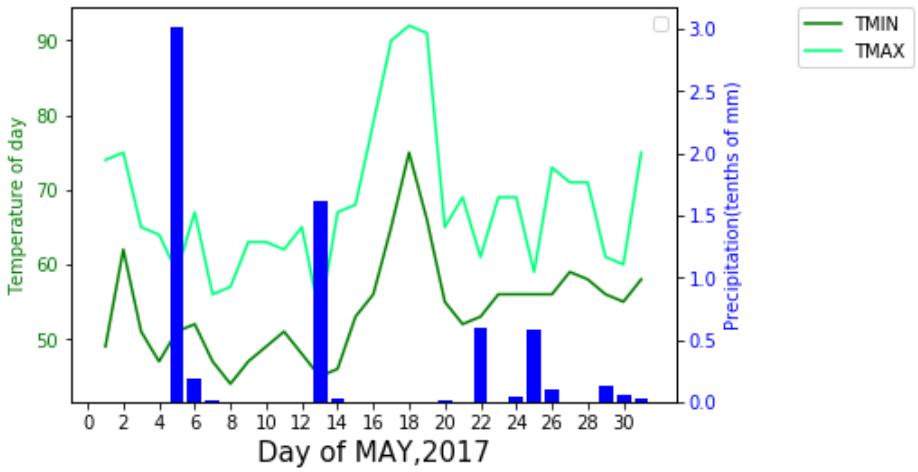






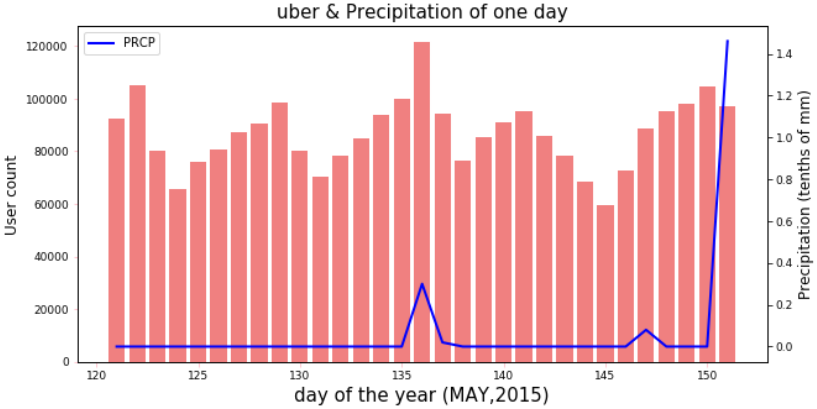
Weather Description

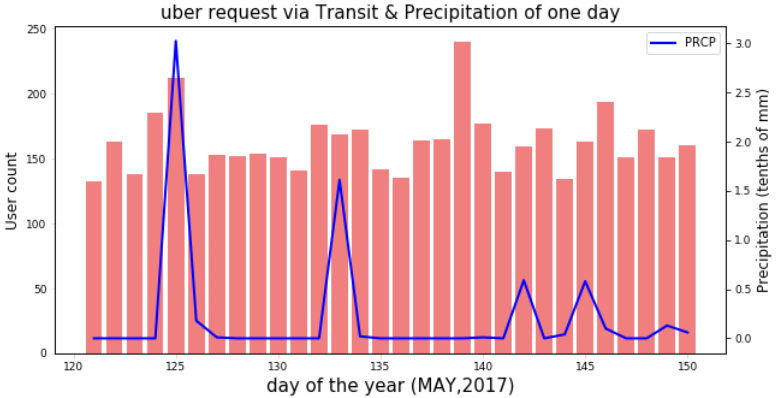


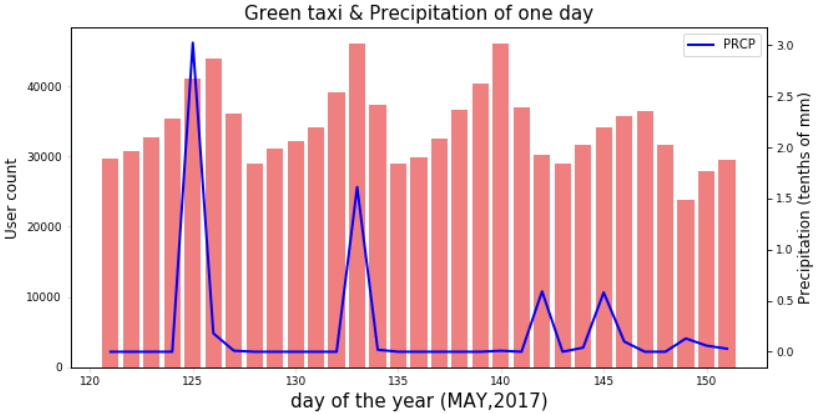


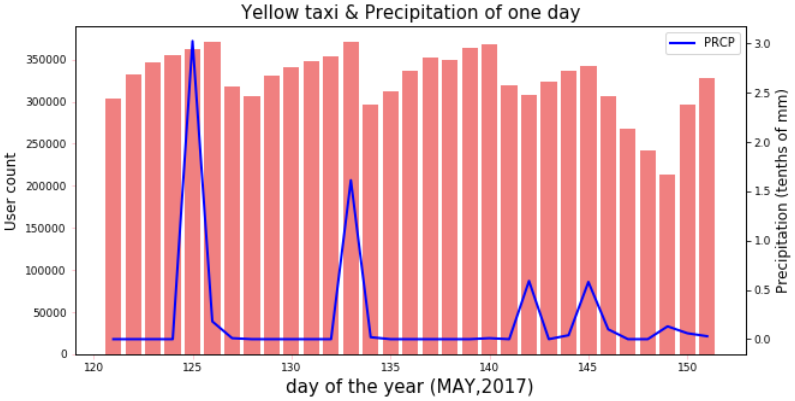
These two green lines represent the temperature of the day. The spring green one represents the maximum temperature of the day and the green one represents the minimum temperature of the day.
The blue histogram represents the amount of precipitation on that day.

Trip & Weather



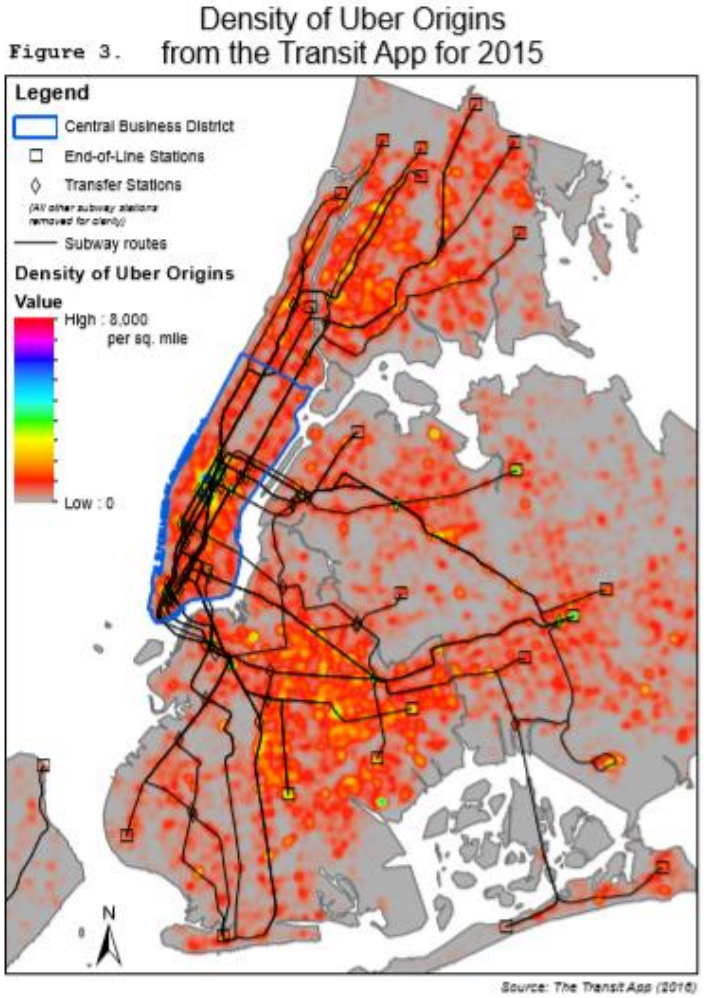






Future Work

- Geographical location analysis:
  - pickup location distribution (density map).
  - The distance of the user's location from the subway station or bus station.



Davidson, Peters and Brakewood (2017). Interactive Travel Modes: Uber, Transit and Mobility in New York City. Proceedings of the *96th Annual Meeting of the Transportation Research Board*, Washington, DC.

Bikeshare Users on a Budget?

A Trip Chaining Analysis of Bike User Groups in Chicago

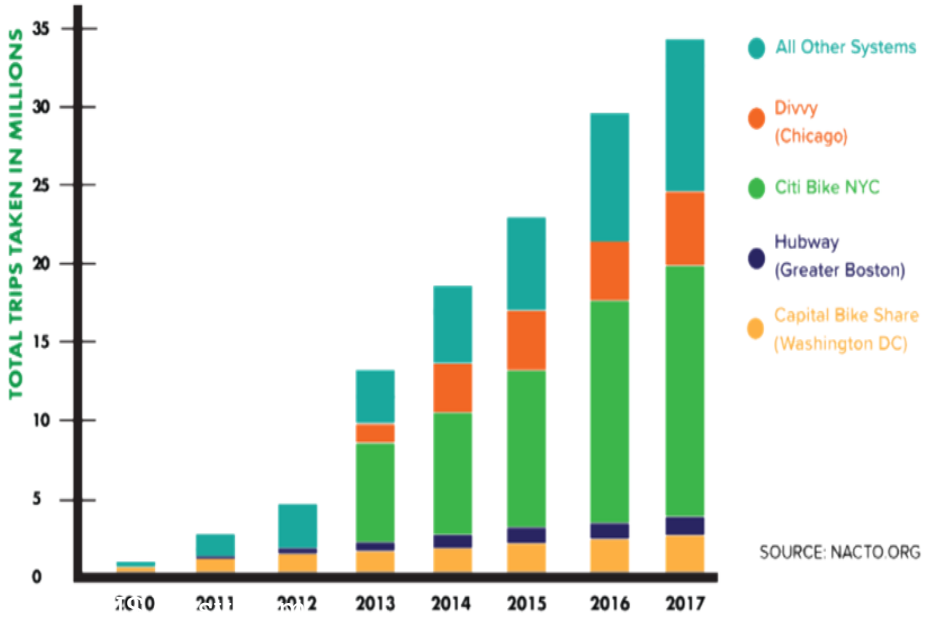
Siyue Yang (Alice)
City University of Hong Kong

Outline

- Introduction
  - Background of bikeshare
  - Bikeshare usage patterns
- Dataset
- Analysis
  - Analysis 1: Exploratory analysis of system trends
  - Analysis 2: Cluster analysis to identify user groups
  - Analysis 3: Trip chaining algorithm to identify “trip chaining” unlocks
- Summary

Background: Rapid growth of bikeshare

Bike share is increasing. When solving management problems, a good comprehension of usage pattern is useful.



Background: Backshare pricing policy changed in 2018

TABLE 1 Pricing Structure for the Divvy Bikesharing System

	2016		2018		
	Annual Membership	24-Hour Pass	Annual Membership	Explore (24-Hour) Pass	Single Ride
Fees	$99	$9.9	$99	$15	$3
0-30 minutes	$0		$0		
31-45 minutes	$1.5	$2	$0	$0	+ $3 for each additional 30 minutes
45-60 minutes			+ $3 for each additional 30 minutes		
61-90 minutes	$4.5	$6			
91-150 minutes	+ $6 for each	+ $8 for each			
Each additional	+$6	+$8	+$3		

Introduction: Bikeshare usage patterns

1. **Usage purposes: potential user groups in bikeshare [1]**

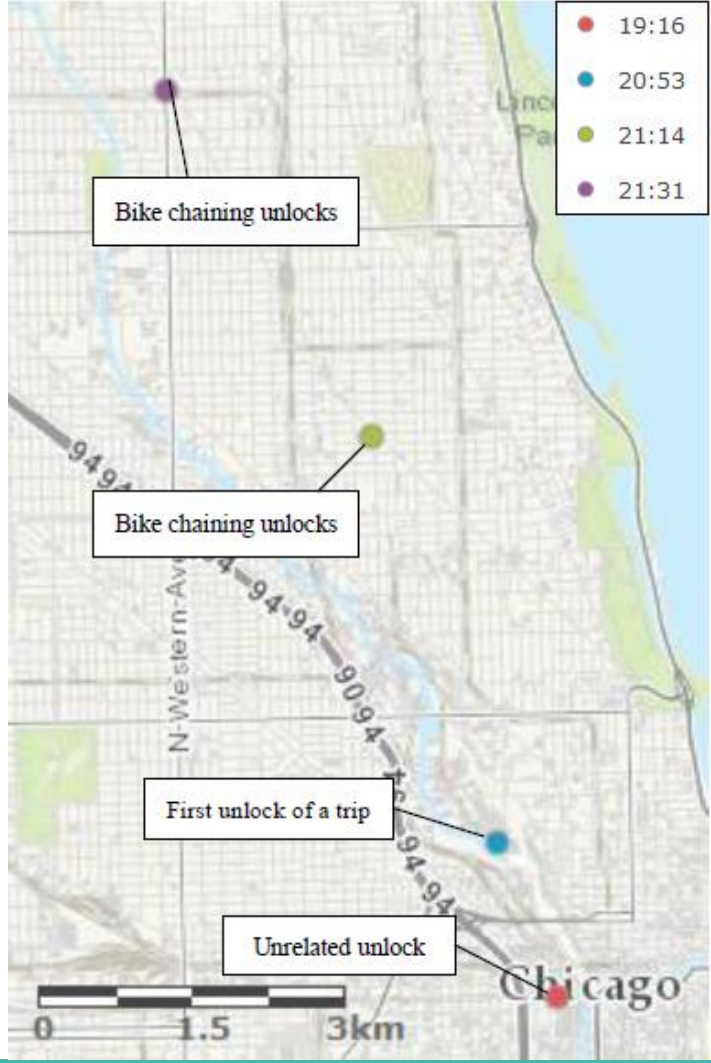
- **Commuters** rent bicycles to travel between home and workplaces, or between home and transit stations on weekdays. The cyclists usually use it during the rush hour (6 – 10 a.m. and 16 – 20 p.m.)
- **Utility users** use bicycles throughout the weekdays for shopping and errands.
- **Leisure users** generally ride at weekends for fun and exercise.
- **Tourist users** use bicycles to the beach, mountain or explore the city.

[1] O'Brien, O., Cheshire, J. and Batty, M., Mining Bicycle Sharing Data for Generating Insights into Sustainable Transport Systems. *Journal of Transportation Geography*, Vol. 34, 2014, pp. 262-273.

Introduction: Bikeshare usage patterns

2. Chaining trips: bike chaining unlocks

The **“bike chaining” unlocks** may occur when a user appears to return a bicycle within 30 minutes and immediately checks out another bicycle to continue the trip; it is likely that the people may avoid paying additional usage fees in this way because they will be charged when a bicycle is rented for more than 30 minutes.



Dataset: Bikeshare in Transit App

Transit & Bikeshare Info

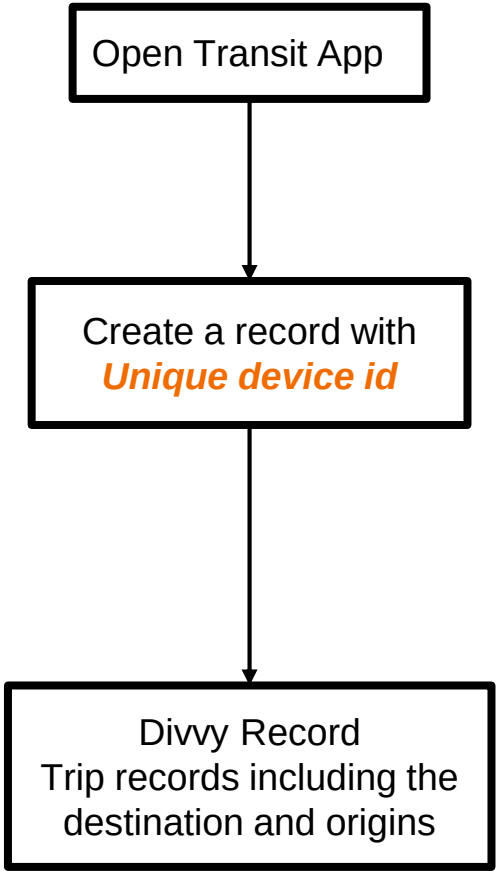
Create bikeshare account

Unlock code for bikeshare

Dataset: Unique Transit dataset

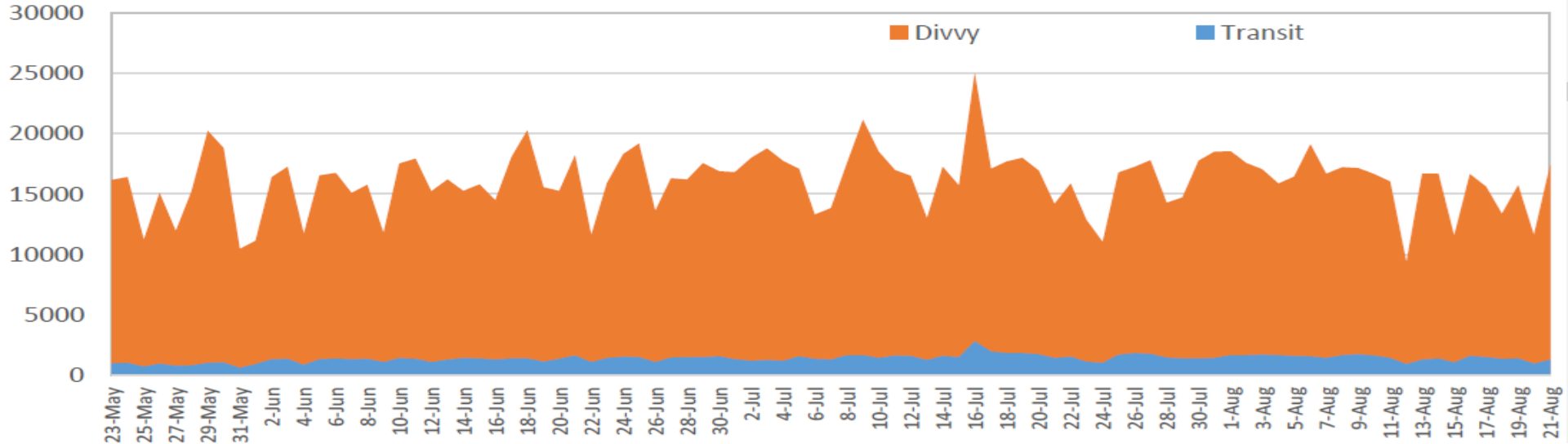
- **Transit App Data**
- **Device information**
- User (phone) location when unlocking a bicycle
- Bikeshare station location **(unlock only)**
- Timestamp

- **Divvy Data**
- Start & end locations of trips by station
- Start & end time & duration of trips
- User type (i.e., annual subscriber or 24hr pass customer), age & gender



Dataset: Area and time analysis

Three months data (May 23 - Aug 21 2016) in Chicago
The use of Transit app represents approximately **8.6%** of all bikesharing trips



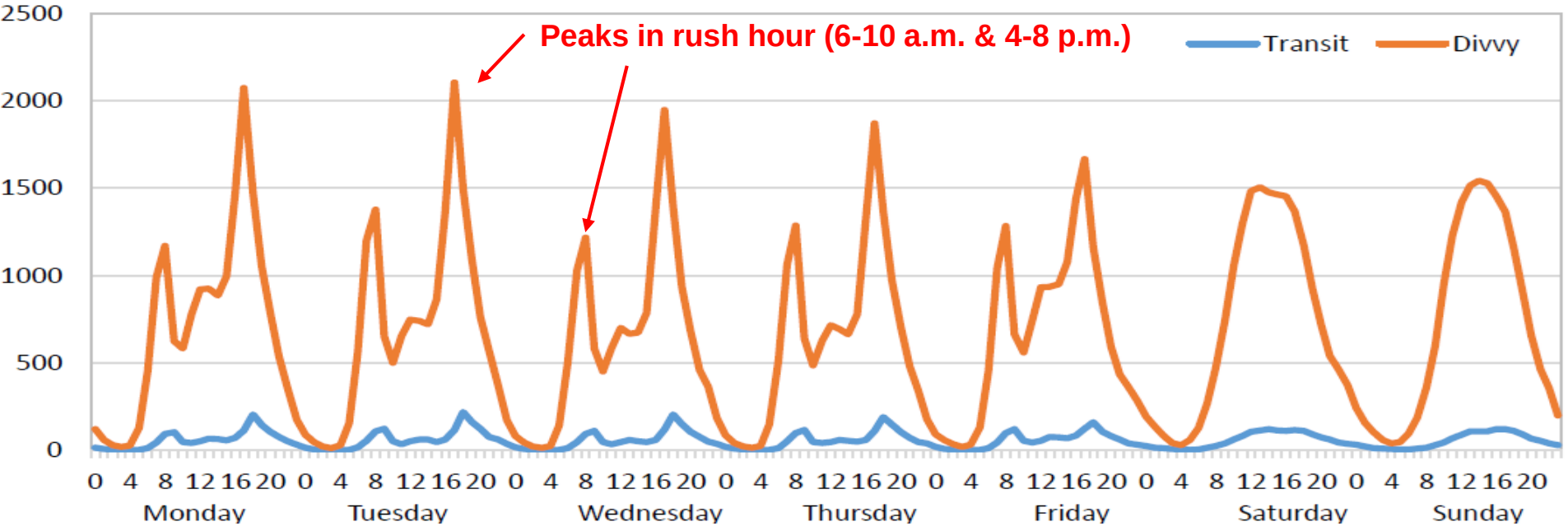
Number of unlocked bikes per day

Analysis

- Analysis 1: Exploratory analysis of system trends
- Analysis 2: Cluster analysis to identify user groups
- Analysis 3: Trip chaining algorithm to identify “trip chaining” unlocks

Analysis 1: Exploratory Analysis of System Trends

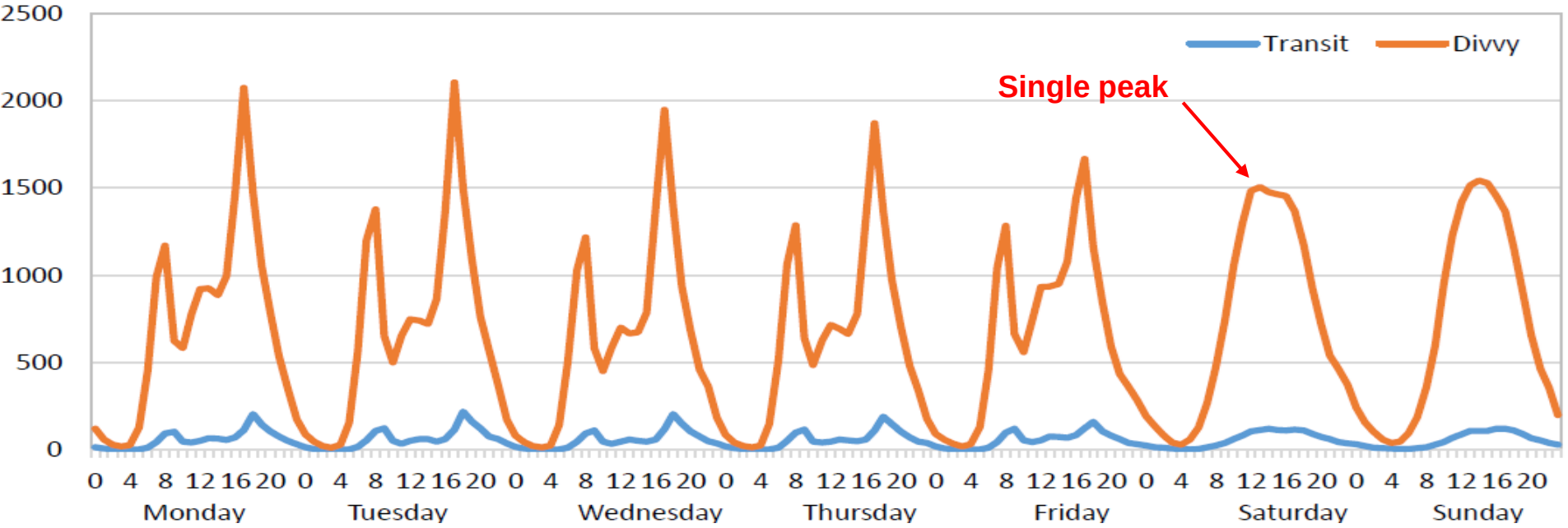
1. Weekday pattern: Commuting



Average number of unlocked bikes per hour in a week

Analysis 1: Exploratory Analysis of System Trends

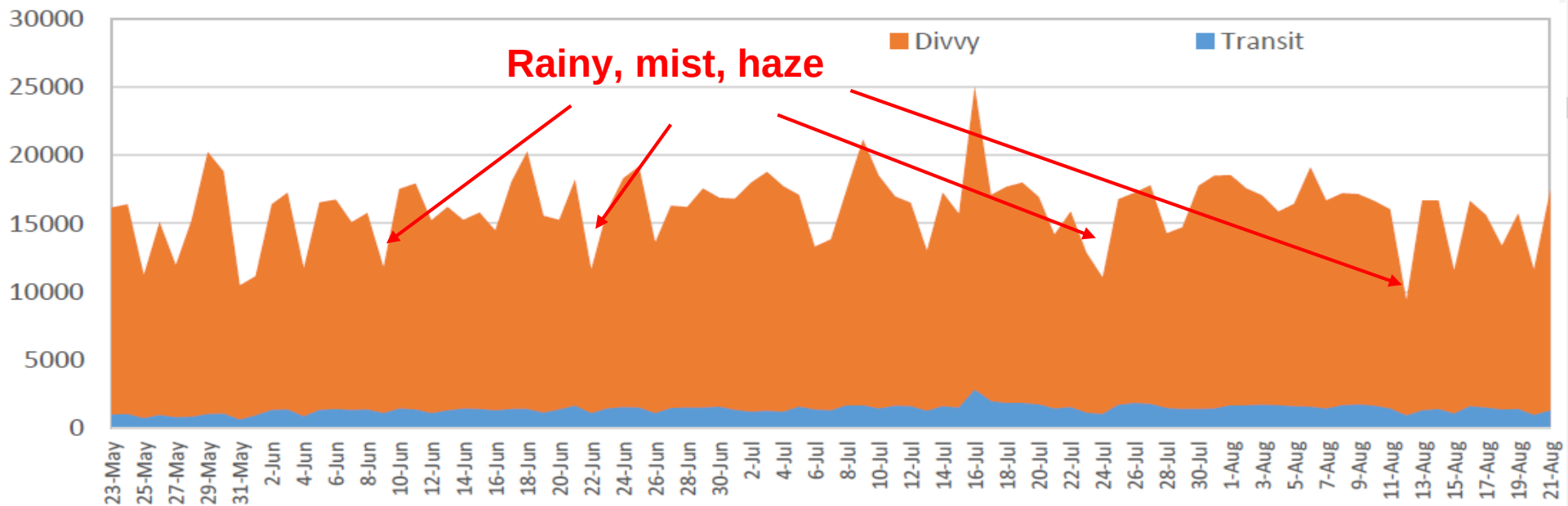
2. Weekend pattern: Tourism and leisure



Average number of unlocked bikes per hour in a week

Analysis 1: Exploratory Analysis of System Trends

3. Ridership corresponds to weather condition

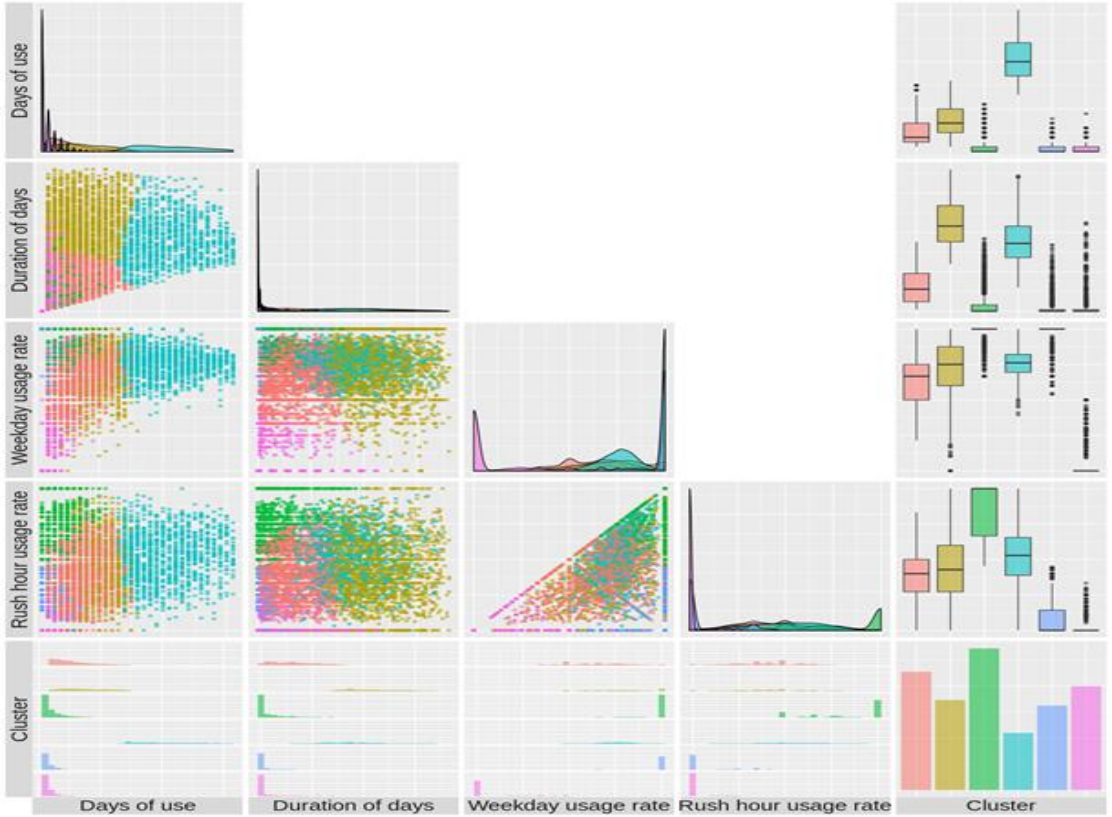


Number of unlocked bikes per day

Analysis 2: Cluster Analysis to Identify User Groups

We used **K-means classification** and the classification variables are as follows:

- Days of use
- Duration of days
- Weekday usage rate
- Rush hour usage rate

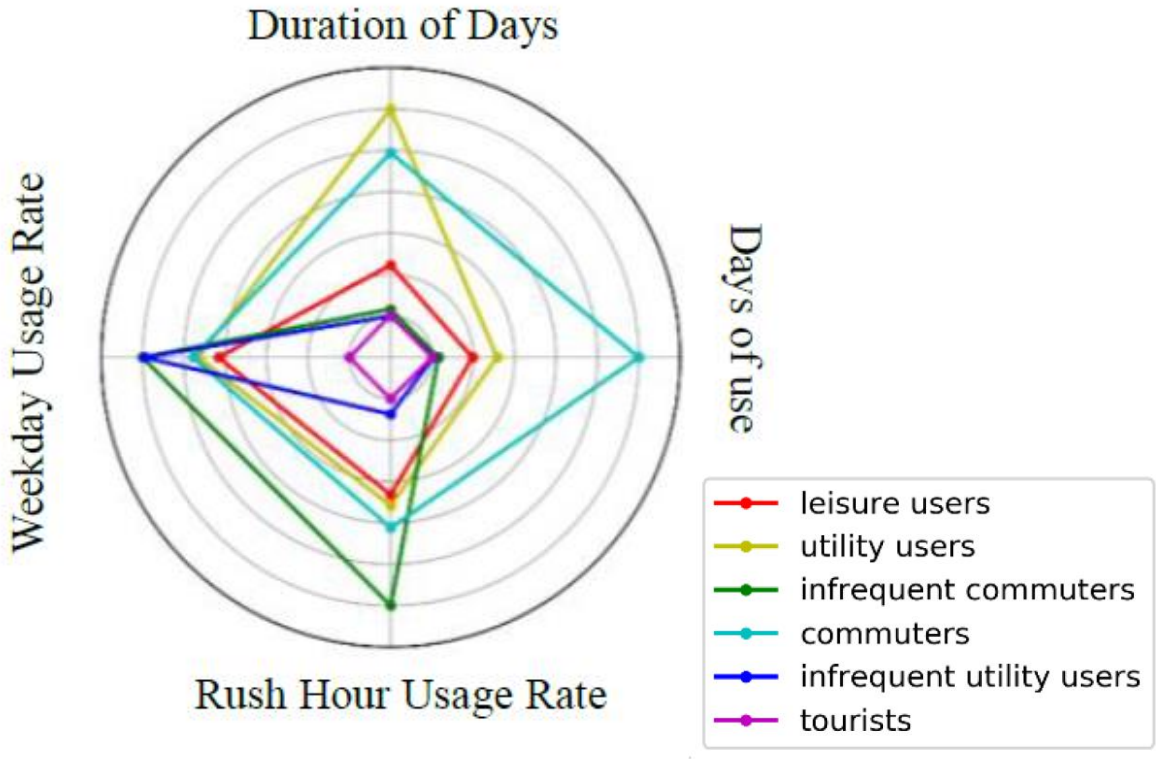


Results of K-means cluster method (using R)

Analysis 2: Cluster Analysis to Identify User Groups

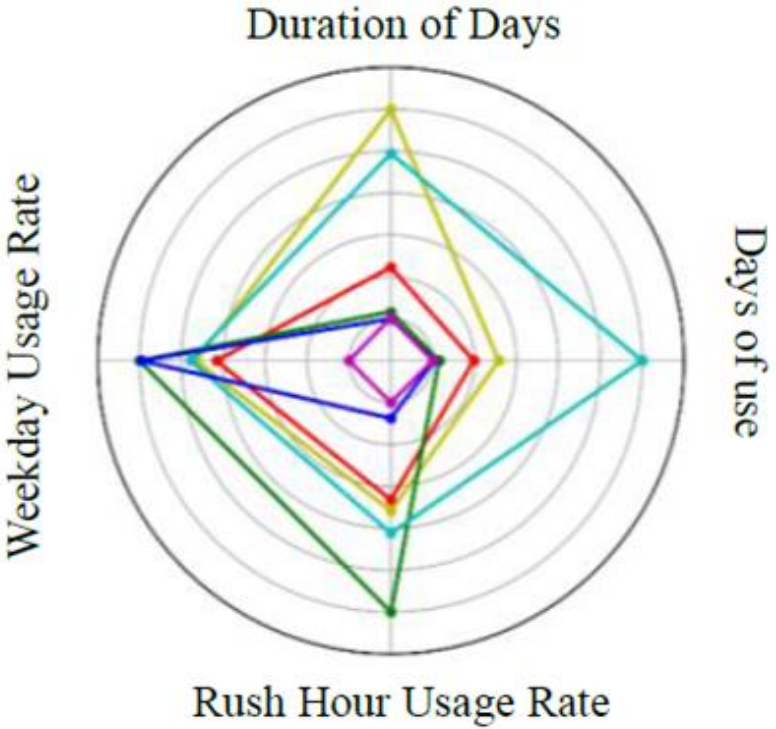
Mean days of use	Mean duration of days	Mean weekday usage rate	Mean rush hour usage rate
20.73	44.81	0.75	0.52
7.51	56.20	0.73	0.44
5.19	15.47	0.64	0.39
2.01	4.06	0.98	0.83
1.54	2.23	0.97	0.07
1.41	2.47	0.06	0.01
5.09	17.56	0.69	0.40

Results of K-means cluster method



Visualization of cluster results (using python)

Analysis 2: Cluster Analysis to Identify User Groups



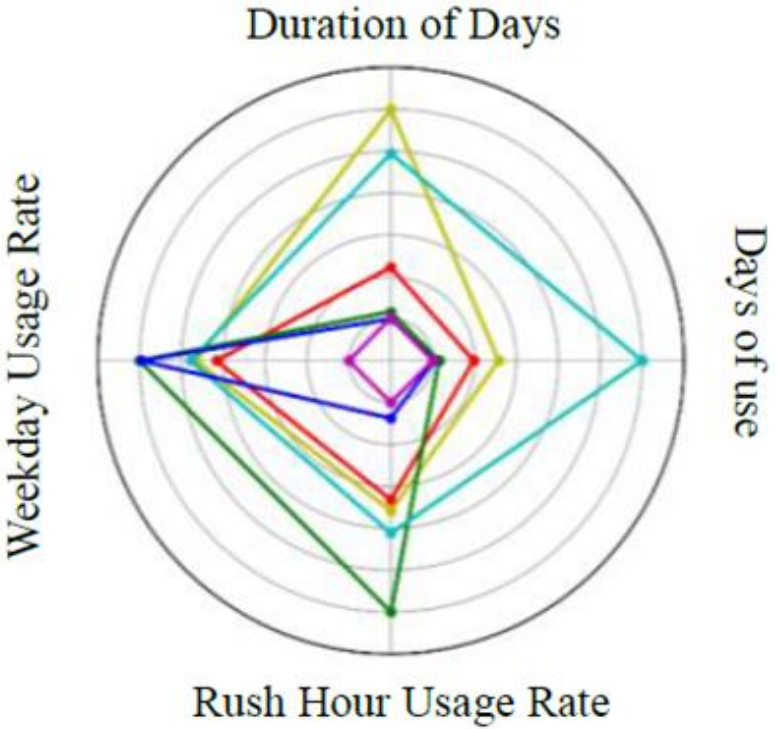
Commuters

- Many days of use
- Most in rush hour on weekdays
- Heavy users
- Potential Chicago residents

Tourists

- Cycle on weekends
- Low duration value
- Few days of use

Analysis 2: Cluster Analysis to Identify User Groups



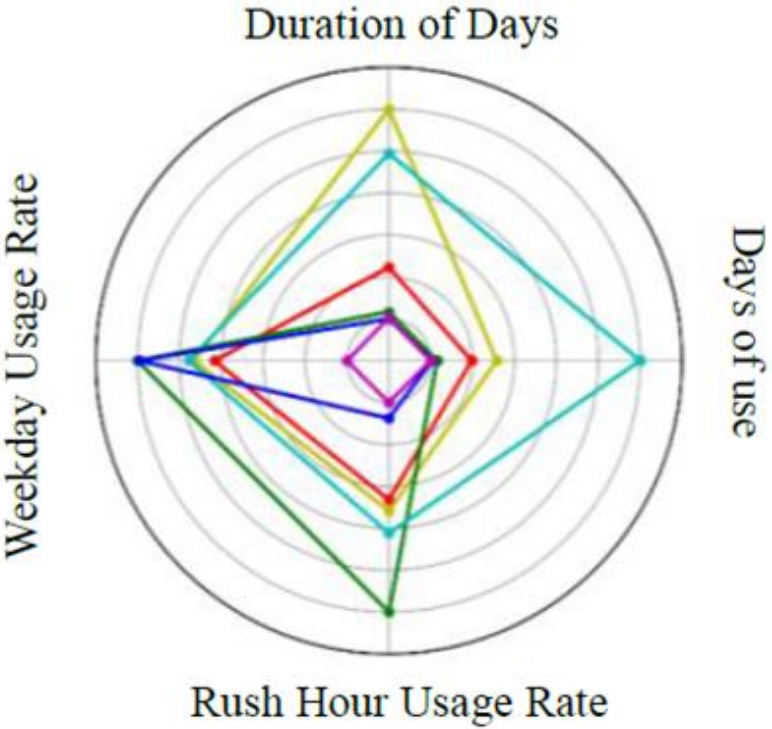
Leisure users

- Low rate of use on weekdays
- Potential Chicago residents
- May use bikeshare to go to the gym or parks

Utility users

- Cycle on weekdays
- Potential Chicago residents
- May cycle for errands or shopping

Analysis 2: Cluster Analysis to Identify User Groups



Infrequent Commuters

- Same pattern as “commuters”
- Less days of use
- Shorter duration value

Infrequent utility users

- Same pattern as “utility users”
- Less days of use
- Shorter duration value

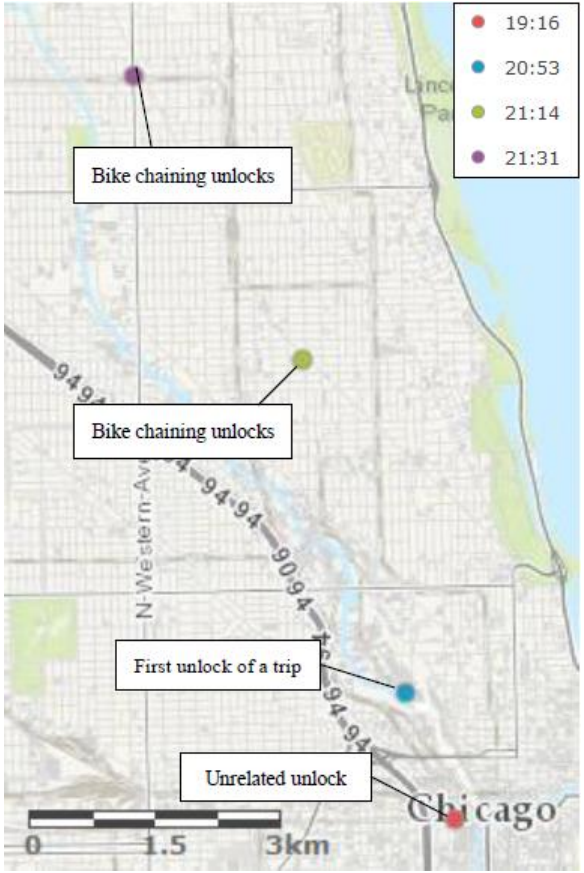
Analysis 2: Cluster Analysis to Identify User Groups

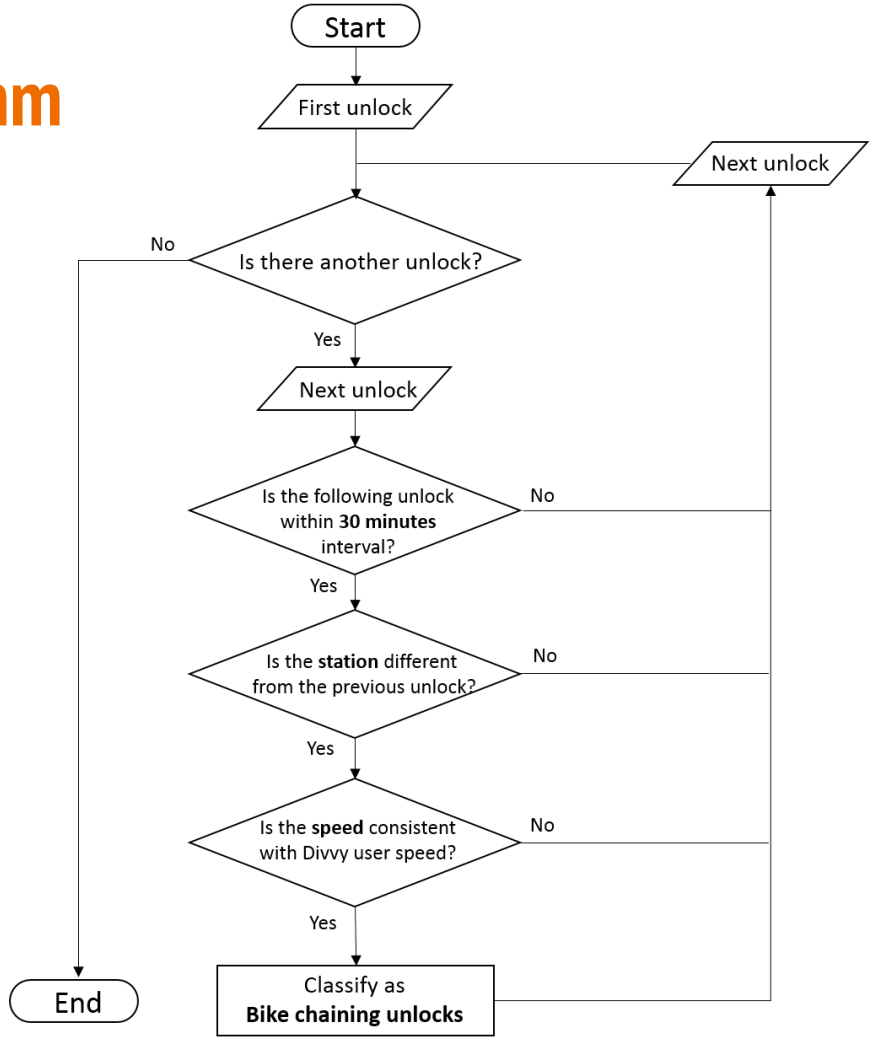
Category	Number of users (% of users)	Number of unlocks (% of unlocks)
Commuters	1094 (10%)	57,232 (45.58%)
Utility users	1729 (15%)	22,737 (18.1%)
Leisure users	2271 (20%)	23,749 (18.9%)
Infrequent commuters	2727 (24%)	10,267 (8.18%)
Infrequent utility users	1627 (14%)	5107 (4.06%)
Tourists	1998 (17%)	6478 (5.16%)
Total	11,446 (100%)	125,570 (100%)

Commuters account for only 10% of the whole users but they complete **45%** of all the unlocks.

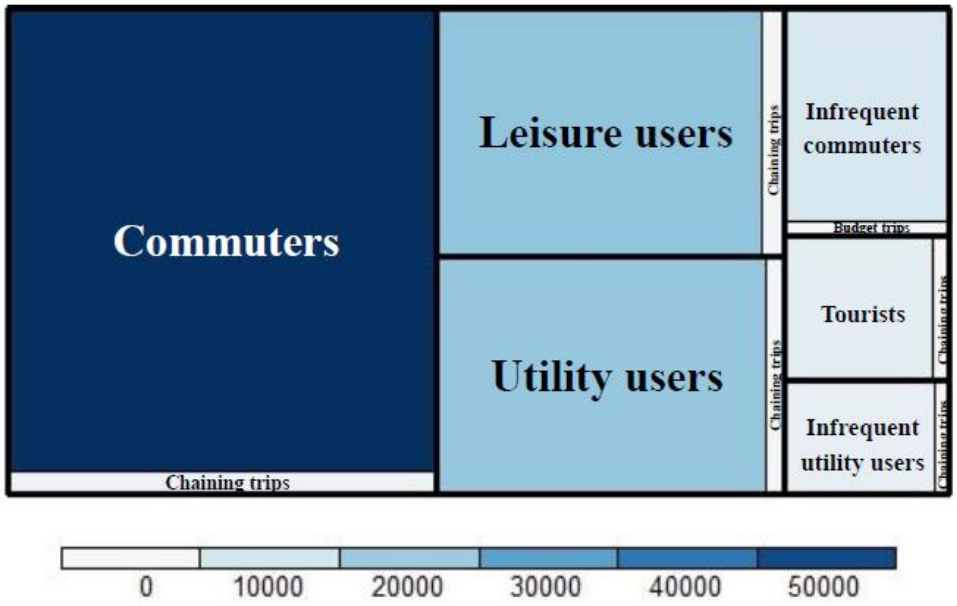
55% of the users purchased the 24-Hour pass.

Analysis 3: Trip Chaining Algorithm

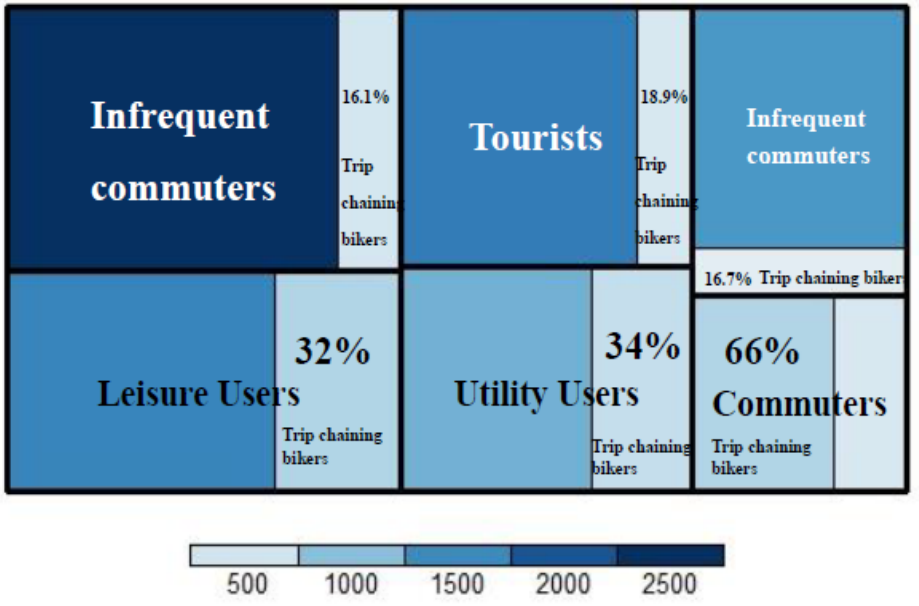




Analysis 3: Trip Chaining Algorithm to Identify Trip Chaining Bikers



Treemap of trip chaining unlocks in each groups



each groups

Treemap of trip

Summary

- Analysis from the system-level:
  - Weekdays: commuting patterns
  - Weekends: recreational usage
- Analysis from individual level:
  - 46% of Transit bikeshare unlocks are unlocked by commuters, but the commuters represents only 10% of Transit bikeshare users
  - 27.3% of Transit app bikeshare users made **“bike chaining”** unlocks to avoid paying additional usage fees
  - 66% of Transit app bikeshare commuters are identified as **“trip chaining bikers”**

Conclusion

- Three analysis have been conducted to address different problems: Home and work inferences of users; relationship between uber, taxi users; bikeshare usage patterns.
- With the big data, we are able to illuminate social processes that were previously undersampled or poorly understood.