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.

Table Names	Description of Contents
Device	Contains a Transit app specific identification number (device ID), device type, model of device, operating system, version of Transit app, and last date of app use
Favorite	Provides information on user designated favorites in terms of transit routes
Feed Download	Provides a summary of activity on the Transit app by day
Installed App	Reports on other installed apps on the user’s device that can impact functionality, such as the Uber app
Location	Includes the location (lat/long) and a unique session ID for each time the app is opened
Nearby View	Contains information about the transit routes presented to a user in each session upon opening the app
Placemark	Includes location data from an optional function that stores places users often go (e.g. home or work)
Session Complete	Provides an event based view of each session, including the beginning and ending location
Sharing System Actions	Provides data on the booking of carshare, bikeshare, and other services, including the location of shared vehicles
Sharing System Purchase	Provides purchase records for shared vehicles, which are primarily bikeshare passes
Trip	Contains information about usage of the trip planning feature, including start and end coordinates (lat/long)
Uber Request	Lists requests for service from Uber, which are then handed off to Uber’s app for fulfillment
User Feed Session	Includes the number of times the app is opened and the different transit agency’s data accessed by the user

Transit Data Utilization: Home/Work Inferences of Users

Danielle Stacy
The University of Alabama

Joint Institute for Computational Sciences

Question

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

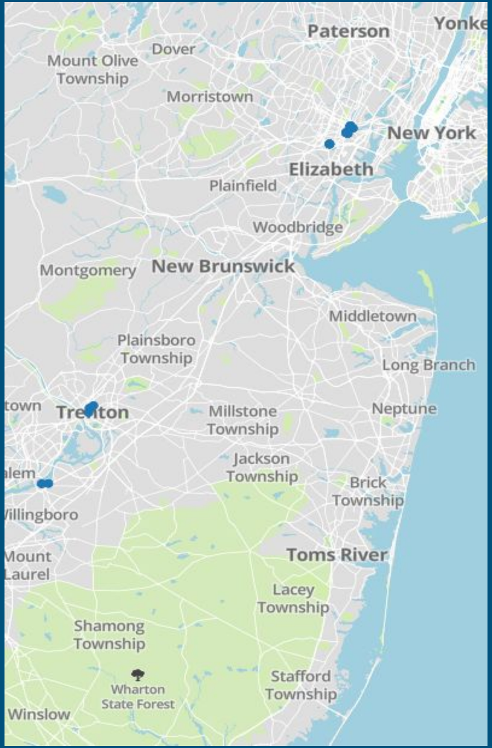
In theory, a user would check the app in the morning at home to check the quickest way to work and then check again in the evening from work to search for the quickest way home. The goal is to check this assumption using the data provided from the app and determine if it is valid to infer home and work locations.

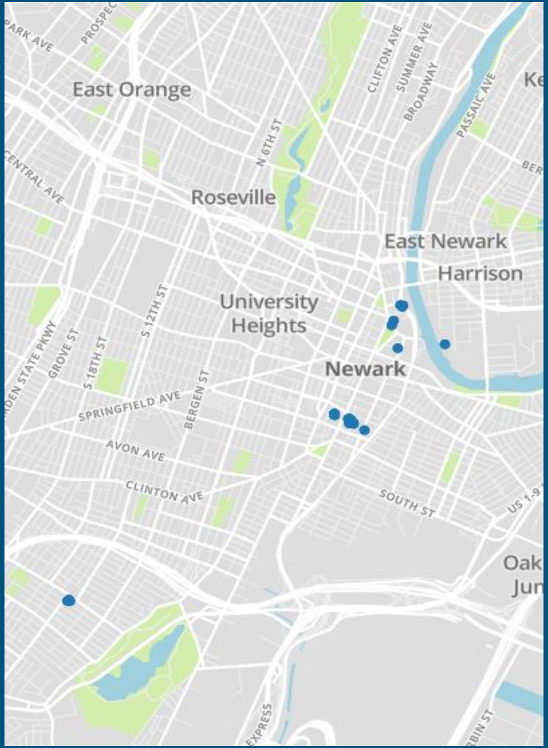
The Plan

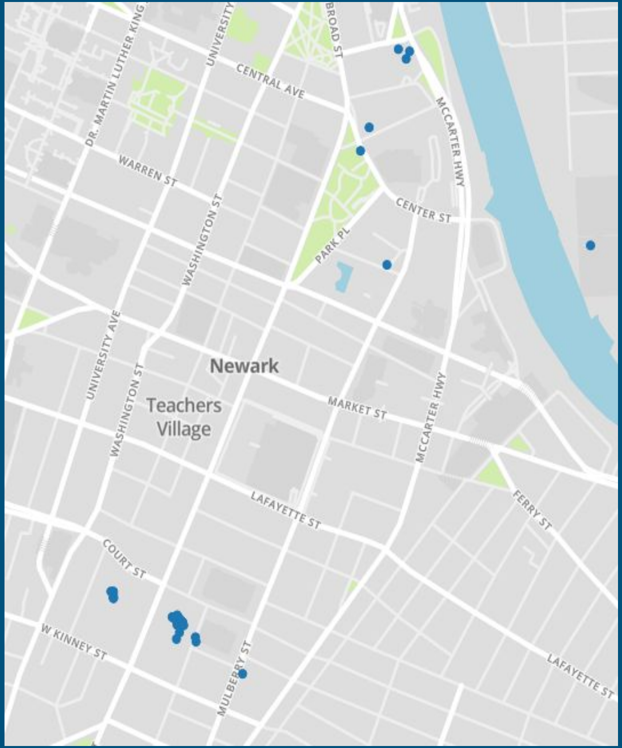
A unique identifier has been assigned to every user, so it is simple to keep track of a specific user across multiple tables. To check a user's location in the morning before work and in the evening after work, I will use the session complete table that provides a timestamp and location coordinates of the user when they opened the app. Specifically, I will check the location of users at 6-9 AM and at 3-7 PM. If there is clustering at specific locations at these times for a user, I would designated those locations as the user's home and work locations respectively.

To validate my chosen locations of the user's home and work, I will make use of the placemark table. From this 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.

Clustering Example







Coordinates have been shifted to protect user identity

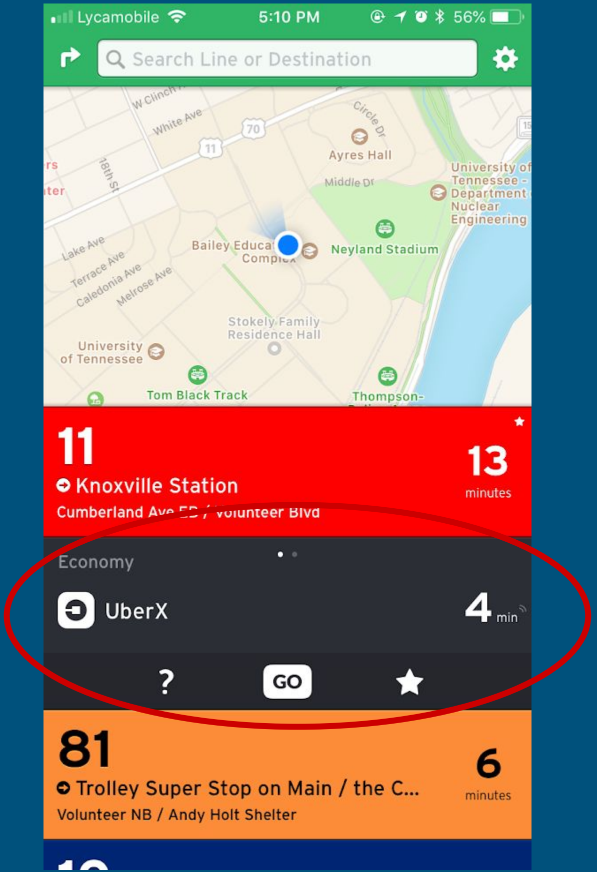
Transit data utilization: Analysis of Uber requests

Jing Guo
Changsha University of Science and Technology

Background

Datasets : Uber request from Transit
(CSV Format) New York central park weather
Uber trip origins in NYC
TLC Trip Record Data (Taxi,FHV)

	Transit users	uber requests from Transit
Amount	17,000,000 per day	2,271 per day



Question

Analysis of Uber request

- What is the difference between the Transit Uber users and Uber app users?
- How public transportation and Uber influence each other ?
- How weather influence Uber?

…….

Plan

Data overview:

- Uber requests data overview
- Uber trip origins data overview

Comparing two datasets

Analysis and Clustering

- New York City

Conclusion

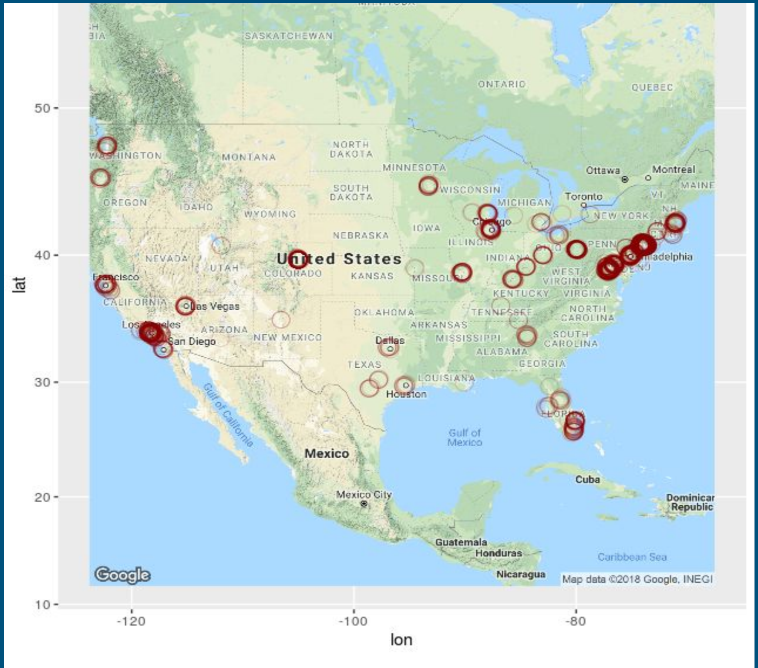
Data Overview

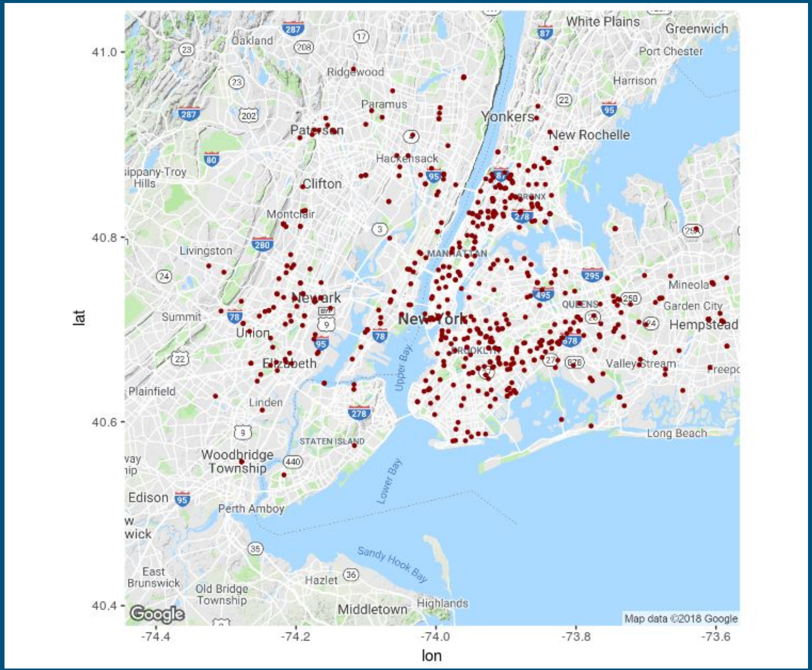
---Uber requests from Transit

- User location distribution
- Uber app installed
- Types of the Uber
- Uber request trends over time
- Weather influence on uber requests

Data Overview

- Distribution (U.S and New York)

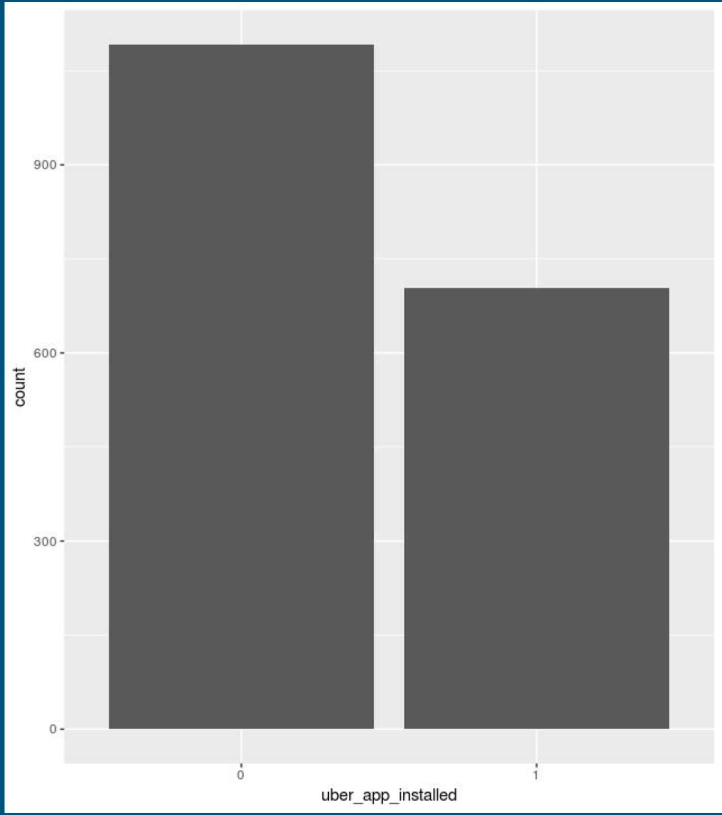




Data Overview

- Uber app installed

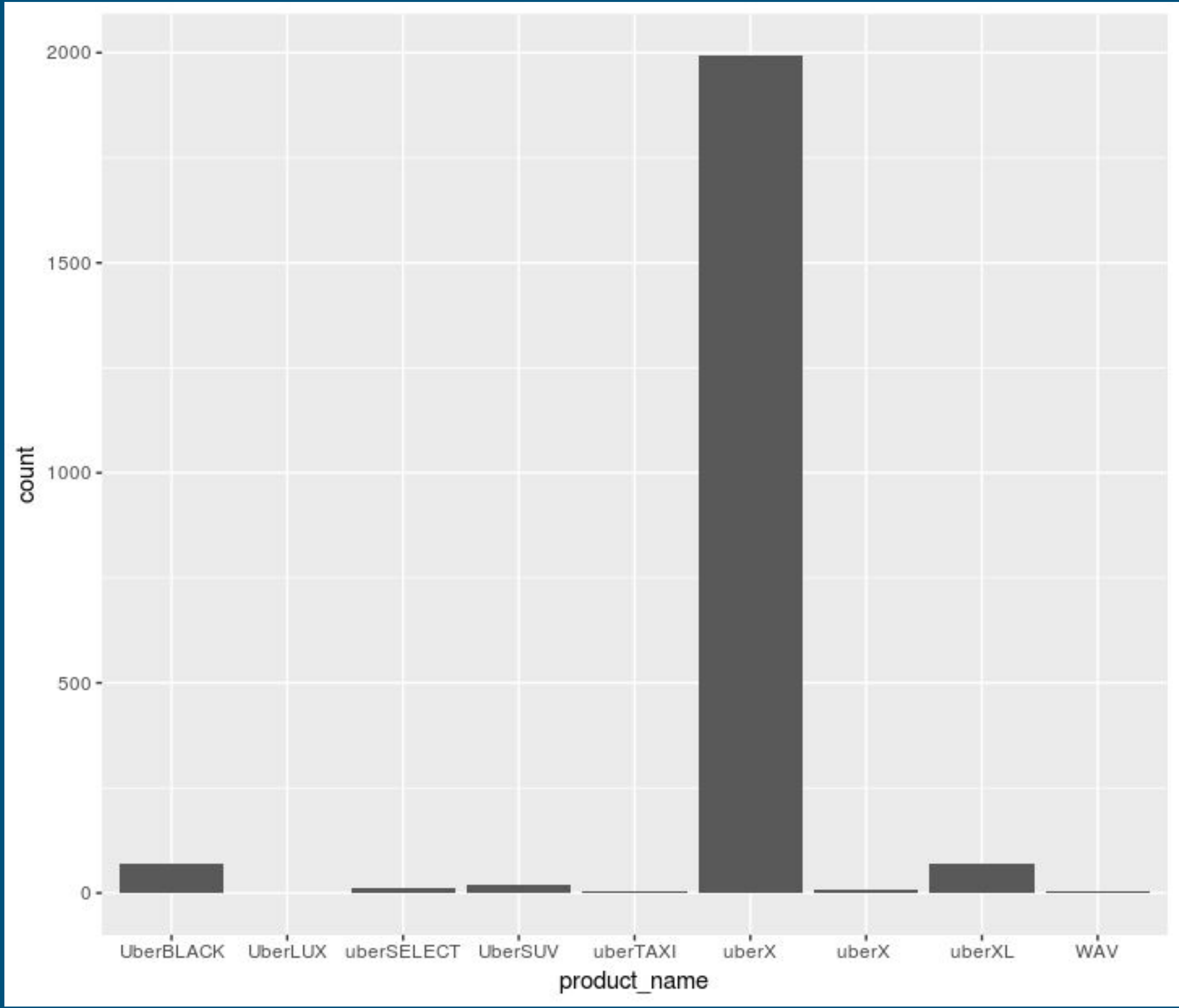
about 60%Transit requests users didn’t installed the app.



Data Overview

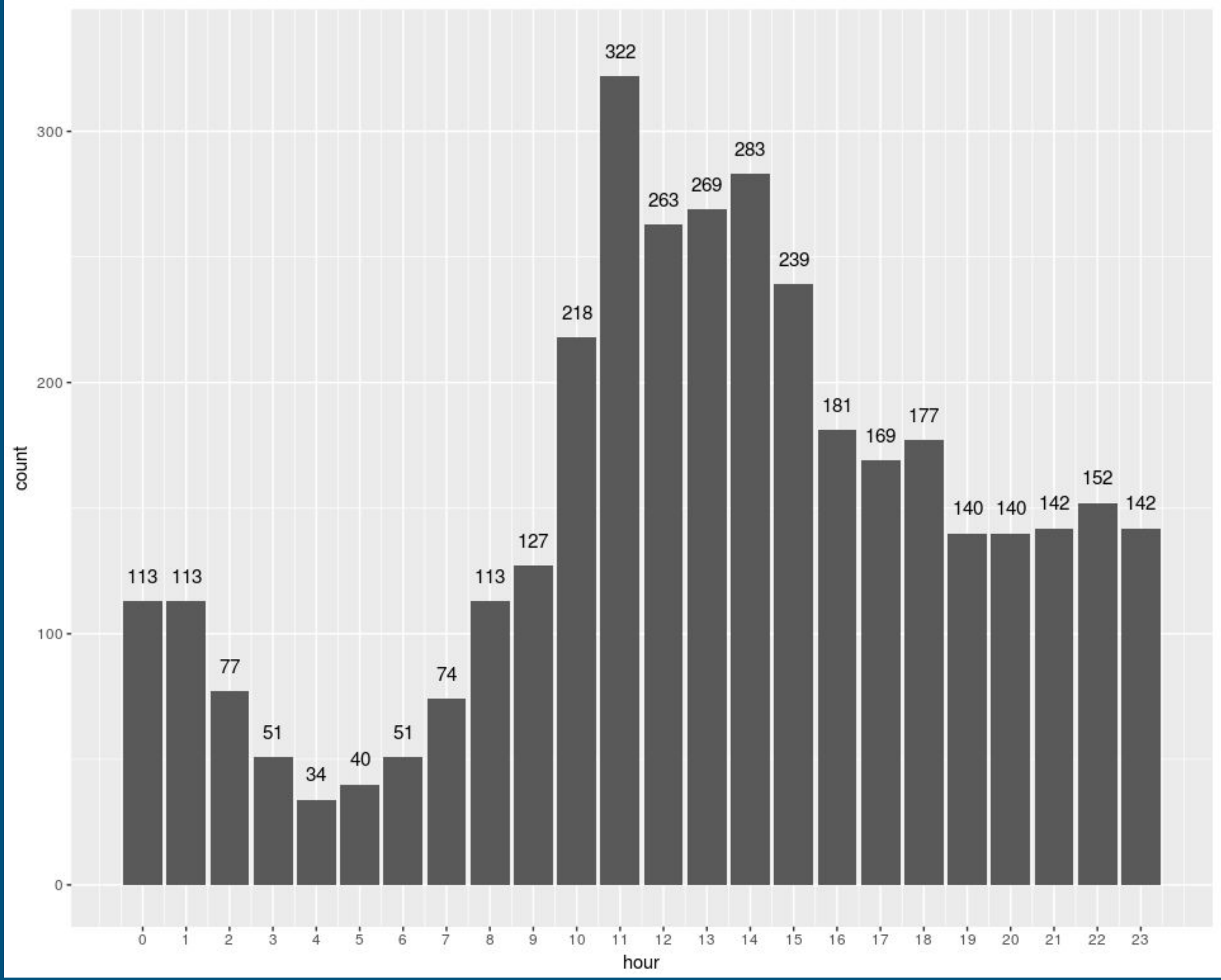
- Types of the Uber

Uberx: 91%



Data Overview

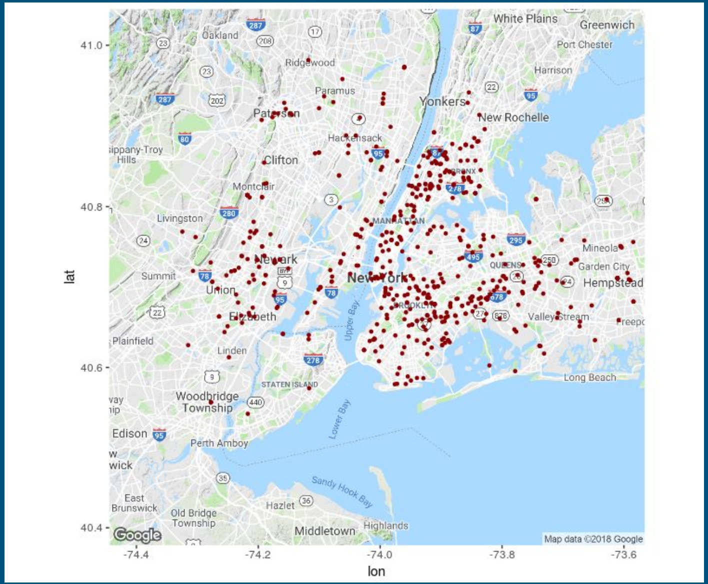
- Uber request trends over time



Future work

Uber and Public Transport in New York City

- Comparing
- Clustering



Transit data utilization: A statistical analysis of usage patterns of bike-sharing service

Siyue Yang
City University of Hong Kong

Outline

- Background
  - Transit data and Divvy data
  - Related work of bike-sharing system
- Research question
- Research Plan
  - Data Overview
  - Comparison between two datasets
  - Clustering on user types
  - Prediction by weather condition
- Progress

Background

	Transit dataset	Divvy dataset
Composition	Information of different bike-share orders on Transit	Information of **all** of the Divvy bike orders
Source	Transit developers	Divvy company website
Amount	101,835 per year	3,595,383 per year
Remark	Devices	Stations, Membership

Usage pattern 1:
Difference between Transit App users and the whole group of users

Usage pattern 2:
Characteristics of different types of users

Background

- Service schedule for bike reallocation
  - trip destination and duration prediction model
  - bike trip demand prediction
  - trip route planning problem for individuals
- Bike flows prediction
  - a hierarchical prediction model predict the bike flows that will be rent from/returned to each station
  - a model-based clustering algorithm to classify bike stations for efficient management

Usage pattern 3:
Factors of
bike-sharing systems

Research Question

Transit data utilization:
A statistical analysis of usage patterns of bike-sharing service

↓

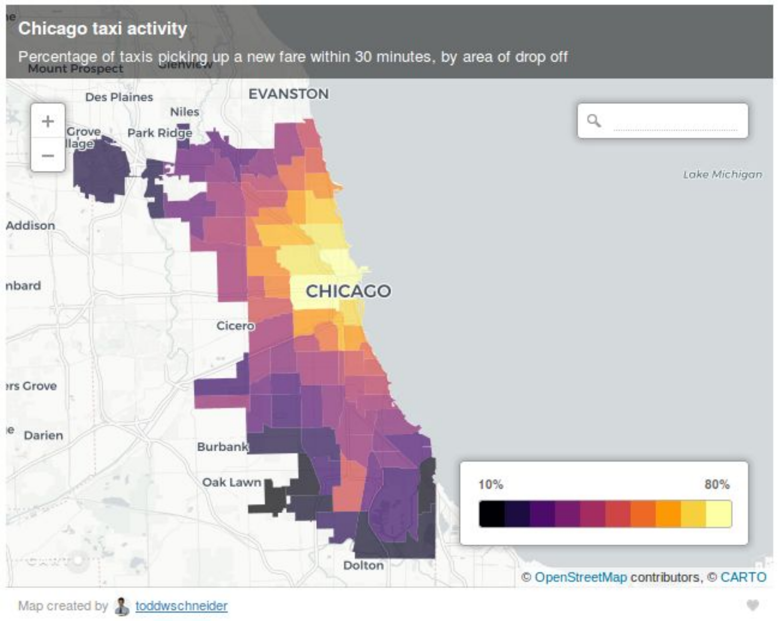
Transit user characteristics + Clustering on user types + Predicting the bike flow

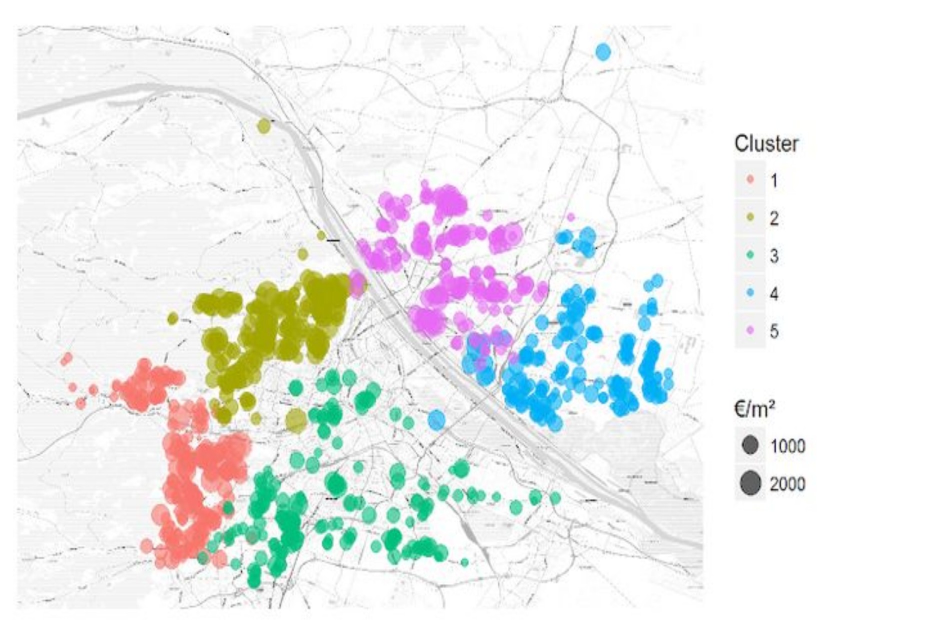
Research Plan

- Data Overview
  - Distribution maps
  - Trends over time graphs
  - Variation between labels
- Comparison between two datasets
- Clustering on user types
  - K-means clustering
  - Analytic Hierarchy Process
  - Latent Subspace Clustering based on deep neural network
- Prediction by weather condition

Data Overview

- Distribution Maps

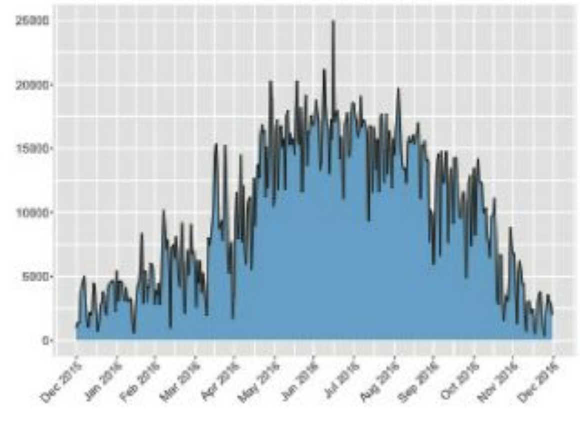




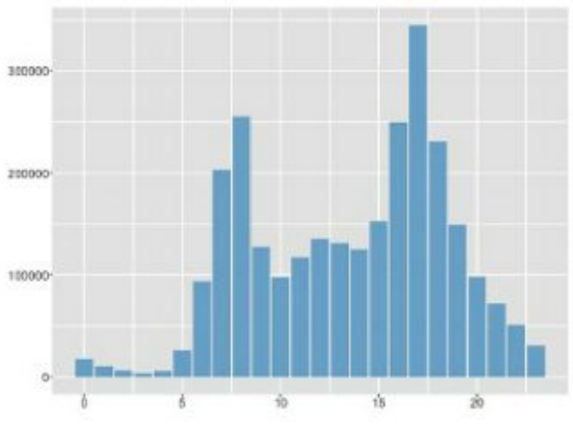
Source: http://www.bernhardlearns.com/2017/10/finding-viennas-house-price-regions.html

Data Overview

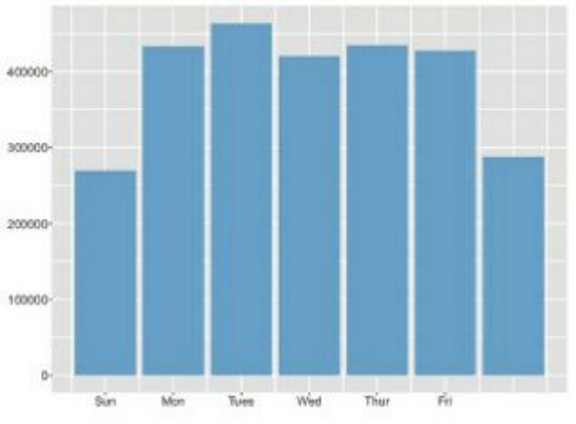
- Trends over time



Number of bikes in a year



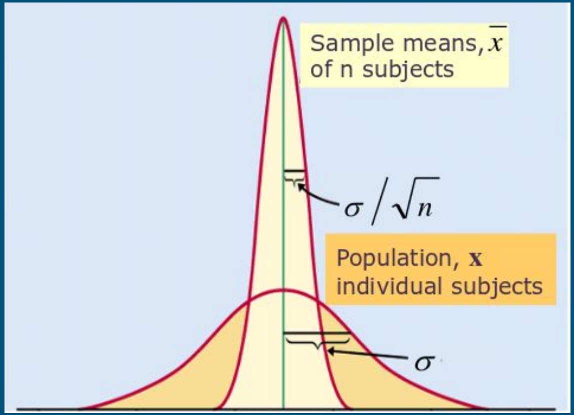
Number of bikes in a month



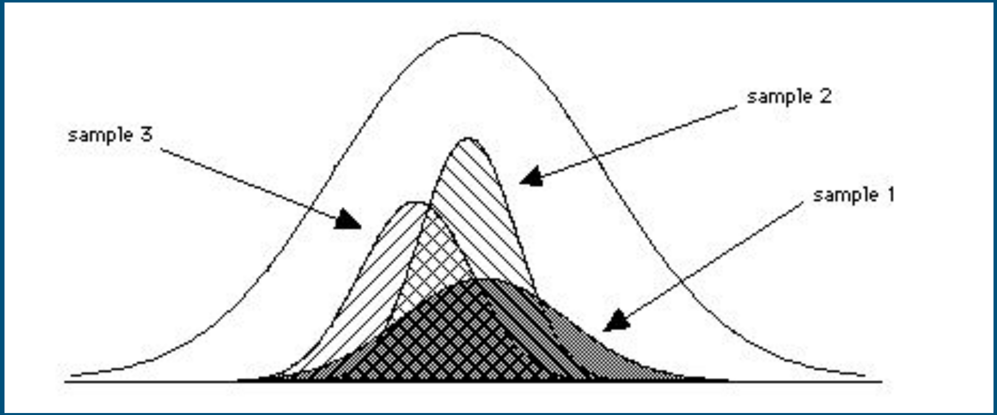
Number of bikes in a week

Comparison

Goal: Transit dataset ~ Divvy dataset (representative sample ~ population)



representative sample

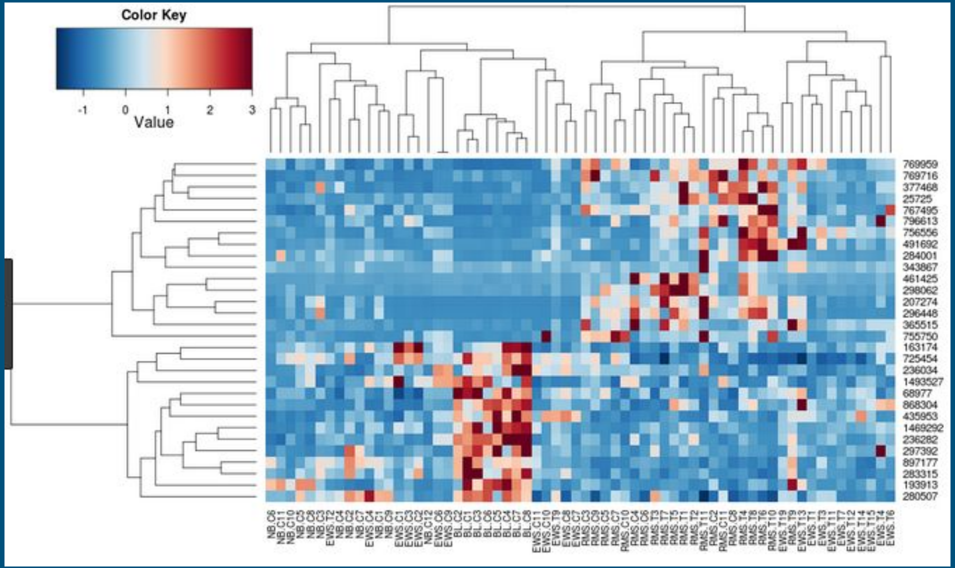


Source: http://psychology.illinoisstate.edu/jccutti/psych240/chpt7.html

Not representative sample

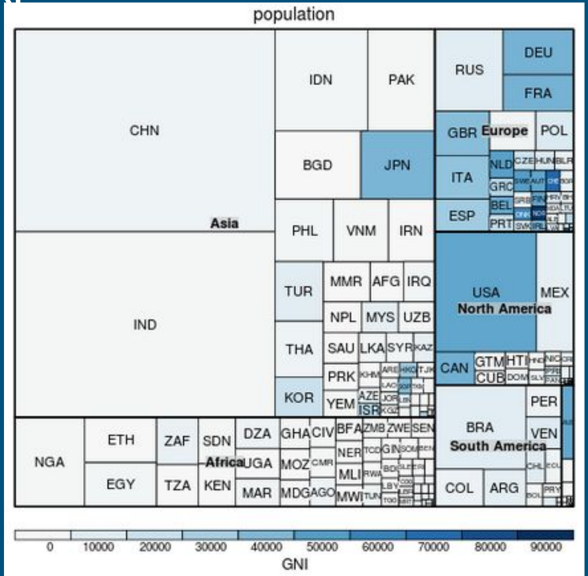
Clustering

- Analytic Hierarchy Process



Source:
http://www.sthda.com/english/articles/28-hierarchical-clustering-essentials/93-heatmap-static-and-interactive-absolute-guide/

- Latent Subspace Clustering by DNN



Source:
https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=2ahUKEwjv04b8zPHbAhXSzVMKHTnwAVgQjxx6BAgBEAI&url=https%3A%2F%2Fwww.stat.auckland.ac.nz%2F~paul%2FReports%2FVoronoiTreemap%2FvoronoiTreeMap.html&psig=AOvVaw2o25lpjSebE3skBYOO9huV&ust=1530111700505776

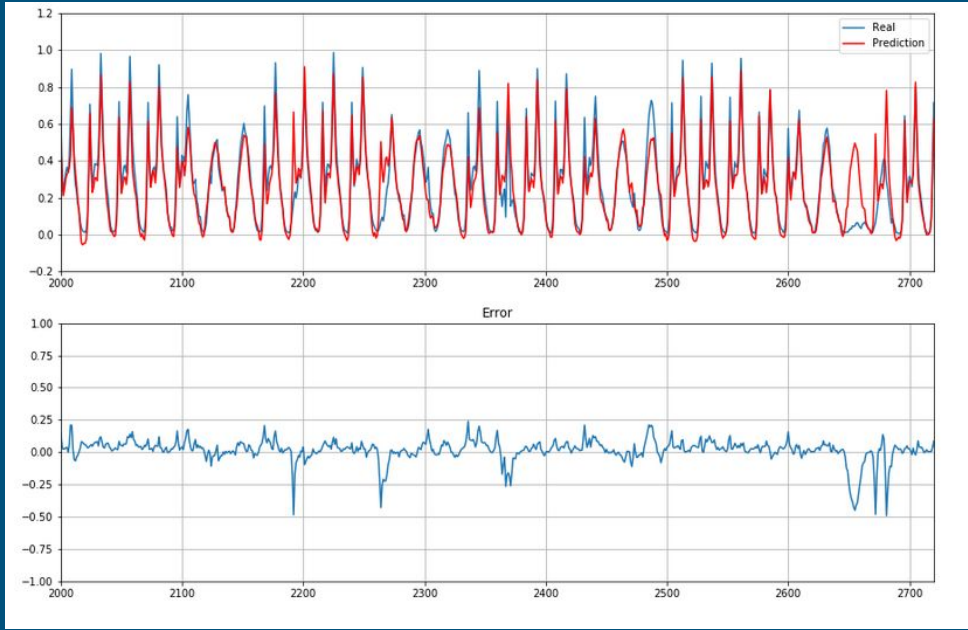
Prediction

Linear regression:

Number of bikes per minute ~ temperature + pressure + wind speed + humidity

Methods:

- Support Vector Regression (SVR) with RBF kernel
- Ridge regression
- Neural network regression



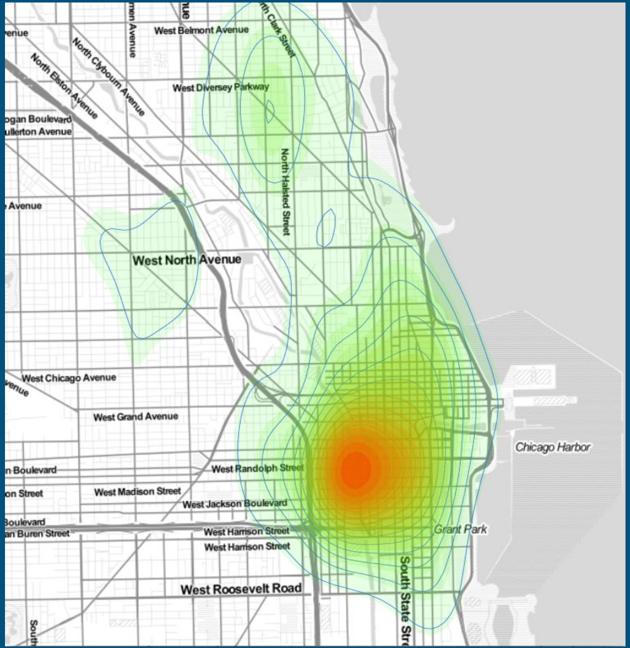
Source: https://www.kaggle.com/marioroos/explore-data-predict-future-trips/data

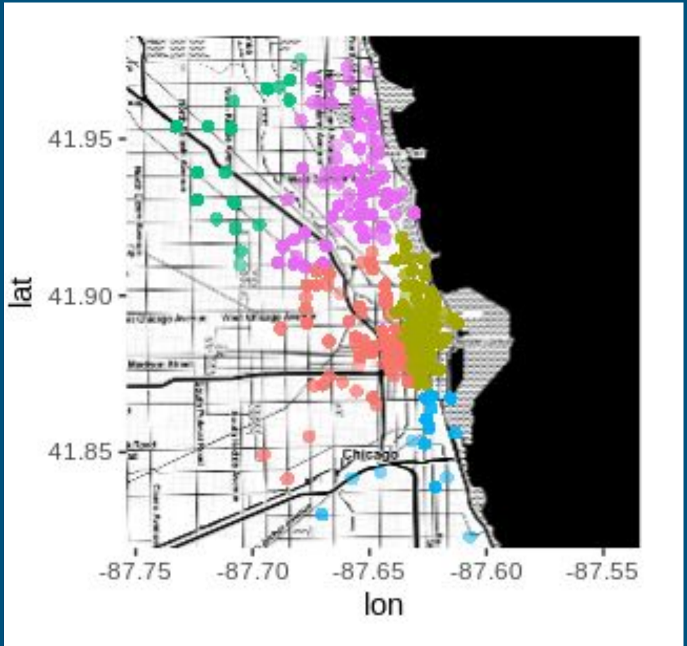
Progress

- Data Overview
    - Distribution maps
    - Trends over time graphs
    - Variation between labels
- Comparison between two datasets
- Clustering on user types
    - K-means Clustering
    - Analytic Hierarchy Process
    - Latent Subspace Clustering based on deep neural network
- Prediction by weather condition

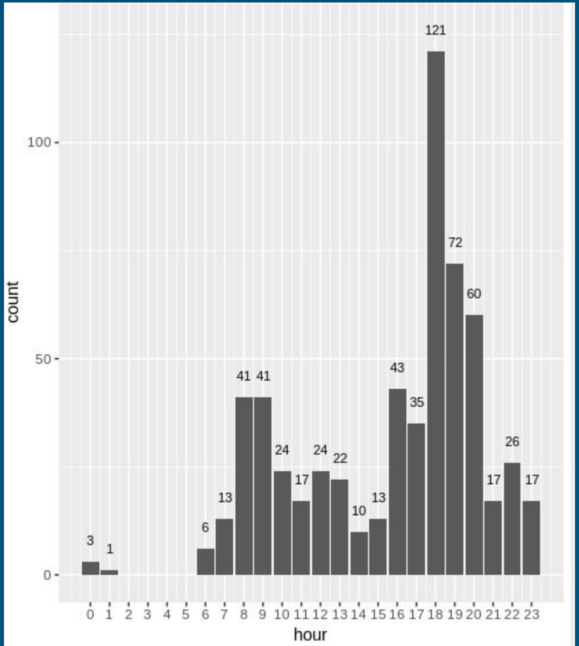
Data Overview

- Distribution Maps





- Trends over time

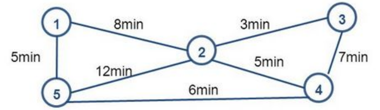


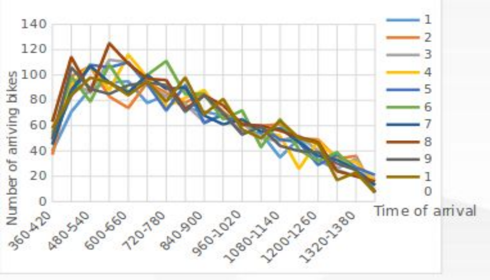
Clustering

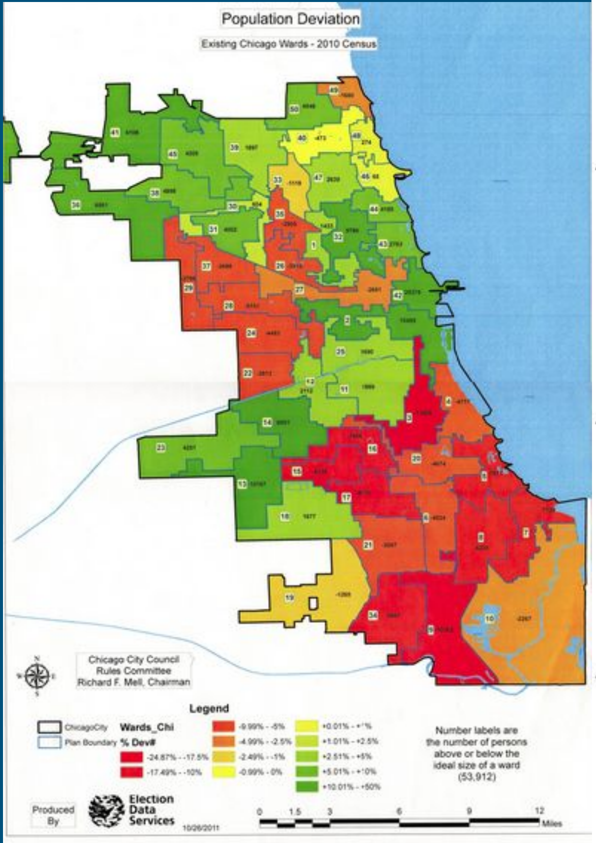
- K-means clustering
- Origin-destination Matrix

Trips Between Zones

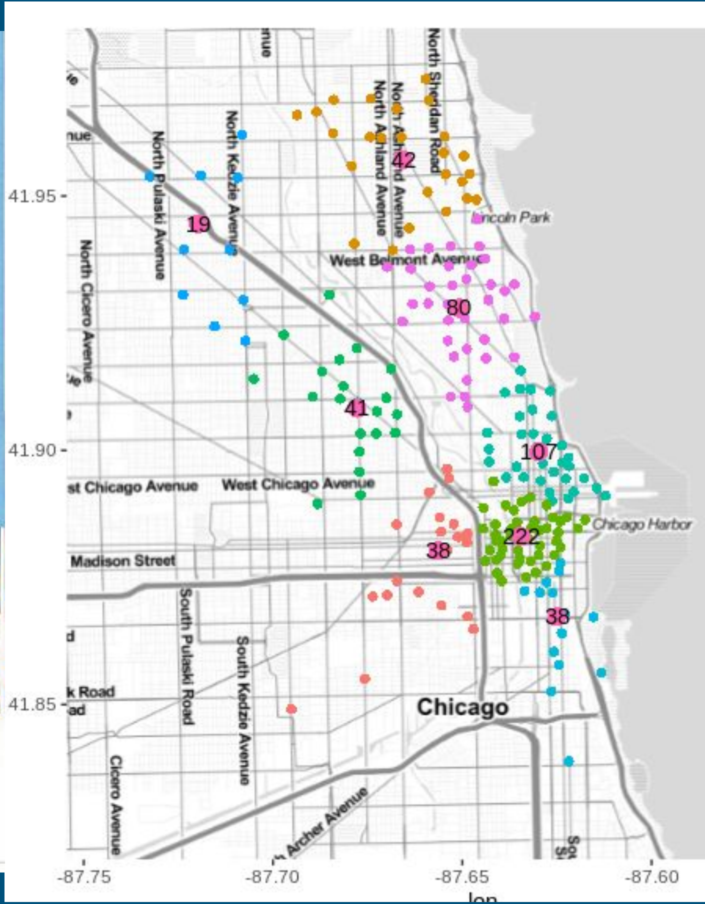
From/To	1	2	3	4	5
1	0	100	100	200	150
2	400	0	200	100	500
3	200	100	0	100	150
4	250	150	300	0	400
5	200	100	5	350	0





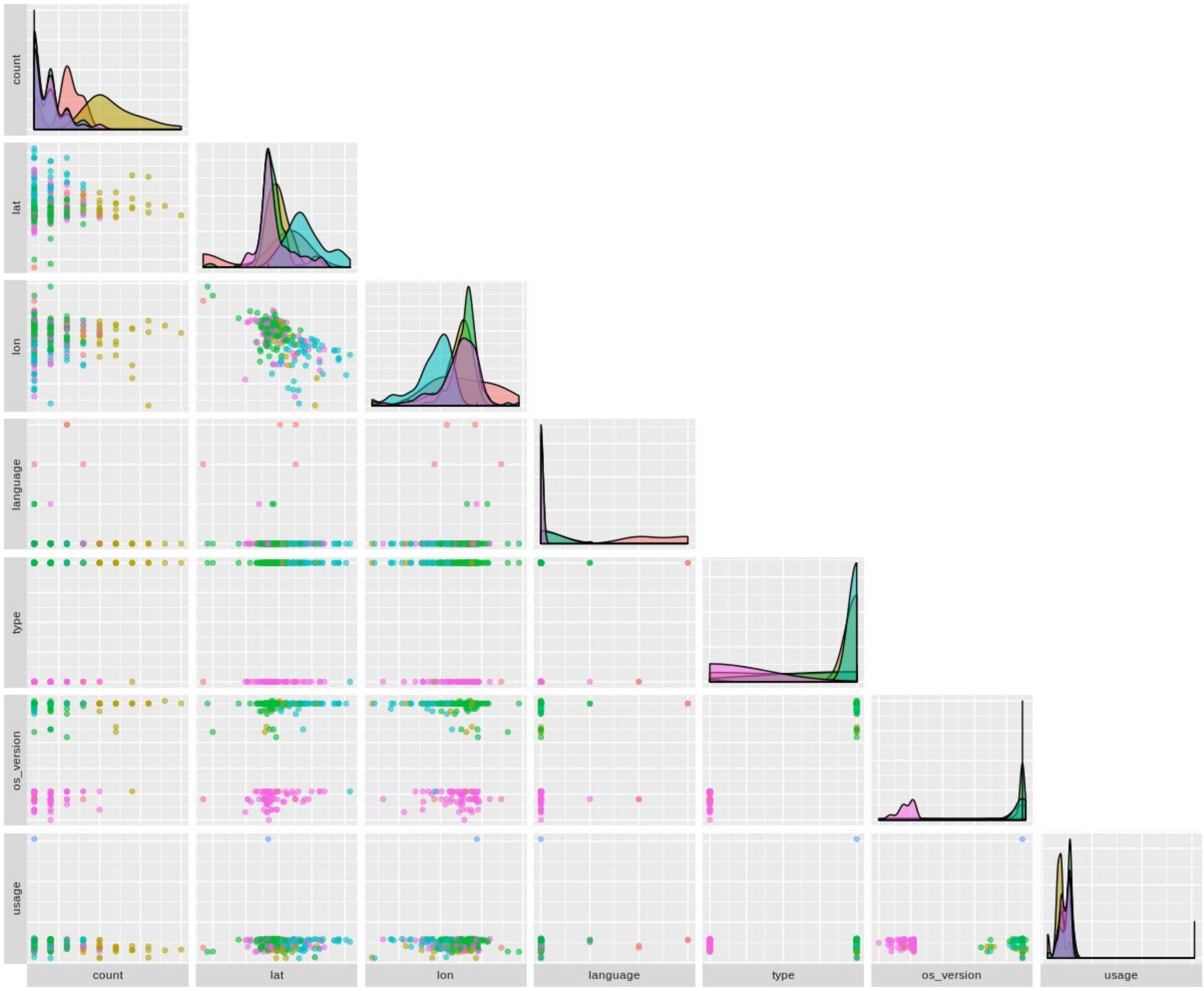


Source: http://ward32.org/news/population-deviation-map-jpeg/



Clustering

- K-means clustering on user types, classifying the users into 6 subgroups



Conclusion

Q & A

Reference

Chang X, Shen J, Lu X, Huang S (2018) Statistical patterns of human mobility in emerging Bicycle Sharing Systems. PLoS ONE 13(3): e0193795. https://doi.org/10.1371/journal.pone.0193795