Citi Bike NYC

By Samantha Lin, Yuritzy Ramos, Matthew Joesoep, & Ruilin Chen

TABLE OF CONTENTS

01

Introduction of the Problem

Background

- Citi Bike is a bikeshare program with 750+ stations across NYC
- Unlock from one station and return to any other station.
- Payment options:
  - Single Rides (30 mins)
  - Day pass (24 hrs)
  - Annual Membership

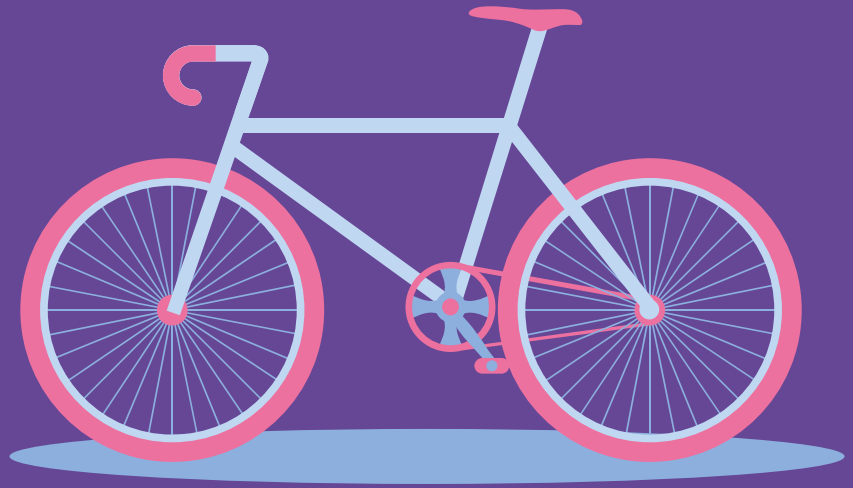
citi bike powered by lyft

FEATURES OF THE DATASET

TRIP DATA

- Trip Duration (seconds)
- Start/Stop Time and Date (NYC local time)
- **Start/End Station Name and ID**
- Start/End Station Lat/Long
- Bike ID
- User Type (Customer = single ride or day pass user; Subscriber = Annual Member)
- Year of Birth of user
- Gender (0=unknown; 1=male; 2=female)

- *16 features*
- *Randomly sampled 0.01 of 2013-2017 period data with about 470,000 rows*



Problem

Problem 1 – Link Analysis

Determine the significance of stations in the network to identify which stations should be retained, expanded or removed.

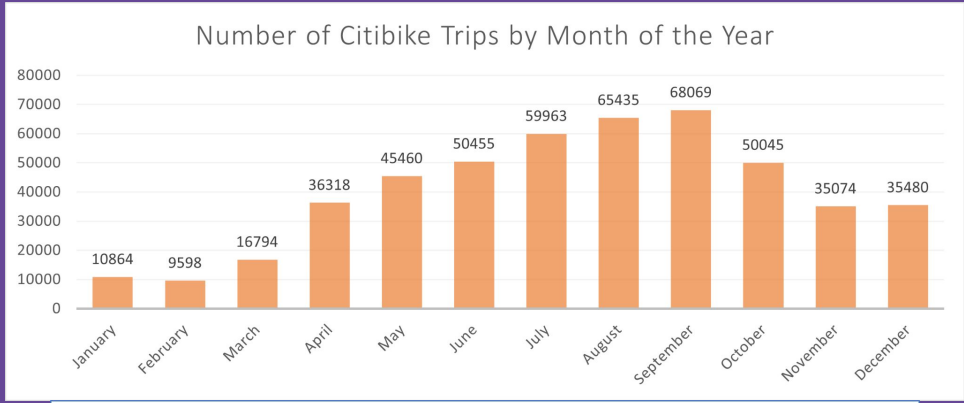
Problem 2 – Cluster Analysis

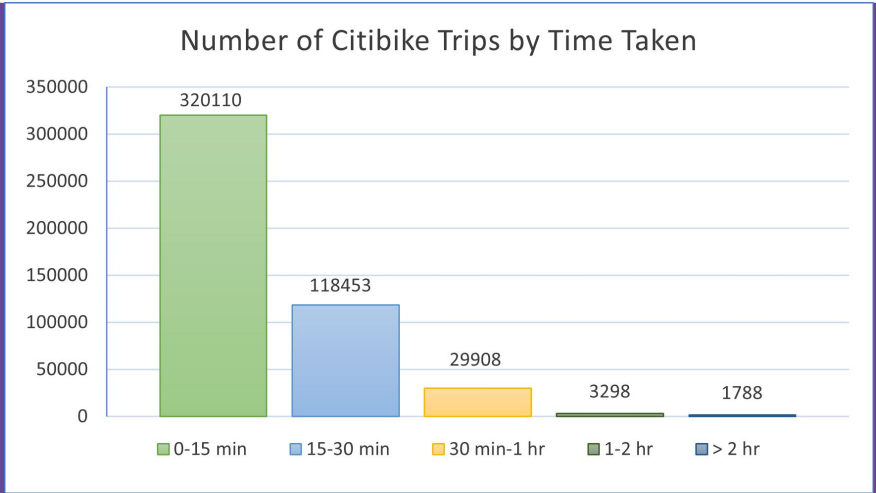
Analyze customer behavior and usage patterns to enhance the user experience and optimize the Citi Bike system.

02

Preprocessing

Visualizing the Dataset





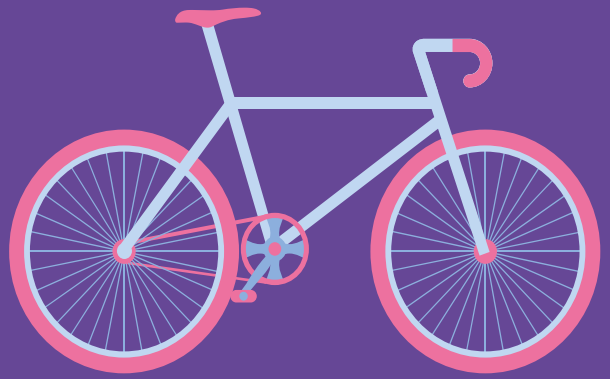
- Observed winter months tend to have the least amount of trips.
- Most users tend to take short trips (0-15 minutes)

Preparing Data for Clustering

- MinMax Normalization of numerical attributes to be in between 0 and 1
- Null values replaced with average
- Boolean attributes encoded to 0=false and 1=true
- Removed all other categorical features and distance

Digraph Construction

- Digraph from the trip data
  - Each station is a vertex
  - Each trip represents an edge: start station → end station
  - To focus on station interconnectivity instead of customer frequency, trips with the same start and end are represented only once.
- Multi-Digraph from trip data but with duplicate edges

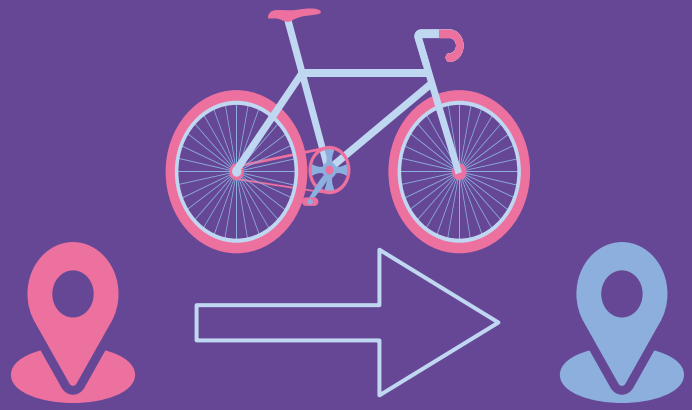


03

Implementation

Link Analysis: PageRank

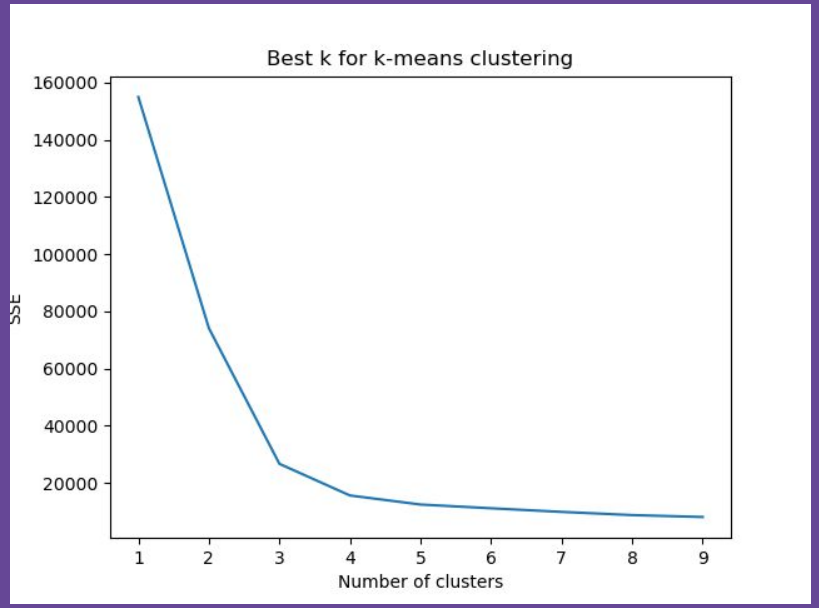
- PageRank algorithm with damping factor 0.15 and 1000 iterations to obtain the ranking of each station
- Input: Digraph .dot files
- Output: Each bike station with individual station's PageRank



Cluster Analysis: K-means

- Used scikit-learn KMeans package
- Tested various k's (elbow method) on the change in SSE
- Tried to calculate silhouette score but calculation was too slow

Finding the Best *k*



The best *k* is k = 3

After k = 3, there is diminishing returns in the SSE reduction as the number of clusters is increased.

04

Results and Analysis

Results – PageRank

Top Ten PageRank Results for Duplicate and Non-duplicate Edges

Station Name	Station ID	PageRank
Cleveland Pl & Spring St	151	0.002991
Pershing Square N	519	0.002913
E 17 St & Broadway	497	0.00289
S 5 Pl & S 4 St	532	0.002761
Broadway & W 60 St	499	0.00267
Centre St & Chambers St	387	0.002626
Mott St & Prince St	251	0.00261
Broadway & E 22 St	402	0.0026
Lawrence St & Willoughby St	323	0.002593
Lafayette St & E 8 St	293	0.002592

Analysis – PageRank

- Station IDs 3468, 3017, 3240, 3014 , 3450, 3485, 3506, 3557, 3607, and 3636 repeatedly appeared as least frequented destinations
  - Usually were close proximity to garages, valet services, and bus stations.
  - Their low ranking out of 840 stations suggests retiring these stations to reduce maintenance costs and allocate resources to more frequented stations.
- Stations with IDs 151, 519, 497, 499, 387, 402, 293, 253, 532, and 323 all had high levels of importance in the CitiBikes network making them are good candidates for expansion and upgrades.

Cluster Analysis Results

Table 2: Statistics for each cluster

	tripduration	birth_year	hour_start	is_subscriber	is_female	N
cluster_0	767.810578	1976.950293	13.889871	1.0	0.000000	316901
cluster_1	892.278741	1978.540183	13.903916	1.0	1.000000	99569
cluster_2	2014.201678	1977.722717	14.490050	0.0	0.022808	57086

cluster_0

~12 minute trips, around 2PM, male subscribers

cluster_1

~15 minute trips, around 2PM, female subscribers

cluster_2

~30 minute trips, around 3PM, mostly male (but some female) users who **ARE NOT** subscribers

Limitations and Future Works

- Limitations:
    - Only could look at a small portion of Citi Bike trip history
    - Only NYC locations
    - Visualization prohibited by Graphviz maximum link threshold (<= 400 links)
- Future works:
    - Incorporating other data sources (e.g. weather, traffic, social media)
    - Analyze whole Citi Bike dataset for historical trends
    - Look at more specific user demographics for the area