

Predicting Citi Bike Availability in Brooklyn

DATA 698 Research Project

CUNY Fall 2022

Philip Tanofsky

2022-11-23

Citi Bike Availability

Predict bike availability to avoid the empty racks



Problem and Objectives

Problem

- Citi Bike offers live map of availability and the Lyft app provides real-time availability for immediate bike rental but the availability of a bike at a location at a future date and time is not guaranteed

Objectives:

- Understand Citi Bike usage and availability patterns across New York City
- Predict the number of available bikes in Brooklyn within a quarter mile of a given location with user-friendly inputs

Previous Work

- Wang, Lindsey, Schoner, and Harrison (2016)
 - Log-linear and negative binomial regression models
- Wang and Chen
 - Zero-inflated negative binomial regression model
- Hyland, Hong, Pinto and Chen
 - Hybrid clustering and regression models
- Médard de Chardon, Caruso, and Thomas
 - Rebalancing evaluation

Data Collection

Citi Bike System Data

- Download CSV files for each month
- API call every 15 minutes for two weeks

NYC Open Data

- Borough shapes via GeoJSON files

Docking Station Elevation

- R library elevatr

Surplus/Shortage Timetable

- Convert trip data to dataframe of bike departure and arrival counts by docking station for a given time interval

Data Statistics

Ridership patterns

- Timeframe: August - October of 2022 (92 days)
- Location: New York City departing trips
- Number of trips: 10,210,102
- Docking Stations: 1,717
- Abandoned trips: 24,887

Bike Availability

- Timeframe: Oct. 31 - Nov. 13 of 2022 (14 days)
- Location: Brooklyn
- Docking Stations: 474

Challenges

Large volume of data

- Over 10 million trips in 3 months
- Over 1700 docking stations in NYC

Rebalancing

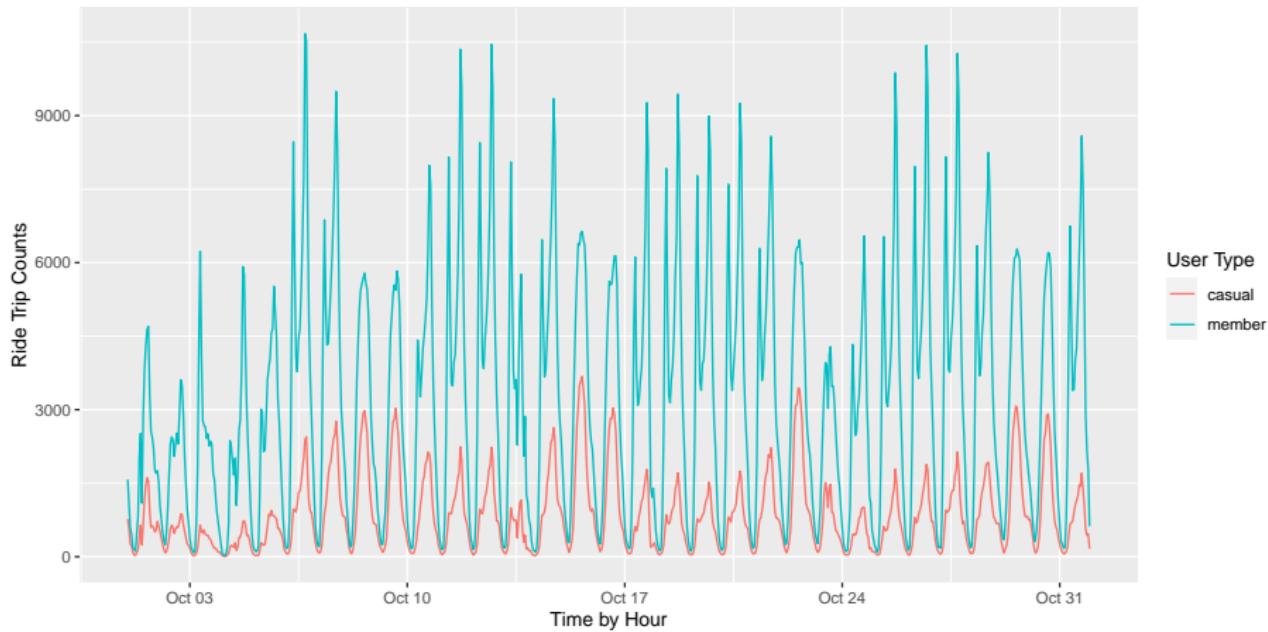
- System action of moving bikes to restock docking stations
- Identification and inconsistency

User-friendly availability prediction

- Independent variables to prediction model
- End-user input values

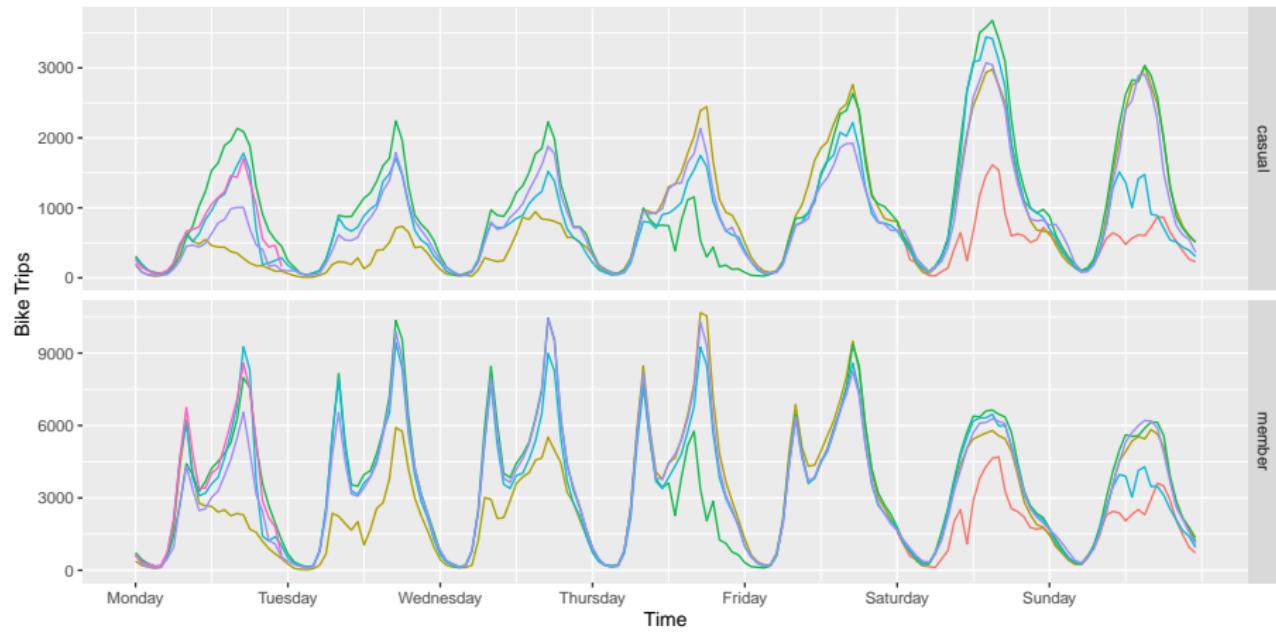
EDA: Bike Trips by Hour

Bike Trips by Hour by User Type for Aug. - Oct. 2022



EDA: Weekly Seasonal Pattern

Weekly seasonal plot of the bike trips for Aug. - Oct. 2022

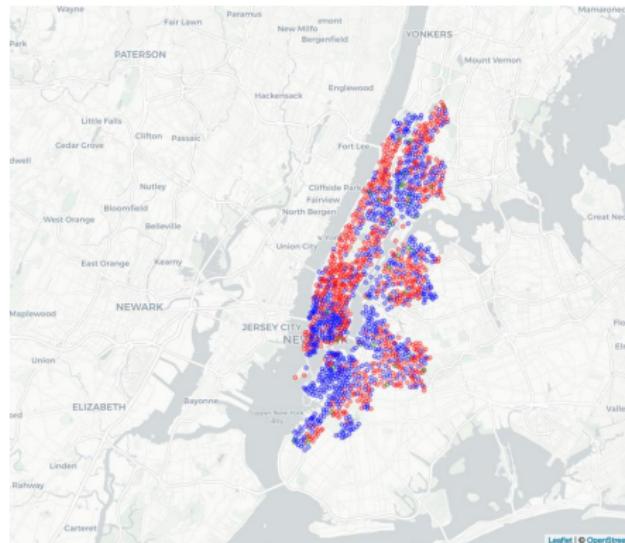


EDA: Three-Month Summary

System-wide view of surplus or shortage for every NYC docking station

Docking Station Color

- Blue: Surplus
- Red: Shortage
- Green: Even

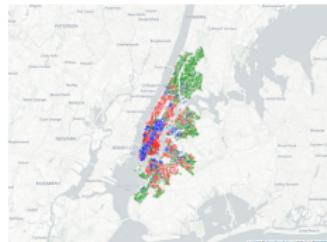


EDA: Daily Pattern

Time lapse of surplus or shortage for every NYC docking station



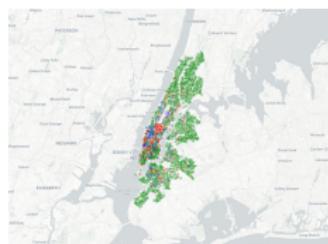
(a) 5 a.m.



(b) 8 a.m.



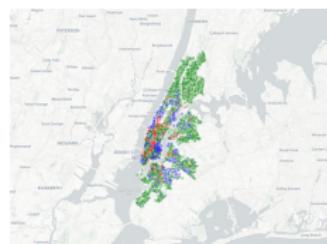
(c) 11 a.m.



(d) 2 p.m.



(e) 5 p.m.



(f) 8 p.m.

Overview of Algorithm Approach

Goal: Predict number of available bikes in Brooklyn within quarter mile

Inputs

- Latitude and longitude (location on map)
- Day of the Week
- Time of Day

Step 1: Hierarchical Clustering

- Cluster all the Brooklyn docking stations
 - Distance: 400m (Quarter mile)

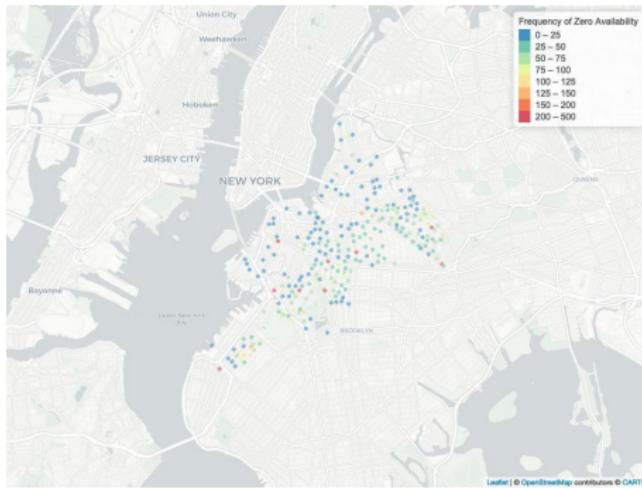
Step 2: Apply Generalized Linear Models

- Predict number of bikes available per cluster
 - Count Models: Poisson, Negative Binomial, Zero-Inflated

Stations with Zero Bike Availability

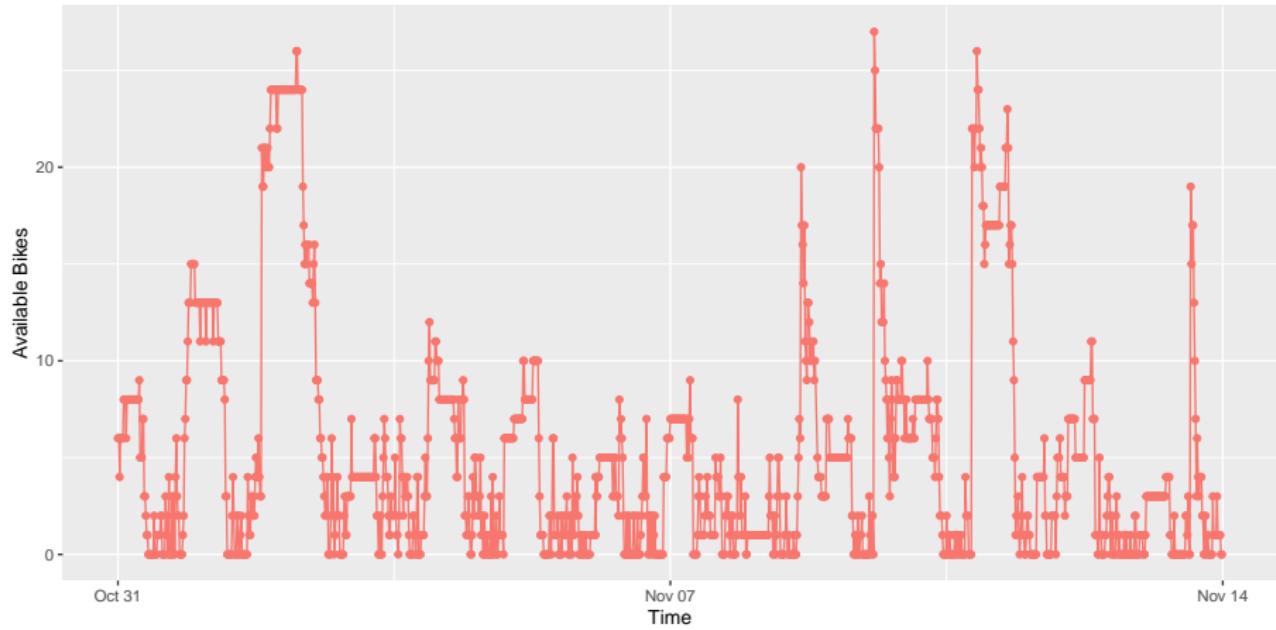
Frequency of zero bikes available per docking station in Brooklyn

- Span: Two weeks
- Instance: 15-minute interval



Rebalancing Example

Inconsistent bike availability at Docking Station 3582 in Brooklyn



Proposed Methodologies

Predictors

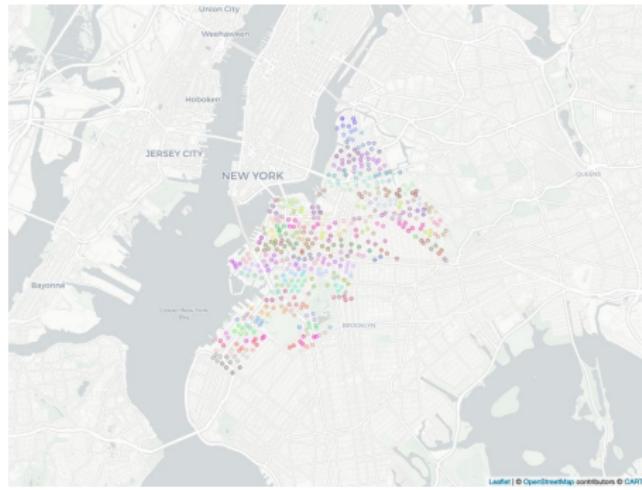
- Latitude and longitude converted to cluster
- Day of the Week converted to 'Weekday' or 'Weekend'
- Time of Day in 1-hour intervals
- Average elevation of cluster omitted

Model Methodologies

- Selected: Generalized Linear Model for Counts
 - Poisson Regression Assumptions:
 - Response variable is count per unit of time
 - Independence: Observations are independent in nature
 - Mean equal to Variance: Not true
- Alternate Consideration: Time Series Model
 - Forecasting does not provide instance prediction
 - Not easily translatable from user-friendly inputs
 - Requires constantly up-to-date info to forecast from given point in time

Modeling Step 1: Clustering Model

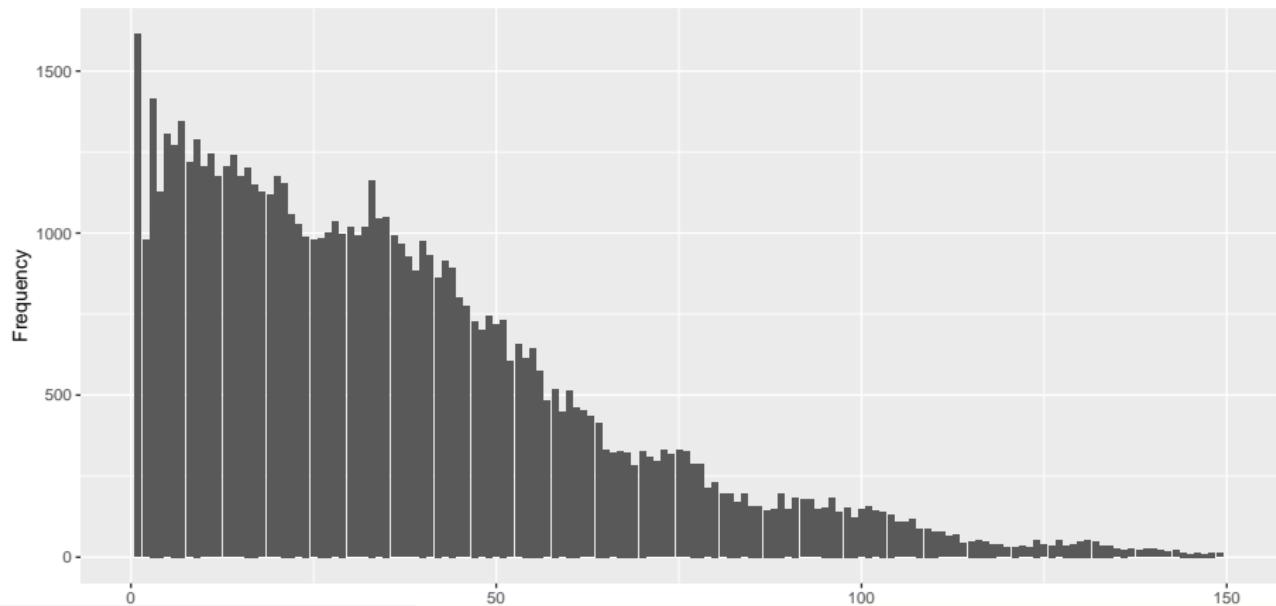
- Borough: Brooklyn
- Docking Stations: 474
- Clusters: 213 (2.23 stations/cluster)



Bike Availability Count Frequency

Histogram of availability based on 1HR intervals for Brooklyn clusters

- High frequency near 0 and 2 and plateaus at docking station capacities



Modeling Step 2: Count Models

Generalized Linear Models for Count Data

5 Models Attempted

- Poisson
 - stats library
- Quasi-Poisson
 - stats library
- Negative Binomial
 - MASS library
- Zero-Inflated
 - Poisson and Negative Binomial
 - pscl library

Model Results 1-Hour Interval

- Certain input to model based on 1 hour intervals below
- Results table

```
## [1] 916.2016
```

```
## [1] 785.3156
```

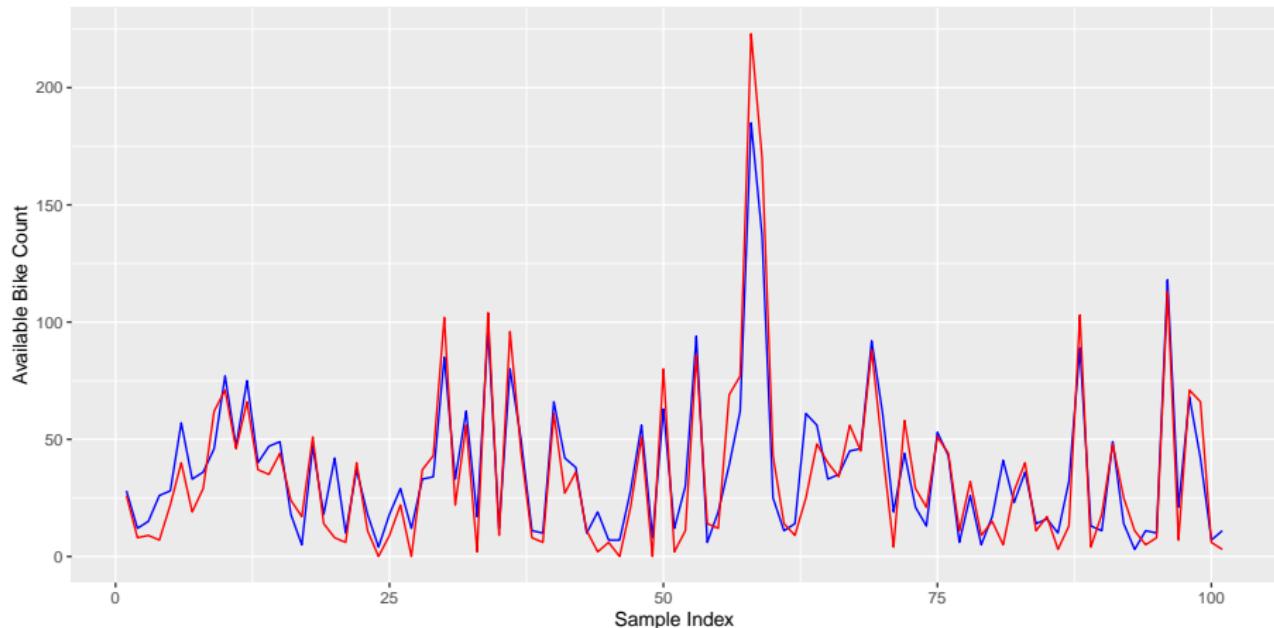
```
## [1] 36.50179
```

Model	RMSE	MAE	Missed.Zero.n	Missed.Zero
Zero (Poisson)	11.50859	8.229184	263	
Poisson	11.56925	8.280158	263	
Quasi-Poisson	11.56925	8.280158	263	
Zero (Neg. Binomial)	12.09165	8.570985	263	
Negative Binomial	12.30029	8.717866	263	

Model Results Visualization

Plot subset of Predicted vs Actual counts for select model - Poisson

- Blue: Prediction; Red: Actual



Availability Prediction

Model prediction inputs

- Longitude and Latitude (GPS location of smartphone)
- Hour of the day (dropdown proposed)
- Weekday or Weekend (dropdown proposed)

```
## [1] "Crown Heights - Available Bikes: 13; Cluster: 58"
```

```
## [1] "Williamsburg - Available Bikes: 57; Cluster: 201"
```

```
## [1] "Sunset Park - Available Bikes: 4; Cluster: 10"
```

Current Achievements

Bike Availability prediction model for Brooklyn

- Hierarchical clustering
 - Location-based
- Poisson Model
 - Simplest count model
 - Better predictor of zero availability

Ridership usage patterns

- Capture patterns of Citi Bike usage
 - System wide evaluation of NYC
 - Weekly and daily bike trip patterns

Future Works

- Deploy web application or smartphone app to use GPS location
- Model of all New York City
- Real-time clustering to ensure all docking stations within quarter-mile of user
- Increase amount of availability data