# Predicting Citi Bike Availability in NYC

## DATA 698 Research Project
## CUNY Fall 2022

Philip Tanofsky

2022-11-21

# Citi Bike Availability

Words go here

# Problem and Objective

Introduce yourselves and describe your problem. Explain your objectives, challenges of your work, proposed methodologies, and the assumptions you made while conducting modeling and/or analysis. Provide an overview of your approach and/or conceptual model (please do not present your code directly). Describe the results you obtain and summarize the current achievements and possibility of future works.

v2 is started at 1:30P on Sat, Nov 19

# Previous Work

- bullet 1
- bullet 2
- bullet 3
- bullet 4

# Challenges and Assumptions

- Too much data
  - Over 10 million trips in 3 months
- Rebalancing identification and not consistent
- User friendly approach
  - Inputs to output

# Data Collection

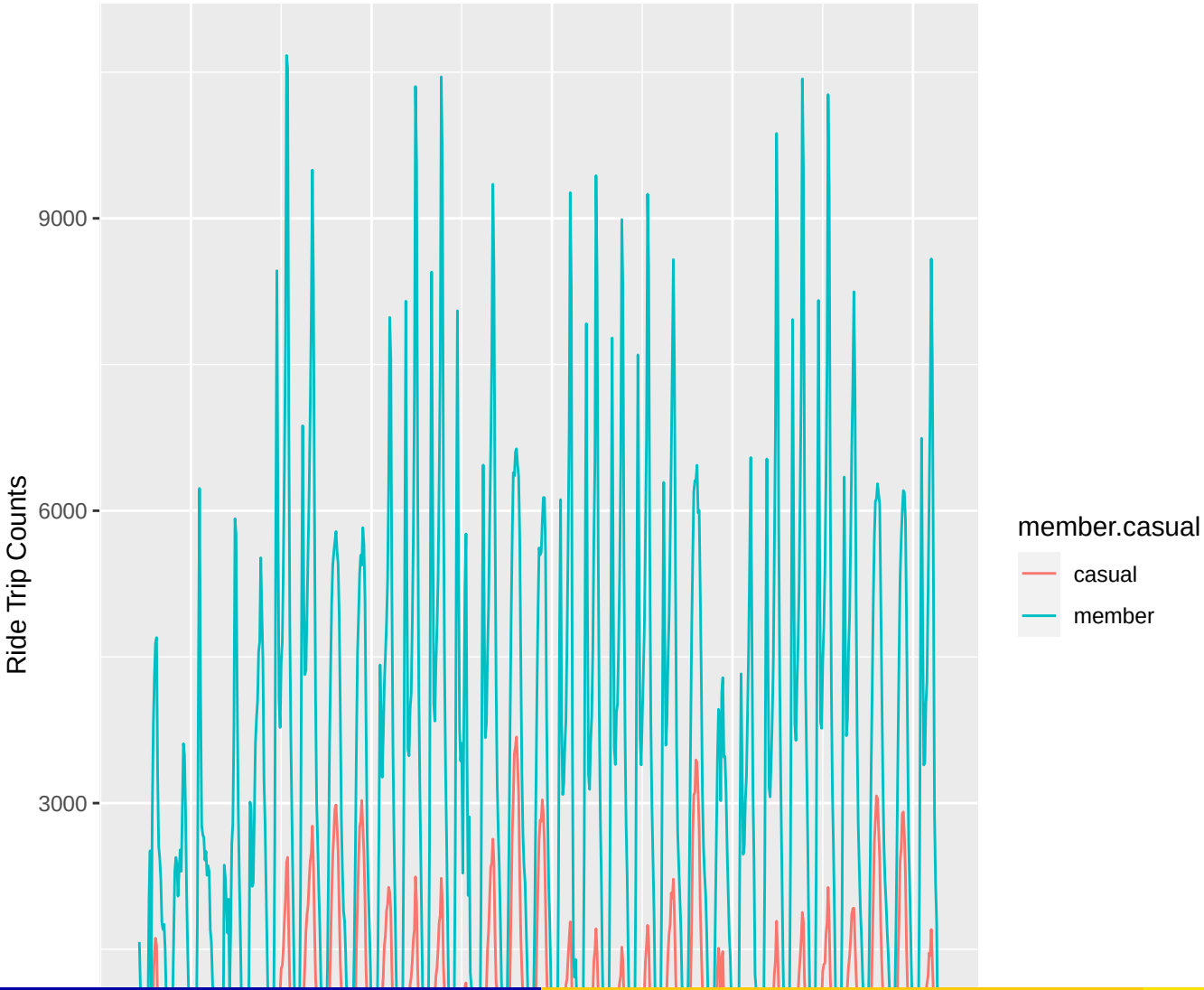
- Citi Bike System Data
  - Download CSV file
  - API call every 15 minutes for two weeks
- NYC Open Data
  - Borough coordinates
- Creating timetable of surplus/shortage

# Data Analysis

- Timeframe: August - October 2022
- Number of trips: 10,210,102
- Stations: XXX
- Abandoned trips: 24,XXX

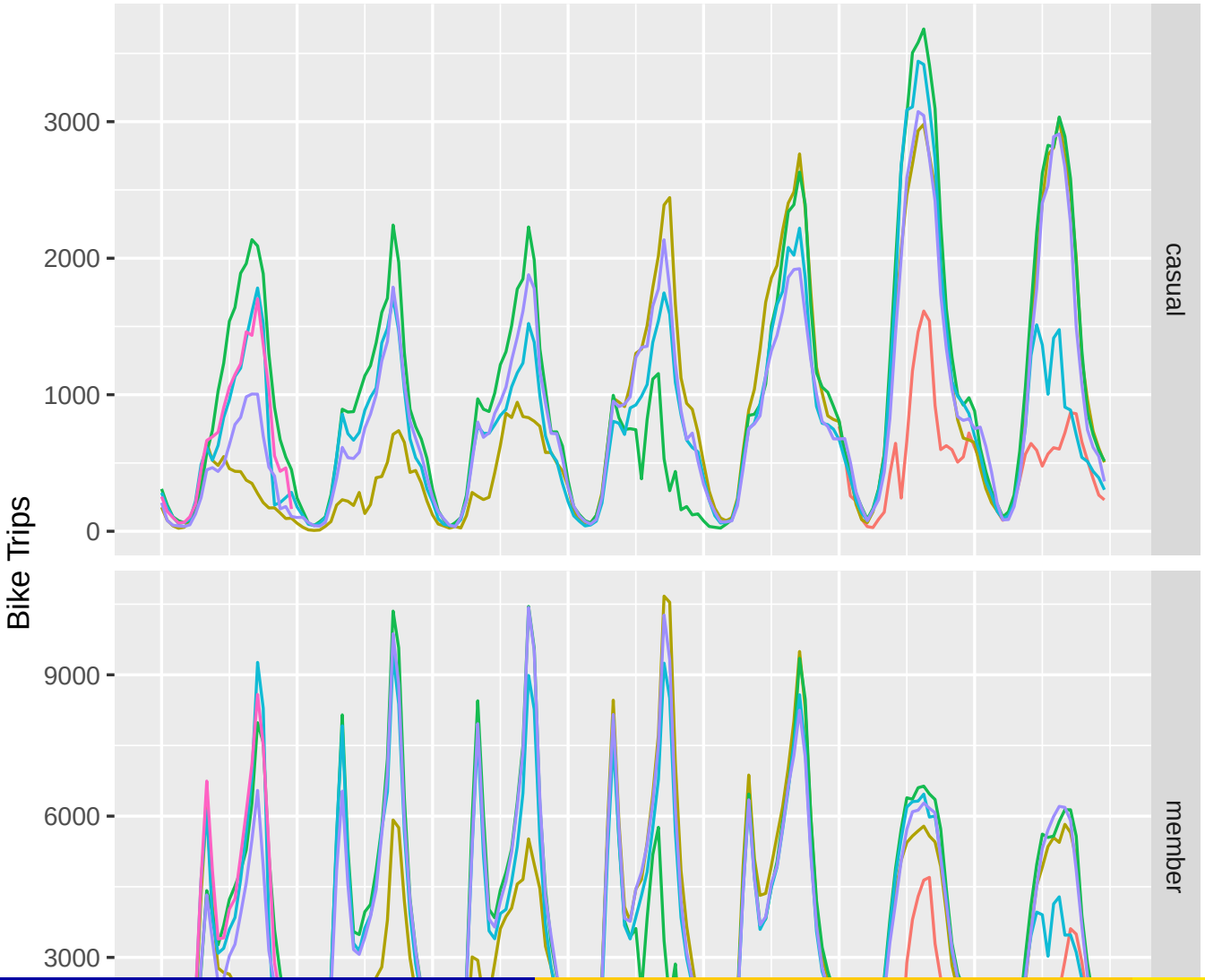
# EDA: Bike Trips by Hour

## Ride Trips by Hour by User Type



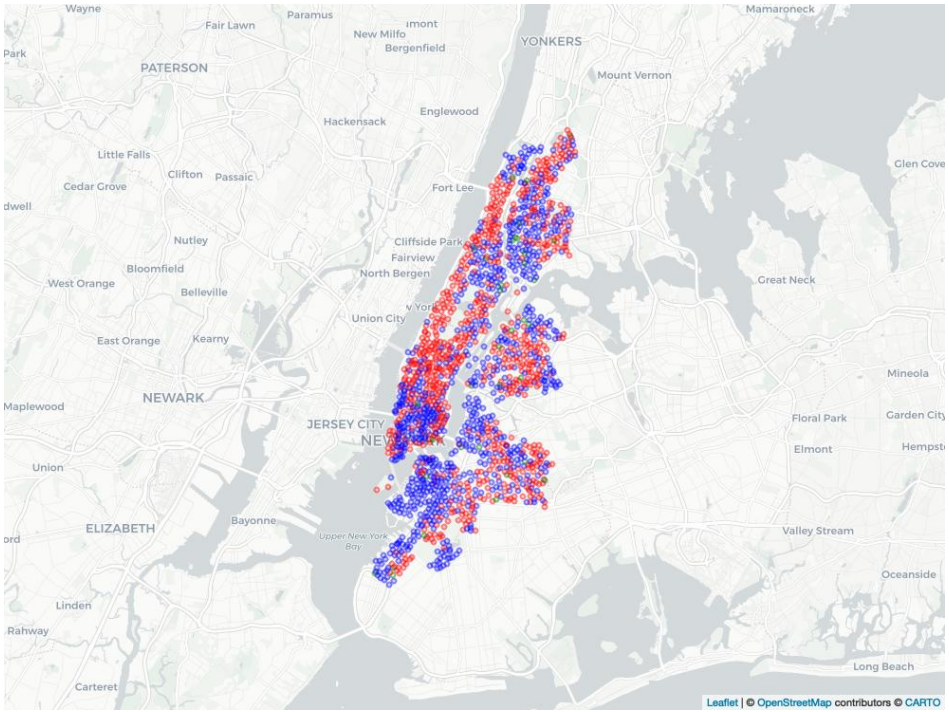
# EDA: Weekly Seasonal Pattern

Weekly seasonal plot of the 10 million bike trips



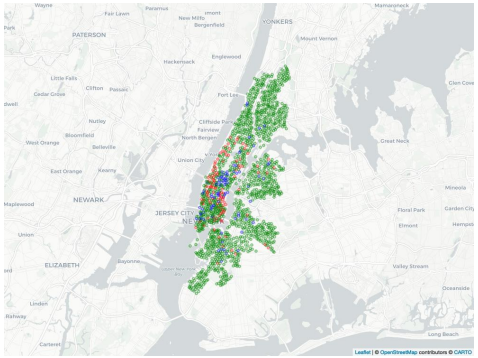
# EDA: Three-Month Summary

System-wide view of surplus or shortage for every docking station - Blue: surplus; Red: shortage; Green: even

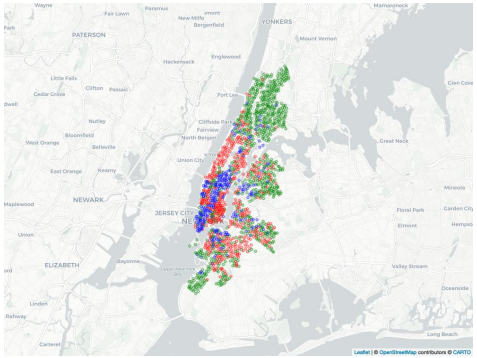


# EDA: Daily Pattern

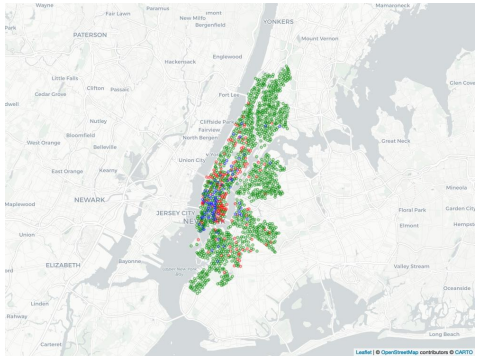
Small multiples of Wednesday averaged for time lapse of 5A to 8P



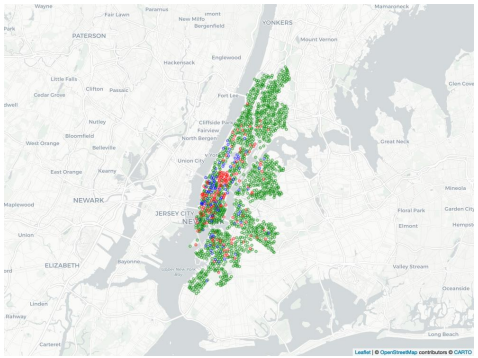
**(a)** 5 a.m.



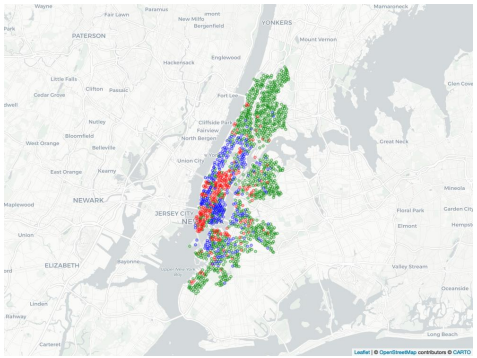
**(b)** 8 a.m.



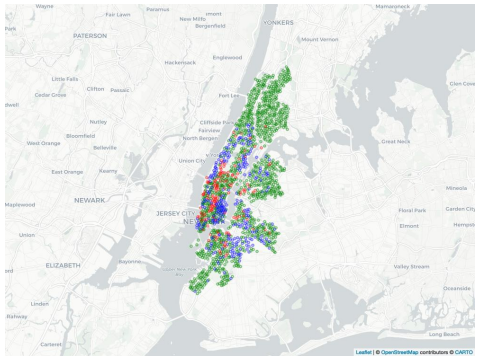
**(c)** 11 a.m.



**(d)** 2 p.m.



**(e)** 5 p.m.

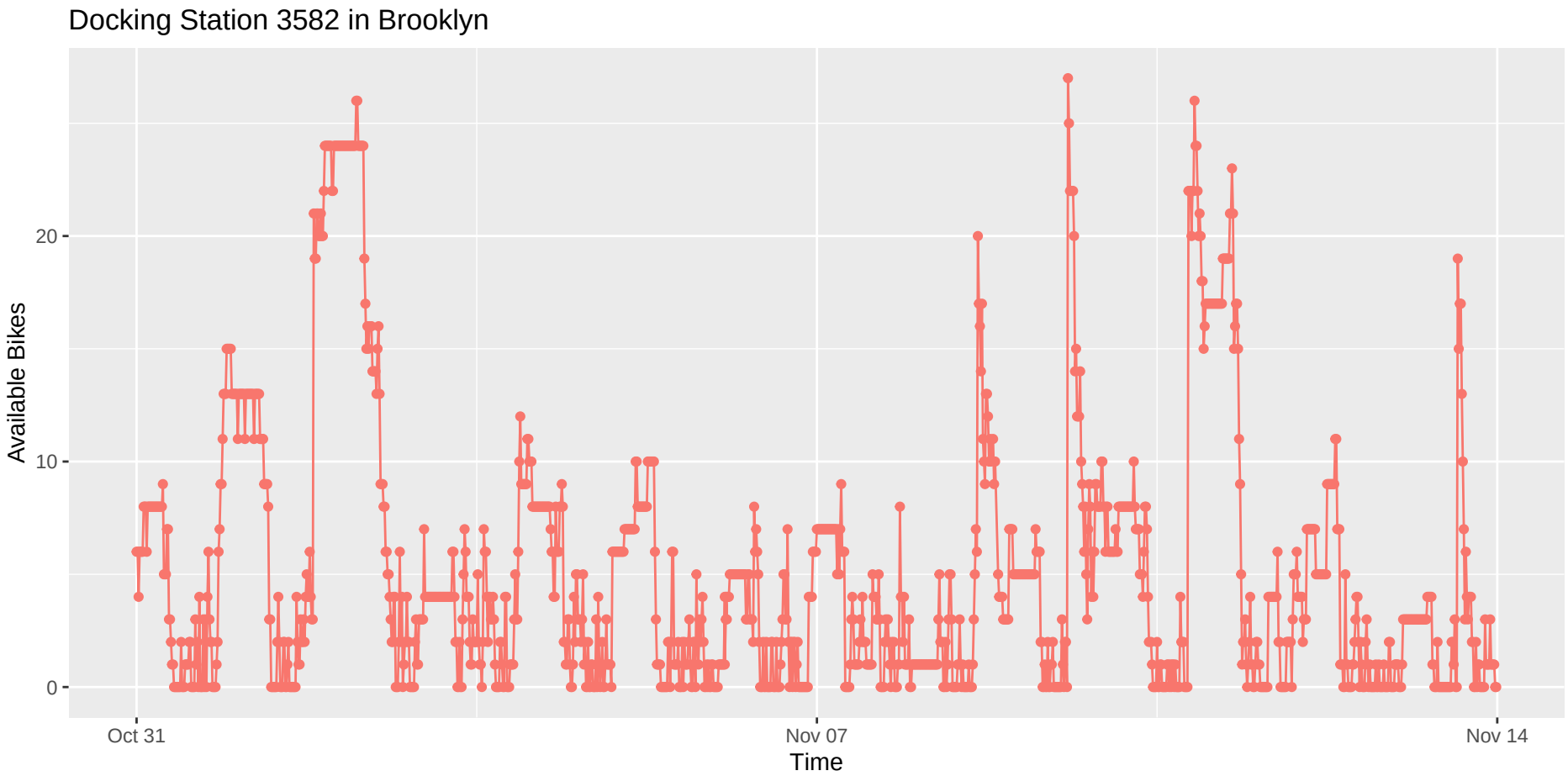


**(f)** 8 p.m.

**Figure 1:** Many figures

# EDA: Rebalancing

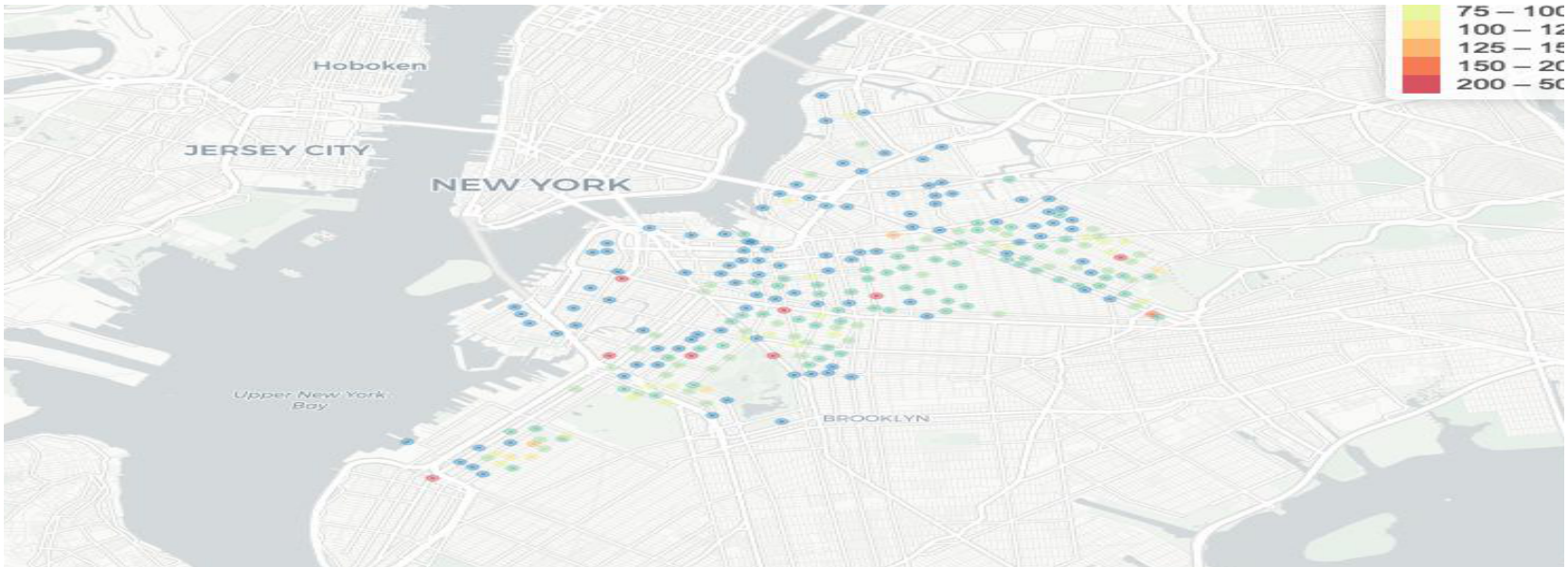
Example of inconsistent nature of rebalancing



# Stations with Zero Bike Availability

Frequency of zero bike availability by docking station - Two weeks - 15-minute intervals

**Brooklyn Docking Stations by Cluster**



# Proposed Methodologies

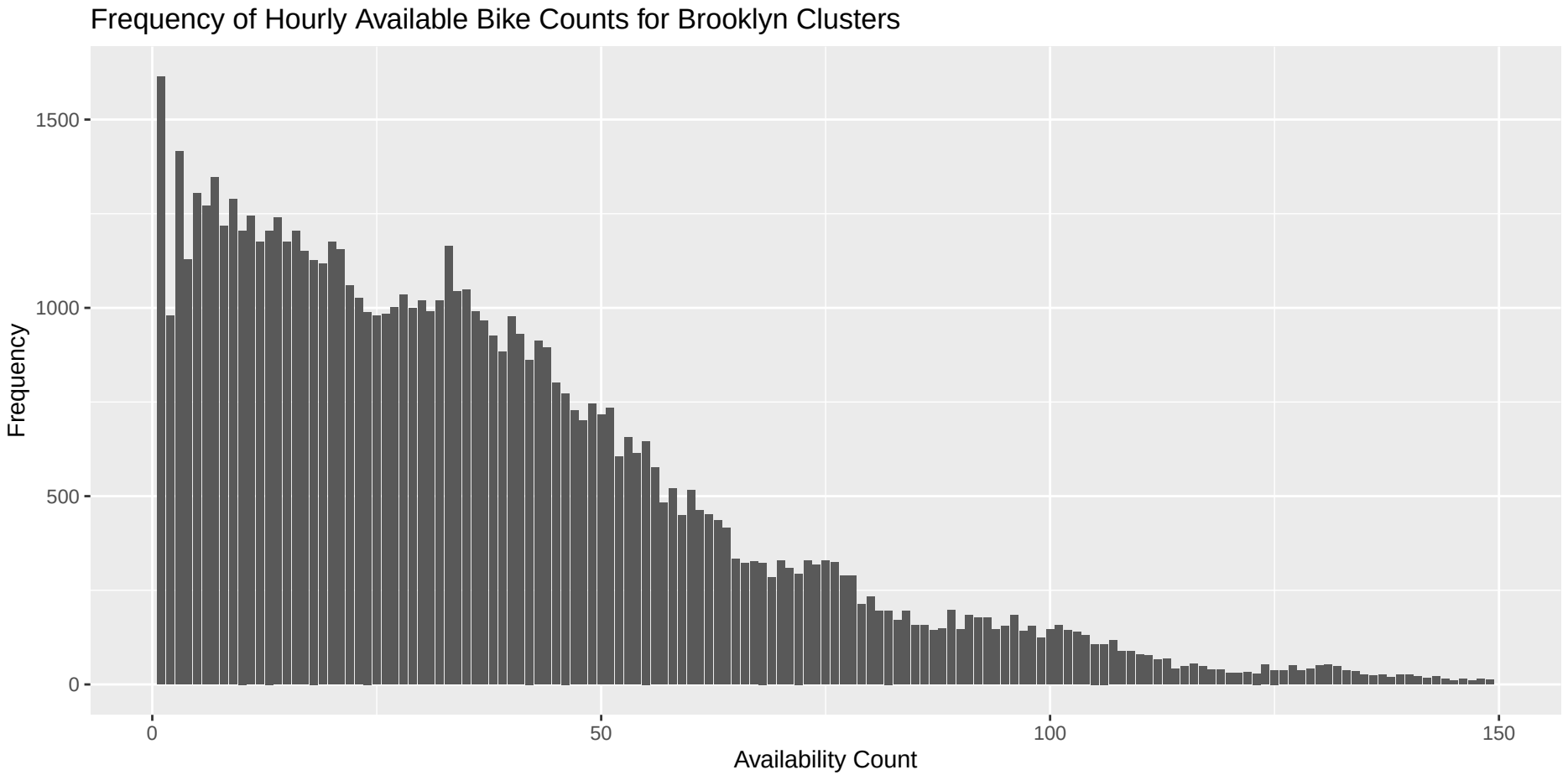
- Inputs
  - Latitude and longitude
  - Day of the Week
  - Time of Day (15M and 1H)
- Citi Bike offers live map of availability
- Lyft provides real-time availability
- Time series model not used explanation (or really, just don't include)
- Poisson distribution and Negative Binomial given the over-dispersion

# Overview of Algorithm Approach

- Data from API call every 15 minutes for two weeks
  - Citi bike availability at each station
  - Two weeks is small interval to predict
    - Valid limitation of model

# Bike Availability Frequency by Cluster (Brooklyn)

Chart of availability based on 1HR intervals . . . shows high frequency near 0 and 2 and plateaus at docking station capacities

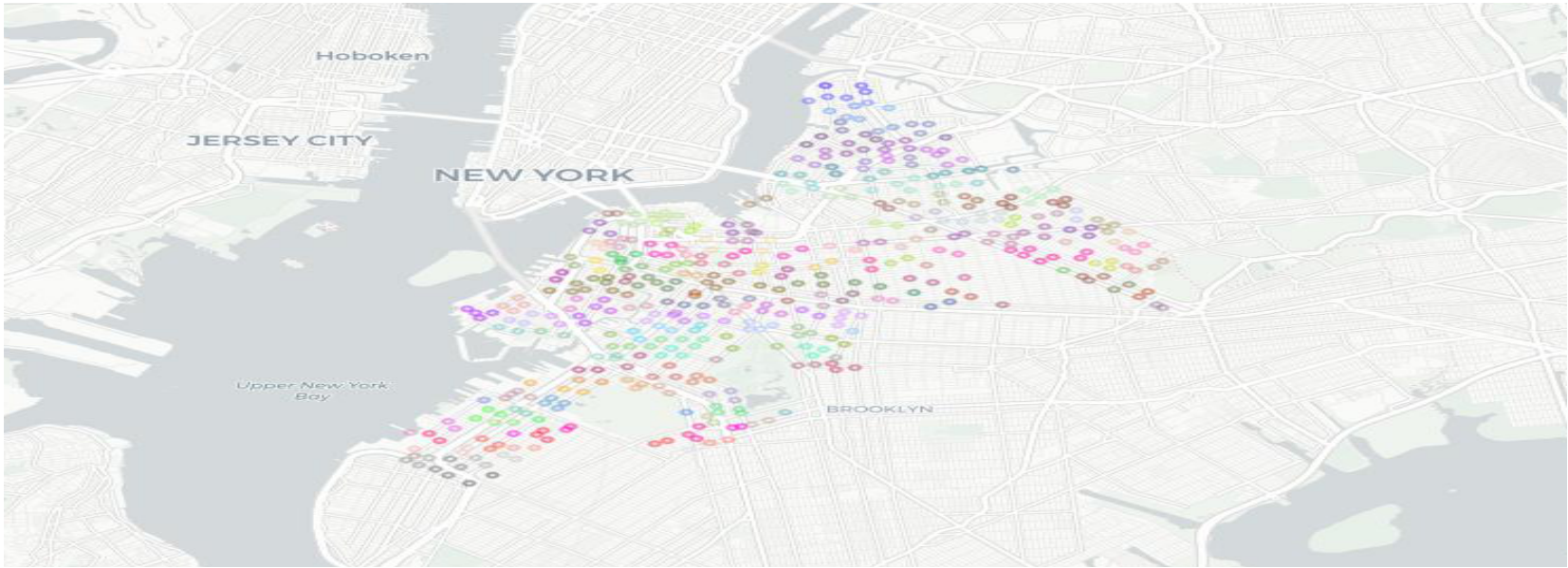


# Modeling Step 1: Clustering Model

Clustering

- Brooklyn
- Docking Stations: 474
- Clusters: 2XX
- Map of Clusters by Color

**Brooklyn Docking Stations by Cluster**



# Modeling Step 2: Count Models

Model Approaches

- 6 models attempted

# Model Results 15-Minute Interval

- Certain input to model based on 15 minute intervals above
- Results table

# Model Results 1-Hour Interval

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

# Model Results Visualization

- Plot of the preds and actuals for best scoring model

# Availability Prediction

- Model prediction
  - Show how it works

# Current Achievements

# Future Works

- Weather . . . actually, can I predict weather? would that really work?
- Subway stations: Citi Bike offers valet
- Model of all NYC
- Real-time clustering would be better