Predicting Citi Bike Availability in NYC

"DATA 698 Research Project"
"CUNY Fall 2022"

Philip Tanofsky

2022-11-20

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 3.5 million trips in month of Sept. 2022
- Rebalancing identification
- User friendly approach
  - Inputs to output

Data Collection

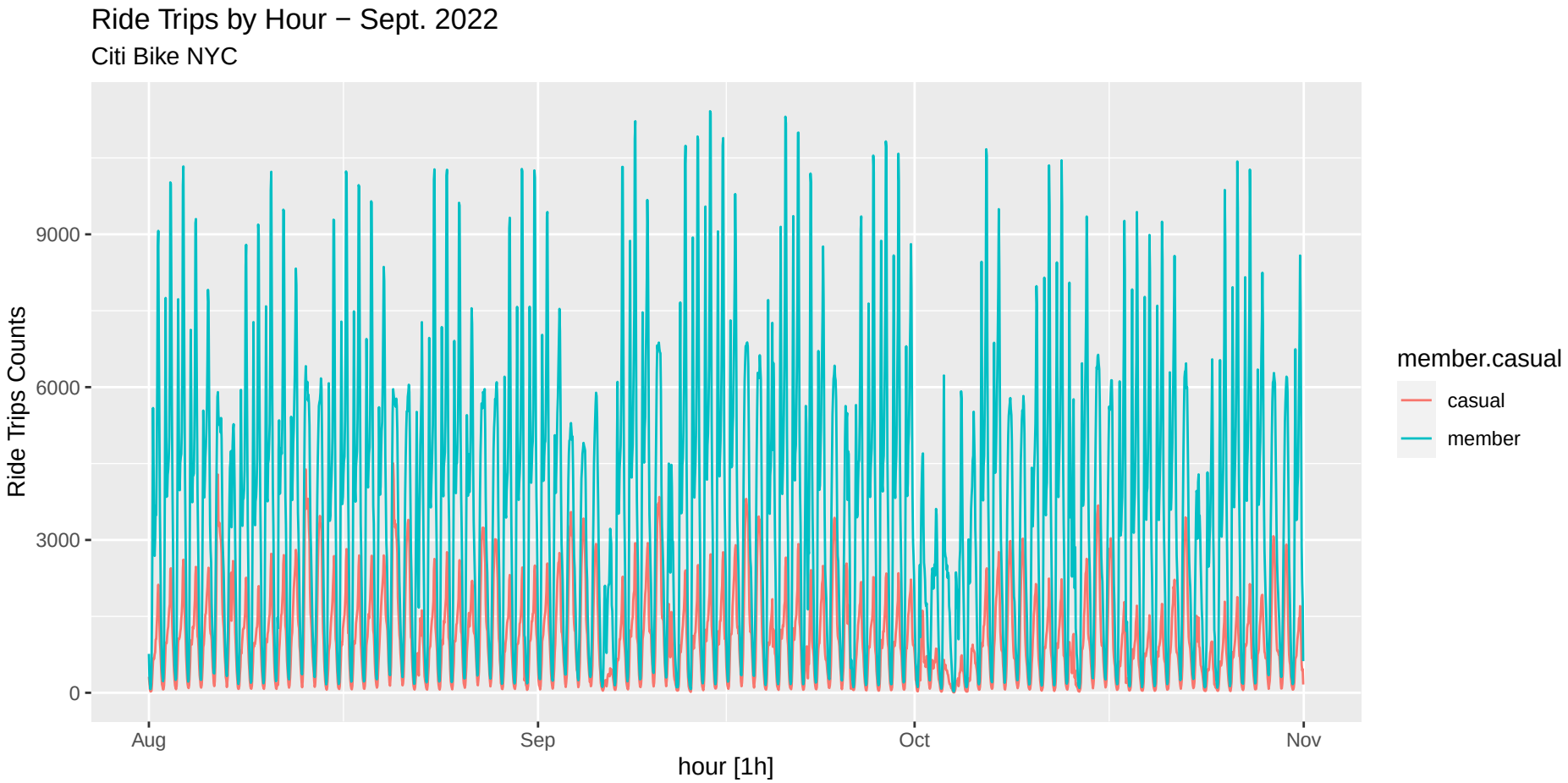
- Creating timetable of surplus/shortage

Data Analysis

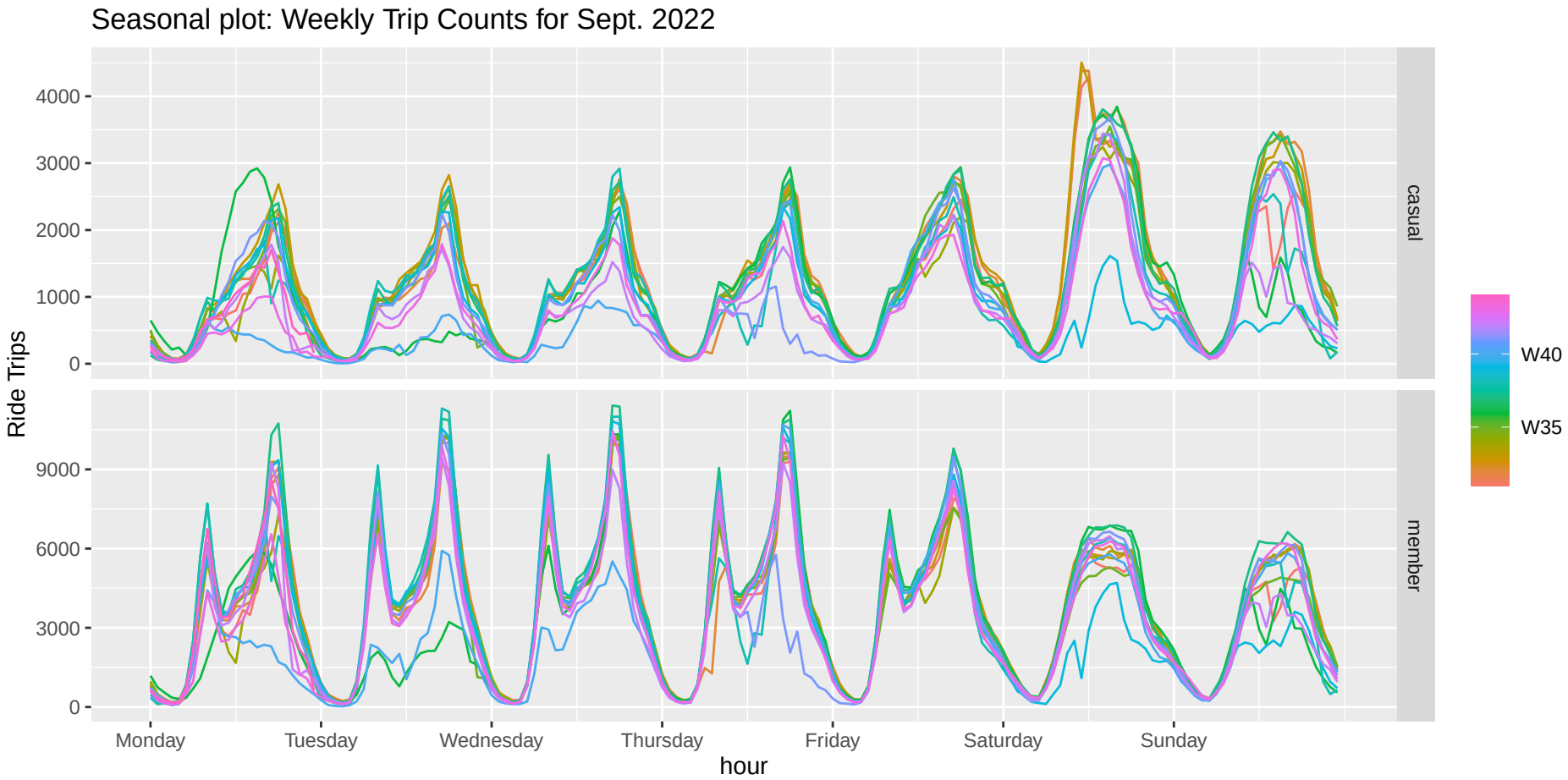
Number of trips Stations XX Abandoned trips

EDA: Bike Trips by Hour

10,210,102 trips for Aug-Oct (3 months of 2022)

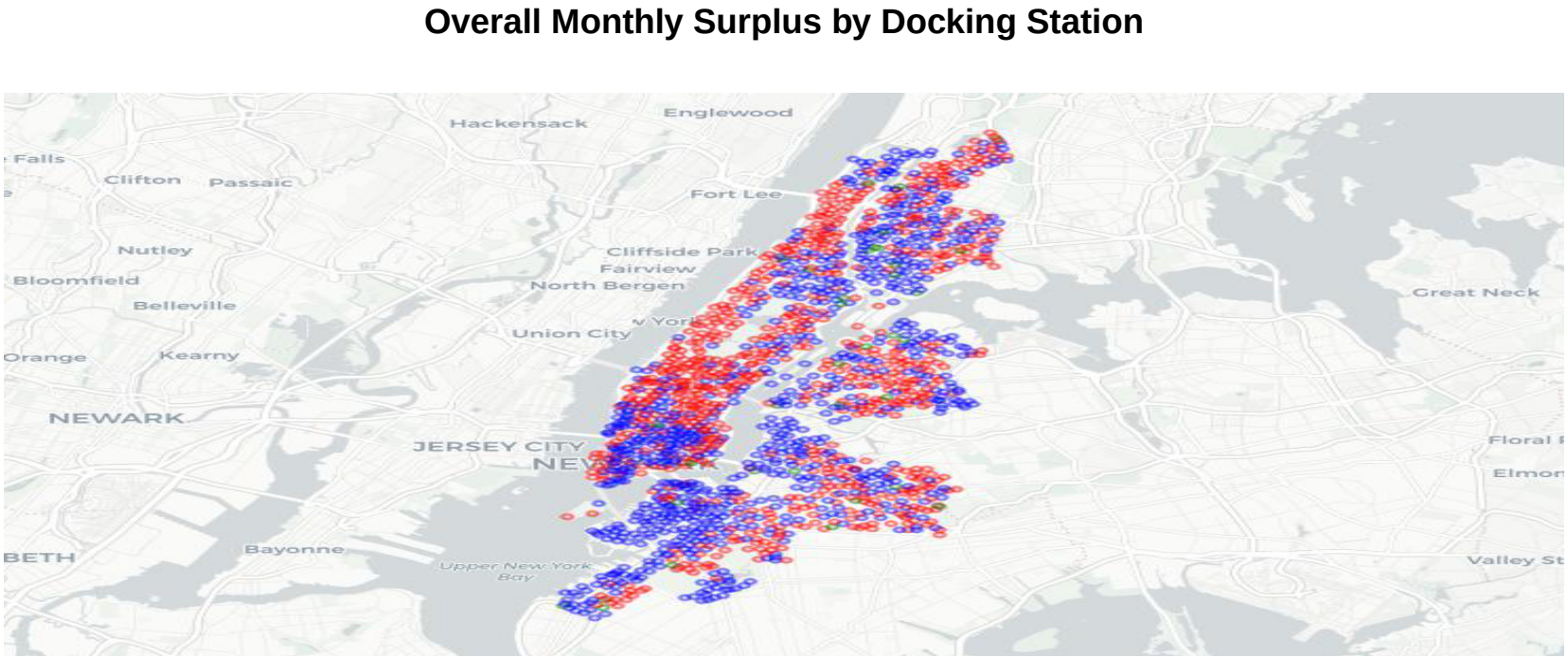


EDA: Weekly Seasonal Pattern



EDA: Three-Month Summary

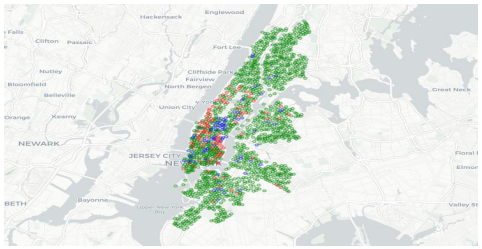
NYC view of surplus or shortage over 3 months



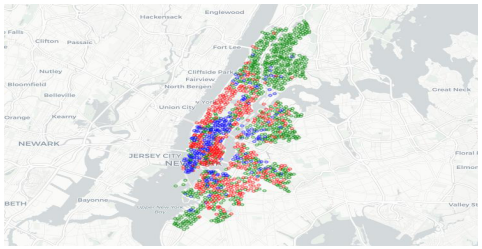
EDA: Daily Pattern

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

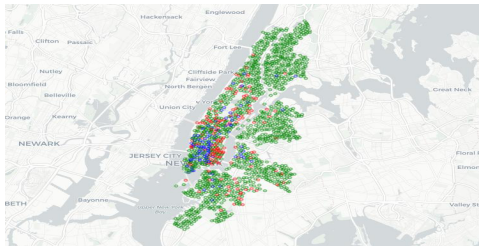
Surplus/Shortage on Wednesdays: 5A



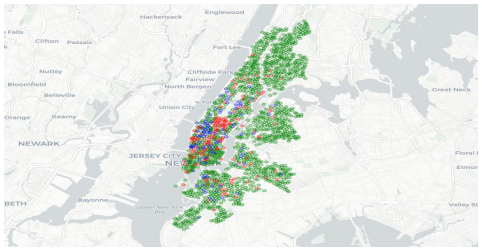
Surplus/Shortage on Wednesdays: 8A



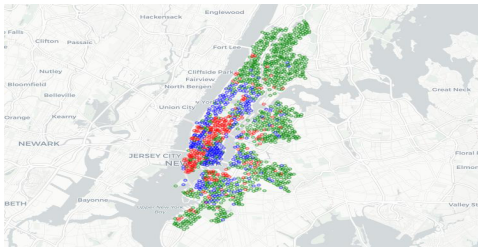
Surplus/Shortage on Wednesdays: 11A



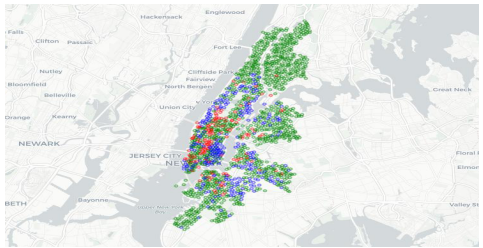
Surplus/Shortage on Wednesdays: 2P



Surplus/Shortage on Wednesdays: 5P

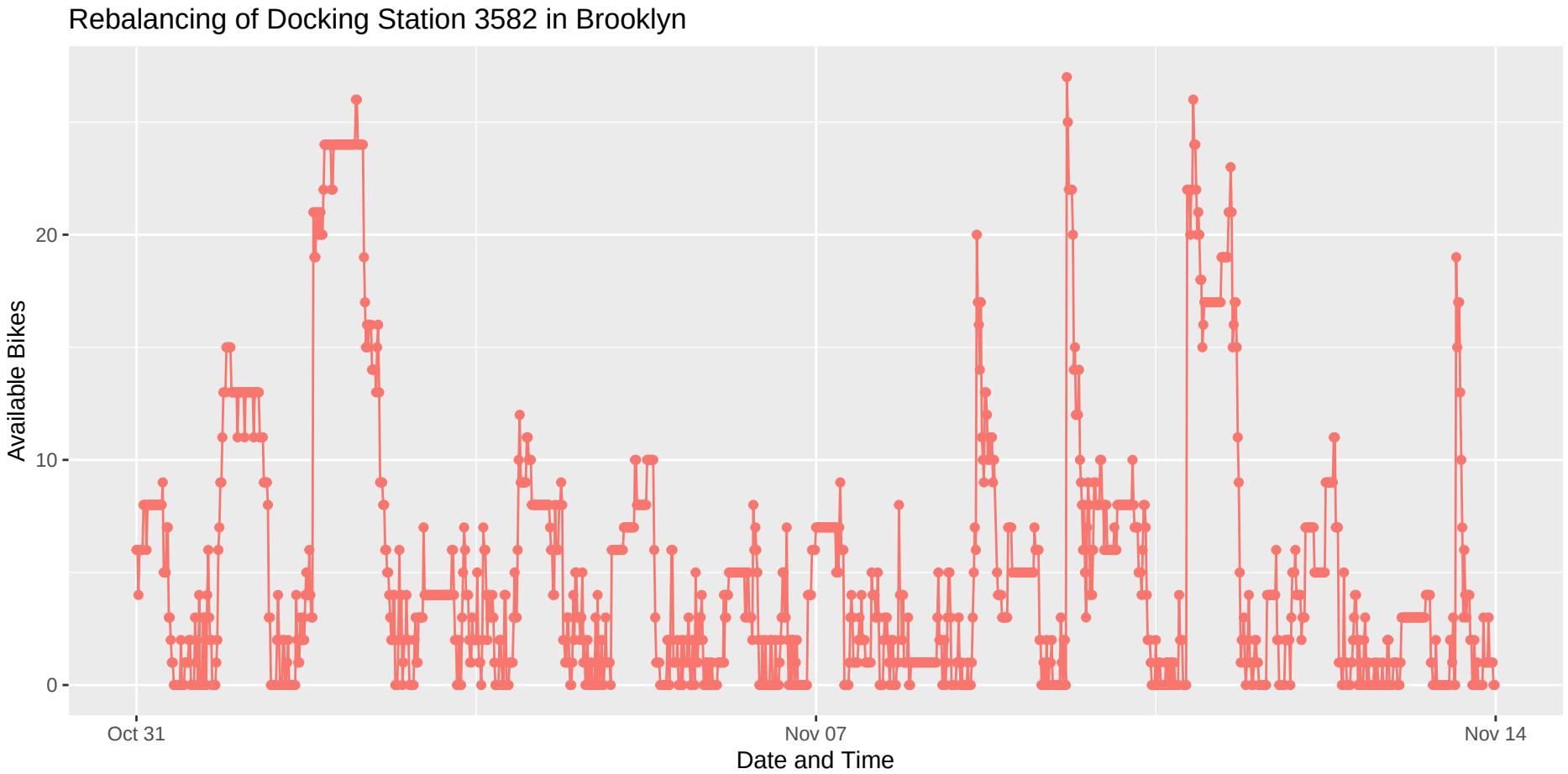


Surplus/Shortage on Wednesdays: 8P



EDA: Rebalancing

Rebalancing . . . in metadata file: Will be plot of a station with availability spikes

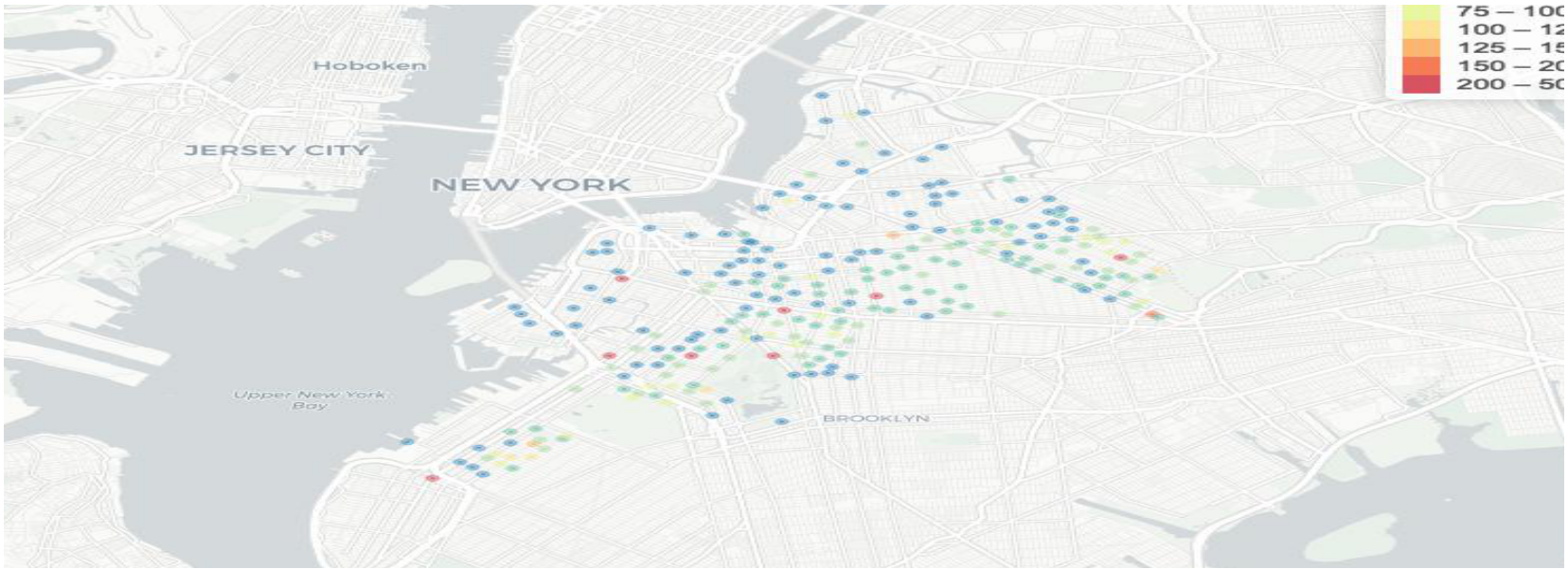


Stations with Zero Bike Availability

Availability Zero for some stations

Which stations have zero availability and how often that occurs over the 2 weeks

Brooklyn Docking Stations by Cluster



Proposed Methodologies

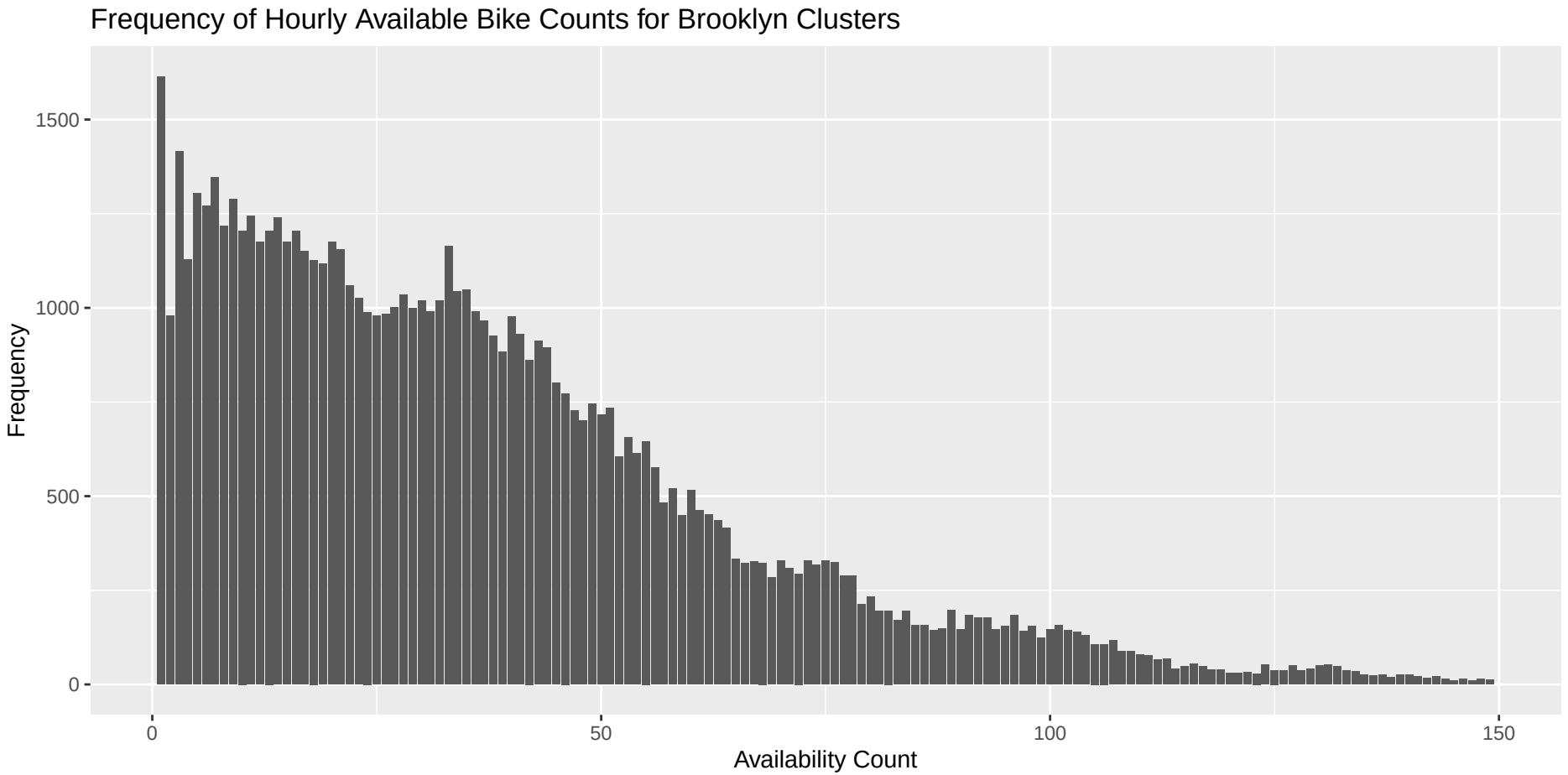
- Inputs
  - Latitude and longitude
  - Day of the Week
  - Time of Day (15M and 1H)
- Citi Bike offers live map of availability
- Lyft provides real-time availability
- Time series model not used explanation (or really, just don't include)
- Poisson distribution and Negative Binomial given the over-dispersion

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

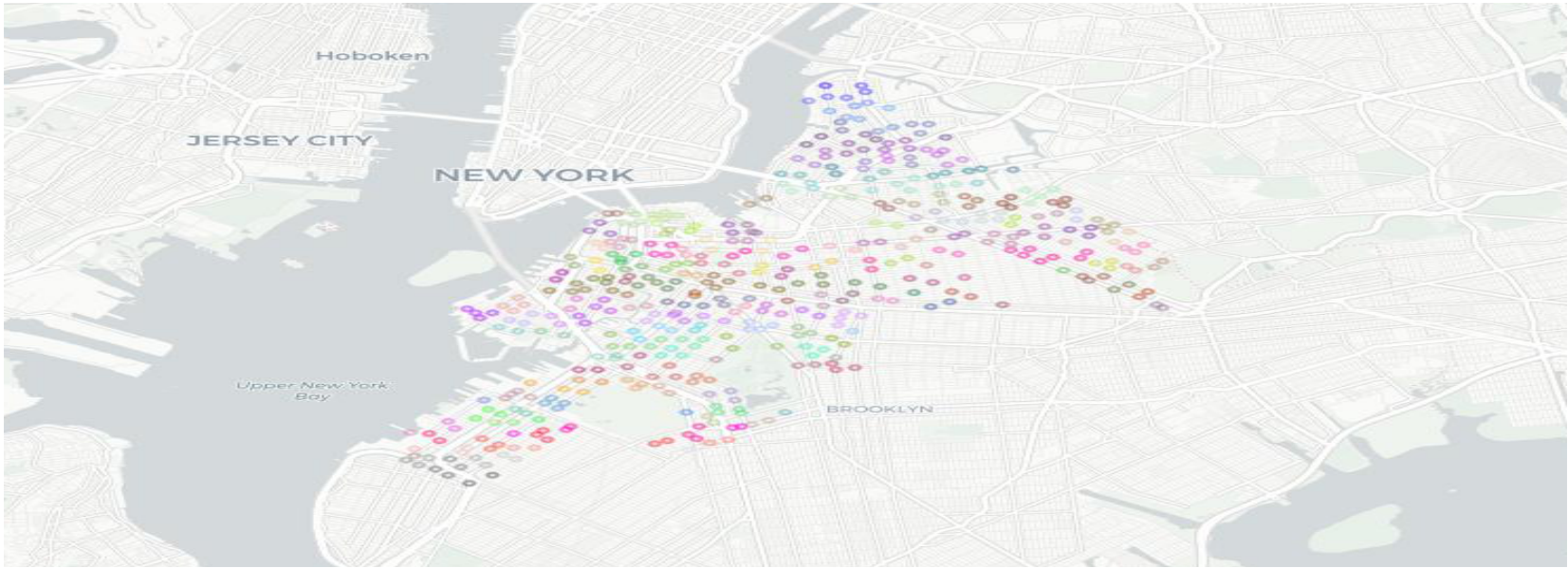


Modeling Step 1: Clustering Model

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