



What is the problem?

- Citi Bike is a bike sharing service in NY with over 24,500 bikes and 1,500 bike stations
- Goal: Provide holistic insights and visualizations for Citi Bike trends and factors impacting ridership behavior to empower city and transit planning



Why is it important?

Unlike previous research, which tends to focus either on a small window of time and/or a specific environmental factor (e.g., COVID), we will provide a holistic view of Citi Bike ridership behavior and station trends as well as factors that impact them.



Who cares?

NYC's population has grown by nearly 8% since 2010, and biking has grown in popularity as a means for transportation, especially considering recent issues with the city's public transit. We will provide insights to support Citi Bike strategy and city planners.

SCRAPE & MANUAL DATA ENTRY FOR SUPPORTING DATASETS



Overview: Daily weather data

Size: <1MB

of Records: 1.5k



Overview: Geographic data for stations

Size: 1MB

of Records: 1k

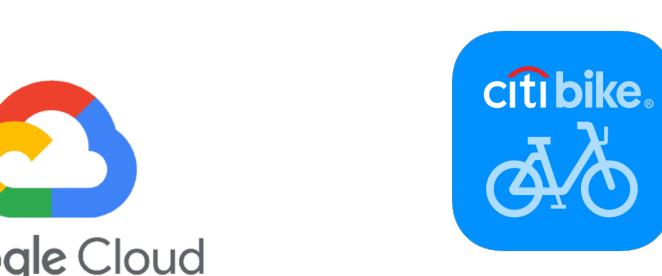


Overview: Citi Bike Membership & Pricing Data

Size: <1MB

of Records: 1k

DATA WAREHOUSING



Storage: Data uploaded to Google Cloud

Overview: Citi Bike Operational Data

Size: 20GB

of Records: 139M
Note: Data available as public dataset on Google Cloud

TRANSFORMATIONS & QUERYING



(NEED TO FILL OUT)



Data Access: Full 20GB transformed and merged (Citi Bike + Weather + Geographic) accessed via SQL queries

ANALYSIS & VISUALIZATION

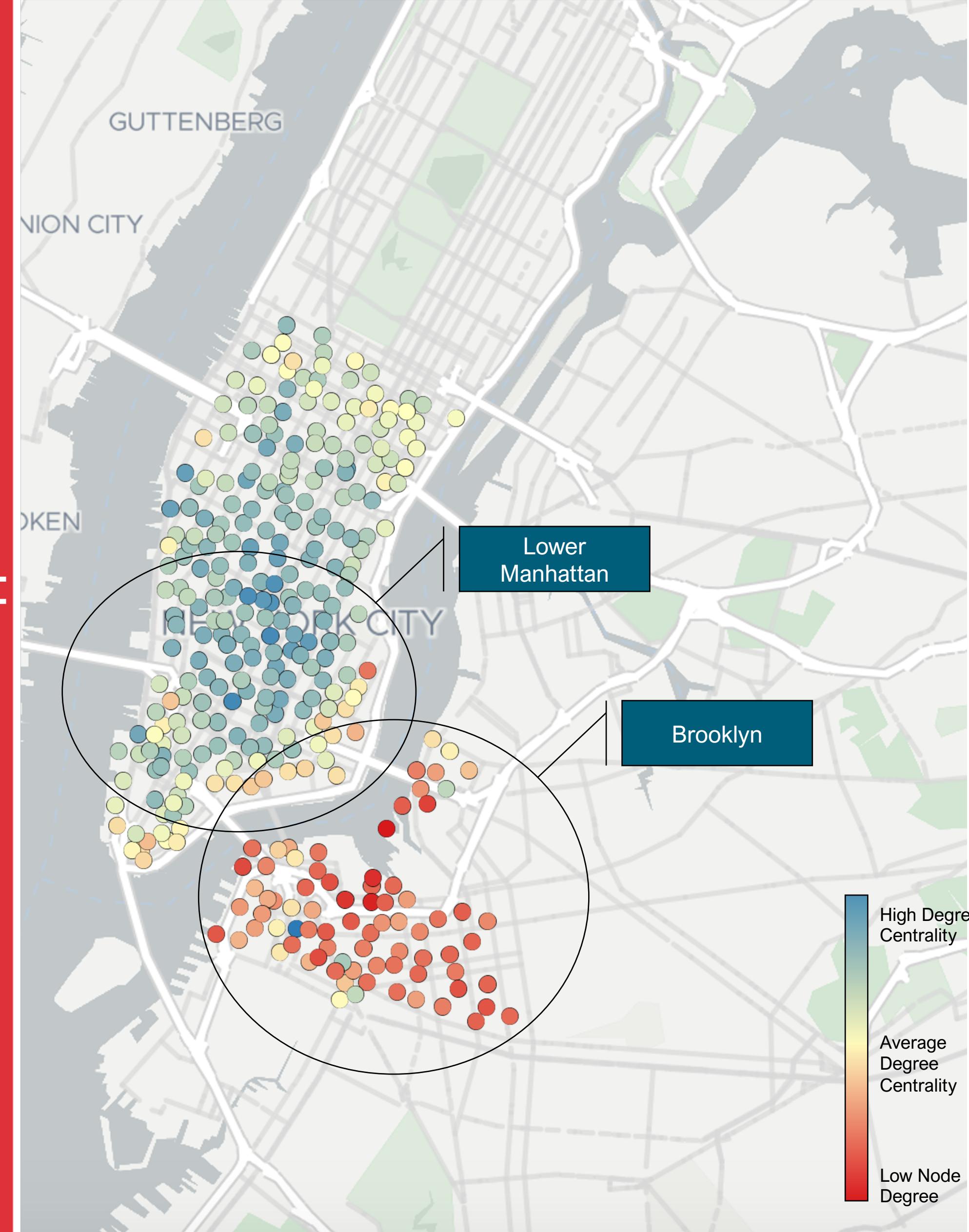


Gephi

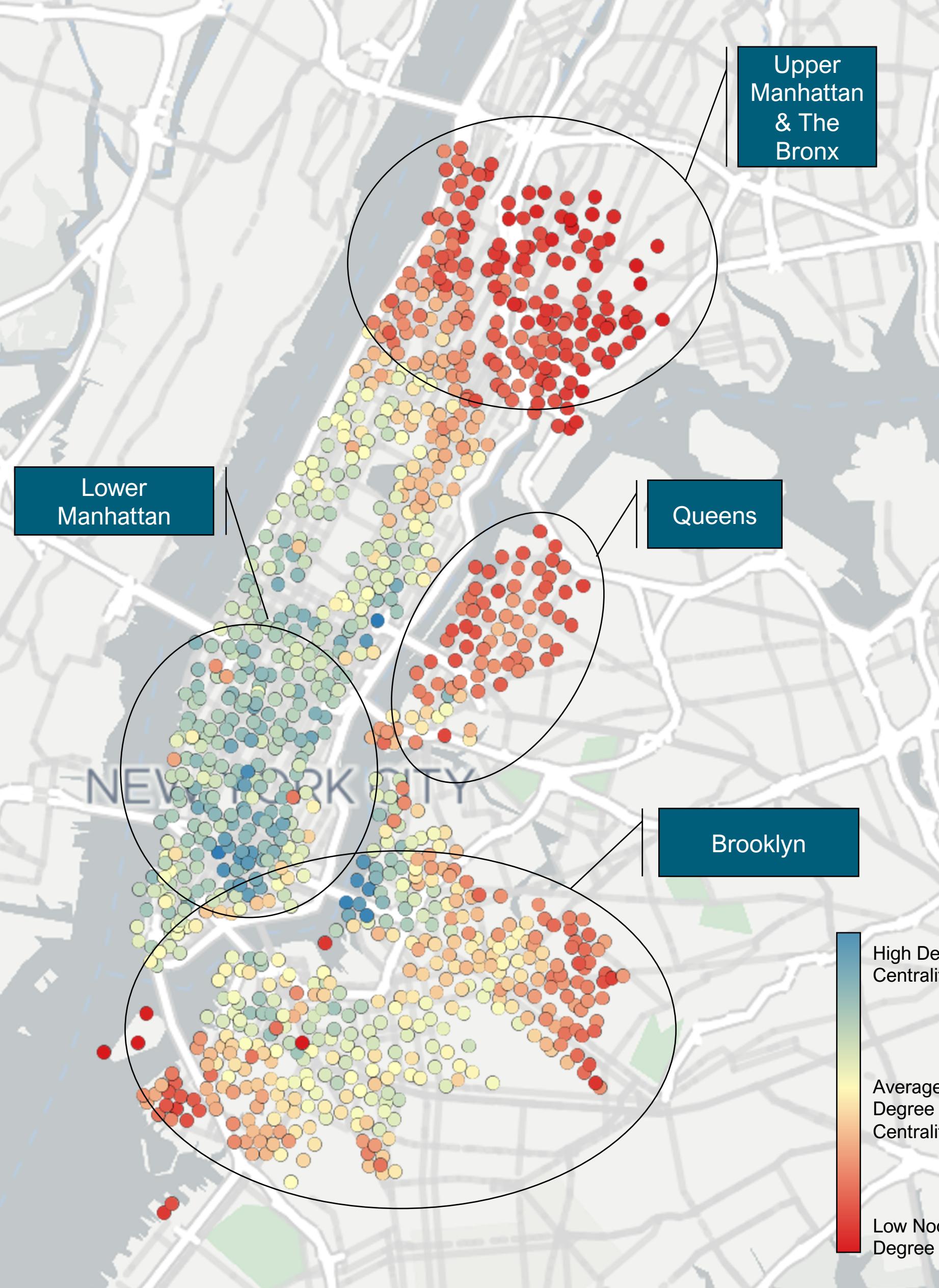


Ridership Insights & Network Analysis

2013 Network Map



2021 Network Map



Network Map Insights

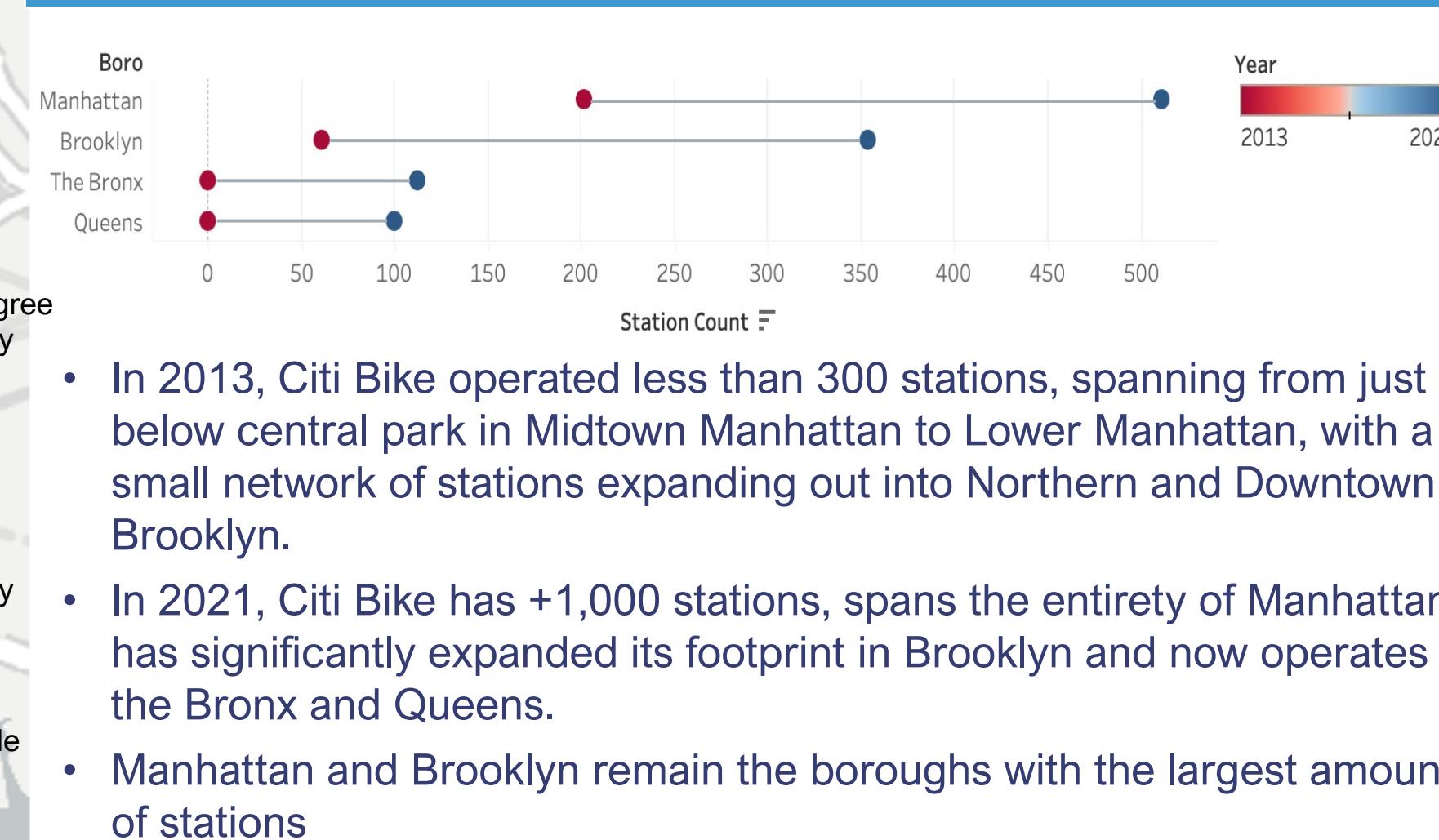
How To Interpret the Maps

- Each bubble represents a Citi Bike station, and its color represents its degree centrality.
- Degree centrality is the measure of the total number of edges connected to a particular node. Essentially, it is the count of bike drop-offs and bike pickups per station, and it provides us with an indication of how "central" a node is within a network.

What Did We Learn?

- Lower Manhattan remains a prime center of activity between 2013 and 2021, with many nodes with high degree centrality (blue nodes).
- We can also observe a cluster of nodes with high degree centrality in Northern Brooklyn.
- Manhattan and Brooklyn have the most active nodes, and they are also the boroughs with the highest density of population.
- There is a clear trend where most activity is occurring in the central area of the network. The degrees with the lowest centrality (red nodes) are all on the "outer ring" of the network.

Station Growth



Factors Impacting Ridership Behavior

Methods for Determining Which Weather Factors Most Impact Rider Behavior

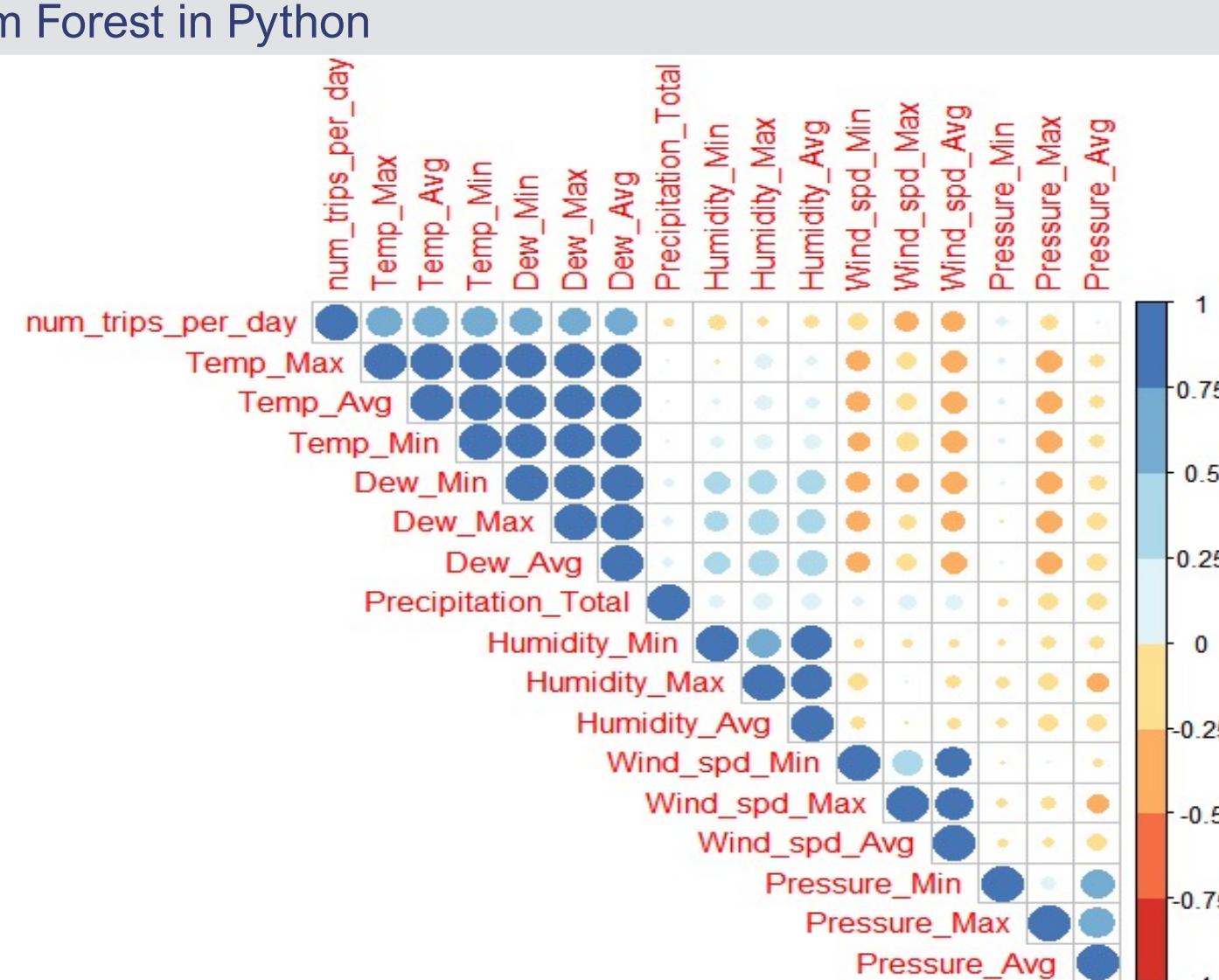
- Correlation Matrix in R (preliminary analysis)
- Linear Regression in R (variable selection via stepwise & p-value significance)
- Random Forest in Python

Pricing, Revenue, and Ridership Changes

Revenue by Membership Type

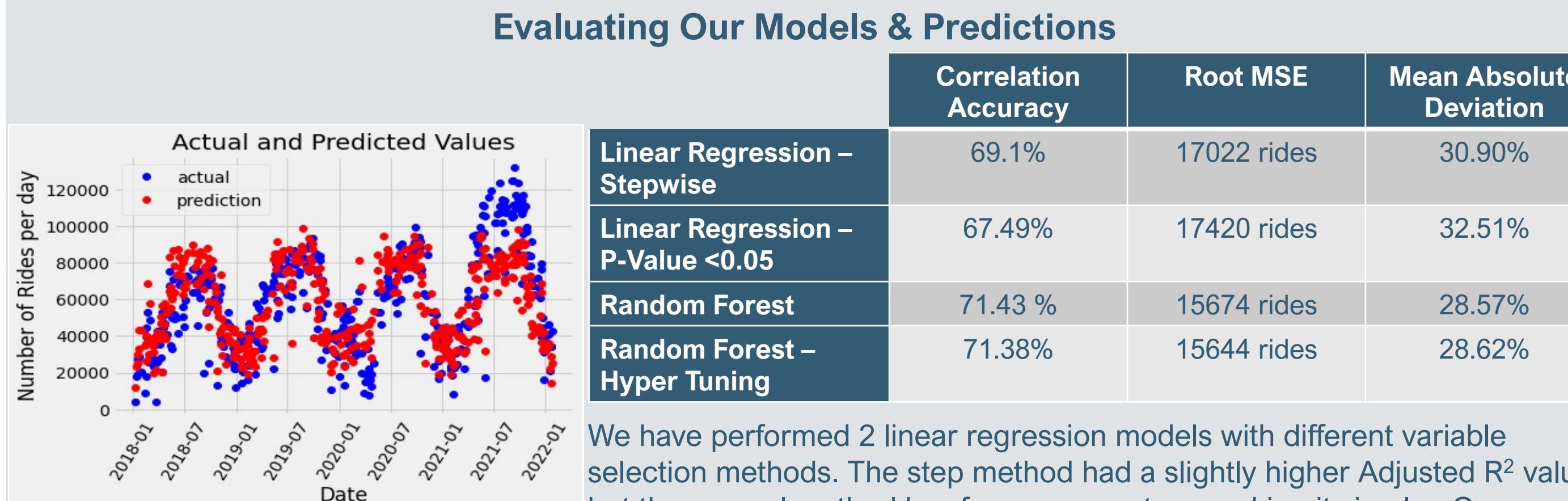
- Annual membership signups have stayed relatively consistent
- Casual member sign-ups have grown significantly, especially in 2020, 2021
- Annual members still account for ~half of membership revenue

Most Significant Factors for Predicting Rider Behavior

**Temperature: Positive Impact****Dew: Positive Impact****Humidity: Slightly Negative Impact****Wind Speed: Negative Impact**

Calculating Price Elasticity of Demand

- Citi Bike has raised its annual membership prices 3 times since 2016: From \$155 to \$169 (March 2018) to \$179 (July 2020) to \$185 (January 2022)
- However, the price increases do not seem to impact demand for annual memberships. Average price elasticity of demand is 0.056
- Demand is inelastic as [% change in quantity]/[% change in price] < 1



We have performed 2 linear regression models with different variable selection methods. The step method had a slightly higher Adjusted R² value but the manual method has fewer parameters, making it simpler. Our Random Forest Model performed better than both linear regression models.

Figure: Random Forest Actual & Predicted

Evaluating Usability

To evaluate the usability of our Citi Bike data visualizations, we conducted a usability survey. Key takeaways:

12 Number of Participants

Rating for Ease of Navigation

XX Rating for Visualization Clarity

XX Overall Rating

All scores are the average provided by the participants, with 1 being the lowest and 5 being the highest. Two key changes we made to our visualizations after the user feedback were: XYZ