Predicting Citi Bike Trip Demand

Rishi Goutam, Srikar Pamidi, James Goudreault

Predicting Citi Bike Trip Demand

Using a neural network model to predict ridership based on time and weather

April 7, 2022 – Rishi Goutam, Srikar Pamidi, James Goudreault

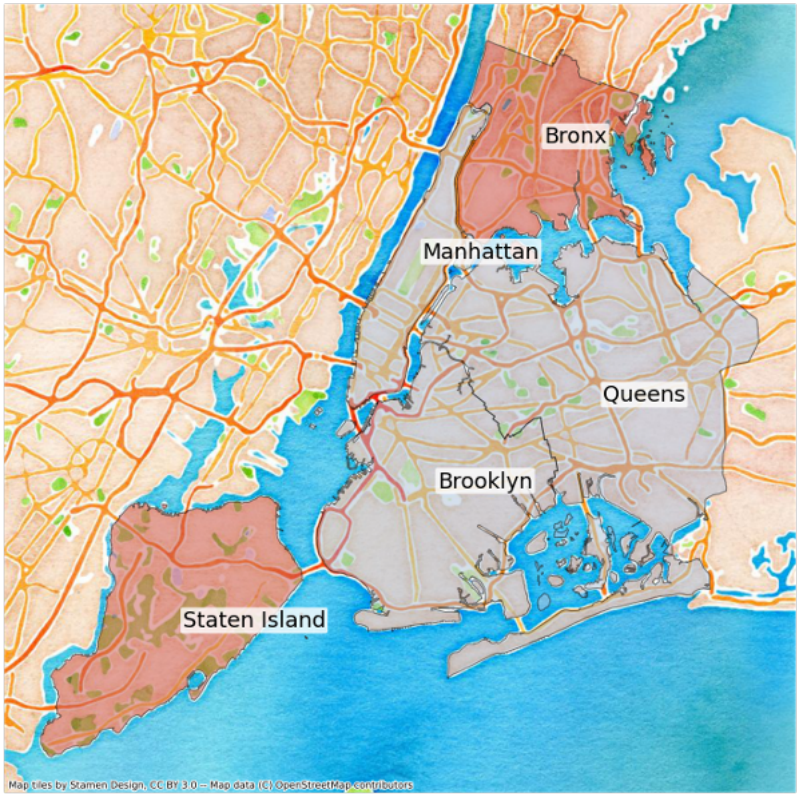
Table of Contents

Predicting Citi Bike Trip Demand

In this article, we show how we analyzed the Citi Bike dataset and built a model to predict ridership based on seasonality and weather.

Introduction

Citi Bike opened in New York City in 2013Citi Bike operates in multiple areas and has been active for several years. We focused our analysis on the primary NYC boroughs Citi Bike operates in (Manhattan, Brooklyn, and Queens). Unless otherwise specified, statistics and graphics in this article are for the year 2019 and these boroughs



and has since grown in ridership, bikes, and bike dock stations. Predicting demand for bikes is important for Motivate, Citi Bike's parent company, in order to both reduce operating costs and increase ridership. Costs are incurred by having under-utilized bikes on the streets by wear-and-tear on bikes due to exposure to the elements or other forms of damage, so it is necessary to warehouse bikes if they are not serving riders. However, having too few a number of bikes available leads to a poor customer experience and loss of

revenue due to dissatisfied customers.

We aim to first understand what features drive demand and then to create a predictive model. Finally, we compare our model against the actual number of trips to evaluate its usefulness.

Analyzing the data

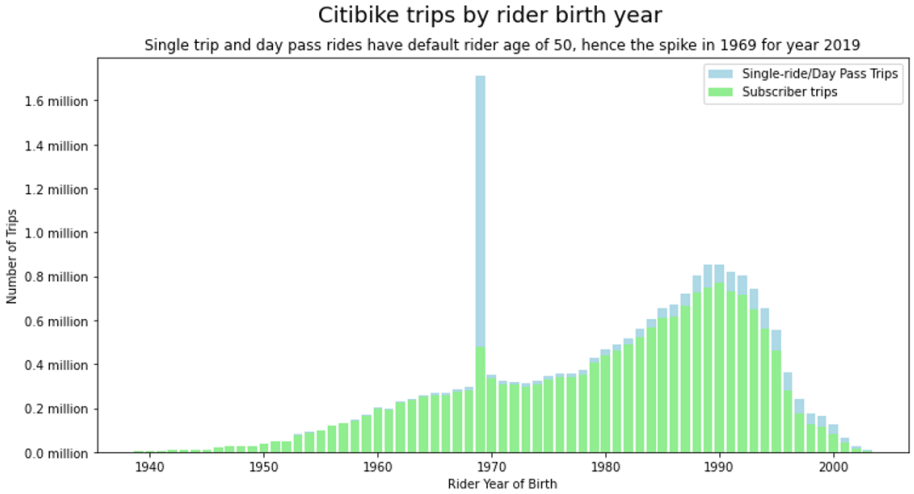
We focused our exploratory data analysis on:

- Rider demographics
- Citi Bike's growth and resilience in the face of COVID-19
- Time (Seasonality)
- Weather

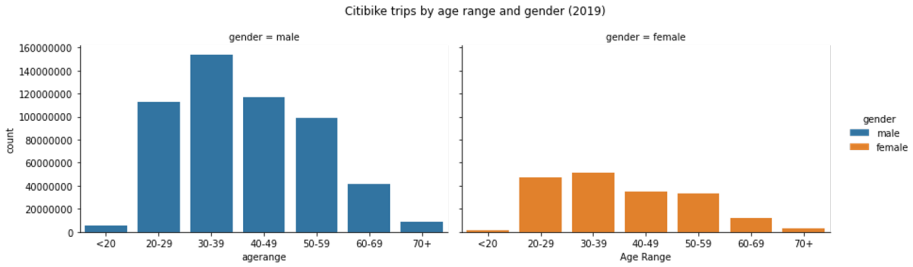
And determined that time and some weather conditions would make for good predictors for a time-series model. Here is that analysis.

Demographics

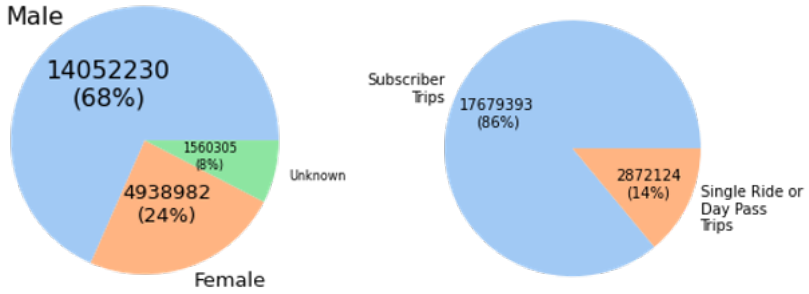
By age, most riders are between 20 and 40 years and are mostly male. Citi Bike has two classes of riders–annual subscribers and single-ride or day pass purchasers. We see a default age of 50 years for riders purchasing one-off trips or passes in the chart below.



TODO FIX THE Y AXIS HERE:

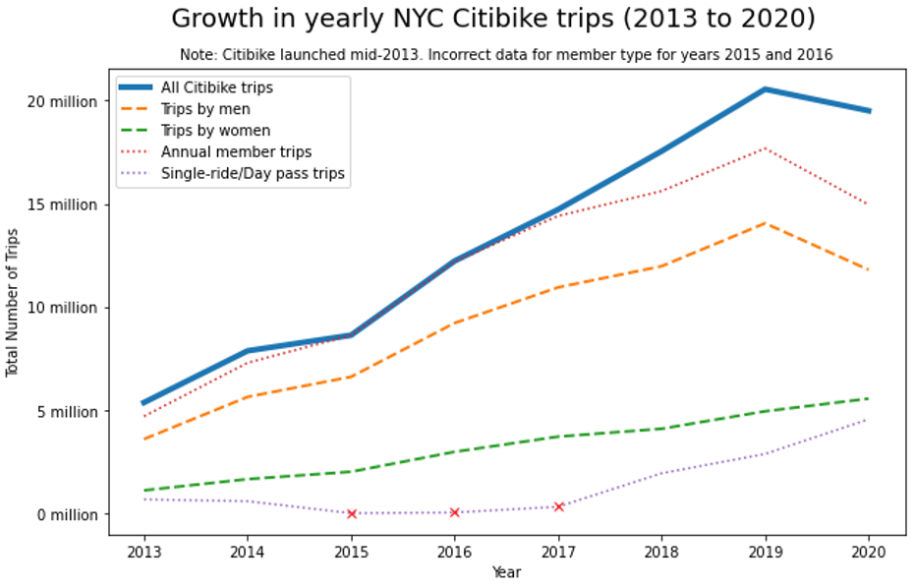


We can see the gender and customer type distribution through the much-maligned pie chart

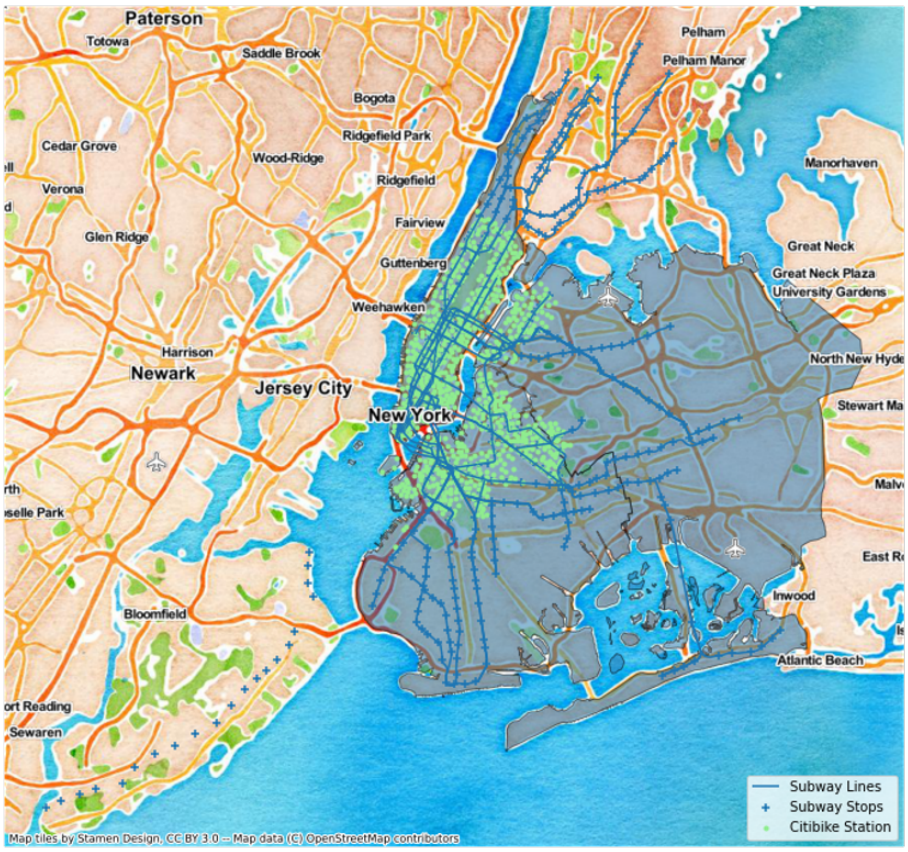


Growth and Resilience of Citi Bike

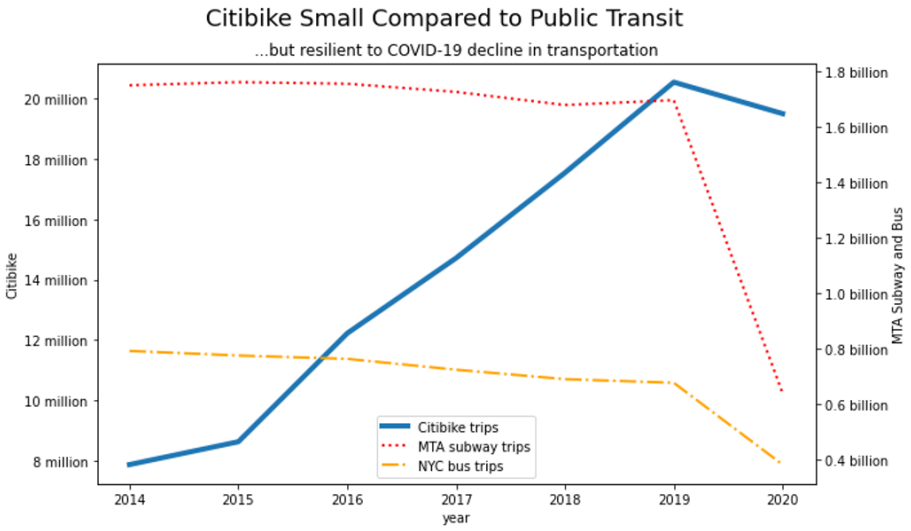
As the number of Citi Bike trips grows, operational efficiency is more important to company finances. In addition, there is need for accurately predicting demand and rebalancing stations effectively



Bike stations appear to expand along subway lines…potential for further expansion into Brooklyn, the Bronx, or Queens?



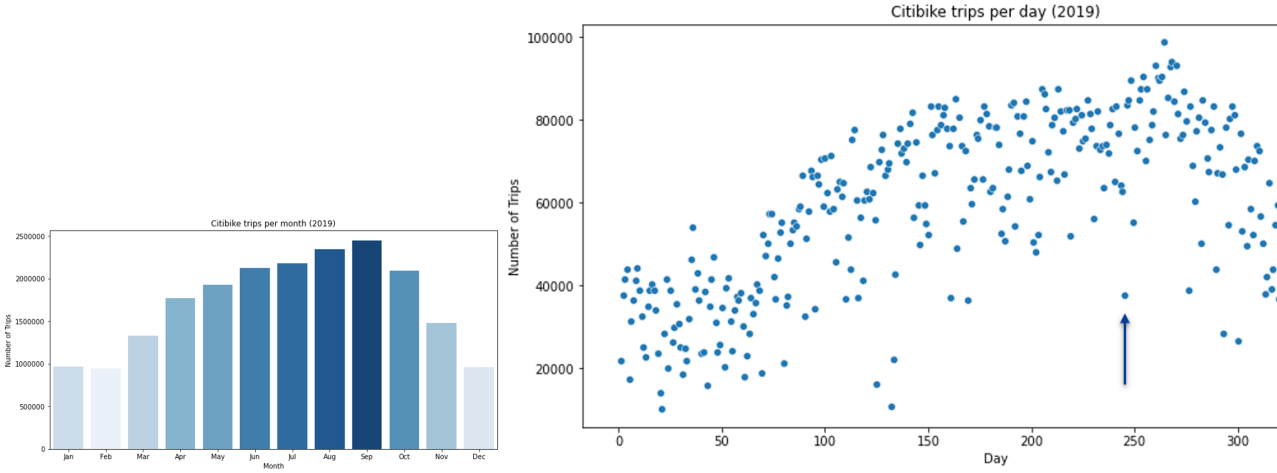
While Citi Bike is not used as much as mass transit in NYC, it offered a way for city residents to move from A to B during the pandemic that wasn't in an enclosed space…perhaps this is why it didn't see as sharp a drop in demand compared to the NYC subway or buses



Temporal Analysis

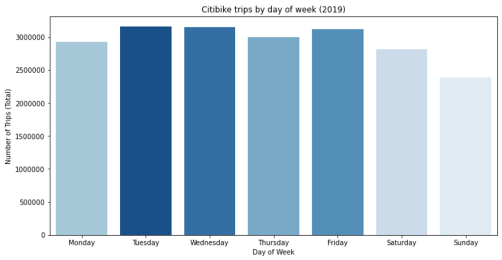
We see increased usage in the summer as one might expect, but not all summer days prove to have high counts.

Labor Day 2019 shows reduced demand...and weather might also play an effect. We examine that later.

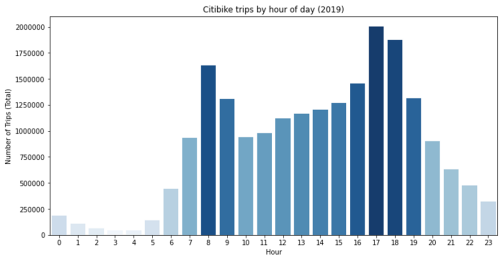


{ width=250px }

Surprisingly, weekends, especially Sunday, seem to have lower trip counts on average than weekday. Sunday truly is the day of rest.

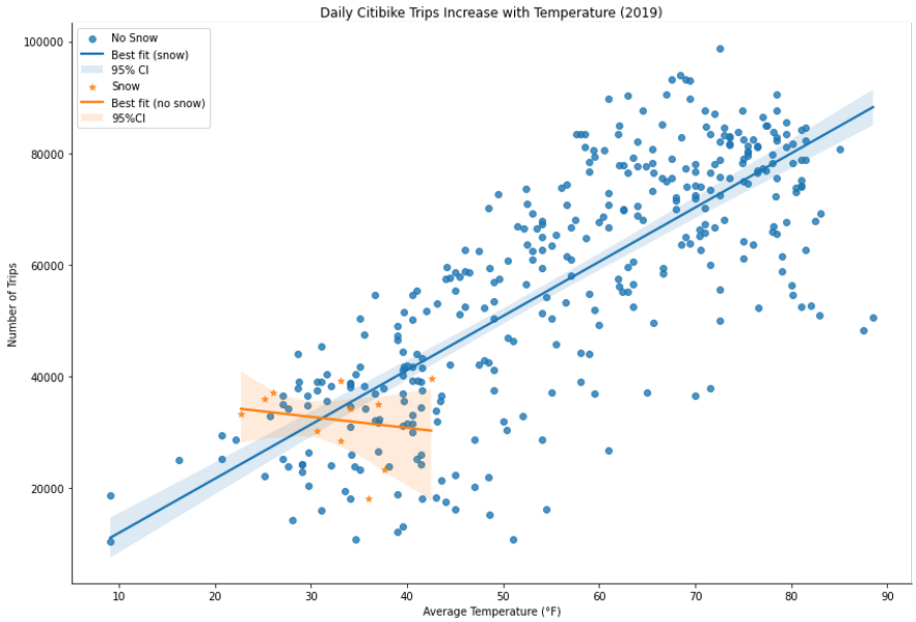


Looks like commuters make up a bulk of trips given the high trip counts around 8am and 5pm. This also explains the reduced number of trips on weekends.



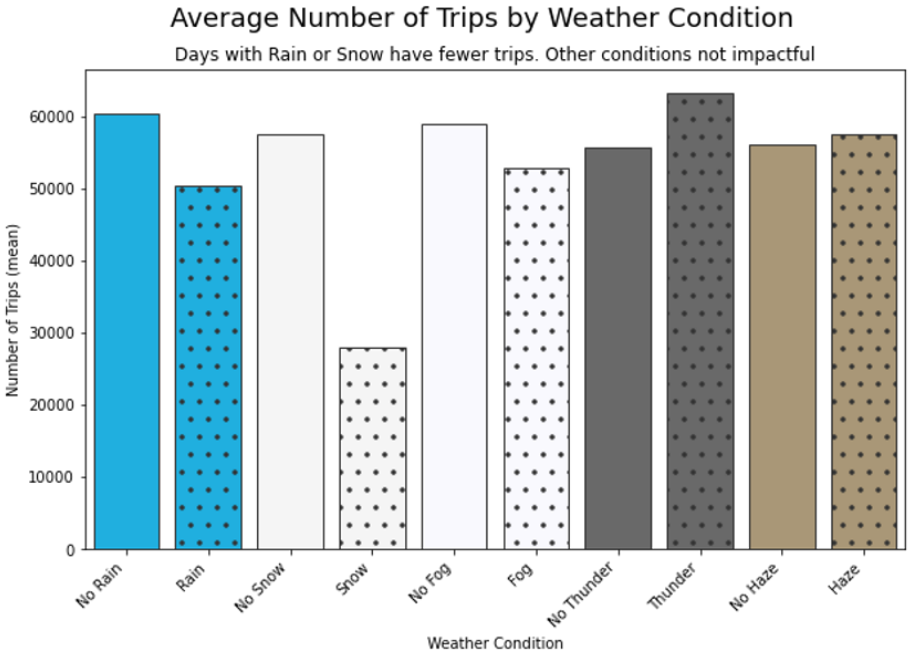
Weather

We began by looking at daily average temperature and found that trip demand is highly correlated with it. We can use this as a model predictor!

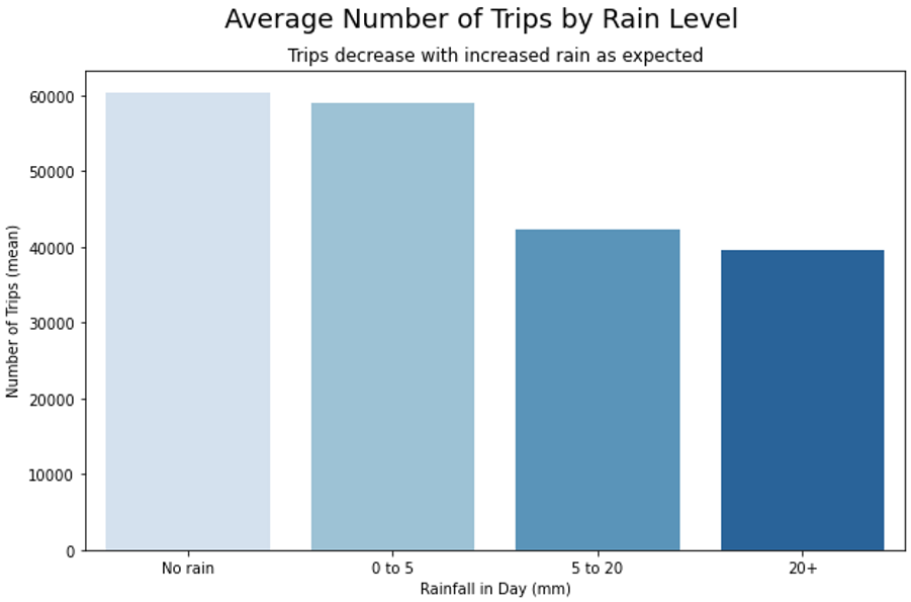


We wanted to investigate whether weather conditions might have an effect on number of trips…however, only saw decrease for days with precipitation (rain/snow)

Conditions like fog (all types), thunder, or haze were low in terms of number of days and did not have as strong an effect (although we were surprised by the positive effect)



Digging deeper into rain, we wanted to see if the amount of rain mattered…and it does! (as expected)



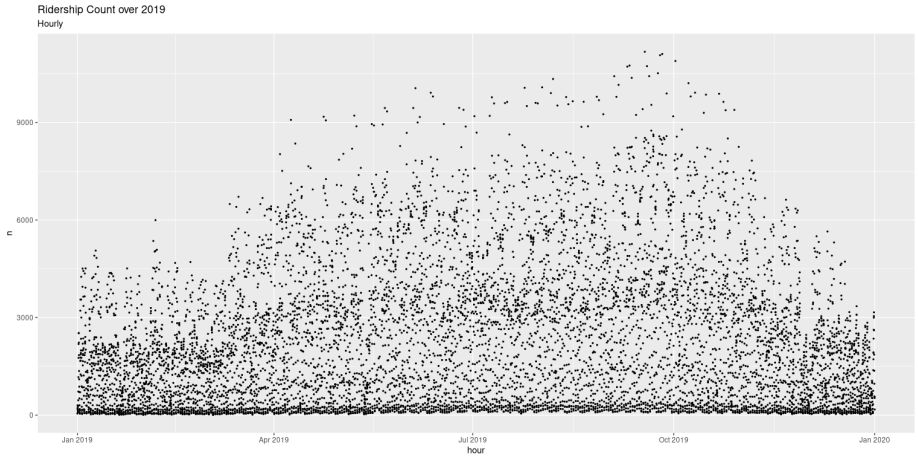
Based on this analysis, we decided to create a model incorporating time, average temperature, and amount of precipitation in order to predict the number of trips.

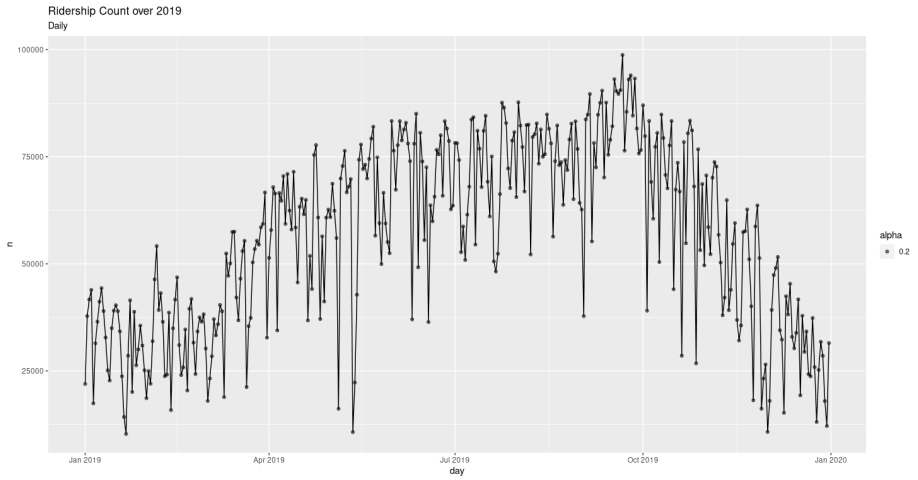
Trip Demand Prediction Models

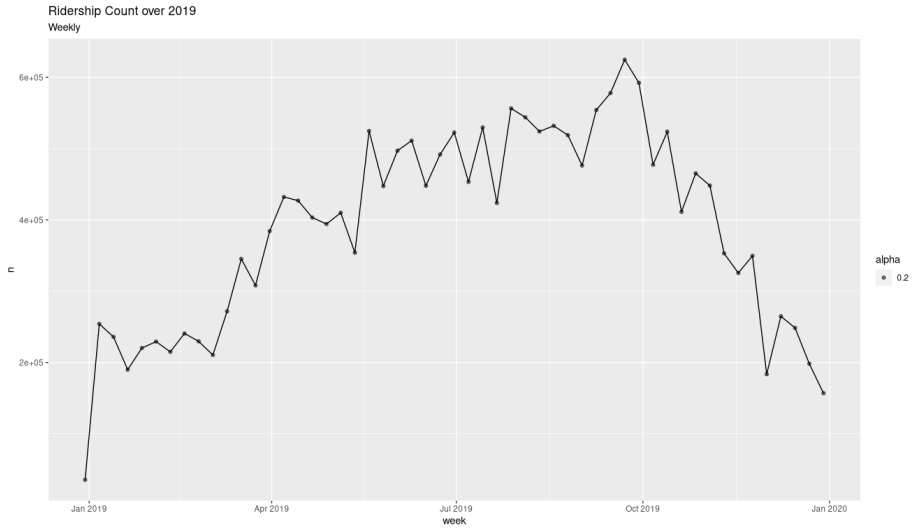
We attempted two models, the first of our models is the traditional SARIMA model, the second was a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). In this, we further distinguish the models by the time resolution, and whether or not the model was including weather data (i.e. had multidimensional inputs). Notably absent from our treatment is a vectorized SARIMA model or SVARIMA; however, this was due to time considerations and the fact that LSTM-RNN models on average demonstrate far better performance on multidimensional time series than traditional models and approaches.

The population data is found in in the seconds resolution; however, at this resolution, observations are not continously present and we have long gaps for many seconds of the year with zero ridership. We take a look at the hourly ridership data from 2019; whatever model we create for one year we can easily generalize to multiple years.

As a preliminary assesment of the data, we look at the ridership over the year along different timescales–in doing so we can begin to understand trends and seasonality relationships within the day. Changing the scale of our data, as we will see, is akin to passing the "wave" of our time series through a low-pass filter, giving us lower frequency relationships. Therefore, if there is high seasonality at low timescale, this will be erased as we increase our timescale and aggregate over larger steps.


Ridership Count over 2019
Hourly
9000
6000
3000
0
n
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Jan 2020
hour


Ridership Count over 2019
Daily
100000
75000
50000
25000
n
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Jan 2020
day
alpha
0.2



We can see from the above graphs that weekly resolution is far too sparse to caputure meaningful relationships. Therefore, we would like to build models that predict at the Hourly timescale if we can, and if not, then use the Daily timescale

At the sub hourly timescale, the data became too unweildy and noisy for a years worth, let alone for the many years of data Citibike has available. However in future extensions of this project we would like to take a second level resolution for one week for one station and predict the ridership at that level.

Our models were thus:

1. Hourly SARIMA, which did not converge to parameters.
2. Daily SARIMA, which converged to parameters, but had low resolution.
3. Daily LSTM-RNN, without weather, which had high RMS error.
4. Daily LSTM-RNN with weather, which had reduced RMS error.
5. Hourly LSTM-RNN without weather, which had great success.
6. Hourly LSTM-RNN for a specific station, with weather.

We compared these models by producing the Daily and Hourly RMSE and comparing them to find the RMSE minimizing model.

Seasonal-differencing autoregressive integrated moving average (SARIMA)

Hourly SARIMA: To begin, we start with a model that does not account for the weather, or uses deep learning, only traditional statistical metrics based on past data.

Differencing

ACF/PACF

Box-Ljung

Seasonal Differencing

Seasonal ACF/PACF

Fitting

- failure to converge to parameters
- hypothesis and next step for model

Daily SARIMA:

- diff, ACF, PACF, Box-Ljung, seasonal diff, SACF, SPACF, and fitting
- Converges to parameters
- Prediction and visualization
- Daily RMSE, avg hourly RMSE
- Hypothesis and next step for model

Long short-term memory recurrent neural network (LSTM)

Intro to RNN and LSTMRNN

Moving from SARIMA to LSTMRNN

Timeseries and Backtesting

Daily LSTMRNN,

- EDA and visuals
- Seasonality, ACF, periodicity,
- Structure of Layers and LSTMRNN
- Backtesting and Visualization
- Predictions and visualization
- Daily RMSE and av hourly RMSE
- Hypothesis and next step to improve

Daily w/weather

- Weather EDA, why we think to add temp and precipitation.
- Seasonality ACF periodicity,
- Layerstructure and adding more dimesions for lag
- Predictions and Vis.
- Daily RMSE and av hourly RMSE
- Hypothesis and next step to improve

Hourly LSTMRNN without weather, great success.

- Hourly EDA, for busiest station.
- Seasonality ACF periodicity,
- Layerstructure and adding more dimesions for lag
- Predictions and Vis.
- Daily RMSE and av hourly RMSE
- Hypothesis and next step to improve

Hourly LSTMRNN for a specific station, with weather, great success.

- Hourly Weather EDA, why we think to add temp and precipitation.
- Seasonality ACF periodicity,
- Layerstructure and adding more dimesions for lag
- Predictions and Vis.
- Daily RMSE and av hourly RMSE
- Hypothesis and next step to improve

Conclusion

We found such and such…

Rebalance Operations

What is rebalancing? Why do we care?

Citi Bike faces a perpetual dilema - how can they ensure there are enough bikes in the right places to satisfy rider demand? Will the natural flow of bikes from rider be enough?

Consider the flow of bikes through a popular commuter station on a weekday:

The blue line represents the cumulative change in quantity of bikes at the station throughout the day. The gray and red lines represent the typical and maximum high-volume dock station capacities respectively.

Even if the station were to naturally start each day fully stocked, the station would be depleted halfway through morning commuter hours with significant remaining demand. Someone needs to bring in more bikes so all riders can get to work - and conversely shuffle them away at the end of the day…

Simply put, rebalancing is the manual movement of bikes from one station to another to: - ensure there are sufficient bikes at each dock station to satisfy predicted demand - ensure there is space at ending dock station to receive incoming bikes - minimize number of underutilized bikes sitting at unpopular locations

How does Citi Bike Rebalance?

Citi Bike employs several methods to manage rebalancing needs:

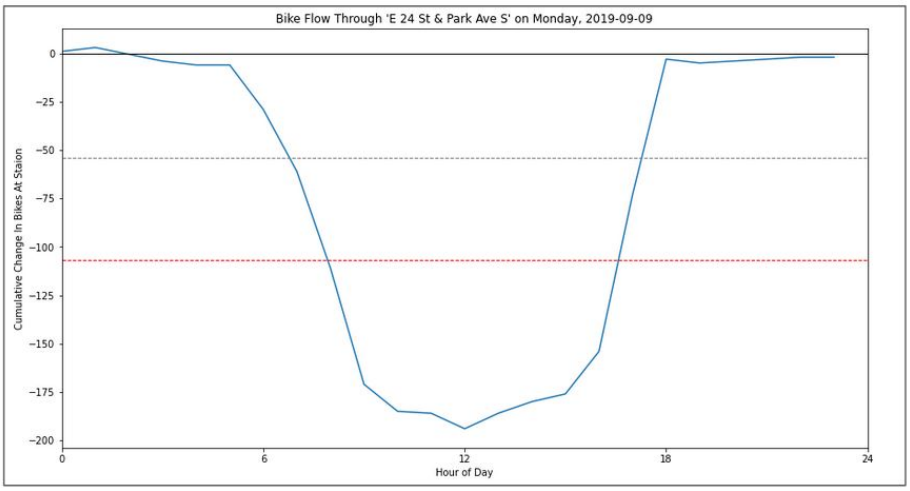


Figure 1: Commuter-Station-Bike-Flow

- *Valet Service* - team members staff popular stations to manage incoming and outgoing bikes, artificially elevating station capacity and temporarily alleviating rush-hour problems
- *Bike Trains* - team members ride ebikes towing carriages of 12-16 bikes, ideal for rebalances in and around tight neghborhood streets
- *Motorized Vehicles* - higher capacity vans operate 24/7 to keep bikes moving to and from the most popular stations
- *Bike Angels* - Launched in 2018, this program grants Citi Bike riders (non-employees) traveling along in-demand rebalance routes a free trip and rewards points

Identifying Rebalance Movements

The Citi Bike dataset does not contain information on individual bike rebalances, nor is that generally available online except as aggregated data in monthly reports. So how can someone tell when a bike was rebalanced, and to where?

The easiest method to do so is (for a given bike) compare the starting station for each trip with the ending station of the previous trip. If the bike appears to have teleported from one station to another between trips, it most likely was rebalanced!

General Rebalance Statistics

Predictably, the number of rebalance operations increases with number of rides. The large drop after 2017 coincides with the introduction of the Bike Angels program and clearly demonstrates how successful this clever initiative is.

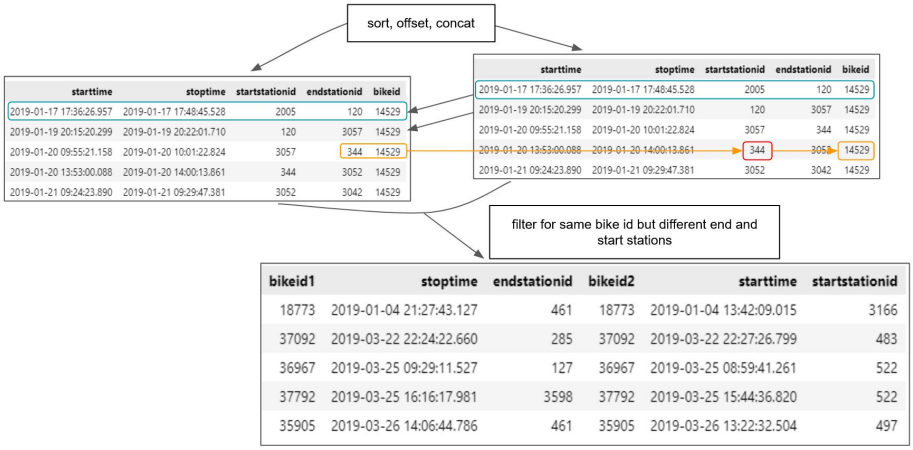
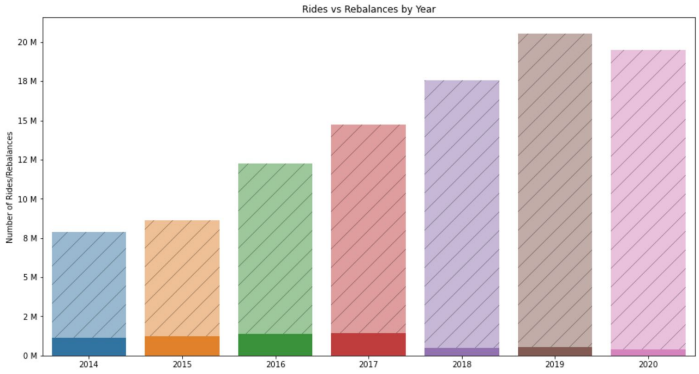
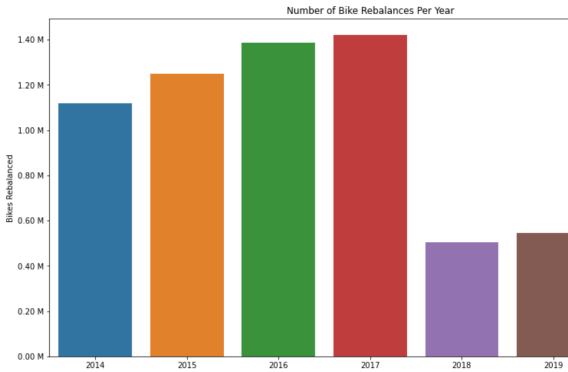
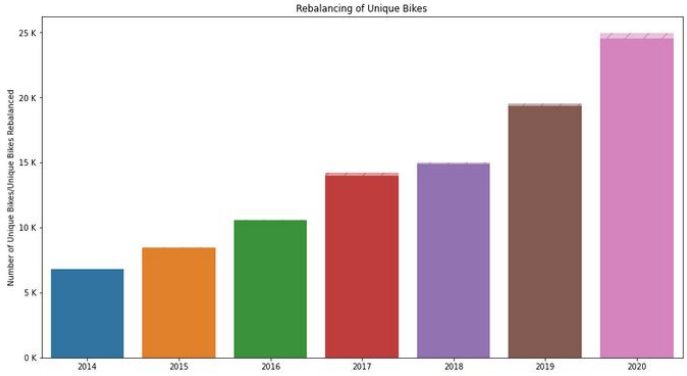


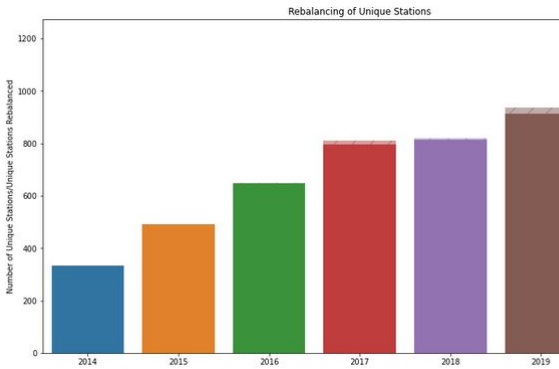
Figure 2: Identifying-Rebalances



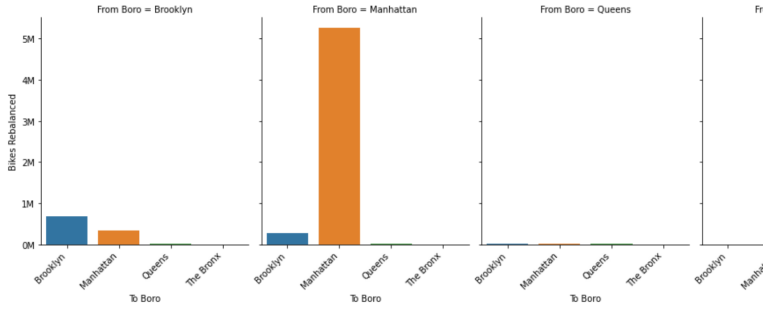


For each year, almost every unique bike in service was rebalanced at least once and almost every station was invovled in rebalancing operations.



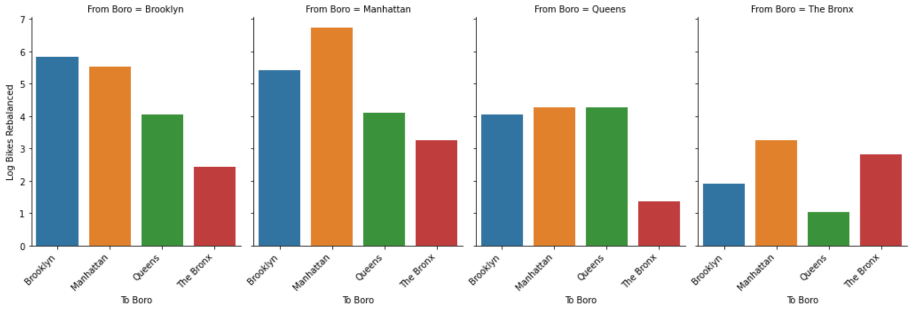


The vast majority of rides occurred in Manhattan, so it is not suprising to see

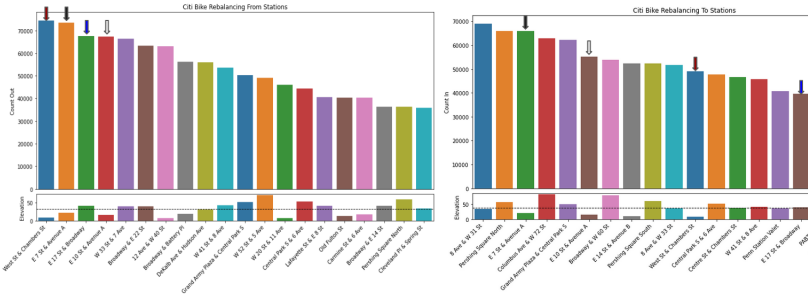


that is where most rebalance activity occurs.

It is important to note that bikes are rebalaced around and across all Boros - confirming Citi Bike actively works to satisfy all rider demand.



Individual stations with the highest rebalanace activity are mostly commuter stations. Many have both a high number of bikes moved to and from them based



on commuter demands as shown above.

Rebalance Routes

We can glean more information about rebalance operations if we look not just at bikes moving to and from individual stations, but consider pairwise movement between two stations.

Not only is the majority of rebalance activity focused on a small number of stations in Manhattan, most of it occurs between a small number of station pairs. This makes sense in the context of commuter bike flow, where there is

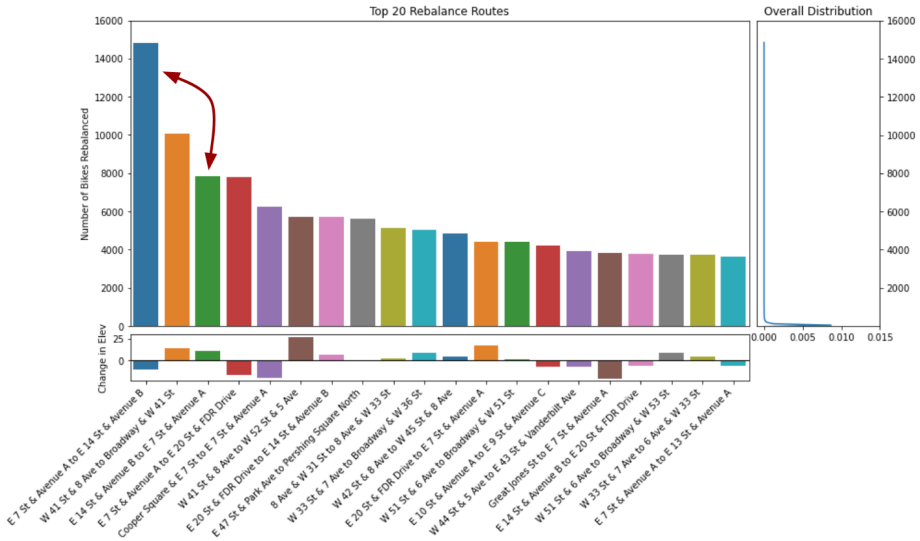


Figure 3: Rebalance-Routes

a need for a constant flow of bikes away from destination stations to origin stations to keep up with demand.

Notice that some of the top routes are just the reverse direction of others. Visualizing bi-directional flow between station pairs further emphasizes how critical rebalance operations are between a small number of station pairs.

It should be noted that rebalance operations are probably more nuanced then bikes being moved along static 'routes' - it's very possible a van will stop at several stations along a circuit - but this aggregate analysis still helps us understand how much effort and expense is required to support the highest volume stations.

Perhaps the most intersting way to look at station related rebalance data is visualized on a map - check out our dash app to see this for 2019!

Rebalance Timing

We don't know precisely when a bike was rebalanced, only when a trip started after the bike was rebalanced These time estimates (in hours) all reflect the maximum time a rebalance operation could take. We can conclusively say a good part of Citi Bike rebalancing is done to alleviate immediate demand, such as during commutes.

Again, it appears that the Bike Angels program and Valet Service had profound

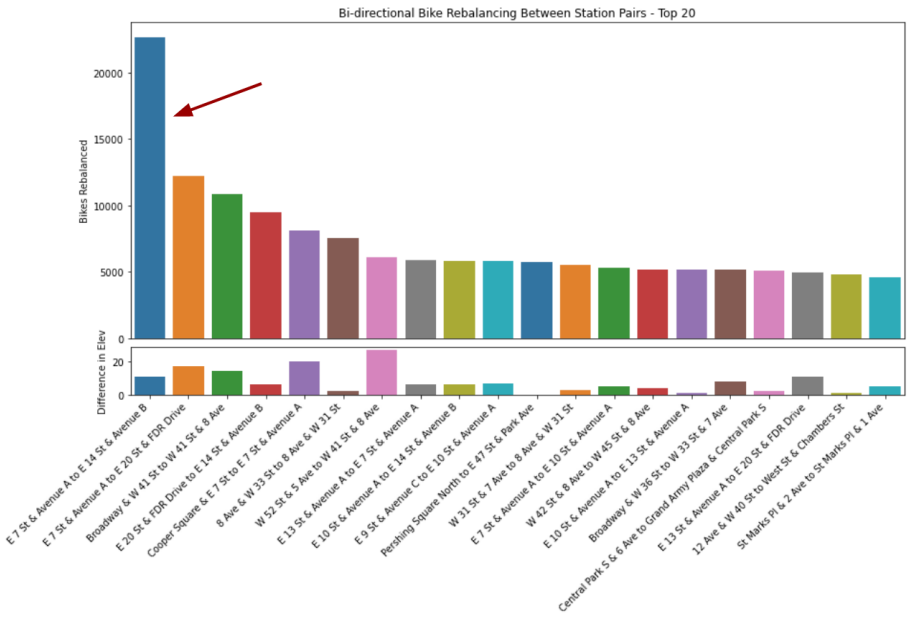


Figure 4: Bi-Directional-Rebalance-Routes

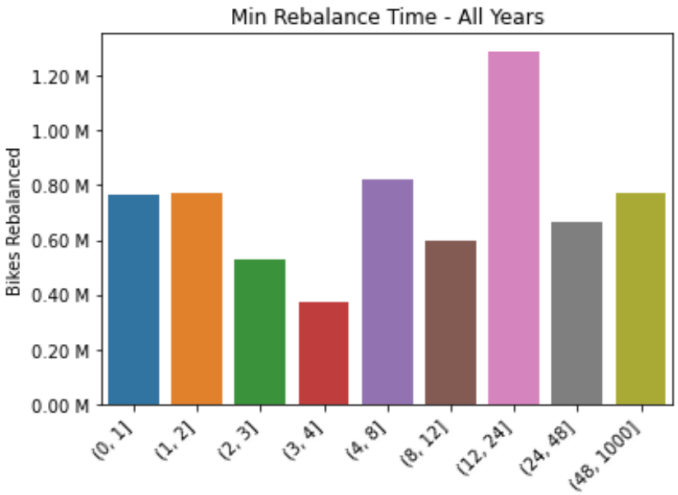
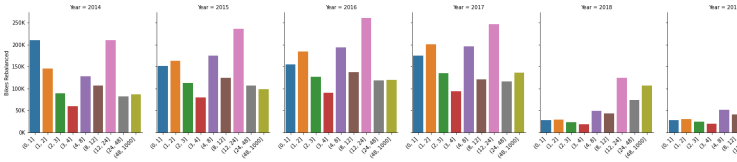
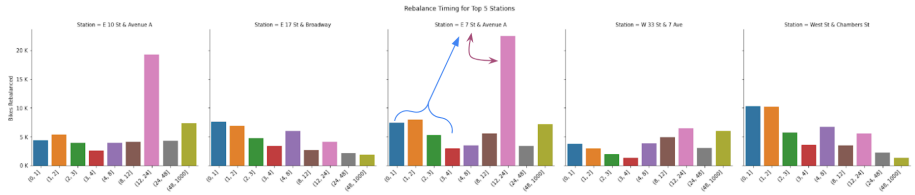


Figure 5: Rebalance-Timing

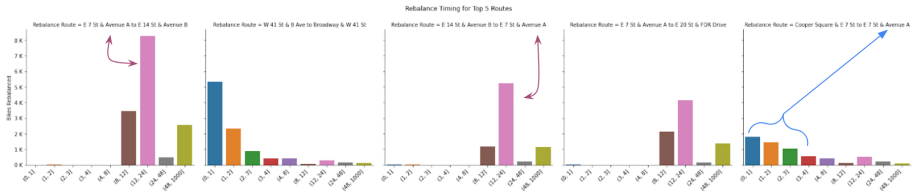


impacts on the need for immediate rebalances.

It appears that there are different reblance strategies used for popular stations. Some seem to have the majority of rebalances done overnight while others seem to be more frequently rebalanced for immediate use. Both modes are necessary to keep up with demand.



Investigating timing for rebalance routes provides deeper insight, where we can see some routes are used to stock stations overnight for the next day's commute and others are used to satisfy immediate demand.



Seasonality

As expected, rebalance activity tracks with ridership where most of the activity occurs in the summer months. The Bike Angel program launched in 2018 had a profound effect on flattening this trend compared to previous years.

Similarly, post Bike Angel rebalance activity tracks with ridership where most

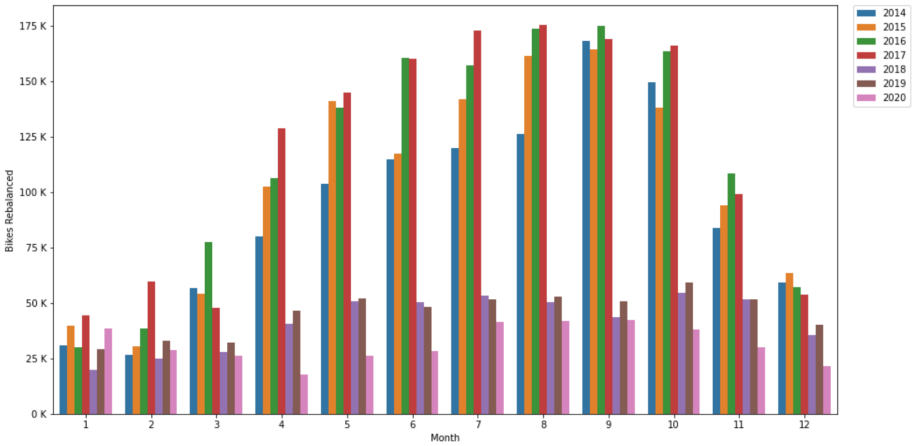
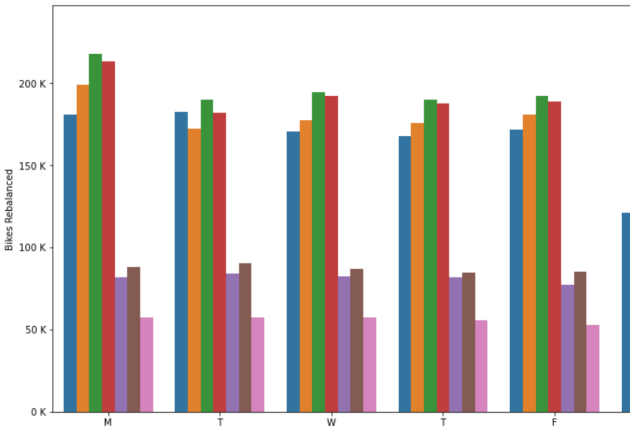


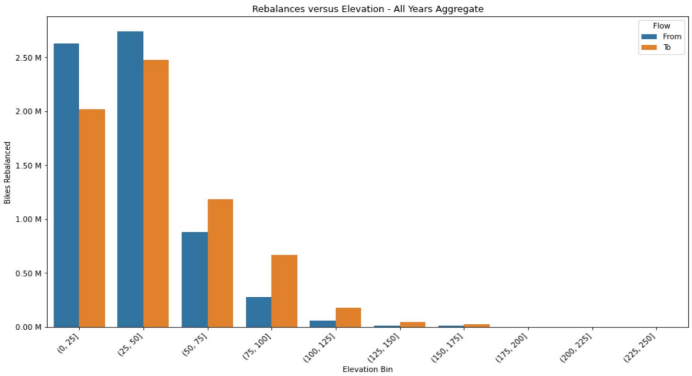
Figure 6: Annual-Rebalance-Season

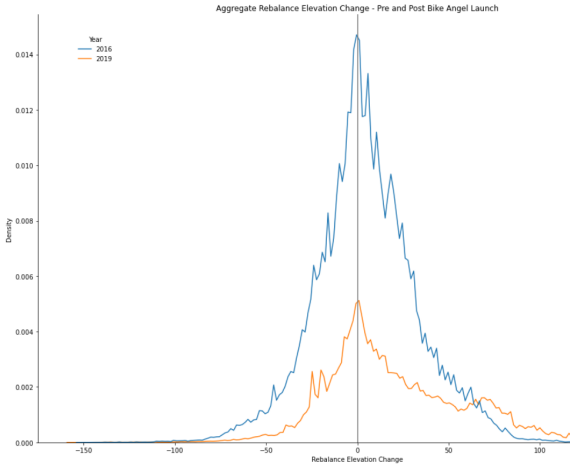


acitivity occurs during the week supporting commuters.

Bikers Dislike Going Uphill

In aggregate more bikes were rebalanced from lower elevation stations than to, and more bikes were rebalanced to higher elevation stations than from. This suggests bikers who ride downhill often find other modes of transport for return trips.





This becomes even more apparent when looking at a density plot of years pre and post launch of the Bike Angel program. In 2016, the difference between station elevations for rebalance movements is somewhat normally distributed. In 2019, there is a long right tail indicating significantly more bikes are rebalanced to higher elevations - even the Bike Angels aren't willing to climb hills.

Future Analysis?

There are nearly endless ways to dive into ridership and rebalance data to understand needs and motivations of bikeshare participants or the general movement of individuals around New York. This analysis focused mostly on data aggregated over the years available (2014-2000) which helps us interpret rebalance operations as a whole. Future work could include:

- Focusing on a specific year, season, or days to understand and predict usage and rebalance needs with respect to changes in overall ridership and events such as holidays and professional sports games
- Comparing rides and rebalance statistics across years as docking stations are introduced, moved, or retired to evaluate past decisions and inform future strategy/goals
- Developing a prective model to inform Citi Bike if a dock stations is on track to be depleted or filled and needs immediate rebalancing outside of normally scheduled operations

Visualizing Bike Stations and Rebalances

look at our dash app!