New York City's Bikeshare System

Citi Bike Analysis: Visualizing Trends & Predicting Ridership

FINAL REPORT

Team: 46
Brent Brewington
Stephanie Chueh
Roshni Mahtani
Kevin Schneider

Introduction

Citi Bike is a public bikesharing system that operates in New York City (and a few areas in New Jersey). Since launching in 2013, with 6000 bikes and around 300 stations, Citi Bike has grown to operate 24,500 bikes and over 1,500 stations. This is timely as New York City's population has grown by nearly 8% since 2010, and biking has grown in popularity as a means for transportation, especially in light of recent issues with the city's public transit (Correal 2021).

Our goal is to provide holistic insights and visualizations for Citi Bike trends and factors impacting ridership behavior to empower city and transit planning. Citi Bike has made system data available, and this is the foundation of our analysis. The company publishes monthly operating reports that include information such as the number of annual membership sign ups and renewals as well as monthly ridership data in csv format. This data includes information on start and end stations and their coordinates, ride duration, and rider type (member or not).

Problem Definition

We will perform in-depth data analysis of the Citi Bike data, incorporating additional analysis by pulling from external sources such as weather and location (geocode) to identify their impact on Citi Bike usage. We will also review Citi Bike membership, revenue, and pricing changes.
We will use this data to derive insights around predicting ridership patterns (e.g., impact of weather on bike usage), analyzing trip types (e.g., most popular start and end points, etc.), understanding users' price elasticity of demand and trends in ridership type (annual members vs. casual riders), and mapping the most popular stations and neighborhoods.

Additionally, we will build a network map of bike stations to trace relationships between stations and neighborhoods. We will leverage visualization tools to create interactive dashboards that show trends over time as well as the impacts of the various factors discussed. By analyzing and clustering trip types and merging additional datasets, we aim to enrich the current understanding of Citi Bike ridership, identify new trends in rider behavior, and provide recommendations for future areas of growth.

Literature Survey

Our team's literature review focused on two key themes: examining the impact of various factors on bike ridership and leveraging analytical models to predict behavior and optimize the system.

What Impacts Bike Ridership? Teixeira (2020) is a case study analyzing the link between bike sharing and subway use during COVID-19. It concluded that bike share ridership was more resilient to the pandemic than subway ridership. There are other key factors that could have impacted ridership to consider, such as new member sign-ups. Wang (2019) examined if different environmental factors had a different impact for male vs. female ridership. Analysis was done on a month of data; we intend to use a wider time range. We will consider these findings as we seek to understand demographics of Citi Bike ridership, trends in bike travel paths, and docking station location. Faghih-Imani (2020) explored factors that influence bike ridership and membership, such as proximity to a station and factors affecting routes like job and restaurant density, through surveys and statistical models. This provides useful background information and feature engineering ideas. The article focuses tuning and optimization around a unique predictive model. An (2019) studied the effect of weather on cycling in NYC, controlling for bulk, environmental, and temporal factors over 12 months. It found that weather impacts cycling rates more than topography, infrastructure, land use mix, calendar events, and peaks. We will be considering weather as a key factor in our analysis. This paper only used 12 months of data and was done pre-COVID, so we will consider how well the effects extrapolate to 2022 onward.

How Can We Optimize the Citi Bike Ecosystem? Faghih-Imani (2016) examined how arrivals and departures at nearby Citi Bike docking stations influenced each other - they found significant dependence on nearby stations. We'd like to incorporate this factor to track trends in station popularity and to explore potential locations for future docking stations. This paper only looked at one month of data, September 2013. For our analysis, we plan to use a more comprehensive data set. Yanocha (2018) contains domain information about the Bikeshare system, which will be useful for knowing the requirements, operations, and considerations of a Bikeshare system to further enhance its utility and experience with data analytics. This guide will serve as a validation tool for the recommendations developed from our analysis. In Hamad (2021), the researchers built several predictive models to forecast bike usage and assess impact of weather. The authors share similar research objectives (understanding ridership habits, prediction) with our aims. The article leans heavily into tuning and optimization behind each predictive model, which is not our focus. In Ford (2019), the authors analyze commuting trends in the Financial District. This is useful for subsetting the data to make assumptions about user types - in this case, commuters. This research focuses on a small subset: commuters in a small region of Manhattan, while our focus is more general. Bouveyron (2015) identified bike sharing subsystems through mixture models - we could use this approach for detection of ridership shifts due to pricing changes, as well as identifying subsystems within and between boroughs. It will be interesting to explore the funFEM model against other time series alternatives like FB Prophet & Uber's Orbit. Faghih-Imani (2017) compared the efficiency of bikeshare vs. taxi trips in New York City, and established a trade-off factor of trip distance vs. station bike capacity (1 km to 19 bikes). O'Mahony (2015) proposed an optimization method for rebalancing bike placement to account for peaks. While rebalancing is not the focus of our project, this paper provides insight into rush hour traffic and bike demand forecasting. Thu (2017) uses Weighted K-Nearest-Neighbor regression and Artificial Neural Networks to predict the weights of several factors on bike demand at various time slots. Given that we will be incorporating weather signals, this will be a useful approach; however, the paper is limited to Citi Bike and weather data and does not incorporate neighborhood or demographic data to cluster stations. It also does not consider other means of transportation like bus or heavy rail.

Proposed Method & Innovations

With our enriched data sets, we will develop innovative insights and visualizations. Unlike previous research, which tends to focus either on a small window of time (e.g., one month of data) and/or a specific environmental factor (e.g., COVID), we will provide a holistic view of both Citi Bike ridership behavior and station trends as well as factors that impact them. Our final result will be an open, end-to-end resource that neighborhood and city planners can leverage.

We also took an innovative approach to structuring our data to make it scalable. We used a variety of tools, including OpenRefine and Python/R, when working with our data and ensuring our dataset is consistent year over year as well as across fields. We also leveraged dbt (Data Build Tool) to transform new data that we upload so it can be appended to the existing dataset, making our work scalable as more data is available. For our analysis, we will employ methods including linear regression and random forest. For our visualizations, we will have a central GitHub page embedded with a series of interactive dashboards and charts, including a choropleth map of NYC with plotly, Gephi network graphs, and interactive charts and Tableau dashboards. Our project will focus on the below three main areas.

Factors Impacting Ridership Behavior: Several external factors will be examined to analyze their impact on ridership behavior. We connected daily weather data from Weather Underground to the Citi Bike ridership data. To determine which factors have the largest impact, we did preliminary analysis with a correlation matrix in R. We also ran two linear regression

models in R (stepwise variable selection and manual variable selection for factors with significance at a p-value <0.05) and random forest models in Python. We found that temperature (positive impact), dew point (positive impact), wind speed (negative impact), and humidity (slightly negative impact) have significant impact on the rides taken on a particular day (precipitation and air pressure did not appear to have a significant impact). The below figure shows the relationship between the temperature and the number of rides taken on a day.

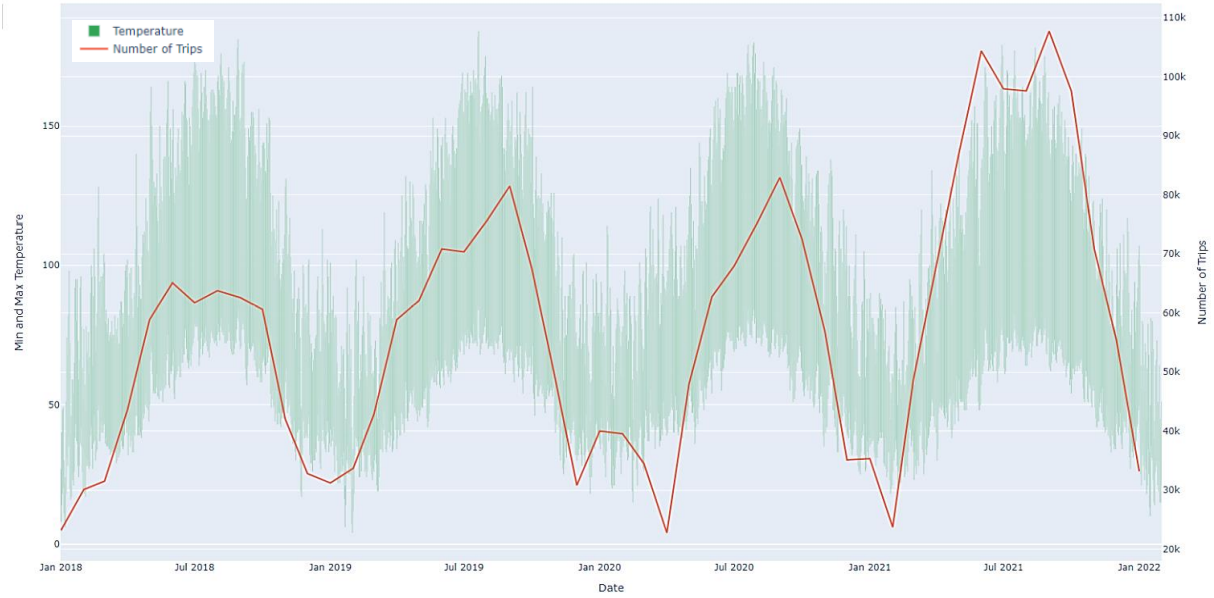


Figure 1: Temperature and Number of Trips Over Time

As the figure demonstrates, the temperature and number of bike trips on a given day have a close relationship; days that have higher temperatures generally having more bike trips taken.

Network Mapping and Trends in Bike Docking Stations: The number of bike docking stations has grown over 350% since Citi Bike launched in New York City. One of our methods was to approach this as a network problem, to better understand how the network has grown and what its characteristics are. Below we show the network growth between 2013 and 2021.

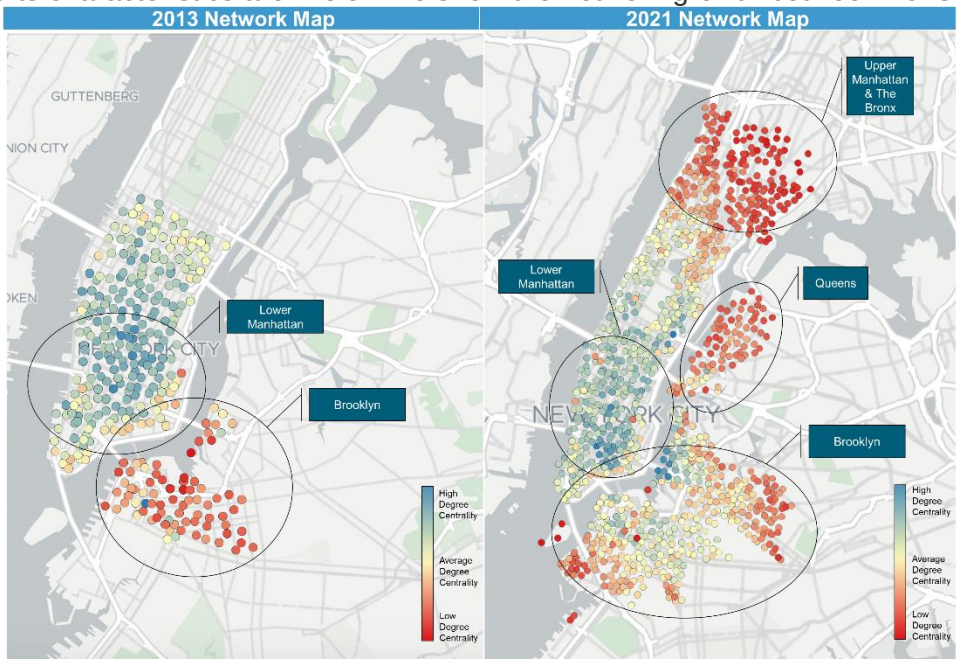


Figure 2: Station Network Map – 2013 vs. 2021

Each bubble in the network maps above represents a Citi Bike station (placed according to its coordinates), and its color represents its degree centrality. Degree centrality is the measure of the total number of edges connected to a particular node (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. We can observe the significant dependence between nearby stations, as acknowledged in the Literature Survey. In 2013, the average degree per node was 168. By 2021, it was 250. The Citi Bike network is dense, with many stations close by to other stations, and most activity occurs within the same neighborhood or neighborhood over.

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 (near the Brooklyn Bridge). Manhattan and Brooklyn have the most active nodes; 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.

Using the Louvain Method for community detection, we saw that communities / bike trips are primarily grouped by borough, except for Brooklyn and Lower Manhattan, likely due to the high bike traffic on the Brooklyn and Manhattan Bridges. This analysis provided an interesting angle of insight into how most trips are occurring locally (in the same or neighboring neighborhoods).

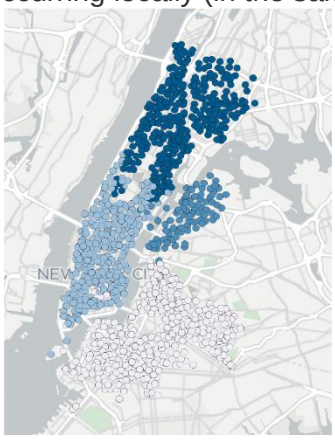


Figure 3: Citi Bike stations colored by their modularity class, derived from the Louvain method

Pricing, Revenue, and Ridership Changes: We also tracked Citi Bike's revenue trends vs. membership signups; Citi Bike offers annual memberships as well as casual ride options (day passes, single rides, etc.). The number of annual membership signups have stayed relatively consistent while casual member sign-ups have grown significantly (Figure 4).

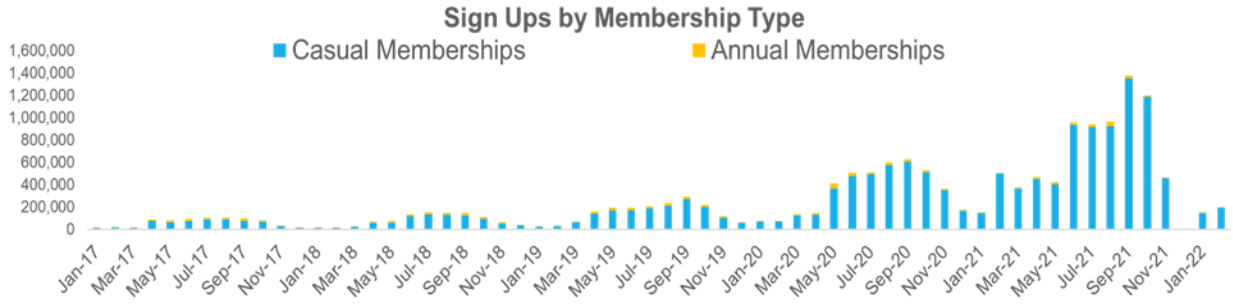


Figure 4: Sign Ups by Membership Type

However, annual members continue to account for approximately half of membership revenue (see Appendix).

In terms of pricing, 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) per year. Yet, 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.

Experiments and Evaluation

To evaluate our proposed approach, we designed experiments to answer three main questions:

1. Is our approach scalable? Our foundational dataset of Citi Bike ridership data includes 139 million rows of data, totaling 20.8 GB in size, hosted on Google Cloud. We aggregated the foundational data using BigQuery into smaller datasets to perform tasks locally or in Google Colab Notebooks. We did not have issues with scaling any machine learning algorithms, as the data required would most likely come from a subset (e.g., predicting daily ridership in 2022 would not require data from 2013), and aggregating the data streamlined the data burden and queries. To ensure our dataset is scalable, we leveraged dbt (data build tool), which transforms uploaded data, merging data views and tests and streamlining the process of adding new data.

2. Is our prediction accurate? We built a statistical model to predict daily ridership using our weather factors (see prediction algorithm in Appendix), which we evaluated through Mean Squared Error (MSE). For our testbed, we split the data into training and test sets (70/30 split) and used that to test accuracy. We improved the model's accuracy through variable selection techniques (stepwise regression and manually with significance using a p-value<0.05). The below table shows comparisons of our rider prediction model performance.

Table 1: Accuracy of Ridership Prediction Model

	Correlation Accuracy	Root MSE	Mean Absolute Deviation
Linear Reg – Stepwise	69.1%	17022 rides	30.90%
Linear Reg – P-Val <0.05	67.49%	17420 rides	32.51%
Random Forest	71.43 %	15674 rides	28.57%
RF – Hyper Tuning	71.38%	15644 rides	28.62%

We performed two linear regression models with different variable selection methods. The step method has a slightly higher Adjusted R^2 value, but the manual method has fewer parameters, making it simpler. Our Random Forest Model performed better than both linear regression models. Below is a visualization of the actual vs. predicted ridership using Random Forest.

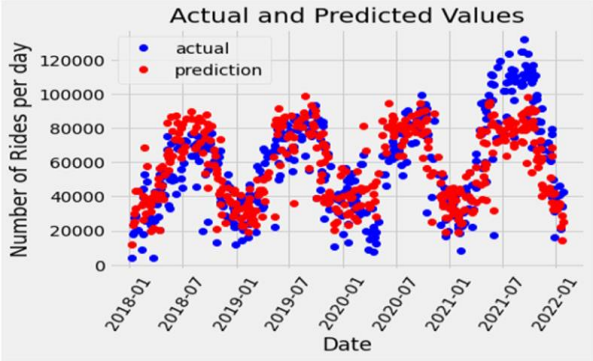


Figure 5: Actual vs. Predicted Ridership with Random Forest Model

We see a significant rise in the number of rides in 2021, post pandemic; the jump is not explained by the weather factors, but likely by other factors such as COVID impacts. As we collect more data post pandemic and include these external factors into our models for analysis, the predictions will get more accurate.

3. Is our visualization usable? To test the usability of our visualizations and dashboards, we created a usability survey in Google Forms with 10 participants, as most experts recommend at least 5 to 10 users (Six, Janet M., and Ritch Macefield 2016). We asked for their impressions and feedback on the dashboards via 1 to 5 scale ratings and open text feedback. Key takeaways below:

Table 2: Usability Survey Key Results

Question	Average Rating
Self-Reported Familiarity with Data Analysis and Reporting	2.1
Rating for Ease of Navigation	4.7
Rating for Visualization Clarity	4.4
Overall Rating	4.9

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 we received the user feedback were providing more high level background and context at the top of the page to orient users to the purpose of the dashboard and clarifying some of the visualization / graph legends.

Conclusions and Discussion

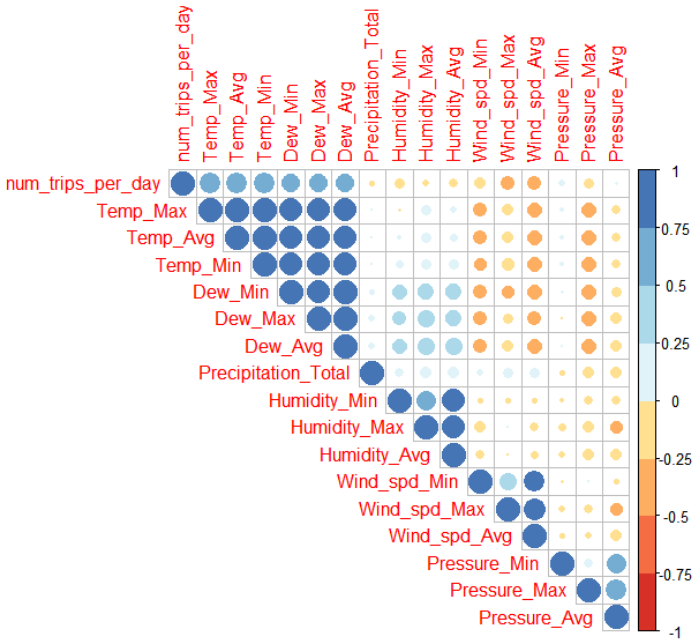
All team members have contributed a similar amount of effort. Our visualizations and analysis will provide business intelligence and bike system utilization insights to help government officials and city planners in infrastructure and financial planning. For example, it appears that annual membership sign ups have stayed relatively resilient to the minor pricing increases that have occurred to date, which is promising as Citi Bike continues its expansion plans. Our network map analysis also shows opportunities for further station and ridership development in areas such as Queens and the Bronx.

Our insights can inform Citi Bike's business strategy for new station development, help optimize operations, enhance biker (customer) experience, and further grow revenue. Metrics for measurement include availability of bikes and docking stations, growth in membership and bike trips, and improvements in customer reviews. If successful, we will improve Citi Bike's availability, increase bike usage, and enhance the user experience

Appendix: Sample Analysis and Visualizations

Factors Impacting Ridership Behavior

To determine which weather factors have the largest impact, we ran a linear regression model as well as a random forest model. Screenshots below.



Appendix Figure 1: Correlation Matrix of Weather Factors

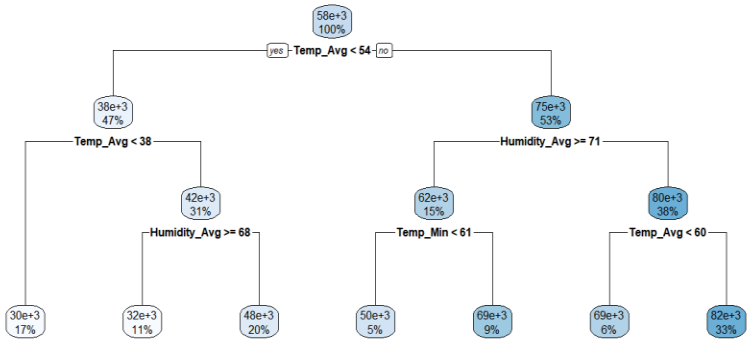
```
Call:
lm(formula = num_trips_per_day ~ Temp_Avg + Wind_spd_Avg + Humidity_Avg +
    Pressure_Max + Precipitation_Total, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-37311  -7397    168   8268  22426

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         194864.76   90031.78   2.164 0.031092 *
Temp_Avg               809.51      38.24  21.170  < 2e-16 ***
Wind_spd_Avg          -583.07     172.69  -3.376 0.000815 ***
Humidity_Avg          -346.31      38.98  -8.884  < 2e-16 ***
Pressure_Max         -5441.77    2930.14  -1.857 0.064106 .
Precipitation_Total  -1159.36    1283.36  -0.903 0.366932
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10820 on 359 degrees of freedom
Multiple R-squared:  0.6946,    Adjusted R-squared:  0.6904
F-statistic: 163.3 on 5 and 359 DF,  p-value: < 2.2e-16
```

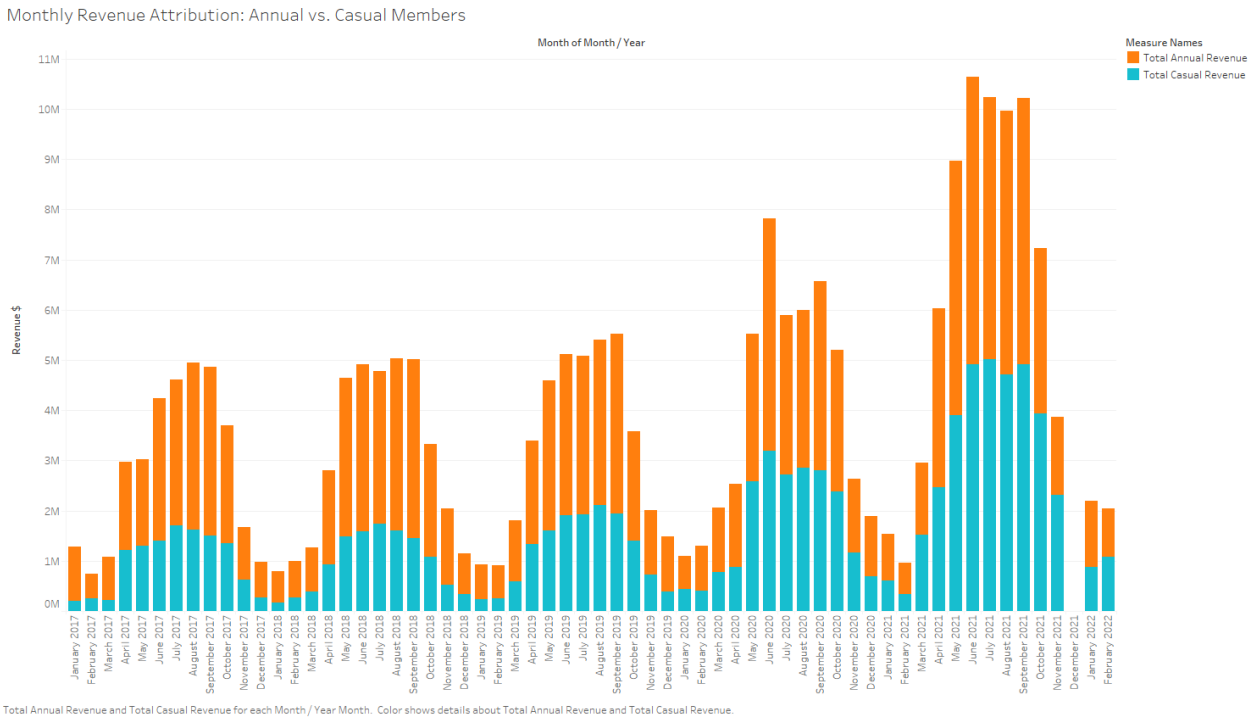
Appendix Figure 2: Linear Regression Output of Weather Factors



Appendix Figure 3: Random Forest Output of Weather Factors

We found that temperature, dew point, wind speed, and humidity have significant impact on the rides taken on a particular day. The factors along with precipitation provides the highest adjusted R^2 value.

Pricing, Revenue, and Ridership Changes: Citi Bike's number of annual membership signups have stayed relatively consistent while casual member signups have grown significantly. However, annual members continue to account for approximately half of the membership revenue despite being a small fraction of the signups (which makes some sense since a casual member sign up could be a single day pass or one ride).



Appendix Figure 3: Monthly Membership Revenue Allocation – Annual vs. Casual Members

References:

An, Ran, et al. "Weather and cycling in New York: The case of Citibike." *Journal of transport geography* 77 (2019): 97-112.

Bouveyron, Charles, Etienne Côme, and Julien Jacques. "The discriminative functional mixture model for a comparative analysis of bike sharing systems." *The Annals of Applied Statistics* 9.4 (2015): 1726-1760.

Correal, A. (2021, August 19). New York City adds 629,000 people, defying predictions of its decline. *The New York Times*.

Faghih-Imani, Ahmadreza, and Naveen Eluru. "A finite mixture modeling approach to examine New York City bicycle sharing system (CitiBike) users' destination preferences." *Transportation* 47.2 (2020): 529-553.

Faghih-Imani, Ahmadreza, et al. "Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City." *Transportation Research Part A: Policy and Practice* 101 (2017): 11-21.

Faghih-Imani, Ahmadreza, and Naveen Eluru. "Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system." *Journal of Transport Geography* 54 (2016): 218-227.

Ford, Weixing, et al. "Riding to Wall Street: determinants of commute time using Citi Bike." *International Journal of Logistics Research and Applications* 22.5 (2019): 473-490.

Hamad, Salma YY, Tao Ma, and Constantinos Antoniou. "Analysis and Prediction of Bikesharing Traffic Flow–Citi Bike, New York." *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, 2021.

O'Mahony, Eoin, and David Shmoys. "Data analysis and optimization for (citi) bike sharing." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. No. 1. 2015.

Six, Janet M., and Ritch Macefield. "How to Determine the Right Number of Participants for Usability Studies." *UX Matters* (2016)

Teixeira, João Filipe, and Miguel Lopes. "The link between bike sharing and subway use during the COVID-19 pandemic: The case-study of New York's Citi Bike." *Transportation research interdisciplinary perspectives* 6 (2020): 100166.

Thu, Nguyen Thi Hoai, et al. "Multi-source data analysis for bike sharing systems." *2017 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2017.

Wang, Kailai, and Gulsah Akar. "Gender gap generators for bike share ridership: Evidence from Citi Bike system in New York City." *Journal of transport geography* 76 (2019): 1-9.

Yanocha, Dana, et al. "The bikeshare planning guide." (2018).