Citi Bike E-Bike Mix and Ridership Revenue

Lab 2 - Benjamin Heuberger, Phillip Hoang, Nicholas Luong

Introduction

Biking as a means of transportation in New York City has blossomed over the last decade, with mayor Eric Adams recently declaring the trend to be “a sign of true progress for our city.”[1] Nowhere has this change been more evident than through the city’s bikeshare program, Citi Bike, which allows the users to rent bikes between docking stations scattered across the city. In a decade of operation, Citi Bike has grown its bike fleet from about 6,000 to 30,000, with weekly ridership multiplying from less than 50,000 rides per week to over 870,000 in a recent, record-breaking week.

A lot of recent growth has been driven by growing demand for electric bikes, which make up 20% of the Citi Bike fleet but account for about half of all rides taken. Recent improvements in e-mobility safety and charging technology have contributed to this growth, and Citi Bike leaders recently stated that they see more of an electric future ahead for the program[2]. However, it is not fully understood how changes to the mix of the Citi Bike fleet (i.e, electric vs. traditional bikes) at stations affect Citi Bike’s revenue from riders. This is a critical question for the company, which, unlike nearly every large public transit system in the country, does not receive any taxpayer funding and relies on revenue from sponsorship and riders. Our study addresses this empirically using publicly-available data on Citi Bike rides: how does the share of rides initiated from stations on electric bikes relate to average revenue per ride? We generate station-level estimates of ride revenue and electric bike availability, and apply regression models to estimate their association.

Data and Methodology

This study uses data from the publicly available Citi Bike system data[3]. This data source compiles a list of all rides across Citi Bike’s trip history and includes information about the rideable type (traditional vs. electric), start/end timestamp, starting and ending station (Name, ID, Latitude/Longitude), and whether the rider is a Citi Bike Member or a casual rider. For the scope of this study, we chose to look at data from May to June 2023 which included 7.1 million observations. We split the data into an exploration set comprising 25% of the observations, which we used to do exploratory analysis and make decisions on features to include in our model, and a confirmation set comprising the other 75% of the data, which we used for our final regressions.

For our data cleaning process, we wanted to focus only on traditional vs electric bikes and remove any rides where there is no ending station. To operationalize this metric, we calculated the revenue per ride based on the member type, duration of ride, and rideable type for each ride in our dataset. With that, we aggregated the dataset by the starting station level and calculated the mean ride time, mean ride revenue, total rides, electric bike percentage in that station, member percentage and days in operation.

[1] NYC DOT Taking New Steps to Expand Bike Infrastructure and Encourage Safe Operation of E-bikes as Overall Bike Ridership Reaches All-time High. (2023, April 24). NYC.gov. Retrieved August 7, 2023, from https://www.nyc.gov/html/dot/html/pr2023/dot-expand-bike-infrastructure.shtml

[2] Barron, J. (2023, May 26). Citi Bike, 10 Years Old and Part of New York’s Street Life. The New York Times. Retrieved August 6, 2023, from https://www.nytimes.com/2023/05/26/nyregion/citi-bike-10-years-old.htm

[3] Citi Bike. (n.d.). System Data. Citi Bike. https://citibikenyc.com/system-data

The aggregated dataset has 1,782 rows, which represents all the stations in NYC. We chose to look at average revenue per ride because it standardizes how we can look at revenue across stations given the differences in how many electric bikes are available at that station for a particular day. To analyze the percentage, we didn't have data on the mix of bikes available per station, so we had to assume the mix of rides taken from a starting station for electric bike percentage. Because cost per ride was not directly available in the dataset, we also had to develop a function to calculate the pricing structure for each ride based on rider type, bike type, and ride time, referencing the Citi Bike site[4]

We used latitude and longitude of each station to zip codes to join in demographic data based on zip code. Our belief is that fields such as income and population density could affect the willingness of riders to pay to use electric bikes, which are more expensive to the rider. We leveraged the ZipcodeR Library and New York State ZIP Codes-County to get demographic data on population density, fixed-borough effects, and income because these factors could impact how electric bikes are used at certain stations: e.g., population density might encourage folks to travel shorter distances and choose not to take an electric bike.

We wanted to explore the relationship between electric bike mix and the average revenue per ride with demographic data as control variables. Looking at the figure below, we can see an initial scatterplot that shows a moderately positive linear relationship between the increase in electric bike and ride price. We opted for a logarithmic transformation on the outcome variable because we wanted to understand the percent change in price, and because the clustering of the data at low levels of revenue-per-ride made interpretations with the nominal price more difficult.

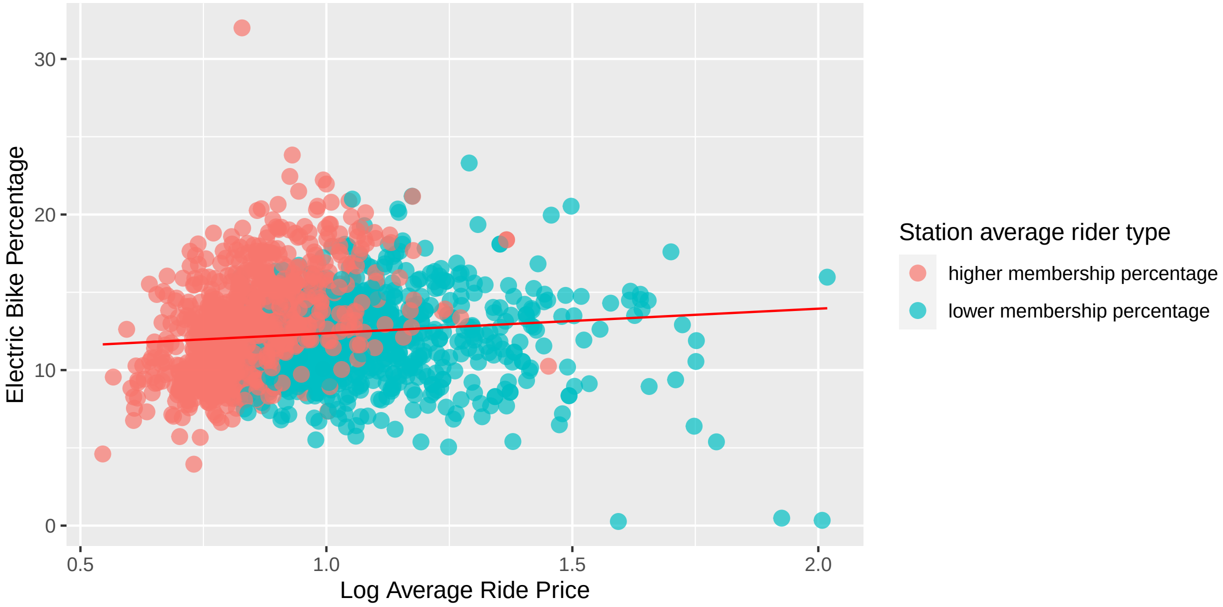

Figure 1: Scatter Plot of Log Ride Price vs. Electric Bike Percentage

This graph reinforces our belief that a linear regression could capture this relationship and so we created a model that would map the following variables against the Ride Price: Electric Bike Percentage, Total Ride, Member Type Percentage, Median Household Income, Population Density, and County. We fit regressions of the form

[4] Citi Bike Membership & Pass Options - NYC. (n.d.). Citi Bike. Retrieved August 6, 2023, from https://citibikenyc.com/pricing

$$\log(\text{Average Ride Price}) = \beta_0 + \beta_1 \cdot \text{Electric Bike \%} + Z\gamma$$

where β_1 represents the percent increase in Average Ride Price for each percent increase in Electric Bikes per station, Z is a row vector of additional covariates, and γ is a column vector of coefficients.

Results

Table 1: Estimated Regressions

	Output Variable: natural log of avg price per ride		
	(1)	(2)	(3)
E-Bike percentage	0.006** (0.002)	0.013*** (0.001)	0.008*** (0.001)
Membership percentage		-0.020^{***} (0.0004)	-0.022^{***} (0.0004)
Total rides		-0.00000^{***} (0.00000)	-0.00001^{***} (0.00000)
Log of Area Income			0.0003 (0.005)
Population Density			-0.00000 (0.00000)
Constant	0.880*** (0.027)	2.408*** (0.036)	2.728*** (0.074)
County fixed effects			✓
Observations	1,782	1,782	1,782
R^2	0.010	0.710	0.751
Residual Std. Error	0.186 (df = 1780)	0.101 (df = 1778)	0.094 (df = 1773)
F Statistic	18.270*** (df = 1; 1780)	1,449.898*** (df = 3; 1778)	668.296*** (df = 8; 1773)

Note: HC_1 robust standard errors in parentheses. Model 3 includes county fixed effects.

Table 1 shows the results of three different regression model specifications. In all of our models, the coefficient on electric bike percentage is positive and highly statistically significant, but modest in magnitude, with point estimates ranging from 0.006 to 0.013. We can interpret this as follows: for every percentage point increase in the portion of a docking station's rides that are taken on electric bikes, there is a .6% to 1.3% increase in the average price per ride. Put into a context, a station that boosted its electric bike availability from an average of around 20% to 25%, might see an approximately 5-6% increase in per-ride revenue. This is a relatively modest change given the average ride price ranges between \$2 and \$3, but given that an average station services over 50 rides per day during the busier months, this marginal added revenue has some tangible practical significance.

We see that electric bike percentage alone explains very little of the variation in average price (Rˆ2 < 0.01 in model 1), but including other covariates dramatically increases the explanatory power of the model (Rˆ2 of the third model = 0.78). Namely, a 1% increase in the percentage of a station's rides taken by members is associated with an approximately 2% lower average ride price, controlling for other factors. Surprisingly, features that capture the demographic attributes of a station's surrounding neighborhood—namely median household income and population density—were not statistically significant, but we do include them in our results for theoretical merit.

Limitations

The model based on Citi Bike data exhibits both statistical and structural limitations. The statistical limitations are rooted in certain assumptions and model specifications. The key assumptions include metric scale, an adequate sample size (n=1782), and the assumption of Independent and Identically Distributed (IID) data. Although our sample size satisfies the n > 30 rule of thumb for the Central Limit Theorem, complete independence may not hold true, as bikes taken from one station may be docked at other stations at the end of a ride, thereby creating inter-station dependence in bike availability. Geographical and temporal scope may also be influenced by factors like county variations, station distribution, and time of day, impacting station usage patterns.

To ensure consistent regression estimates, it is essential that the population distribution is represented by a single best linear predictor. This is supported by using the Variance Inflation Factor (VIF) where we were able to demonstrate an absence of collinearity. Using the QQ Plot, we were also able to observe a normal distribution for residuals. The Residual plots and Shapiro tests were utilized to assess homoskedasticity and the variance was found to be consistent across the data. Model specification limitations stem from data quality and availability, as well as sampling bias. Importantly, our key independent variable of interest is the percentage of rides from a station that originate from electric bikes, which we take to be a proxy for electric bike availability at each station. This decision was made because the true bike mix at each station cannot be observed through the available data. However, our operationalized definition of this variable may be more reflective of the heterogeneous preferences of riders rather than the true bike mix at stations. Additional data collection on station bike mix could help better address this issue.

The data also lacks comprehensive rider profile data and full visibility into ride costs, leading to the need for assumptions in estimating revenue. The absence of information on riders outside of membership further limits the analysis. While we try to account for some level of variation in rider profile by including a zip-code-level feature related to income, many rides may originate outside of the immediate geographies where riders live. We are unable to observe rider age, wealth, or other attributes that may affect their willingness to ride electric bikes and spend money on Citi Bike. Structural limitations encompass omitted variable bias, where the limited data might not capture all relevant variables causing bias, and reverse causality, wherein increased revenue from electric bike trips might lead to more electric bike purchases by Citi Bike.

Conclusion

Our study examines the relationship between the electric bike mix at stations on Citi Bike ridership revenue and provides insights into the dynamics of the bikeshare program. The empirical results indicated a positive and statistically significant relationship between electric bike availability and average ride price at the station level. While the model with additional covariates showed an impressive explanatory power, the research faced statistical limitations related to assumptions, model specifications, and data quality. Structural limitations, such as omitted variable bias and reverse causality, may also have influenced the findings. Addressing these limitations in future research will be crucial to enhance the validity and broader applicability of the findings, leading to a deeper understanding of Citi Bike usage patterns and pricing dynamics.