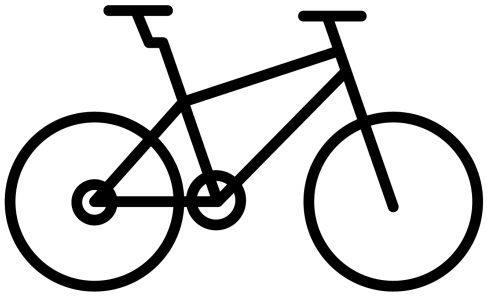


How Citi Bike Meets Demand

Citi Bike Technical Report
MGT585: Fundamentals of Business Analytics
Jason Wang, Jenny Yang, Rosaura Ortiz, Zachary Boehler

EXECUTIVE SUMMARY

This report addresses the issue of bike allocation between stations by conducting a thorough analysis using the four-step analytics process. Analysis methods include examining the available dataset, building a linear regression, and optimizing models. Detailed results are in section 3.

Our analysis shows that Citi Bike:

- Heavily used by subscribers on the weekdays.
- The mean for demand is 45.37 bikes with a standard deviation of 32.53.
- The mean for the trip duration is 12.97 minutes with a standard deviation of 9.54.

We used these predictors to predict the future demand for bikes over five randomly selected stations:

- Start stations
- End stations
- Time of the day
- Start per capita income
- Start percentage of households with no vehicle
- Trip duration in minutes
- Miles traveled

The optimization model shows that Citi Bike is more popular in the evenings in East Village and Lower East Side stations. Our recommendation for Citi Bike is to allocate 95 at Murray Hill, 66 at Lincoln Square, 0 in Brooklyn, 224 in East Village, and 114 in the Lower East Side.

1. OBJECTIVE AND QUESTIONS

1.1. Project Purpose

1.1.1. Objective of the project

Find the number of bikes to stock in each station at the beginning of the day to maximize the number of daily bike trips.

1.1.2. Questions

There are two questions that we focus on to achieve the project objective:

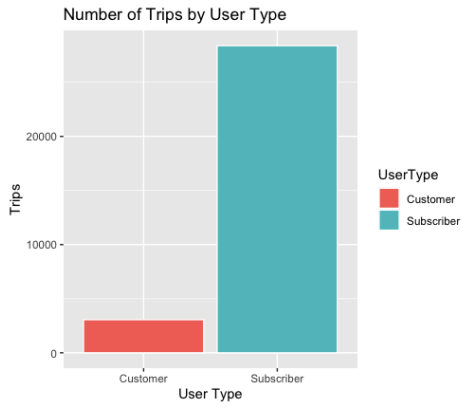
1. How many daily trips between stations (daytime and evening)?
2. How many bikes do we have to allocate to each station?

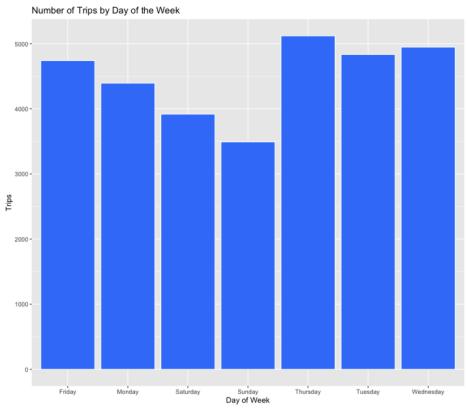
*We are defining trips in terms of users' demand and the number of trips between two stations on that day.

2. DATA ANALYSIS AND RESULTS

2.1. Descriptive Analytics

There were several interesting variables in this data. First, we noticed that a significant number of riders were subscribers. 90.25% were subscribers, and the remaining 9.25% were customers. It shows that a majority of riders use these bikes regularly. We also looked at the number of trips compared to the days of the week. Since most users are subscribers, it makes sense that weekdays and weekends had fewer trips. We also looked at the mean and standard deviation of the continuous variables we used in our predictive analysis. For demand, the mean was 45.37, and the standard deviation was 32.53. The mean for start per capita income was $82,560, and the standard deviation was $31,106. The mean for end per capita income was $82,790, and the standard deviation was $31,025. The mean for trip duration minutes was 12.97, and the standard deviation was 9.54. The last variable was distance miles with a mean of 1.12 and a standard deviation of .88. We also looked at the total amount of missing values. There were 10,681 missing values. However, we concluded that the missing values were insignificant because the data set has a total of 859,294 values. It is only 1.22% of the overall data.





2.2. Predictive Analytics

2.2.1. Regression analysis

Explanatory variables include the time of the day, start stations, end stations, start per capita income, start percentage of households with no vehicle, trip duration in minutes, and miles traveled as predictors. The response variable was Citi Bike's demand.

```
Call:
lm(formula = Demand ~ StartStationId + EndStationId + DemandTime +
    StartPerCapitaIncome + StartPctHouseholdsNoVehicle + TripDurationMinutes +
    DistanceMiles, data = citibikeDemand)

Residuals:
    Min      1Q  Median      3Q     Max
-54.298 -17.330  -3.803  11.186 121.944

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -6.001e+01  1.961e+00 -30.599  < 2e-16 ***
StartStationId             -5.179e-03  1.280e-04 -40.461  < 2e-16 ***
EndStationId               -1.206e-03  1.199e-04 -10.061  < 2e-16 ***
DemandTimeevening           2.029e+01  3.080e-01  65.880  < 2e-16 ***
StartPerCapitaIncome        2.505e-04  4.881e-06  51.323  < 2e-16 ***
StartPctHouseholdsNoVehicle 1.072e+02  2.374e+00  45.139  < 2e-16 ***
TripDurationMinutes        -8.753e-02  2.288e-02  -3.826  0.00013 ***
DistanceMiles               6.676e-01  2.461e-01   2.713  0.00667 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.13 on 30737 degrees of freedom
  (706 observations deleted due to missingness)
Multiple R-squared:  0.363,     Adjusted R-squared:  0.3628
F-statistic:  2502 on 7 and 30737 DF,  p-value: < 2.2e-16
```

R-squared is 0.363. We can explain 36.30% of the variance in our data with this model. Moreover, P-value for F-statistics is < 0.05. Therefore, at least one of the coefficients is non-zero, and the model is valid.

Regression Equation:
Demand = -60.01 - 0.0052*StartStationId - 0.001*EndStationId + 20.29*DemandTimeevening + 0.0003*StartPerCapitaIncome + 107.2*StartPctHouseholdsNoVehicle - 0.088*TripDurationMinutes + 0.67*DistanceMiles

Results of Regression Analysis:

Variable	**Regression coefficient**	**Significance**
Intercept	-60.01	$p < 0.05$
StartStationId	-0.0052	$p < 0.05$
EndStationId	-0.001	$p < 0.05$
DemandTimeevening	20.29	$p < 0.05$
StartPerCapitaIncome	0.0003	$p < 0.05$
StartPctHouseholdsNoVehicle	107.2	$p < 0.05$
TripDurationMinutes	-0.088	$p < 0.05$
DistanceMiles	0.67	$p < 0.05$

2.3. Prescriptive Analytics

To predict the number of bikes needed for our objective, we used specific predictors from the overall data to find out the demand between five random stations: Murray Hill (id=519), Lincoln Square (id=3164), Brooklyn (id=3423), East Village (id=326), and Lower East Side (id=473). We decided to use a sample date of 08/01/2018 since that would be a time when the Citi Bike program would be used very often. Also, we took a practical approach to our predictive model by using 10 minutes of trip duration to go .865 mi, as these values represent how many New Yorkers are using the program for daily routines. Once we calculated the predicted demand for all combinations between the five stations, we placed the values into an optimization model to find the maximum number of trips with 500 bikes. We found that the East Village and Lower East Side had the most bicycle demand while Brooklyn had almost no demand. The other two stations had lower demand in the daytime and increased demand in the evening.

Morning Demand		Destination Station					
		Murray Hill	Lincoln Square	Brooklyn	East Village	Lower East Side	Total
Origin Station	Murray Hill	6	3	3	6	6	24
	Lincoln Square	2	0	0	2	2	6
	Brooklyn	0	0	0	0	0	0
	East Village	22	19	19	22	22	104
	Lower East Side	24	21	21	24	24	114
	Total	54	43	43	54	54	248
Evening Demand		Destination Station					
		Murray Hill	Lincoln Square	Brooklyn	East Village	Lower East Side	Total
Origin Station	Murray Hill	26	23	23	27	26	125
	Lincoln Square	22	19	18	22	22	103
	Brooklyn	0	0	0	0	9	9
	East Village	42	39	39	43	42	205
	Lower East Side	44	41	41	45	45	216
	Total	134	122	121	137	144	658

3. RECOMMENDATIONS AND CONCLUSION

Our analysis shows that for these five stations, we recommend that Citi Bike allocate the initial 500 bikes in the following manner: 95 at Murray Hill, 66 at Lincoln Square, 0 in Brooklyn, 224 in East Village, and 114 on the Lower East Side. This action would result in 715 trips and satisfy the demand for those using the bike system within a mile. Although our model is limited in scope, it proves that Citi Bike can use the optimization model to figure out all sorts of combinations of which stations to allocate bikes. Citi Bike must maintain a practical approach when predicting the demand for bicycles so that the people who need them most can use them.