ISyE 6414: Regression Final Project

Use of Regression Analysis to Predict the Usage of Bike-Sharing Systems

Team 12:
April Zhouting Yu
Eunbee Kim
Nevin Thomas
Simon Kwon

1 Abstract

In recent years, bike sharing systems have been prevalent worldwide, proving an affordable, convenient, and sustainable transportation mode to citizens. The driving forces of such a growth are attributed to a desire to reduce pollution, increased urbanization, and the need for integrated mobility in city environments. However, the bike sharing usage is often spatiotemporally imbalanced, leading to many operation issues such as rebalancing challenges. Stations becoming empty or full is the main cause of users failing to rent or return bikes. Such an issue necessitates efficient bike allocation strategies which highly depends on accurate demand prediction. This study uses data from the CitiBike system in New York City to analyze possible factors that impact the bike sharing usage. Multiple linear regression analysis models were proposed to investigate the influence of temporal (day, month, etc) and weather features (temperature, precipitation, snowfall, and wind speed) on the bike sharing usage and additionally extended with times series trend incorporated. Finally, another regression model was proposed to identify the impact of demographic features (age, income, and living areas) on the bike sharing usage.

2 Introduction

Emerging as an innovate and green and low carbon mode of public transport, public bike sharing systems have brought people enormous convenience and grown rapidly in popularity during the past decade. Such systems provide access to pick-up and drop-off public bikes at numerous bike stations for an affordable fee with an aim to reduce congestion, noise, and air pollution. Unlike traditional fixed-route public transportation system, services provided by bike-sharing systems are more flexible and can meet different categories of users with different travel needs. However, such high transport flexibility gives rise to problems for both the bike-sharing users and operators. For stations located at different places in the city, bike usage can be imbalanced and skewed. For example, some stations which users like to borrow bikes from lack available bikes for interested users while some stations may have too many incoming bikes and thus be jammed often without enough docks for returned bikes, leading to customer loss. System operators try to solve this issue by constantly repositioning bikes between stations. Therefore, accurate study for how bike usage is impacted and prediction on daily usage of bike-sharing system are necessary, based on which the allocation of bikes can be conducted in advance and optimized.

This study uses the data from the CitiBike system in New York, the nation's largest bike-sharing system, as a case study and aims to understand how bike usage is affected. In this study, the following remarks will be discussed in the next sections:

1.Factors that influence the usage of the CitiBike system

2.Significnat factors that determine the usage

3.Different behavior of individuals toward using the CitiBike system

The objective of our study is to identify factors that affect the CitiBike usage and build an accurate prediction model that estimates bike sharing trips. The study will introduce two models. The first model predicts the number of the CitiBike daily usage based on temporal and weather features while the second model predicts annual usage based on demographic features. The rest of the report is structured as follows. The next section provides a brief presentation of data used in the study and data sources. This is followed by the description of the proposed methodology. Analysis and results based on this methodology are described and discussed in the fourth section. Finally, concluding remarks and future research perspectives are presented in the last section.

3 Data Sources

CitiBike was launched in 2013 in New York with 6,000 bikes and 330 bike docking stations in Manhattan and Brooklyn and has expanded to over 800 stations and 13,000 bikes. It is the nation's largest bike-sharing system and this study uses it as a case study to examine the bike usage in a deeper layering of understanding and to propose a prediction model.

The trip data archived from 06/01/2013 to 10/31/2019 were used in following analysis. The historical trip data was obtained from the official webpage of CitiBike [1]. The archived data provide detailed information on trips such as duration, start and end time, start and destination station codes, BikeID, user type (one-time or subscriber), and user information (age and gender) as shown in Table 1.

1	tripduration	starttime	stoptime	start_station_id	end_station_id	bikeid	usertype	birthyear	gender
2	2059	6/1/2013 0:00	6/1/2013 0:35	406	406	19599	0		0
3	1521	6/1/2013 0:01	6/1/2013 0:26	2008	310	15567	1	1983	1
4	2028	6/1/2013 0:01	6/1/2013 0:35	485	406	18445	0		0
5	1829	6/1/2013 0:03	6/1/2013 0:34	265	436	15234	1	1984	1
6	899	6/1/2013 0:09	6/1/2013 0:24	494	494	15539	1	1967	1
7	395	6/1/2013 0:11	6/1/2013 0:18	312	410	19477	1	1970	1
8	424	6/1/2013 0:11	6/1/2013 0:18	494	519	18489	1	1957	1
9	1077	6/1/2013 0:12	6/1/2013 0:30	482	480	16963	1	1982	1
10	689	6/1/2013 0:12	6/1/2013 0:24	439	317	16164	1	1982	1
11	1997	6/1/2013 0:14	6/1/2013 0:48	531	531	16443	1	1969	2
12	226	6/1/2013 0:15	6/1/2013 0:18	358	358	20306	1	1980	2
13	1221	6/1/2013 0:15	6/1/2013 0:35	355	303	16937	0		0
14	624	6/1/2013 0:15	6/1/2013 0:26	271	485	17317	1	1972	2
15	846	6/1/2013 0:15	6/1/2013 0:29	477	223	19116	1	1975	2
16	200	6/1/2013 0:16	6/1/2013 0:19	293	475	18414	1	1983	1
17	415	6/1/2013 0:19	6/1/2013 0:25	517	2009	19765	1	1977	2
18	921	6/1/2013 0:19	6/1/2013 0:34	2004	294	18451	1	1985	2
19	1493	6/1/2013 0:19	6/1/2013 0:44	358	528	20306	1	1980	2
20	1622	6/1/2013 0:19	6/1/2013 0:46	3019	478	15536	0		0

Table 1: Trip data of CitiBike

[1] https://www.citibikenyc.com/system-data

Other external data sought in this study include weather and demographic information since they have been found to be tightly related to bike usage [1]. These external data are obtained from National Weather Service[2] and United States Census Bureau[3] respectively. The weather data include temperature, wind speed, precipitation, snow depth in terms of daily and monthly, time of sunrise and sunset and so on as shown in Table 2.

1	DATE	REPORT	DailyAver	DailyAver	DailyAver	DailyAver	DailyMax	DailyMini	DailyPea	DailyPrec	DailySno	DailySno	DailySust	HourlyDe	HourlyDry	HourlyPre	HourlyWe	HourlyWi	HourlyWi
2	1/1/2013 0:51	FM-15												24	39	0	33	22	10
3	1/1/2013 1:51	FM-15												26	38	0	33	17	3
4	1/1/2013 2:51	FM-15												27	39	0	34		7
5	1/1/2013 3:51	FM-15												27	39	0	34		9
6	1/1/2013 4:51	FM-15												27	39	0	34		7
7	1/1/2013 5:51	FM-15												26	39	0	34	21	7
8	1/1/2013 6:51	FM-15												26	39	0	34	21	10
9	1/1/2013 7:51	FM-15												27	38	0	34		7
10	1/1/2013 8:51	FM-15												27	39	0	34	21	7
11	1/1/2013 9:51	FM-15												27	39	0	34	17	10
12	1/1/2013 10:51	FM-15												27	39	0	34		6
13	1/1/2013 11:51	FM-15												26	39	0	34	16	5
14	1/1/2013 12:51	FM-15												25	37	0	33	22	9
15	1/1/2013 13:51	FM-15												25	37	0	33		0
16	1/1/2013 14:51	FM-15												24	38	0	33		7
17	1/1/2013 15:51	FM-15												21	38	0	32		7
18	1/1/2013 16:51	FM-15												18	36	0	30	21	10
19	1/1/2013 17:51	FM-15												16	35	0	29		5
20	1/1/2013 18:51	FM-15												15	33	0	27		6
21	1/1/2013 19:51	FM-15												14	32	0	26		5
22	1/1/2013 20:51	FM-15												13	30	0	25		7
23	1/1/2013 21:51	FM-15												12	29	0	24	22	11
24	1/1/2013 22:51	FM-15												9	27	0	22		
25	1/1/2013 23:51	FM-15												10	26	0	21		6
26	1/1/2013 23:59	SOD	22	33	31	7	40	26	26	0	0	0	15						
27	1/2/2013 0:51	FM-15												10	25	0	21	18	10

1	MonthlyD	MonthlyD	MonthlyD	MonthlyD	MonthlyD	MonthlyD	MonthlyG	MonthlyG	MonthlyG	MonthlyG	MonthlyG	MonthlyG	MonthlyM	MonthlyM	MonthlyM	MonthlyT	MonthlyT	Sunrise	Sunset
2																			
3																			
4																			
5																			
6																			
7																			
8																			
9																			
10																			
11																			
12																			
13																			
14																			
15																			
16																			
17																			
18																			
19																			
20																			
21																			
22																			
23																			
24																			
25																			
26																		720	1639
27																			

Table 2: NYC Weather Data

The demographics data include population age and income distribution in each NTA (Neighborhood Tabulation Area).

1	nta_name	borough	nta_code	populatic	under_5_	5-9_years	10-14_yea	15-19_yea	20-24_yea	25-29_yea	30-34_yea	35-39_yea	40-44_yea	45-49_yea	50-54_yea	55-59_yea	60-64_yea	over_65_y	median_a	househol	less_thar	10000_to_	15000_to_
2	Bath Bea	Brooklyn	BK27	33163	2210	1826	1922	1657	1726	2850	2771	2442	2459	2226	2114	2165	1907	4888	38.3	11428	895	618	1223
3	Bay Ridge	Brooklyn	BK31	79134	5166	4513	3379	3551	4347	6827	6876	5459	5453	5579	5938	4908	4758	12380	39.5	33487	2313	1573	3235
4	Bedford	Brooklyn	BK75	75318	7480	6580	5378	4635	6479	8683	6984	5586	4331	3625	3603	3452	2686	5816	29.1	26059	3673	2317	3385
5	Bensonhu	Brooklyn	BK29	64267	4151	3563	2891	3470	4182	4939	4634	4677	4431	4391	4663	4566	3602	10107	39.6	22187	2356	1413	2718
6	Bensonhu	Brooklyn	BK28	91646	6498	5162	4529	4648	5793	8036	7766	5917	6409	5840	6679	5866	5432	13071	37.8	29909	2330	1746	3383
7	Borough F	Brooklyn	BK88	105913	14350	11602	10077	8104	7754	8329	6880	5720	5402	3715	4110	4375	4749	10746	25.6	27422	2208	2630	4619
8	Brighton I	Brooklyn	BK19	34157	1938	1897	1582	1489	1733	2267	2210	2139	2122	2048	2508	2113	2307	7804	44.3	14097	2781	1404	1522
9	Brooklyn	Brooklyn	BK09	24140	1592	775	564	1055	1527	2358	3073	2532	2112	1646	1205	1252	1235	3214	37.2	11152	451	391	579
10	Brownsvil	Brooklyn	BK81	60124	5246	3895	4882	4243	4548	5420	4341	3883	3914	3793	3868	3240	2503	6348	32.1	21552	5683	2174	3048
11	Bushwick	Brooklyn	BK77	60834	3975	3532	4100	3757	6793	8065	6503	5143	4145	3527	3231	2483	1734	3846	30.1	19207	2150	1294	2440
12	Bushwick	Brooklyn	BK78	75003	4586	4748	4145	4169	8147	10173	7000	5097	4343	4628	3847	3658	2967	7495	31.1	26256	4307	2249	3187
13	Canarsie	Brooklyn	BK50	88367	5225	5125	6027	6175	6834	6716	5879	5298	5971	6005	6671	6398	5716	10327	37.1	28677	2310	1301	2437
14	Carroll Ga	Brooklyn	BK33	42463	3379	2481	1730	1607	1591	3269	5861	4452	3473	3211	2187	2592	2092	4538	36.5	18111	1988	695	1069
15	Clinton H	Brooklyn	BK69	38655	2430	2092	1890	1767	3766	4550	4608	3283	2828	2190	1896	1890	1662	3803	33.1	15797	1358	774	1085
16	Crown He	Brooklyn	BK61	107134	6831	5585	5616	6544	9246	12060	10311	7883	7192	6358	6854	6305	4910	11439	33.7	43221	6535	3340	5335
17	Crown He	Brooklyn	BK63	41962	3140	2777	2585	2463	3814	4311	4027	2609	2016	2427	2490	2305	2240	4758	32.3	14971	1225	1164	1820
18	Cypress H	Brooklyn	BK83	48899	3709	3222	3212	3890	4724	3891	3479	3332	3369	3537	3596	2933	2295	3710	32.6	14159	2004	652	1571
19	DUMBO-V	Brooklyn	BK38	42245	3154	2044	1277	1492	2561	5275	6419	4343	3214	2751	2183	1966	1778	3788	34.1	17585	1515	874	1405
20	Dyker Hei	Brooklyn	BK30	45877	2926	2721	2674	2676	2723	3164	3421	3336	3409	3214	3201	3073	2556	6783	38.9	14907	1066	800	1907
21	East Flatb	Brooklyn	BK91	53535	3113	3318	2991	3517	3540	4076	3552	3223	3932	3318	3532	3910	3235	8278	39.1	18556	1352	1015	1675
22	East New	Brooklyn	BK82	94448	7665	7515	6630	6850	7624	7707	6976	6005	5468	6336	6196	5463	5024	8989	32.3	32744	6174	2386	4051

Table 3: NYC Demographics

4 Methodology

The CitiBike trip data were assembled with the temporal and weather related data and demographic data respectively to create the final data sets to analyze each model. The first model used regression analysis to predict the daily bike-sharing usage based on the temporal

[2] https://w2.weather.gov/climate/xmacis.php?wfo=okx
[3] https://www.census.gov/quickfacts/newyorkcitynewyork

and whether characteristics. The following model is considered first:

$$Y_i = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 D_{weekend} + \beta_6 LD + \epsilon$$

, where Y refers to the number of bike share usages on i^{th} day (to which NOR(Number of rides) will be used to refer for the rest of the report), W_js refer to the weather related variables (W_1: Temperature (°F), W_2: Precipitation (in inches to hundredths), W_3: Snowfall(in inches to tenths), and W_4: Wind speed (m/s)), $D_{weekend}$ indicates whether a given day is weekend or weekday, and LD refers to the length of the day, which is computed by sunset time - sunrise time. Three questions were through the model analysis:

> 1.How much variance in the number of daily bike-sharing usage can be explained by the temporal and weather vectors?
>
> 2.What predictors are found to be the most statistically significant in determining the number of daily usage?
>
> 3.When controlling other variables, which variable has the most statistical significance?

Through the analysis of the model, several models were implemented by taking appropriate actions including the exclusion of insignificant factors, checking assumptions and corresponding Y transformation to update the model. As an expansion of the first model, we wanted to examine whether there was a time series trend in the response variable (i.e. the bike-sharing usage on i^{th} day depends on the previous day or the previous week). In order to do so, we conducted ARIMA model (AutoRegressive Integrated Moving Average) by modelling the response variable as a time series. Detailed analysis and results of these models were provided in the next section.

The second model used regression analysis to predict the number of annual bike-sharing usage across different areas in the city of Manhattan which are termed as Neighborhood Tabulation Areas (NTA). The city is assumed to consist of 46 sections. In this model, demographic features including the size of each NTA, users' age, and income distribution will be considered to predict the annual number of bike-sharing usage. The following model was considered:

$$Y_j = \beta_0 + \beta_1 \text{Area}_j + \beta_2 \text{Age}_j + \beta_3 \text{Income}_j + \epsilon \quad \forall j \in \text{NTA}$$

, where Y_j refers to the number of annual bike-sharing usage in jth NTA and Area refers to the size of jth NTA. Age is comprised of 3 different groups: Teenage (0-24), Mid-age (25-59), and Older (>60) and refers to the number of population of each age group in jth NTA. In this model, the following income groups are considered: Low (\$0-\$35,000), Mid (\$ 35,001 - \$ 100,000), and High ($\geq$ \$ 100,001). The Income factor refers to the percentage of each income group in jth NTA. Similarly to the first model, the following questions were addressed through the analysis:

> 1.How much variance in the number of annual bike-sharing usage can be explained by the demographics across different areas?
>
> 2.What predictors are found to be the most statistically significant in determining the number of daily usage?
>
> 3.When controlling other variables, which variable has the most statistical significance?

5 Analysis and Results

5.1 Model 1: Daily number of rides

5.1.1 Exploratory Data Analysis

We see from Figure 1 that the number of rides (NOR) displays a generally increasing trend over years and fluctuates with a cycle of one year. The weather variables - temperature, speed of wind, precipitation and length of day also show a very similar behavior with a systematic cycle of one year. This suggests that these variables might have good predictive power and also at the same time we need to be wary of possible high correlations among them.

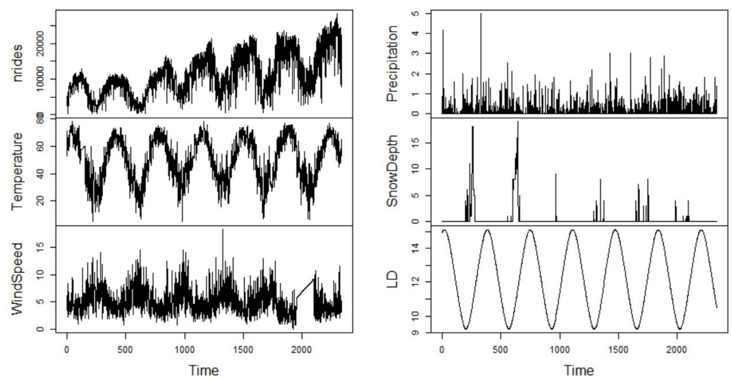

Figure 1: Time series of all the variables

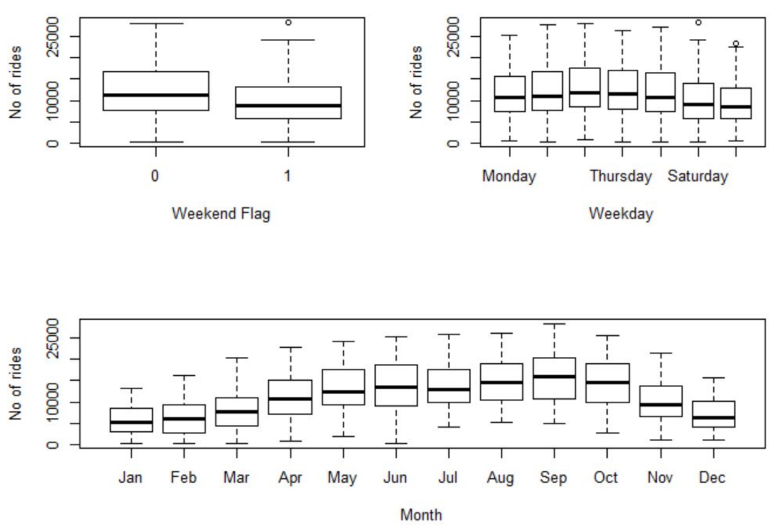

Figure 2: Box plots showing the variation of the number of rides based on weekday-weekend difference, day of the week and month.

The seasonal nature of the number of rides is further confirmed by the box-plots in Figure 2. The box-plot based on weekend flag indicates that there is a considerable difference in the median number of rides on a weekday (0) compared to weekend (1). This is further explored in detail for each day of the week in the top right plot in Figure 2. The highest median is observed on Wednesday whereas the lowest median is observed on Sunday. The number of rides on weekdays are higher due to the fact that many people tend to use bikes to commute to their work place. The bottom plot in Figure 2 demonstrates the monthly seasonality of the number of rides. The results show that the NORs are higher during the months of April to October. Highest median is observed in September whereas the lowest median is observed in January. This is consistent with the observation that most people are more inclined to use bikes during the dry summer months when the weather is conducive to riding.

5.1.2 Modelling

We split the dataset into a train and test split so that we could check the accuracy of the forecasts given by the developed model. The train set consisted of data from 6/1/2013 to 10/31/2018 while test set contained one year of data from 11/1/2018. We used the RMSE calculated on the test set as the evaluation metric for the implemented models.

We started out by fitting a linear regression model on all the available explanatory variables (LR1). While it gave a decent R^2 value of 0.826 and an RMSE of 3204.474 as mentioned in Table 5, the p-values for multiple coefficients were less than 0.05. Hence, there was not enough proof to reject the null hypothesis of 0 coefficient. This can happen very often due to multicollinearity among the explanatory variables. So, we built a second linear regression model (LR2) on all but the weekend flag and length of day variables. We did an ANOVA test which gave a p-value of 0.07 indicating that we should favor the simpler model (LR2). The second model was not only more parsimonious but also resulted in a slightly better RMSE (as can be seen in Table 5). We then proceeded to do a residual analysis for the LR2 model and the resulting plots are attached in Appendix Figure 9. We observe that there is significant deviation from the normal assumption we made as can be seen from the residuals vs fitted and the QQ plots.

Therefore, we did a Box-Cox transformation to account for the non-constant variance. The Box-Cox results indicate that λ value is approximately 0.6 (plot attached in Appendix Figure 10). We decided to do a square-root transformation on the NOR variable for convenience and better interpretation. When we fitted a linear regression model after the square-root transformation of the dependent variable, the performance of the model improved significantly as can be seen from the improved RMSE value in Table 5. The R^2 value also increased to 0.84. The behavior of the residual plots of this transformed model conforms better with the normal assumption as is demonstrated by Appendix Figure 11. The coefficients of the model indicates that NOR on any particular day increases with an increase in temperature and decreases with an increase in wind speed, precipitation, and depth of snow.

In addition to the above mentioned models, we also a fitted Seasonal Auto Regressive model on the NOR variable as it can be treated as a time series. We used the time points to estimate the trend. The month and weekday variables were used to estimate the seasonality. Then, the stationary residuals (R_i) were estimated by subtracting the fitted trend+seasonality from the NOR series. The residuals were then modelled as a Seasonal Auto-Regressive Model with weather related attributes as the exogenous variables. We assumed an AR order of 1 and period of 7. The resulting model indicated that NOR on any particular day has a positive dependence on the NOR on the previous day and the NOR on the same day in the previous week.

Model equations for all the implemented models are depicted in Table 4.

Model	**Equation**
LR1	$NOR = -4109 + 99 * Temp - 181 * Wind - 3989 * Precipitation - 85 * Snow$ $+318 * DayLength + day + month + 12356 * timept(trend)$
LR2	$NOR = -1135 + 102 * Temp - 181 * Wind - 4009 * Precipitation - 85 * Snow$ $+day + month + 12345 * timept(trend)$
Transformed	$\sqrt{NOR} = 37 + 0.59 * Temp - .97 * Wind - 22 * Precipitation - 1.56 * Snow$ $+day + month + 58 * timept(trend)$
SAR	$R_i = -809.6826 + 0.5438 * R_{i-1} + 0.1485 * R_{i-7} + 37.5 * Temperature$ $-121.6382 * Windspeed - 3603.1890 * Precipitation$

Table 4: Model equations

Method	LR1	LR2	Sq-root transformed LR	Seasonal AR
RMSE	3204.474	3203.249	2850.121	3154.850

Table 5: RMSE values obtained for the implemented models

5.1.3 Forecasting

Since the square-root transformed model gave the best results on the testing set, we used it do the forecasting task. The forecasting was done for the period from 11/1/2018 to 10/31/2019/ (corresponding to the points 1970 to 2334 on the x-axis of Figure 3). We can see from Figure 3 that the forecast seems to be pretty good as the green series almost overlaps the black one.

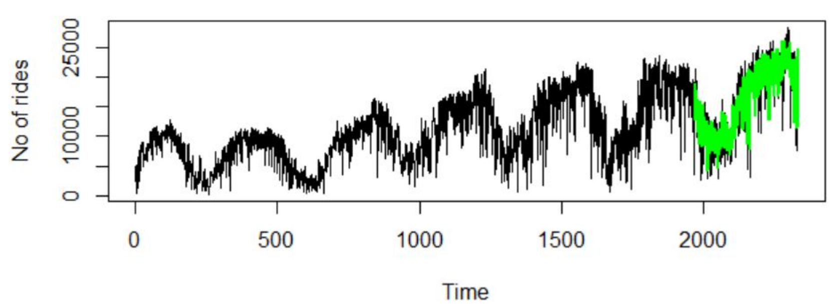

Figure 3: The green series shows the number of rides forecasted for the next year using the square-root transformed model trained on all the previous year's data

Figure 4: NYC NTA Boundaries

5.2 Model 2: Number of rides from/to each NTA

In this section, we focused on the number of rides from/to each neighborhood tabulation areas (NTA). The NYC Department of City Planning created boundaries of NTAs using whole census tracts from the 2010 Census as building blocks. As shown in Figure 12 (`https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq`), there are 195 NTAs in New York City in total. The complete NTA_id and NTA_name table is attached in the appendix Figure 12.

From our data, we have the start_station_id and end_station_id for each trip. For each record, we labeled it with starting_nta_id and end_nta_id. We then have the total number of rides starting from(ending in) each NTA during the whole time horizon as shown in Figure 5 and Figure 6. Note that according to our data, there are 46 NTAs in total with positive number of NOR. By applying two-sample Kolmogorov-Smirnov test, we have $D = 0.043478$ and p-value $= 1$, indicating that these two samples are from the same distribution, i.e., there is no significant differences between the numbers of rides starting from/ending in a specific NTA. The hot spots with more than 2×10^6 rides in total include Hudson Yards-Chelsea-Flatiron-Union Square, Midtown-Midtown South, West Village and SoHo-TriBeCa-Civic Center-Little Italy.

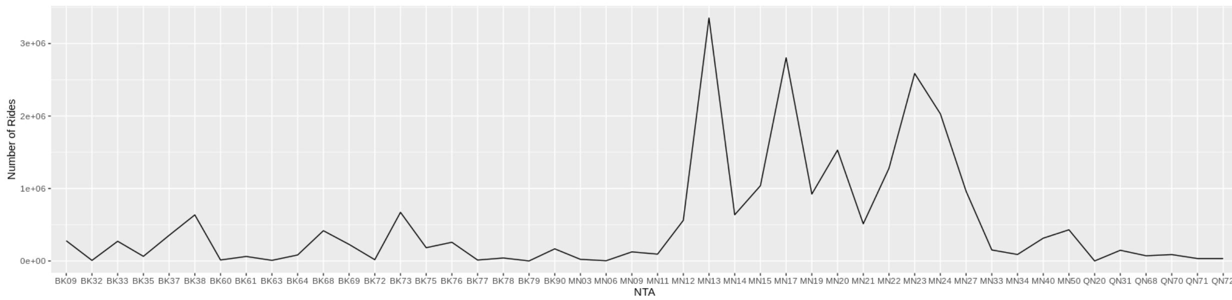

Figure 5: Number of Rides Starting from Each NTA

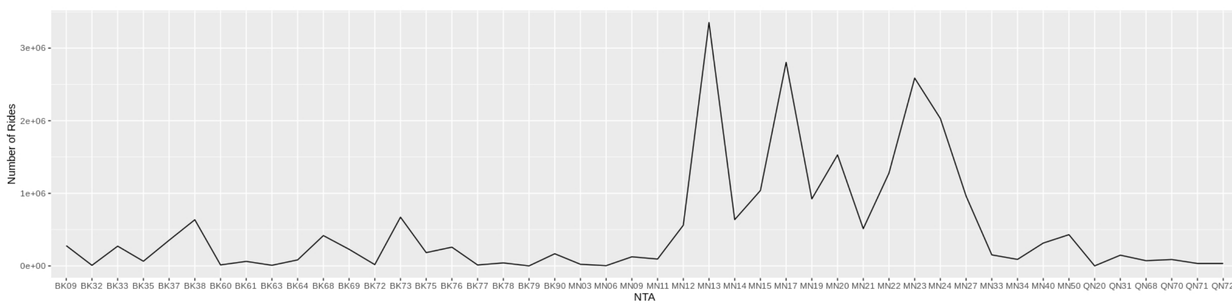

Figure 6: Number of Rides Ending in Each NTA

	belong_to_which_nta	n	area	midage	teenage	older	low_income	mid_income	high_income
1	BK09	278547	1.1958302	10568	4940	5701	4277.317	6296.921	13565.762
2	BK32	7628	5.9922645	21986	14307	9814	22376.825	23078.883	11921.293
3	BK33	271755	5.3403947	19184	6467	9222	10806.249	11783.946	19872.806
4	BK35	63842	3.7611480	22761	17518	14665	30352.108	24356.760	12662.132
5	BK37	353291	5.0894248	33204	14069	13731	9554.457	22951.570	41374.973
6	BK38	635747	3.4105188	18910	9328	7532	11158.830	10810.492	20275.678
7	BK60	13250	3.7877643	22936	15495	18158	28407.787	29414.712	11540.501
8	BK61	61836	6.1803981	38598	27850	22654	48452.102	40477.967	18203.931
9	BK63	7991	1.9142576	13569	10588	9303	16133.409	19457.632	6370.959
10	BK64	82556	1.2249781	9517	3752	4291	4489.089	6733.634	9270.277
11	BK68	417156	1.9739818	11784	6831	7052	10378.016	10698.011	9211.974
12	BK69	228657	2.4583563	14805	10083	7355	10446.173	14762.652	13446.175
13	BK72	17482	1.3882879	6219	7877	5395	22302.870	8714.851	2749.279
14	BK73	671132	3.4585660	24039	14810	6967	15784.211	17923.686	20622.103
15	BK75	182867	3.9077843	24129	19797	11954	34348.176	26298.725	14671.099
16	BK76	257800	4.2346147	15065	8560	6862	8499.051	13273.988	11577.961
17	BK77	12310	2.9858076	22549	18615	8063	24042.843	24660.464	12130.693
18	BK78	41009	4.8111004	24915	22489	14120	34342.096	28825.993	11834.911
19	BK79	63	2.4054608	11515	7870	6618	16841.364	13607.305	3576.331
20	BK90	167076	4.6894701	14167	9888	7987	13439.149	12713.317	9832.534
21	MN03	21924	3.0483018	31652	20723	17146	42359.073	30887.236	12625.692
22	MN06	3056	1.2775068	8429	7127	4796	12940.055	8063.843	2662.102
23	MN09	125127	2.4184456	14785	21256	11561	20633.229	14647.869	18425.902
24	MN11	94049	1.7317983	19737	11225	10169	18348.048	17477.389	14269.563
25	MN12	559886	4.1234962	51027	21583	43336	30219.370	37134.992	66664.637

rotating

Figure 7: NOR of each NTA and Demographic data

Given the similarity between the number of rides starting from/ending in each NTA, we take the number of rides starting from each NTA as the response (y variable) in our model. To analyze the relationship between NOR and demographics, we considered the area, population age, population income of each NTA. Part of the data we used in our analysis is shown in Figure 7. For age factor, we divided the population into three groups according to age; for income factor, we also divided the population into three groups based on their income level. We started from the fitted linear regression model with the formulation

$$\begin{aligned} y_i =& 343,503 + 62,637 \cdot \text{Area}_i + 313.32 \cdot \#\ \text{teenager}_i + 231.50 \cdot \#\ \text{mid-age}_i + 114.06 \cdot \#\ \text{elder}_i \\ &- 122.82 \cdot \#\ \text{low-income}_i - 324.65 \cdot \#\ \text{mid-income}_i - 91.80 \cdot \#\ \text{high-income}_i. \end{aligned}$$

By selecting the significant variables and re-running the linear regression model, we have the modified formulation

$$\begin{aligned} y_i =& 379,757 + 62,637 \cdot \text{Area}_i + 238.69 \cdot \#\ \text{teenager}_i + 145.33 \cdot \#\ \text{mid-age}_i \\ &- 62.60 \cdot \#\ \text{low-income}_i - 237.28 \cdot \#\ \text{mid-income}_i. \end{aligned} \tag{1}$$

With the equation (1), we observed the positive correlation between the teenager population size/mid-age population size and NOR in each NTA and negative correlation between the low-income and mid-income population size and NOR.

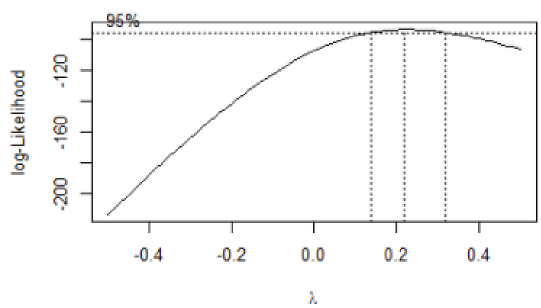

Figure 8: Box-Cox transformation of Model 2

We then did a Box-Cox transformation as shown in Figure 8. Given that $\lambda \approx 0.2$, we decided not to transform y.

6 Conclusions

We have studied how regression can be used to make models that can predict the usage of bike sharing systems. Particularly, we formulated two regression models with predictors of interest (i.e., one with temporal and weather characteristics and the other with demographic information of stations) to see how the bike sharing usage can be accounted for by the predictors. With respect to temporal and weather variables, we compared four regression models (see Section 4) and found out log-transformed model is the best performing model with least errors. The results show that the bike sharing usage (NOR) differs by days of the week, months, and weather features through model 1. Also, model 2 showed that NOR is accounted for by demographic features of area which each station (NTA) is located in (see

Section 4). We observed a similar pattern between the number of rides staring from each NTA and the number of rides ending at each NTA. By using the total NOR that starting from and ending at each NTA as a dependent variable, we showed that the relationship between NOR and demographic information of NTAs can be expressed by a linear regression model (see the formula provided in section 5.2). The current study clearly showed that a significant relationship exists between the bike sharing usage and the predictors of interest. The current study can serve as an evidence for prediction regarding how the demand will evolve according to seasonality, weather, and demographics. It will help administrators in practice better deal with the common issues regarding redistribution of bikes to some extent. For future work, this rebalancing issue can be addressed in a more detailed manner with an extended optimization-incorporated or machine learning-incorporated research that predicts future demand among bike stations and designs a bike redistribution algorithm that instructs operators about what stations to rebalance and how to do so.

7 Appendix

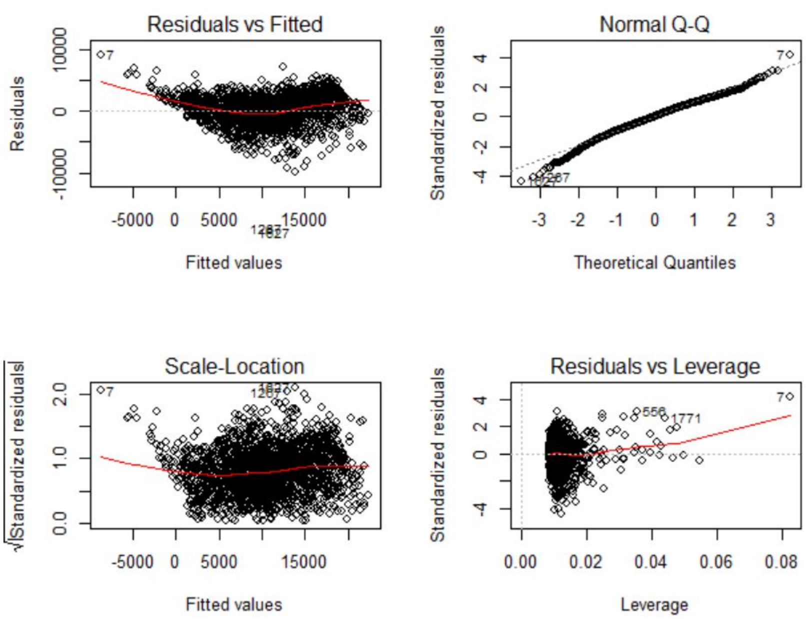

Figure 9: Residual analysis for the second linear regression model (LR2)

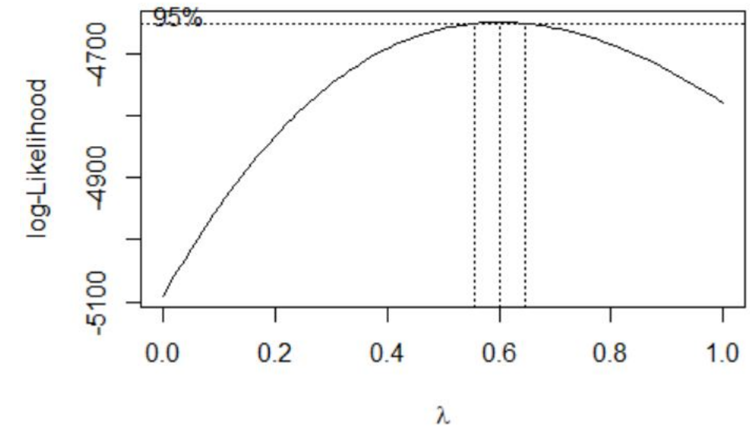

Figure 10: Box-Cox transformation for LR2 model

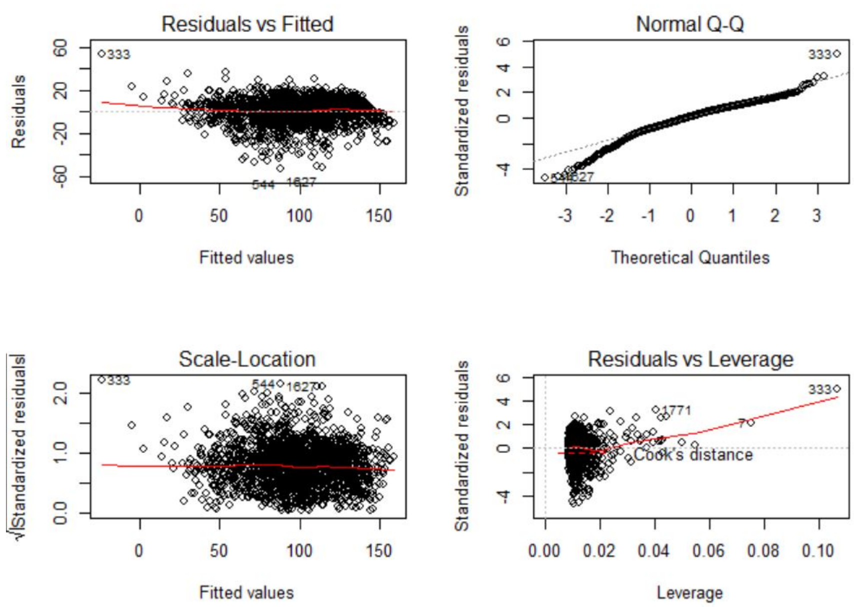

Figure 11: Residual analysis for the square-root transformed linear regression model

NTA code	NTA name	Borough	NTA code	NTA name	Borough
MN15	Clinton	Manhattan	BK35	Stuyvesant Heights	Brooklyn
MN24	SoHo-TriBeCa-Civic Center-Little Italy	Manhattan	BK78	Bushwick South	Brooklyn
MN27	Chinatown	Manhattan	BK72	Williamsburg	Brooklyn
BK68	Fort Greene	Brooklyn	BK90	East Williamsburg	Brooklyn
MN13	Hudson Yards-Chelsea-Flat Iron-Union Square	Manhattan	BK76	Greenpoint	Brooklyn
BK75	Bedford	Brooklyn	QN31	Hunters Point-Sunnyside-West Maspeth	Queens
MN23	West Village	Manhattan	QN68	Queensbridge-Ravenswood-Long Island City	Queens
MN17	Midtown-Midtown South	Manhattan	MN12	Upper West Side	Manhattan
BK09	Brooklyn Heights-Cobble Hill	Brooklyn	MN33	East Harlem South	Manhattan
BK38	DUMBO-Vinegar Hill-Downtown Brooklyn-Boerum Hill	Brooklyn	MN09	Morningside Heights	Manhattan
MN20	Murray Hill-Kips Bay	Manhattan	BK64	Prospect Heights	Brooklyn
MN19	Turtle Bay-East Midtown	Manhattan	MN34	East Harlem North	Manhattan
MN22	East Village	Manhattan	MN11	Central Harlem South	Manhattan
BK69	Clinton Hill	Brooklyn	QN71	Old Astoria	Queens
BK32	Sunset Park West	Brooklyn	QN72	Steinway	Queens
BK73	North Side-South Side	Brooklyn	MN03	Central Harlem North-Polo Grounds	Manhattan
MN21	Gramercy	Manhattan	MN06	Manhattanville	Manhattan
BK33	Carroll Gardens-Columbia Street-Red Hook	Brooklyn	QN70	Astoria	Queens
MN14	Lincoln Square	Manhattan	BK63	Crown Heights South	Brooklyn
BK37	Park Slope-Gowanus	Brooklyn	BK60	Prospect Lefferts Gardens-Wingate	Brooklyn
MN50	Stuyvesant Town-Cooper Village	Manhattan	BK77	Bushwick North	Brooklyn
MN40	Upper East Side-Carnegie Hill	Manhattan	BK79	Ocean Hill	Brooklyn
BK61	Crown Heights North	Brooklyn	QN20	Ridgewood	Queens

Figure 12: Table of NTA code and NTA name

8 References

1. Gebhart, K. and R.B. Noland, 2013. The impact of weather conditions on capital bikeshare trips. In: 21 Proceedings of the Transportation Research Board 92nd Annual Meeting