12-2021

A Study of the Impact of Infrastructures on Public Bicycle-Sharing System Demand

Lijie Zhou

A STUDY OF THE IMPACT OF INFRASTRUCTURES ON PUBLIC BICYCLE-SHARING SYSTEM DEMAND

THESIS

Presented in Partial Fulfillment of the Requirements for

the Master of Science Degree in the Graduate School

of Texas Southern University

By

Lijie Zhou, B.S.

Texas Southern University

2021

Approved By

Mehdi Azimi

Chairperson, Thesis Committee

Gregory H. Maddox

Dean, The Graduate School

Approved By:

Mehdi Azimi	10/28/2021
Ph.D., Chairperson of Thesis Committee	Date
Yi Qi	10/28/2021
Ph.D., Committee Member	Date
Fengxiang Qiao	10/28/2021
Ph.D., Committee Member	Date
Yachi Wanyan	10/28/2021
Ph.D., Graduate School Representative	Date

A STUDY OF THE IMPACT OF INFRASTRUCTURES ON PUBLIC BICYCLE-SHARING SYSTEM DEMAND

By

Lijie Zhou, M.S.

Texas Southern University, 2021

Assistant Professor Mehdi Azimi, Advisor

A bicycle-sharing system (public bicycle system, or bike-share scheme) is a service in which bicycles are made available for shared use to individuals on a short-term basis for a price or free. To study the impacts of bike infrastructures (particularly bike lanes and bike paths) on bicycle sharing system demand, autoregressive integrated moving average (ARIMA) models with exogenous factors are proposed to capture the relationships among the variables in a longitudinal analysis. The results show that every mile of bike lane added to the Houston bike share system will result in 38 average daily ridership increase in seven days of a week, 37 average daily ridership increase in weekdays (5 days of a week), and 82 average daily ridership increase on weekends. By adding every station to the public bikeshare system, the average daily ridership will increase by 16. The increase of average daily ridership will be 17 for weekdays, and 34 on weekends. It means that adding one bike station have less impact on average daily ridership of seven days of a week or five weekdays compared to the weekends. The impact of built environment on average daily ridership of weekends are almost

two times more when compared to its impact on the average weekday ridership, resulting in more demand during weekends. Regarding the influence of the weather variables, temperature and wind speed had no impact on the average daily trip counts in Houston. However, precipitation displayed a significant negative impact on the average daily ridership.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

VITA

2014-2018 .... B.E. in Construction Cost of Civil Engineering

Xuchang University (China)

2018-2021 .... Graduate Research Assistant

Department of Transportation Studies

Texas Southern University

Major Field .... Transportation Planning & Management

ACKNOWLEDGEMENT

This thesis was proposed based on the project funded by the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE @ UNC Charlotte), one of the Tier I University Transportation Centers that were selected in a nationwide competition by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT), under the FAST Act.

I would like to express my deep gratitude to the professors Dr. Qi and Dr. Azimi for providing me the opportunity to join the research project team. As my advisor, Dr. Azimi has taught and instructed me a lot on how to do the research, which helped me to have a better understanding about transportation engineering. Also, I would like to thank other committee members for participating in my thesis defense and Houston BCycle, City of Houston, and Harris County Precinct One for their assistance in data collection.

CHAPTER 1

INTRODUCTION

1.1 Background

Cycling has become an important transportation mode for work and recreational activities in major cities. Cities are increasingly promoting cycling as a valuable transportation alternative to driving, arguing that mode shift away from the private auto provides region-wide congestion, environmental, and health benefits. Whether due to health benefits, environmental factors, or financial reasons, more people are becoming bicycle commuters in big cities. In fact, cycling has grown in popularity as a primary means of transportation throughout the past decade. Between 1999 and 2011, total United States federal and state government funding on bicycling and pedestrian infrastructure exceeded $7 billion. The U.S. Federal Highway Administration, FHWA completed the Non-motorized Transportation Pilot Program in 2012, which allocated $25 million to each of four pilot cities over 5 years to measure the impacts of new infrastructure on mode shift to bicycling and walking. Among all the programs, bike-sharing, or public bicycle programs, have received increasing attention in recent years with initiatives to increase cycle usage, improve the first mile/last mile connection to other modes of transit, and lessen the environmental impacts of our transport activities. In 2016, Houston BCycle, a bikeshare service in Houston, secured a $3.5 million grant from the Federal Highway Administration, delivered through the Texas Department of Transportation, to fund expansion of the system. In 2017, the Houston City Council approved a biking “vision statement.” The document laid out a vision for 700 miles of on-street bike lanes and 450 miles of off-street trails. The council dedicated $1.1 million per year over four years to begin implementation. These were important shifts in one of the country’s most car-dependent cities, which sprawls over 627 square miles. According to the last Census count, only about 7 percent of city residents walk, bike, or take public transit to work. For years the city only had one protected bike lane cutting through the Central Business District. However, Houston residents seem responsive to the increase in bicycling resources. Bikeshare ridership has been increasing in the recent years, with riders mainly using bikes to replace short car trips. As a result, the installation of bicycle infrastructure is growing with more cyclists on the road. Houston BCycle is now working to add more new stations to underserved communities, and City and County are adding more bike lanes and infrastructure.

The previous research conducted by the research team has demonstrated the effect of the bike lane for bikeshare system in Houston. It measured the length of bicycle routes in a buffer around each bike station, which means the bike lane impacts only on the riders from/to that station. Therefore, it has not considered the bike infrastructure impact at the system level. The city planning agencies expect to know the future cycling patterns before carrying out bike lane expansion plan, and the bikeshare operations would like to forecast the system demand as the new bike infrastructure planned. It is necessary to explain how much bike-share ridership across the city will increase as a result of installing extra bike lanes. In order to measure the marginal cost of building bike lanes or bike paths on bikeshare demand at a network-wide level over time, time series models are needed to capture system-wide bike ridership

1.2 Research Objectives

The goal of this research is to identify how the bike infrastructure, especially bike lane investment, benefits bike-sharing system and demonstrate how the proposed method can provide new insights into system-level casualty and temporal lag characteristics with public agency infrastructure decisions. Specific project objectives include:

- Develop a longitudinal model to investigate the relationship between the number of the added bike lane miles and the bikeshare ridership;
- Illustrate how the public bike infrastructure benefits bike-sharing system in Houston; and
- Demonstrate how the proposed method can provide new insights into system-level casualty and temporal lag characteristics.

CHAPTER 2

LITERATURE REVIEW

In recent years, an increasing attention has been paid on how various factors, such as weather, built environment, and transportation infrastructure have impact on bikeshare ridership. This chapter provides a review on the previous research studies related to the topic and the results obtained from those studies. The studies related to built environment factors, public transit factors, socio economic factors, temporal factors, spatial factors, and weather factors has been among the popular studies in the past several years.

2.1 Bike Share

The role of cycling in the city transportation systems has attached increasing attention in recent years due, at least in part, to climate change, unstable fuel prices and concerns about global motorization. The increasing environmental problems have caused many decision makers and planners to closely examine the need for more sustainable transportation options. Bike share, a shared use of a bicycle fleet, stands out among different types of sustainable transportation modes. In the past decade, this evolving concept has gained increasing interests across the world. There are an estimated 500 cities in 49 counties that operate bikeshare systems. The global bikeshare fleet is estimated approximately at 550,000 bicycles by 2012 (Larsen, 2013). As a public bicycle system, it is initiated with the idea of increasing cycle usage, improve the first mile/last mile connection to other modes of transportation, and lessen the environmental impacts of transportation activities.

2.2 Built Environment Affecting Bike Share

The purpose of the current study is to explore the relationship of bikeshare ridership and bike infrastructures. Bicycle infrastructures include bike stations and bike lanes. Bike station provides equipment that users can check in and out bicycles. The development of bike lanes seems to have impacts directly on the usage of bikeshare system.

Xu and Chow (2019) conducted a longitudinal study of the relationship of bikeshare infrastructure and bikeshare system performance in New York City. They proved that the extension of bike lanes and building new bike station will lead to the increase of bikeshare daily trip. Autoregressive conditional heteroscedasticity (ARCH) model with autoregressive (AR) disturbance was proposed in their research to capture the relationship among variables. The dependent variable was average daily trip counts and independent variables were weather factor and building environment factors. Weather factors included precipitation, snow depth, temperature, and wind speed. Build environment factors included bike lane length and active station. The results showed that the installation of additional mile of bike lanes in New York City resulted in an average increase of 102 bikeshare daily trip. Moreover, there were 135 and 13 more trips generated per day when one more bikeshare station was added in Manhattan and non-Manhattan, respectively. In addition, weather factors could affect bikeshare daily trip, as well.

Tien et al. (2019) used robust linear regression model to model bike sharing demand at station level by using built environment factors in the city of Lyon. Independent variables were composed of public transportation factors, socio-economic factors, topographic factors, bikesharing network factors, and leisure factors. Public transportation variables included the number of metro, and tramway and railway stations. Socio-economic variables consisted of population, number of jobs, number of students in campus, and number of student residences near a bikeshare station. Topographic variable was altitude. Bike sharing network variable was made up of number of bike sharing stations and the capacity of each station. Leisure variables referred to number of restaurants, number of cinemas, and the presence of embankment road. The results showed that bike sharing was mainly used for commuting purpose by long term subscribers while short term subscriber's trip purposes were more changeable. And there seemed to be an important inter-modality that was the combination between the train and bike sharing. A fascinating finding was that students were important users of bike sharing. In addition, various bikesharing type were impacted by socio-economic factors, which relied on the period within the day and type of subscribers. Furthermore, the methodology used in the research could be applied to plan and operate existing bikeshare system and make an estimation of car-share demand in the future.

Built environment factors influence the usage of public bikesharing, taking account of the spatial correlation between nearby stations as well as the ratio of demand to supply (D/S) at bike station (Zhang et al., 2017). A multiple linear regression model was employed to analysis the relationship among various variables. A buffer of 300 meter around each bike station was considered as a reasonable walking distance. Descriptive variables included station attributes and accessibility, cycling infrastructure, public bus stops within 300 m buffer, and land use types within 300 m buffer. Station attributes and accessibility consisted of capacity of the bike station, number of other bike stations within 300 m buffer, distance to city government, population within 300 m buffer. Cycling infrastructure included bike lane within 1000 m buffer, main road within 300 m buffer, secondary road within 300 m buffer, and branch road within 300 m buffer. Public transport facilities were made up of public bus stops within 300 m buffer, distance to the closest public bus stop, whether closest stop is a bus terminal, whether closest stop is a transportation hub. Land use characteristics referred to land use types within 300 m buffer, whether near a shopping mall, whether near a residential community, whether near a recreational place, whether near a park. The results showed that trip demand and the ratio of demand to supply at bike station were positively impacted by population density, length of bike lanes and branch road, and diverse land-use types near the station, and were negatively influenced by the distance to the city center and the number of nearby bikeshare station.

Spatial and temporal factors, in a certain extent, can affect bikeshare ridership. Different time periods of a day and spatial layout could be other important factors that can influence bikeshare ridership. Faghih-Imani et al. (2016) employed spatial error and spatial lag models to accommodate for the influence of spatial and temporal interactions by using the data from New York City's bicycle-sharing system. A buffer of 250 meter around bikeshare station was adopted in their research considering the distance and bikeshare station density in the city. Descriptive variables included weather factors, temporal variables, and spatial variables. Weather variables referred to rainy and humidity factors. Considering the start time of the trips for departures and end time of the trips for arrivals, five time periods were created: AM (7:00–10:00), Midday (10:00–16:00), PM (16:00–20:00), Evening (20:00–24:00), and Night (0:00–7:00), which were temporal variables. Spatial variables included population density, employment density, length of bicycle routes and streets, presence of subway and path train stations, the number and capacity of CitiBike stations (excluding the origin/destination station), the number of restaurants (including coffee shops and bars), and area of parks that had been calculated at 250 meters station buffer level. The results showed that different time period variables significantly influenced bikeshare ridership. Annual members preferred to use bikeshare on weekday while short term user tended to ride bicycles on weekend. Furthermore, placing bikeshare station close to bike facilities could increase the usage of short-term users. However, because cyclists' movement was impacted by railway, the distance from bikeshare stations to railway station negatively influenced the usage of bikeshare. Moreover, the arrival and departure rate of short-term users were less sensitive to population and job density but more sensitive for long-term users.

Daddio (2012) employed OLS regression (Demand Estimation) and Raster Analysis to identify the determinants of bikeshare usage and estimate the demand of bikeshare ridership. The buffer around the bikeshare station was 400m. Trip generation, trip attraction, and transportation network are three primary impact factors that affected the bikeshare demand. Trip generation variables include age 20-39, non-white population, low-vehicle household prevalence, income, hotel rooms, and alternative commuters. Trip attractions referred to attractors, retailers, colleges, and parks. Transportation network factors referred to bus stops, metro rail, bike infrastructure, and distance from system center. The results of multi-regression revealed that five determinants significantly affect bikeshare ridership, which included: population (aged 20-39), non-white population, retail density (using alcohol licenses as a proxy), metro rail stations, and the distance from the center of the bicycle sharing system. In addition, Raster Analysis showed that about 13% bikeshare station experienced fewer than 18 trip each day, which suggested planners to consider avoiding low-needed bikeshare areas when they made decisions.

Wei Ding (2016) demonstrated the relationship of built environment and weather with bike sharing ridership by adopting polynomial regression and multiple linear regression. The principal purpose was to explore the function of temporal factors and weather and nearby built environment factors concerning station-level ridership. To investigate the effect of the scale, Wei used 0.25-mile, 0.5-mile, and 0.75-mile buffers to evaluate 5-min, 10-min, and 15-min walking distances and measured the transportation infrastructure, built environment, and socio demographic features of each bikeshare station. Weather factors, temporal variables, and built environment factors were among the variables. Weather factors referred to wind speed and temperature as well as four different level precipitation which included light, moderate, heavy, and violent. Built environment factors included origin, short-term holder, weekend, holiday, distance to park, distance to waterfront, distance to bike station, distance to light rail, job density, housing density, bike lane density, and bus density. The results showed that temperature and wind speed were not linearly associated with daily ridership. And rain, weekend, and holiday reduced the ridership of bikesharing. Various impacts of weather and temporal factors on annual members and short-term users were obtained, as well. Rainfall could more affected annual members than short-term users. Annual members seemed to use bikesharing on weekend while short-term users liked to use bikesharing on weekends and holidays. Moreover, bikeshare station-level ridership is negatively impacted by job density, proximity to park, and proximity to waterfront in all three buffers. All those can help planners to predict future bikeshare station and optimize the bikeshare system.

Beside of built environment factors, socio-economic, psychological, temporal factors, etc. can also impact ridership over time. Seasonal trend is a common factor that affects bikeshare ridership over time. Gallop et al. (2011) used seasonal autoregressive integrated moving average (ARIMA) analysis to account for the relationship of time serial correlation and bikeshare ridership. Weather variables included temperature, relative humidity, wind speed, clearness, fog, three level precipitation (drizzle, rain, and snow). Temporal variables referred to holidays, Olympics, Saturday, and Sunday. The results displayed that temperature, rain, rain in the previous 3 hours and humidity were significant, and clearness was found to be marginally significant at the 10% level. The influence of rain and its lags was close to 24% of the average hourly bicycle traffic counts. Bicycle accounts would increase 16.5% with each degree Celsius rise from the average level. The humidity and clearness were less significant on bikeshare account. There was 0.08% decrease in bicycle traffic per unit change in relative humidity; and 0.62% bike traffic increased at each of the four transitions between cloudy and clear skies. Snow was not significant based on their study.

2.3 Summary

Existing studies have been mostly on the impact of bike infrastructures on ridership at station level. They measured the length of bicycle routes in a buffer around each bike station, which means the bike lane impacts only on the riders from/to that station. Therefore, they have not considered the bike infrastructure impact at the system level in a city such as Houston. The city planning agencies expect to know the future cycling patterns before carrying out bike lane expansion plan, and the bikeshare operations would like to forecast the system demand as the new bike infrastructure planned. It is necessary to explain how much bike-share ridership across the city will increase as a result of installing extra bike lanes. In order to measure the marginal cost of building bike lanes or bike paths on bikeshare demand at a network-wide level over time, time series models are needed to capture system-wide bike ridership.

CHAPTER 3

DESIGN OF THE STUDY

3.1 Data Description

In this chapter, the data required for the analysis will be introduced in two sections. Section 3.1 describes the bikeshare ridership data as well as the data related to the bike lanes that were added to the bike infrastructure in Houston. To do the analysis, it is required to select the model variables. The dependent variable and independent variables will be described in detail in Section 3.2 of this chapter.

3.1.1 Data Processing

The bikeshare ridership data used in this study was provided by Houston BCycle. Houston BCycle, which owns and operates bike-sharing system in Houston, was first launched in May 2012 with only three bike station and 18 bikes and have been expanding to 1000 bikes at 100 stations by 2020. The dataset of 2019 bikeshare records includes trip ID, user program name, user ID, user role, user city, user state, user zip, user country, membership type (annual membership, single-use pass, and monthly membership), bike ID, bike type, check-out kiosk name, return-kiosk name, duration minutes, adjusted duration minutes, usage fee, adjustment flag, distance, estimated carbon offset, estimated calories burned, check-out data location, return time location, trip over 30 minutes, local program flag, trip route category (round trip or one way), and trip program name. Among those variables, the ones required for the study were selected. Besides, the data with the trip duration less than one minute were removed from the analysis since a one-minute trip duration could not be considered as a valid trip.

Ridership was counted by average per week, because people need time to realize the new bike infrastructure near them. Active bike share station counts was based on if there are ridership in the bike station; in some area, it might happen that no ridership all day In the selected dataset, daily trip counts and daily active stations were calculated from trip IDs and check-out kiosk names, individually. It should be noted that more information could also be obtained from the data, such as which bikeshare station is the most popular, which month has the most rider/user, etc. We didn’t include that information here since it was beyond the scope of this study. The average daily bikeshare trips per week was generated from January 1st to December 31st, 2019. There was a total of 53 observations (weeks) in 2019, with an average weekly ridership of 646 trips and 77 active stations. The 53rd week (the last week of the year) had only two days. Therefore, the data corresponding to that week were eliminated and 52 weeks of the dataset were chosen for the analysis.

Figure 1 shows the sum of ridership counts and the average daily ridership counts per week in 2019. From Figure 1, one can see that there are higher bikeshare activities between the 11th and 19th weeks, which were the second half of March, the whole month of April, and the first week of May. The highest ridership trip count was in the 16th week, which was more than 7,000 trips. And less bike sharing demands occurred in the period of the 3rd to 11th weeks (middle of January to middle of March). The lowest bikeshare ridership demand appeared in the 9th week as well as 51st weeks, which approached 2,500 trips. A steady trend describes the change of the bike-sharing ridership demands for the other weeks.

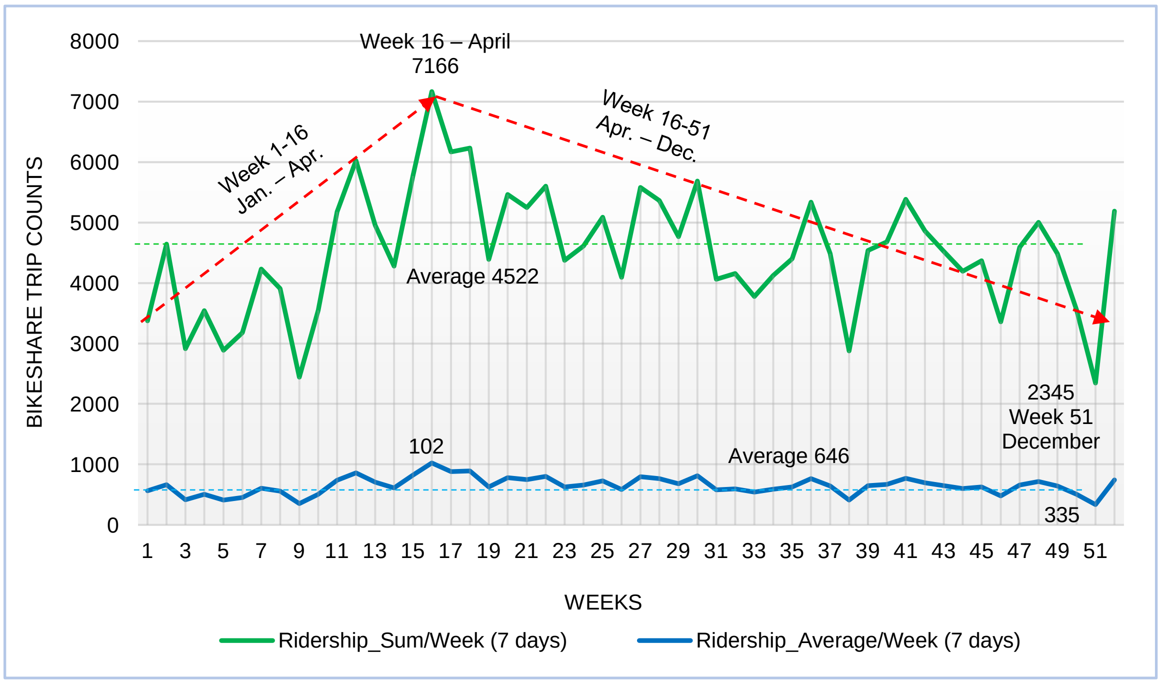


Figure 1. Houston Bikeshare Trip Counts per Week in 2019

Figure 2 shows the bike sharing ridership daily counts per weekday (5 days) and daily bike sharing ridership counts per weekend (2 days) in 2019. From Figure 2, one can see that the daily bike sharing ridership counts per weekend is higher compared to the daily bike-sharing counts per weekday, which means more people used the bike sharing facilities at the weekends. There are higher bike-sharing demands between the 11th and 19th weekends, and the highest daily trip count was approximately 1500 at the 16th weekend. Lower bikeshare demands occurred in the period of the 3rd to 7th weekends, and the lowest daily trip demands emerged at the 6th weekend and 51st weekend, which were close to 500 trips per day. For the weekday bikeshare ridership demands, there are higher bike-sharing demands between 10th and 36th weeks, and the highest daily trip count is approximately 900 at the 16th week. Lower bikeshare demands occurred in the period of the 1st to 10th weeks and also 37th to 52nd weeks. Furthermore, the lowest daily trip demands emerged at 38th week and 51st week, which were close to 300 trips per day. The other weeks displayed a stationary trend.

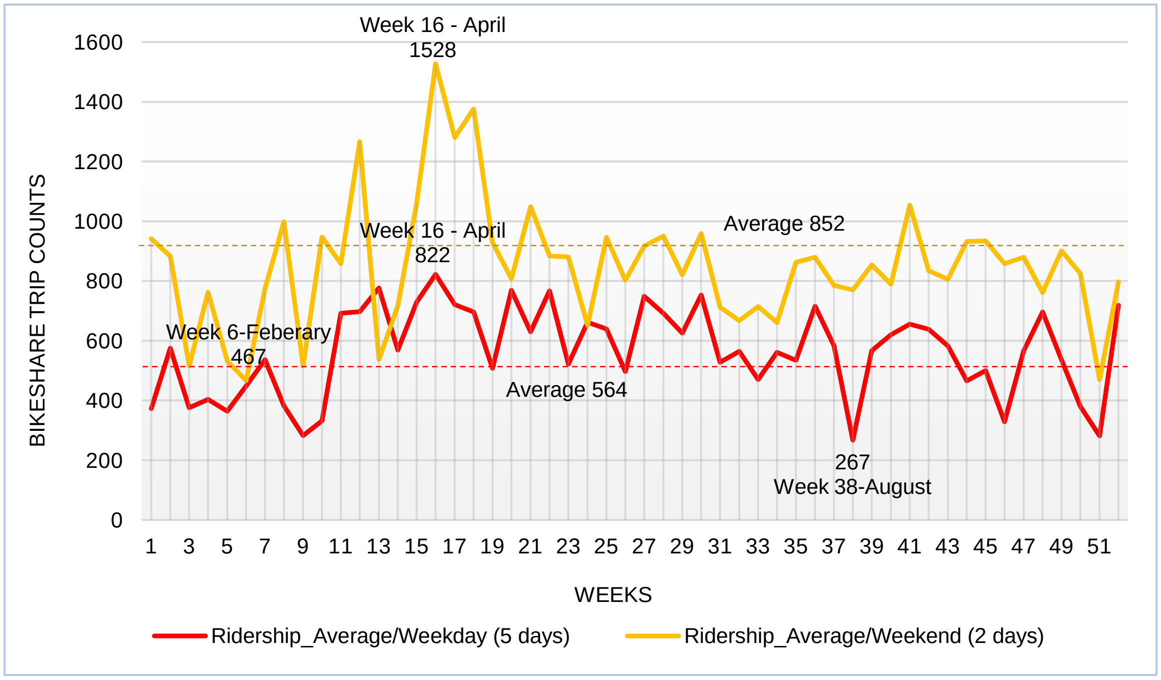


Figure 2. Houston Bikeshare Daily Trip Counts per Week in 2019

Figure 3 shows the sum of bike-sharing ridership counts per weekday (5 days) and the sum of bikeshare ridership counts per weekend (2 days) in 2019. From Figure 3, one can see that the total bike-sharing ridership counts per weekday are higher than the total bike-sharing counts per weekend. The average sum of weekday ridership is 2820, while the average sum of weekend ridership is

1704. There are higher bike-sharing demands between the 11th and 35th weeks, and the highest trip count per weekday was approximately 4100 at the 16th weekday. Lower bikeshare demands per weekday occurred in the period of the 1st to 10th weekdays and 37th to 52nd weekdays. The lowest total trip demands emerged at the 38th weekday, which was close to 1300 trips total. For the weekend bikeshare ridership demands, there are higher bike-sharing demands between the 11th and 19th weekends, and the highest daily trip count was approximately 1500 at the 16th weekend. Lower bikeshare demands occurred in the period of the 3rd to 7th weekends, and the lowest daily trip demands emerged at the 6th and 51st weekends, which were close to 900 trips per day. The other weeks displayed a stationary trend.

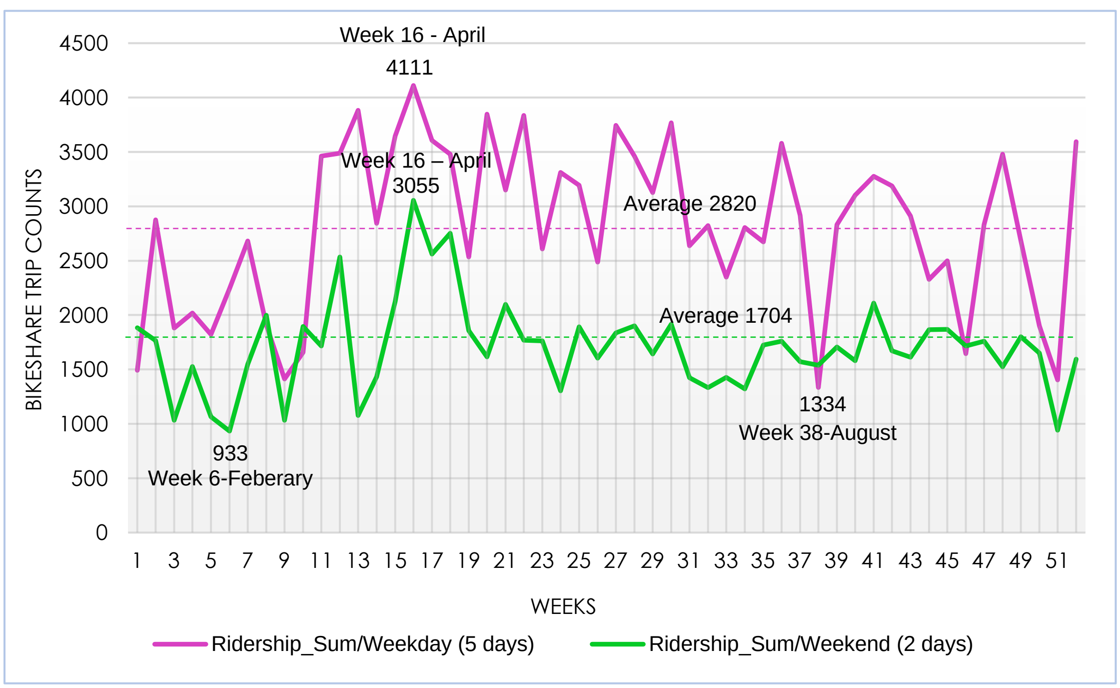


Figure 3. Houston Bikeshare Sum of Daily Trip Counts per Week in 2019

From Figure 2, and Figure 3, one can that the ridership of weekday and weekend are clearly different. In order to explore and compare the ridership between weekday and weekend, three scenarios are considered in the paper:

Scenario 1: 7 days from Monday to Sunday (week)

Scenario 2: 5 days from Monday to Friday (weekday)

Scenario 3: 2 days from Saturday to Sunday (weekend)

In 2019, around 15.5 miles of bike lane were added to the bike system in Houston. For this study, the data representing the characteristics of each segment of the added bike lanes were received from Harris County and City of Houston. The data included the lengths, types, locations, and dates of completion of each bike lane segments (Table 1). By checking the location of the constructed bike lanes, it was found that the bike lane projects were mainly located in either Near Northside or Third Ward of the Greater Houston Area. The projects included two types of bike lane facilities: dedicated protected bike lanes and shared use streets.

Table 1. Bike Lanes Added to the System in Houston in 2019

Date	**Location**	**Type**	**Miles**
01/30/19	Hardy to Maury	Dedicated Protected Bike Lane	0.10
01/30/19	Maury to Freight Rail	Dedicated Protected Bike Lane	0.05
01/30/19	Freight Rail to US 59 SB Frontage Road	Dedicated Protected Bike Lane	0.55
03/29/19	Hardy to Elysian	Dedicated Protected Bike Lane	0.05
03/29/19	Elysian to Maury	Dedicated Protected Bike Lane	0.05
03/29/19	Lyons to McKee	Dedicated Protected Bike Lane	0.30
03/29/19	Hardy/Sterrett to Runnels	Dedicated Protected Bike Lane	0.55
04/11/19	Kelley to Orr	Dedicated Protected Bike Lane	1.65

Date	Location	Type	Miles
04/12/19	Kelley to Euel	Dedicated Protected Bike Lane	0.05
04/12/19	Euel to Orr	Dedicated Protected Bike Lane	1.60
04/12/19	Orr to Line	Dedicated Protected Bike Lane	0.30
04/12/19	Line to Campbell	Dedicated Protected Bike Lane	0.15
04/12/19	Campbell to Sumpter	Dedicated Protected Bike Lane	0.20
04/12/19	Sumpter to Lorraine	Dedicated Protected Bike Lane	0.10
10/22/19	HARDY TO LBJ HOSPITAL	Dedicated Protected Bike Lane	2.48
03/29/19	Lorraine to Lyons	Shared Use Street	0.20
05/30/19	CAROLINE TO LIVE OAK	Dedicated Protected Bike Lane	0.81
06/26/19	CLEBURNE TO GRAY	Dedicated Protected Bike Lane	1.06
12/18/19	HOLMAN TO COMMERCE	Dedicated Protected Bike Lane	2.00
04/10/19	AUSTIN TO ENNIS	Shared Use Street	1.00
04/15/19	BLODGETT TO ALABAMA	Shared Use Street	0.47
04/15/19	SOUTHMORE TO BLODGETT	Shared Use Street	0.31
04/15/19	HERMANN TO SOUTHMORE	Shared Use Street	0.38
05/30/19	LIVE OAK TO SAUER	Shared Use Street	0.12
05/30/19	SAUER TO NETTLETON	Shared Use Street	0.21
		Total	**15.5**

Table 2 displays the total added bike lane length in miles in Houston from January 2019 to December 2019, and Table 3 shows the total added bike lane length in miles by location in the city during the same time. The total length of the dedicated protected bike lanes is 12.81 miles (see Table 2), which accounts for 83% of the total constructed bike lane length.

Table 2. Total Added Bike Lane Length (in miles) by Type in Houston from January to December 2019

Miles added by type	Calendar Year 2019	% of total
Dedicated Protected Bike Lane	12.81	0.83
Shared Use Street	2.69	0.17

The bike lane length of the shared use street is 2.69 miles, which is 17% of the total added bike lane length. The total lengths of bike lanes constructed in Near Northside and Third Ward are 9.14 and 6.37 miles, respectively, accounting for 59% and 41% of the total added bike lanes (Table 3).

Table 3. Total Added Bike Lane Length (in miles) by Location from January to December 2019

Miles added by location	Calendar Year 2019	% of total
Near Northside	9.14	0.59
Third Ward	6.37	0.41

Figure 4 illustrates the pattern of increased bike lane in Houston from January 2019 to December 2019. In the first 12 weeks, the rate of the increase in bike lane length is quite slow. However, from 13th week to 16th week, we can see a major change in the length of bike facilities; approximately 7 miles of new bike lane was constructed and added to the system. Then, a slower growth of bike lanes occurred from weeks 17 to 43 (around 3 miles bike lane increasing during 26 weeks). A relatively long total segment of bike lanes was added to the bicycle infrastructure within only one week (between the week 43 and 44). At the final weeks of 2019, about two miles of bike facility was constructed and opened to the users. Besides, the dotted line shown in Figure 4 reveals the increasing trend of bike lane length in 2019.

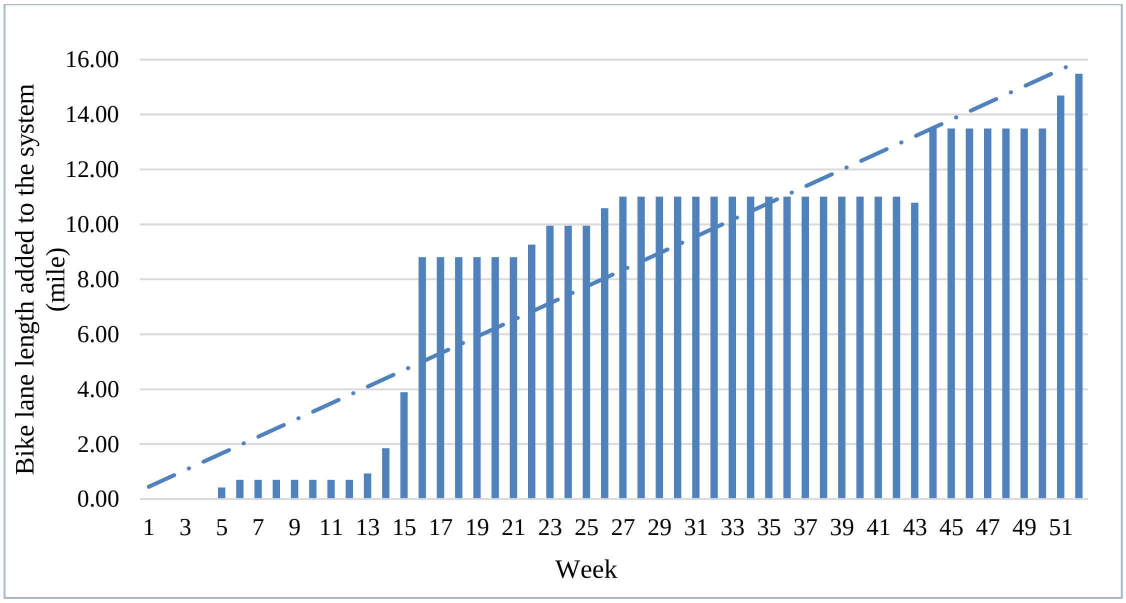


Figure 4. The Increased Bike Lane in Houston from January 2019 to December 2019

In order to investigate and compare the impact between weekday and weekend, the data were prepared for the analysis in three different formats: the data corresponding to 7 days of a week, the data corresponding to 5 days of a week (weekdays), and the data corresponding to 2 days of a week (weekend).

The variables studies in the research including daily ridership per week, added bike lane length (total length of added bike lane length), active station (means that the station has ridership), temperature, wind speed, and precipitation. Table 4 lists bikeshare variables data for each week (7 days) from week 1 to 52 in

2019. It shows that the average weekly ridership of 52 weeks is 646, the average weekly added bike lane is 0.28 miles; the average weekly active station is 77; the average temperature is 73.54 F; the average wind speed is 8.86 mph; and the average precipitation is 0.142 inch.

Table 4. Bikeshare Variables for Each Week (7 Days) from Week 1 to Week 52 in 2019

Week	Daily Trip Counts per Week	Bike Lane Length (mile)	Active Stations	Temperature (F)	Wind Speed (mph)	Precipitation (inch)
1	563	0	60	83.20	6.98	0.050
2	663	0	71	86.86	8.93	0.000
3	416	0	68	85.69	9.19	0.040
4	506	0	66	83.50	8.46	0.010
5	412	0.5	68	74.33	7.87	0.100
6	454	0.7	70	60.14	10.53	0.010
7	604	0.7	74	62.86	9.17	0.060
8	558	0.7	72	57.57	10.76	0.060
9	349	0.7	62	59.29	8.39	0.290
10	507	0.7	68	58.16	9.94	0.010
11	740	0.7	75	65.21	10.80	0.060
12	860	0.7	79	64.07	7.21	0.000
13	708	1.19	80	65.39	10.17	0.010
14	611	1.85	77	65.29	9.00	0.010
15	825	4.96	81	71.00	11.89	0.140
16	1024	8.81	80	68.00	11.23	0.130
17	881	8.81	81	73.27	9.73	0.030
18	890	8.81	86	76.97	11.50	0.180
19	627	8.81	84	74.31	10.03	1.250
20	781	8.81	81	77.79	9.70	0.000
21	749	8.81	79	82.26	14.03	0.000
22	800	9.46	78	81.83	8.11	0.170
23	624	9.95	76	82.89	7.49	0.390

Week	Daily Trip Counts per Week	Bike Lane Length (mile)	Active Stations	Temperature (F)	Wind Speed (mph)	Precipitation (inch)
24	659	9.95	77	82.96	9.26	0.020
25	727	9.95	77	84.26	10.69	0.190
26	584	10.71	73	79.86	7.31	0.440
27	797	11.01	76	83.54	6.93	0.040
28	766	11.01	78	86.61	9.03	0.040
29	681	11.01	76	85.89	9.14	0.000
30	812	11.01	78	83.09	8.69	0.010
31	580	11.01	77	84.96	6.17	0.000
32	594	11.01	78	87.00	8.24	0.000
33	540	11.01	75	86.77	7.50	0.290
34	589	11.01	80	83.50	7.37	0.580
35	628	11.01	80	84.93	6.19	0.140
36	763	11.01	84	85.90	6.89	0.000
37	641	11.01	81	84.83	7.20	0.050
38	411	11.01	70	80.19	9.97	1.580
39	648	11.01	82	82.44	7.46	0.010
40	669	11.01	84	83.56	6.97	0.000
41	769	11.01	86	73.63	9.86	0.070
42	694	11.01	87	74.77	7.66	0.090
43	646	11.56	83	66.60	10.26	0.370
44	599	13.49	82	58.70	9.96	0.070
45	624	13.49	80	62.56	8.19	0.240
46	480	13.49	76	49.01	8.63	0.040
47	656	13.49	85	64.49	8.26	0.030
48	715	13.49	78	69.89	12.81	0.000
49	641	13.49	82	60.79	6.43	0.000
50	507	13.49	74	61.59	9.97	0.080
51	335	14.92	69	51.66	9.00	0.000
52	741	15.49	73	61.63	6.79	0.000
Average	646	0.28	77	73.54	8.86	0.142

Table 5 includes bikeshare variables data for weekdays (5 days) from week 1 to 52 in 2019. The table shows that the average weekly ridership of 52 weeks (weekdays) is 564, the average weekly added bike lane is 0.28 miles; the average weekly active station is 77; the average temperature is 74 F; the average wind speed is 9.07 mph; and the average precipitation is 0.174 inch.

Table 5. Bikeshare Variables Data for Weekdays (5 Days) from Week 1 to Week 52 in 2019

Week	Daily Trip Counts per Week	Bike Lane Length (mile)	Active Stations	Temperature (F)	Wind Speed (mph)	Precipitation (inch)
1	373	0.00	55	82.33	7.25	0.070
2	575	0.00	71	86.80	7.36	0.000
3	376	0.00	69	85.44	9.30	0.058
4	404	0.00	65	82.94	8.12	0.014
5	364	0.42	68	79.40	8.10	0.086
6	449	0.70	72	63.94	10.98	0.018
7	536	0.70	75	62.34	9.36	0.082
8	382	0.70	70	54.78	11.02	0.082
9	282	0.70	62	58.80	7.70	0.408
10	331	0.70	65	52.52	10.62	0.014
11	692	0.70	75	68.68	12.26	0.072
12	698	0.70	78	62.72	6.30	0.000
13	776	0.93	83	68.10	8.92	0.000
14	569	1.85	78	62.80	8.32	0.012
15	730	3.89	83	72.68	10.94	0.166
16	822	8.81	80	68.38	12.74	0.188
17	721	8.81	81	73.24	9.96	0.042
18	696	8.81	85	77.70	13.12	0.000
19	507	8.81	82	75.54	11.14	1.492
20	770	8.81	83	76.24	8.40	0.000
21	630	8.81	82	82.36	15.06	0.000
22	767	9.27	79	81.32	9.50	0.198
23	522	9.95	76	81.34	8.10	0.548
24	662	9.95	80	83.08	7.70	0.000

Week	Daily Trip Counts per Week	Bike Lane Length (mile)	Active Stations	Temperature (F)	Wind Speed (mph)	Precipitation (inch)
25	639	9.95	79	83.40	9.32	0.228
26	497	10.59	74	80.38	7.80	0.564
27	749	11.01	78	82.84	7.36	0.056
28	692	11.01	80	87.64	8.58	0.000
29	625	11.01	78	85.88	9.72	0.000
30	753	11.01	80	82.58	8.50	0.014
31	528	11.01	80	85.04	6.64	0.002
32	564	11.01	80	86.72	8.12	0.004
33	470	11.01	75	86.88	7.02	0.372
34	561	11.01	81	84.08	7.64	0.578
35	535	11.01	81	84.74	6.32	0.186
36	716	11.01	84	86.52	7.26	0.000
37	583	11.01	82	84.62	6.86	0.064
38	267	11.01	66	79.28	10.56	2.210
39	566	11.01	82	81.84	6.44	0.000
40	620	11.01	85	83.84	7.80	0.000
41	655	11.01	87	76.34	10.30	0.000
42	638	11.01	88	73.16	8.06	0.132
43	582	10.79	84	68.96	11.28	0.486
44	466	13.49	81	59.60	11.52	0.094
45	500	13.49	79	64.64	9.32	0.332
46	329	13.49	72	47.24	11.02	0.050
47	566	13.49	84	68.02	8.64	0.008
48	696	13.49	78	70.52	12.56	0.002
49	537	13.49	83	60.58	6.54	0.000
50	380	13.49	72	59.16	9.94	0.116
51	281	14.69	69	51.56	9.30	0.000
52	719	15.49	74	59.44	5.10	0.000
Average	564	0.28	77	74.02	9.07	0.174

Table 6 lists bikeshare variables data for weekends (2 days) from week 1 to 52 in 2019. The table shows that the average weekly ridership of 52 weeks (weekends) is 852, the average weekly added bike lane is 0.28 miles; the average weekly active station is 76; the average temperature is 73.77 F; the average wind speed is 8.54 mph; and the average precipitation is 0.064 inch.

Table 6. Bike Sharing Variables Data for Weekends (2 Days) from Week 1 to Week 52 in 2019

Week	**Daily Trip Counts per Week**	**Bike Lane Length (mile)**	**Active Stations**	**Temperature (F)**	**Wind Speed (mph)**	**Precipitation (inch)**
1	942	0.00	72	84.95	6.45	0.000
2	883	0.00	71	87.00	12.85	0.000
3	517	0.00	64	86.30	8.90	0.000
4	763	0.00	69	84.90	9.30	0.000
5	532	0.70	70	61.65	7.30	0.145
6	467	0.70	65	50.65	9.40	0.000
7	775	0.70	73	64.15	8.70	0.005
8	999	0.70	78	64.55	10.10	0.010
9	516	0.70	63	60.50	10.10	0.000
10	947	0.70	76	72.25	8.25	0.000
11	858	0.70	75	56.55	7.15	0.025
12	1267	0.70	81	67.45	9.50	0.000
13	538	1.85	70	58.60	13.30	0.050
14	717	1.85	74	71.50	10.70	0.005
15	1062	7.65	76	66.80	14.25	0.080
16	1528	8.81	80	67.05	7.45	0.000
17	1280	8.81	82	73.35	9.15	0.000
18	1376	8.81	88	75.15	7.45	0.620
19	929	8.81	88	71.25	7.25	0.640
20	808	8.81	77	81.65	12.95	0.015
21	1049	8.81	73	82.00	11.45	0.000
22	884	9.95	74	83.10	4.65	0.110
23	881	9.95	74	86.75	5.95	0.000
24	651	9.95	70	82.65	13.15	0.080

Week	Daily Trip Counts per Week	Bike Lane Length (mile)	Active Stations	Temperature (F)	Wind Speed (mph)	Precipitation (inch)
25	946	9.95	74	86.40	14.10	0.090
26	803	11.01	71	78.55	6.10	0.130
27	918	11.01	71	85.30	5.85	0.000
28	950	11.01	73	84.05	10.15	0.145
29	821	11.01	72	85.90	7.70	0.000
30	959	11.01	75	84.35	9.15	0.000
31	713	11.01	69	84.75	5.00	0.000
32	667	11.01	74	87.70	8.55	0.000
33	714	11.01	73	86.50	8.70	0.085
34	661	11.01	77	82.05	6.70	0.590
35	863	11.01	80	85.40	5.85	0.010
36	880	11.01	85	84.35	5.95	0.000
37	786	11.01	79	85.35	8.05	0.000
38	770	11.01	80	82.45	8.50	0.000
39	854	11.01	83	83.95	10.00	0.040
40	790	11.01	82	82.85	4.90	0.000
41	1055	11.01	86	66.85	8.75	0.235
42	835	11.01	85	78.80	6.65	0.000
43	806	13.49	79	60.70	7.70	0.090
44	933	13.49	84	56.45	6.05	0.000
45	935	13.49	83	57.35	5.35	0.010
46	858	13.49	86	53.45	2.65	0.000
47	880	13.49	85	55.65	7.30	0.080
48	762	13.49	76	68.30	13.45	0.000
49	901	13.49	80	61.30	6.15	0.000
50	825	13.49	79	67.65	10.05	0.000
51	471	15.49	72	51.90	8.25	0.005
52	798	15.49	72	67.10	11.00	0.015
Average	852	0.28	76	73.77	8.54	0.064

3.1.2 Variable Selection

The goal of this research is to study the impact of bike lanes on bikeshare ridership. Therefore, the dependent and independent variables will be the ridership and the length of bike lanes that are added to the bike system, respectively. In addition to the length of bike lanes, active stations and weather factors are the other independent variables that must be considered in the study. Weather variables refer to precipitation, wind, and temperature.

In the process of selecting variables, it is very important to perform a multicollinearity test to reduce the errors that may occur in the model and to obtain more precise results. Variance Inflation Factor (VIF) is a value that estimates the correlation among independent variables and identifies how much error or variance of a regression coefficient would be inflated caused by multicollinearity in the process of the modeling. When the numeric VIF is smaller than 2.5, it means independent variables are not correlated. A VIF between 2.5 and 10 implies a moderate correlation among independent variables. A high correlation will be indicated if the VIF is bigger than ten.

Bikeshare related data was tested for three time periods: 7 days of a week, 5 days of a week (weekdays), and 2 days of a week (weekend). Table 7 presents the results of the multicollinearity tests for these three time periods. As seen in the table, the biggest VIF of the independent variables for each time period is 2.08, 1.83, and 1.78. Most of the VIFs are smaller than 2.5, which means that there are no significant correlations among those independent variables. In other words, all these independent variables are independent of each other.

Table 7. Multicollinearity Test Results (Variance Inflation Factor)

Time-Period	Bike Lane Length	Active Stations	Temperature	Wind Speed	Precipitation
week (7d)	**2.08**	1.92	1.20	1.18	1.09
weekday (5d)	**1.83**	1.81	1.31	1.26	1.20
weekend (2d)	1.58	**1.78**	1.16	1.15	1.14

Table 8 shows the characteristics of the variables used in the analysis, including the definition of the variables, the source of the data, and the statistical summary of the variables (minimum, maximum, mean, and standard deviation). The dependent variable is the average daily ridership per week. There are two types of independent variables: built environment factors and weather factors. The built environment factors include the total length of bike lanes length added to the bike system (in miles) and average daily active stations per week. The weather factors include the average daily temperature per week in Fahrenheit, the daily average wind speed per week (in miles per hour), and the average precipitation per week (in inches). The data of bikeshare ridership and active stations were received from Houston BCycle. Harris County provided the lengths of bike lanes and the Houston weather data were downloaded from the Weather Underground website.

Table 8. Characteristics of Variables used in the Analysis

Variables	Definition	Source	Min	Max	Mean	Std. Dev
DEPENDENT						
Average Daily Trip Counts	Average daily trips per week	Houston BCycle	335	1024	647	143
INDEPENDENT						
Built Environment Factors						
Bike Lane Length	Total length of bike lanes (miles) added to the bike system	Harris County Precinct One and City of Houston				
Active Stations	Average number of daily active stations per week	Bikesharing company: Houston BCycle	60	87	76.87	6.05
Weather Factors						
Temperature	Average daily temperature per week (F)	Weather Underground website: *https://www.wunderground.com/about/data*	49	87	73.96	10.92
Wind Speed	Daily average wind speed per week (mph)	Weather Underground website: https://www.wunderground.com/about/data	6.2	14	8.927	1.71
Precipitation	Average precipitation per week (inch)	Weather Underground website: https://www.wunderground.com/about/data	0	1.58	0.142	0.29

3.2 Methodology

A time series regression analysis method was applied to develop the model(s) using the data provided by Houston BCycle, Harris County, and City of Houston. *R* language was the main coding tool used in the study to help in statistical analysis and modeling. Time series models are usually used to estimate the variance of time-dependent variables based on their previous data. To better compare and understand the bikeshare ridership changes between weekdays and weekends, the dataset was divided into three models; the specifications of those three models are exhibited as below:

Model 1: Ridership on average over seven continuous days per week

Model 2: Ridership on average over five continuous weekdays per week

Model 3: Ridership on average over two continuous days of weekends per week

There are many types of time series models for analysis of large datasets; and model selection should rely on various characteristics of the respective and appropriate test results. Figure 5 displays the procedures of how to select time series models, including stationary test, Breusch-Pagan test, ACF & PACF, model selection, and model validation.

Figure 5. Procedure of Selecting a Time Series Model

The total observations of data are 52 weeks, which will be divided into two sets of data: training data and testing data. Training data were used to develop a fitted model, and it was the data corresponding with the 1st week to 48th week. Testing data were utilized to check the accuracy of the model, ad it was the data corresponding to the 49th to 52nd weeks. The training data were used in the process of developing and selecting the model, while the testing data were implemented only in the process of model validation.

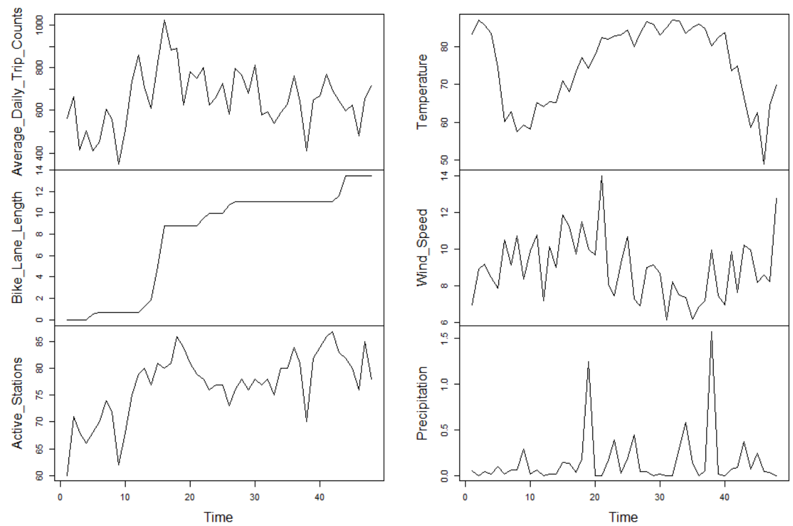
3.2.1 Stationary Test

The stationarity of the data should be checked for the time series analysis, in the first step. Stationarity is an important concept in time series analyses. Time series models usually assume that each point is independent of one another. The best indication of this is when the dataset is stationary. For data to be stationary, the statistical properties of a system should not change over time. The null hypothesis of the Dickey-Fuller test is the variables contain a unit root in a time series model, which means that the variables are not stationary. And 5% significant level is the threshold for the test. The results of the stationary test for three proposed models are shown in Table 9. The p-values of the variables in the three models are greater than 0.05, which indicates that the original variables are not stationary (the null hypothesis cannot be rejected). To eliminate non-stationary, differences method should be applied to the variables. For example, the original dataset are y1, y2, y3, y4, …; new dataset of first difference (d=1) are y'1=(y2-y1), y'2=(y3-y2), y'3=(y4-y3), …; the dataset of second difference (d=2) are y"1=(y'2-y'1), y"2=(y'3-y'2), y"3=(y'4-y'3), … After the first difference, the p-values of some of the variables are still greater than the 0.05 significant level, as shown in Table 9. Therefore, the second difference is necessary to be implemented. The results of the second difference for the three models are ideal, having the p-values yield to 0.01. Therefore, all the variables became stationary after two times difference.

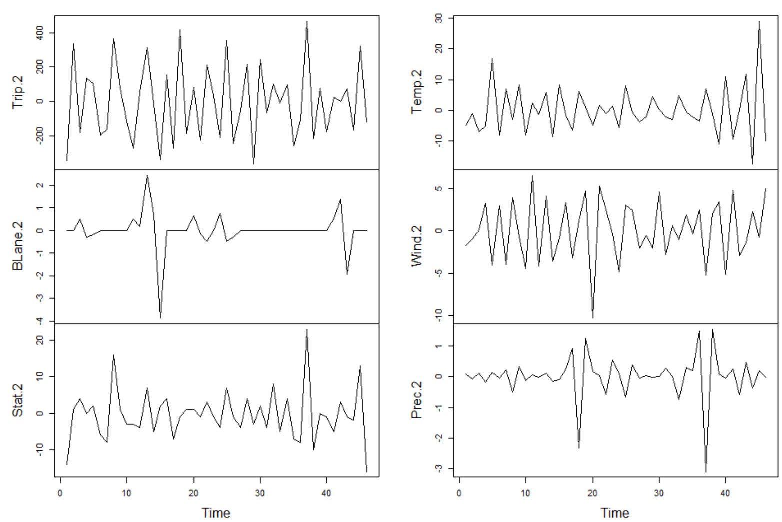
Table 9. Stationary Test Results (Dickey-Filler Test) for the Proposed Models

Model	Output				Null Hypothesis	Result
	Variable	***p*-value**				Original
		Original	**d=1**	**d=2**		Not reject
Model_1 (7 days a week)	Average Daily Trip Counts	0.58	<0.01	<0.01	The variable contains a unit root in an autoregressive model	
	Bike Lane Length	0.67	0.07	<0.01		d=1
	Active Stations	0.42	<0.01	<0.01		Not reject
	Temperature	0.13	0.29	<0.01		
	Wind Speed	0.7	<0.01	<0.01		**d=2**
	Precipitation	0.14	<0.01	<0.01		**Reject**
	Variable	***p*-value**				Original
		Original	**d=1**	**d=2**		Not reject
Model_2 (5 days a week)	Average Daily Trip Counts	0.41	<0.01	<0.01	The variable contains a unit root in an autoregressive model	
	Bike Lane Length	0.71	0.08	<0.01		d=1
	Active Stations	0.29	<0.01	<0.01		Not reject
	Temperature	0.25	0.42	0.02		
	Wind Speed	0.6	<0.01	<0.01		**d=2**
	Precipitation	0.08	<0.01	<0.01		**Reject**
	Variable	***p*-value**				Original
		Original	**d=1**	**d=2**		Not reject
Model_3 (2 days a week)	Average Daily Trip Counts	0.44	<0.01	<0.01	The variable contains a unit root in an autoregressive model	
	Bike Lane Length	0.7	0.08	<0.01		d=1
	Active Stations	0.49	0.03	<0.01		Not reject
	Temperature	0.47	0.05	<0.01		
	Wind Speed	0.04	<0.01	<0.01		**d=2**
	Precipitation	0.05	<0.01	<0.01		**Reject**

Figure 6, Figure 7, and Figure 8 show the change of the time series plot of original variables and stationary variable plot with two times difference for Model 1, Model 2, and Model 3, respectively. From Figure 6 (a), Figure 7 (a), and Figure



(a)

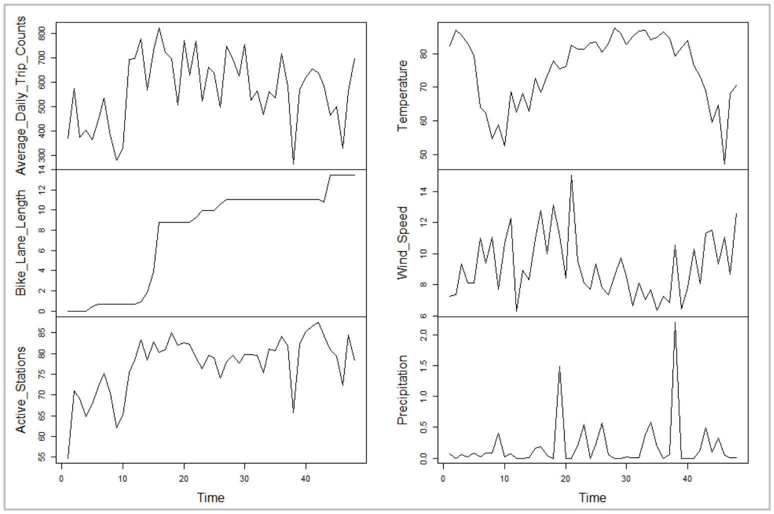


(b)

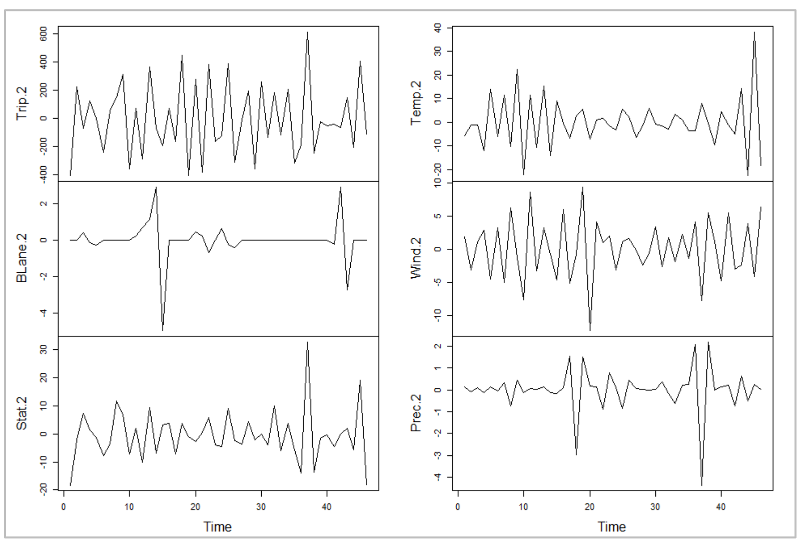
Figure 6. (a) Time Series Plot of Original Data, (b) Time Series Plot of Stationary Data after the Second Difference for Model 1

8 (a), by checking the trend in the plot, one can see that those six variables (average daily trip counts, bike lane length, wind speed, precipitation, and temperature) are not stationary for three models. However, the very steady change

shown in Figure 6 (b), Figure 7 (b), and Figure 8 (b) indicate that those six variables are stationary after the second difference.

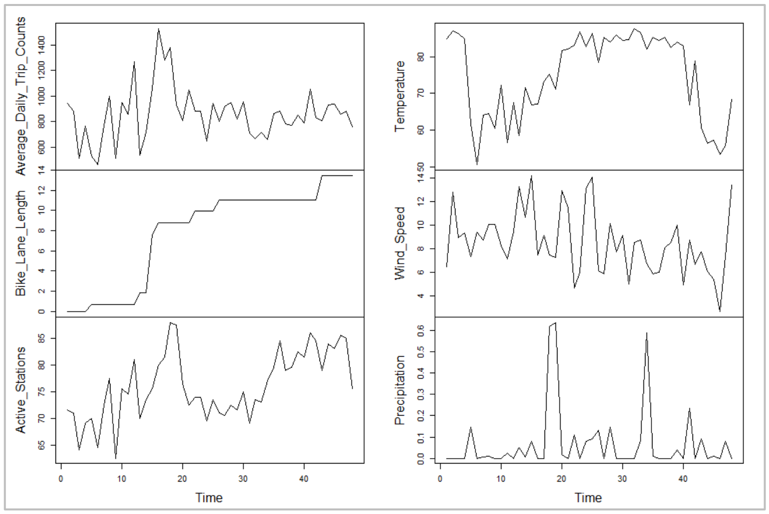


(a)

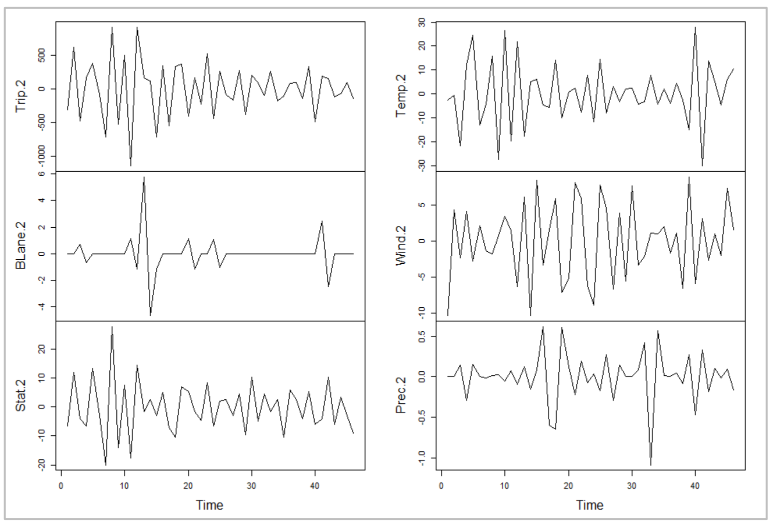


(b)

Figure 7. (a) Time Series Plot of Original Data, (b) Time Series Plot of Stationary Data after the Second Difference for Model 2



(a)



(b)

Figure 8. (a) Time Series Plot of Original Data, (b) Time Series Plot of Stationary Data after the Second Difference for Model 3

3.2.2 Breusch-Pagan Test

The second step for the time series model selection is conducting the Breusch-Pagan test, a method used to check if the variables are homoscedasticity. If the *p*-value of the test is smaller than the 5% significance threshold, the null hypothesis of homoscedasticity would be rejected and heteroskedasticity is present. Table 11 shows the results of Breusch-Pagan test for the three proposed models. The *p*-value of the heteroskedasticity test for Model 1 is 0.6283 (>0.05), which identifies that the null hypothesis is not rejected, and the variables are homoscedasticity. Also, the *p*-values of Model 2 and Model 3 are 0.9934 and 0.934, respectively, meaning that all the models are homoscedasticity.

Table 10. Results of Heteroskedasticity Test (Breusch-Pagan Test) for the Proposed Models

Model	***p*-value**	**Null Hypothesis**	**Results**
Model_1 (7 days a week)	0.6285 > 0.05	homoscedasticity	Not rejected
Model_2 (5 days a week)	0.9934 > 0.05	homoscedasticity	Not rejected
Model_3 (2 days a week)	0.9341 > 0.05	homoscedasticity	Not rejected

3.2.3 ACF & PACF

The ACF (autocorrelation function) tells how the variable is correlated with its lagged value while the PACF (partial autocorrelation function) conveys how the residual of the variable is correlated with its lagged value. Figure 9, Figure 10, and Figure 11 display the ACF and PACF for Model 1, Model 2, and Model 3, respectively. As Figure 9, Figure 10, and Figure 11 show, the lags in both the ACF and PACF are out of the blue dash lines that is 95% confidence interval margin. It proves the correlations of the variables with their lags and indicates how many lags can be used in the model. If all lags are inside the 95% confidence interval, no correlation will be in the variables.

Based on the characteristic of ACF and PACF for three models, the lags of ACF for three models show cut off, while the lags of PACF for three models show decay. Therefore, ARIMA (p, d, q) (Autoregressive Integrated moving average model) was selected to conduct the analysis. And p, d, q, are three important parameters for ARIMA model.

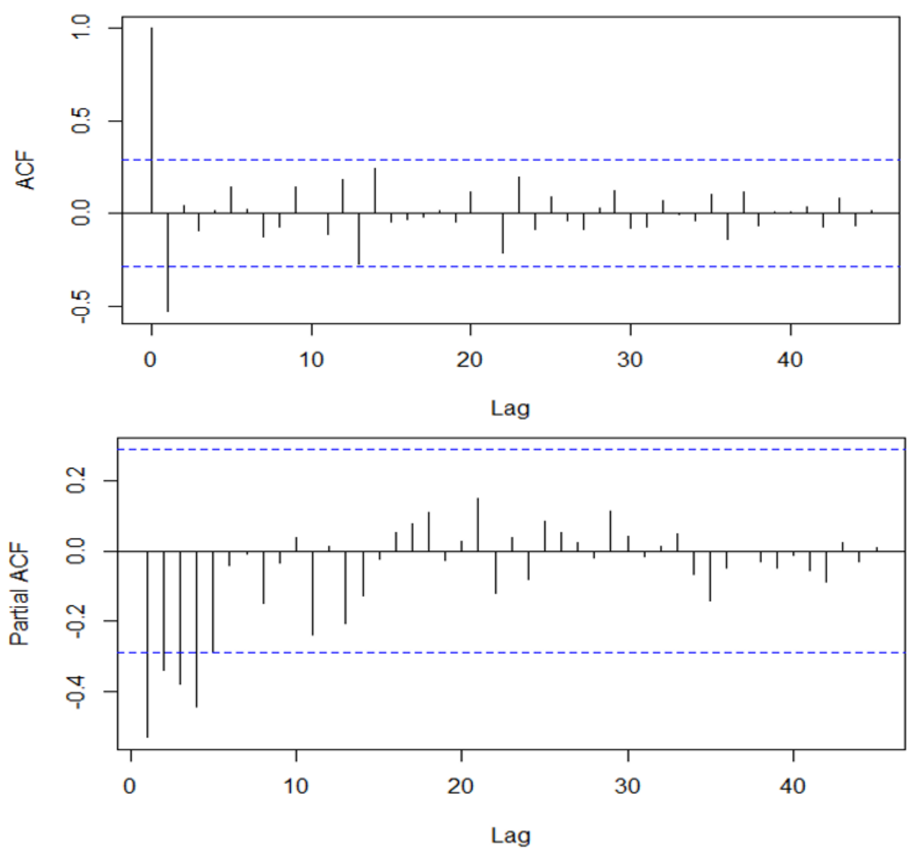


Figure 9. Autocorrelation and Partial Autocorrelation of Average Daily Trip Counts for Model 1

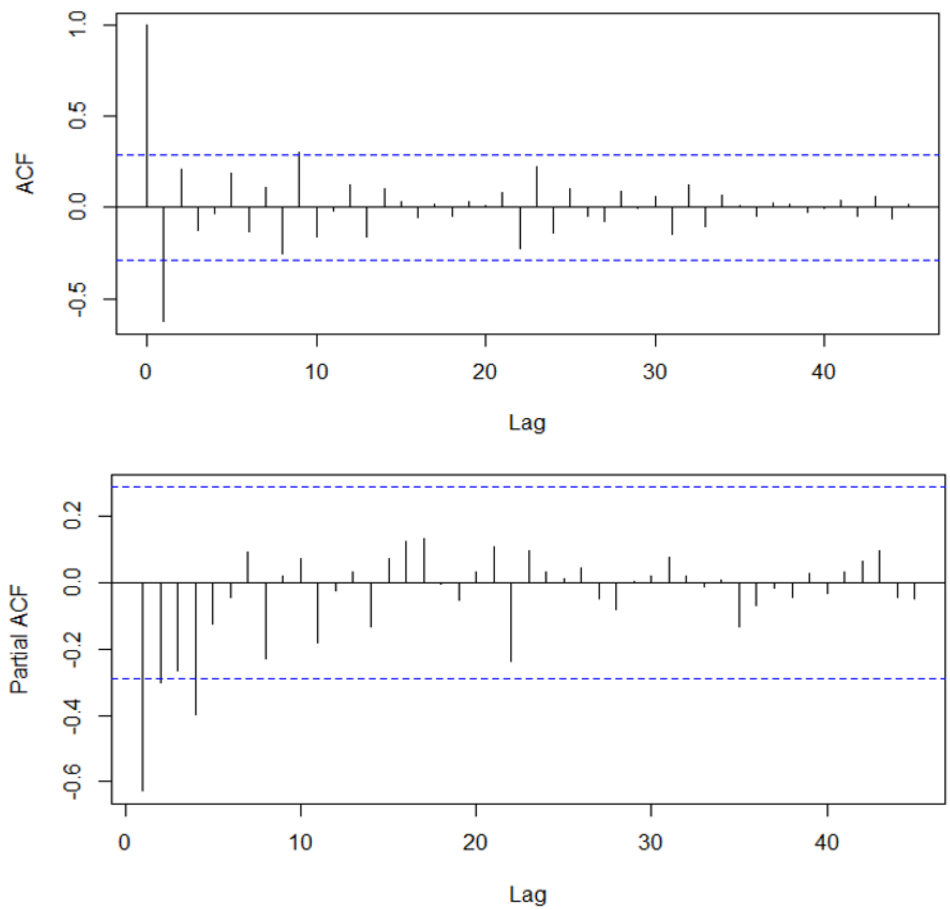


Figure 10. Autocorrelation and Partial Autocorrelation of Average Daily Trip Counts for Model 2

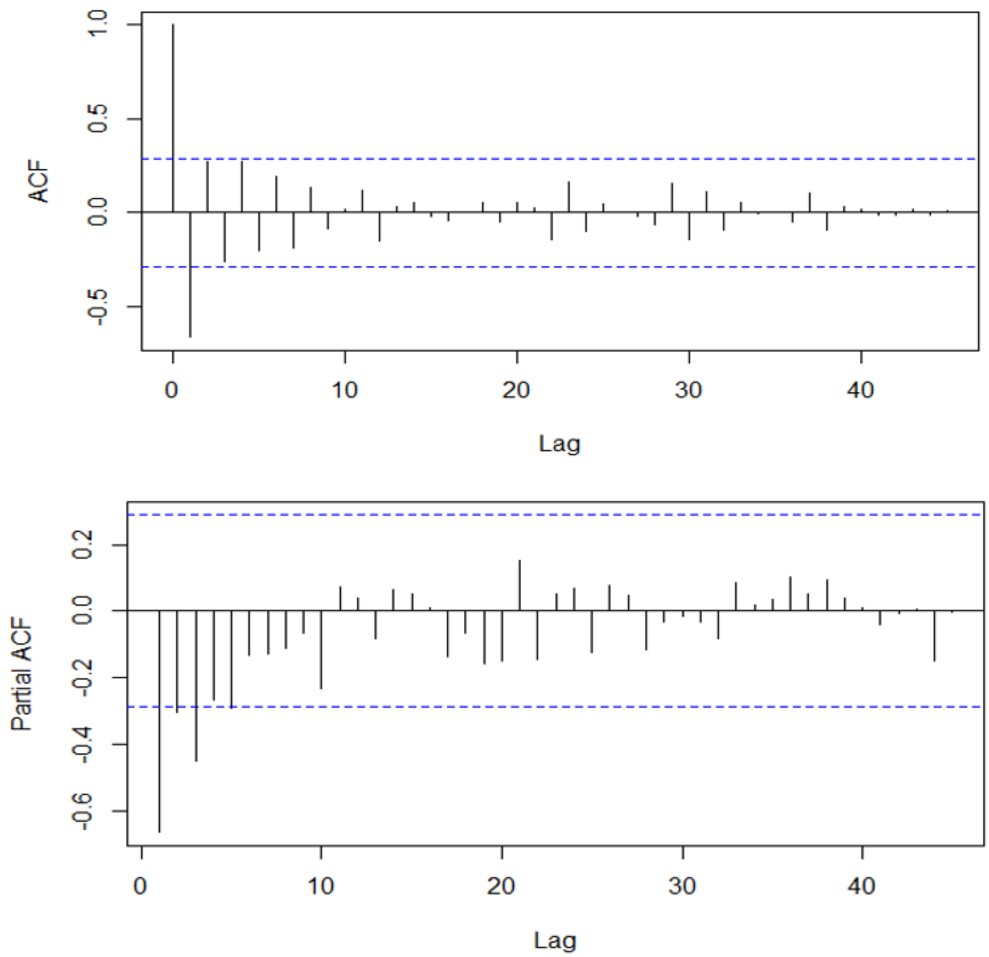


Figure 11. Autocorrelation and Partial Autocorrelation of Average Daily Trip Counts for Model 3

3.2.4 Model Selection

ARIMA (p, d, q) model is a time series model that can be implemented to the non-stationary variables. “AR” in ARIMA refers to “autoregressive” model, which involves the regression on its lags (p). The formula of the AR model can be expressed as equation (*1*). “I” in ARIMA stands for “integrated”, referring to the times of difference (d). “MA” in ARIMA is moving the average model, which indicates that regression errors have a function on previous error lags (q) in the different period. The formula of the MA model has been expressed in equation (*2*).

And the ARIMA model is the combination of the AR model and MA model, the equation is shown as (*3*). The model that was selected is the ARIMA model with exogenous factors; the equation is expressed in equation (*4*).

Autoregressive (AR) Model

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon \tag{1}$$

Moving Average (MA) Model

$$Y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{2}$$

Autoregressive Moving Average (ARIMA) Model

$$Y_t = \alpha + \underbrace{\boxed{\beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p}}}_{\text{AR Model}} + \varepsilon_t + \underbrace{\boxed{\theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}}}_{\text{MA Model}} \tag{3}$$

Autoregressive Moving Average (ARIMA) Model with exogenous factors

$$Y_t = \alpha + \underbrace{\boxed{\beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p}}}_{\text{AR Model}} + \varepsilon_t + \underbrace{\boxed{\theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}}}_{\text{MA Model}} + \varphi_n X_n \tag{4}$$

where:

Y_t= dependent variables, which is daily average ridership counts per week;

X_n = independent variables, including bike lane length, active stations, precipitation, wind speed, and temperature;

ε_t= error term;

α = constant term;

β = coefficient in AR;

θ = coefficient in MA

p = the order of the AR;

q = the order of the MA;

φ_n = coefficient of independent variables.

Normally, p and q can be obtained from ACF and PACF. But In order to achieve the best model that can fit the data, ‘auto.arima’ function was use in the *R* language to explore the proper values of p and q.. The outputs of ‘auto.arima’ function include all possible models along with the Akaike Information Criterion (AIC). AIC is a criterion that identifies how well the model fits the data. The smaller the AIC is, the better the model fit the data. Figure 12, 13 and 14 show the process of the model selection and the best models that fits to the data for three proposed models.

```
> fit_arima <- auto.arima(training_y, xreg = x_metrix_training, d=2, trace = TRUE,
 seasonal= FALSE, stepwise=FALSE, approximation=FALSE)

 ARIMA(0,2,0)                    : 603.6416
 ARIMA(0,2,1)                    : Inf
 ARIMA(0,2,2)                    : Inf
 ARIMA(0,2,3)                    : Inf
 ARIMA(0,2,4)                    : Inf
 ARIMA(0,2,5)                    : Inf
 ARIMA(1,2,0)                    : 581.9314
 ARIMA(1,2,1)                    : Inf
 ARIMA(1,2,2)                    : Inf
 ARIMA(1,2,3)                    : Inf
 ARIMA(1,2,4)                    : Inf
 ARIMA(2,2,0)                    : 575.8938
 ARIMA(2,2,1)                    : Inf
 ARIMA(2,2,2)                    : Inf
 ARIMA(2,2,3)                    : Inf
 ARIMA(3,2,0)                    : 577.8014
 ARIMA(3,2,1)                    : Inf
 ARIMA(3,2,2)                    : Inf
 ARIMA(4,2,0)                    : 574.815
 ARIMA(4,2,1)                    : Inf
 ARIMA(5,2,0)                    : 576.9001



 Best model: Regression with ARIMA(4,2,0) errors
```

Figure 12. Model Selection Result for Model 1

```
> fit_arima <- auto.arima(training_y, xreg = x_metrix_training, d=2, trace = TRUE,
 seasonal= FALSE, stepwise=FALSE, approximation=FALSE)

 ARIMA(0,2,0)                    : 616.8333
 ARIMA(0,2,1)                    : Inf
 ARIMA(0,2,2)                    : Inf
 ARIMA(0,2,3)                    : Inf
 ARIMA(0,2,4)                    : Inf
 ARIMA(0,2,5)                    : Inf
 ARIMA(1,2,0)                    : 582.9786
 ARIMA(1,2,1)                    : Inf
 ARIMA(1,2,2)                    : Inf
 ARIMA(1,2,3)                    : Inf
 ARIMA(1,2,4)                    : Inf
 ARIMA(2,2,0)                    : 573.7857
 ARIMA(2,2,1)                    : Inf
 ARIMA(2,2,2)                    : Inf
 ARIMA(2,2,3)                    : Inf
 ARIMA(3,2,0)                    : 576.6916
 ARIMA(3,2,1)                    : Inf
 ARIMA(3,2,2)                    : Inf
 ARIMA(4,2,0)                    : 576.4183
 ARIMA(4,2,1)                    : Inf
 ARIMA(5,2,0)                    : 575.4362



 Best model: Regression with ARIMA(2,2,0) errors
```

Figure 13. Model Selection Result for Model 2

```
> fit_arima <- auto.arima(training_y, xreg = x_metrix_training, d=2, trace = TRUE,
 seasonal= FALSE, stepwise=FALSE, approximation=FALSE)

 ARIMA(0,2,0)                    : 652.352
 ARIMA(0,2,1)                    : Inf
 ARIMA(0,2,2)                    : Inf
 ARIMA(0,2,3)                    : Inf
 ARIMA(0,2,4)                    : Inf
 ARIMA(0,2,5)                    : Inf
 ARIMA(1,2,0)                    : 630.171
 ARIMA(1,2,1)                    : Inf
 ARIMA(1,2,2)                    : Inf
 ARIMA(1,2,3)                    : Inf
 ARIMA(1,2,4)                    : Inf
 ARIMA(2,2,0)                    : 626.1956
 ARIMA(2,2,1)                    : Inf
 ARIMA(2,2,2)                    : Inf
 ARIMA(2,2,3)                    : Inf
 ARIMA(3,2,0)                    : 617.736
 ARIMA(3,2,1)                    : Inf
 ARIMA(3,2,2)                    : Inf
 ARIMA(4,2,0)                    : 618.0983
 ARIMA(4,2,1)                    : Inf
 ARIMA(5,2,0)                    : 617.7638



 Best model: Regression with ARIMA(3,2,0) errors
```

Figure 14. Model Selection Result for Model 3

The results of model selection for three proposed models are summarized in Table 11. The best model for Model 1 is ARIMA (4,2,0) with the smallest AIC among 21 possible ARIMA models. ARIMA (2,2,0) and ARIMA (3,2,0) are the best models among 21 possible models for Model 2 (5 days of a weekday) and Model 3 (2 days of a weekend), respectively.

Table 11. Model Selection Results of the Three Models

'auto.arima' Function in R	**Possible Model Counts**	**Best Model**	**AIC**
Model 1	21	ARIMA (4,2,0)	574.815
Model 2	21	ARIMA (2,2,0)	573.786
Model 2	21	ARIMA (3,2,0)	617.736

3.2.5 Model Validation

In order to evaluate the accuracy of the model, the Ljung-Box test was applied to check if the fitted model is white noise. White noise is a kind of time series that the mean error is close to zero and has no correlation among the variables. The outputs of the residual test for three selected models are displayed in Figure 15, 16, and 17.

Figure 15 (a), Figure 16 (a), and Figure 17 (a) shows the residual plot from week 1 to 48 of Model 1, Model 2, and Model 3, respectively. Figure 15 (b) Figure 16 (b), and Figure 17 (b) present residuals' autoregressive function, and the range between the blue dot line is 95% confidence interval. Figure 15 (c) Figure 16 (c), and Figure 17 (c) display the histogram of the residuals, and the red line is the boundary of the 95% confidence interval.

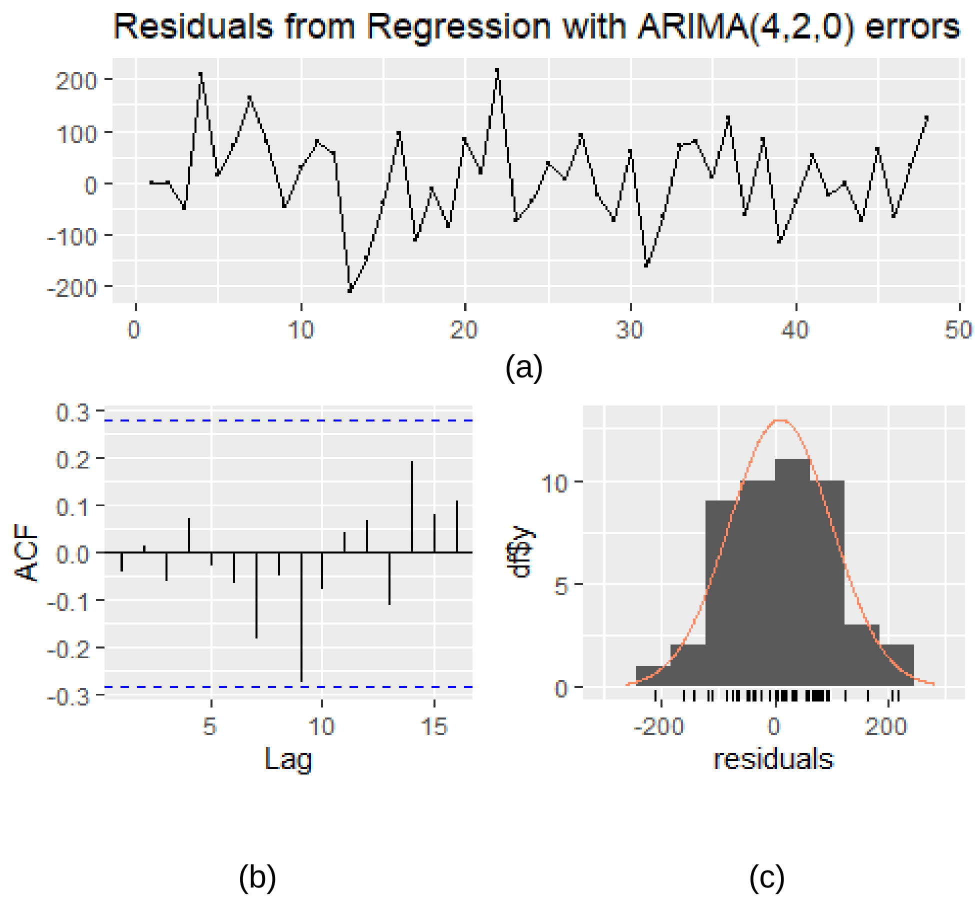


Figure 15. Ljung-Box Test Results of Estimated Model for Model 1: (a) Residuals Plot from Week 1 to 48, (b) Autocorrelation Function of Residuals, and (c) Histogram of the Residuals

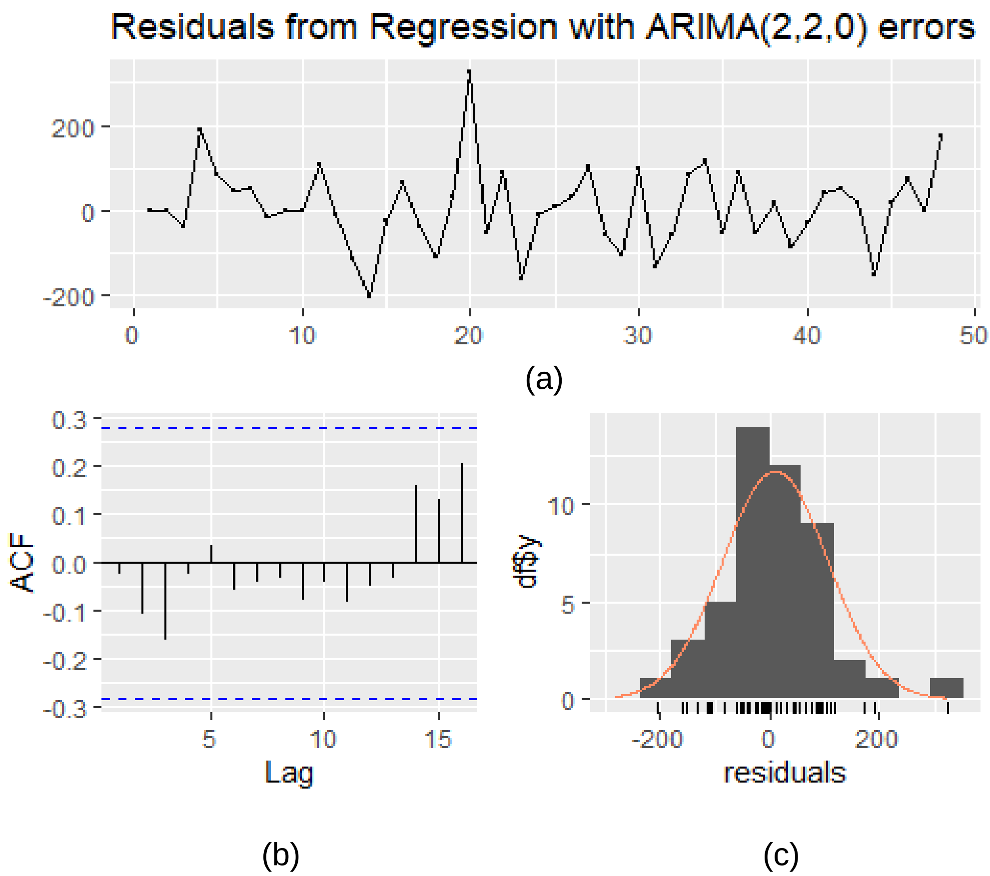


Figure 16. Ljung-Box Test Results of Estimated Model for Model 2: (a) Residuals Plot from Week 1 to 48, (b) Autocorrelation Function of Residuals, and (c) Histogram of the Residuals

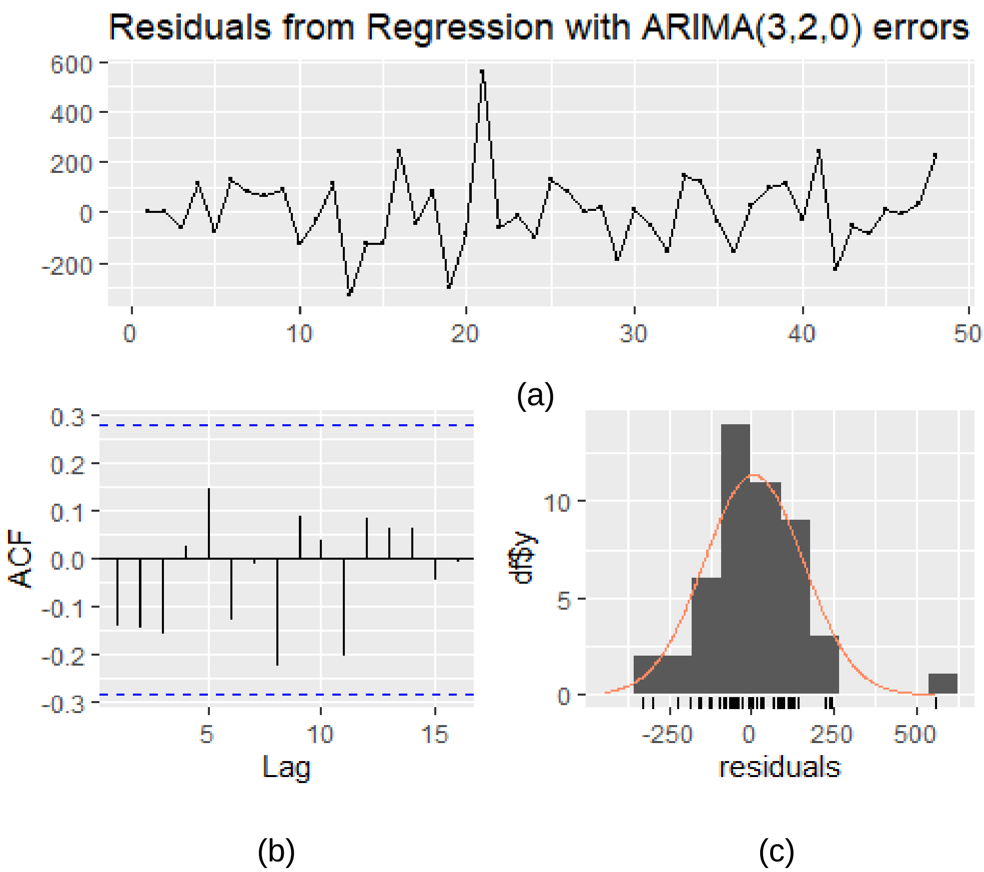


Figure 17. Ljung-Box Test Results of Estimated Model for Model 3: (a) Residuals Plot from Week 1 to 48, (b) Autocorrelation Function of Residuals, and (c) Histogram of the Residuals

Table 12 summarizes the results of the outputs for three models. The mean errors of Model 1, Model 2, and Model 3 are 9.47, 9.31, and 6. 19, respectively. The mean errors of the models are quite small and can be able to a certain extent. According to Figures 15 (b), Figure16 (b), and Figure 17 (b), there are no residuals out of the dotted line, which means all residuals are within the 95% confidence interval. It also indicates that no correlations exist among variables. The histograms shown in Figures 15 (c), 16 (c), and 17 (c) convey that most residuals are under the 95% confidence interval. All evidence proves that three estimated models are white noise, which means that the models represent a good fit for the data and can be used to forecast the future.

Table 12. Results of Ljung-Box Test for Three Estimated ARIMA Models

Ljung-Box Test	ME	ACF	P-value	Result
Model 1 ARIMA (4,2,0)	9.47	residuals within 95% confidence interval	< 0.05	White Noise
Model 2 ARIMA (2,2,0)	9.31	residuals within 95% confidence interval	< 0.05	White Noise
Model 3 ARIMA (3,2,0)	6.19	residuals within 95% confidence interval	< 0.05	White Noise

CHAPTER 4

RESULTS AND DISCUSSION

Modeling and parameter estimation should be carried out in the following step after the model selection is completed for the three proposed models. In this chapter, the model parameters are estimated for each of the three models. Furthermore, the forecast ridership for the last four weeks of the years (weeks 49 to 52) will be calculated from the three models and will be compared to the corresponding actual ridership. At the end of the chapter, methods and criteria will be proposed for assessing the accuracy of forecasts.

4.1 Estimation Results

Training dataset (week 1 to 48) were used to estimate the parameters of the three models. Table 13 shows the outputs of the model estimation process. The dependent variable is average daily trip counts, and the independent variables are bike lane length, active stations, temperature, wind speed, and precipitation. ARIMA (4,2,0) is designed for Model 1 with seven consecutive days per week; ARIMA (2,2,0) and ARIMA (3,2,0) are developed for Model 2 with five consecutive weekdays per week and Model 3 with two consecutive days of weekends per week, respectively. The coefficient and p-value of each variable are listed in the table. The threshold of significance for the independent variables is 5%. The star (*) marks for the coefficient of variables mean they are significant and would impact on dependent variable, average daily ridership. Single star mark (*) indicates a low significance level, and double star mark (**) and triple star mark (***) convey a moderate significance level and a high significance level, respectively.

Table 13. ARIMA Model Outputs Based on the Training Data from Week 1 to 48

Sample: 1-48; Number of observations=48						
Model	**7 days/week (Model 1)**		**5 days/week (Model 2)**		**2 days/week (Model 3)**	
ARIMA	ARIMA (4,2,0)		ARIMA (2,2,0)		ARIMA (3,2,0)	
Variables	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**
Average Daily Ridership						
Bike Lane Length	38.61**	0.010	37.84**	0.004	82.72***	0.000
Active Station	16.28***	0.000	17.28***	0.000	34.15***	0.000
Temperature	2.54	0.339	3.72	0.126	4.11	0.158
Wind Speed	7.40	0.335	3.85	0.514	-3.87	0.554
Precipitation	-114.07**	0.009	-49.83*	0.054	-421.79***	0.002
AR						
ar1	-1.157***	0.000	-1.165***	0.000	-1.248***	0.000
ar2	-1.034***	0.000	-0.548***	0.000	-0.987***	0.000
ar3	-0.701**	0.002	-	-	-0.495***	0.000
ar4	-0.414**	0.008	-	-	-	-
MA	N/A					
Log Likelihood	- 274.26		-276.95		- 297.37	

p* < 0.05 (low significance level)
p** < 0.01 (moderate significance level)
p*** < 0.001 (high significance level)

Table 14 shows the summary of the ARIMA model outputs. Built environment factors, referring to the bike lane length and the active stations, show a high significance level to average daily ridership with very low p-values for three models. And the impact of those two variables is positive. Every mile bike lane being added to the bike infrastructure will generate 38 more trips per day for Model

1. For Model 2 and 3, for each mile of bike lane is built, 37 and 82 more trips are generated, respectively. By adding one bikeshare station to the system, the average daily ridership will increase by 16 more trips every week, 17 more trips every weekday, and 34 more trips on the weekends. Those results show that the built environment factors have high impacts on the ridership. The impact caused by the built environment on the average daily ridership of the weekends is double than the impacts on the weekday ridership and weekly average ridership. Regarding the weather variables, temperature and wind speed don't present significant impacts on the average daily trip counts with relatively high p-values for the three models; however, precipitation displays a negatively significant result on the average daily ridership. Every inch of precipitation would result in 114 ridership decrease per week, 49 ridership reduction per weekday, and 421 ridership decrease on the weekends. Therefore, precipitation would lead to a drastic influence on the weekend ridership.

Table 14. ARIMA Model Outputs Summary

Variables	**Model**		
	Model 1 (7 days of the weeks)	Model 2 (weekdays)	Model 3 (2 days of the weekends)
Built Environment			
Bike Lane Length	38 ↑	37 ↑	**82** ↑
Active Station	16 ↑	17 ↑	**34** ↑
Weather Factors			
Precipitation	114 ↓	50 ↓	**422** ↓

The entire AR lag terms of the ARIMA model show a highly significant level to the average daily ridership, but the MA lag term is not suitable to be applied to the model. Log-likelihood implies how well the predicted model fits the data whose function is similar to AIC; the smaller the absolute value of log-likelihood is, the better the model would fit the data.

4.2 Forecast

The training data from weeks 1 to 48 was utilized to develop the model, while the testing data from weeks 49 to 52 were used to check the accuracy of the estimated variables. In this section, considering the ARIMA models presented in the last section, forecasting the ridership for the following four weeks (week 49 to week 52 which are the last weeks of the year) will be studied. Figure 18 (a), (b), and (c) displays the forecasted ridership in the next four weeks based on the testing data from week 1 to week 48 for Model 1, Model 2, and Model 3, respectively. The dark shade shown in the plot is the range of 80% confidence interval, and the light shade represents 95% confidence interval. The accuracy of the forecast can be told by that the ridership for the three models are in the range of 80% confidence interval.

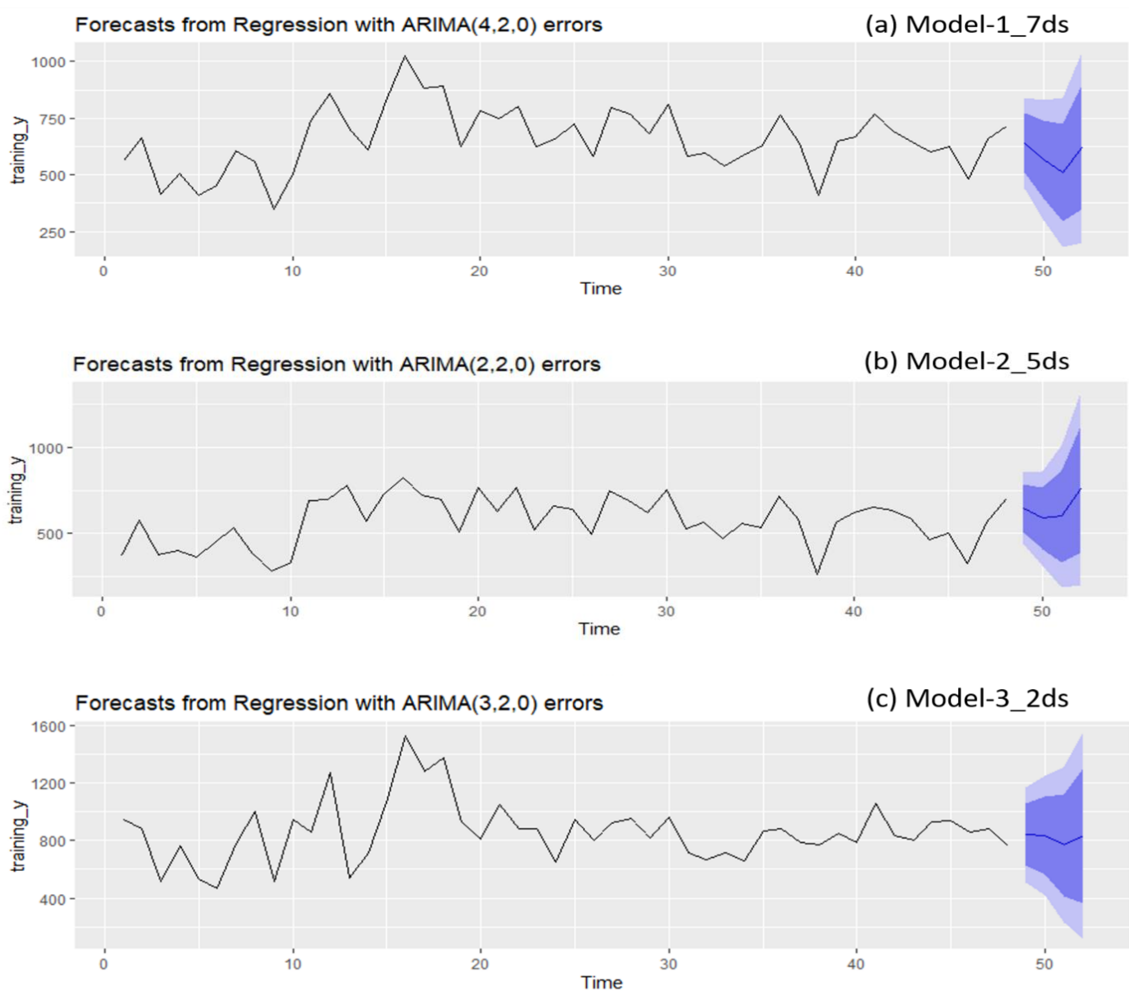
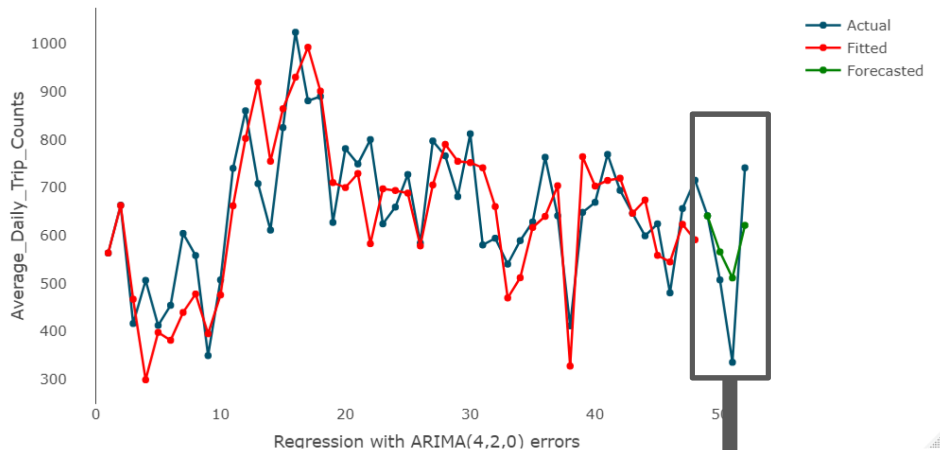
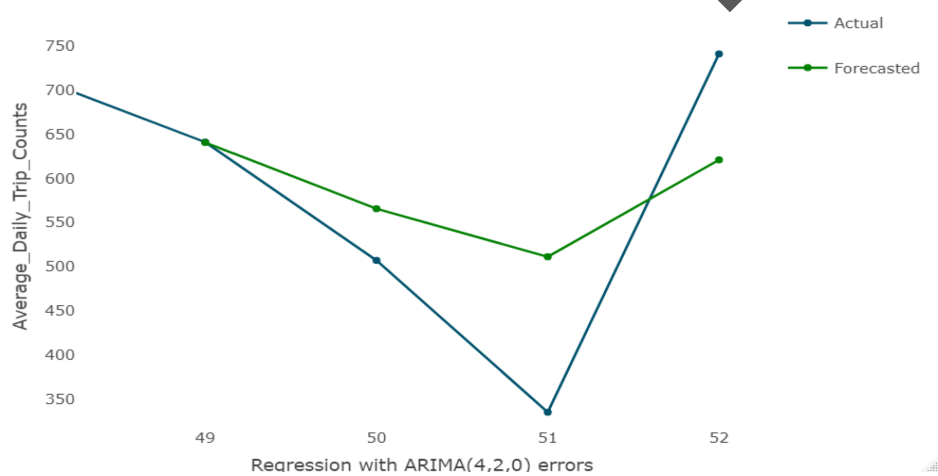


Figure 18. Forecast for the Three Fitted Models: (a) ARIMA (4,4,0) for Model 1, (b) ARIMA (2,2,0) for Model 2, and (c) ARIMA (3,2,0) for Model 3

Figure 19, Figure 20, and Figure 21 show the actual daily trip counts from week 1 to 52 (in blue) versus in-sample forecast average daily trip counts from week 1 to 48 (in red) and out-of-sample forecast average daily ridership from week 49 to 52 (in green) for Model 1, Model 2, and Model 3, respectively. Figure 19 (a), Figure 20 (a), and Figure 21 (a) mainly display the difference between actual average daily ridership per week and in-sample forecast daily ridership. Figure 19 (b), Figure 20 (b), and Figure 21 (b) show the actual and forecast ridership, corresponding to the last four weeks of the year, for three models and in a larger scale (zoomed view). It is not difficult to see that the model provides a good fit for the actual ridership.

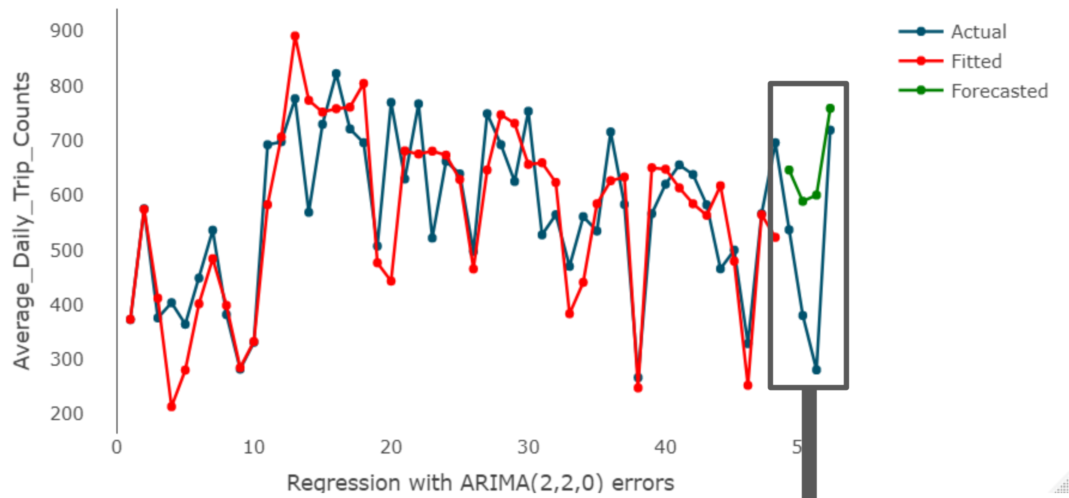


(a)

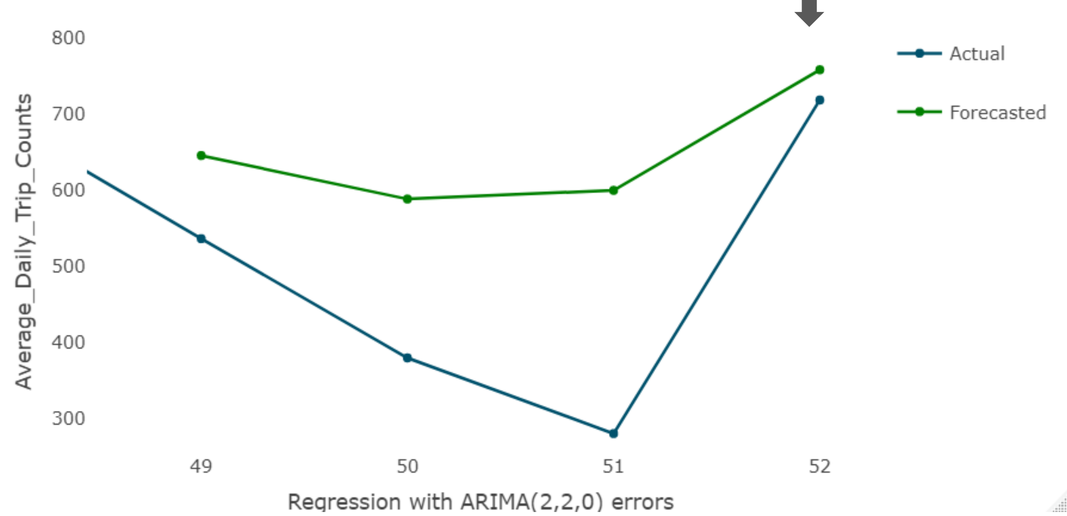


(b)

Figure 19. Daily Trip Counts from Week 1 to 52. (a) Actual Daily Trip Counts from Week 1 to 52 (Blue) versus In-Sample Forecast Average Daily Trip Counts from Week 1 to 48 (Red), (b) Out-of-Sample Forecast Average Daily Trip Counts from week 49 to 52 (Green) for Model 1

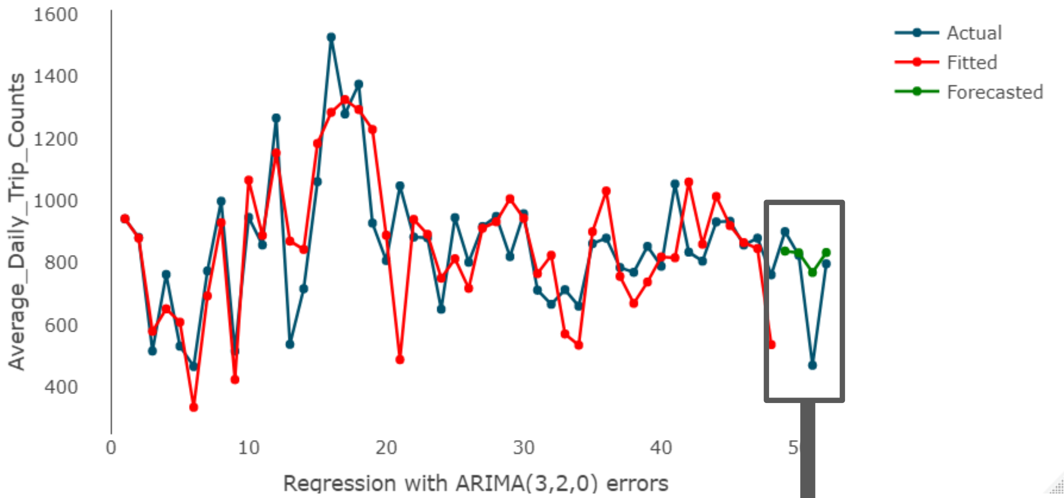


(a)

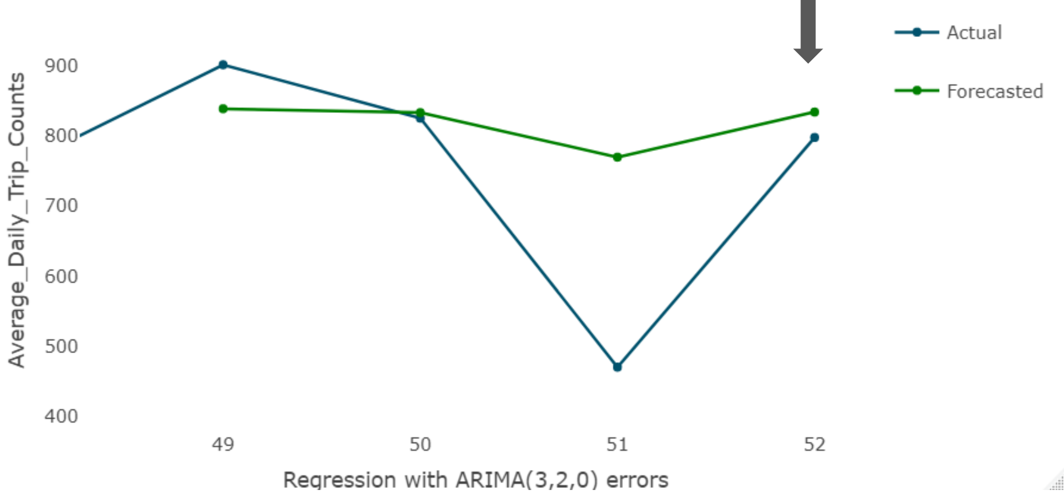


(b)

Figure 20. Daily Trip Counts from Week 1 to 52. (a) Actual Daily Trip Counts from Week 1 to 52 (Blue) versus In-Sample Forecast Average Daily Trip Counts from Week 1 to 48 (Red), (b) Out-of-Sample Forecast Average Daily Trip Counts from week 49 to 52 (Green) for Model 1



(a)



(b)

Figure 21. Daily Trip Counts from Week 1 to 52. (a) Actual Daily Trip Counts from Week 1 to 52 (Blue) versus In-Sample Forecast Average Daily Trip Counts from Week 1 to 48 (Red), (b) Out-of-Sample Forecast Average Daily Trip Counts from week 49 to 52 (Green) for Model 1

4.3 Accuracy Test

There are many methods and criteria for assessing the accuracy of forecasts. For this study, Root Mean Square Error (RMSE) was used to evaluate the accuracy of the out-of-sample forecast. RMSE for training and testing set in three models show as Table 15, which measures the difference between observed value (sample value) and predicted value, and lower value indicates less residuals or errors occurs in the model. Equation (5) express the formula of RMSE, $\hat{y}_t$ represent the predicted value and y_t is sample value over the T time periods (1, 2, …, t).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}} \tag{5}$$

Table 15. RMSE for Training and Testing Sets in Three Models

Accuracy Test		RMSE
Model 1 ARIMA (4,2,0,)	Training set (Fitted Model, week 1 to 48)	90.05
	Test set (Forecast, week 49 to 52)	110.54
Model 2 ARIMA (2,2,0,)	Training set (Fitted Model, week 1 to 48)	95.92
	Test set (Forecast, week 49 to 52)	199.48
Model 3 ARIMA (3,2,0,)	Training set (Fitted Model, week 1 to 48)	148.80
	Test set (Forecast, week 49 to 52)	153.85

To better understand and evaluate the forecast accuracy of the models, Normalized Root Mean Square Error (NRMSE) was generated. NRMSE explains the errors of the models and forecasts in percentage; and it is calculated by dividing

RMSE by the difference between the maximum sample value and minimum sample value or the average sample value:

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} = \frac{RMSE}{\bar{y}} \quad (6)$$

A model with a small NRMSE (for example 10%) is considered a good fit, which means only 10% samples don't fit the model well. Table 16 shows the NRMSE for training and testing sets in three Models. For the training set, the three models have a relatively small NRMSE (13%, 17%, and 14%), which again proves that the estimated models fit the data perfectly. For the test set, three models have moderate NRMSE (20%, 31%, 21%), indicating a moderate fit. Model 2, i.e. ARIMA model (2,2,0), implies a relatively higher error. However, the model will be still helpful for predicting and making plans for future projects. Using more and precise data for the model development possibly can reduce the error and improve the model.

Table 16. NRMSE for Training and Testing Sets in Three Models

NRMSE	**Training set (Fitted Model, week 1 to 48)**	**Test set (Forecast, week 49 to 52)**
Model 1 ARIMA (4,2,0,)	13%	20%
Model 2 ARIMA (2,2,0,)	17%	31%
Model 3 ARIMA (3,2,0,)	14%	21%

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

Previous studies have been mostly on the impact of bike infrastructures on ridership at station level. They measured the length of bicycle routes in a buffer around each bike station, which means the bike lane impacts only on the riders from/to that station. Therefore, they have not considered the bike infrastructure impact at the system level in a city such as Houston. In this study, longitudinal analyses were conducted to identify the effects of bike infrastructures, including added bike lanes, on bicycle sharing system demand. Three ARIMA models were generated through several statistical analyses: ARIMA (4,2,0) for Model 1 (7 days of the week), ARIMA (2,2,0) for Model 2 (5 days of the week), and ARIMA (3,2,0) for Model 3 (2 days of the weekend). The model results showed that each mile bike lane added to the bikeshare system would result in 38 average daily ridership increase per week (7 days), 37 average daily ridership increase per weekday (5 days), and 82 average daily ridership increase per weekend (2 days). Furthermore, by adding every station to the public bikeshare system, the average daily ridership will increase by 16. The increase of average daily ridership will be 17 for weekdays, and 34 on weekends. The impact of built environment on average daily ridership of weekends are almost double when compared to its impact on the average weekday ridership, resulting in more demand during weekends. Regarding the weather variables, temperature and wind speed have no major impact on average daily trip counts. However, precipitation displayed negatively significant impacts on average daily ridership. Every inch of precipitation would cause 114 ridership decrease per week, 49 ridership reduce per weekday, and 421 ridership decrease on weekend. The numbers indicate that precipitation would lead to a drastic influence on the weekend ridership.

City planning agencies are interested to know the future cycling patterns before carrying out any bike lane expansion plan; and bikeshare operators would like to forecast the system demand as the new bike infrastructures are planned. It is very important to identify how much bikeshare ridership in a system will increase as a result of installing extra bike lanes. The proposed models in this study are able to capture system-wide bike ridership and can be used to measure the marginal cost of building bike lanes or bike paths on bike share demand at a network-wide level over the time.

Ridership can be impacted in many ways, including population density, the distribution of bike station, the distance to the bike station, job density, and bike infrastructure, etc. In my study, bike lane length, active station and weather factors were considered in the model, and the result is acceptable. However, there may are some factors impact our accuracy like less data, incomplete data. More data might be collected and implemented in the future research to optimize the model. Furthermore, more variables such as population density, distance to bikeshare stations, user type, etc. can be considered in the future studies to investigate the impact of the built environment on the ridership

REFERENCES

Daddio, D.W. (2012). Maximizing bicycle sharing: an empirical analysis of capital bikeshare usage. University of North Carolina at Chapel Hill. DOI: 10.17615/qv32-b860.

DeMaio, P. (2003). Smart bikes: Public transportation for the 21st century. Transportation Quarterly. Volume 57, No. 1, pp. 9–11.

DeMaio, P. (2004). Will smart bikes succeed as public transportation in the United States? Journal of Public Transportation. Volume 7, No. 2, pp.1-15.

Ding, W. (2016). The Relationship of Built Environment and Weather with Bike Share – Evidence from the Pronto Bike Share System in Seattle. University of Washington.

Faghih-Imani, A., and Eluru, N. (2016). Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system. Journal of Transport Geography. Volume 54, pp. 218-227. DOI: 10.1016/j.jtrangeo.2016.06.008.

Gallop, C.W., Tse, C.H., & Zhao, J. (2011). A Seasonal Autoregressive Model of Vancouver Bicycle Traffic Using Weather Variables.

Houston Bike Plan. Huitt-Zollars. https://www.huittzollars.com/expertise/planning/houstonbike-plan

Houston Bcycle (2020). Wikipedia. https://en.wikipedia.org/wiki/Houston_Bcycle

Houston Bcycle Program. City of Houston Transportation Planning & Development. https://www.houstontx.gov/planning/transportation/BCycle.html#:~:text=Houston%20BCyce%20initially%20began%20as,to%20encourage%20biking%20in%20Houston

Larsen, J., (2013) Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide. Earth Policy Institute.

DeMaio, P. and MetroBike, LLC (2009). Bike-sharing: History, Impacts, Models of Provision, and Future. Journal of Public Transportation. Volume 12, No. 4, pp. 41-56. DOI: http://doi.org/10.5038/2375-0901.12.4.3.

Tran, T.D., Ovtracht, N., and Faivre d'Arcier, B. (2015). Modeling bike sharing system using built environment factors. ScienceDirect. Volume 30, pp. 293-298. DOI: 10.1016/j.procir.2015.02.156.

Xu, S.J. and Chow, J.Y.J (2019). A longitudinal study of bike infrastructure impact on bikesharing system performance in New York City. International Journal of Sustainable Transportation. DOI: 10.1080/15568318.2019.1645921.

Ying Zhang, Y, Thomas, T., Brussel, M., and van Maarseveen, M (2017). Exploring the impact of built environment factors on the use of public bikes at bike station: case study in Zhongshan, China. Journal of Transport Geography. Volume 58, pp. 59-70. DOI: 10.1016/j.jtrangeo.2016.11.014.