Analyzing and Forecasting Usage Trends of Capital BikeShare

1. Introduction

Capital BikeShare is a bicycle-sharing system company mainly operating in Washington D.C. and some areas of Maryland and Virginia. The company started off in 2010 with just 400 bicycles at 49 different locations; since then, they have expanded to more than 4,300 bikes available at over 500 stations. Users can either pay for an annual membership of $7/month (with the first 30 minutes of every ride being free), $2/30 minutes for non-members, or even $8/day for people visiting D.C.

Capital BikeShare has been on a steady incline in usage ever since it was launched. However, like almost all businesses, Capital BikeShare's business was significantly impacted in 2020 when the world was devastated with a global pandemic. This project aims to **analyze the impact of COVID-19 on Capital BikeShare's usage trends and forecast ridership in 2021**.

2. Audience

The main audience of this project will mainly be the **key decision makers** and **operation managers** of Capital BikeShare. The key business decision makers will need to decide how they will accommodate for the forecasted usage trends for 2021, while the operation managers will need to ensure that bikes and stations are ready to meet the demands of the customers according to the forecasts.

3. Data

The data for this project was obtained from the publicly available files on the Capital BikeShare website. The files contain every trip taken since 2010, with monthly reports starting in 2018. These datasets include data on trip ID, start/end station, start/end latitude and longitude, duration of trip (not in recent datasets), ride type (regular vs. electric bike), and membership type.

One of the main challenges of working with this data was having to find efficient ways of handling the massive amount of data. Because these files contained all trips ever taken, the total amount of instances came out to over 26 million rows. With so much data and limited amount of memory for use, it was imperative to find different ways of handling the data during wrangling and pre-processing.

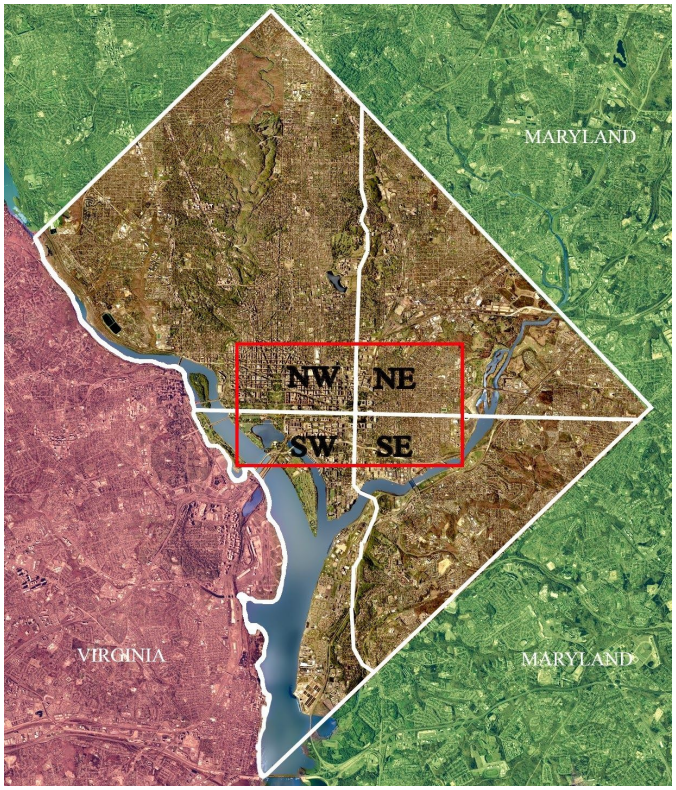
When all the data were compiled together, there were a total of around 85,000 entries with null values. Most of the null values came from missing station names and IDs. I assumed that these entries had latitude and longitude values that were used in other entries, however this was not the case. Instead of removing 85,000 entries for null values, I figured I could use the latitude and longitude coordinates listed on these rows and extract street names or any other significant location identifier as their station names. To accomplish this, I used GeoPy and their Nominatim geocoder. GeoPy is a Python package that provides easy access to

several geocoding web services. I chose the Nominatim geocoder as it was the only one that was open source and free.

In the more recent data files, the trip duration of each trip was not reported. This was an easy fix; calculate the duration by subtracting the trip start time from the end time. In the older files, there were some entries with missing latitude and longitude coordinates. Thankfully, the station names associated with these entries existed in the recent files and so I was able to obtain the coordinates from there.

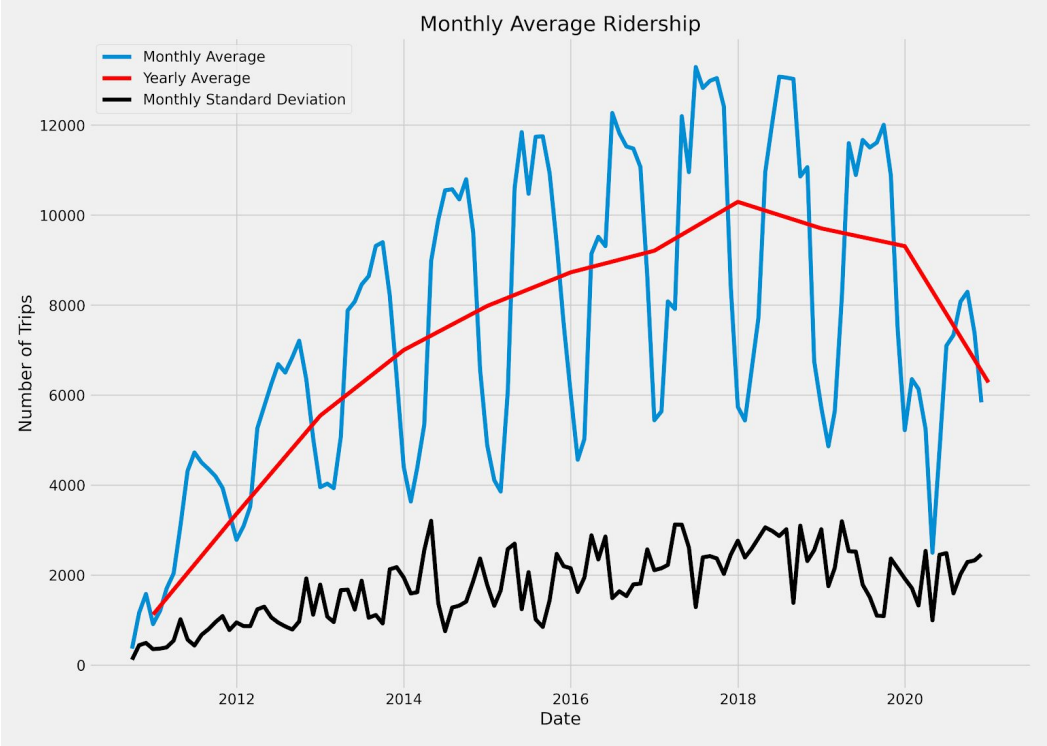
For my main analysis, I wanted to analyze the overall usage trend of Capital BikeShare and forecast its usage for 2021. To do this, I needed the daily count of trips. Instead of loading up the huge combined dataset and storing in memory every time I ran the analysis, I created a separate file containing the amount of trips taken per day. This significantly reduced the operation times during analyses.

In the second part of the modeling section, I wanted to analyze the differences of usage trends between central and outer D.C. I arbitrarily defined central D.C. as the area within the red rectangle in the image below. This grid was bounded by (38.88, 38.92) latitudinally and (-77.05, -76.97) longitudinally. Every station outside of the red rectangle was considered as outer D.C. Like before, I created separate files for the daily count of trips taken in both central and outer D.C.



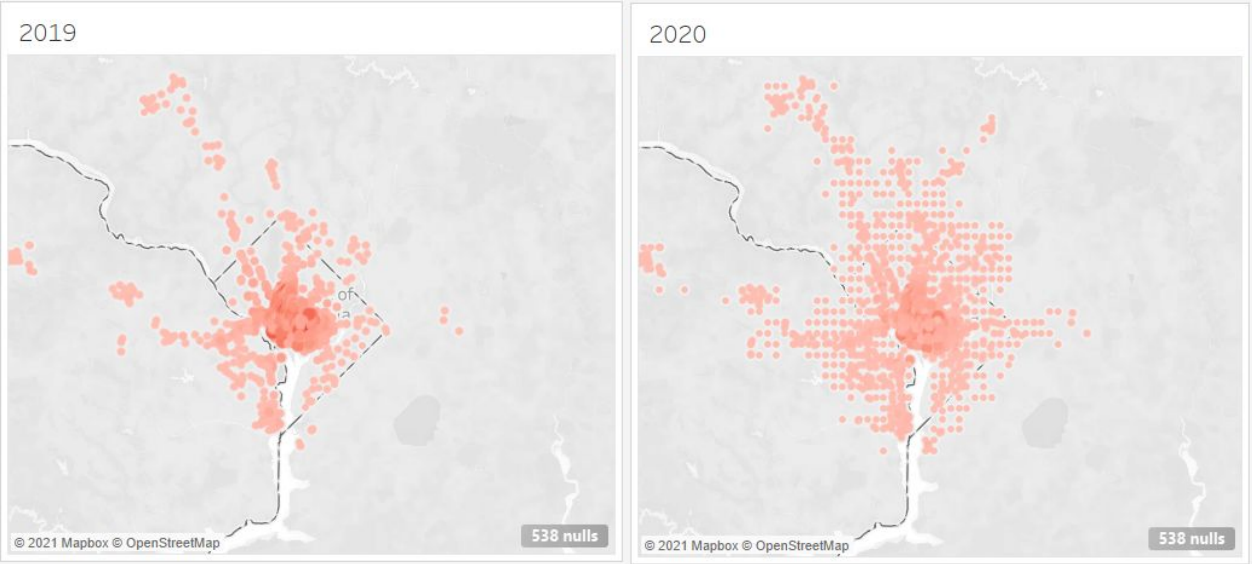
4. Exploratory Data Analysis

The main portion of EDA was done on Tableau for this project. The first things I wanted to examine with this data was what the overall trend looked like year-to-year, and whether there was an element of seasonality every year. The figure below shows that Capital BikeShare saw an almost linear increase every year since its inception until 2018, and between 2018 and 2020, the company saw a slight decline in its usage. From a cursory investigation on why this occurred, I was unable to pinpoint a specific reason however, it could possibly be due to Capital BikeShare expanding its territory to certain areas of Virginia and Maryland. This expansion could cause a decrease in the budget attributed to maintaining and adding stations in the central D.C. area, the area with most usage every year. Furthermore, there is an obvious element of seasonality in the time series. Daily trip count starts off low every year, with peaks occurring in the middle of the year, and once again falling back down near the end of the year. This makes sense as it gets a little cold for bike rides in Washington D.C. during the start and end of each year, with average temperatures going down to as low as 25°F in January. Capital BikeShare experiences peak usage every year during the summer times as many people will look for an eco-friendly form of transportation for work and leisure.



Expectedly, Capital BikeShare experienced its worst decline in usage in 2020. Due to the global pandemic forcing many workplaces to mandate a Work-from-Home policy starting in mid-March/April, the usual surge of usage in spring 2020 did not happen; instead the average usage fell to its lowest since 2011. During the summer of 2020, the usage of the bikes

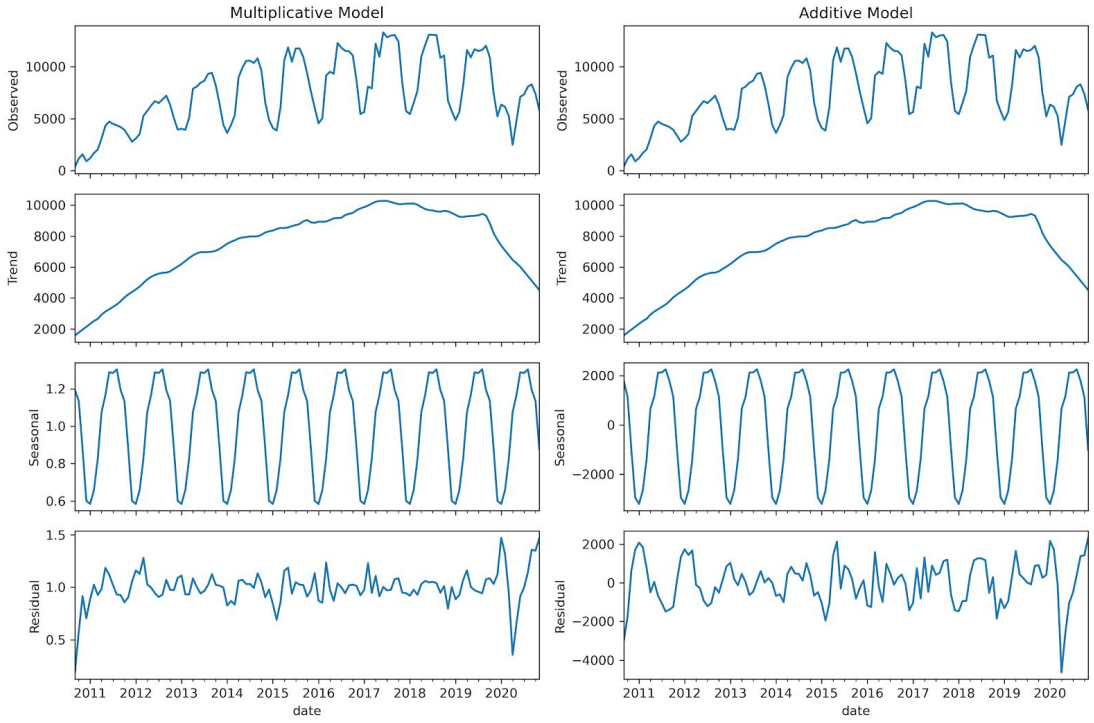
increased, but not back its normal levels. However, from the two maps shown below, the usage didn't increase in just central D.C. Instead, there was a relatively significant increase in usage in outer D.C. which could be attributed to more people using these bikes for leisure and relieving quarantine stress with fresh air instead of using these as a mode of transportation between buildings in downtown D.C. For the two maps below, size is directly proportional to the amount of trips taken at the station and the darker colors also indicate higher usage.



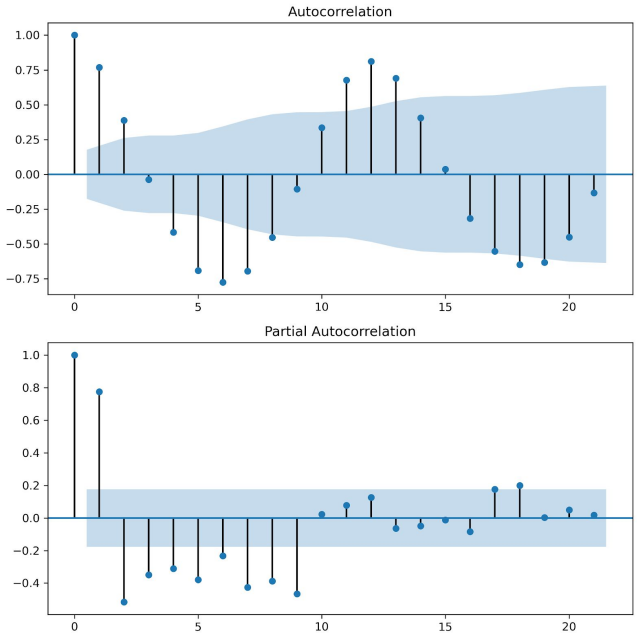
5. Modeling

To forecast Capital BikeShare's usage for 2021, I opted to use an Autoregressive Integrated Moving Average (ARIMA) model, specifically the Seasonal ARIMA model. Simply put, an ARIMA model is one that utilizes its lags (past values) and lagged forecast errors to model and forecast a time series. The first step in utilizing an ARIMA model is decomposing the time series to better understand the different components of a time series. From the graphs below, it can be seen that the time series can be interpreted as a multiplicative model composed of a clear trend and seasonal components. A multiplicative model is a better representation of the time series than an additive model as there is a slight element of seasonality visible in the residuals of the additive model.

The next step of ARIMA modeling is to make the time series stationary to gauge what order of differencing will be required for the model. An ARIMA model can be characterized by 3 different terms: *p, d,* and *q*. The *d* term is determined by the order of differencing required to make the time series stationary. For this time series, taking the first difference stastionarized the time series rather effectively. With the Dickey-Fuller test, the first differenced time series achieved a *p*-value of 0.001695, indicating that the data has become stationary. We can expect the ARIMA model's *d*-value to be 1 due to only requiring one order of differencing.



The next step to ARIMA modeling is to plot the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) plots; doing so will help get an idea of what *p* and *q* values the model will use. The ACF represents the total correlation between different lag functions while the PACF is the correlation between two lags irrespective of other lags. The gradual decrease in the ACF plot indicates that $q = 1$ would be appropriate and the sudden drop after 2 lags in the PACF plot indicates a *p* value of 2 would be appropriate.



While performing the tests and analyses above give us a general understanding of what sort of model to use, *pmdarima* is a Python package that implements R's *auto.arima*. By defining minimum and maximum orders to test for *p, d,* and *q* values, *pmdarima* automatically discovers the optimal orders for an ARIMA model. The table below shows the final orders of the resulting model and its performance metrics.

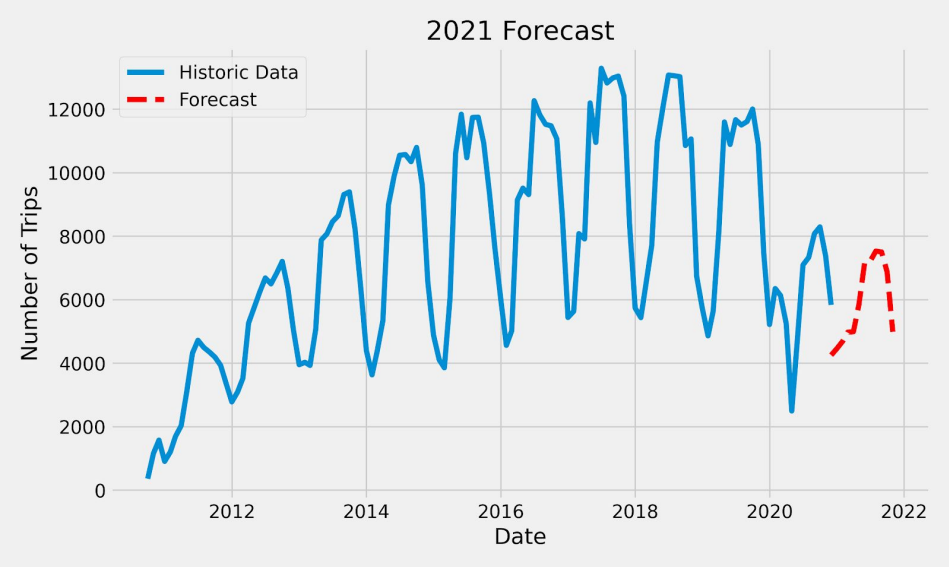
SARIMAX Results

Dep. Variable:	y	**No. Observations:**	123
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 12)	**Log Likelihood**	-1038.172
Date:	Tue, 09 Feb 2021	**AIC**	2086.343
Time:	00:15:24	**BIC**	2100.363
Sample:	0	**HQIC**	2092.038
	- 123		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6921	0.161	4.292	0.000	0.376	1.008
ma.L1	-0.9038	0.110	-8.228	0.000	-1.119	-0.689
ar.S.L12	0.8699	0.079	10.966	0.000	0.714	1.025
ma.S.L12	-0.4584	0.169	-2.709	0.007	-0.790	-0.127
sigma2	1.358e+06	1.21e+05	11.248	0.000	1.12e+06	1.59e+06

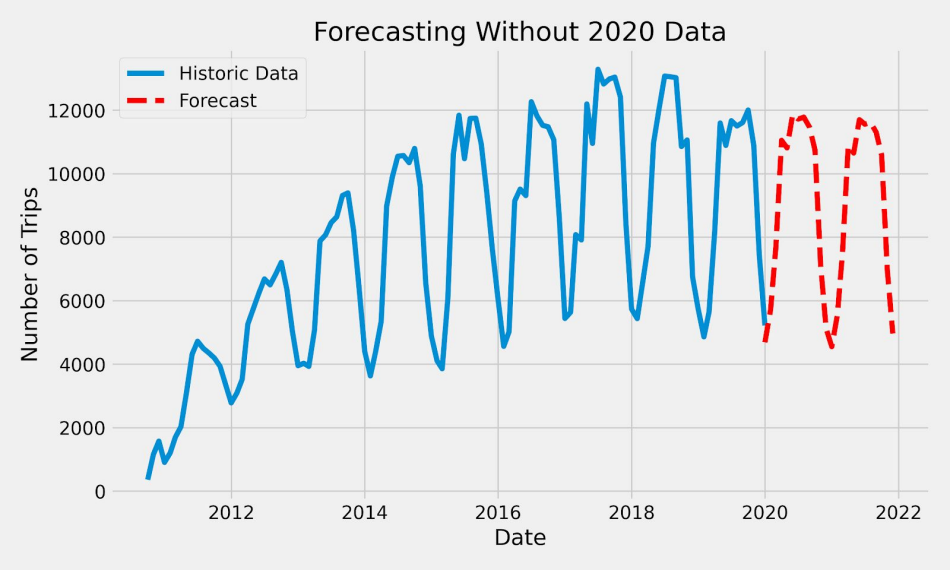
Ljung-Box (Q):	57.54	**Jarque-Bera (JB):**	114.31
Prob(Q):	0.04	**Prob(JB):**	0.00
Heteroskedasticity (H):	2.97	**Skew:**	-0.89
Prob(H) (two-sided):	0.00	**Kurtosis:**	7.40

With this model, I forecasted the Capital BikeShare's usage trends for 2021 shown below in red dashed lines. While the model performed as expected (following the downward trend seen in 2019 and 2020, along with the seasonal lows and peaks), this is not a forecast

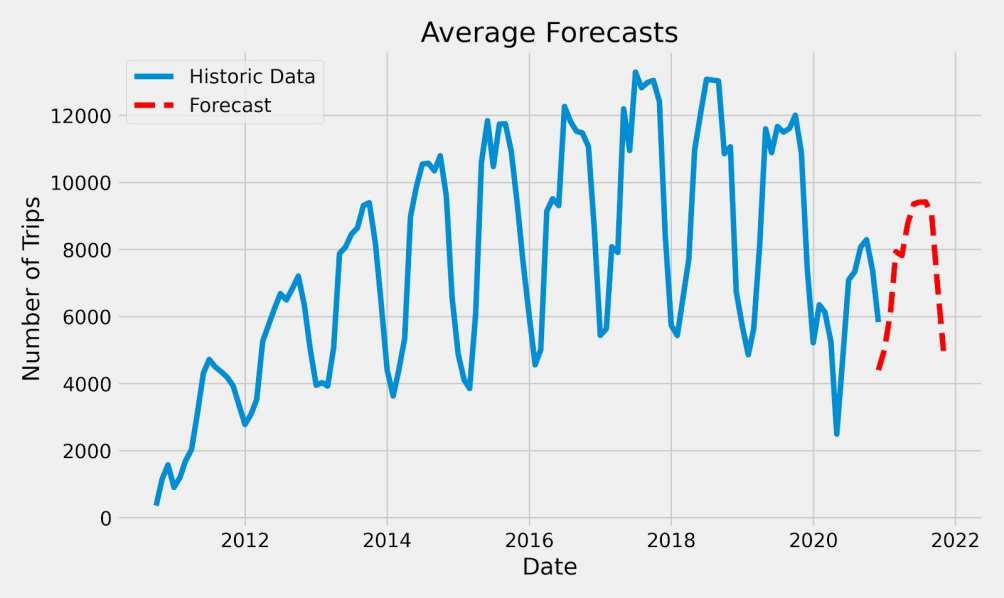


that we can expect to come true. With vaccines being rolled out and a better understanding of the importance of safety procedures and guidelines, it is naive to assume that people will be in quarantine as strictly as when the pandemic first broke out last year. More employees will return to offices, and people are more likely to use public services. Because of this, we can expect to see an increase in Capital BikeShare usage. To account for this flaw in the model, I wanted to see how forecasts will look if we excluded 2020 data.

I performed the same modeling steps using *pmdarima*'s auto_arima and trained the model without 2020 data. The forecast shows that throughout 2020 and 2021, there will be a plateau in usage. This would be a reasonable forecast had there not been a global pandemic.

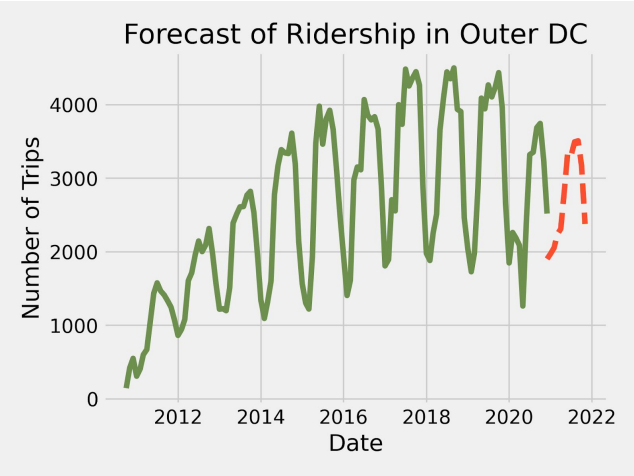


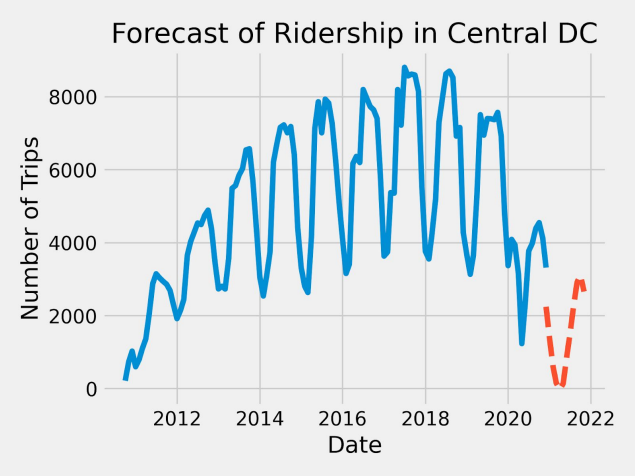
But this does not solve the flaw of having 2020 as an outlier. In order to fix this, I decided to average out the two forecasts for 2021. The average of the two forecasts presents a



more believable forecast, and one that we can expect to actually observe in real life. Usage will definitely increase compared to last year, but it also would not completely go back to normal levels due to the fact that the pandemic is not yet over.

During the EDA process, I found that usage definitely decreased in downtown D.C. areas with outer D.C. stations seeing a lot more trips. Because of this, it would be incorrect to assume that both areas will see an equal increase in usage in 2021. I forecasted 2021 usage for both central and outer D.C. and found that while the outer D.C. areas will see an increase in usage, the central D.C. areas will see a significant decrease in usage. This is not a fair forecast for downtown D.C. as the forecast is heavily dependent on its past values and usage downtown diminished drastically in the past year.





6. Conclusions

2020 has drastically impacted any model to accurately make forecasts due to it being a drastic outlier from every other year. People were forced to stay home, businesses were forced to shut down due to safety measures or financial reasons, and almost everyone's lives went through a drastic change throughout 2020. However, now that we as a society have a better understanding of the importance of safety precautions and vaccines are slowly being given out, we can expect to see some level normalcy return to our lives.

From the ARIMA models and analysis performed in this project, Capital BikeShare can expect an increase in usage from last year. People will be more likely to use the bikes outside the central D.C. areas than in the past, with stations in parks seeing more usage as people will look towards riding the bikes more for leisure and to relieve pandemic fatigue rather than for getting from one building to another. It will be important for Capital BikeShare to maintain its docking stations in downtown areas but also maintain and increase the availability of bikes in outer D.C.