Analyzing and Forecasting Capital BikeShare Usage After COVID-19

By: Ji Ho Lee
Mentor: Ken Cavagnolo

What is Capital BikeShare?

- Bicycle-sharing system company focused in Washington D.C. and some areas of Virginia and Maryland
- Started in 2010 with 400 bikes and 49 stations
- Expanded to 4,300+ bikes in 500+ stations
- Offers multiple ways to ride
    - $7/month membership
    - $2/30 min. for non-members
    - $8/day for visitors

What is the goal?

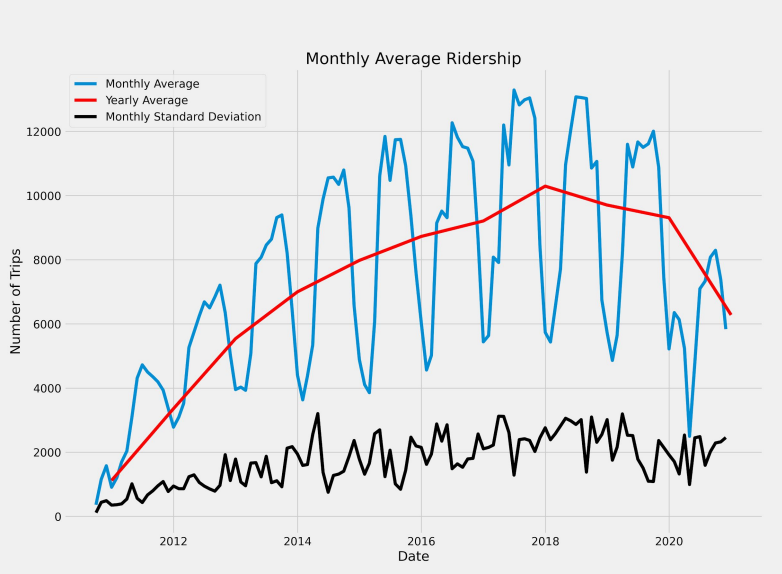
- COVID-19 has had drastic effects on every business
    - Not always obvious

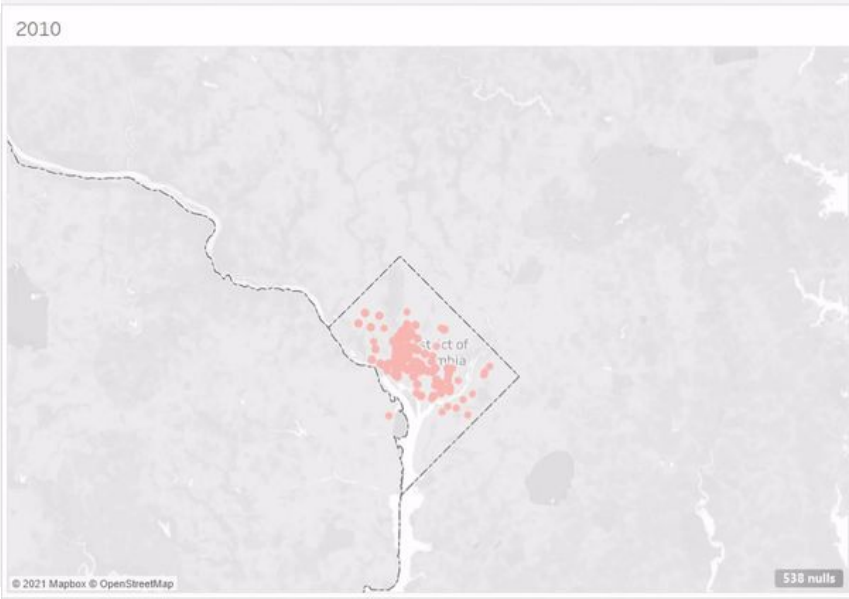
1. Analyze usage trends to find underlying effects/changes
2. Forecast 2021 usage using ARIMA modeling
3. Provide actionable insights for Capital BikeShare based on analysis/forecasting

What does the data look like?

- Data obtained from Capital BikeShare's website
  - Contains information on all trips taken starting from 2010 until now (over 26 million entries)
  - Start/end date/time, start/end station, start/end latitude/longitude, bike number (not included in recent data), membership status, ride type (regular vs. electric bike) and trip duration (older datasets)
- Most null values came from missing station name/id
  - Used latitude/longitude coordinates to obtain street names or significant identifiers to fill in missing station names
- Calculated trip duration for recent datasets
- Removed unnecessary columns (trip ID, bicycle ID, and ride type)

What does the data look like? (Part 2)

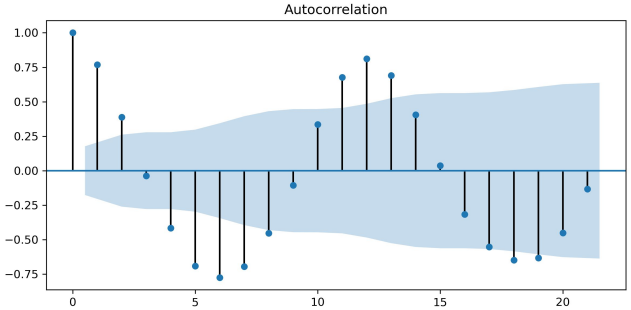




How can we forecast usage for 2021?

- Autoregressive Integrated Moving Average (ARIMA) model
  - Explains and forecasts a time series based on its own past values (lags) and lagged forecast errors
  - Seasonal ARIMA for time series with seasonality component
  - Characterized by 3 main components *p*, *d*, and *q* values
- Steps to ARIMA modeling:
  - Stationarize time series with differencing
  - Determine orders of AR and MA (plot ACF and PACF, use auto_arima in Python)
  - Fit data to model and make forecasts

Modeling Results



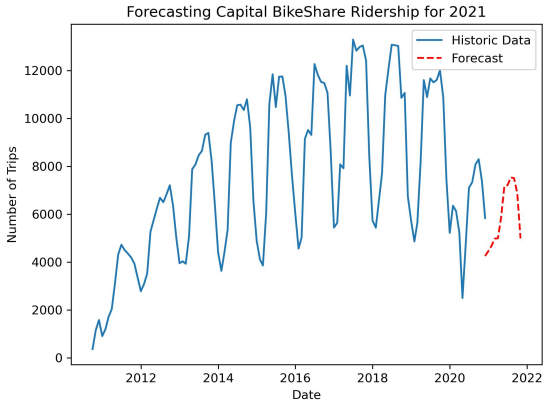
SARIMAX Results

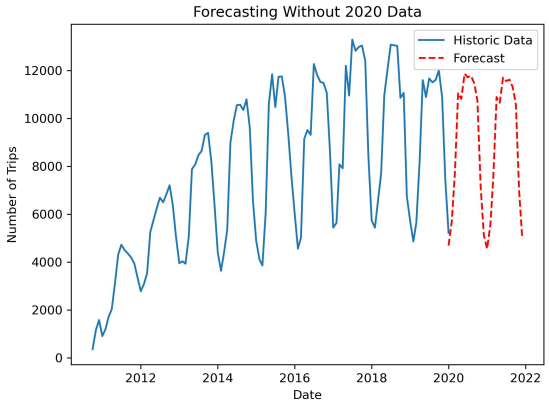
Dep. Variable:	y	**No. Observations:**	123
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 12)	**Log Likelihood**	-1038.172
Date:	Wed, 27 Jan 2021	**AIC**	2086.343
Time:	22:07:33	**BIC**	2100.363
Sample:	0	**HQIC**	2092.038
	- 123		
Covariance Type:	opg		

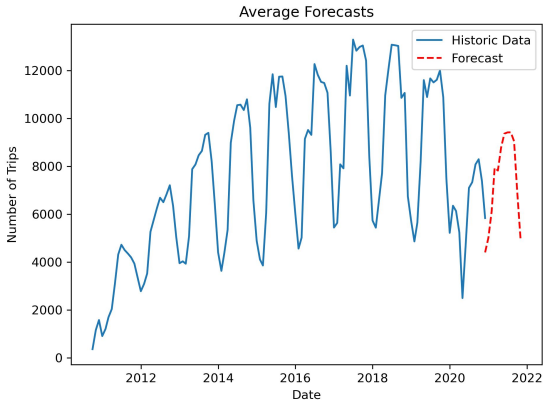
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6921	0.161	4.292	0.000	0.376	1.008
ma.L1	-0.9038	0.110	-8.228	0.000	-1.119	-0.689
ar.S.L12	0.8699	0.079	10.966	0.000	0.714	1.025
ma.S.L12	-0.4584	0.169	-2.709	0.007	-0.790	-0.127
sigma2	1.358e+06	1.21e+05	11.248	0.000	1.12e+06	1.59e+06

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

Forecast Results

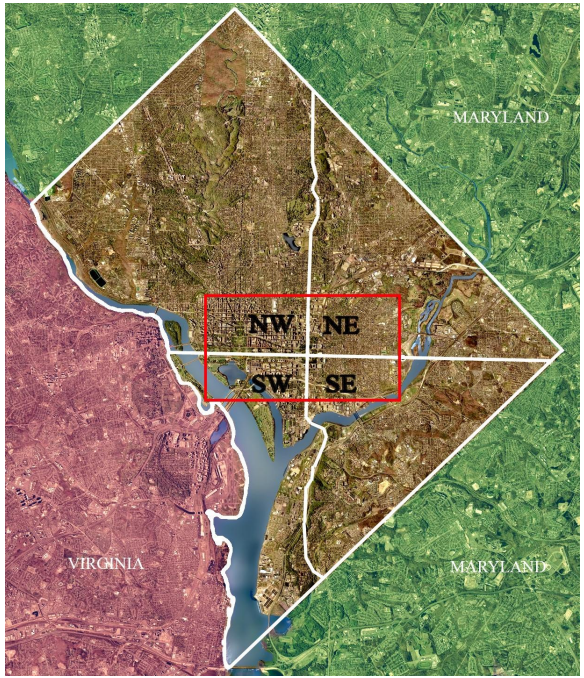




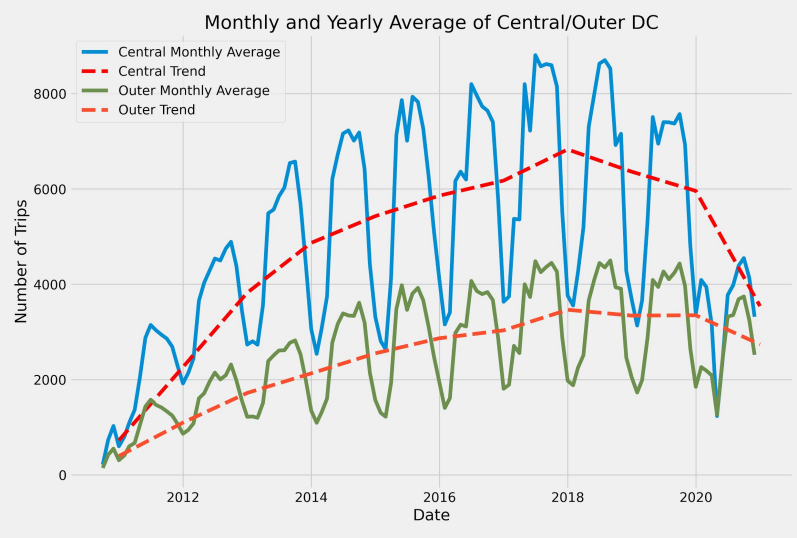


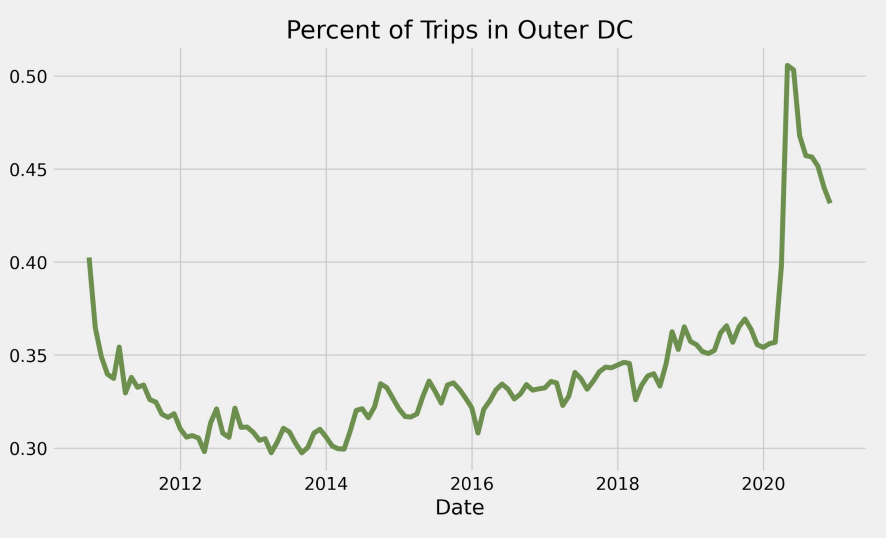
Can we expect a rise in usage everywhere?

- Usage became less concentrated in central D.C. in 2020
  - Work-from home
- Increased usage in areas outside of central D.C.
  - Mostly stations located in local parks
  - Could be used more for leisure than a mode of transportation
- Define downtown D.C. as every station within red rectangle and outer D.C. as everything else



Central/Outer D.C. Usage Trends





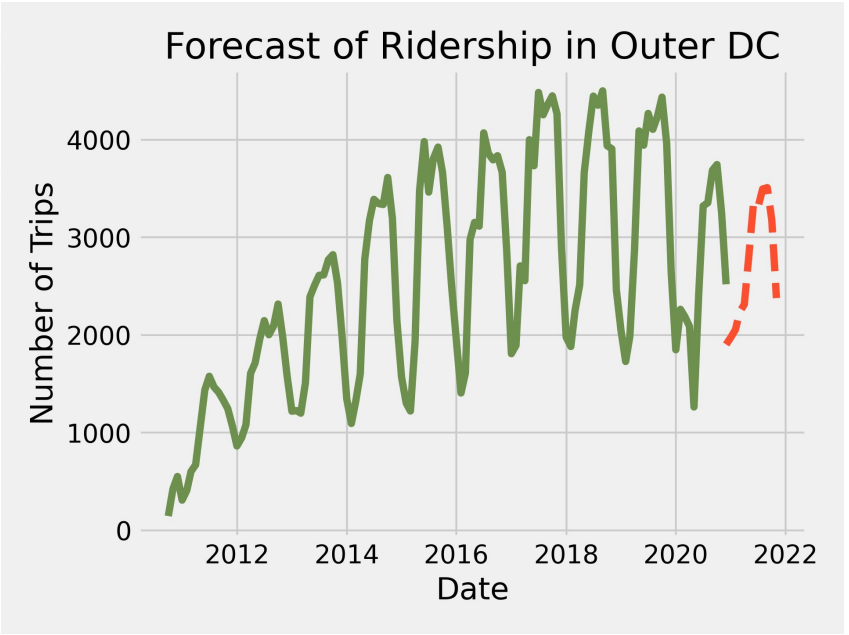
Forecasting Outer D.C. Usage

SARIMAX Results

Dep. Variable:	y	No. Observations:	123
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 12)	Log Likelihood	-903.969
Date:	Tue, 09 Feb 2021	AIC	1817.938
Time:	00:33:36	BIC	1831.958
Sample:	0	HQIC	1823.632
	- 123		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6209	0.138	4.515	0.000	0.351	0.890
ma.L1	-0.9126	0.082	-11.190	0.000	-1.072	-0.753
ar.S.L12	0.8961	0.074	12.181	0.000	0.752	1.040
ma.S.L12	-0.4658	0.173	-2.693	0.007	-0.805	-0.127
sigma2	1.477e+05	1.3e+04	11.355	0.000	1.22e+05	1.73e+05

Ljung-Box (Q):	62.76	Jarque-Bera (JB):	145.50
Prob(Q):	0.01	Prob(JB):	0.00
Heteroskedasticity (H):	3.17	Skew:	-1.04
Prob(H) (two-sided):	0.00	Kurtosis:	7.93



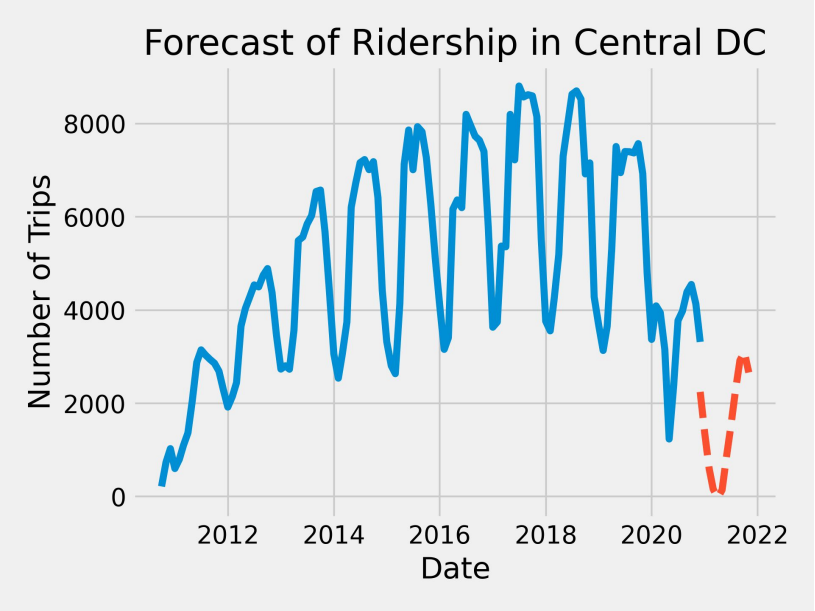
Forecasting Central D.C. Usage

SARIMAX Results

Dep. Variable:	y	No. Observations:	123
Model:	SARIMAX(2, 1, 2)x(2, 0, [], 12)	Log Likelihood	-988.140
Date:	Tue, 09 Feb 2021	AIC	1990.281
Time:	00:34:32	BIC	2009.909
Sample:	0	HQIC	1998.253
	- 123		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.7031	0.068	24.884	0.000	1.569	1.837
ar.L2	-0.9198	0.070	-13.123	0.000	-1.057	-0.782
ma.L1	-1.8410	0.146	-12.629	0.000	-2.127	-1.555
ma.L2	0.9615	0.147	6.525	0.000	0.673	1.250
ar.S.L12	0.1627	0.148	1.095	0.273	-0.128	0.454
ar.S.L24	0.1897	0.155	1.223	0.221	-0.114	0.494
sigma2	8.928e+05	1.52e+05	5.868	0.000	5.95e+05	1.19e+06

Ljung-Box (Q):	41.41	Jarque-Bera (JB):	33.43
Prob(Q):	0.41	Prob(JB):	0.00
Heteroskedasticity (H):	3.51	Skew:	-0.46
Prob(H) (two-sided):	0.00	Kurtosis:	5.39



Conclusion

- COVID-19 caused 2020 to be a huge outlier in all time series data, making forecasting much more difficult
- Expect to see an increase of usage compared to 2020, but not completely back to normal levels
- Keep maintaining bikes and stations in central D.C. with more offices opening back up
- Outer D.C. will still have increased usage as not everyone will be back to work, more people will use these bikes to relieve pandemic fatigue