Exploring Micromobility and Factors in Growth and Success of the Programs

DATA 606 Capstone
Spring 2022, Professor Wang
David Fahnestock

What is Micromobility?

Shared-use simple methods of transportation, including bikeshare and scooter programs

- Help combat traffic congestion and pollution in cities
- Global micromobility market projected to grow from $44 billion in 2020 to over $214 billion by 2030
- Large cities have implemented micromobility programs (New York, Chicago, San Francisco)

Data Overview – Micromobility Usage

- Unit of Analysis is at the city level
- Each row represents one trip from point A to point B
- Each city can have millions of trips per year
- Not all city programs track the same data elements
- Data from other sources to be used in conjunction for analysis (city unemployment, demographics)

Standard Data Fields
Date/Time of Departure
Departure Station
Date/Time of Arrival at Destination
Destination Station
Type of Transport (docked/undocked bike, e-bike)
Type of Customer (daily, monthly pass)

Extra Data Fields (not always present)
Departure Geocoordinates
Destination Geocoordinates
Gender
Year of Birth

Data Derived Elements

Derived fields created to provide additional features for use in analysis

Derived Data Fields
Duration of Trip
Month
Year

Data Overview – Supplemental Data

- Historic Unemployment Rates from the U.S. Bureau of Labor Statistics.
- Historic daily weather conditions for each of the cities from the U.S. National Oceanic and Atmospheric Administration (NOAA)

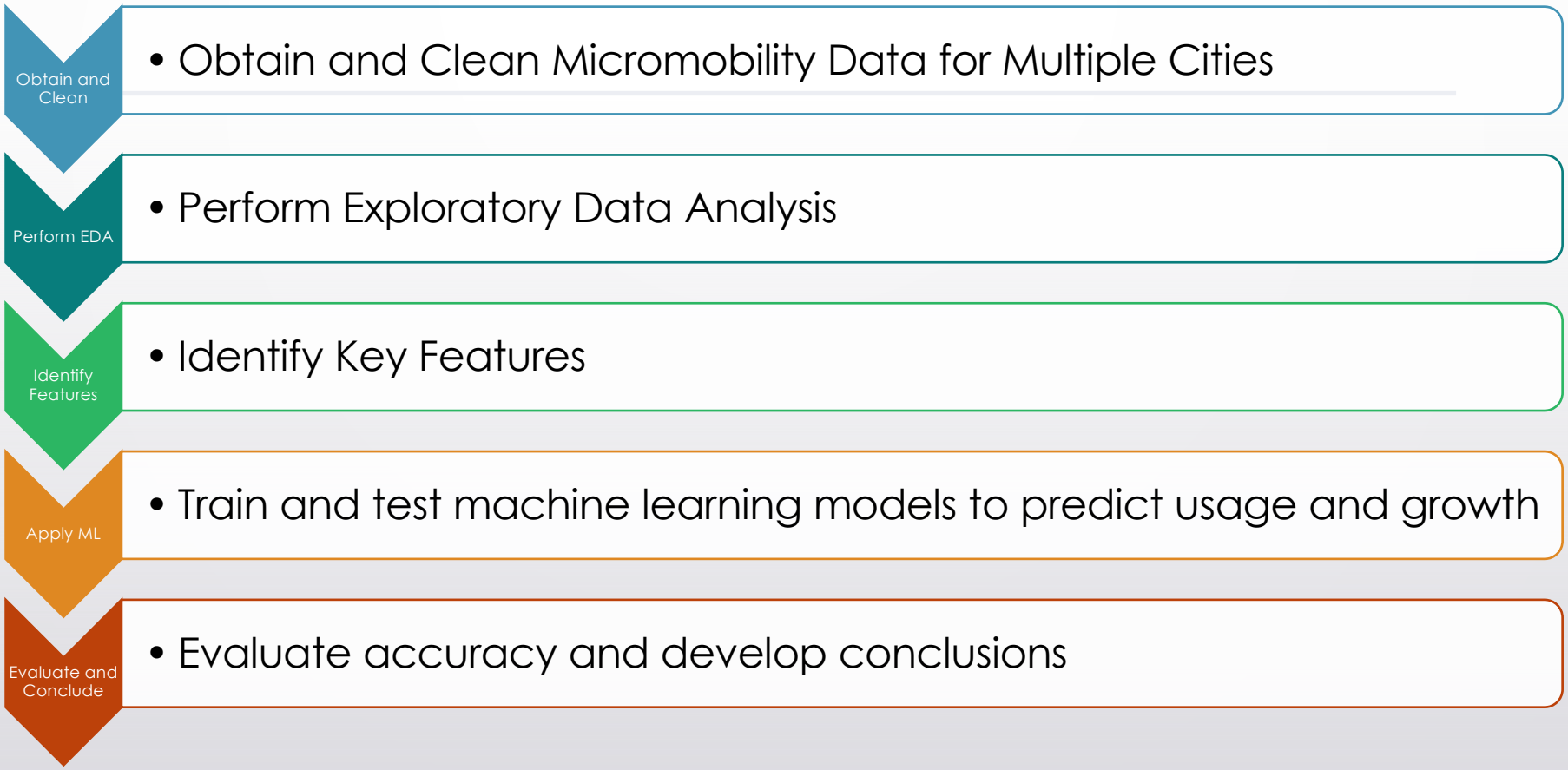
Unemployment Rate Data
State
Year
Month
Unemployment Rate

NOAA Daily Historic Weather Data
City
High Temperature
Average and Peak 5 second Wind speed
Rain and Snow Precipitation

Research Questions

- What factors impact usage and growth of micromobility?
- Can machine-learning be used to accurately:
  - Predict growth or usage patterns in micromobility?
  - Identify other cities where micromobility programs would have a high likelihood of success?

Research Process



Data Cleansing Challenges

- Merging data from different time periods
- Field layout changes within each city over time
- Cities track different field elements
- Significant number of nulls impact usefulness
- Sheer volume of data

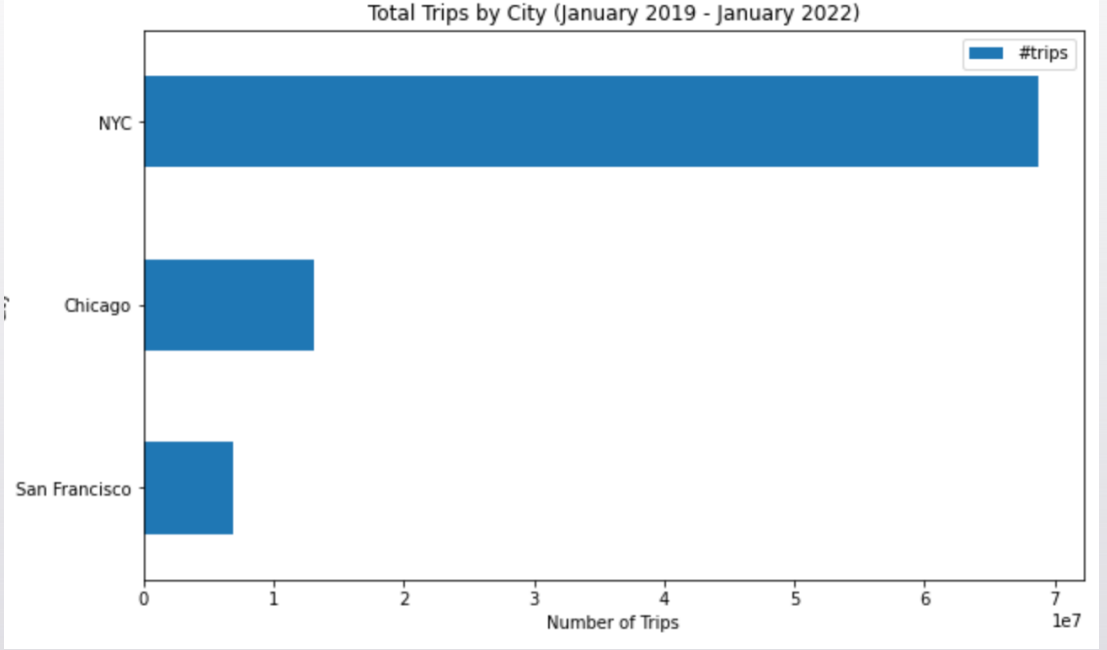
Overview of Cities

Metric	Chicago	New York City	San Francisco
Population (2020)	2,746,388	8,804,190	873,965
Median Household Income	$62,097	$67,046	$119,136
Median Age	34.8	36.9	38.3

Source: census.gov (2020 census)

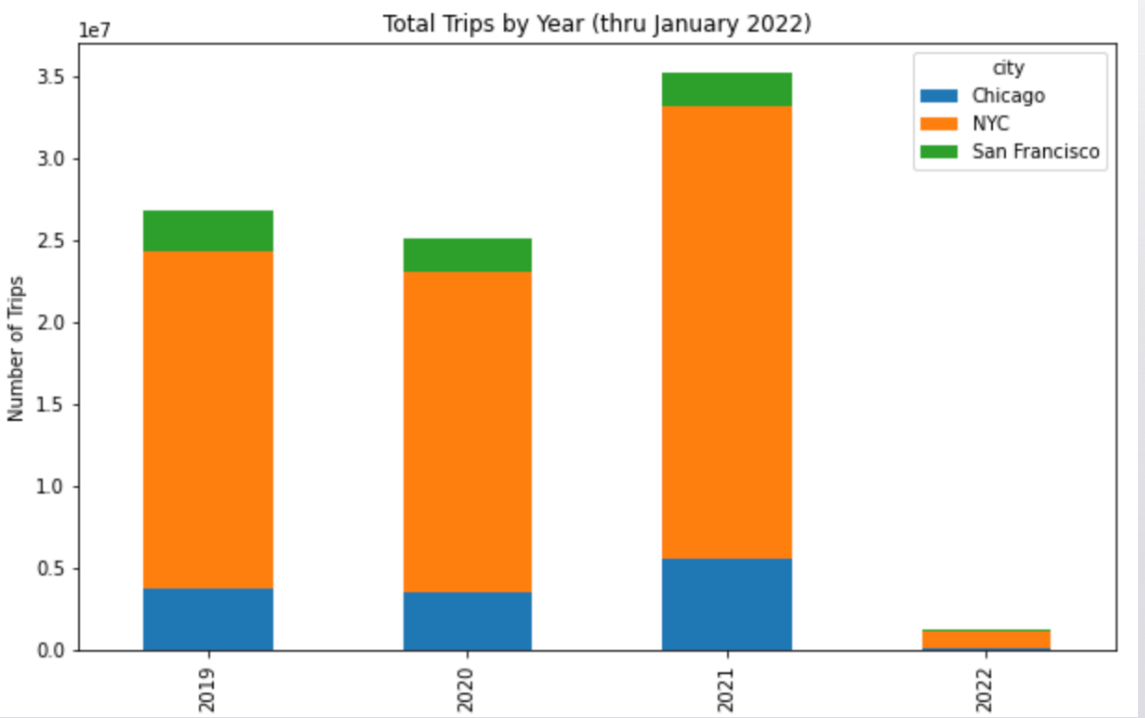
EDA – By City for January 2019 through January 2022

- Included three cities in our analysis
- NYC had nearly 70 million trips
- Chicago totaled 13 million trips
- San Francisco had about 7 million trips

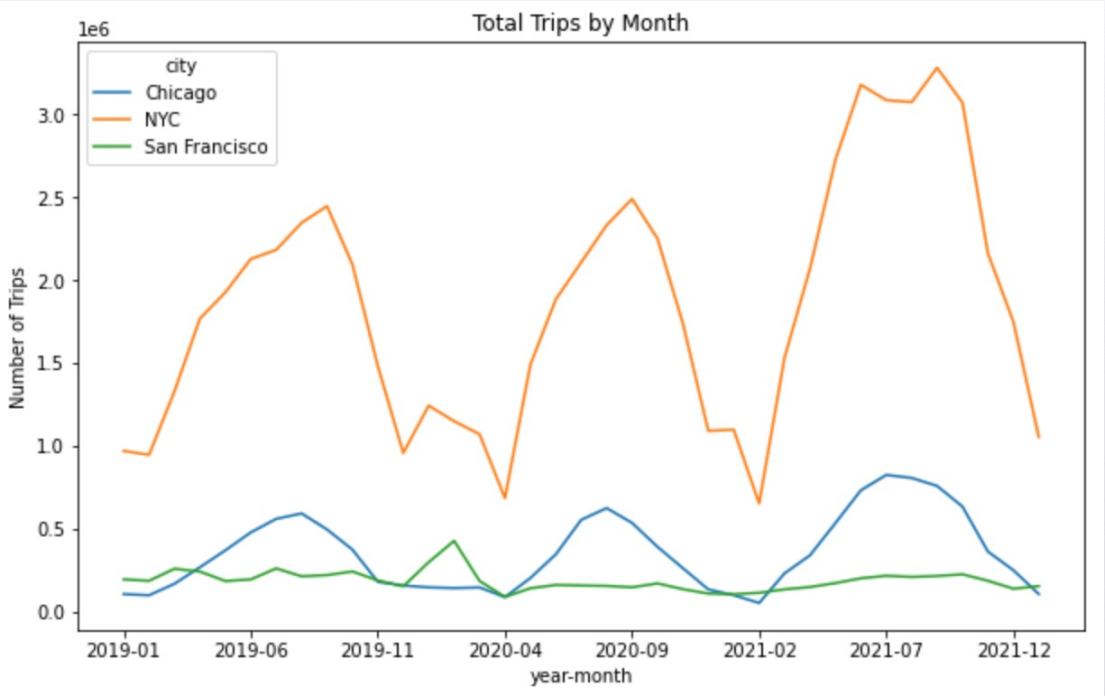


Summary of Trips by Year

- Decrease in usage in 2020 with the Covid-19 pandemic
- 2021 trips in total exceeded that of 2019 before the pandemic
- San Francisco's usage remained relatively steady

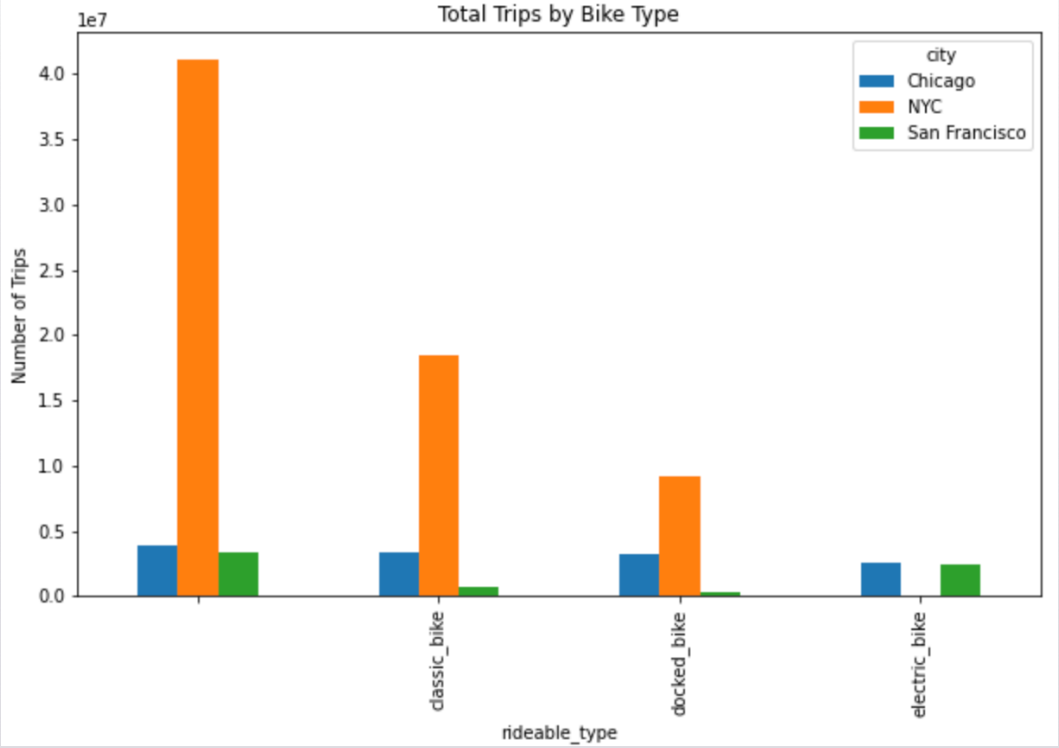


Seasonal Patterns in Usage



- Usage is clearly seasonal for NYC and Chicago
- San Francisco is relatively consistent throughout the seasons
- Chicago's usage most consistent over the three year period
- Usage in NYC has skyrocketed in 2021

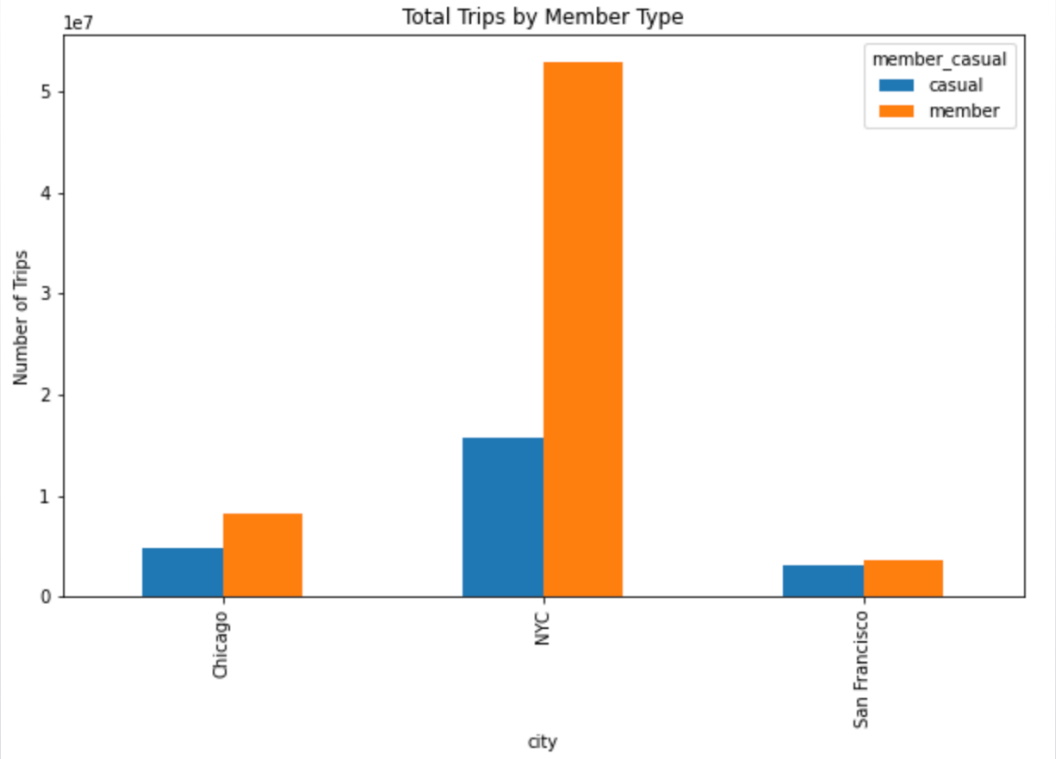
Types of Bikes

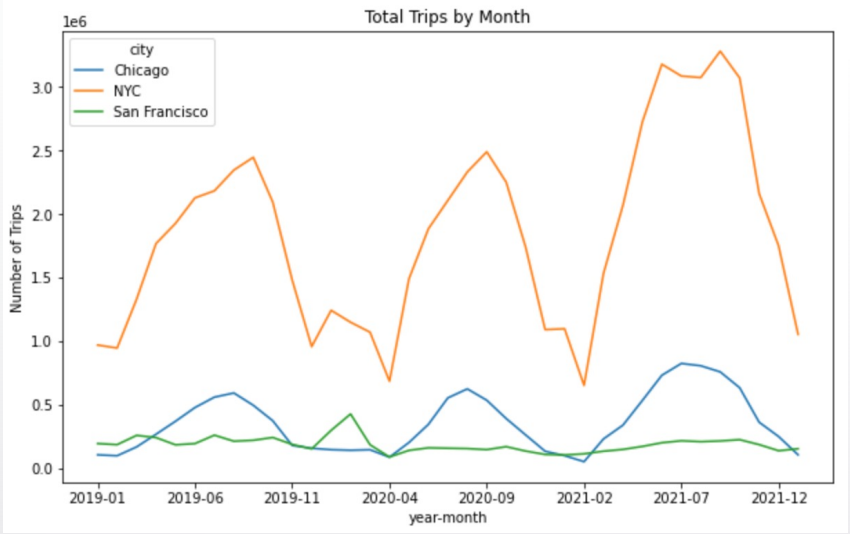


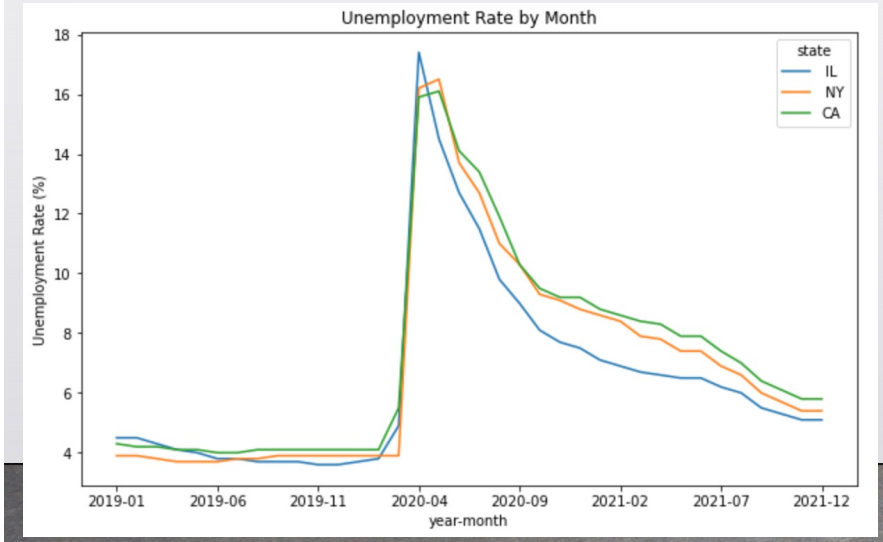
- Large number have no type specified
- E-Bike usage minimal in NYC but significant in San Francisco
- Classic dockless bikes are predominant in NYC

Types of Users

- All three cities have more trips by members than casual riders
- NYC's usage by members far exceed casual riders
- San Francisco is nearly even between members and casual riders







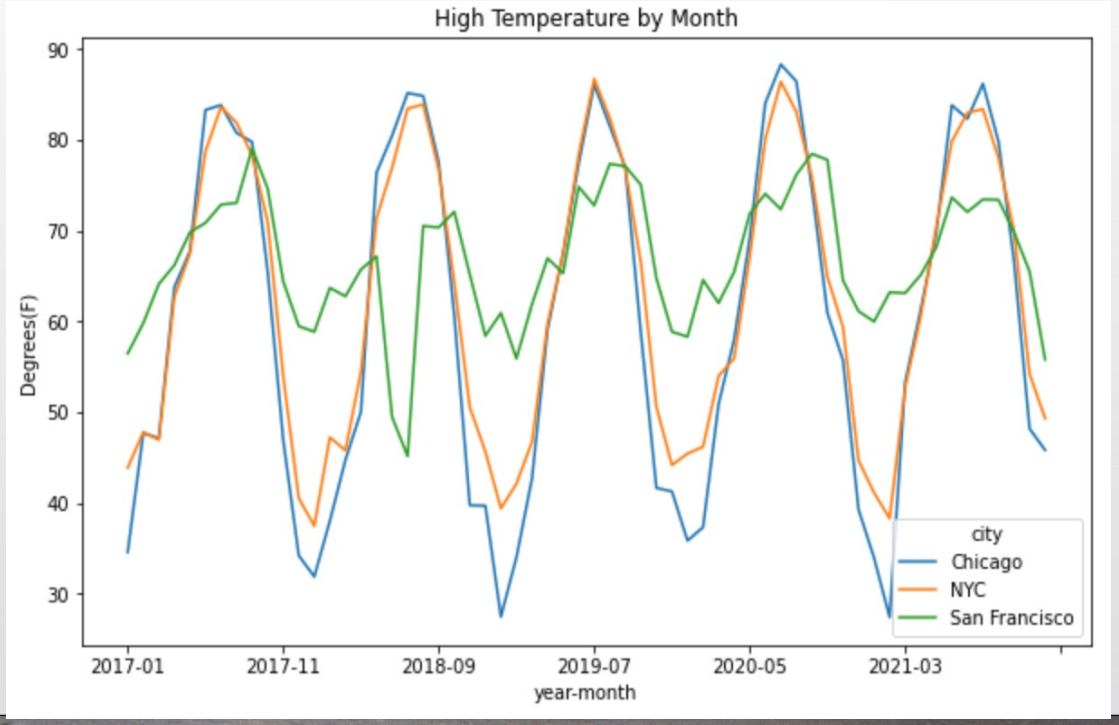
Unemployment Rates

- Unemployment rates across all three states had a similar trend over the period. The sharp spike in March 2020 marks pandemic shutdowns.
- No correlation evident between micromobility usage (top) and unemployment rates (bottom)

Source Data: Bureau Labor Statistics (bls.gov)

Weather - Temperature

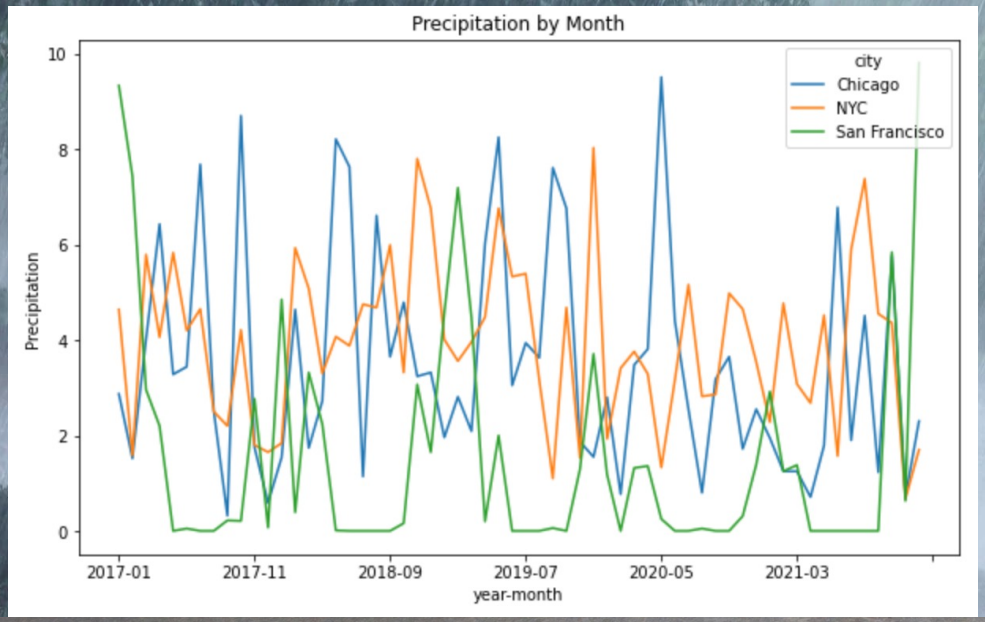
- San Francisco fluctuates less than Chicago and New York City over the seasons.
- Chicago and New York City have relatively similar temperatures and seasonal patterns.
- These fluctuations are similar as what was seen in micromobility usage for the cities.

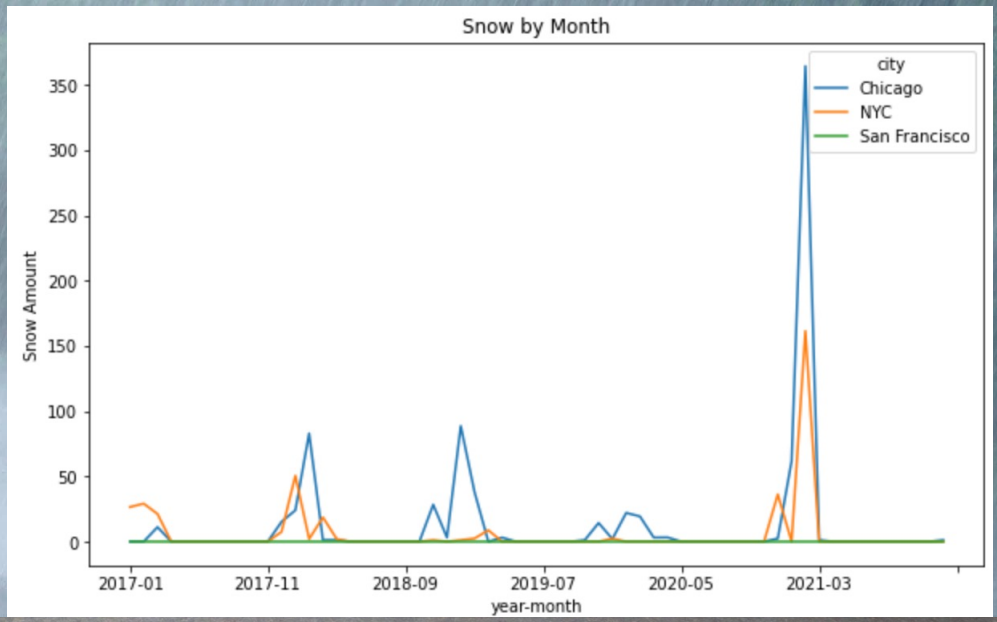


Data Source: NOAA

Weather - Precipitation

- Precipitation levels fluctuate significantly in each city, San Francisco has dry summers.
- Chicago has the most snowfall. San Francisco has essentially no snow.

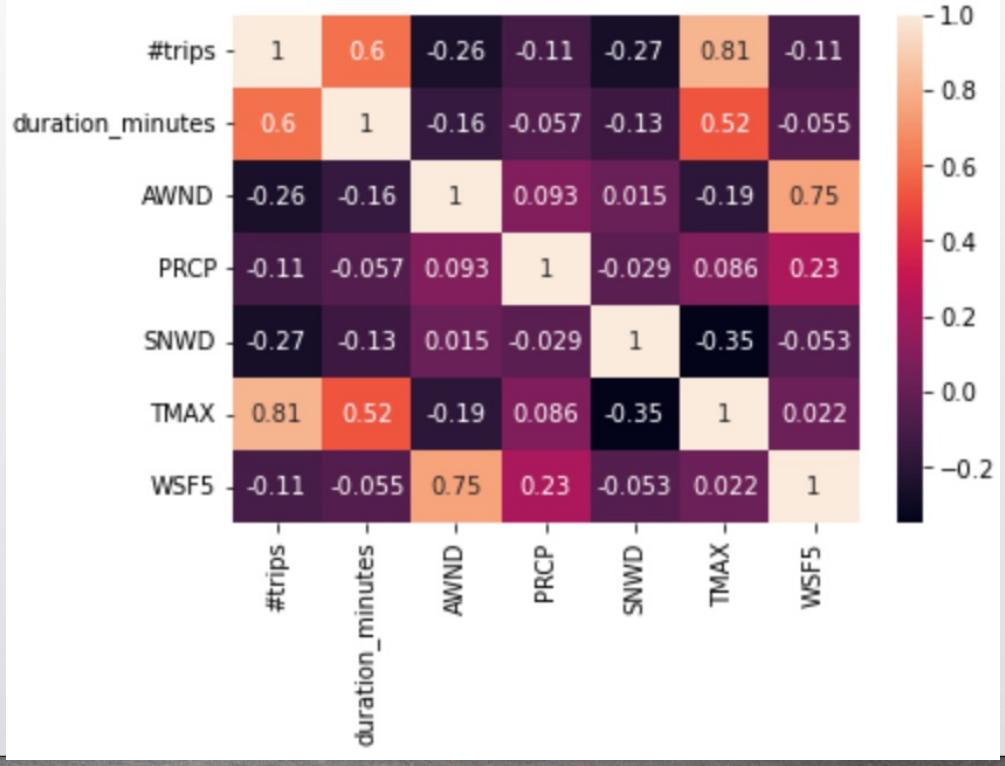




Data Source: NOAA

Correlations between Weather and Micromobility Usage

- There are strong correlations between #Trips/Duration and High Temperature (TMAX).
- As expected, there are negative correlations between #Trips/Duration and precipitation, snow(SNWD), and wind (AWND, WSF5)



Matrix for Chicago for 2017 - 2021

Feature Selection Considerations

From exploratory data analysis, the following feature considerations were determined:

Feature(s)	To be Used for Machine Learning?	Reasoning
Median Income Level	No	Changes in Median income occur over long periods of time. This analysis is for a relatively short period of time. Perhaps once there are decades of data, this would be suitable.
Subscriber Type, Type of Bike, Age, Gender	No	A considerable portion of the data had nulls for these features. Therefore, they will not be reliable for ML.
Unemployment Rates	No	The sharp increase in unemployment during the pandemic does not appear to impact micromobility usage. Including this would likely negatively impact the accuracy of ML models.
Trip Date	Yes	The date of trips will be a key element in ML models for time series based predictions.
Trip Count and Duration	Yes	Trip count and duration will be target features for prediction.
Weather Conditions	Yes	Weather conditions appear to impact micromobility usage.

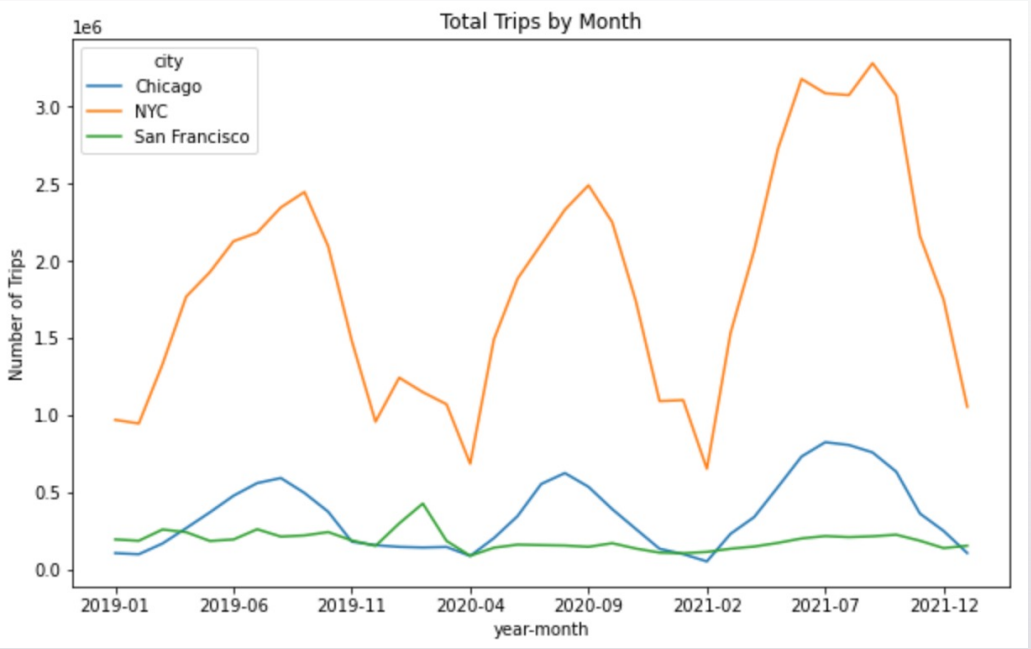
Machine Learning Models Applied

ARIMA – Univariate auto-regressive model, useful for time-series forecasts. It has been used effectively to forecast inflation in Ireland. Seasonal ARIMA (SARIMA) is a variation that accounts for seasonal patterns.

Long Short-Term Memory (LSTM) Neural Network – Recurrent Neural Network found to be suitable for time series based forecasting with multivariate input.

Machine Learning - ARIMA

- **AutoRegressive Integrated Moving Average (ARIMA)** is a univariate analysis useful for prediction when using time series data.
- **Seasonal ARIMA (SARIMA)** is an extension of ARIMA that accounts for time series data that contain a seasonal element. From EDA, it appears there is a seasonal component to micromobility usage for NYC and Chicago.



ARIMA – Target Feature Selection

As ARIMA is a univariate analysis, the only input will be the time (month) and then our target feature will be the number of trips

Number of Trips

- To predict the total number of trips in the future

SARIMA Methodology

- We will use SARIMA to predict usage for NYC and Chicago in terms of number of trips (our Target Variable)
- We will need to determine the below 7 elements to build our model

Elements of SARIMA	
SARIMA(p,d,q)(P,D,Q)m	
p: Trend autoregression order. **d**: Trend difference order. **q**: Trend moving average order.	**P**: Seasonal autoregressive order. **D**: Seasonal difference order. **Q**: Seasonal moving average order. **m**: The number of time steps for a single seasonal period.

SARIMA – Determine Parameter Value ‘p’

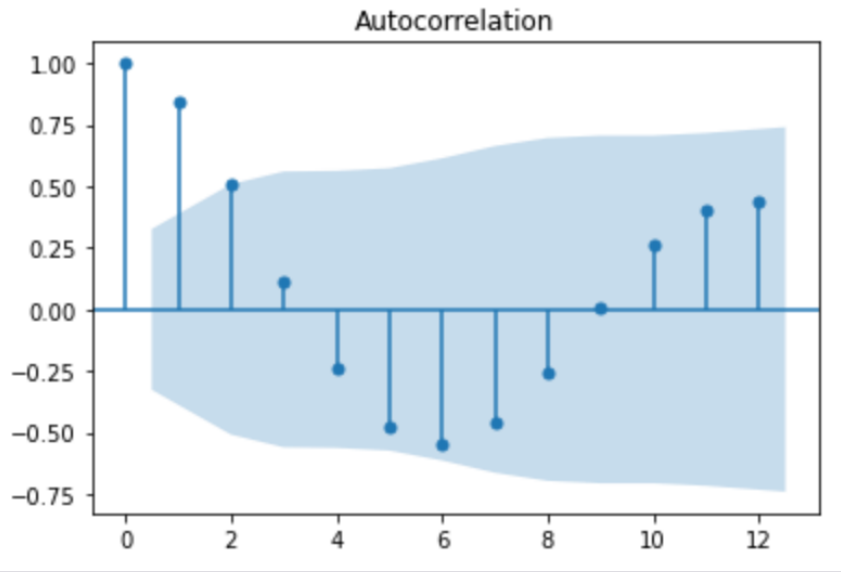
SARIMA(**p**,d,q)(P,D,Q)m

- Parameter ‘p’: The ACF plot is used, and it shows the autocorrelations measuring the relationship between an observation and its previous one.

- **p will be 2** as it is the maximum lag with a value in the ACF plot external to the confidence interval shaded in blue.

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
```

```
plot_acf(df_chicago['#trips'], lags=12)
plt.show()
```



SARIMA – Determine Parameter Value ‘d’

SARIMA(**2**,**d**,q)(P,D,Q)m

```
from statsmodels.tsa.stattools import adfuller
```

- Parameter ‘d’: The Dickey-Fuller test is used to test for Stationarity of the dataset. For stationarity, the p value returned by the Dickey-Fuller test must be <= 0.05.

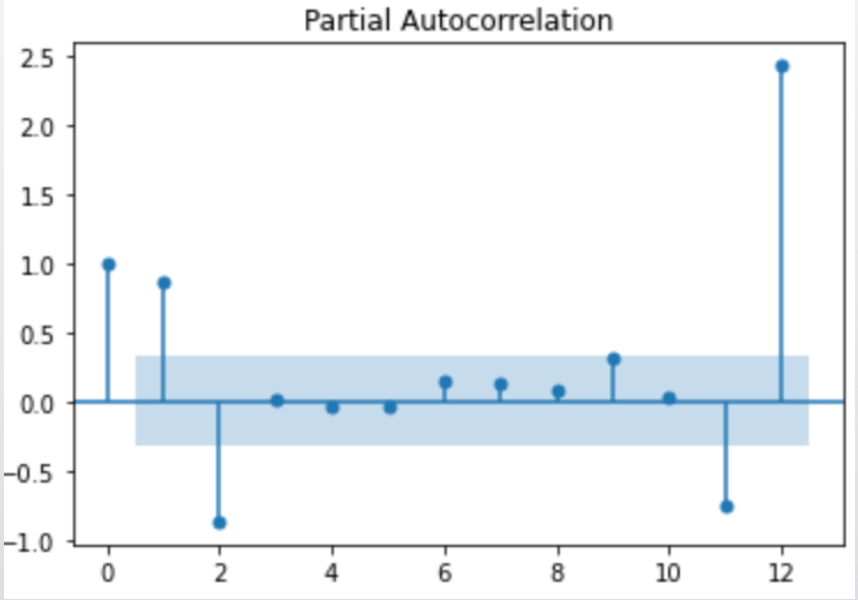
- **d will be 0** as the data is stationary without the need for any transformations. The p value was `5.9412047326864 06e-05`

SARIMA – Determine Parameter Value ‘q’ and ‘m’

SARIMA(**2**,**0**,**q**)(P,D,Q)**m**

- Parameter ‘q’: The PACF plot is used to determine the value for q.
- **q will be 2** as it the maximum lag with a value in the PACF plot external to the confidence interval shaded in blue.
- **m = 12** as the seasonality is annual (12 months)

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
plot_pacf(df_chicago['#trips'], lags=12)
plt.show()
```

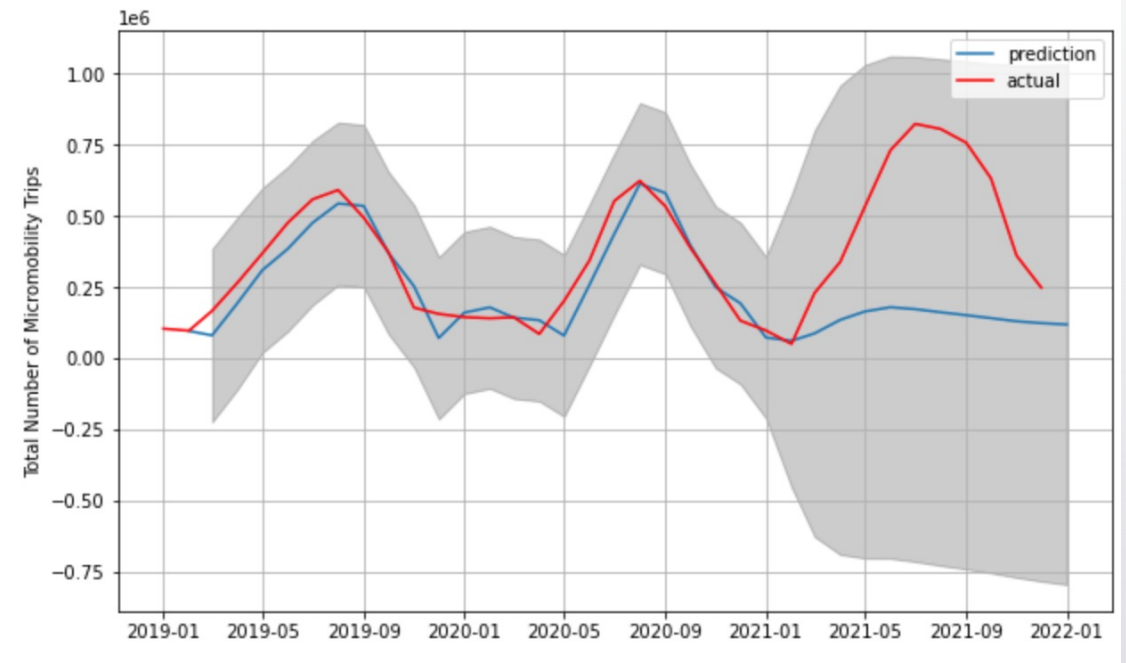


SARIMA – Results for Chicago

```
from statsmodels.tsa.statespace.sarimax import SARIMAX
```

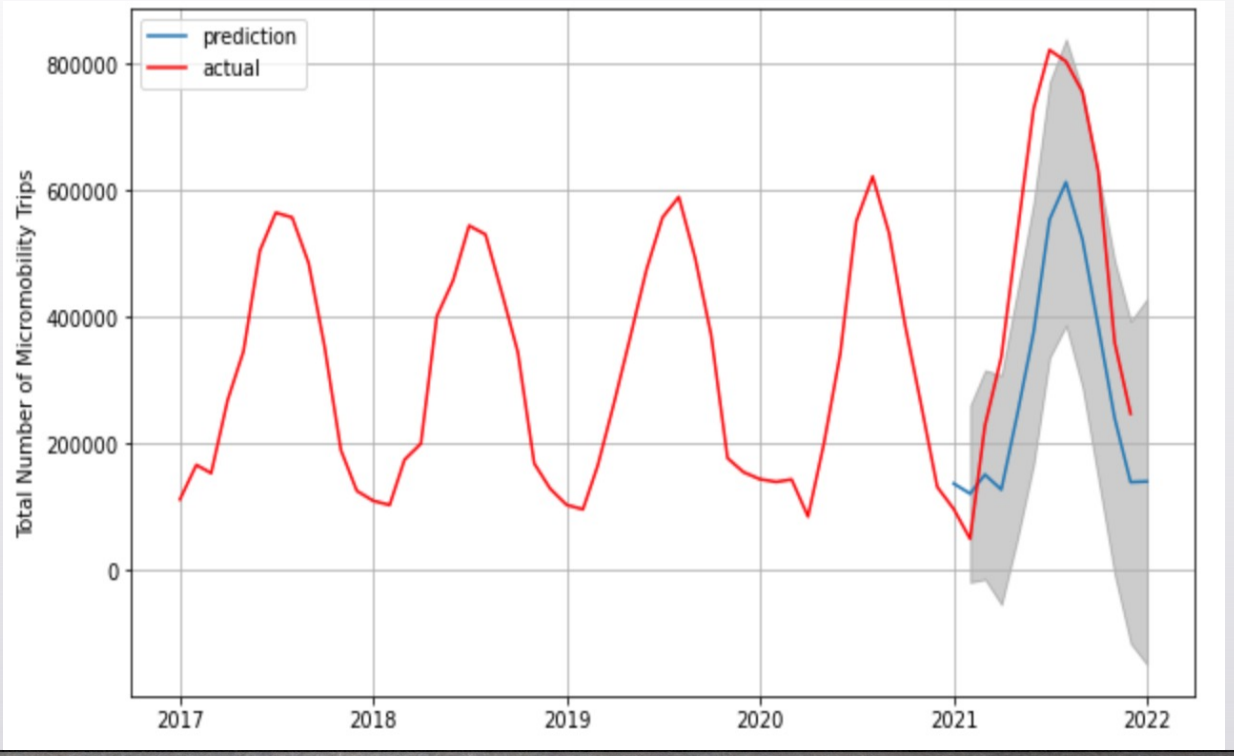
SARIMA(**2,0,2**)**12**

- Using the SARIMAX Python library, we use the parameters.
- Trained the model on the first 24 months of data. Prediction on the testing set is the last 12 months (starting January 2021).
- As shown, the performance of the model was poor.
- It is clear a longer span of data is needed in addition to potential additional parameter tuning.



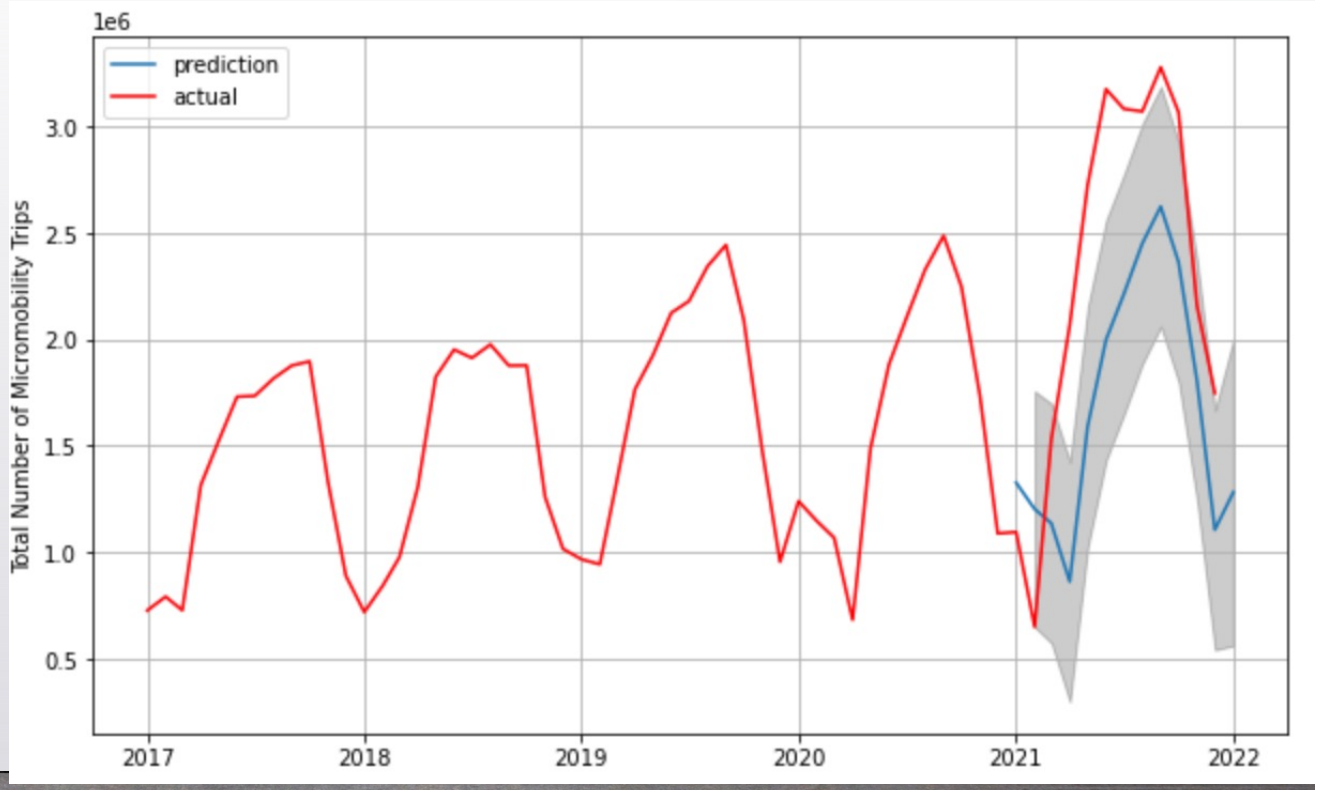
SARIMA – Results for Chicago (increased to include 5 years)

- With the poor performance using 3 years of data, I expanded the data to 5 years.
- The parameters were updated based on this new data span to SARIMA(2,1,2)(1,1,1)12.
- Training was done using the first 4 years and testing was on the final year.
- As can be seen in the chart for the final year, this generally resulted in a much better prediction.
- The normalized root mean squared error is 0.2789. This is decent performance but not great.



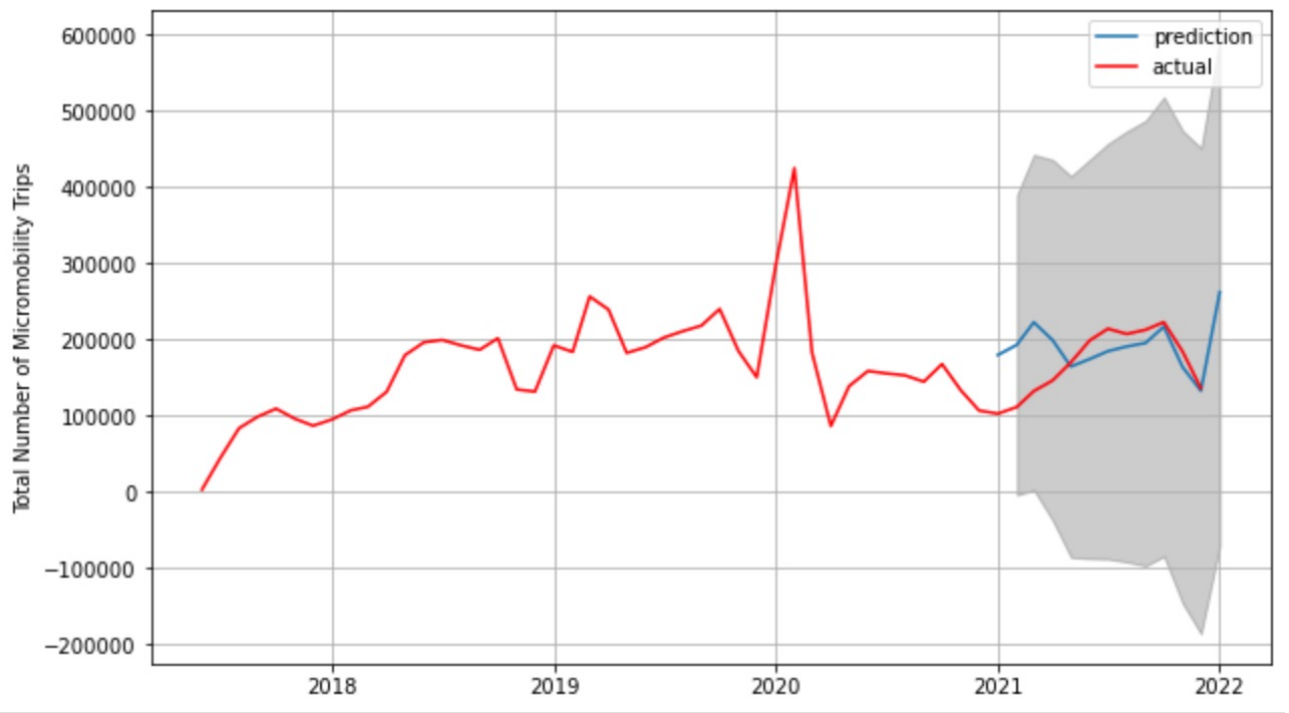
SARIMA – Results for New York City

- The parameters were updated for NYC by applying the same methodology resulting in SARIMA(2,0,2)(1,1,1)12.
- Similar to Chicago, the prediction was consistent with the prior years but did not account for the sudden increase in usage in 2021.
- The normalized root mean squared is 0.2849. This performed a little worse than Chicago.



SARIMA – Results for San Francisco

- Parameters: SARIMA(1,1,1)(1,1,1)12.
- There was an unusual spike at the beginning of 2020 that may have impacted model performance.
- The normalized root mean squared error is 0.5279. This is the worst performance of the three cities.
- As this city does not appear as seasonal, ARIMA was also used but that resulted in even worse performance (NRMSE was 0.5743)



ARIMA Summary of Results – Best to Worst

Rank	City	NRMSE	Notes
1	Chicago	0.2789	Seasonal usage with growth in 2021
2	New York City	0.2849	Similar usage pattern to Chicago
3	San Francisco	0.5279	Less seasonal pattern with unusual spike in usage in early 2020

LSTM Machine Learning – Feature Selection

- LSTM uses a recurrent neural network (RNN) and is useful for time series multi-variate analysis, in contrast with ARIMA which is univariate analysis.
- Based on our EDA, we will utilize these features:

Feature	Notes
Date of Trips	Analysis will be done with micromobility usage aggregated by date
Weather conditions	Weather conditions from NOAA on the same day within the city, such as temperature, precipitation, and wind conditions.
Target Feature: Number of Trips	The model will be used to predict the number of trips.

LSTM Machine Learning

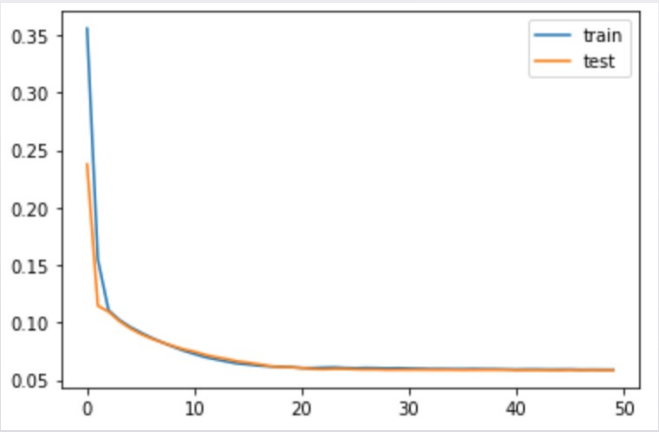
- The micromobility usage data aggregated by day was merged with the NOAA weather data by day.
- The final dataframe was then used to train and test LSTM.

city	DATE	#trips	AWND	PRCP	SNWD	TMAX	WSF5
Chicago	2017-01-01	1727	4.92	0.00	0.0	40.0	19.0
Chicago	2017-01-02	1960	6.26	0.11	0.0	40.0	19.0
Chicago	2017-01-03	4537	10.07	0.00	0.0	39.0	31.1
Chicago	2017-01-04	3269	17.00	0.00	0.0	19.0	34.0
Chicago	2017-01-05	2917	12.75	0.00	0.0	13.0	23.9

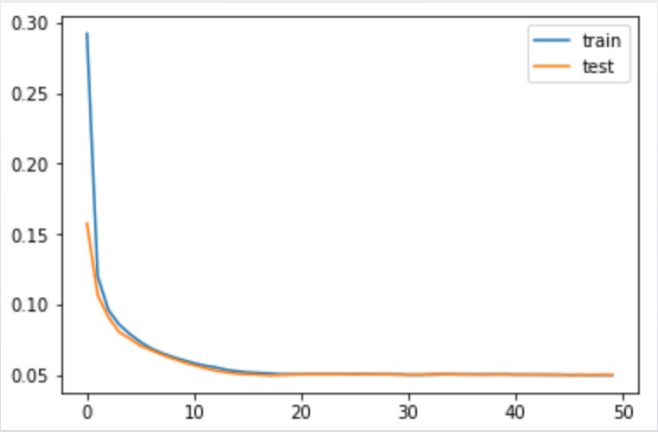
LSTM Machine Learning

- Training was done of the first 4 years and testing was on the final year.
- Below depicts the loss for each of the 50 epochs for each city.

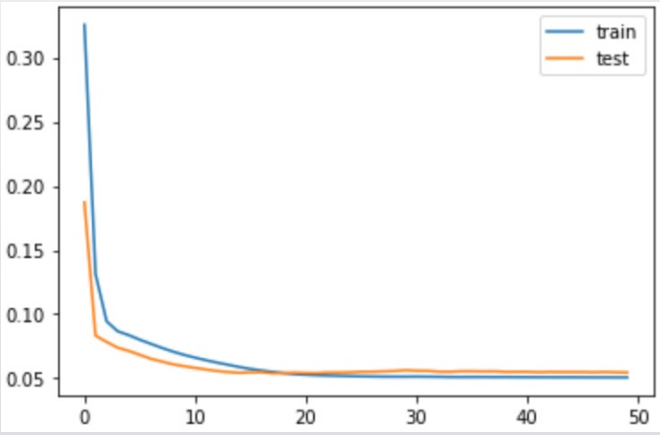
Chicago



New York City

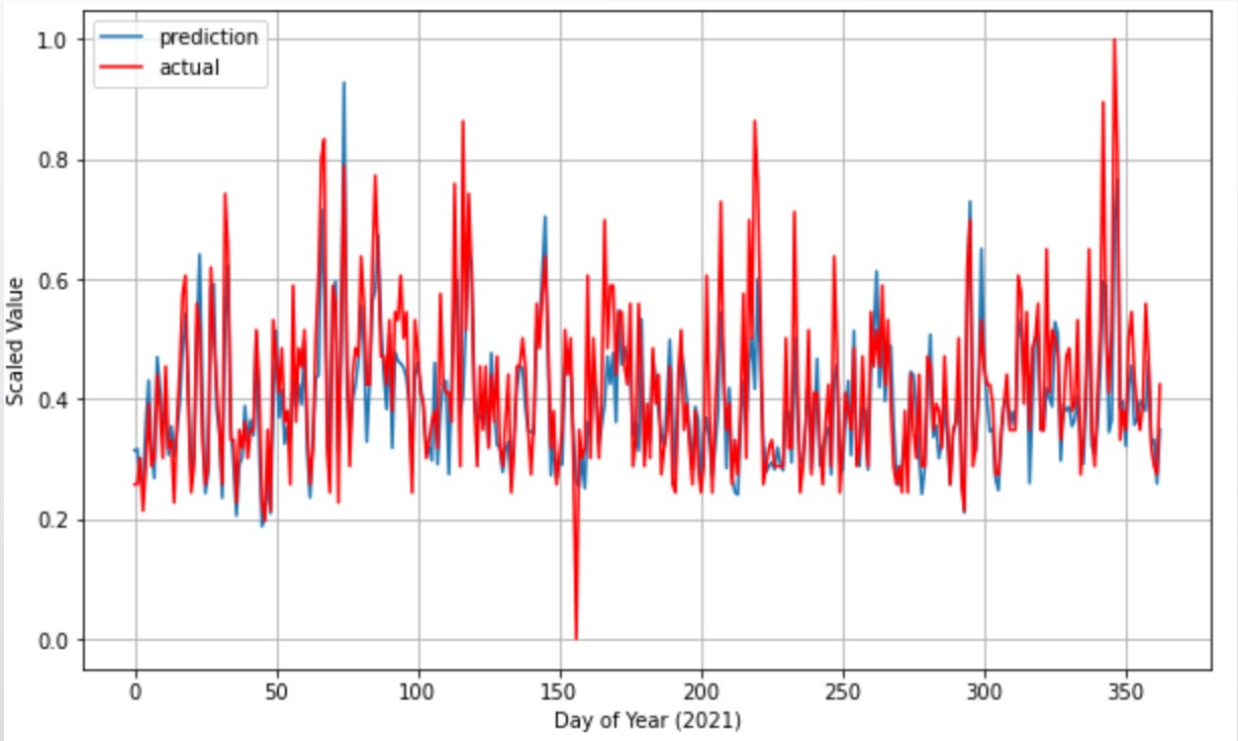


San Francisco



LSTM Machine Learning – Chicago Results

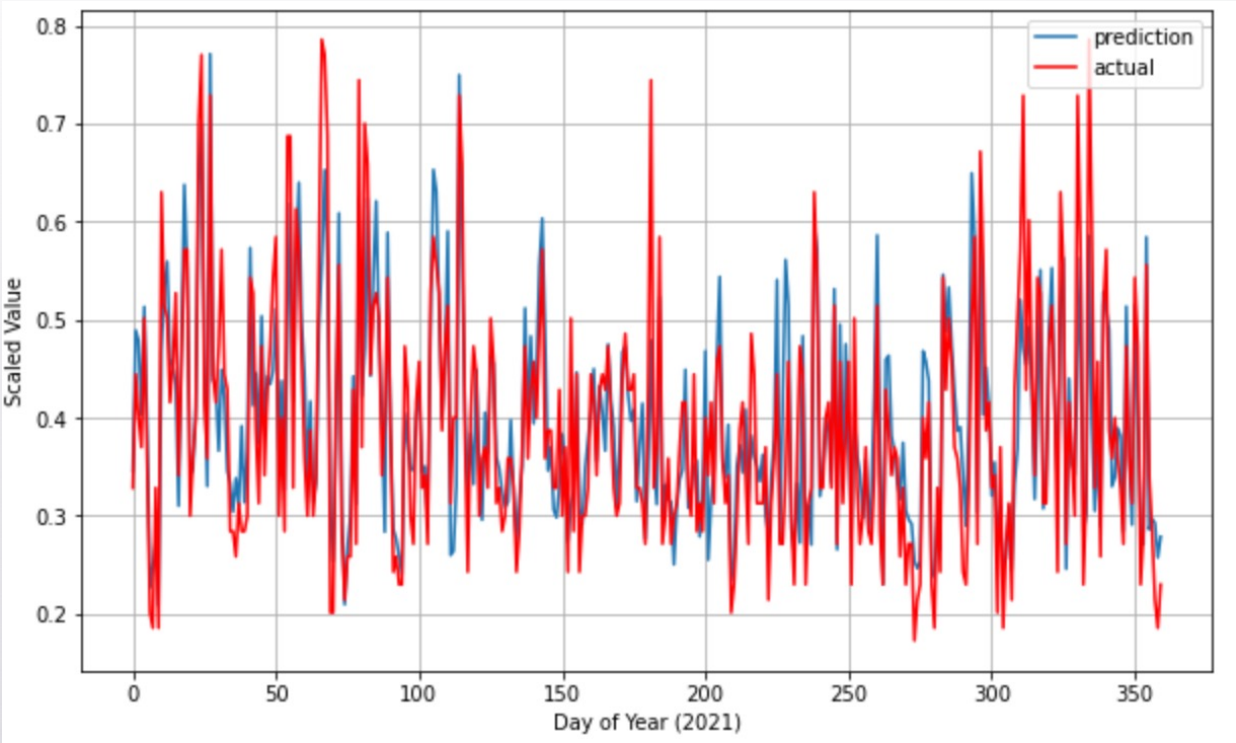
- The model performed well for Chicago.
- Normalized Root Mean Squared was: 0.0847



Scaled Predicted vs Actual for the test period (365 days of year 2021)

LSTM Machine Learning – NYC Results

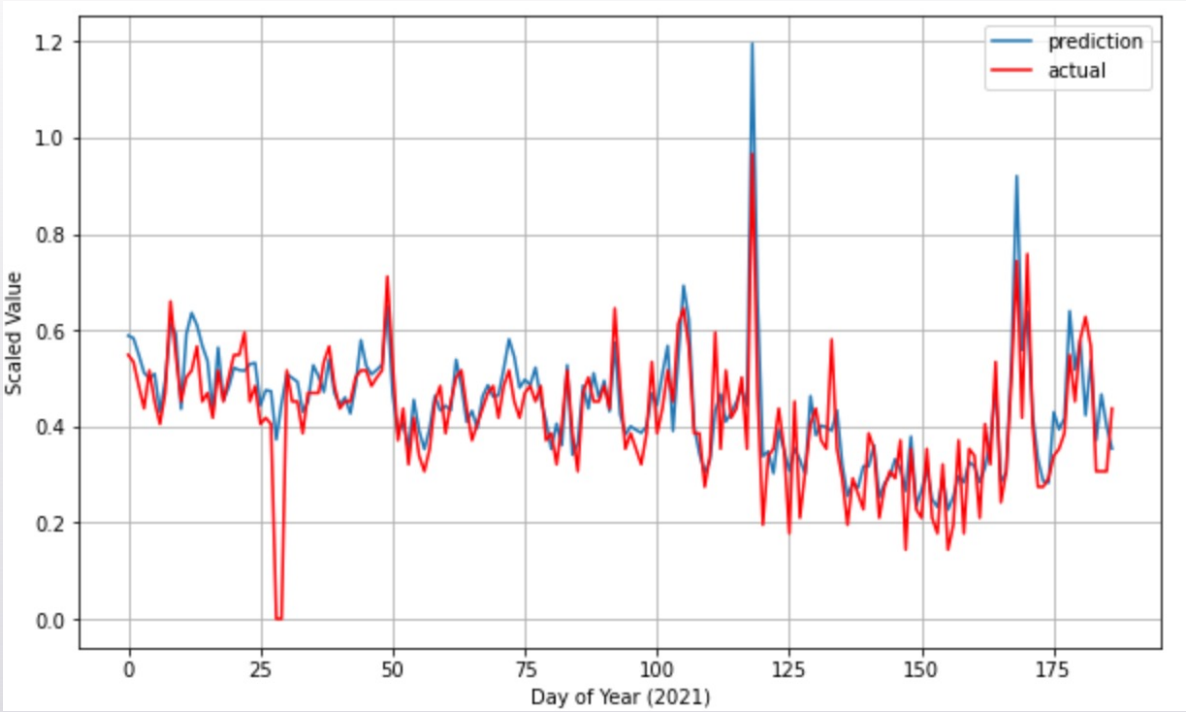
- The predictions were very good for NYC
- Normalized Root Mean Squared was: 0.1095



Scaled Predicted vs Actual for the test period (365 days of year 2021)

LSTM Machine Learning – San Francisco Results

- The model also performed well for San Francisco
- Normalized Root Mean Squared was: 0.0807



Scaled Predicted vs Actual for the test period (6 months of year 2021)

LSTM Summary of Results – Best to Worst

Rank	City	NRMSE
1	San Francisco	0.0807
2	Chicago	0.0847
3	New York City	0.1095

- There is a risk of overfitting. Additional train/tests could be done to reduce the portion of the data used for training versus testing.
- The portion used in this case (4 of the 5 years for training and 1 year for testing) was consistent with that used for ARIMA for comparative purposes. LSTM performed far better than ARIMA.
- Note that LSTM as a multivariate analysis was able to utilize many features for training (primarily weather related) that seem to have contributed to the higher performance of the LSTM model over the ARIMA model.

Conclusions

What factors impact usage and growth of micromobility?

- Past usage can be used to predict future usage.
- Weather conditions are correlated with usage. Generally, warmer and nicer weather results in higher usage.
- Unemployment rates do not appear to significantly impact usage.

Can Machine Learning be used to predict growth or usage patterns?

- Yes, application of SARIMA and LSTM have demonstrated that ML can be used effectively to predict usage or growth of micromobility.

References

Data:

- Chicago Bikeshare: https://ride.divvybikes.com/system-data
- NYC Bikeshare: https://ride.citibikenyc.com/system-data
- San Francisco Bikeshare: https://www.lyft.com/bikes/bay-wheels/system-data
- BLS Unemployment Data: https://www.bls.gov
- NOAA Weather Data: https://www.noaa.gov

General:

- Market background: https://www.alliedmarketresearch.com/micro-mobility-market-A11372#:~:text=The%20global%20micromobility%20market%20was,17.4%25%20from%202021%20to%202030

Data Models:

- J. Du, Q. Liu, K. Chen and J. Wang, "Forecasting stock prices in two ways based on LSTM neural network," 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019, pp. 1083-1086, doi: 10.1109/ITNEC.2019.8729026.
- Meyler, Aidan, Geoff Kenny, and Terry Quinn. "Forecasting Irish inflation using ARIMA models." (1998): 1-48.
- https://towardsdatascience.com/understanding-the-seasonal-order-of-the-sarima-model-ebef613e40fa
- https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/