Improving NYC Rideshare Prediction Accuracy Through Automation

WGU Capstone

Javier Lopez
Student ID: 000697446
BSDMDA

Table of Contents

Project Overview

A. Summary of Elements

This project aimed to answer multiple research questions regarding rideshare demand in New York City (NYC) before, during, and after the beginning of the Covid-19 pandemic. Our research questions are:

- Which rideshare company had the most growth in NYC since the beginning of the Covid-19 pandemic?
- What is the relationship between weather, Covid-19 cases, hospitalizations, deaths, vaccinations, and rideshare demand?
- Which predictive model and exogenous variables have the most impact on rideshare demand and forecasting?
- What is the forecasted growth of rideshare demand for each company?

We aimed to answer these questions by exploring rideshare data and multiple exogenous variables, including weather temperature and precipitation, and Covid-19 cases, hospitalizations, deaths, and vaccinations from February 1, 2019, through March 31, 2022. Despite Juno, Via, Uber, and Lyft operating in NYC throughout the specified period, we limited the scope to the two largest rideshare providers, Uber and Lyft. This limited scope allowed us to model most of the rides in NYC while focusing on half the number of providers. Using negative binomial regression models, we quantified the effects of the abovementioned variables on ride volume for each provider. We tested multiple hypotheses using SARIMAX, Prophet, and XGBoost predictive models to compare the impact of the exogenous variables on ride demand forecasts. Ultimately, we used the best-performing models to forecast rideshare demand for each provider in NYC for the second quarter of 2022.

Project Plan

B. Summary of Execution

We planned this project to answer the four abovementioned questions through data exploration and predictive modeling. The goals and objectives we outlined to answer these questions are as follows:

- Determine which rideshare company, Uber or Lyft, had the highest increase in demand before, during, and after the pandemic's beginning.
- Determine the relationship between temperature, precipitation, Covid-19 cases, hospitalizations, deaths, vaccinations, and rideshare demand.
- Determine what predictive model performed best when forecasting rideshare demand for NYC as a whole.
- Forecast future ride volume for NYC in the second quarter of 2022 using the best-performing predictive model.

This project followed the Waterfall method: requirements, design, implementation, verification, deployment, and maintenance. Because the deliverable for this project was not an application, we did not have to execute the maintenance phase. During the requirements phase, we iterated on our research questions. We then continued to the design phase, determining which data sets and sources were best suited to help us answer our research questions and what predictive models were best to forecast time-series data. We also decided on the project's scope to ensure a desirable result.

During the implementation phase, we gathered our data, the high-volume for-hire trip records and Covid-19 cases, hospitalizations, and deaths from the NYC Open Data portal, the weather data from the National Oceanic and Atmospheric Administration (NOAA), and New York state’s vaccination data from the Centers for Disease Control (CDC) database. We then

conducted our data exploration and analysis on the Jupyter Notebook programming environment using the Python programming language, linear regression and predictive models, and multiple libraries, including Pandas, Numpy, Matplotlib, Seaborn, and Scipy. Following the implementation phase, we tested our predictive model by forecasting ride demand in NYC for the second quarter of 2022. Finally, during the deployment phase, we gathered our results and concisely recorded them to disseminate and present to stakeholders. There were not many differences in the initial project plan we set except for the addition of the deployment stage that we did not initially account for where we recorded all findings.

Our project timeline allotted two weeks, ten business days totaling eighty payroll hours, for completion. We also set a total of seven milestones that we ultimately met. Shown in Table 1 are the milestones and deadlines.

Table 1. Milestones and Deadlines

Milestone	**Projected Start Date**	**Projected End Date**	**Duration**
Establish analytics process requirements	7/18/22	7/18/22	1 Day
Develop predictive model designs and specifications	7/19/22	7/19/22	1 Day
Gather data	7/20/22	7/20/22	1 Day
Explore rideshare data	7/21/22	7/22/22	2 Days
Create and test predictive models	7/25/22	8/05/22	10 Days
Predict future demand	8/08/22	8/10/22	3 Days
Report findings	8/11/22	8/12/22	2 Days

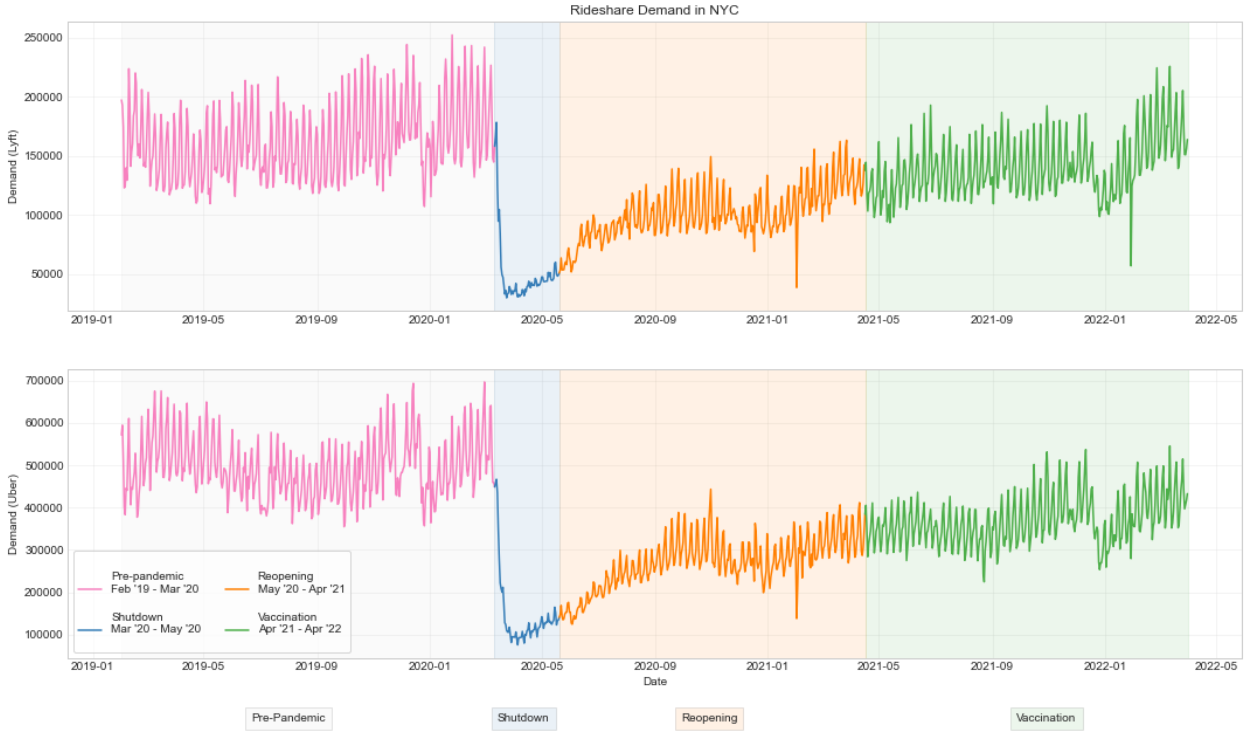
Methodology

C. Data Selection and Collection Process

When selecting our data, we retrieved the high-volume for-hire (rideshare) trip records for NYC for the study period of February 1, 2019, to March 31, 2022, from the NYC Open Data portal. This data set is available as individual monthly files containing historical trip records served by Uber, Lyft, Juno, and Via. The data set contains multiple variables for each observation, including pickup and dropoff times, locations, and financial variables.

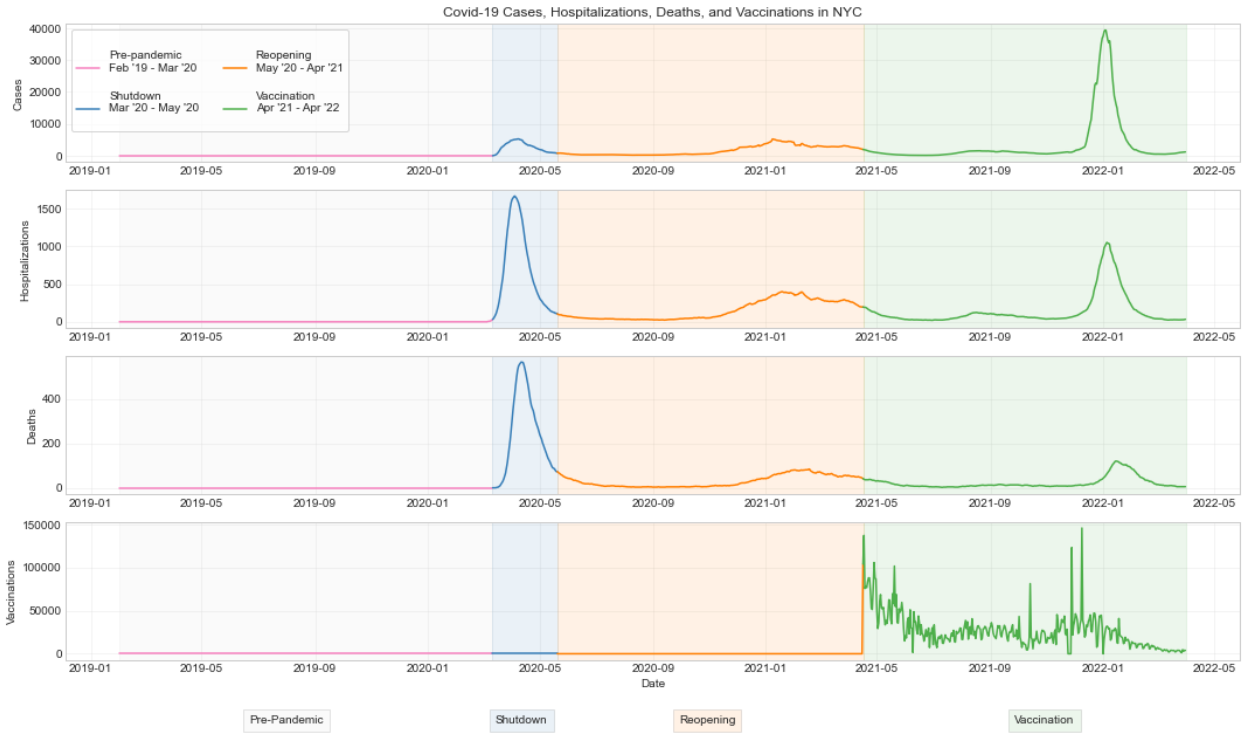
Using Pandas, we extracted rides from Uber and Lyft from the data set, as they represent most of the daily rides conducted in the city. From those records, we used the pickup location to identify the borough each trip originated from and the pickup time variable to aggregate the daily rides in each borough for each rideshare provider, as shown in Figure 1.

Figure 1. Rideshare Demand in NYC



We selected weather data for the Central Park weather station from NOAA that includes temperature (°F) and precipitation (in.). As one of the objectives of this project is to determine the relationship between weather, Covid-19 cases, hospitalizations, deaths, vaccinations, and rideshare demand; we incorporated Covid-19 variables in two ways: (1) temporal phases of the Covid-19 pandemic and (2) number of daily Covid-19 cases, hospitalizations, deaths, and vaccinations (shown in Figure 2).

Figure 2. Covid-19 Cases, Hospitalizations, Deaths, and Vaccinations in NYC



Using Pandas and Python, we created dummy variables for these temporal phases to ensure that our models consider them. Shown in Table 2 are the temporal phases of our study. These key dates caused a significant shift in how society operated during that time. Thus, they will play an essential role in our modeling and were not something we originally accounted for in our plan. However, something we saw would add value to our predictive models and analysis during the data cleaning.

Table 2. Temporal Phases of the Covid-19 Pandemic

Temporal Phase	Date	Event
Phase 0	Feb. 1, 2019 - Mar. 10, 2020	Pre-pandemic
Phase 1	Mar. 11, 2020 - May 19, 2020	Stay at home shutdown
Phase 2	May 20, 2020 - Apr. 15, 2021	States reopening
Phase 3	Apr. 16, 2021 - Mar. 31, 2022	Vaccines administered

One of the most significant obstacles we had to overcome when collecting our data was the nuanced format of the governmental agencies' databases. Finding the correct data was very time-consuming and significantly impacted our timeline. However, once all data was collected, we were delighted to see that it was very well-formatted, making data cleaning a smooth process.

C1. Advantages and Limitations of the Data

As we touched on earlier, the most significant advantage of the data set we used was the lack of null values. This lack of null values saved us much time of not having to impute any data and allowed us to ensure that all data was accurate. We were also very fortunate to be working with data sets that already had the rideshare trip records and Covid-19 data broken down by borough. One of the most significant obstacles we had to overcome when cleaning the data was handling the thirty-eight individual files of historical trip records from February 2019 to March 2022. As these files combined totaled over 15GB of data and would be too time-intensive to clean at once, we decided to condense the files into two individual ones: (1) pre-pandemic trip records and (2) trip records for pandemic phases one through three.

Table 3. Summary of Variables

Variable	Mean	Type
Rides		
Daily Lyft rides	131,525	Continuous
Daily Uber rides	371,898	Continuous
Covid-19 cases, hospitalizations, deaths		
New cases (7-day rolling average)	2,582	Continuous
New hospitalizations (7-day rolling average)	195	Continuous
New deaths (7-day rolling average)	45	Continuous
Covid-19 vaccinations		
Vaccines administered (7-day rolling average)	26,712	Continuous
Covid-19 pandemic temporal phases		
Phase 0: Pre-pandemic		Dummy
Phase 1: Stay at home shutdown		Dummy
Phase 2: States reopening		Dummy
Phase 3: Vaccines administered		Dummy

D. Data Extraction and Preparation Processes

Once we downloaded all of the data, we began our data cleaning and preparation processes. The data cleaning process involved handling many files as the rideshare data is only available as an aggregation of historical trip records by month. To load these files into Jupyter Notebook's Python environment, we used the OS library to walk through the data directories and Pandas to load in the parquet format files. We then loaded one of the files and explored the data frame to ensure no missing values. As there were no missing values, we moved on to prepare our data by standardizing column names to lowercase names with underscores for spaces and indexing each ride observation by its pickup timestamp. The preparation of the rideshare data consisted of a two-part process where we split our data by the pickup taxi zone location and their respective boroughs.

When preparing our data by pickup taxi zone, we kept only two relevant columns from the data frame, the rideshare company's license number and the pickup location for each ride. Using the license number column, we split the data frame into two individual data frames, the license number for rides provided by Uber (HV0003) and the license number for rides provided by Lyft (HV0005). To account for memory usage, we deleted the original data frame from memory before we began working on our rideshare provider data frames. We then categorized our data by creating dummy variables for the pickup locations. As these locations were integer data types, we converted them to string data types and prefaced them with 'zone_' once they became column names. Three taxi zones did not belong to a borough, and thus we removed those from our variables.

Once we indexed the data frame by timestamps and each observation belonged to one individual variable, we resampled the data by day and aggregated the ride counts. As all thirty-nine parquet files were of identical format, we created a program using the three functions we created and had it loop through each file. This data preparation process allowed us to aggregate each taxi zone variable by the borough for our analysis. We achieved this by using the taxi zones dictionary provided by the NYC Open Data portal that lists each taxi zone and its respective borough.

As one of our research questions asks about the relationship between weather, Covid-19 cases, hospitalizations, deaths, vaccinations, and rideshare demand, we moved on to prepare our remaining data sets. The NOAA weather data set was already in a straightforward format but only gave us the minimum temperature and maximum temperature for the day. We calculated the daily average temperature and dropped the other temperature columns from the data frame to arrive at one single temperature value. We then stored these values in a parquet file containing the historical temperature and precipitation data.

Preparing our covid data was simple as we only had to adjust the column names and select the columns we needed for our analysis and models. The same was the case for the vaccination data gathered from the CDC. Once we finished cleaning all the data sets, we tested the data stationarity to ensure it was ready for our models. We further tested the data by plotting each set's autocorrelation, partial autocorrelation, and seasonal decomposition. After all the testing was complete, we aggregated all of our data sets into one data frame by borough and stored it for later use on our predictive models.

E. Data Analysis Process

E1. Methods

When analyzing the data, we first looked into the five number summary statistics for the two rideshare companies in NYC. We then split the data into temporal phases, allowing us to see each company's mean daily rideshare demand using Python's Describe function and assess which had the highest increases from one phase to the next. The split showed us that Uber had more ride demand on average in Manhattan than Lyft had in all of NYC during Phase 0, and they were very close to achieving the same benchmark by Phase 3.

Table 4. Mean Daily Rideshare Demand in NYC and boroughs

	Phase 0: Pre-Pandemic	**Phase 1: Shutdown**	**Phase 2: Reopening**	**Phase 3: Vaccination**	**Overall Demand**
Lyft	162,514	50,058	92,472	114,854	131,525
Uber	496,266	144,887	248,692	304,995	371,898

We used the mean daily rideshare demand values shown in Table 4 to calculate the growth percentage for each company through the temporal phases. Through exploratory data analysis and descriptive statistics, we arrived at the data that gave us the insight needed to answer our first research question. We were then able to visualize the data using Matplotlib's plot function for each pandemic phase. We carried through the use of descriptive statistics into

research question two, where we explored the relationships between Lyft and Uber rideshare demand and the weather, covid cases, and vaccinations.

Before plotting our variables into a scatterplot, we calculated the Pearson correlation coefficient, P-value, and linear regression slope for each variable regarding rideshare demand. We then normalized the ride counts for Lyft and Uber by calculating the Z-score and plotted each variable to rideshare rides in a scatter plot to visualize their correlation. We then moved on to our third objective, to determine what combination of the predictive model and exogenous variables performed best when forecasting rideshare demand in NYC. We began testing the three predictive algorithms with just the Lyft records and logged the mean absolute error and root mean square error for each model in a data frame. These two metrics will help us determine what model has the lowest error rate.

We first tested the SARIMAX algorithm using auto-ARIMA to find the p, d, and q variables, followed by Facebook's Prophet algorithm and the XGBoost algorithm. After logging the metrics for these models using endogenous data, we determined the different models needed to compare our exogenous variables best. Shown in Table 5 are the combinations of exogenous variables we used to model each predictive algorithm. We then compared the metrics for each model and plotted them in a bar graph.

We used the best-performing model to forecast the exogenous variables for the second quarter of 2022. This process consisted of forecasting the covid cases, hospitalizations, deaths, temperature, and precipitation. We then used the exogenous figures to forecast the Lyft and Uber ride demand in NYC. Finally, we explored the forecasted data for each rideshare provider to determine the estimated growth for the forecasted timeframe.

Table 5. Predictive Models Tested

		Weather, Temperature	Weather, Precipitation	Covid-19, Cases	Covid-19, Hospitalized	Covid-19, Deaths	Covid-19, Vaccinations
Lyft	**Model 1**	X					
	Model 2		X				
	Model 3			X			
	Model 4				X		
	Model 5					X	
	Model 6						X
	Model 7					X	X
	Model 8	X					X
	Model 9	X				X	X
Uber	**Model 1**	X					
	Model 2		X				
	Model 3			X			
	Model 4				X		
	Model 5					X	
	Model 6						X
	Model 7	X	X				
	Model 8	X					X
	Model 9	X	X				X

E2. Advantages and Limitations

The Python libraries were the most valuable tools utilized for our project. We used many Python libraries to explore, model, and analyze our data throughout our project. These libraries include Pandas, Numpy, Matplotlib, Seaborn, SKLearn, StatsModels, SciPy, PMDARIMA, Prophet, and XGBoost. Pandas and Numpy allowed us to manipulate our data and view it in a formatted manner, while Matplotlib and Seaborn allowed us to visualize the data. We used the

StatsModels library to decompose trend and seasonality from our data. The same library provided the function necessary for the Augmented Dickey-Fuller test to check stationarity.

Regarding modeling our data, SKLearn provided us with functions to calculate the mean absolute error and root mean square error of our predictive models. The StatsModels, PMDARIMA, Prophet, and XGBoost libraries provided the necessary algorithms to model our data. The most profound limitation in our project was using XGBoost as a candidate predictive algorithm as it is a gradient boosting algorithm. Gradient boosting algorithms are more efficient than their predecessor ARIMA algorithms; however, they are not optimal for forecasting time series data. We scaled our data to adapt it for the XGBoost models, confirming our previous statement once we compared the RMSE and MAPE for each algorithm.

E3. Analytical Method Application

When analyzing the correlation between rideshare demand and the exogenous variables, we began by calculating the Pearson correlation coefficient and the linear regression slope. These two calculations measured the linear relationship between rideshare demand and each variable. Our null hypothesis was that there was no correlation between the rideshare demand and the variable.

The correlation coefficient can vary from a negative one to a positive one, with zero implying no correlation. A correlation of one or a negative one implies an exact linear relationship. A positive relationship implies that as x increases, so does y. In contrast, a negative linear relationship implies that as x increases, y will decrease. The results of the Pearson correlation coefficients, p-values, and linear regression slope, as shown in Figure 3 above, were in line with our observations once we visualized the correlation between each variable in a scatter plot.

Table 6. Pearson Correlation Coefficient and P-value

		Weather, Temperature	Weather, Precipitation	Covid-19, Cases	Covid-19, Hospitalizations	Covid-19, Deaths	Covid-19, Vaccinations
Lyft	**r-value**	-0.13	0.02	-0.17	-0.48	-0.50	0.05
	p-value	0.00	0.49	0.00	0.00	0.00	0.11
Uber	**r-value**	-0.17	0.02	-0.02	-0.50	-0.51	-0.02
	p-value	0.00	0.58	0.00	0.00	0.00	0.44
Rideshare Overall	**r-value**	-0.17	0.02	-0.20	-0.50	-0.51	-0.13
	p-value	0.00	0.55	0.00	0.00	0.00	0.01

We then tested the multiple predictive algorithms and measured their performance by calculating the mean of the root mean square error (RMSE) and the mean absolute percentage error (MAPE). These metrics told us how close each model's observed and forecasted figures were to each other. To ensure we selected the model with the lowest error rate, we weighted RMSE more than MAPE when analyzing performance metrics. RMSE penalizes more significant errors, while MAPE penalizes all errors equally. In the cases of predicting rideshare volume, more significant errors affect forecasts more than minor errors do. Once we forecasted rideshare demand for the second quarter of 2022, we tested the statistical significance of the forecasts using a z-test to get the z-score and p-value between the observed mean and the forecasted mean.

Results

F. Evaluate Success

F1. Statistical Significance

The Pearson coefficient that gave us how correlated each variable is to rideshare demand was calculated using SciPy's regress function. The correlation coefficient R helped us understand how strong the relationship between each variable was, as shown in Table 6. The results showed a negative correlation between overall rideshare demand and the weather temperature, Covid-19 cases, hospitalizations, deaths, and vaccinations. Covid-19 hospitalizations and deaths had the strongest negative correlation at -0.50 and -0.51, respectively, with p-values at precisely zero. Precipitation was the only variable with a very weak positive relationship with overall rideshare demand with an R-value of 0.02. However, the p-value of 0.55 rendered this relationship statistically insignificant.

When looking at the relationship of each variable with the individual rideshare providers, the only difference was on Covid-19 vaccinations, where Lyft had a weak positive relationship with an R-value of 0.05 and Uber had a weak negative relationship of -0.02. Both results for the Covid-19 vaccination variable were below the 95% confidence threshold. Therefore, we considered them to be statistically insignificant as well. However, because R is symmetric, it does not explain which variable affects the other. Therefore, we plotted each variable's regression line to understand their relationship with the overall rideshare demand in NYC, as shown in Figure 4.

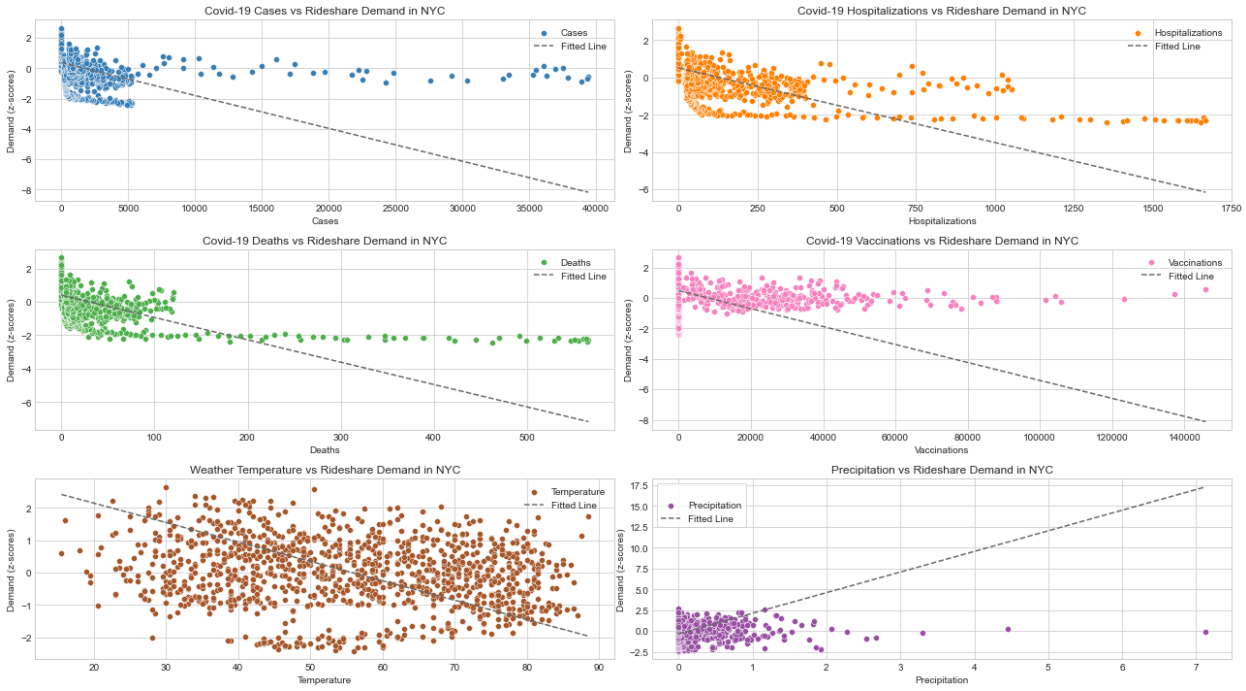
In the Pearson Correlation Coefficient formula:

- r is the correlation coefficient
- n is the sample size

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

- x is the rideshare demand
- y is the exogenous variable
- "∑xy" is the sum of the products of each value in x, and y
- "∑x" is the sum of x
- "∑y" is the sum of y
- "$\sum x^2$" is the sum of the squared x values
- "$\sum y^2$" is the sum of the squared y values

Figure 4. Correlation Between Rideshare Demand and Covid-19 Cases, Hospitalizations, and Deaths in NY



When testing the quality of our model's forecast, we used the SKLearn library to calculate the RMSE and MAPE, and the StatsModels library to calculate the z-score and p-value. We based our model performance on the RMSE to ensure that we selected the model with the smallest error while considering the MAPE to ensure that the selected model had one of the fewest errors. We then used the z-score and the p-value to test our null hypothesis that

the sample's mean was identical to the population's mean. Since we were forecasting rideshare demand, the size of the errors for the model needed to be minimal to ensure that providers could meet demand. We tested each model with the Lyft and Uber data sets. We found that Facebook's Prophet algorithm rendered the lowest RMSE scores with score variance between models similar to the SARIMAX models. While XGBoost returned RMSE scores with the least variance between models, it did not achieve the best RMSE or MAPE scores. For the complete model performance metrics, reference Table 7.

In the RMSE formula:

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

- RMSE is the root mean square error
- f is the model's forecasted value
- o is the observed value
- The bar above the squared difference between the model's forecasted value and the observed value is the mean.

In the MAPE formula:

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

- M is the mean absolute percentage error
- n is the sample size
- At is the observed value
- Ft is the forecast value
- $\sum | (A_t - F_t) / A_t |$ is the sum of the absolute value of the difference between the observed value and the model's forecasted value divided by the observed value

In the z-test formula:

$$Z = \frac{(\bar{X} - \mu_0)}{s}$$

- Z is the z-score
- X-bar is the sample mean
- u_0 is the population mean

- s is the standard deviation

The results depicted in Figure 7 showed that different models had the best RMSE score for the Lyft and Uber data sets. The Lyft data was best predicted by the Prophet Model 6, using the Covid-19 vaccination data, with the lowest RMSE of 15,790 rides, an MAPE of 8.89%, and the lowest z-score of -2.60 standard deviations below the observed mean. A p-value of 0.0094 gave us added certainty of the statistical significance of the model's performance metrics. The Uber demand was best predicted by the Prophet Model 1, using the weather temperature data, with the lowest RMSE score of 38,032 rides and an MAPE of 7.75%. The z-score showed the test forecast to be 3.98 standard deviations above the observed mean with a p-value of 0.0001.

Table 7. Predictive Model Performance Metrics

		Lyft				Uber			
		RMSE	**MAPE**	**Z-score**	**P-value**	**RMSE**	**MAPE**	**Z-score**	**P-value**
SARIMAX	**Control**	23,911	11.99	5.95	0.0000	73,182	13.99	12.24	0.0000
	Model 1	85,993	50.06	29.41	0.0000	60,777	11.59	9.23	0.0000
	Model 2	24,465	12.23	6.05	0.0000	62,094	12.21	7.03	0.0000
	Model 3	39,889	18.73	0.13	0.8985	175,354	27.25	-1.04	0.2968
	Model 4	23,209	11.68	5.88	0.0000	74,231	14.73	12.20	0.0000
	Model 5	23,846	11.96	5.88	0.0000	71,957	13.81	11.74	0.0000
	Model 6	40,447	22.61	20.80	0.0000	121,530	26.38	28.20	0.0000
Prophet	**Control**	23,774	14.73	-9.74	0.0000	39,542	8.07	4.83	0.0000
	Model 1	22,778	13.96	-9.09	0.0000	38,032	7.75	3.98	0.0001
	Model 2	24,259	15.06	-10.00	0.0000	39,468	7.91	3.39	0.0010
	Model 3	77,357	39.40	-3.10	0.0019	205,198	26.45	3.81	0.0001
	Model 4	26,006	16.16	-10.50	0.0000	47,045	10.34	6.90	0.0000
	Model 5	21,178	12.77	-7.72	0.0000	44,773	9.66	1.05	0.0000
	Model 6	15,790	8.89	-2.60	0.0094	43,337	9.06	5.72	0.0000
XGBoost	**Control**	24,092	12.38	6.64	0.0000	54,353	11.50	5.79	0.2933
	Model 1	24,920	12.66	8.81	0.0000	55,135	11.04	12.20	0.0000

Model 2	24,926	12.65	8.88	0.0000	55,458	11.16	5.86	0.0000
Model 3	24,151	12.32	6.99	0.0000	57,027	11.82	3.73	0.0035
Model 4	24,924	12.59	8.84	0.0000	55,458	11.16	7.57	0.0000
Model 5	24,926	12.65	8.88	0.0000	66,038	13.48	8.07	0.0000
Model 6	24,926	12.65	8.88	0.0000	55,311	11.11	-0.92	0.0000

F2. Practical Significance

Given the unpredictability of forecasting models during the Covid-19 pandemic due to ever-changing external influences on rideshare demand, many data scientists needed to develop new forecasting methods that did not just rely on historical observations. By introducing already known exogenous variables, data scientists could better predict demand. The practical significance of using external variables, such as Covid-19 cases, hospitalizations, deaths, vaccinations, and the weather, could be seen through the Pearson Correlation Coefficient observed for each variable in Table 6. From there, we could see that Covid-19 hospitalizations and deaths significantly negatively correlated to rideshare demand with over 95% certainty. This information could significantly impact how data scientists handle their predictive modeling, not just during the Covid-19 pandemic but during any large-scale event.

F3. Evaluate Effectiveness

Our project set forth the five goals and objectives reiterated below. We used these goals as milestones to answer each of the four research questions.

- Determine which rideshare company, Uber or Lyft, had the highest increase in demand before, during, and after the pandemic's beginning.
- Determine the relationship between temperature, precipitation, Covid-19 cases, hospitalizations, deaths, vaccinations, and rideshare demand.
- Determine what predictive model performed best when forecasting rideshare demand for NYC as a whole.

- Forecast future ride volume for NYC in the second quarter of 2022 using the best-performing predictive model.

To achieve the first objective, we utilized the growth rate formula, the difference between past observations and current observations divided by the past observations multiplied by one hundred to get a percentage. This formula gave us the information needed to answer our first question and meet the first objective. To achieve the second objective, we calculated and plotted the Pearson Correlation Coefficient and p-values to determine the relationship and strength between the two rideshare providers and each variable. We then achieved the third objective by building and testing a predictive model for each variable using the three predictive algorithms. After testing and measuring the effectiveness of the twenty-one models, we compared the performance metrics and chose the best-performing model to forecast demand for the second quarter of 2022. Once we forecasted the demand for Lyft and Uber, we compared them to determine which rideshare provider was forecasted to have the highest growth. After meeting all four objectives to arrive at an answer for each research question, we deemed our project successful.

G. Key Takeaways

G1. Conclusions

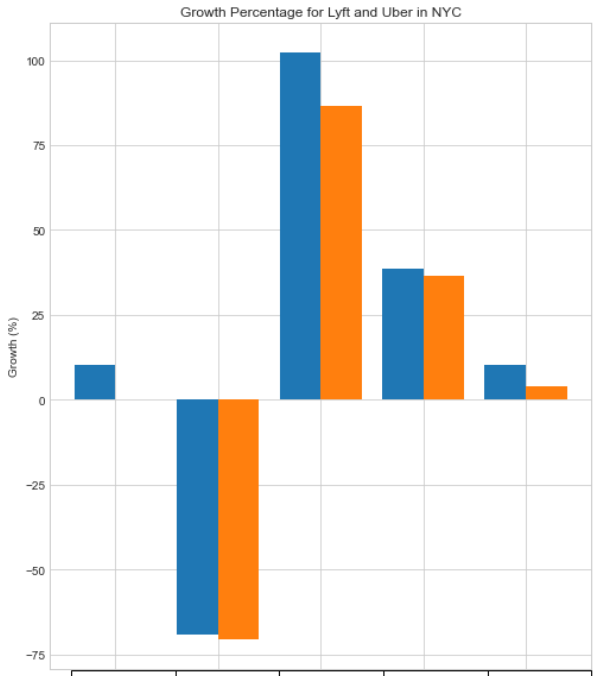
Our project sought to answer the four key research questions reiterated below. Through the use of descriptive and predictive analytics, we were able to answer these questions.

- Which rideshare company had the most growth in NYC since the beginning of the Covid-19 pandemic?
- What is the relationship between weather, Covid-19 cases, hospitalizations, deaths, vaccinations, and rideshare demand?
- Which predictive model and exogenous variables have the most impact on rideshare demand and forecasting?

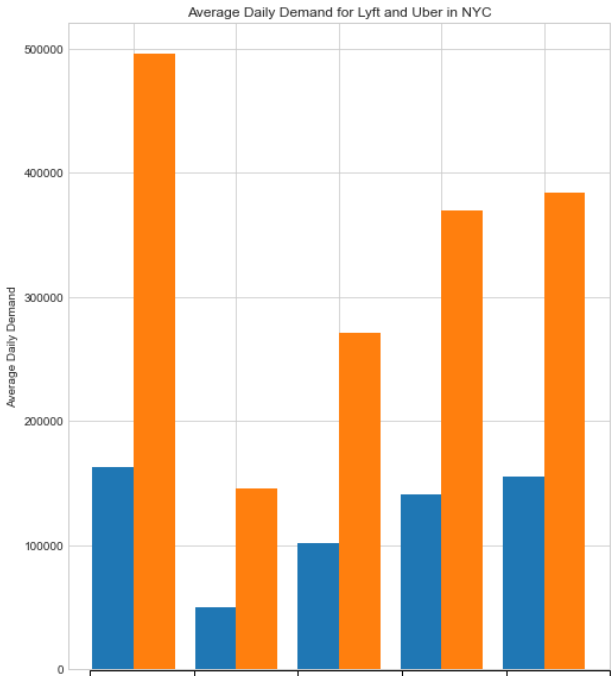
- What is the forecasted growth of rideshare demand for each company?

We approached our first research question by dividing our data into four key phases: pre-pandemic, Covid-19 shutdown, state reopening, and vaccination. During these phases, many businesses were affected due to stay-at-home orders that made it nearly impossible for customers to get to them. Rideshare was no exception; thus, we deemed it essential to split the data. We aggregated the daily rideshare demand for each rideshare provider (Lyft and Uber) during each phase and evaluated their growth. We found that pre-pandemic, Lyft was averaging a 10.29% increase in demand, while Uber was averaging a 0.17% increase.

Figure 5. Growth Percentage and Average Daily Demand for Lyft and Uber in NYC



	Pre-Pandemic	Shutdown	Reopening	Vaccination	Forecast
Lyft	10.29	-69.06	102.26	38.44	10.23
Uber	0.17	-70.68	86.61	36.4	3.86

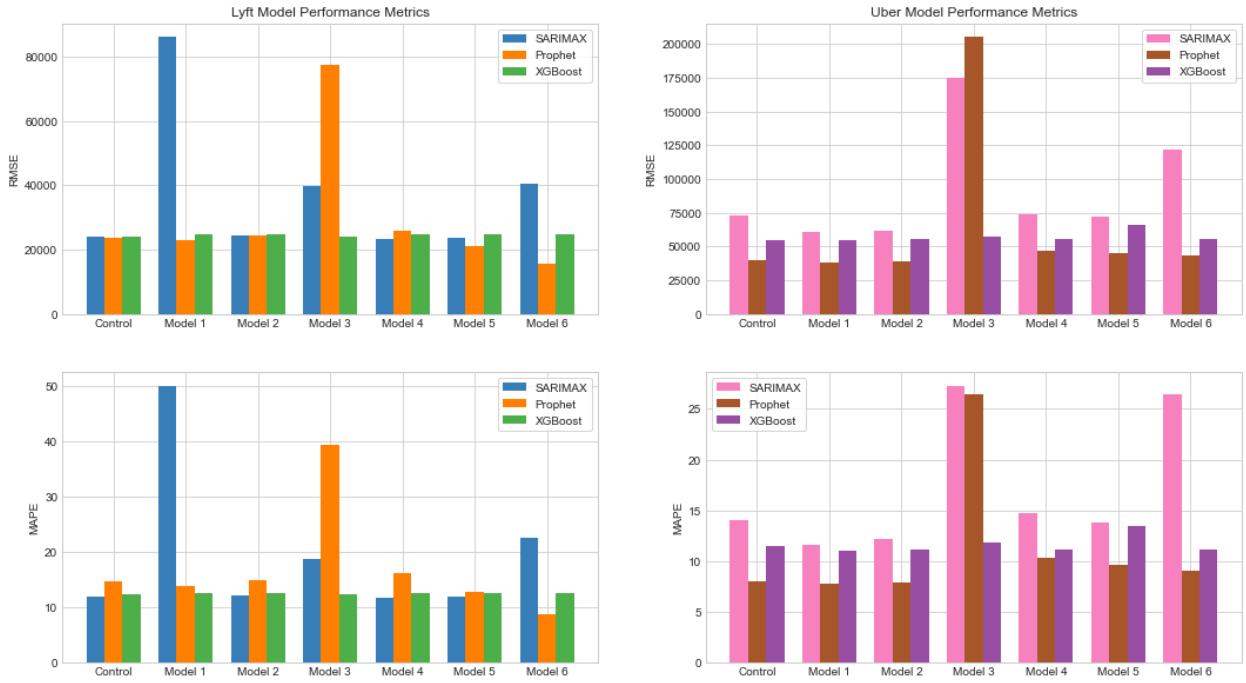


	Pre-Pandemic	Shutdown	Reopening	Vaccination	Forecast
Lyft	162242.84	50201.86	101539.71	140569.78	154949.37
Uber	495462.41	145248.5	271047.79	369715.34	383989.84

During the Covid-19 shutdown, both providers saw significant decreases in demand. Lyft had a 69.19% decrease in demand, while Uber saw a 70.81% decrease. During the state

reopening, rideshare demand for Lyft rose by 102.26%, and demand for Uber rose by 86.61%. The vaccination phase saw a steady increase in demand at 38.44% for Lyft and an increase of 36.40% for Uber. Overall, we found that Lyft had the most growth since the beginning of the Covid-19 pandemic seeing more significant increases in demand and smaller decreases during the Covid-19 shutdown phase, as shown in Figure 5.

Figure 6. Predictive Model Performance Metrics

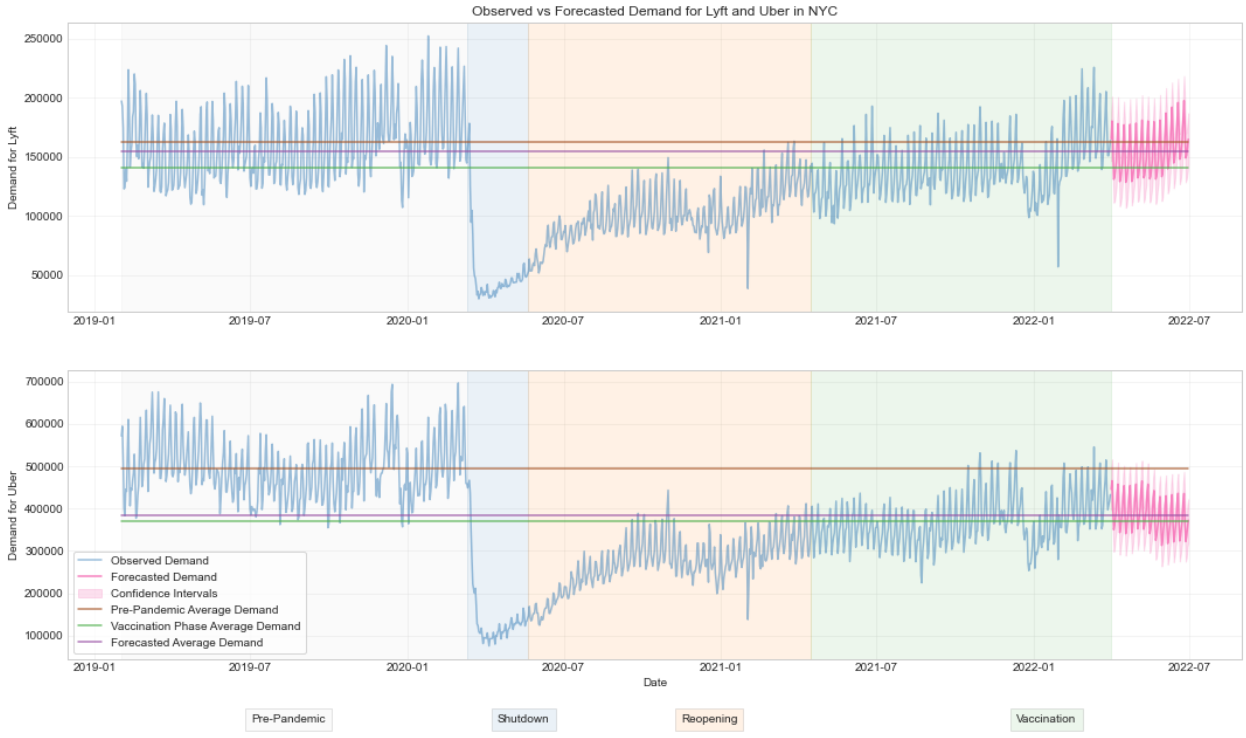


We then tested the relationship between rideshare demand and each variable and found that all but precipitation negatively correlated with rideshare demand in NYC. Upon further testing, we found that the 0.55 p-value did not provide much certainty as there was little to no precipitation in NYC between February 2019 and March 2022. All other p-values, however, provided over 99% certainty of the negative relationships overall. After investigating the relationships, we tested the three predictive algorithms (SARIMAX, Facebook's Prophet, and XGBoost) by modeling each variable individually. The RMSE, MAPE, and z-score for the

forecasts on these models, as shown in Figure 6, showed that the Prophet algorithm performed best overall, but Covid-19 vaccinations were the best predictor for Lyft, while the temperature was the best predictor for Uber.

To answer our final research question, we used Prophet Model 6, Covid-19 vaccinations, to forecast demand for Lyft from April 1, 2022, through July 31, 2022. We repeated the same process for Uber using Prophet Model 1, temperature, to forecast demand for Uber during the same timeframe. The models forecasted a 10.23% increase in demand for Lyft and a 3.86% increase in demand for Uber compared to the previous phase. This forecast represented an increase in average daily demand for Lyft of 14,379 rides and an increase of 14,275 rides for Uber.

Figure 7. Observed vs. Forecasted Demand for Lyft and Uber in NYC



G2. Storytelling Analysis

When deciding on the best methods to represent the data from our analysis, we decided on a combination of tables and charts depending on the type and amount of data we were trying to convey. The first step in our data analysis was to explore and learn as much as possible about our data through the five number summary statistics and by visualizing the data using the Matplotlib and Seaborn libraries. As we were working with multiple time series data sets, we determined that the best way to convey change over time was through a line graph, as shown in the line graphs in Figures 1 and 2. We further broke down our data into different pandemic phases and calculated the average demand for each rideshare provider during each phase, as shown in Table 2 and Table 4.

After exploring our data and extracting as much information about each rideshare provider as possible, we began looking into the relationship between each provider and the external variables. We calculated the Pearson Correlation Coefficient and regression line for each variable compared to Lyft and Uber. Once calculated, we plotted the R-values and regression lines into the scatter plots pictured in Figure 4 and listed them along with the p-values in Table 6. We used a combination of scatter plots and a table to allow for a thorough understanding of the relationship between each variable and their statistical significance.

Once we had the information necessary, we moved on to creating and testing our models. This process involved planning what models would be needed, as shown in Table 5, and what combination of variables we needed for further testing pending the model performance metrics. When comparing model performance, we used a combination of a table and a bar chart, shown in Table 7 and Figure 6, to assist us in visualizing the data from over twenty different models. We then forecasted future demand using the best-performing model and visualized this data in a line graph picture in Figure 7. The completed metrics led to our final comparison of forecasted and observed demand through the pandemic phases.

We determined that the best way to convey our findings was through a bar chart with an adjoining table listing the growth in demand for each provider throughout the different phases as a percentage and average daily rides, as shown in Figure 5. All in all, through the use of tables, bar charts, line graphs, and scatter plots, we could convey the performance of each rideshare provider from pre-pandemic times through the forecasted demand for the second quarter of 2022.

G3. Recommended Courses of Action

Our findings for Lyft and Uber show two different stories from pre-pandemic times through the first quarter of 2022. While our initial data exploration showed a stark difference in demand, with Uber averaging 495,462 daily rides and Lyft averaging about a third of the demand at 162,243 daily rides, we found that Uber's growth was nearly stagnant at just 0.17%. Lyft, however, was experiencing an increase in demand of 10.29% during the same period. The trend continued in a similar direction throughout the pandemic, with Lyft gaining more demand during each phase and Uber trailing behind yet maintaining a higher volume altogether.

Based on these findings, we recommend that Uber follow one of two courses of action to reverse the trend potentially. We are basing our first recommendation on the correlation between temperature and demand. Given the negative correlation, we recommend that Uber increase their marketing spending during the warmer seasons to attract more riders and stay top of mind for the cooler seasons where demand is likely to increase. During the cooler months, customers will choose between the two providers, and staying top of mind is likely to steer them to the first provider that comes to mind. The same recommendation can also be applied to Lyft to continue the upward trend in demand.

We are gearing our second recommendation towards Lyft, where the best-performing model used the Covid-19 vaccination variable. As we move further away from the beginning of

the Covid-19 pandemic, we must take away the importance of forecasting demand using more than just historical endogenous data and incorporating multiple exogenous variables into our models. We recommend that Lyft look further into potential economic variables that may impact the business and test multiple models for NYC and their other markets. Both of these recommendations can be applied equally to Lyft and Uber, as the forecast models show that they are not yet forecasted to reach a pre-pandemic level of demand.

Appendices

---

1. Hawkins, A. J. (2020, March 19). Uber is doing 70 percent fewer trips in cities hit hard by coronavirus. The Verge. Retrieved July 12, 2022, from https://www.theverge.com/2020/3/19/21186865/uber-rides-decline-coronavirus-seattle-sf -la-nyc

2. Henke, N., Puri, A., & Saleh, T. (2022, April 28). Accelerating analytics to navigate COVID-19 and the next normal. McKinsey & Company. Retrieved July 12, 2022, from https://www.mckinsey.com/business-functions/quantumblack/our-insights/accelerating-an alytics-to-navigate-covid-19-and-the-next-normal

3. TLC Trip Record Data. TLC Trip Record Data - TLC. (n.d.). Retrieved June 21, 2022, from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

4. National Centers for Environmental Information (NCEI). (n.d.). Daily summaries station details. Daily Summaries Station Details: NY CITY CENTRAL PARK, NY US, GHCND:USW00094728 | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Retrieved June 21, 2022, from https://www.ncdc.noaa.gov/cdo-web/ datasets/GHCND/stations/GHCND:USW00094728/detail

5. COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths | NYC Open Data. (n.d.). NYC Open Data -. Retrieved July 12, 2022, from https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-a n/rc75-m7u3

6. COVID-19 Vaccinations in the United States,Jurisdiction | Data | Centers for Disease Control and Prevention. (2021, May 24). CDC Data Sets. Retrieved July 12, 2022, from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/un sk-b7fc/data