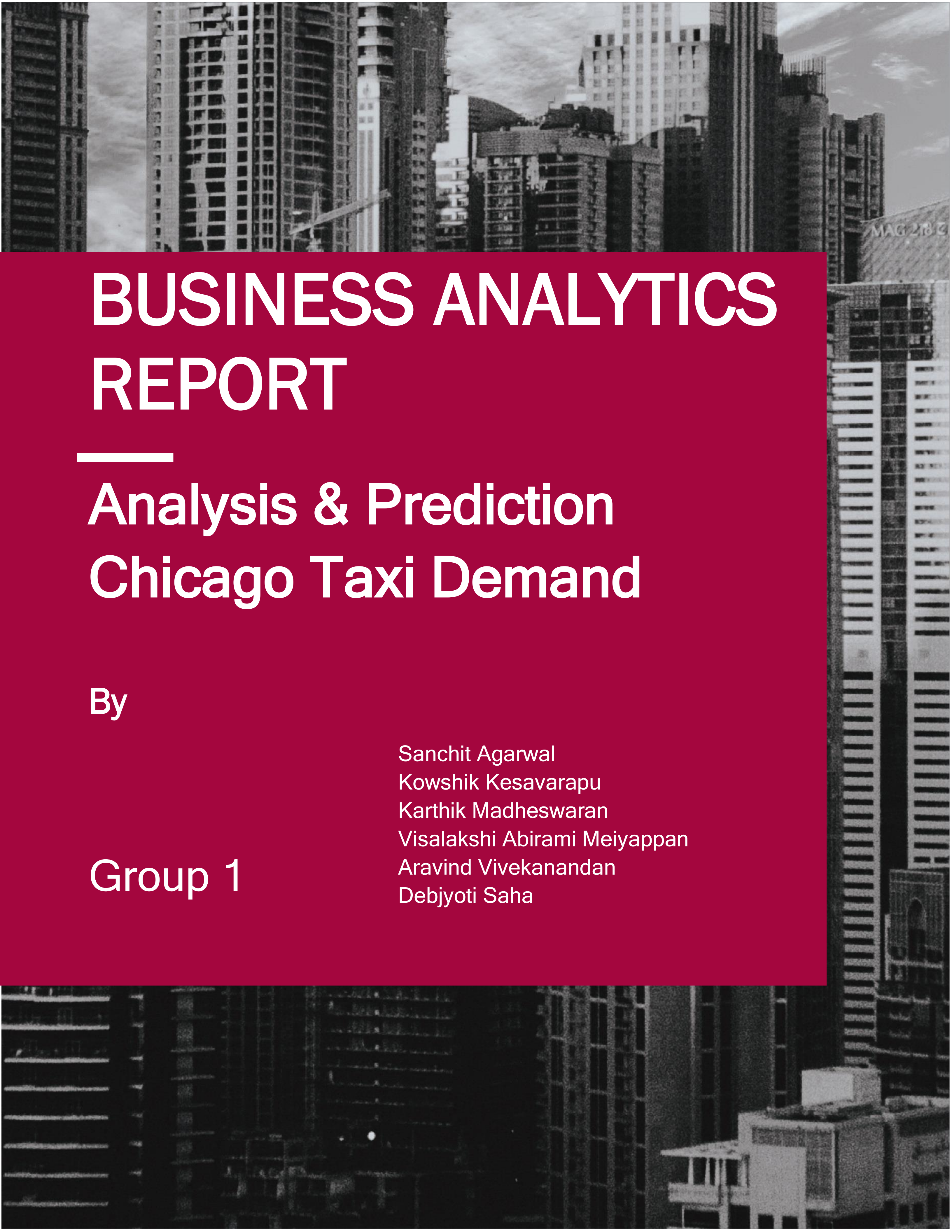


BUSINESS ANALYTICS REPORT

Analysis & Prediction Chicago Taxi Demand

By

Group 1

Sanchit Agarwal
Kowshik Kesavarapu
Karthik Madheswaran
Visalakshi Abirami Meiyappan
Aravind Vivekanandan
Debjyoti Saha

Table of Contents

Overview

In this project, we analyze the taxi industry in Chicago, specifically the taxi trips that took place in the city of Chicago for two consecutive years - 2019 and 2020. For this project, we pulled the data from the Chicago Open Data Portal and used R as our programming language. Using CRISP-DM methodology, we aim to answer the following questions:

1. What were the trends in the Taxi industry before the COVID-19 pandemic, and how did it change a year later?

2. Which Machine Learning algorithm will be best suited for building a model to predict the taxi demand?

Business Overview

The taxi industry in Chicago has a rich history, starting from 1853 when the first taxi company was found to cater to the transportation needs of people using the railway [[1]]. In the 1920s, the two biggest taxi companies in Chicago, Yellow Cab Company and Checker Taxi were engaged in a full-out rivalry which included shooting in broad daylight and the taxi drivers engaging in tank warfare maneuverers to beat the competition [[2]]. In recent years, the prevalence of ride-hailing companies like Uber has started to take over a significant share of the market, with Uber claiming to have profited $46,380,000, and creating 25,000 incremental rides in 2013 [[3]].

The COVID-19 pandemic wreaked havoc on the taxi industry, with the traditional taxi company hit the hardest, since they were already dealing with the blowback from the emergence of rail-hailing industries. Considering the significance of the taxi industry we will be analyzing the taxi trips reported by the cab companies in Chicago. Also, we will be analyzing data from 2019 and 2020, years enveloping the pandemic, to better capture and report trends in the demand for taxis.

By building a reasonably accurate predictive model, we hope that it can be used by the taxi companies to schedule their taxi fleet more efficiently, resulting in reduced passenger waiting time, better utilization of the fleet resources and an increase in incremental rides. We will be building the model using various machine learning algorithms and comparing their performances.

Dataset Overview

The city of Chicago hosts a publicly accessible datastore which contains around 600 datasets containing information on city departments, public services and their performances, for the benefit of researchers. The taxi trip dataset is populated using information captured from the two biggest payment processors in service for the taxi companies. In the dataset, rides from ride-hailing companies such as Uber and Lyft are not being recorded and thus not in the scope of this project. The dataset is periodically updated by the city of Chicago and it contains around 198 million records. However, the observations during 2019 and 2020 account for approximately 20.8 million.

The dataset contains 23 columns, being:

S NO	COLUMN NAMES	DESCRIPTION
1	**Trip ID**	UUID for each trip
2	**Taxi ID**	UUID for each taxi
3	**Trip Start Timestamp**	Trip start time (rounded to the nearest 15 minutes).
4	**Trip End Timestamp**	Trip end time (rounded to the nearest 15 minutes).
5	**Trip Seconds**	The trip duration is in seconds.
6	**Trip Miles**	The total distance covered in miles.
7	**Pickup Census Tract**	The Census Tract from where the trip began.
8	**Dropoff Census Tract**	The Census Tract from where the trip ended.
9	**Pickup Community Area**	The Community Area from where the trip began.
10	**Dropoff Community Area**	The Community Area from where the trip ended.
11	**Fare**	The fare for the trip.
12	**Tips**	The tip for the trip. Cash tips are not recorded.
13	**Tolls**	The tolls for the trip.

14	**Extras**	Any extra charges incurred during the trip.
15	**Trip Total**	Total cost of the trip.
16	**Payment Type**	Type of payment used.
17	**Company**	The company under which the taxi is registered for.
18	**Pickup Centroid Latitude**	The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.
19	**Pickup Centroid Longitude**	The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.
20	**Pickup Centroid Location**	Tuple of pickup geo-coordinates.
21	**Dropoff Centroid Latitude**	The latitude of the center of the drop-off census tract or the community area if the census tract has been hidden for
22	**Dropoff Centroid Longitude**	The longitude of the center of the drop-off census tract or the community area if the census tract has been hidden for privacy.
23	**Dropoff Centroid Location**	Tuple of drop-off geo-coordinates

To maintain privacy and ensure that this public data is not being exploited for malicious use, certain provisions have been taken:

1. The trip start and end timestamps have been rounded to the nearest 15 minutes.
2. The taxi license number is masked using a UUID (Universally Unique Identifier).
3. Census Tracts having less than 3 trips in the relevant 15-minute time slot are not shown.

Other outliers are automatically not updated in the dataset, such as:

1. Trip times less than 0 and more than 86,400 seconds.

2. Trip Miles less than 0 and more than 3500 miles.

3. Trip cost less than $0 and more than $10,000.

Methodology

Our project is broadly divided into two main objectives:

1. Comparing pre-covid and post-covid trends for the Chicago taxi industry using Exploratory Data Analysis.

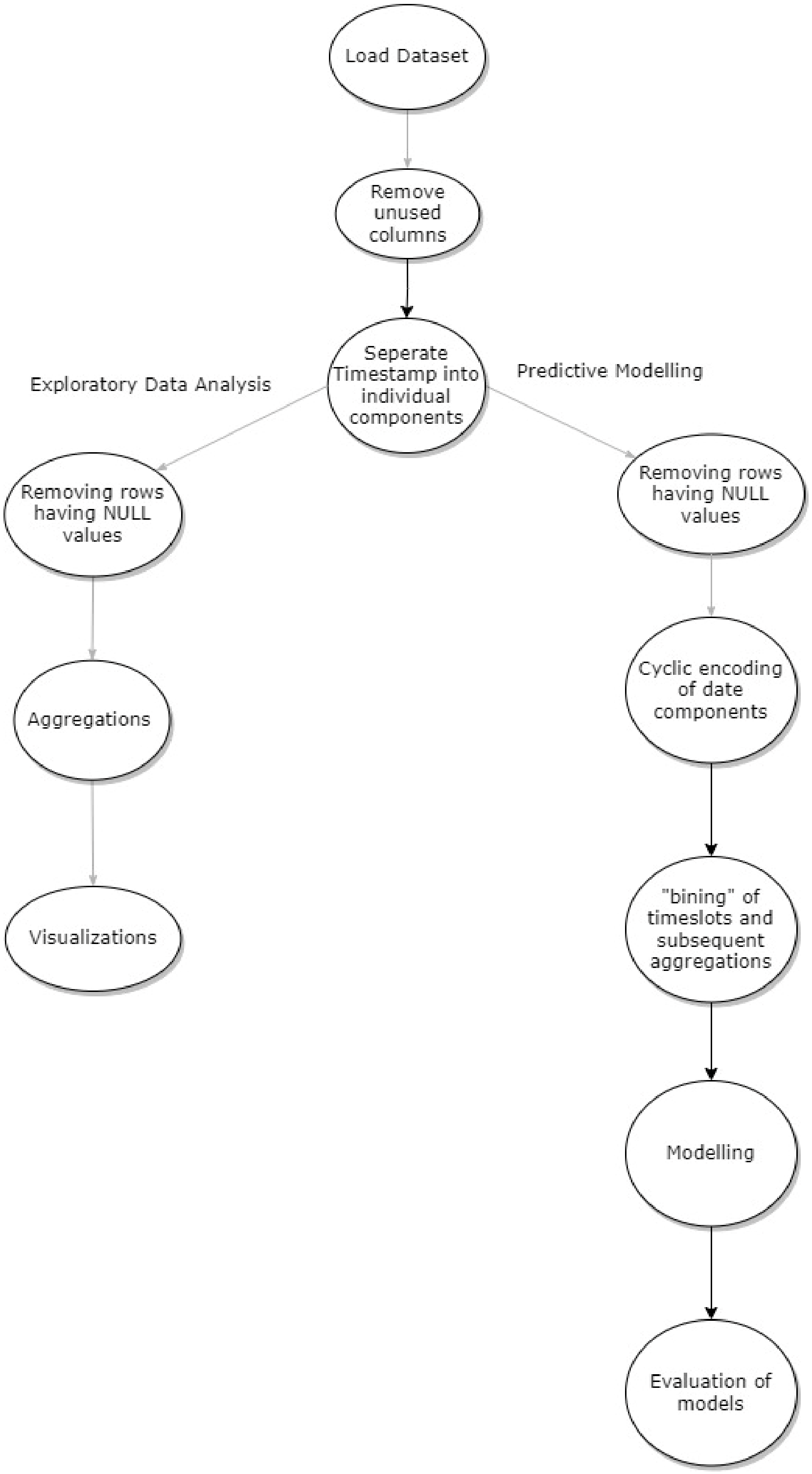
2. Use a set of preselected Machine Learning algorithms to build a model for predicting taxi demand and compare their performance.

For the effective performance of our models, we are only using the data from 2020 so that our models can understand the trends in the covid scenario better.

Since we are processing a large dataset, we ensured that proper error handling is in place, including batch processing of the original dataset.

For the modeling part, we split the pre-processed dataset into 70% training dataset and 30% testing dataset. The dataset splitting was NOT randomized to preserve the trend over time.

The figure below depicts the flowchart of the various processes involved in our project.


Load Dataset
Remove unused columns
Seperate Timestamp into individual components
Exploratory Data Analysis
Predictive Modelling
Removing rows having NULL values
Aggregations
Visualizations
Removing rows having NULL values
Cyclic encoding of date components
"bining" of timeslots and subsequent aggregations
Modelling
Evaluation of models

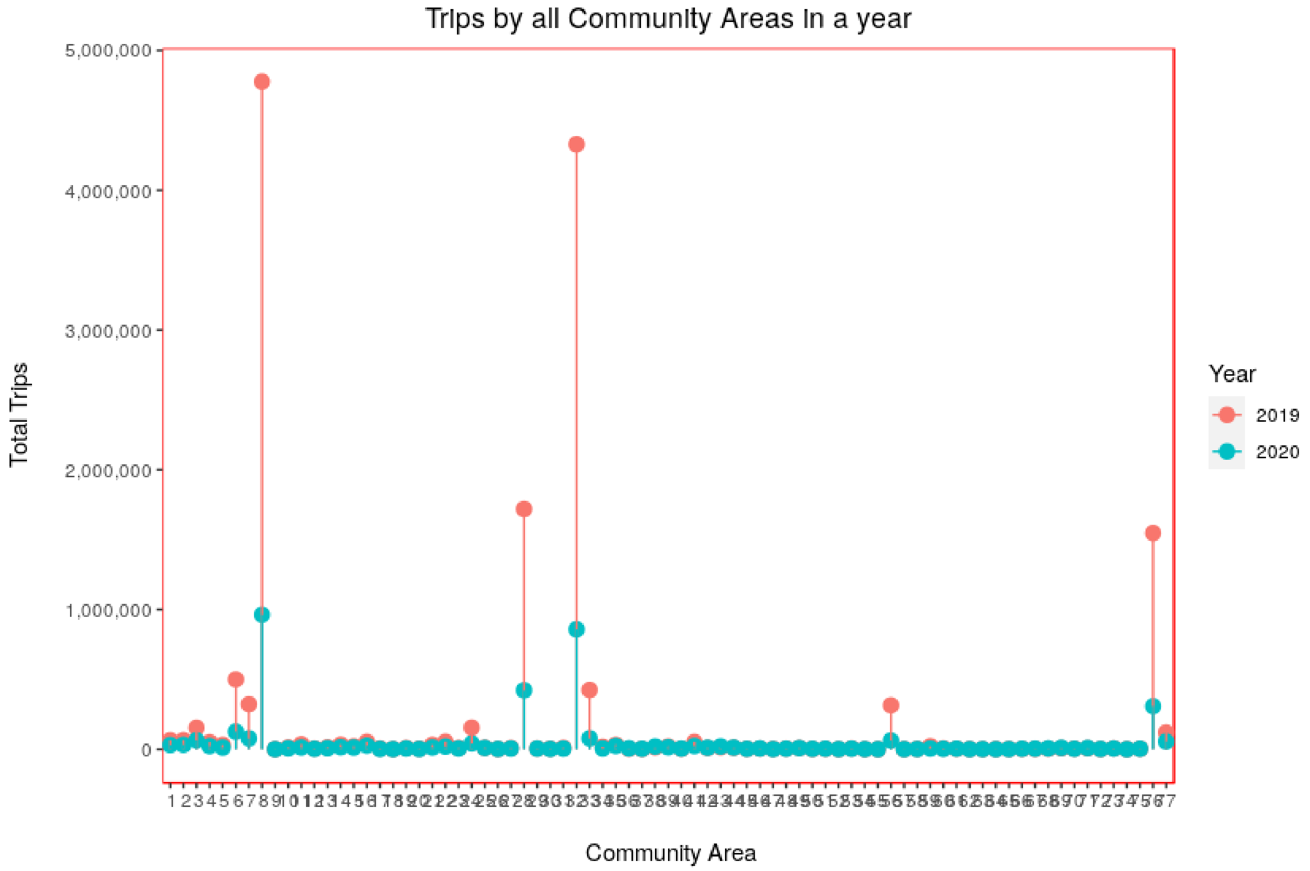
Exploratory Data Analysis

As a first step, we explored the Chicago Taxi dataset and we had our observations recorded. We compared various attributes of the data and analyzed the relationships between them for 2019 as well as 2020.

We have recorded our observations through graphs and inferred the various patterns reflected in the existing data.

A. Taxi Trips by Community Areas in Chicago

The city of Chicago is divided into 77 community areas. The below graph shows the number of taxi trips availed in a community area before and after Covid.

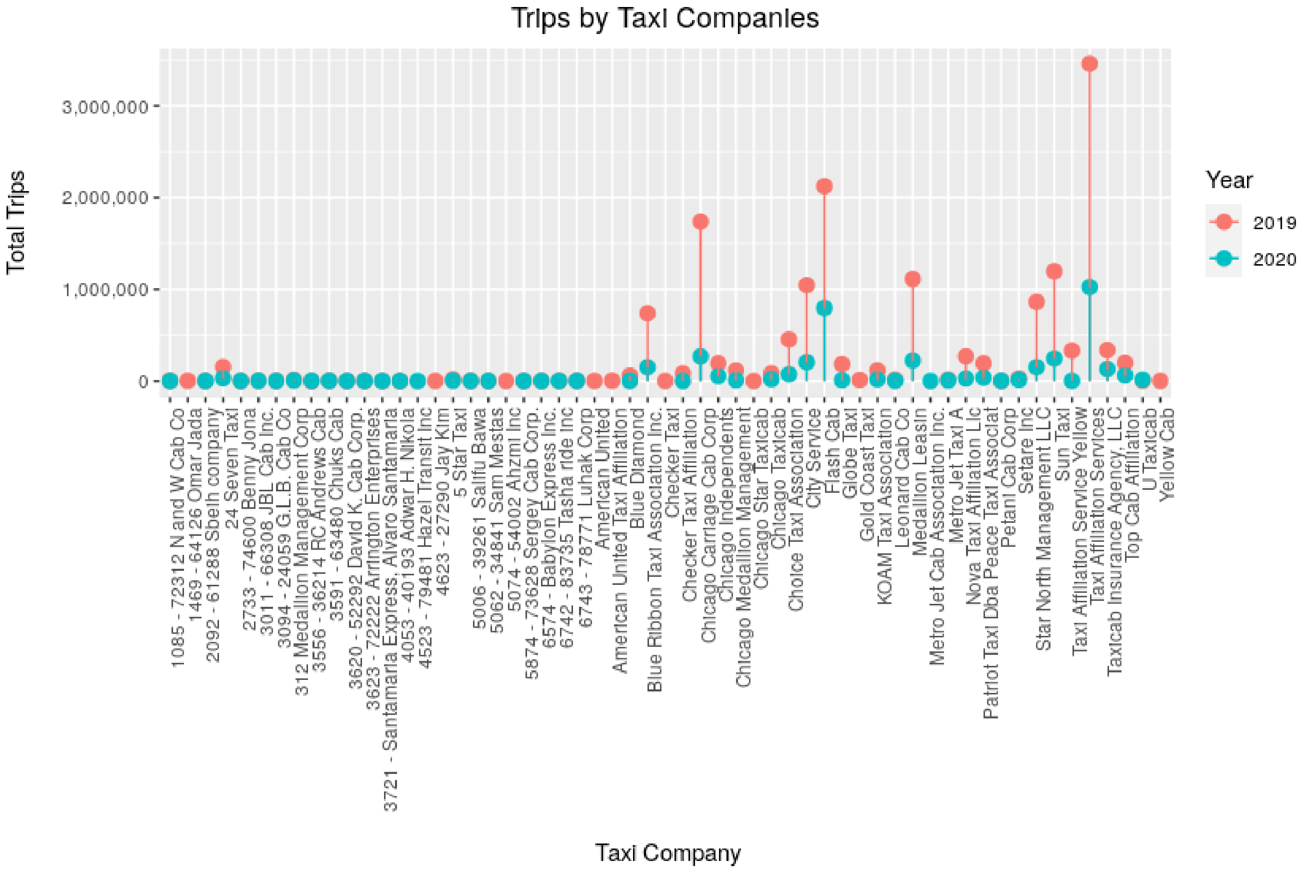


From the data we can infer that the following community areas have the highest number of taxi trips availed before and after the pandemic:

Area #	Community Area	Home to
8	Near North Area	Wrigley Field, Lincoln Park Zoo, Lincoln Park, Chicago History Museum, Boystown, North Avenue Beach, etc.
28	Near West Side	Garfield Park Conservatory, United Centre, National Museum of Mexican Art, Wicker Park, etc.
32	The Loop	The Art Institute of Chicago, Cloud Gate, Willis Tower, Grant Park, Chicago Cultural Centre, Chicago Theatre, etc.
33	Near South Side	Adler Planetarium, Shedd Aquarium, Field Museum, Northerly Island, Glessner House, etc.
76	O'Hare	O'Hare International Airport, Rotunda Tower Garden, Schiller Woods, Skydeck Chicago, Museum of Contemporary Art, etc.

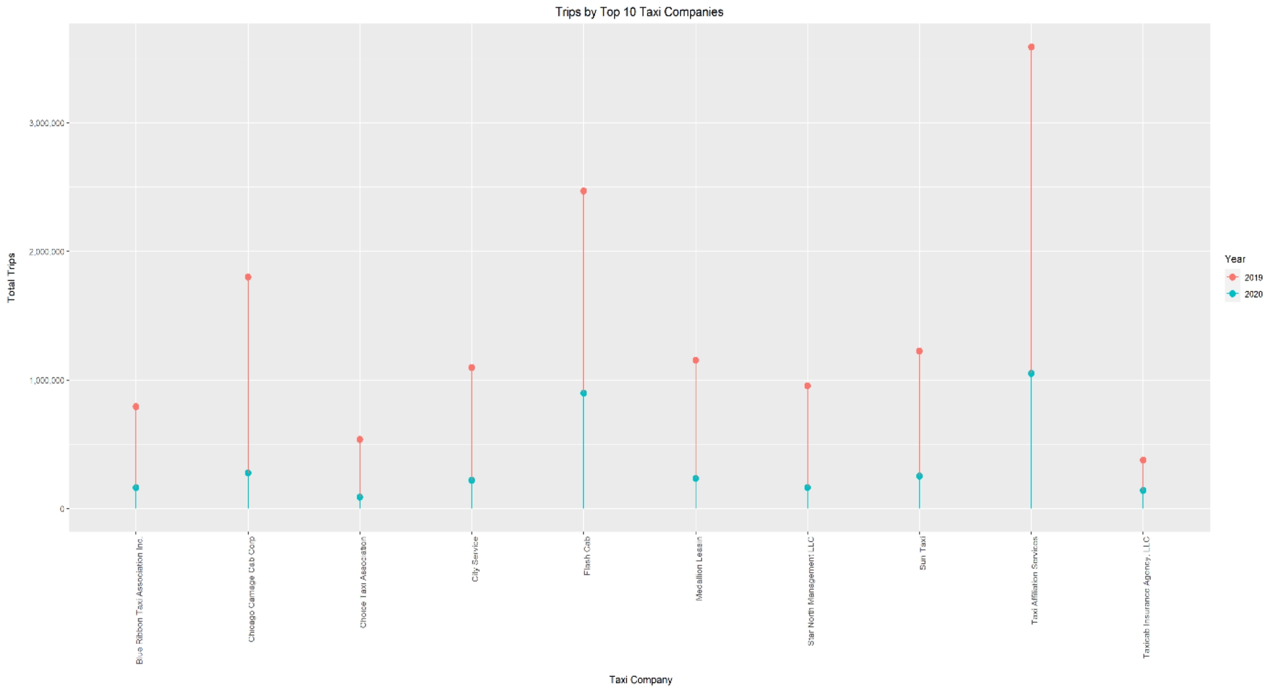
B. Taxi Trips offered by Taxi Companies

The data contains details of the companies of the various taxi trips made in Chicago.



From the above graph, it can be observed that there were more than 50 taxi service providers in the city during these 2 years.

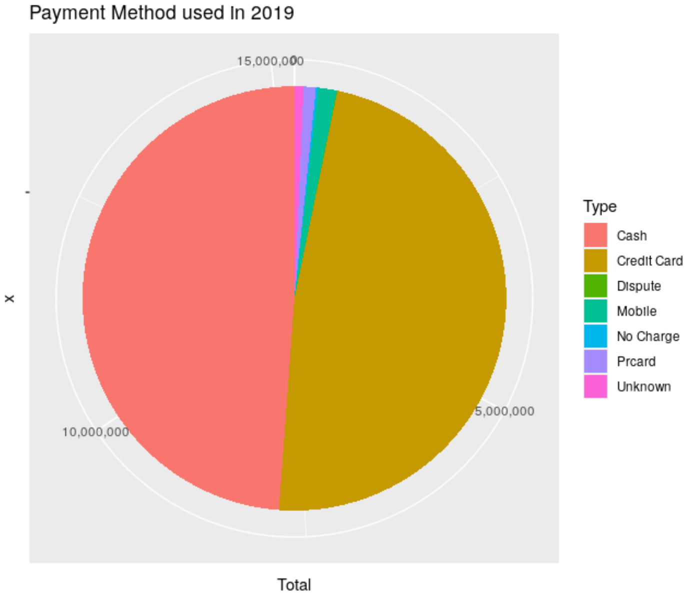
Also, it can be observed that a few taxi companies have very minuscule or no taxi trips in 2019 and a few companies who did not have a taxi trip in 2019 have managed to offer services post-pandemic.

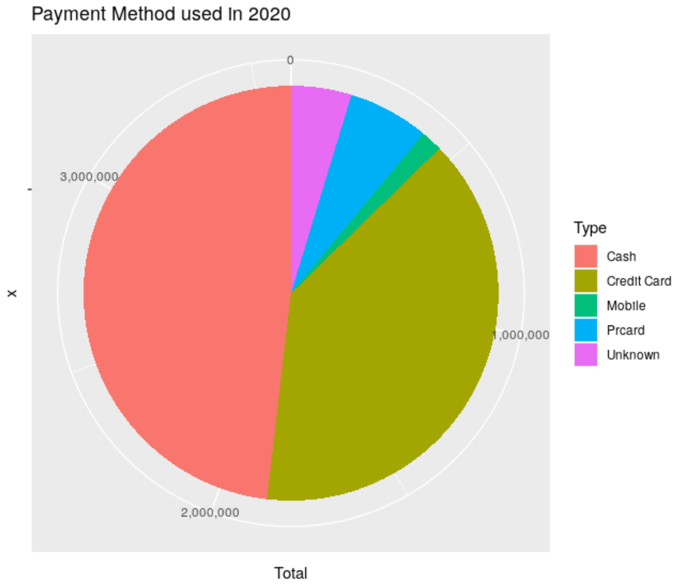


Though the number of taxi trips reduced considerably, the following taxi companies consistently offered the highest number of rides before and after the pandemic.

- Taxi Affiliation Services, LLC
- Flash Cab
- Chicago Carriage Cab Corporation

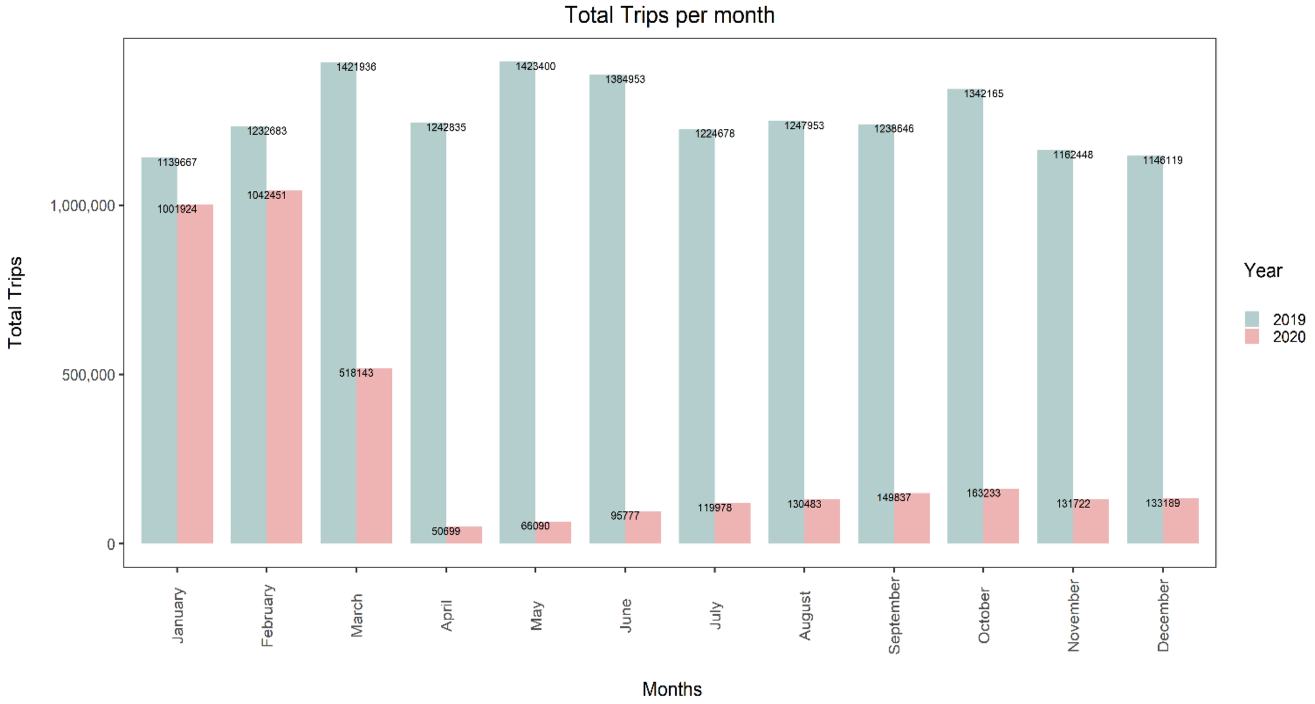
C. Mode of Payment (2019 vs 2020)

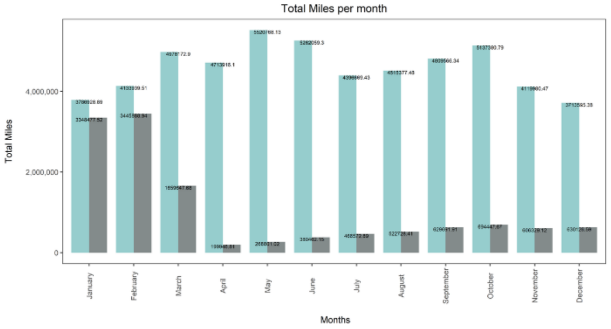


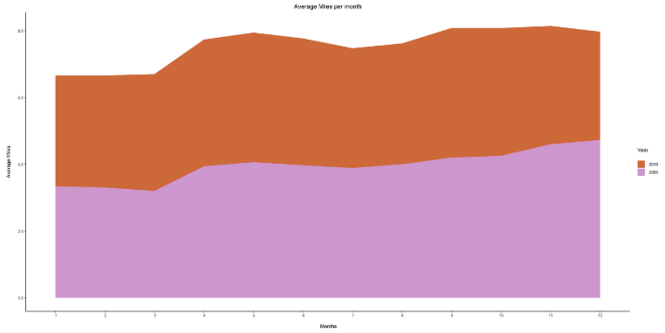


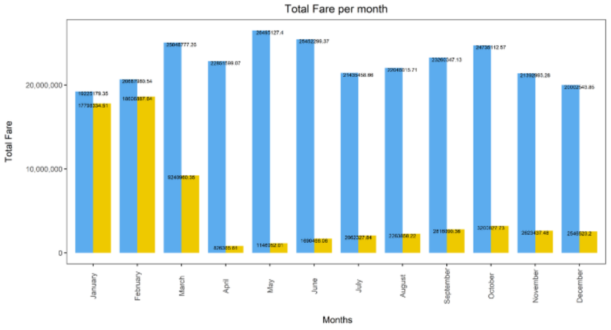
While cash and credit cards have been the preferred methods of payment in both the years 2019 and 2020, there is a slight dip in the number of credit card payments after the pandemic. It looks like digital apps have started gaining momentum post-pandemic.

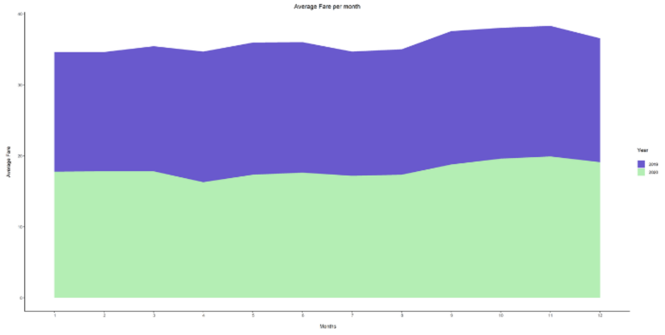
D. Number of taxi trips, total miles and total fare across months (2019 vs 2020)









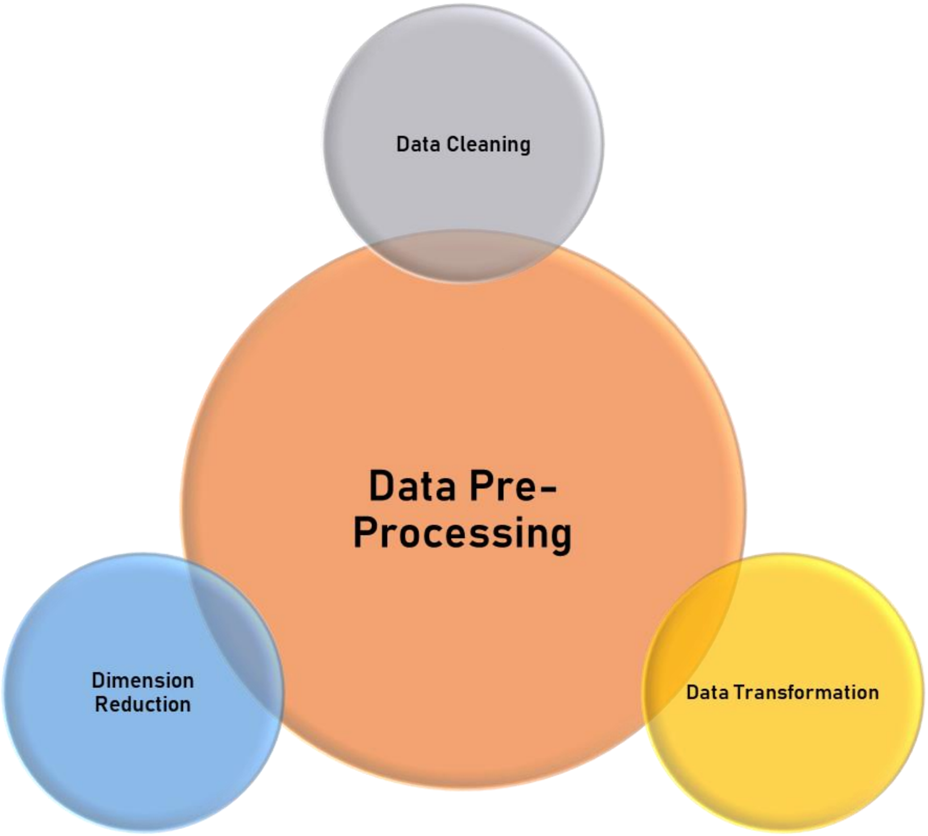


In the year 2019, the total number of trips taken every month remains stable with March, May and June being the chart-toppers. On the other hand, the year 2020 shows a downward trend due to the outbreak of COVID- 19 as it hits rock-bottom in April and plateaus through the rest of the year. The plots for total fare and total miles follow the same trend as the number of taxi trips is directly proportional to the distance covered on-road and revenue of the taxi industry.

An interesting point to note here is even though the business for the taxi industry has been dull in 2020, the mean number of trips taken and the mean trip cost continue to remain stable. On digging deep into this, we can infer that the taxi fare has either not been increased or the hike is insignificant in the year 2020.

Data Pre-processing

Raw data from an entity isn't very useful as it is often ambiguous. Data Pre-Processing is the step wherein data is encoded, manipulated, or dropped so that it matches the requirements of a machine learning model. Owing to the limited processing speed of our laptops and the enormous size of the dataset (around 8 GB), data is divided into batches and processed separately. Various Data Pre-Processing measures were implemented to ensure that any irrelevant or discrepant attributes don't disturb the functioning of the models.



A. Data Cleaning

The Chicago taxi trip dataset consists of a fairly high number of records with data that is incomplete or missing. These values are not suitable to the model that has been employed, and hence have to be removed. We used R programming to remove records containing null and missing values.

B. Dimension Reduction

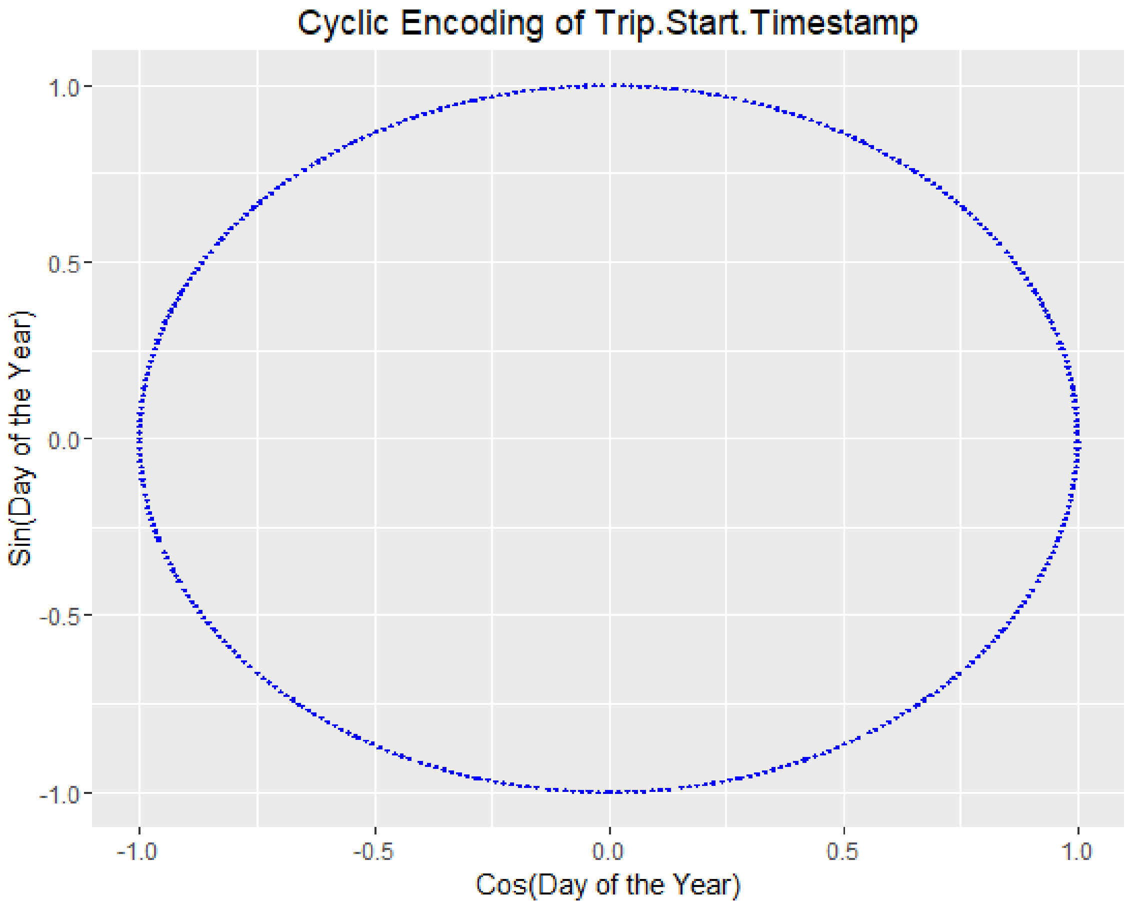
The higher the dimensionality of a dataset, the harder it is to fit it into a model. A total of 23 attributes forms this dataset, hence it is essential to omit the redundant fields as it reduces the size of the dataset. We removed attributes that don't serve any functionality to our business objective. Reducing the dimensionality enabled us better computational time to process high data. The following are some attributes that were removed from the chosen dataset.

- Pickup Census Tract
- Dropoff Census Tract
- Pickup Centroid Latitude
- Pickup Centroid Longitude
- Pickup Centroid Location
- Dropoff Centroid Latitude
- Dropoff Centroid Longitude
- Dropoff Centroid Location

C. Data Cleaning

Input attributes to the model in the relatively clean data set have to be modified so that the model understands it and discovers the underlying patterns to produce useful insights. The presence of time variables in the taxi dataset calls for cyclic encoding. To achieve this, the start date is first extracted from the start timestamp of the type character.

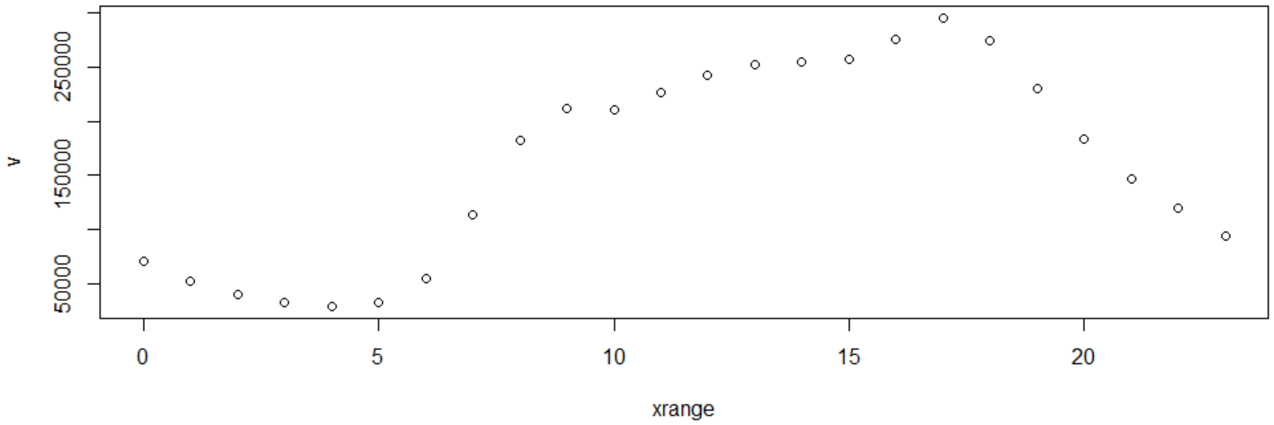
A function is called to number the days of the year from 1 to 365 (or 366, in case of a leap year). The sine and cosine values corresponding to every day of the year are computed using the functions sin () and cos (). A plot against the sine and cosine counterparts of a day of the year looks like the graph below:



An ideal way to encode a time series object is by using the “cyclic_encoding()” function from a package called “lubridate”. Given the enormous size of the dataset and the incorrect datatype of time variables, the said object didn’t prove to be an optimal method.

A new set of attributes are introduced to the dataset as a result of manipulating some existing attributes, namely - TimeBin, Avg_Miles and Avg_Fare.

- TimeBin - It categorizes every record based on the start time. We plotted the start time for all the trips and divided the day into 5 parts.



i.Bin ‘1’ - 12:00 AM to 4:59 AM

ii.Bin ‘2’ - 5:00 AM to 9:59 AM

iii.Bin ‘3’ - 10:00 AM to 2:59 PM

iv.Bin ‘4’ - 3:00 PM to 5:59 PM

v.Bin ‘5’ - 6:00 PM to 11:59 PM

This classification is done taking into consideration the hour at which a trip begins. This value is obtained by using an in-built function called “strp-time()”.

- Total_Trips - This field represents the total number of taxi trips in a given community area on a given date during a specific time interval.
- Total_Miles - This field represents the summation of the distance (in miles) covered during all taxi trips in a given community area on a given date during a specific time interval.
- Total_Fare - This field represents the summation of the cost of every trip (in US Dollars) for all taxi trips in a given community area on a given date during a specific time interval.
- Avg_Miles - This field represents the mean distance (in miles) covered during all taxi trips in a given community area on a given date during a specific time interval.
- Avg_Fare - This field represents the mean cost of every trip (in US Dollars) for all taxi trips in a given community area on a given date during a specific time interval.

Modeling

One of the biggest challenges the traditional taxi company faces is improving the efficiency of its taxi fleet. A predictive model capable of predicting taxi demand will help the companies plan ahead of the schedule of their taxis which will help in reducing wait time for the passengers, improving utilization of the resources and also help in predicting traffic “hot-spots”.

For this project, we decided to apply and compare three modeling algorithms, namely: Vector Auto Regression, Multi-Layer Perceptron and Recurrent Neural Networks. These algorithms have been extensively studied and applied to predict time series data.

We also decided to create separate models for each community area since it can be inferred that each community area has its time series and inherit patterns. In the future, data features extracted from other community areas can also be used to better model the taxi demand prediction.

Vector Auto Regression (VAR)

VAR models help capture relationships between different features in a time series dataset. It uses the past values of the variable, called lag, past values of other dependent variables and error terms to predict the next variables in the series.

VAR models have been used in natural sciences, finance sectors and economies as they mostly deal with time-series data. Other areas of applications include:

1. Data description
2. Forecasting
3. Structural inference
4. Policy analysis

Advantages of VAR

1. Forecasting a related variable collection where no extra interpretation is required.

2. Testing is done if only one variable is useful for forecasting.

Deep Neural Networks- Multilayer Perceptron

A deep neural network is a type of artificial neural network that has multiple hidden layers of units between its input and output layers. DNNs, like shallow ANNs, can model complex non-linear relationships. DNN architectures produce compositional models in which the object is represented as a layered composition of image primitives. The additional layers allow for the composition of features from lower layers, allowing for the modeling of complex data in fewer units than a similarly performing shallow network.

Neural networks were the prime choice of modeling algorithms for this project due to they being universal approximators, especially for non-linear relationships. Due to their nature, neural networks are specially used to model time series data.

Architecture

The development of MLP networks has 2 main agendas: Architecture & Training. The architecture of MLP is very important because a smaller number of connections may not be able to capture all the relationships within the data features, whereas more connections may cause an over-fitting of the training data.

The multilayer perceptron is one of the most prevalent types of deep neural networks. Deep feedforward networks or multilayer perceptrons are the foundational deep learning models.

These models are named feedforward because information flows via the function being evaluated from x, through the intermediate computations necessary to define f, and finally the output y. There are no feedback links, therefore the model's outputs are not transmitted back into it.

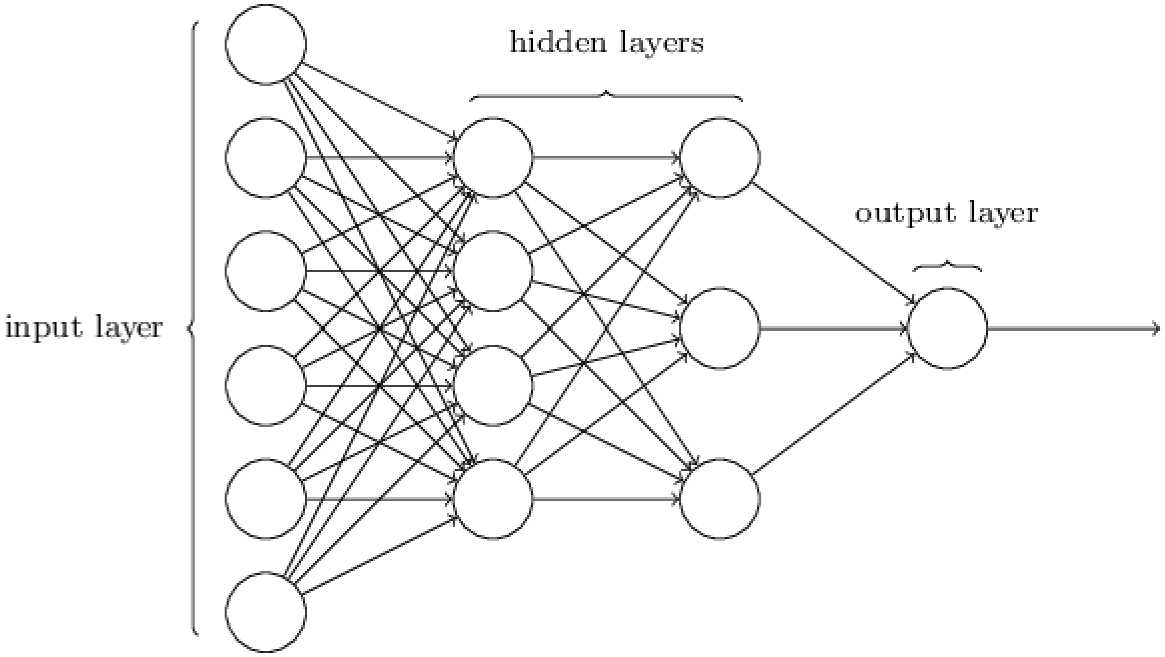


Fig 1: Multilayer Perceptron Architecture

Disadvantages of MLP

1. Computations are very time-consuming and sometimes become complicated because of complex connections.

2. Neural networks are black boxes; It's hard to interpret their learned network and how the training algorithm arrived at that particular structure.

Recurrent Neural Network (RNN)

Recurrent neural networks are a class of neural networks that allow previous outputs to be used as inputs while having hidden layers. RNNs models are extensively used in natural language processing, speech recognition, sequential prediction, etc.

Recurrent neural networks follow backpropagation through the time algorithm to determine the gradients. The principles of backpropagation time are the same as the general backpropagation. The model trains itself by calculating errors from its output layers and shifts back to its input layer.

Types of RNN

1. One-to-One
2. One-to-Many
3. Many-to-One
4. Many-to-Many

Working of RNN

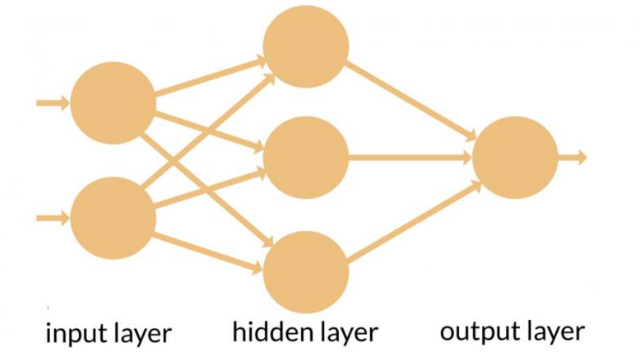


Fig 2: RNN

In a feed-forward network, when the training data is given to the model it goes through the hidden layers evaluating through the activation layers and displays the output through the output layers. This network doesn’t touch the nodes once again after going through it.

Feed forward networks are not efficient in predicting the new values because it considers only the input which comes from train data.

RNN completes its information in a cycle through a loop. When producing the output, RNN learns from its current input and considers the input value which was given previously.

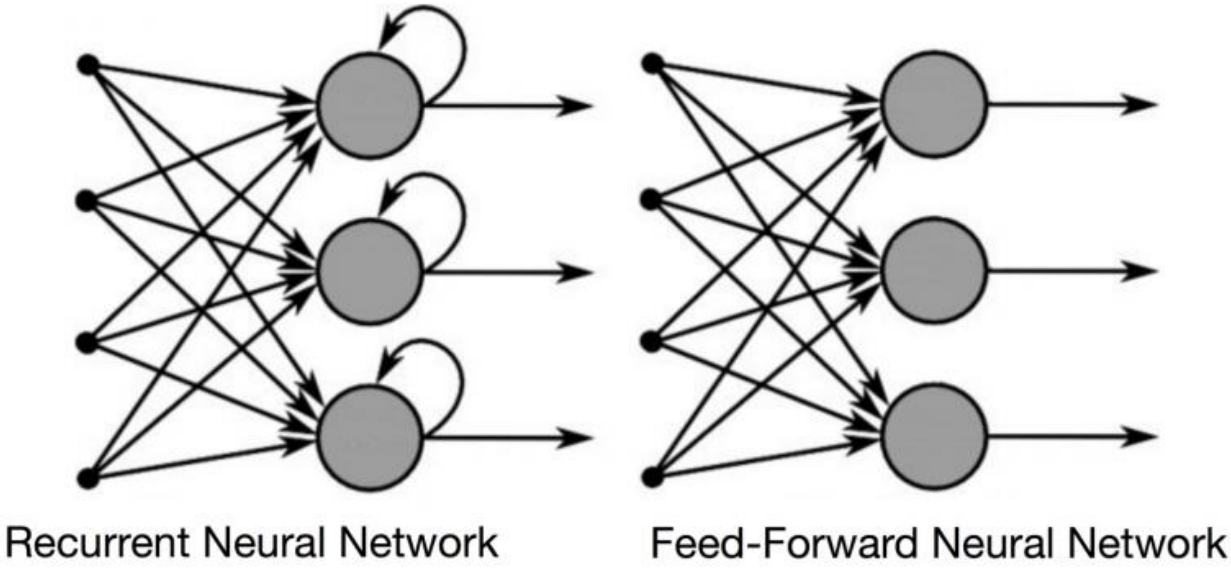


Fig: Comparison between RNN & Feed forward network

Advantages of RNN

1. Can process the input of any length.
2. Model size remains constant.
3. Weights are shared according to time.

Disadvantages of RNN

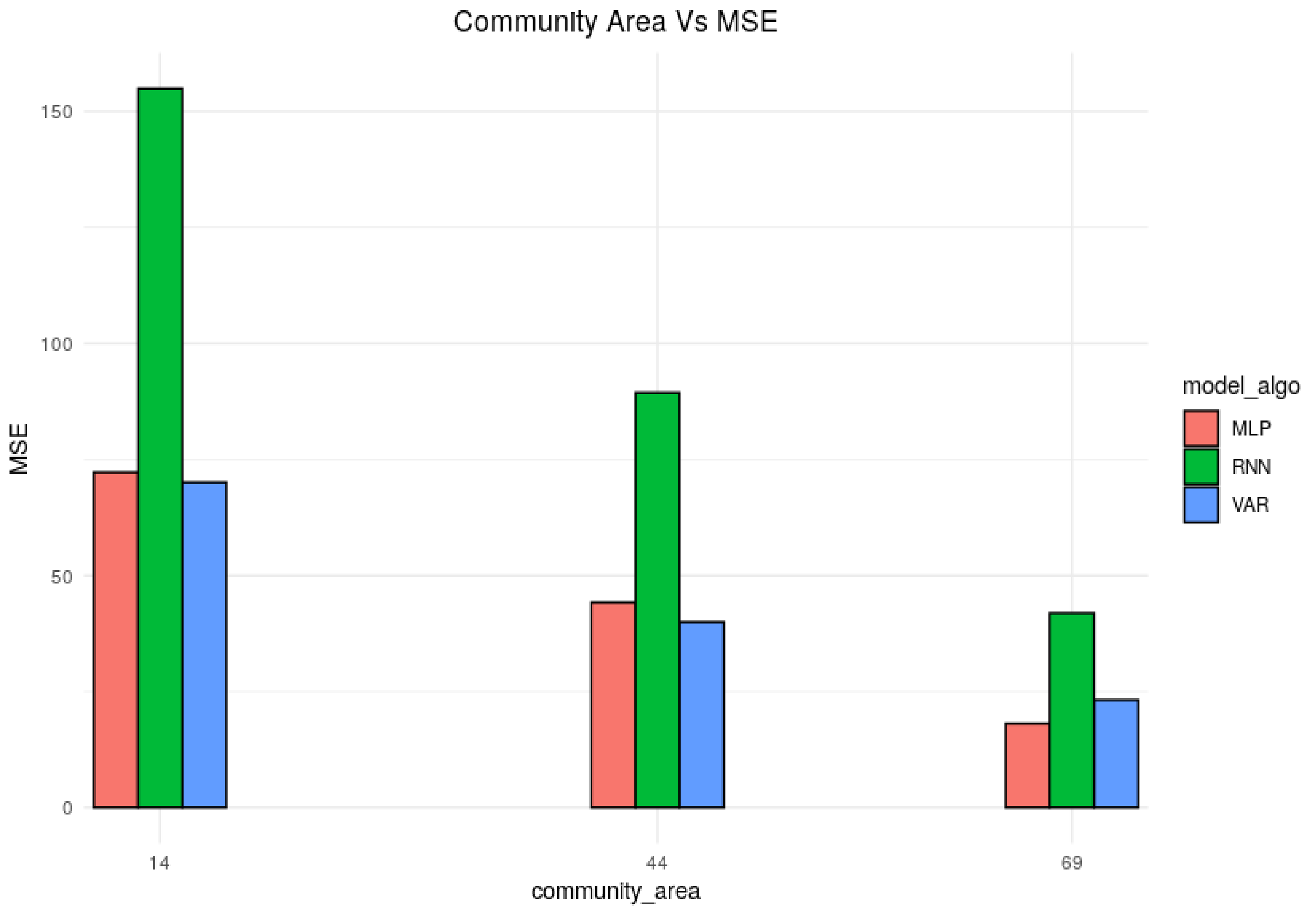
1. Training RNN's can be very computationally expensive.
2. Cannot consider any future input for the current state.

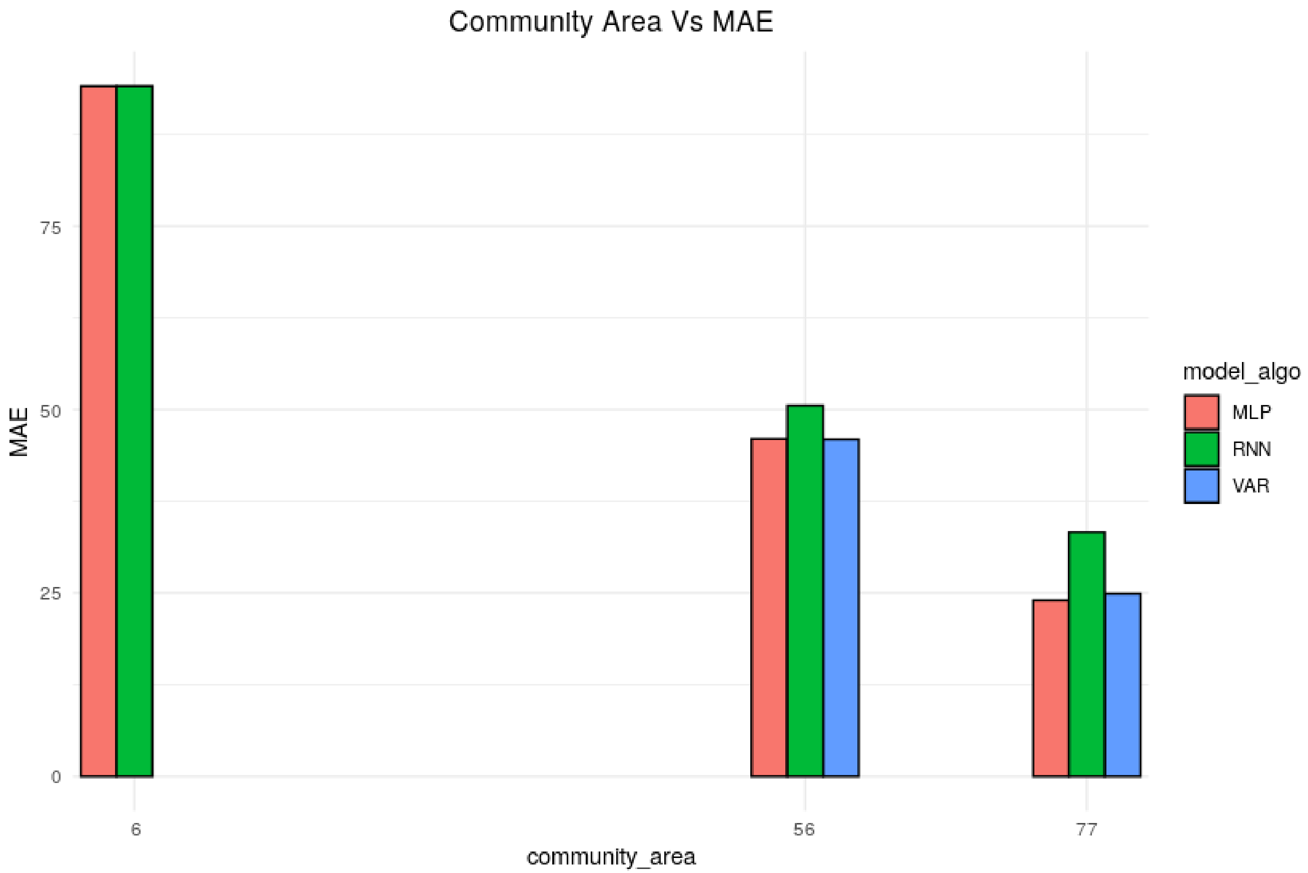
MLP vs RNN vs VAR

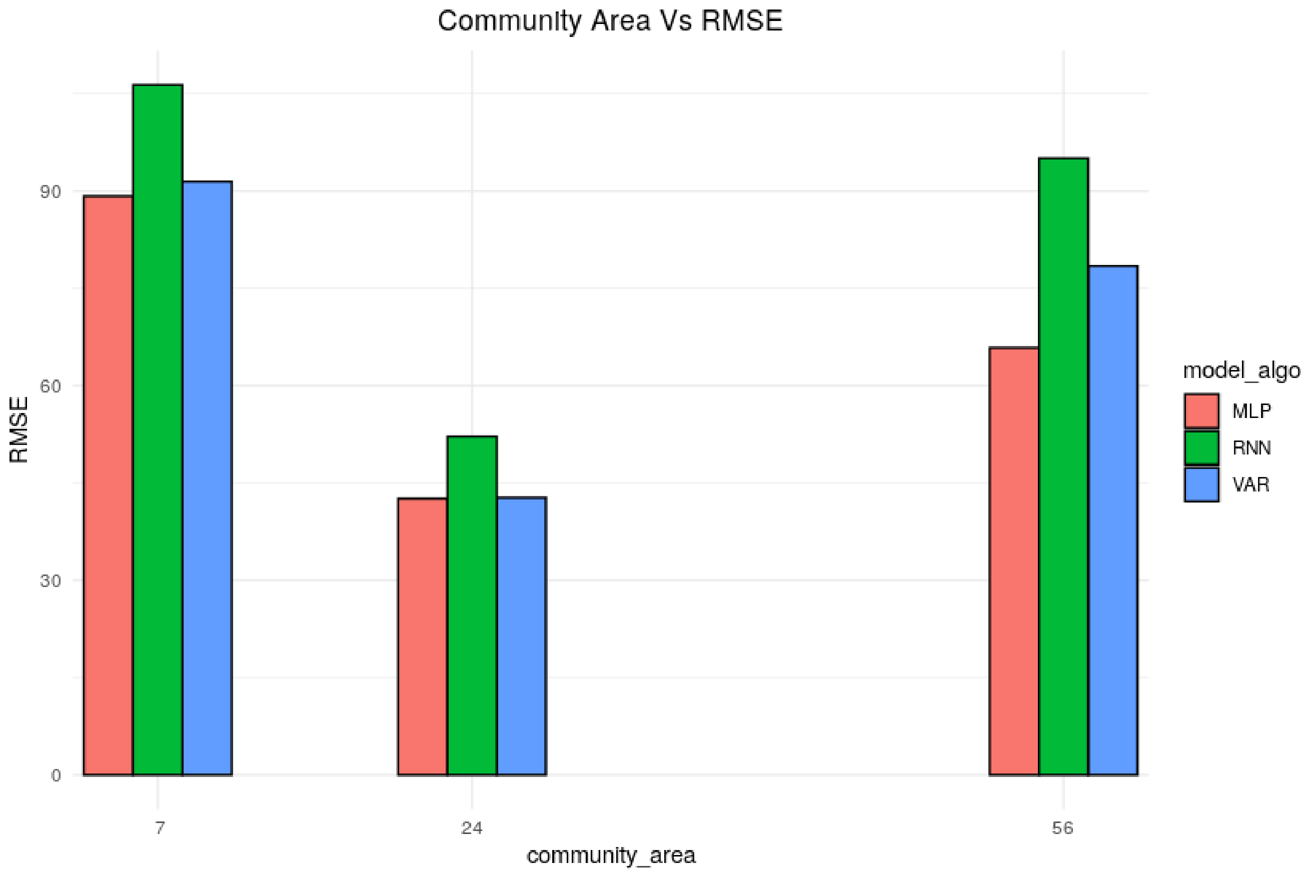
Multi-Layer Perceptron	Recurrent Neural Network	Vector Auto Regressions
MLP is the foundation of more complex deep neural networks	RNN is designed for sequence-related problems.	VARs are generally used to model linear relationships between variables over time.
Can model non-linear relationships	Can model non-linear relationships as well as cyclic patterns	Not useful for modeling non-linear relationships
Does not take previous values into account	Runs a feedback loop, as a result, previous value can influence present values	Uses lag variables to account for cyclic patterns.

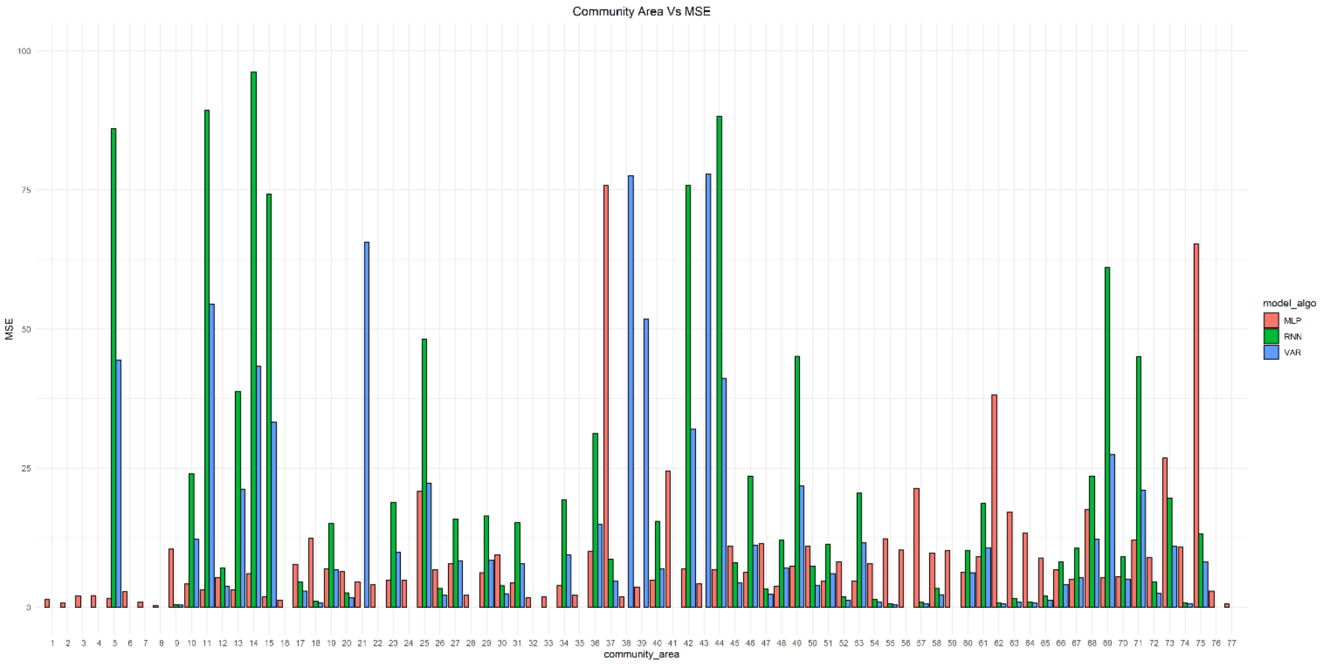
Evaluation

We have calculated Mean Absolute Error, Mean Squared Error and Root Mean Squared Error for 3 algorithms for all 77 community areas. Also, for each evaluation metric, we are comparing the modeling algorithms for only the top 3 community areas which show the greatest variance, to better visualize the model's comparative performances.




Community Area Vs MAE
MAE
0
25
50
75
6
56
77
community_area
model_algo
🟥 MLP
🟩 RNN
🟦 VAR


Community Area Vs RMSE
RMSE
0
30
60
90
7
24
56
community_area
model_algo
🟥 MLP
🟩 RNN
🟦 VAR



The graphs clearly show that the metrics are generally on the higher side for RNNs, followed by VAR and MLP. Interestingly, we expected Recurrent Neural Networks to perform better than the other two algorithms, however, the metrics show a different story. However, this shouldn't be deduced to the fact that RNNs are a weaker selection. Complex neural networks like those require a precise approach to data preparation and hyperparameter selection, success in achieving them will result in a significant accuracy improvement.

MLP, being a simple deep neural network, performed better than RNN, which requires a more precise approach and VAR, since VAR can only recognize linear relationships.

Reflection

One of the biggest challenges we faced as a team was working with a large dataset. Our approach had to be modified in the later stages of the project due to our task of handling large datasets. We made use of a library in R named “data.table” which aids in fast processing of large data, limited to available RAM.

We also observed that using complex neural networks like RNNs requires a much careful approach to data preparation, feature selection and hyperparameter selection. In a practical business scenario, a simple deep neural network such as MLP can satisfy business requirements whilst reasonably compromising on prediction accuracy.

Conclusion and Future Scope

We performed a statistical project on the Chicago taxi dataset using CRISP-DM. Through Exploratory Data Analysis, we discovered and compared trends in pre-COVID and post-COVID eras. We pre-processed and modeled the dataset to predict taxi demand using 3 modeling algorithms, and compared their performances.

Lastly, we had experience working on a very large dataset and adapted our code to work on them. Even though the dataset we worked on was fairly large (around 8 GB), it was just a small sample of the overall dataset, starting from 2016.

In future projects, we would like to further refine our approach towards modeling Recurrent Neural Networks, to fully utilize their potential. We could also combine the dataset with datasets to capture external signals which might be affecting the demand such as weather, events and festivals, traffic data, etc.

References

1. Lupkin, J. M. (n.d.). Encyclopedia of Chicago. Retrieved from http://www.encyclopedia.chicagohistory.org/pages/1232.html

2. Tribune, C. (n.d.). 1920–Taxi Wars. Chicago: Chicago Tribune. Retrieved from https://chicagology.com/notorious-chicago/1920taxiwars/

3. Uber. (n.d.). UberDATA: Uber's Economic Impact on the City of Chicago. Retrieved from https://www.uber.com/blog/chicago/uberdata-ubers-economic-impact-on-the-city-of-chicago/