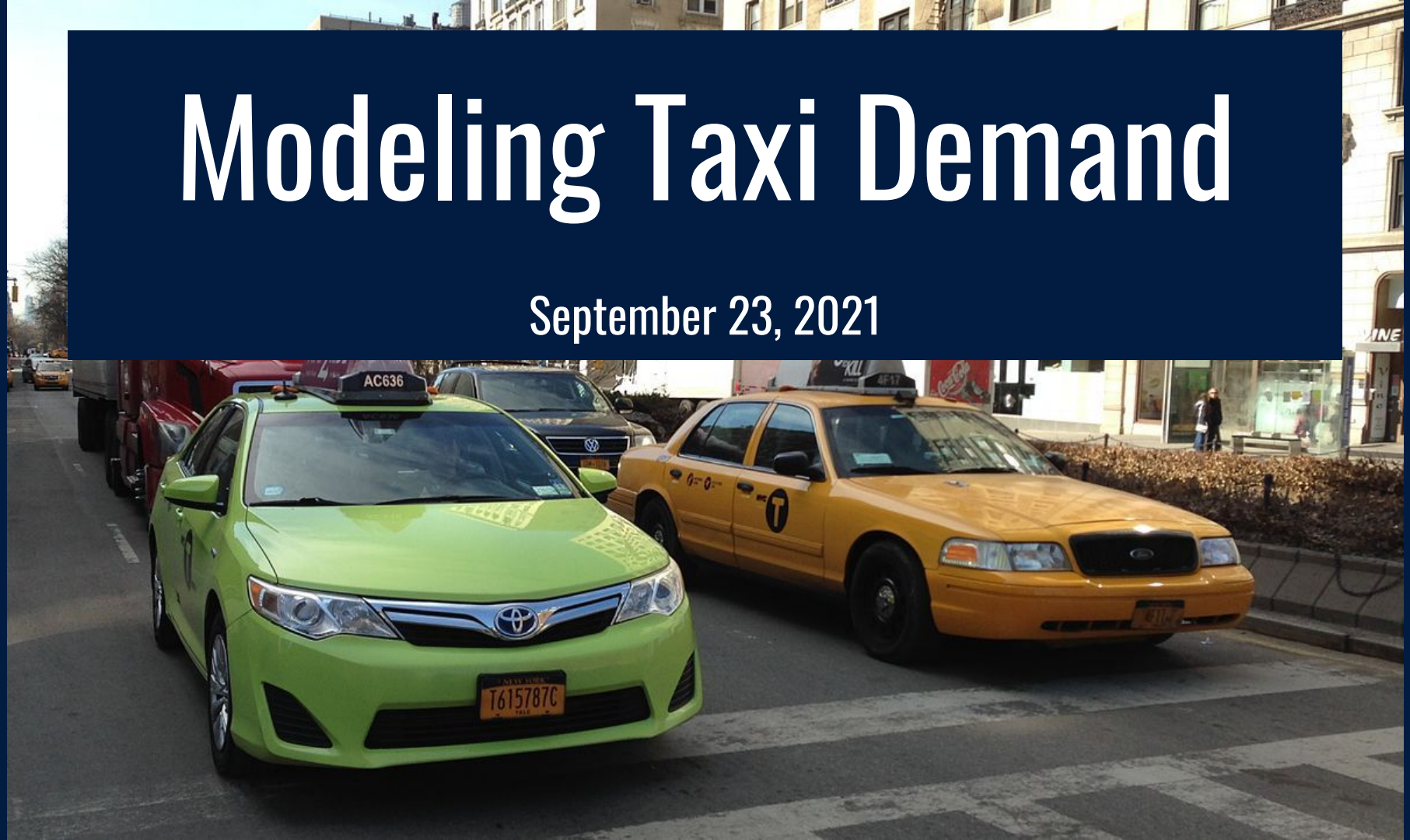
Taxi Fare Prediction

September 23, 2021

Modeling Taxi Demand

September 23, 2021

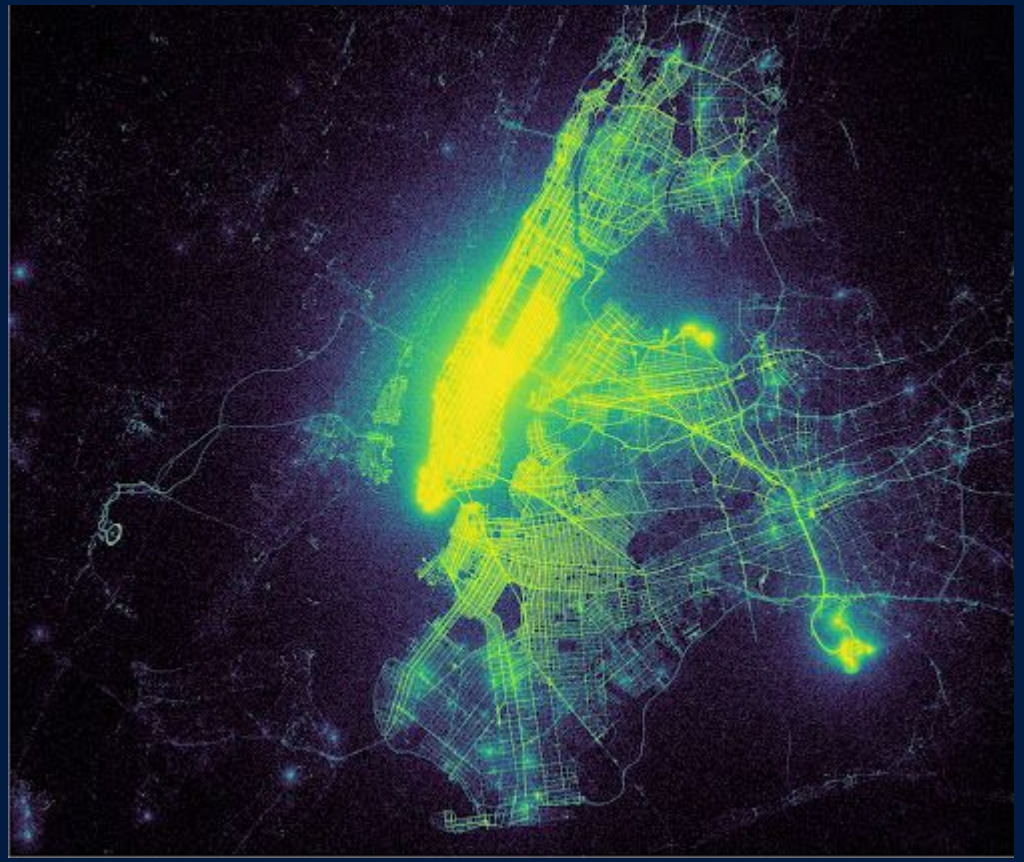


PREDICT TAXI DEMAND

What's in it?

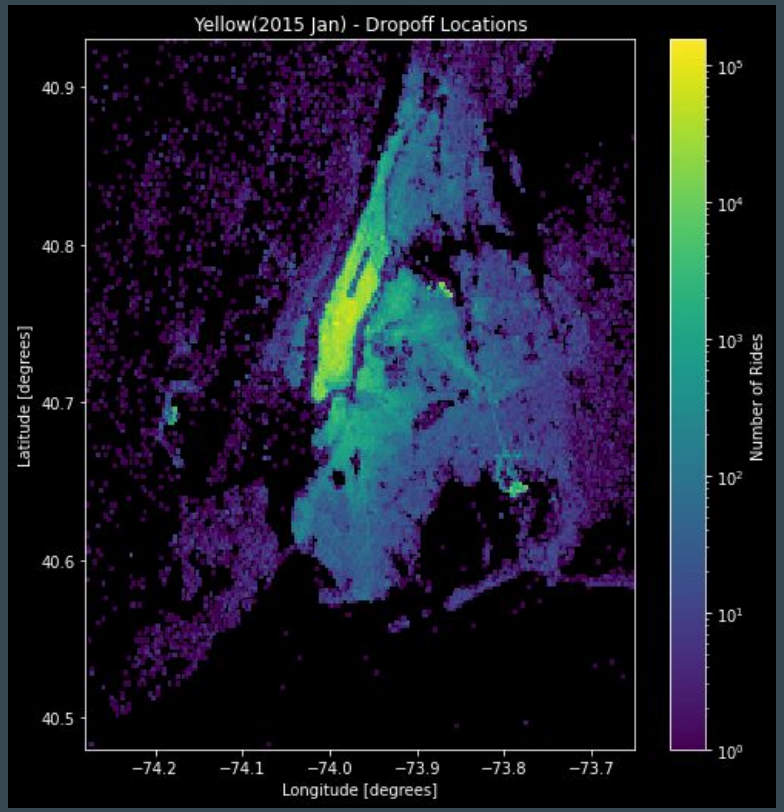
Company:

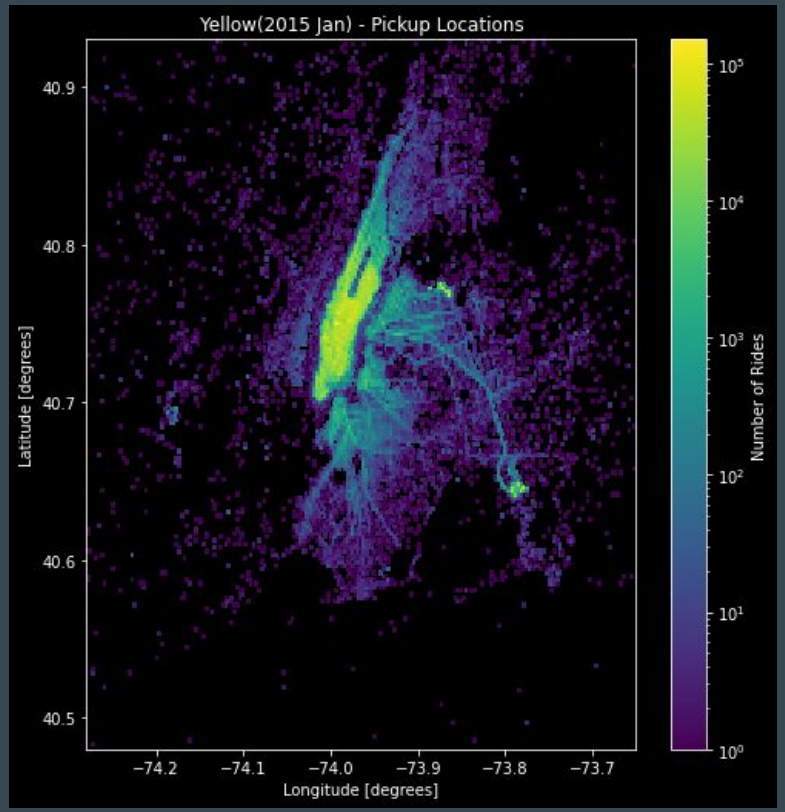
- Predicting demand and deploy taxis accordingly there by efficiently making use of resources
- Use the demand prediction model to offer discounts or surcharges, thereby increasing the revenue.



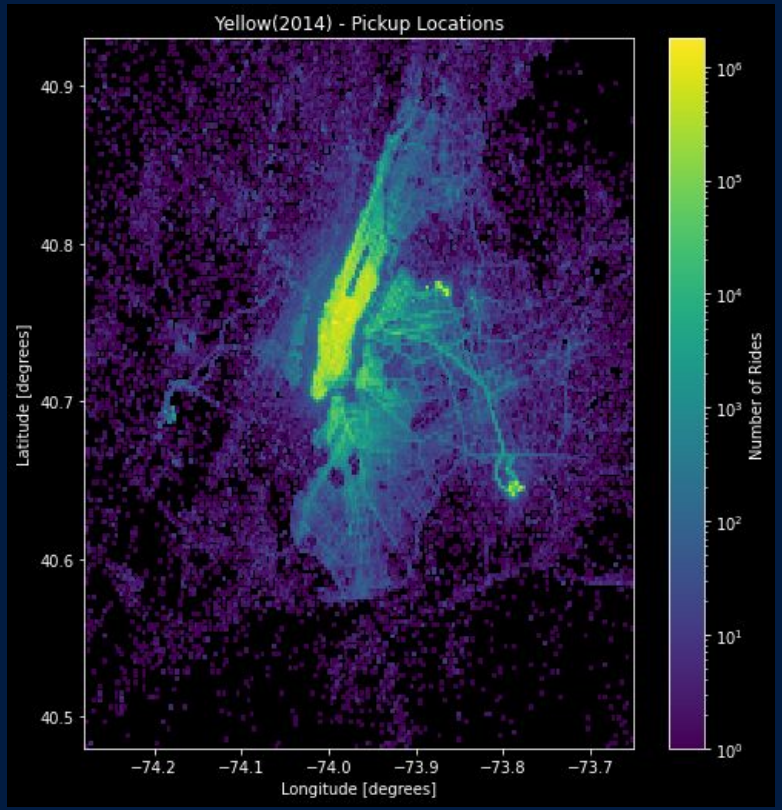
YELLOW TAXI

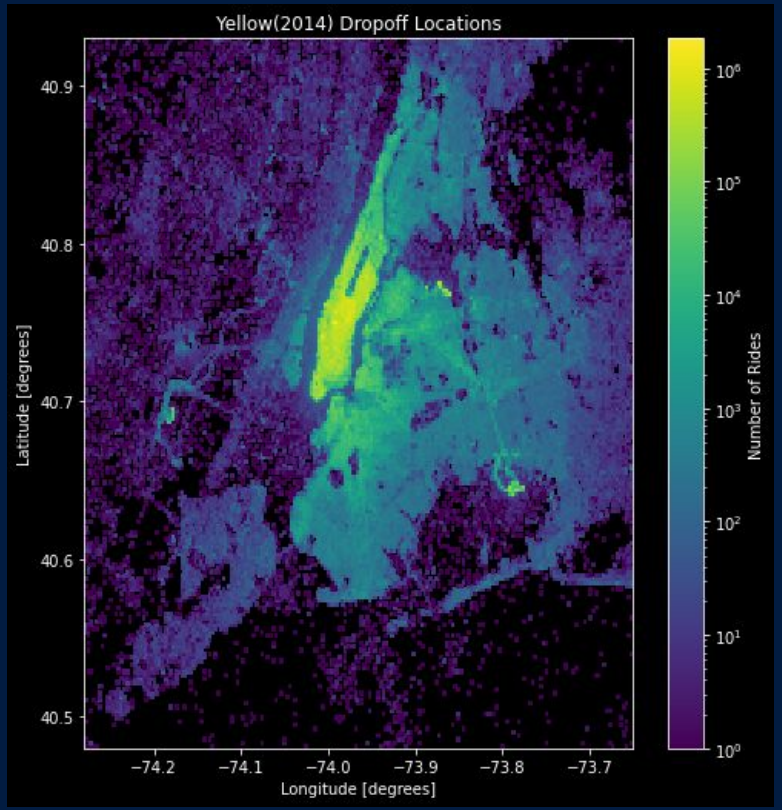
2015 - Pickup/Drops By Months



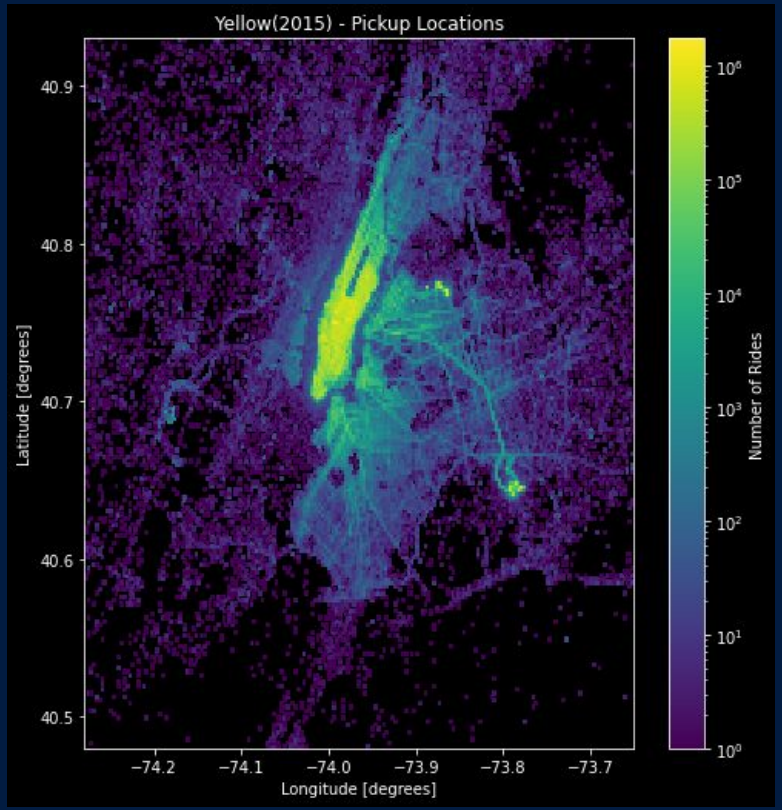


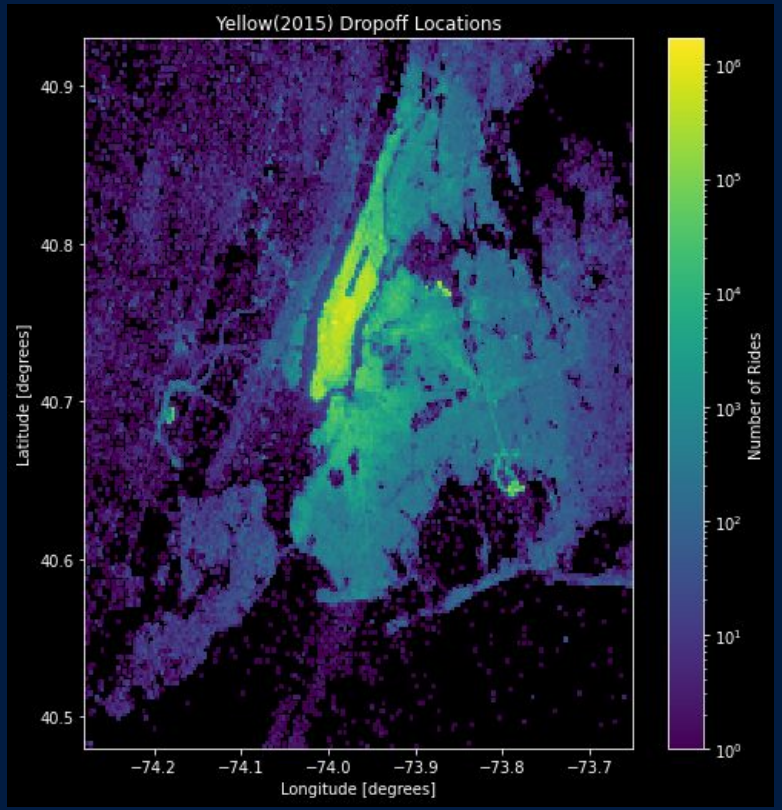
2014


Yellow(2014) - Pickup Locations
Latitude [degrees]
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude [degrees]
Number of Rides
10^6
10^5
10^4
10^3
10^2
10^1
10^0


Yellow(2014) Dropoff Locations
Latitude [degrees]
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude [degrees]
Number of Rides
10^6
10^5
10^4
10^3
10^2
10^1
10^0

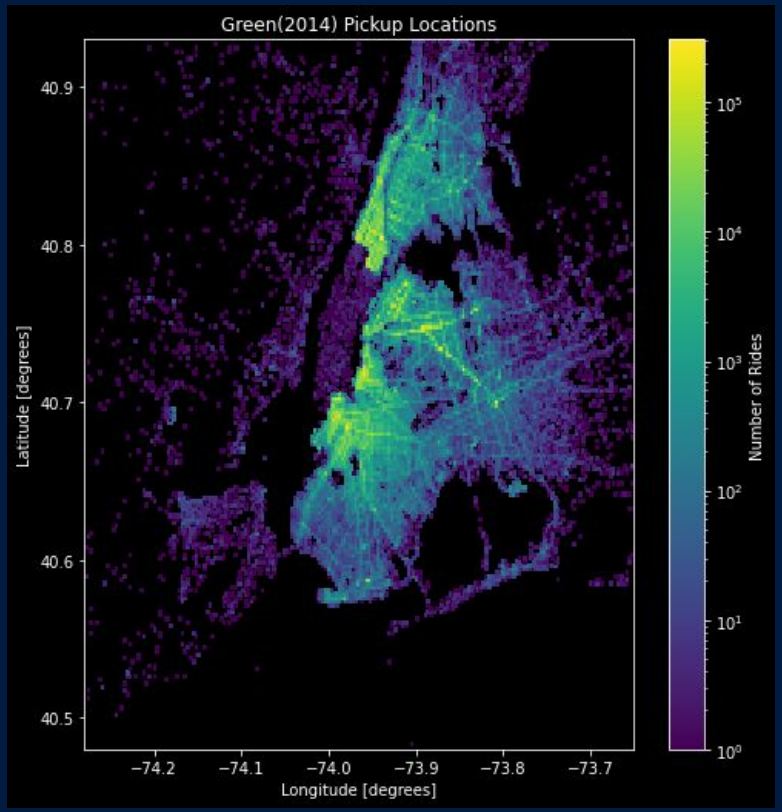
2015

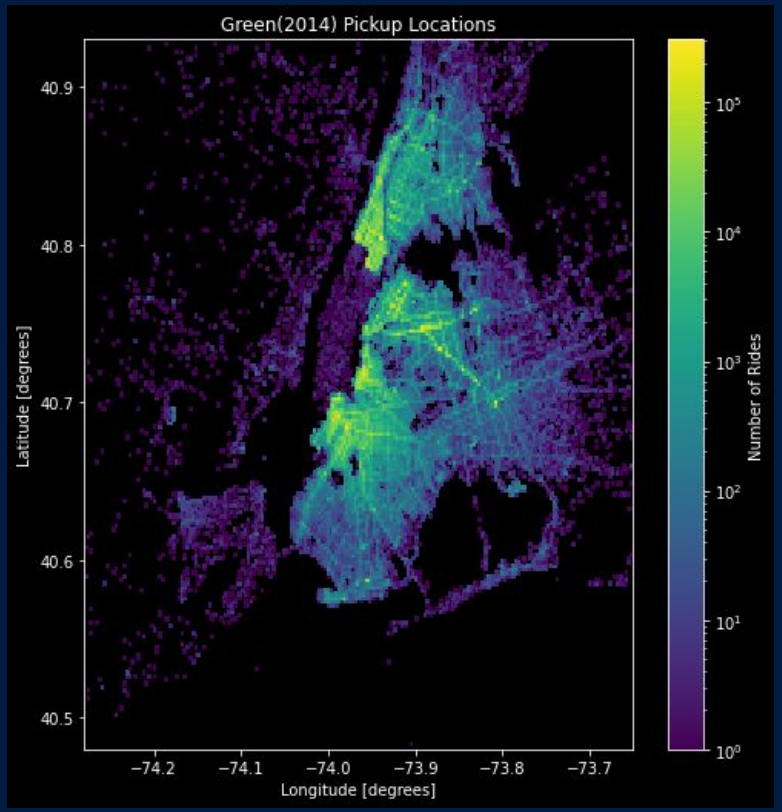

Yellow(2015) - Pickup Locations
Latitude [degrees]
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude [degrees]
Number of Rides
10^6
10^5
10^4
10^3
10^2
10^1
10^0


Yellow(2015) Dropoff Locations
Latitude [degrees]
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude [degrees]
Number of Rides
10^6
10^5
10^4
10^3
10^2
10^1
10^0

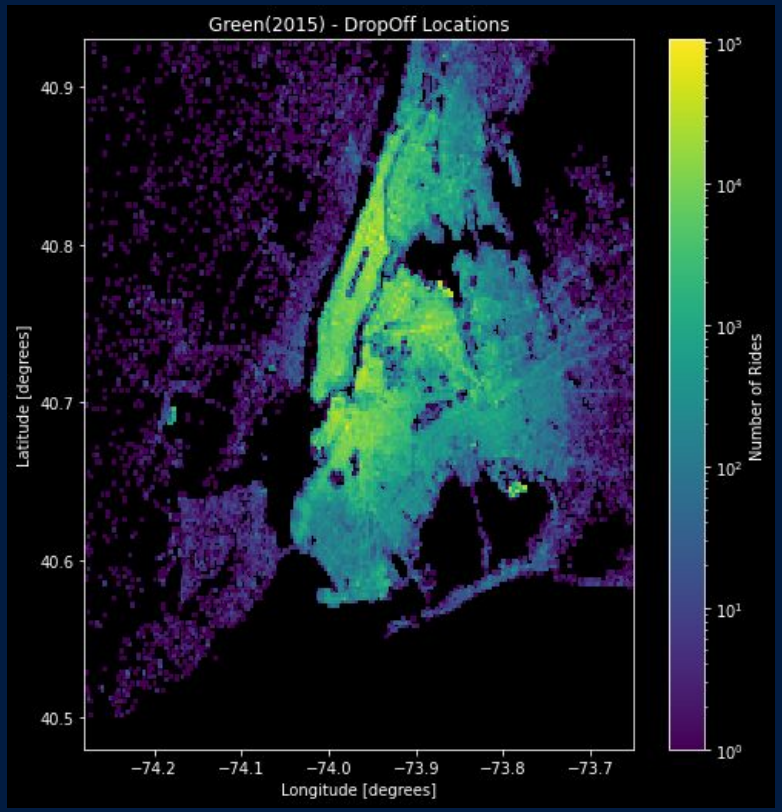
GREEN TAXI

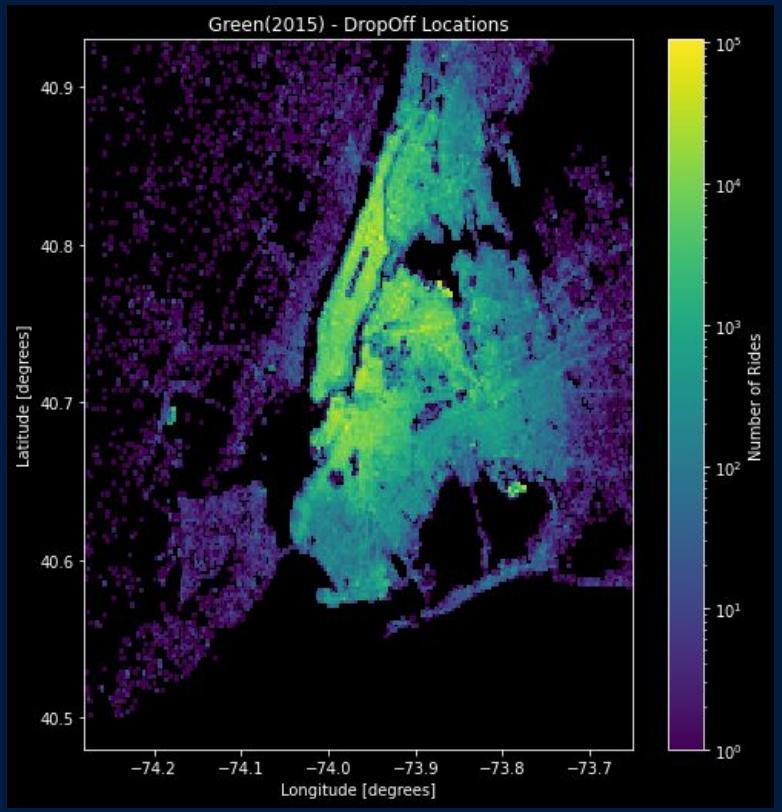
2014


Green(2014) Pickup Locations
Latitude [degrees]
Longitude [degrees]
Number of Rides
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
10^5
10^4
10^3
10^2
10^1
10^0


Green(2014) Pickup Locations
Latitude [degrees]
Longitude [degrees]
Number of Rides
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
10^5
10^4
10^3
10^2
10^1
10^0

2015


Green(2015) - DropOff Locations
Latitude [degrees]
Longitude [degrees]
Number of Rides
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
10^5
10^4
10^3
10^2
10^1
10^0


Green(2015) - DropOff Locations
Latitude [degrees]
Longitude [degrees]
Number of Rides
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
10^5
10^4
10^3
10^2
10^1
10^0

What are the options?

Probabilistic Models

- Linear regression
- Multiple Linear Regression.

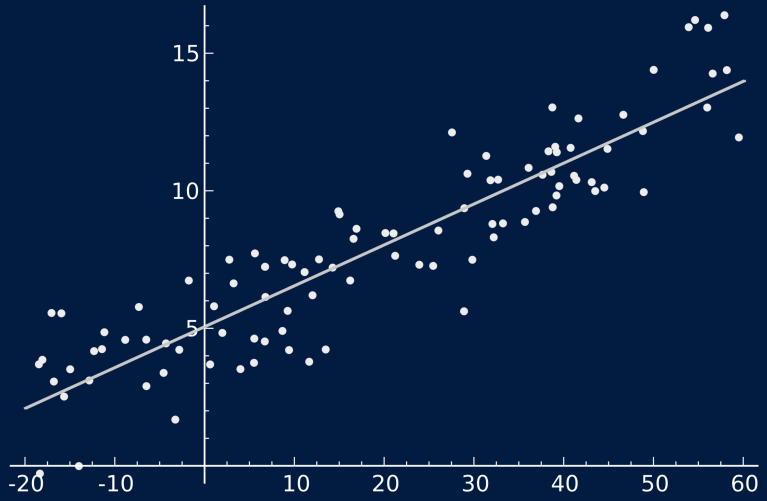
Machine Learning Models

- Artificial Neural Networks
- Decision Tree
- Clustering
- Ensemble Methods

Time Series Models

- Autoregressive(AR)
- Vector AR
- Moving Average(MA)
- ARIMA
- STARIMA

Multiple Regression Model



$$Y = \sum_{i=0}^{n} \beta_i X_i + \epsilon$$

Use MLE(Least Squares Estimation) to determine the coefficients for the explanatory variables.

Step 1: Check Correlation Coefficients

Strong correlation b/w X_i and Y indicates X_i is important

Strong correlation b/w X_i and X_j - multicollinearity

Step 2: Stepwise Selection

Step 3: Best Subsets Regression

$$y = b + W^T X$$

$$f(x) = \sigma(b + W^T X)$$

$$W = \begin{bmatrix} w1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} X = \begin{bmatrix} x1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

ARTIFICIAL NEURAL NETWORK

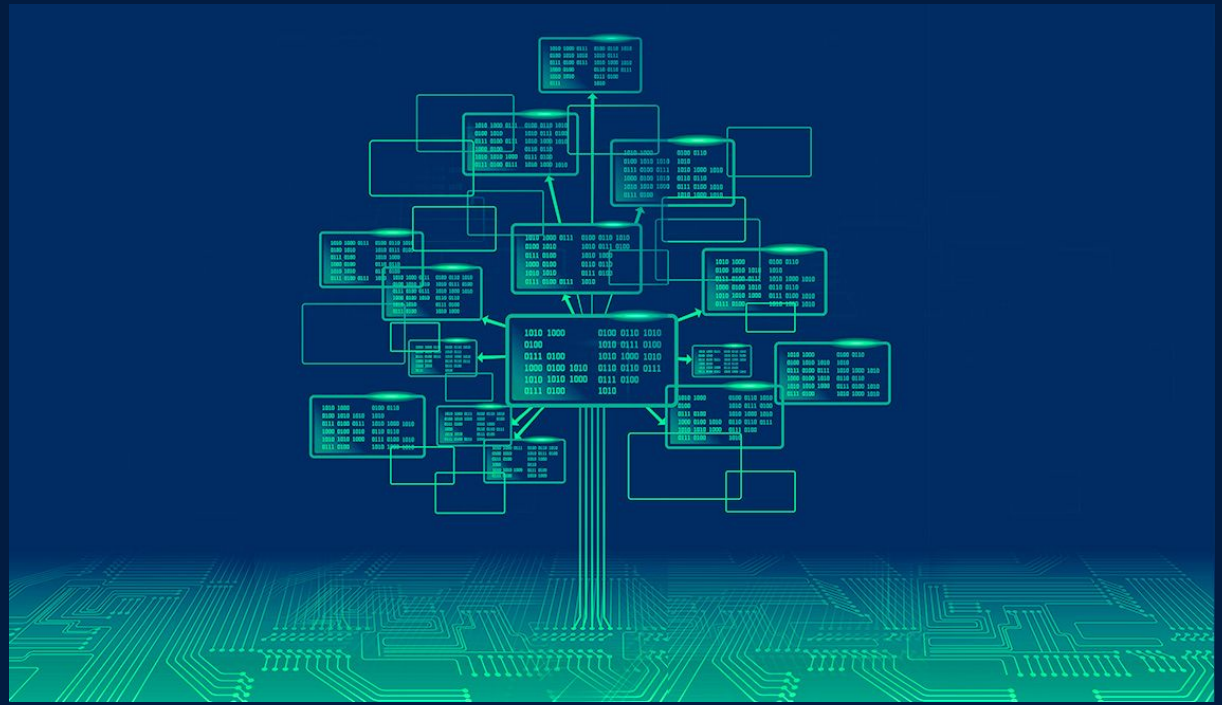
Pros:

- Fast Predictions
- Good with non-linear data
- Can be used for both regression and classification

Cons:

- Computationally intensive
- Needs lot of training data
- Overfitting and generalization

Decision Trees



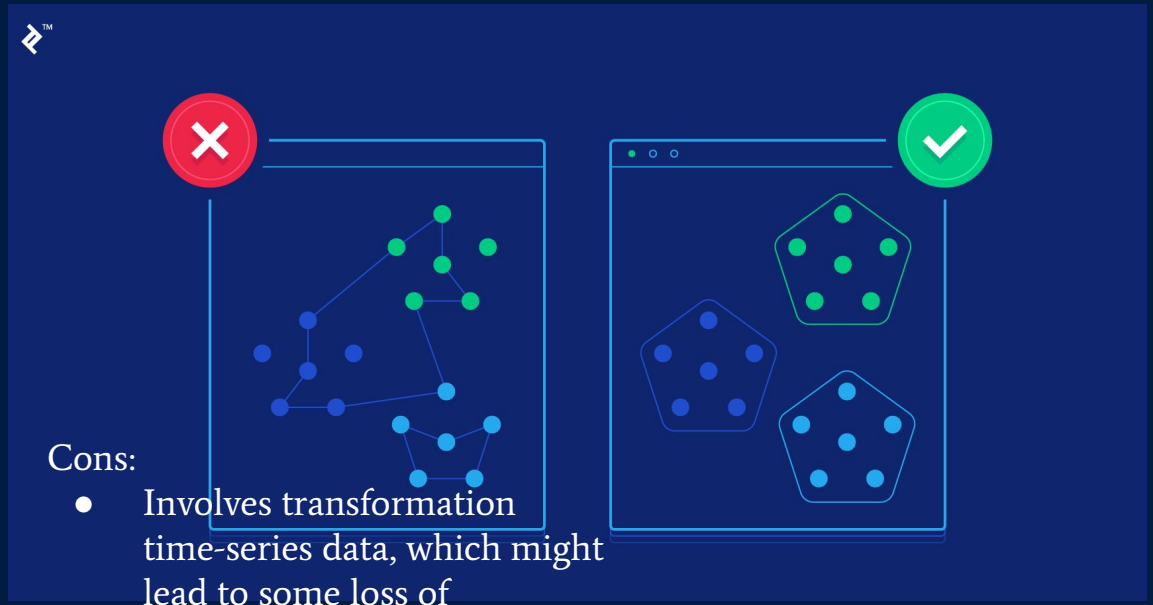
Pros:

- Understandability
- Resistant to Outliers

Cons:

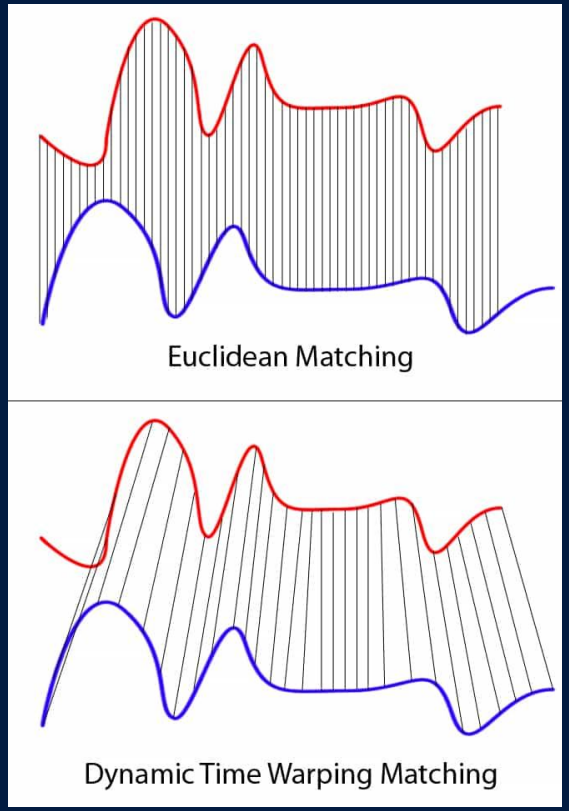
- Prone to overfitting
- Needs careful Parameter tuning
- Biased Tree can be created if there is an imbalance

CLUSTERING WITH DYNAMIC TIME WARPING

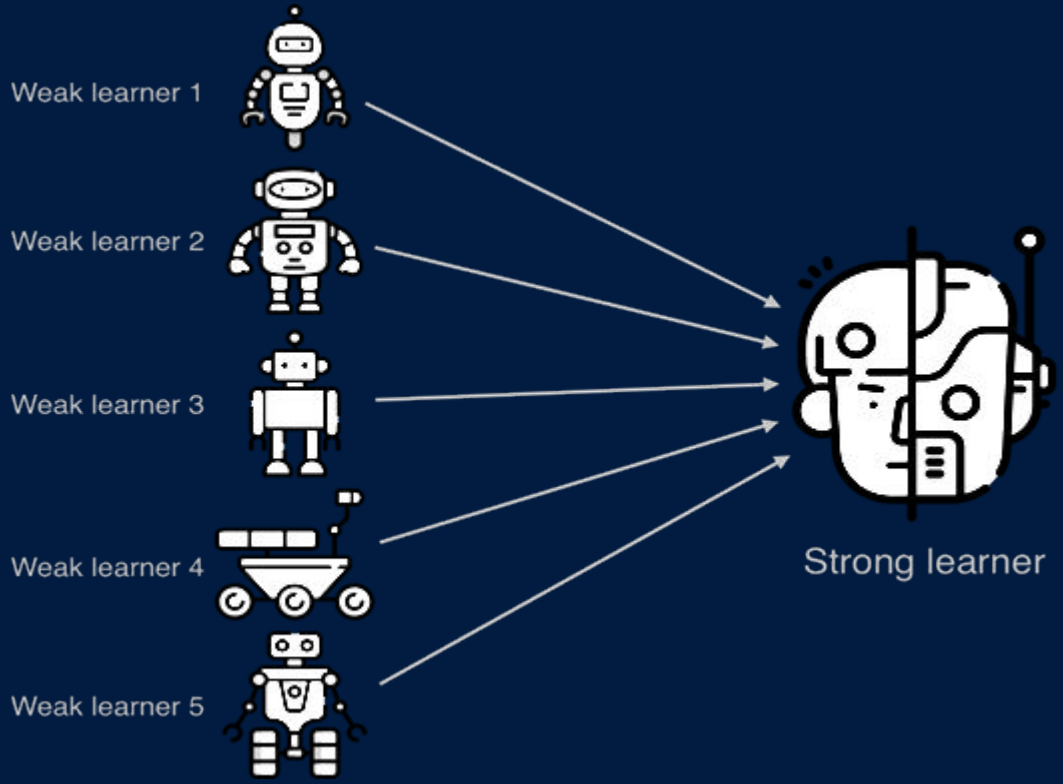


Cons:

- Involves transformation time-series data, which might lead to some loss of information
- Incorrect Parameter setting due to assumptions
- Slow processing time



ENSEMBLE


Weak learner 1
Weak learner 2
Weak learner 3
Weak learner 4
Weak learner 5
Strong learner

TIME SERIES

What kind of Forecasting?

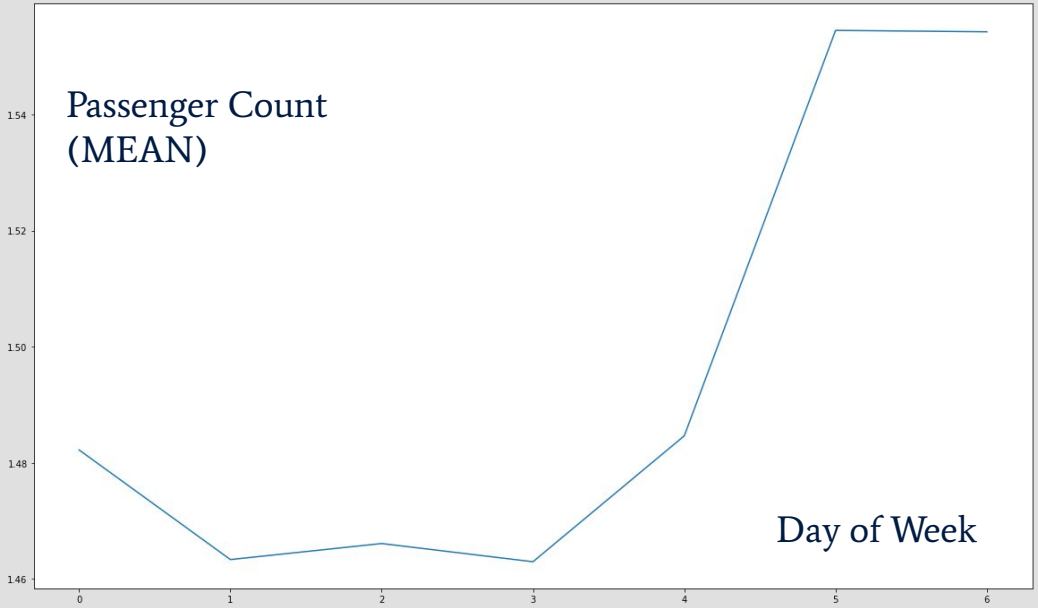
- Point forecast
- Interval forecast
- Density forecast

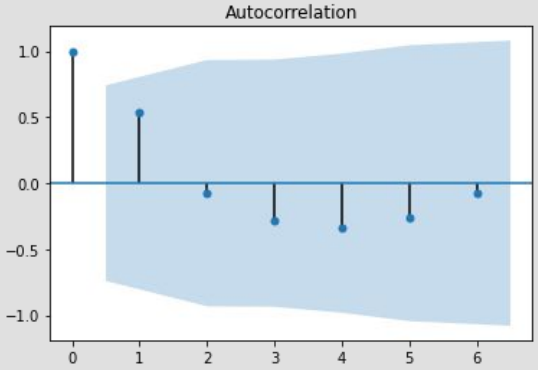
What kind of data needed?

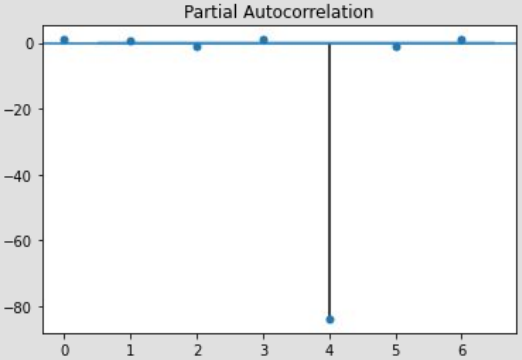
- Data collected at a single point of time
- Observations of data made over time

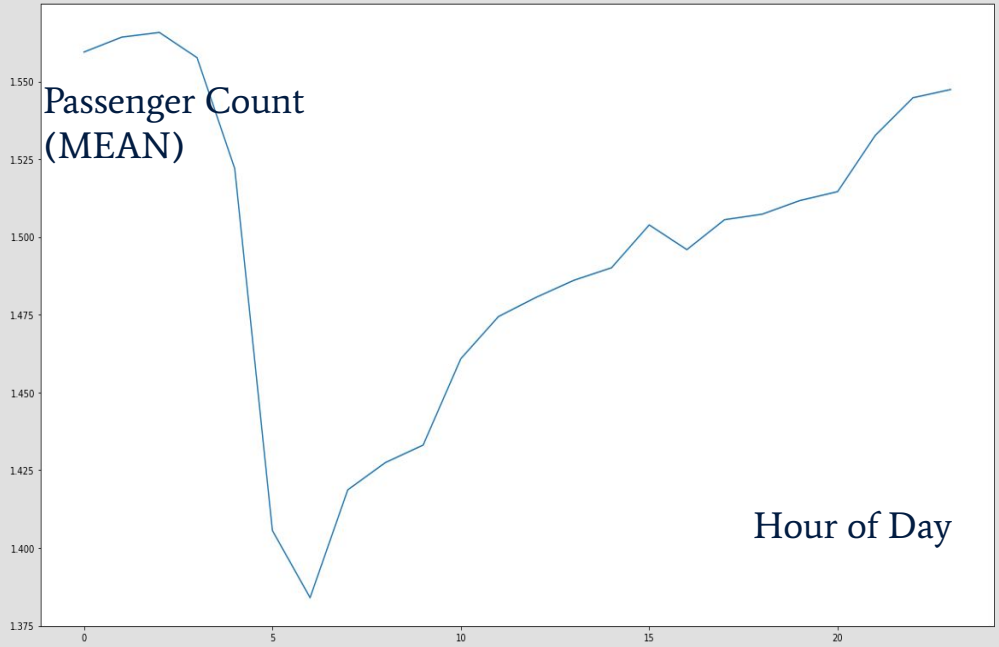
Identify the patterns in data:

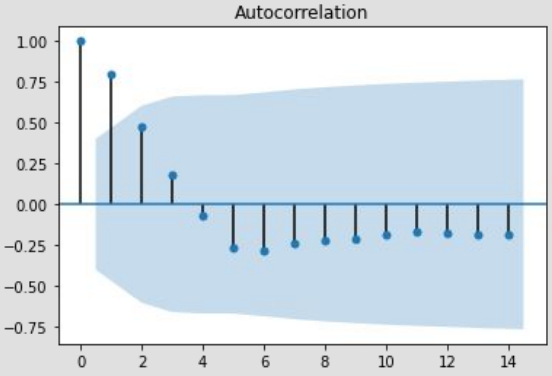
- Horizontal (Stationary)
- Trend
- Seasonal
- Cyclical

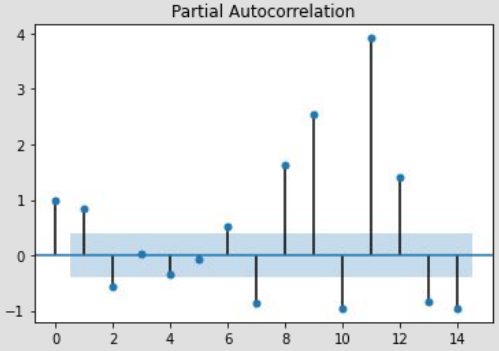

Passenger Count
(MEAN)
Day of Week
1.54
1.52
1.50
1.48
1.46
0
1
2
3
4
5
6


Autocorrelation
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5
6


Partial Autocorrelation
0
−20
−40
−60
−80
0
1
2
3
4
5
6


Passenger Count
(MEAN)
Hour of Day
1.550
1.525
1.500
1.475
1.450
1.425
1.400
1.375
0
5
10
15
20


Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
2
4
6
8
10
12
14


Partial Autocorrelation
4
3
2
1
0
−1
0
2
4
6
8
10
12
14