Predict Future Number of Taxi Transactions in New York City

Yun-Chung Liu

Summary

- The goal of this project is to accurately predict number of transactions in future time points with fine temporal and spatial resolution.
- The task is to predict hourly number of transactions in December 2019 in 8 taxi zones in New York City.
- The best performing model, Random Forest Regression, reduced prediction error of baseline model exponential average by **80 %**.

	Exponential Smoothing	**Autoregressive Moving Average**	**Random Forest Regression**	**Long-Short Term Memory**
RMSE*	47.6	11.0	**9.6**	195.3

RMSE: root mean squared error

Objective

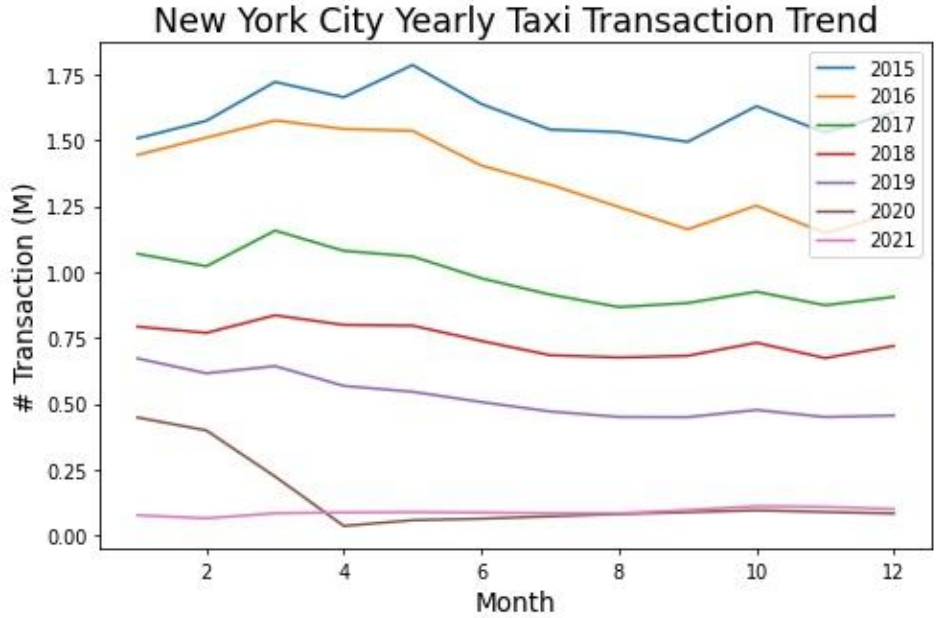
- Accurate prediction of future number of taxi transaction enables hailing company to more efficiently allocate drivers, design incentives and set prices accordingly.
- This presentation will walk through exploratory analysis (looking for patterns and useful features for precision), model buildup (time series, machine learning and deep learning models) and performance evaluation.

Data Source

This demo uses New York City Taxi and Limousine Commission (TLC) Trip Record Data for exploratory analysis and model buildup. The trip data is publicly available on their website updated until June 2022 as of this writing.

Taxi Ride Trend in Different Months of Year

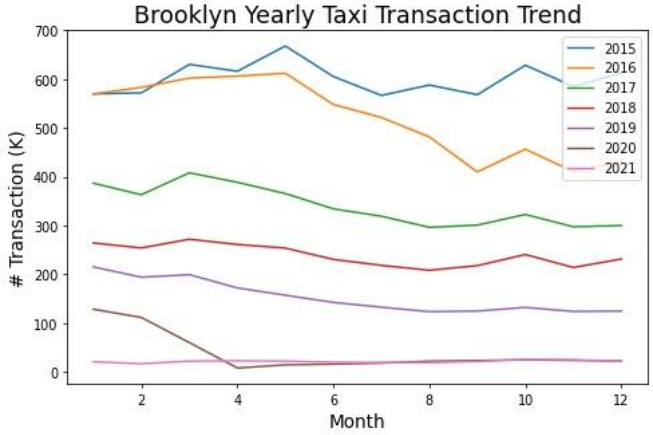
Taxi transactions shared similar pattern from 2015-19, with total amount of transaction decreasing gradually. The number of transactions dropped dramatically since April 2020, presumably due to the Covid pandemic.

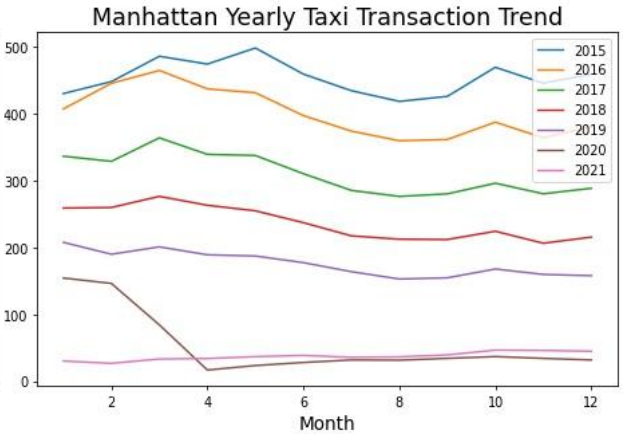


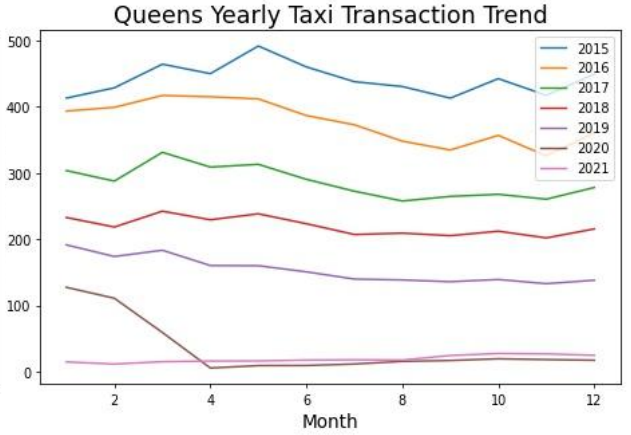
Boroughs

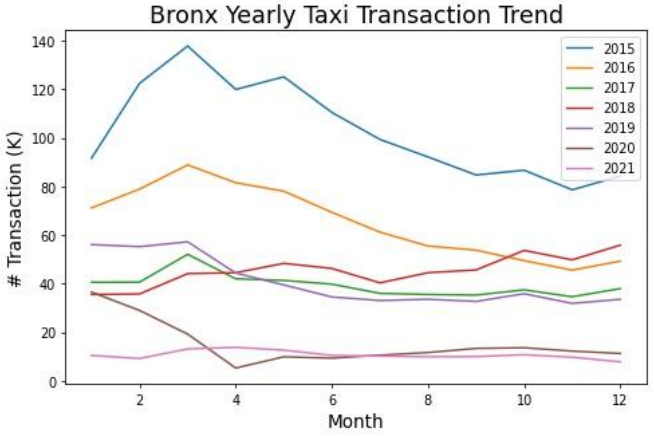
There are five boroughs in New York city, Brooklyn, Queens, Bronx, Manhattan and Staten Island as shown below. From the figures in the next page, we can see that there are more taxi transactions in Brooklyn, Queens and Manhattan, the pattern of monthly transactions resembles more the pattern of New York City as a whole.

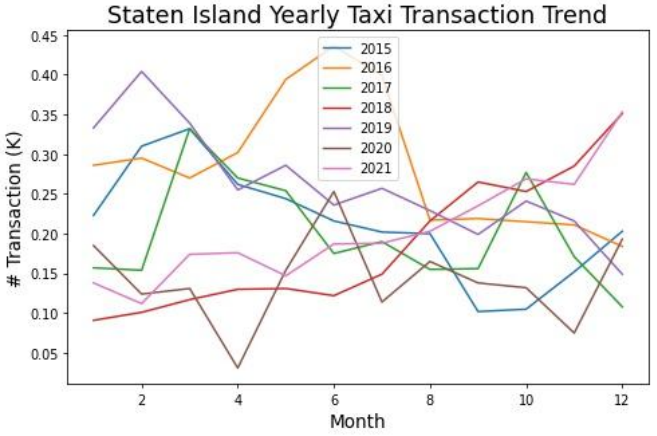
(from New York City Taxi and Limousine Commission)


Brooklyn Yearly Taxi Transaction Trend
Transaction (K)
Month
2015
2016
2017
2018
2019
2020
2021

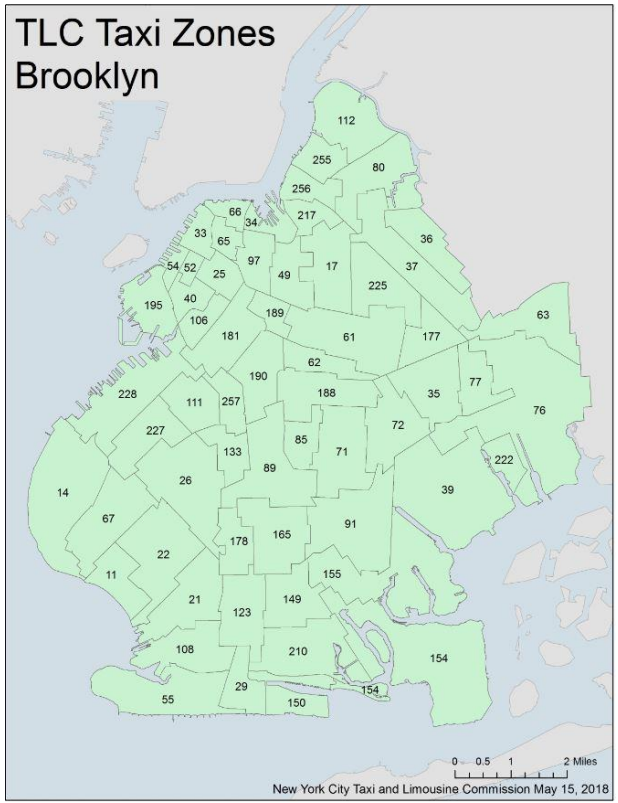

Manhattan Yearly Taxi Transaction Trend
Month
2015
2016
2017
2018
2019
2020
2021

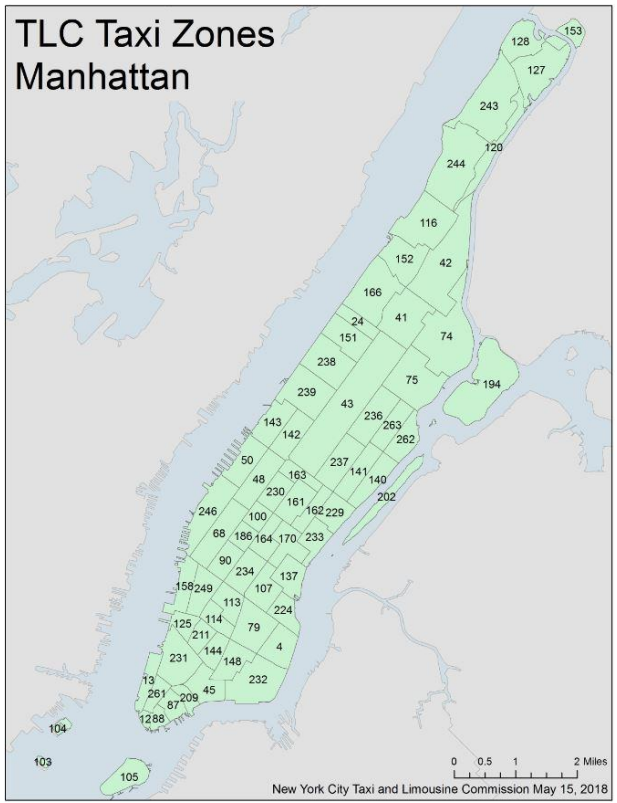

Queens Yearly Taxi Transaction Trend
Month
2015
2016
2017
2018
2019
2020
2021

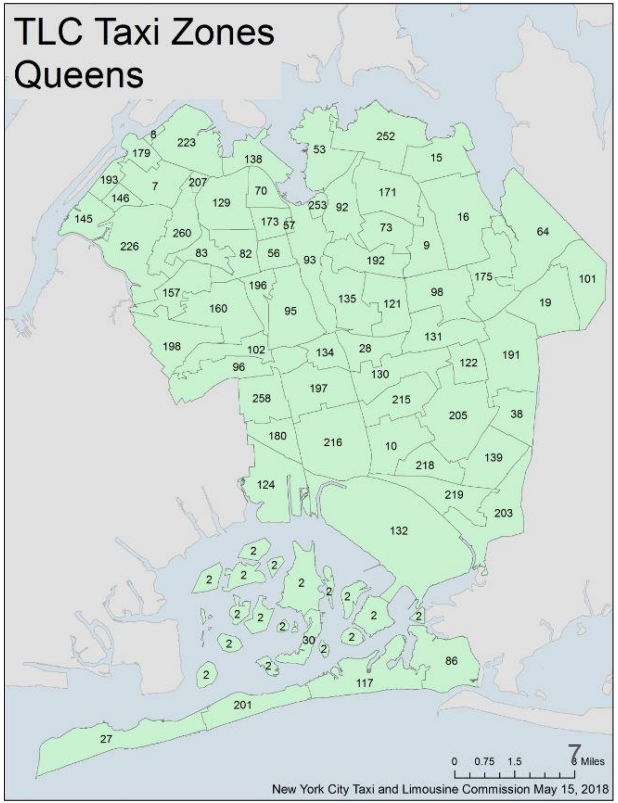

Bronx Yearly Taxi Transaction Trend
Transaction (K)
Month
2015
2016
2017
2018
2019
2020
2021


Staten Island Yearly Taxi Transaction Trend
Transaction (K)
Month
2015
2016
2017
2018
2019
2020
2021

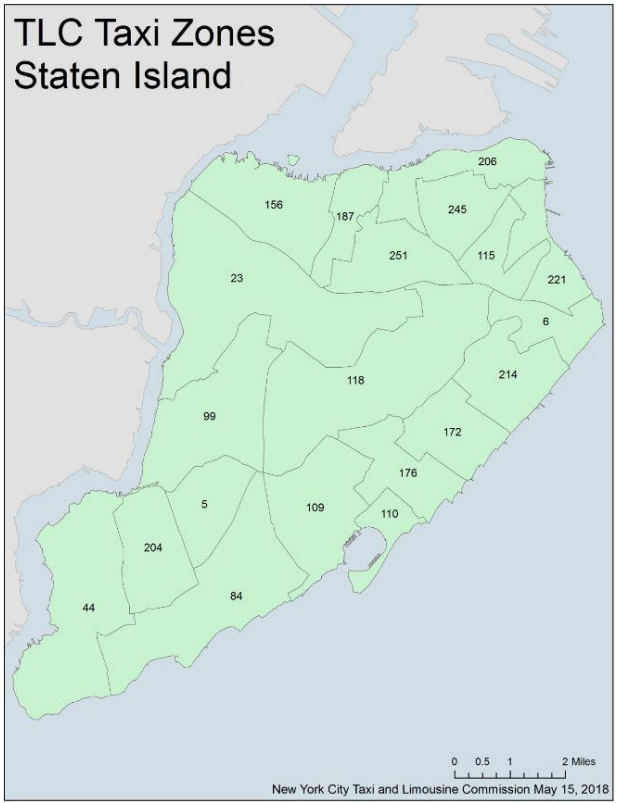
Taxi Zones I

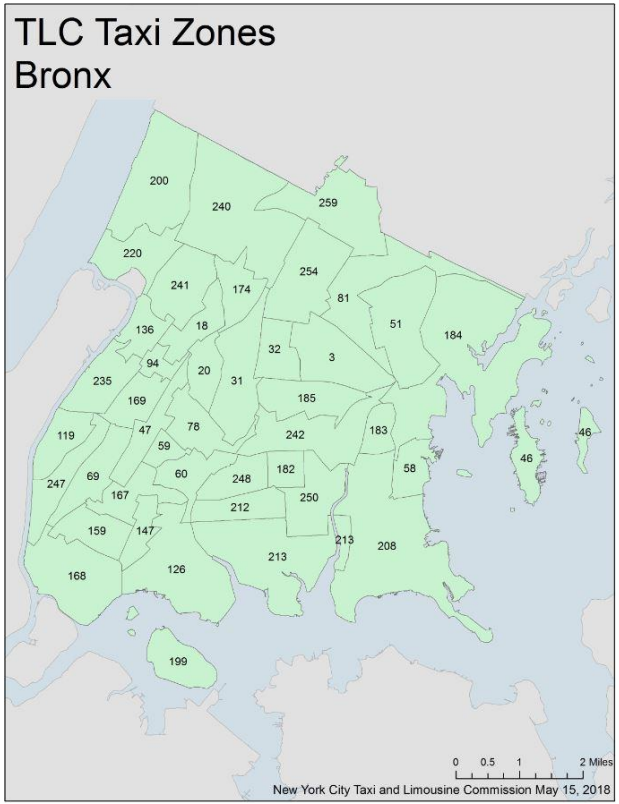

TLC Taxi Zones
Brooklyn
0 0.5 1 2 Miles
New York City Taxi and Limousine Commission May 15, 2018


TLC Taxi Zones
Manhattan
0 0.5 1 2 Miles
New York City Taxi and Limousine Commission May 15, 2018


TLC Taxi Zones
Queens
0 0.75 1.5 3 Miles
New York City Taxi and Limousine Commission May 15, 2018

Taxi Zones II





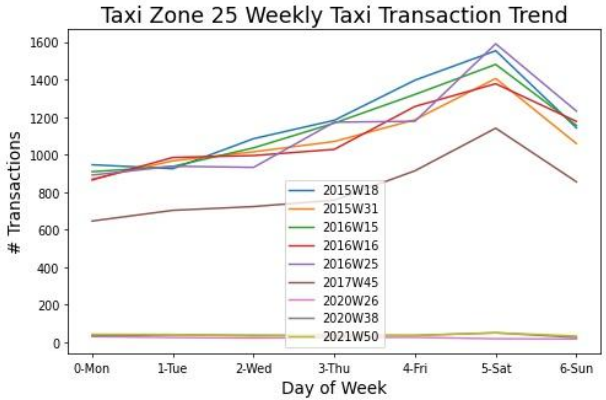
Categorizing Taxi Zones

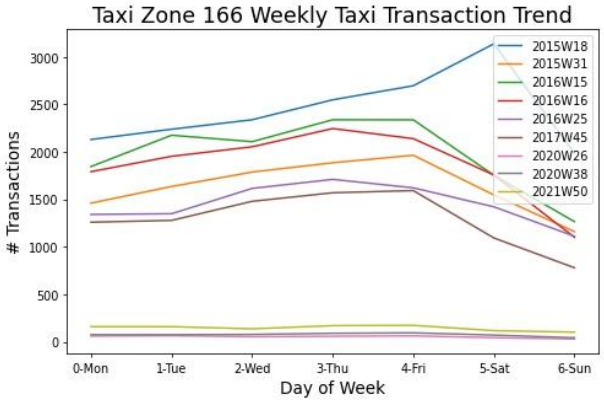
Taxi zones are categorized into 4 tiers according to average annual taxi transactions. Taxi zones above 100,000 transactions annually are categorized as Tier 1, above 20,000 transactions as Tier 2, above 10,000 as Tier 3 and the rest are considered Tier 4. Top Tier zones have greater influence on hailing companies revenue. **2 zones were randomly selected for from each Tier for further analysis and future transaction prediction task.**

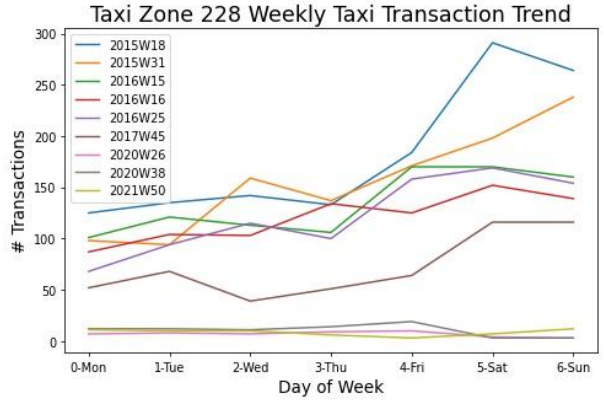
	Tier 1	Tier 2	Tier 3	Tier 4	Total
Brooklyn	12	15	11	23	**61**
Queens	9	8	4	48	**69**
Manhattan	7	6	0	54	**67**
Bronx	0	6	17	20	**43**
Staten Island	0	0	0	20	**20**
Total	**28**	**35**	**32**	**165**	**260**

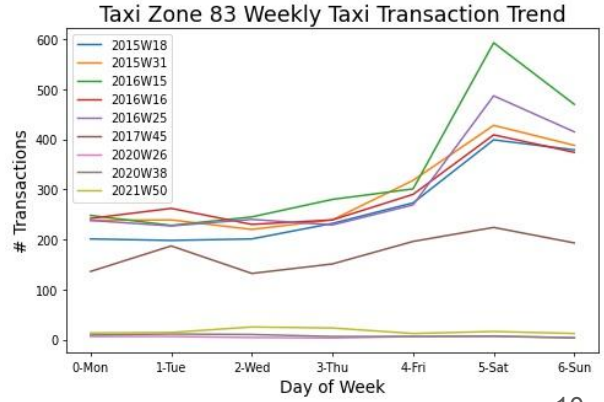
Weekly Trend I

From Tier 1 (25, 166) and Tier 2 (228, 83) zones, we can observe some weekly trends (peaking on Friday and Saturday). Again, the trend disappeared almost completely in 2020-21.





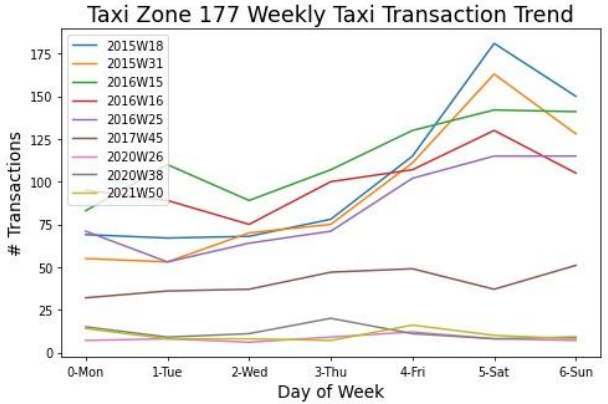


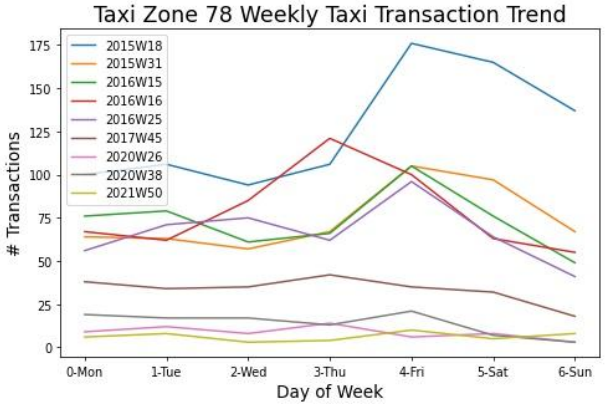


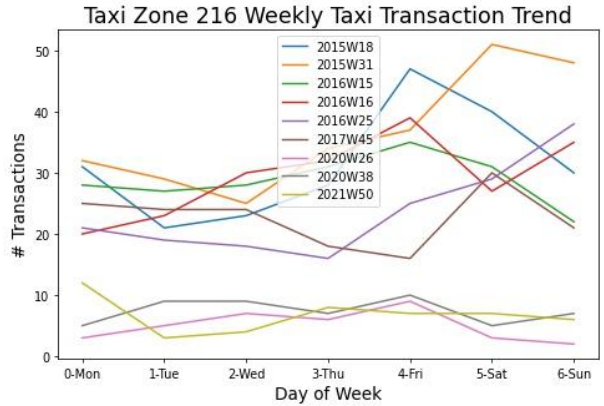
The trend is observed from 10 randomly selected weeks from 2015-2021.

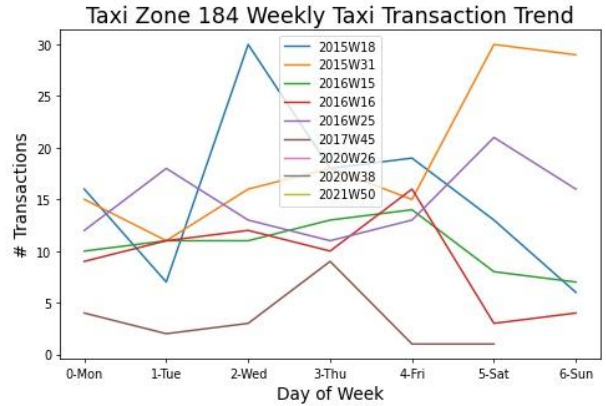
Weekly Trend II

Trends from zones in Tier 3 (177, 78) and Tier 4 (216, 184) are less clear. However, on can still observe similar trends of transactions every week in Tier 3 zones.









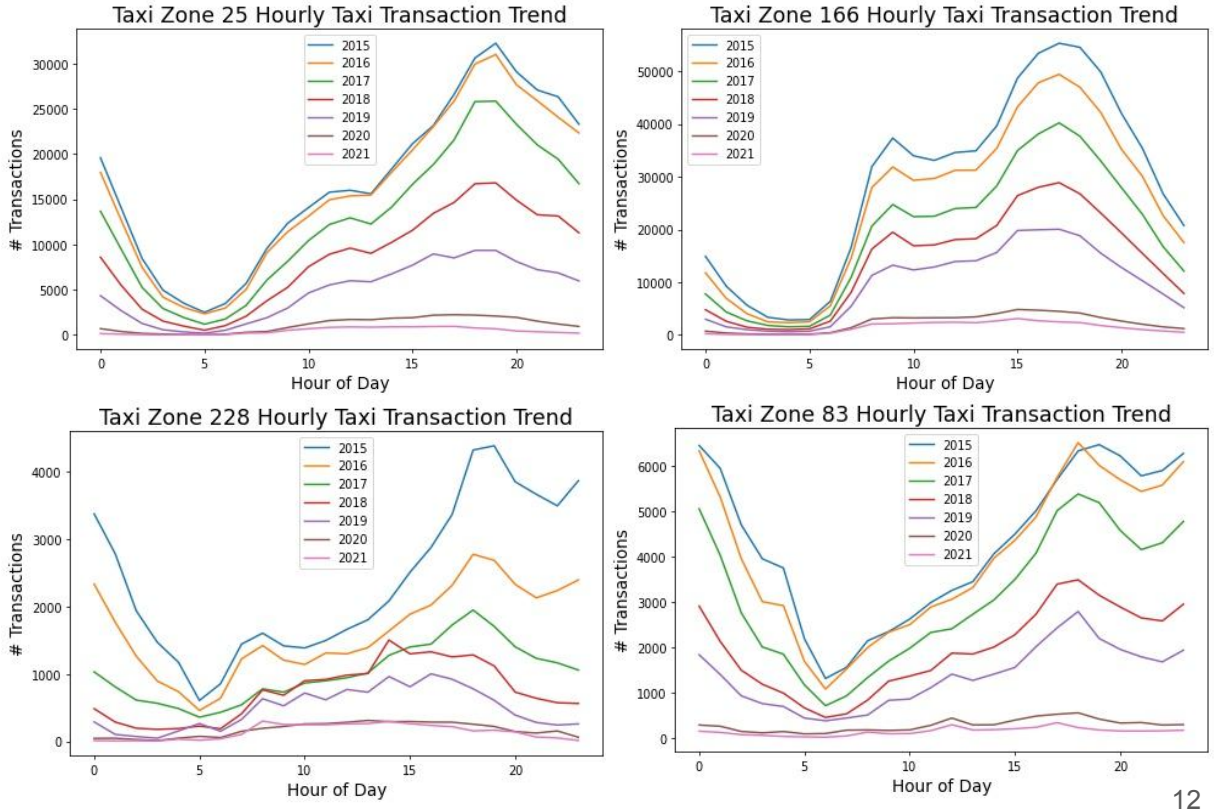
The trend is observed from 10 randomly selected weeks from 2015-2021.

Daily Trends I

If we aggregate daily transactions data every year, trends can be more easily seen.

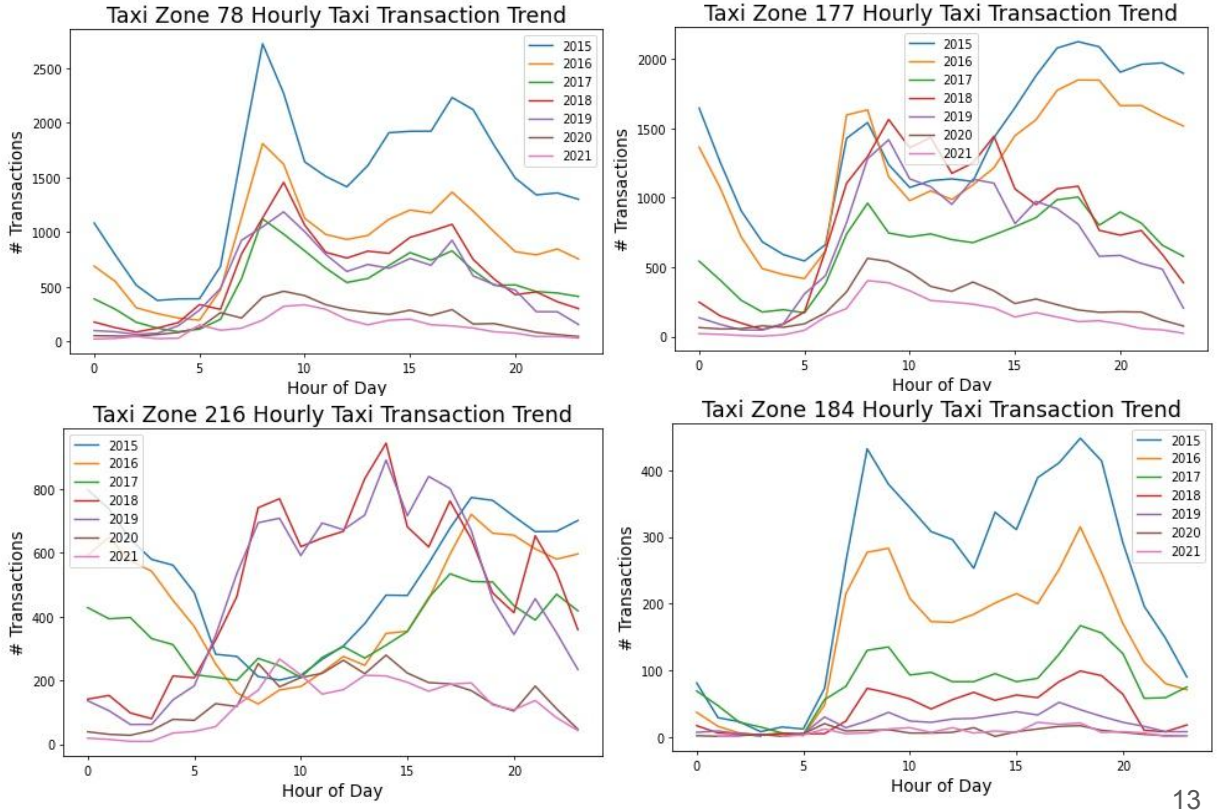
For zone 25, 228 and 83, the peak is in the evening.

For zone 166, there seems to be peaks in the morning and evening.



Daily Trends II

Daily trends can even be seen on Tier 4 zones. For example, Zone 184 show similar trends of hourly transaction in different years. Again, trends almost disappear in 2020, 2021.



Future Prediction Task Framed

- **Target**

  To predict hourly number of transactions in 8 taxi zones (2 from each Tier) using December 2019 data. Transaction records before time points to be predicted are used by models to form predictions.

- **Models**

  A. **Time Series Models**: use previous trends to predict future number of transactions.
  B. **Machine Learning (ML) Models**: use previous trends and other features (found during exploratory analysis) to predict future number of transactions.

Time Series Models

- **Exponential Smoothing**: predict value at time *t* based on values previous data, weighing less to values at earlier time points.

$$s_t = \alpha x_t + (1-\alpha)s_{t-1} = s_{t-1} + \alpha(x_t - s_{t-1}).$$

- **Autoregressive (AR)** model: use numbers from t-p to t-1 to predict the value at time t.

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

- **Moving Average (MA)** model: puts weights on deviation (error) from average (from t-q to t-1) to predict the deviation from average at time t.

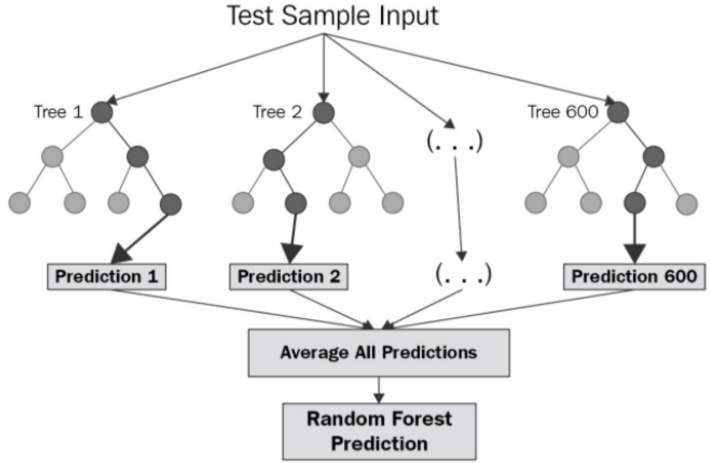
$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

- **Autoregressive Moving Average (ARMA)**: combines AR and MA models.

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

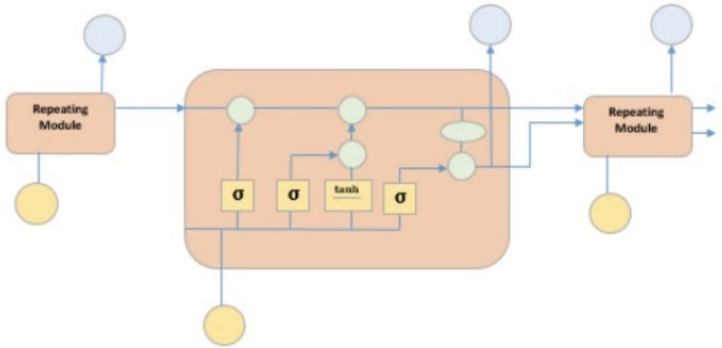
Machine Learning Models

- **Random Forest Regression**: use bagging to derive multiple tree combinations to generate multiple prediction values. The values averaged would the final prediction.



Source: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

- **Long Short Term Memory (LSTM)**: train weights to specify the importance of features at previous time points on prediction target. The model can incorporate both static and dynamic (time series) features for prediction.



Source: https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm

Performance

Random forest regression yielded the least RMSE* on prediction hourly number of transaction in the 8 chosen taxi zones. Overall, the RMSE for Tier 1 zones are the largest.

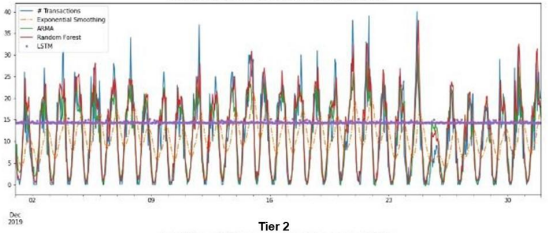
	Exponential Smoothing	Autoregressive Moving Average	Random Forest Regression	Long-Short Term Memory
Overall	**47.6**	**11.0**	**9.6**	**195.3**
Tier 1	183.0	38.8	32.5	250.2
Tier 2	4.3	2.9	3.1	157.6
Tier 3	2.4	1.9	2.0	176.6
Tier 4	0.7	0.6	0.8	196.7

* root mean squared error $\sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$

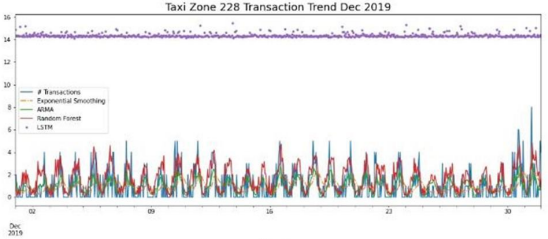
Results and Interpretations

- **Pattern of taxi transaction exists**: From exploratory analysis, we see that Overall Annual transactions decrease from 2015 to 2019 around 15% - 30%. Since 2020, number of transactions dropped drastically. Before 2019, number of taxi transactions fluctuates monthly, daily and hourly with a pattern.
- **Random forest regression improved prediction:** leveraging features besides transaction numbers at previous time points, random forest regression algorithm lowered prediction error (measured by RMSE) by 80 % from the baseline model of exponential smoothing. The improvement is more significant in taxi zones with larger amount of transactions.
- **Issue of LSTM**: LSTM did not perform well may be due to the nature of data. LSTM allocates weights on number of transactions at previous time points. However, previous transactions can be positively and negatively correlated with the number of transactions now, depending on the time. For week, day and hour data, one hot encoding makes data sparse, which can lead to poor performance on the prediction task.

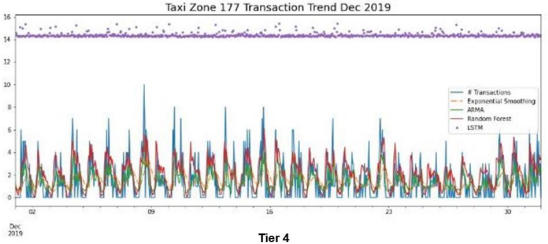
Tier 1



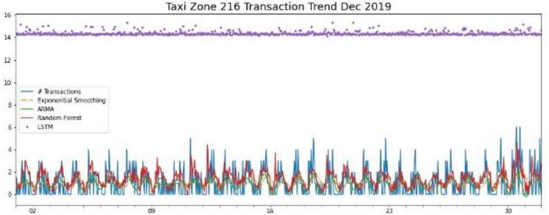
Tier 2



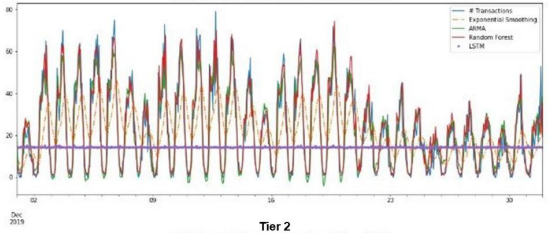
Tier 3



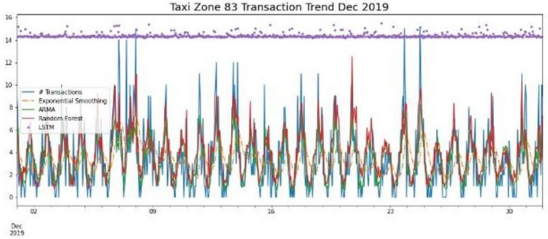
Tier 4



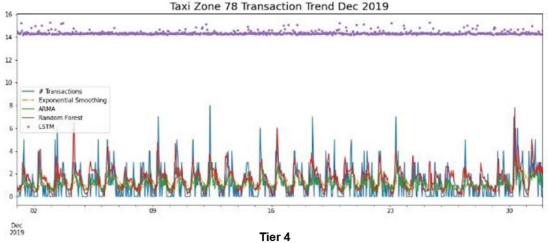
Tier 1



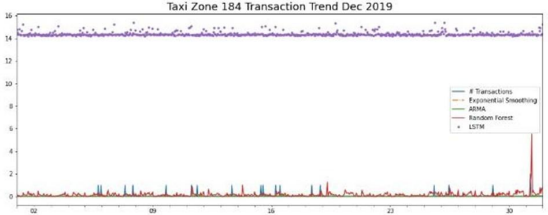
Tier 2



Tier 3



Tier 4



Next Steps

- **Include other features**: e.g. weather, temperature, day/ night etc.
- **Leverage other algorithms**: Other models can be leveraged or combined for future taxi transaction prediction.
- **Scale up**: Include more data for model training, predict all locations
- **Modify model on the go:** Different patterns might occur as time goes. Thus, the model and weights should be constantly adjusted to optimize prediction results.

Backup

Features Used

Month, Day of week, hour of day, number of transactions in the taxi zone and 2 neighboring taxi zones (t-1 ~ t-6).