"FORECAST PROTOTYPE FOR STATION CAPACITY "

Time Series and Forecasting
University of Chicago
MSCA
2022

Akhir Syabani
Ryan Liao
Mauro Ballesteros

Index

Today we will cover

1. Problem Statement
2. Assumptions and Hypothesis about Data
3. Data Properties and Exploratory Data Analysis
4. Data Processing and Transformations
5. Feature Engineering
6. Proposed Models
7. Results

1. Problem Statement

The Bike sharing Rebalancing Problem with Stochastic Demands is a variant of the one-commodity many-to-many pickup and delivery vehicle routing challenge, and station location strategy; where demands at each station are represented by random variables, with:

- Associated probability distributions, that depend on stochastic scenarios.
- The Modelling of Outlier Situations

.

Use Case

- To Provide real-time short term (*within hours, accurate to 10min*) forecasting of station dock balance to help users make decisions when docking/ renting.

By providing a real-time short-term forecast of traffic and dock availability, we eliminate the major pain points of current users:

- Can't find bike to ride when arrived, and
- No space to park.
  - *This has being the #1 reason people are not subscribing to divvy's services, and a high churn rate.*

- To Provide Middle and Long-term (Within years, accurate to days) forecast of station dock balance to help divvy make decisions to:
  - Implementing Temporary Stations
  - Routine Intervention Designs

By having a mid-term forecast, divvy can pre-arrange resources, thus cut down costs maneuvering the bikes. More importantly, by having long term forecast, divvy can implement a more efficient growth planning based on the city's usual mobility challenges and user demand.

2. Assumptions and Hypothesis about data

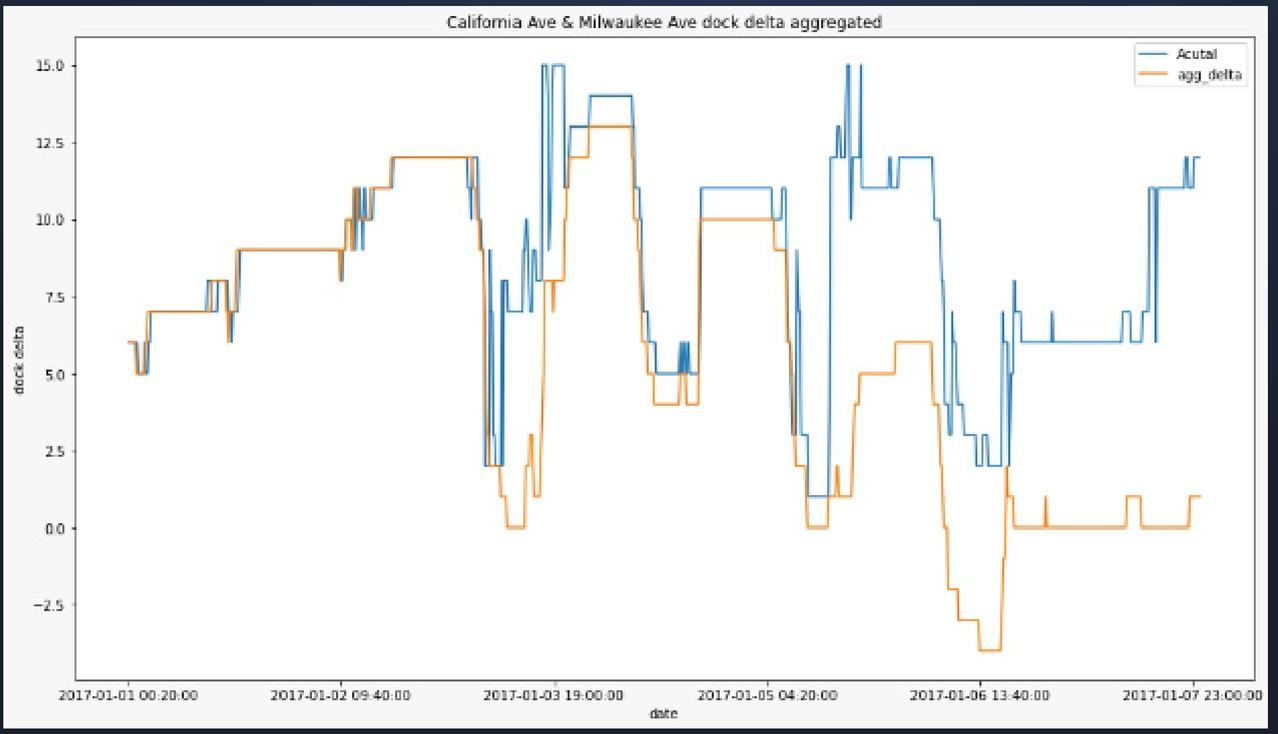
The data we got to model is Station-Dock Availability

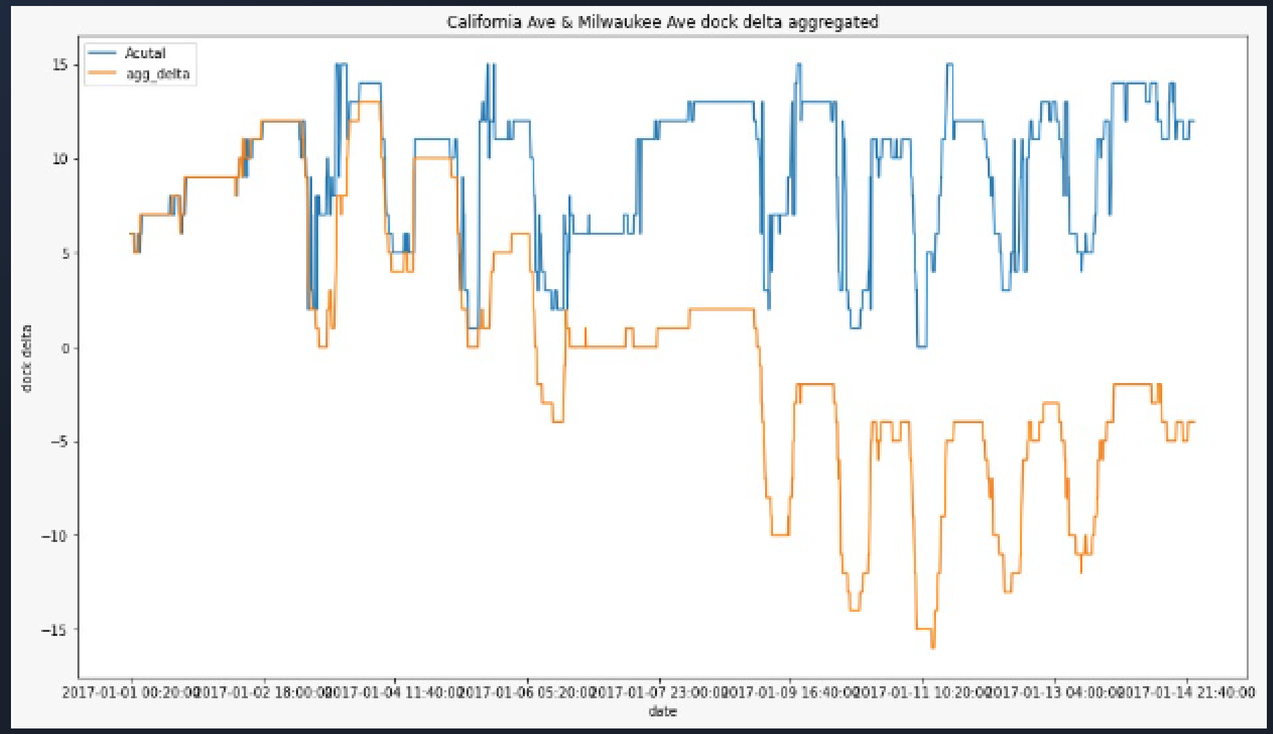
- However, it was not available (Except for 1 station), so we decided to enginer those by simulating over the Travel History
- Modeling Assumptions:
  - **St = S0+ Cb**
- Definitions:
  - St: True Station Dock, at time t
  - S0: The initial dock availability at time 0
  - Cb: Cumulative sum of Balance change

3. Data Properties and Exploratory data analysis

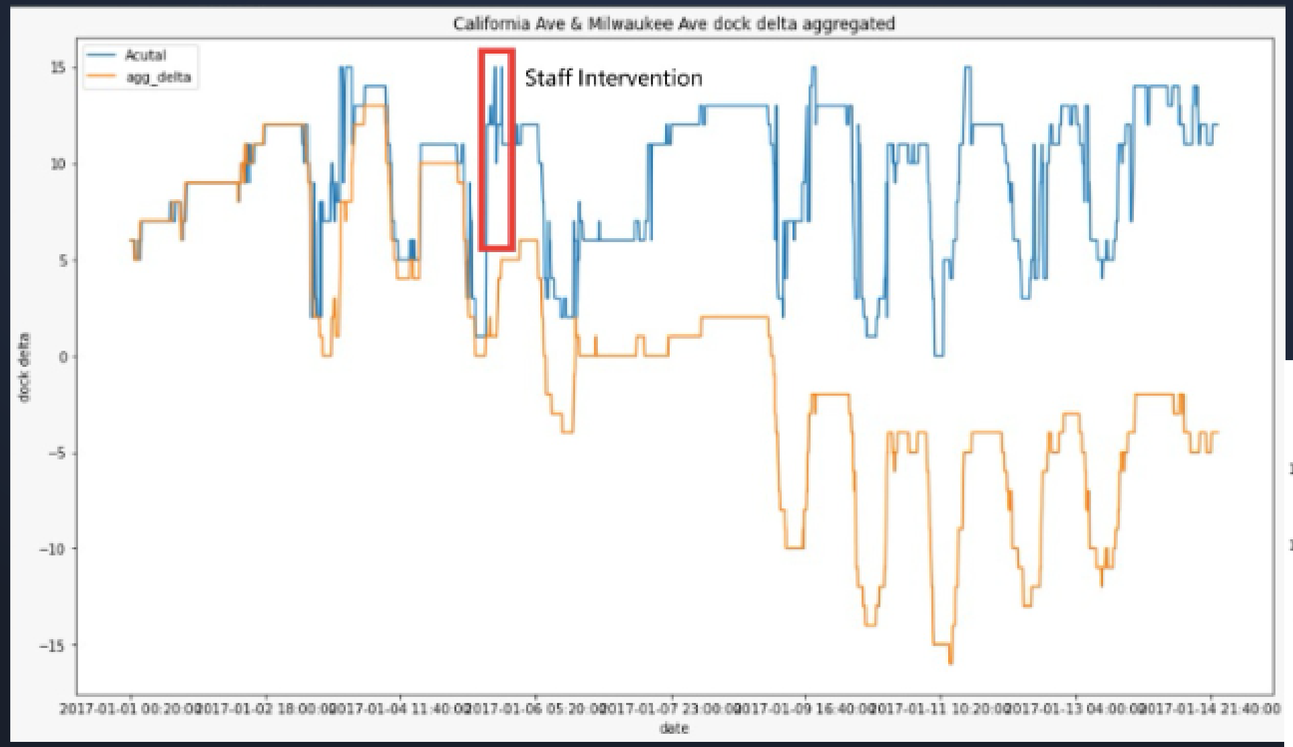
We speculate that Divvy staff are manually adding/removing bikes from the station to mitigate the balancing, however that's not captured by the travel history data. We further decided to ignore the Staff Intervention because that's something Divvy has full control over, and we didn't need to include into our model.

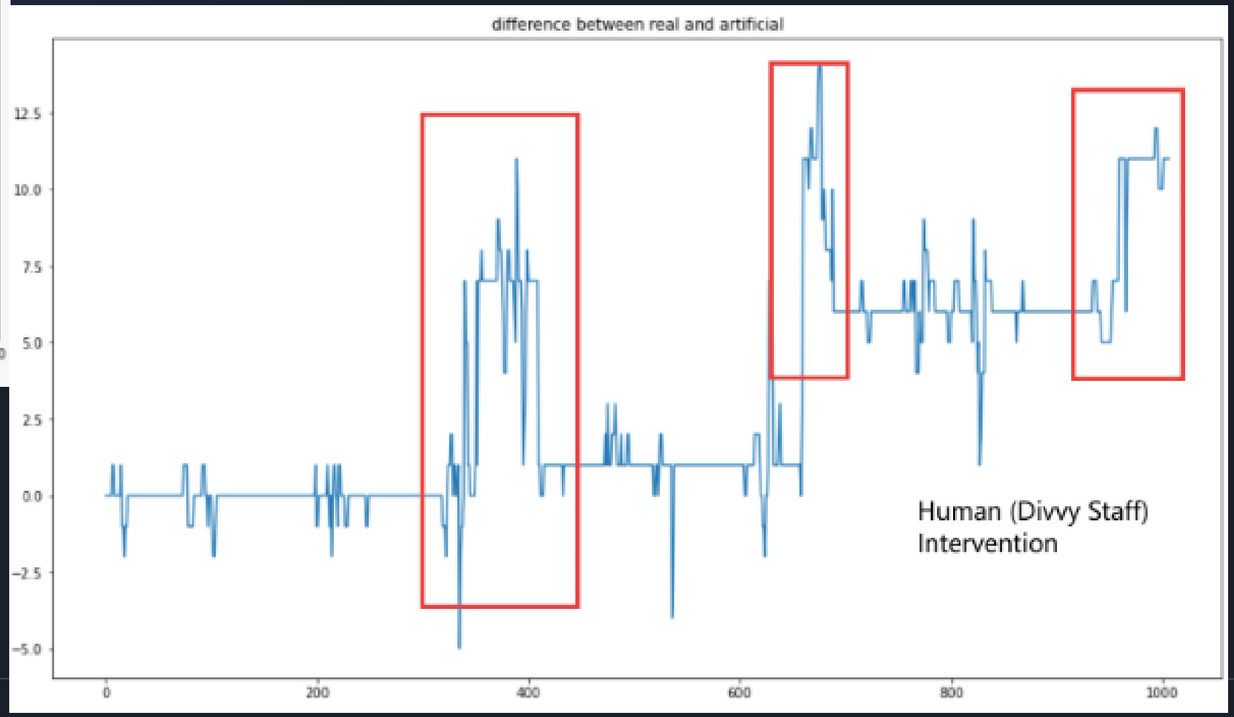
It’s reassuring to observe that our artificial data is parallel with the actual data.





The staff intervention



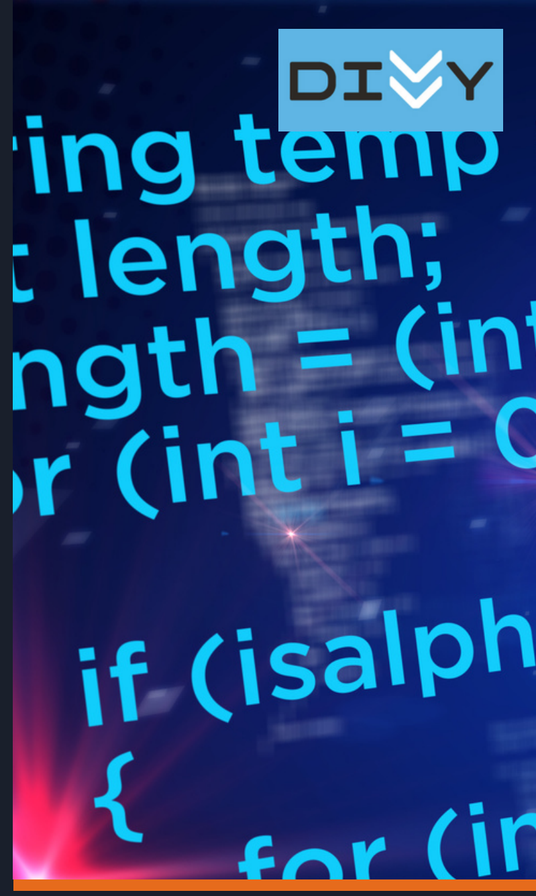


Updated Assumptions and Hypothesis about data

Upon observing the staff intervention, we adjusted our model assumption

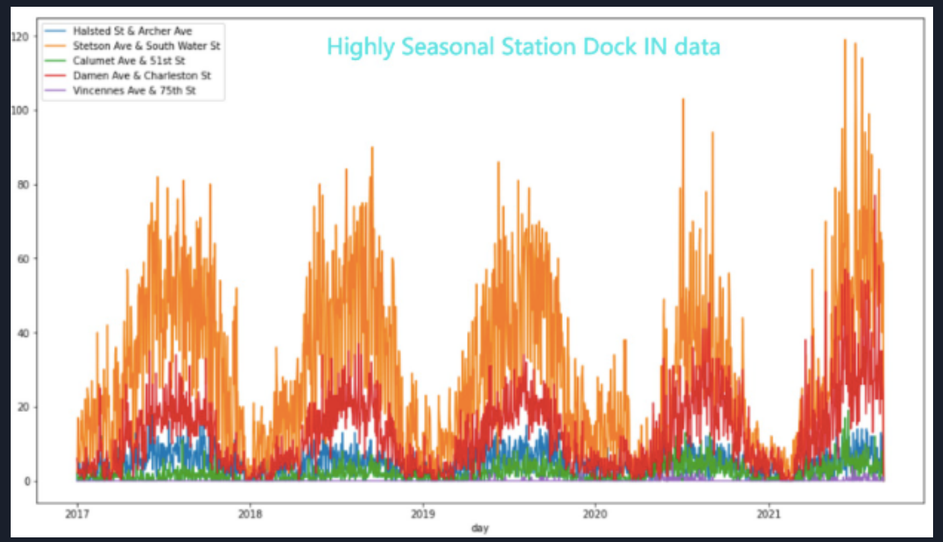
- Modeling Assumptions:
  - St = S0+ Cb
  - St = S0+ Cb + **Ht + Eps**
- New Definitions:
  - Ht: Human (Staff) Intervention at time t
  - Eps: Error
- Decision:
  - Since divvy has full control of Ht, there is not point of forecasting that
  - So we decided to only focus on Cb (cumulative sum of balance)

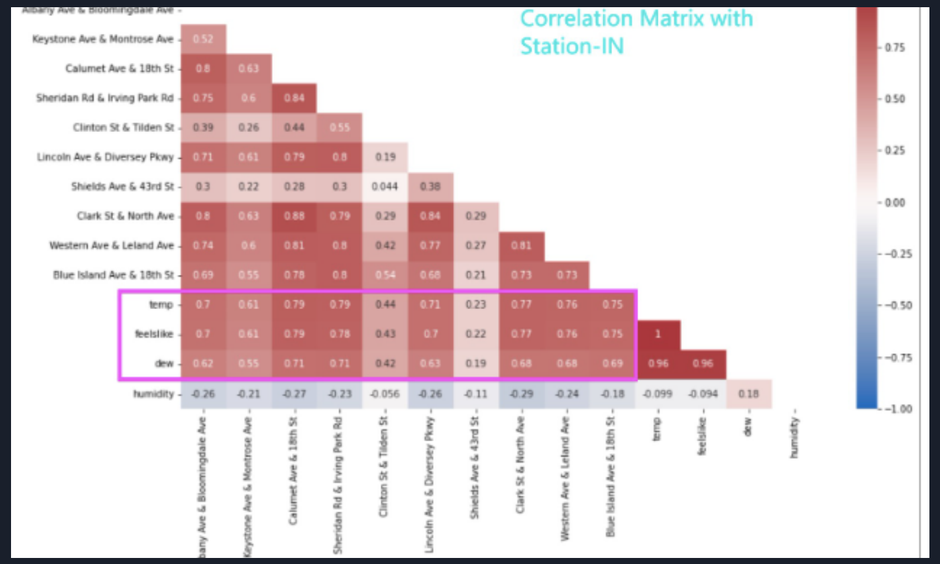
4. Data Transformation & Feature Engineering

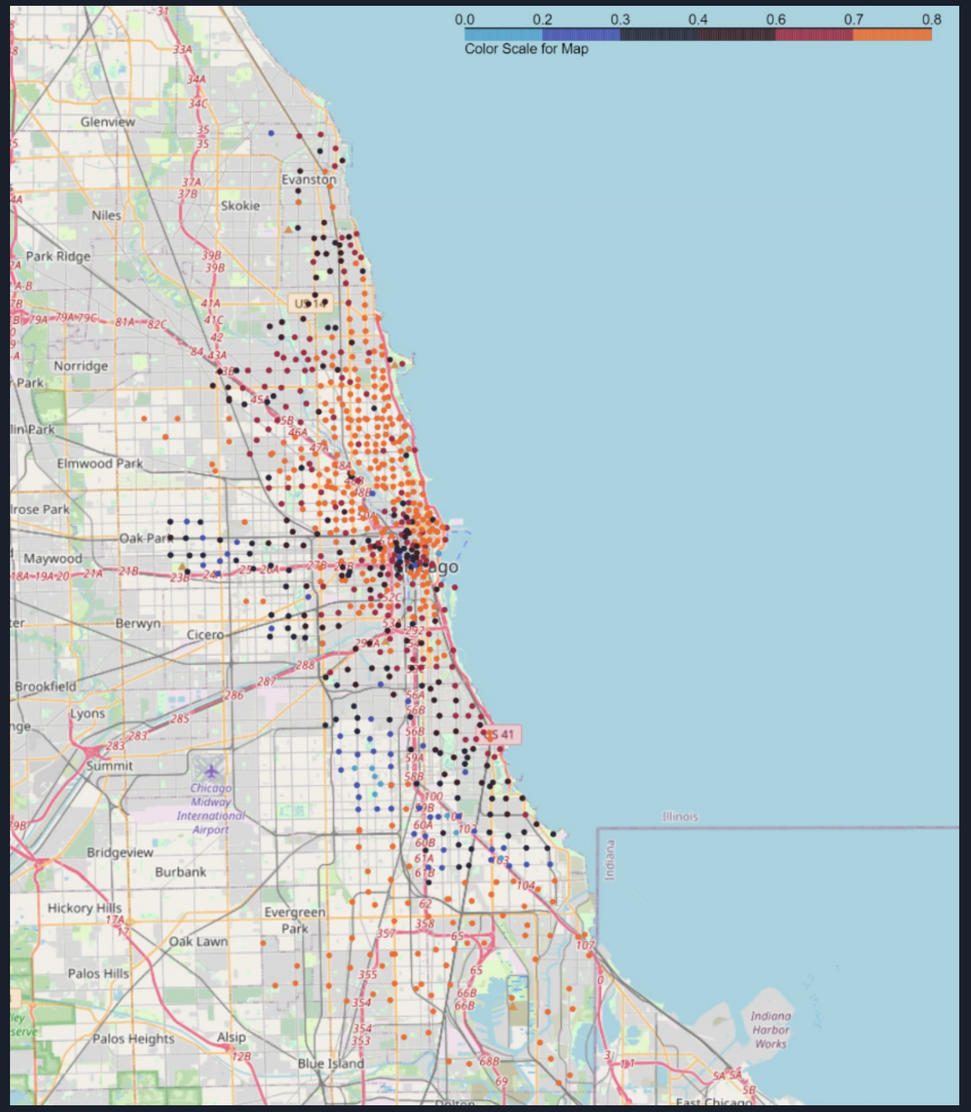


- Data Transformation :
  - Transform structured Rides Data -> Time Series
    - TS-Bike-IN
    - TS-Bike-OUT
    - Balance := In - Out
      - It was agreed to model IN, OUT separately
  - Aggregation (To improve Forecastability)
    - Horizontally (Stations -> Neighborhoods)
    - Vertically (10Min Time Frame -> Hourly)

Exploring Correlation with Weather (When Aggregated Daily):





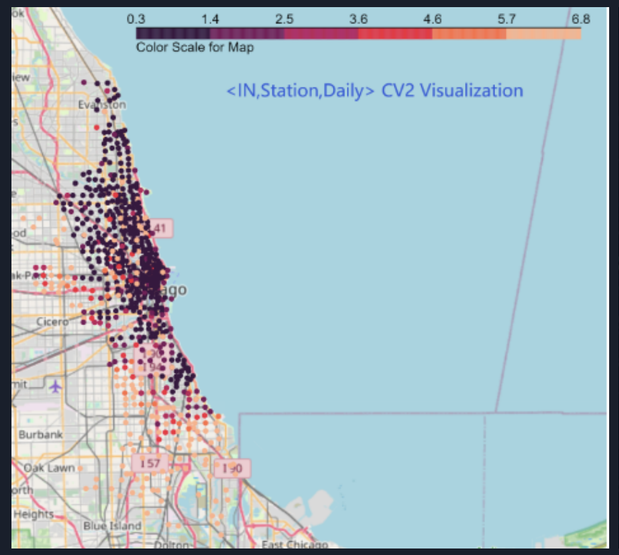


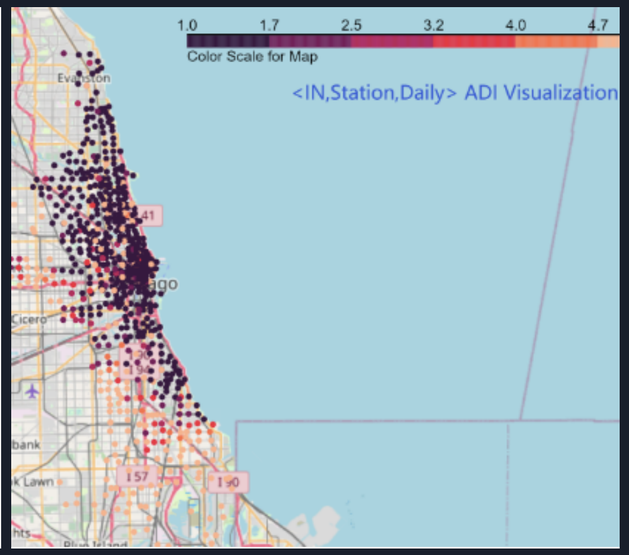
Forecastability (First Look)

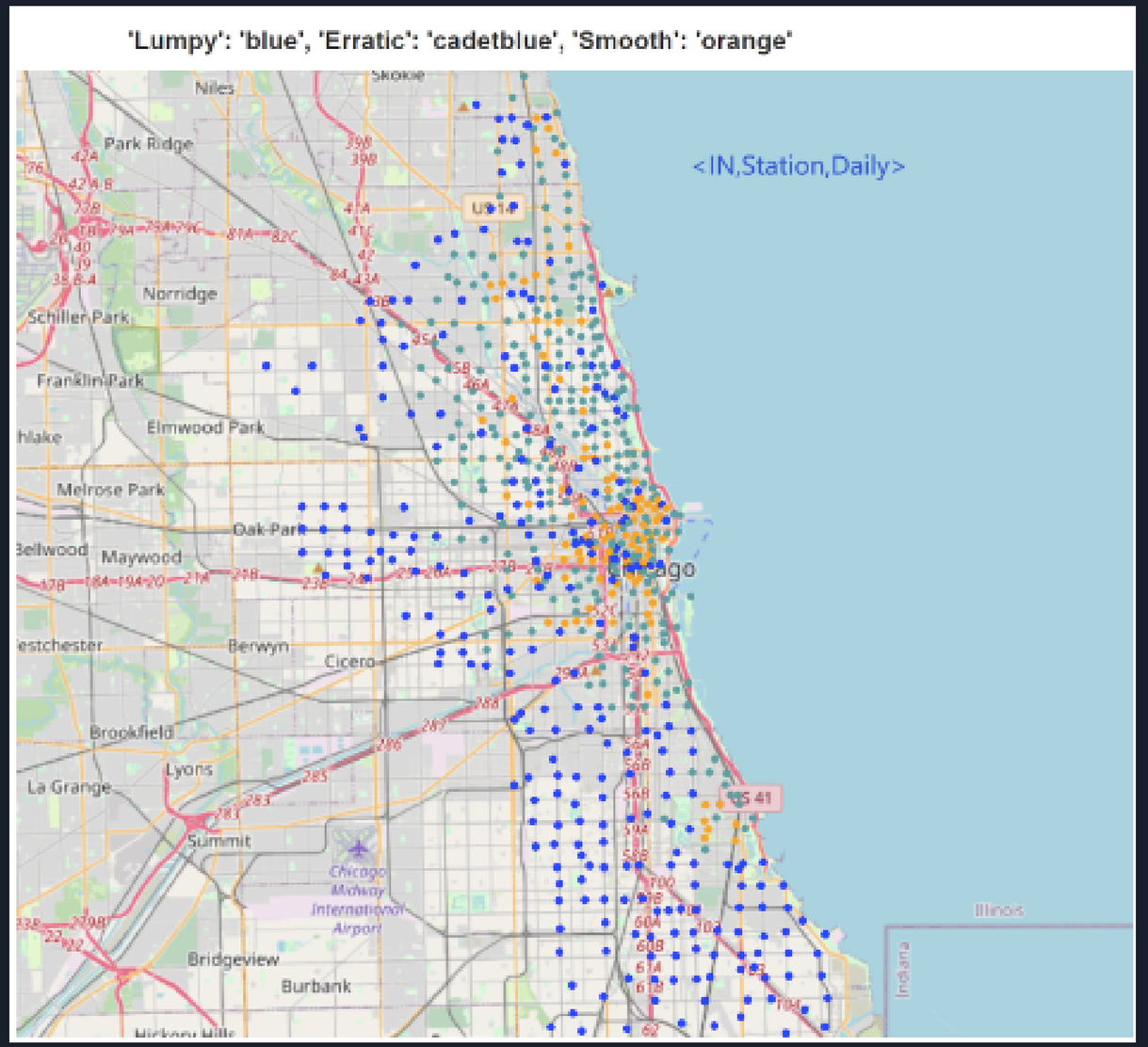
How we attempted to improve TS forecastability:

- Horizontal (Station < Nbh < Community)
- Vertical (10Min < 1 Hour < Day)

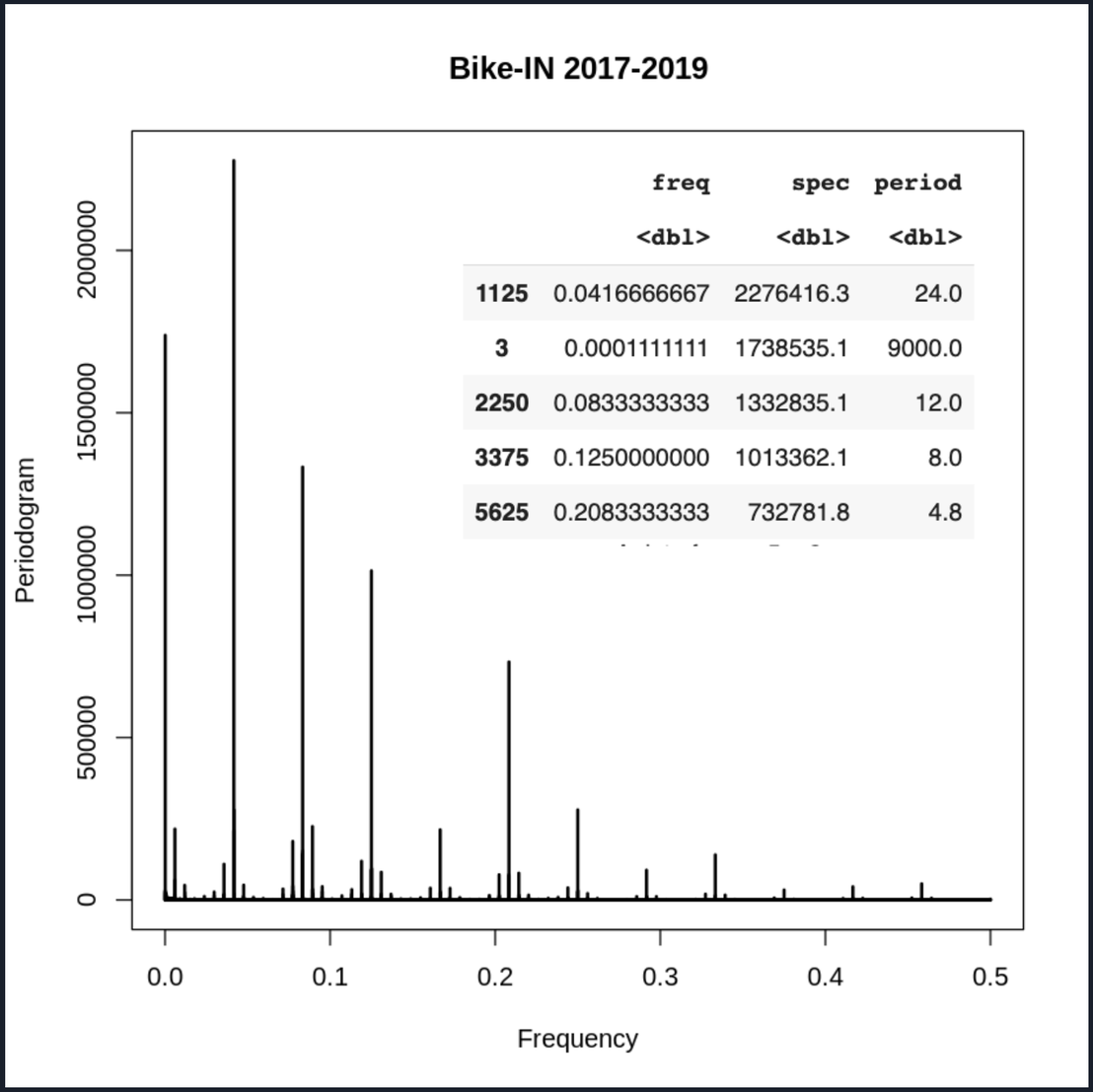
Horizontal / Vertical	Station	Neighborhood	Community
10 Min	Lumpy	Lumpy	Lumpy
1 hour	Lumpy	Erratic	Erratic
1 Day	Smooth	Smooth	Smooth



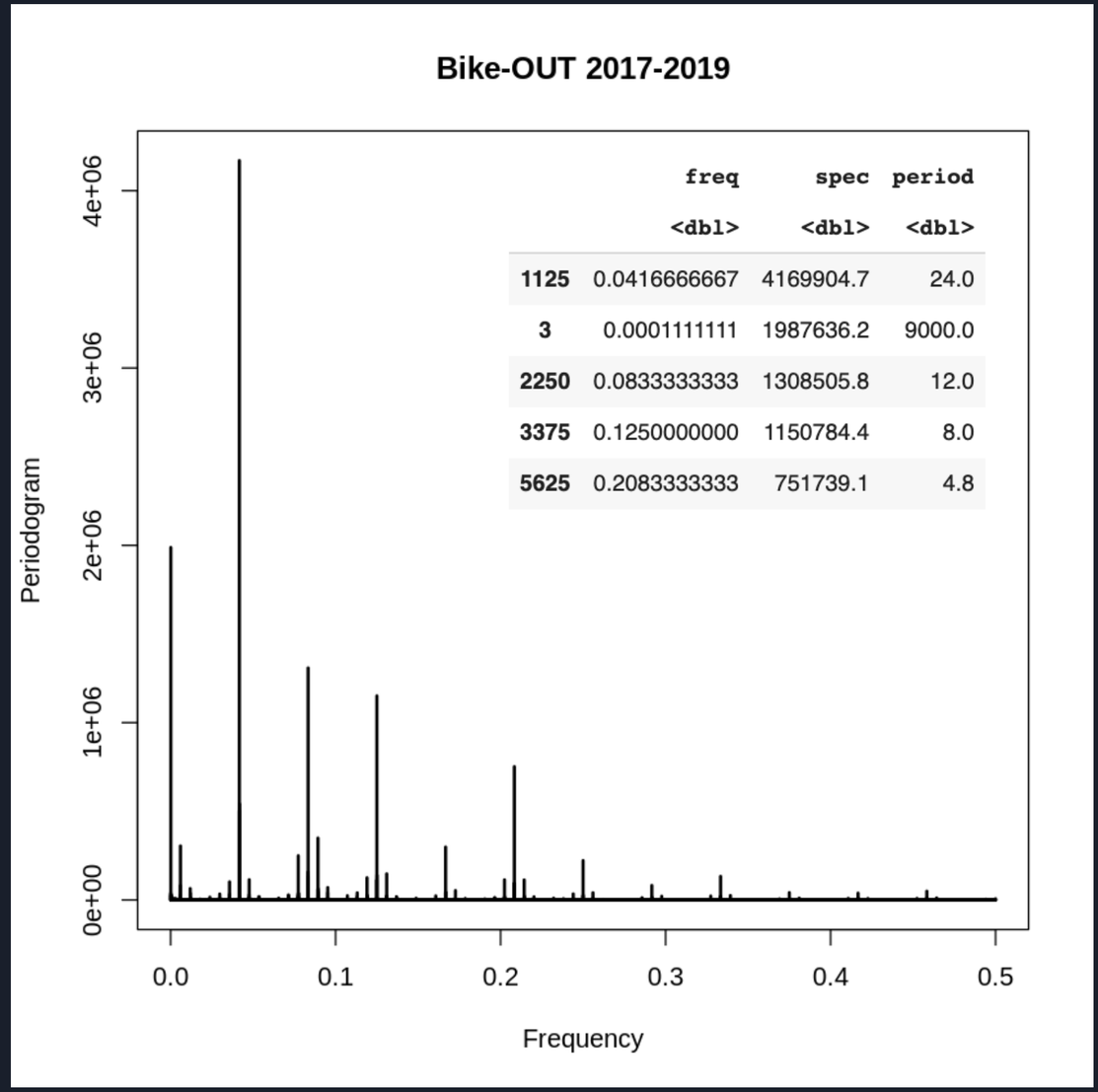




Erratic Data <Neighborhood, Hourly, IN&OUT>

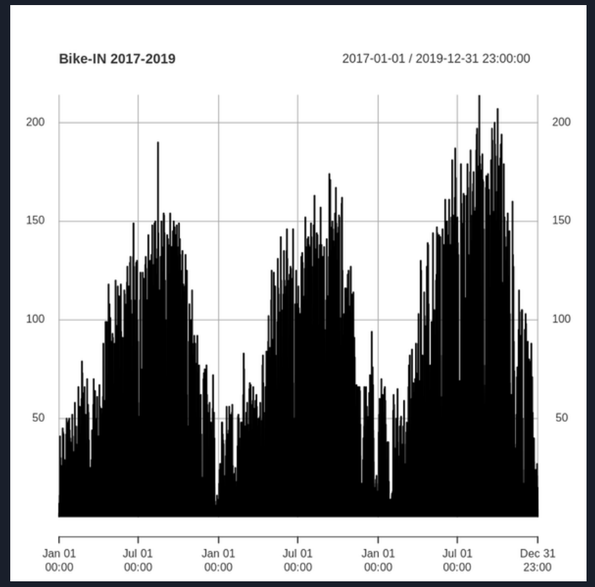


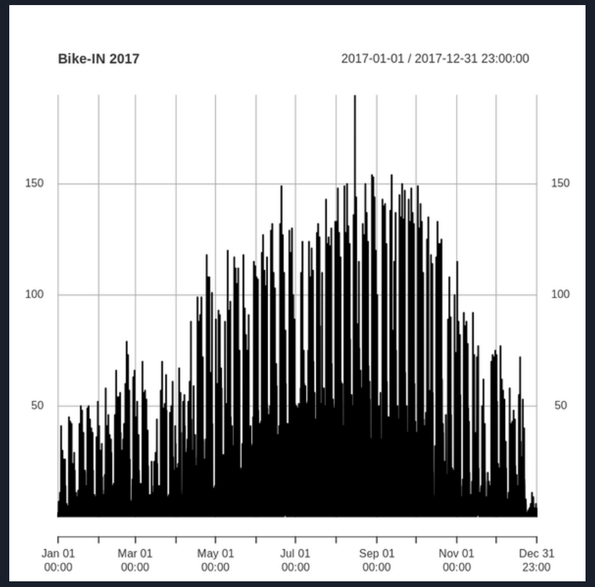
	freq	spec	period
	<dbl>	<dbl>	<dbl>
1125	0.0416666667	2276416.3	24.0
3	0.0001111111	1738535.1	9000.0
2250	0.0833333333	1332835.1	12.0
3375	0.1250000000	1013362.1	8.0
5625	0.2083333333	732781.8	4.8

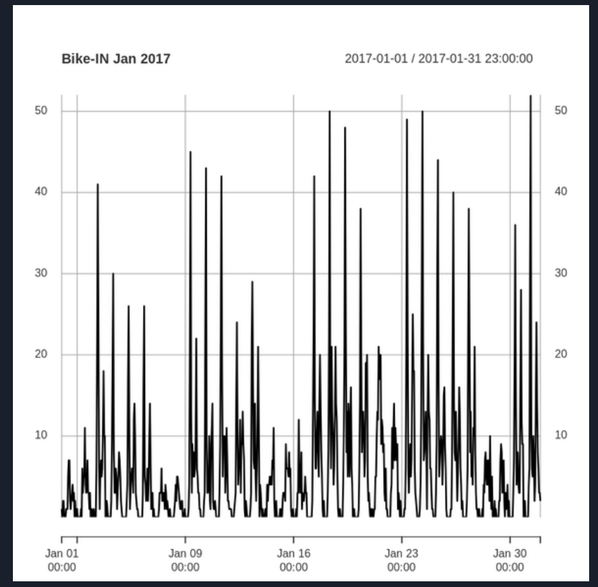


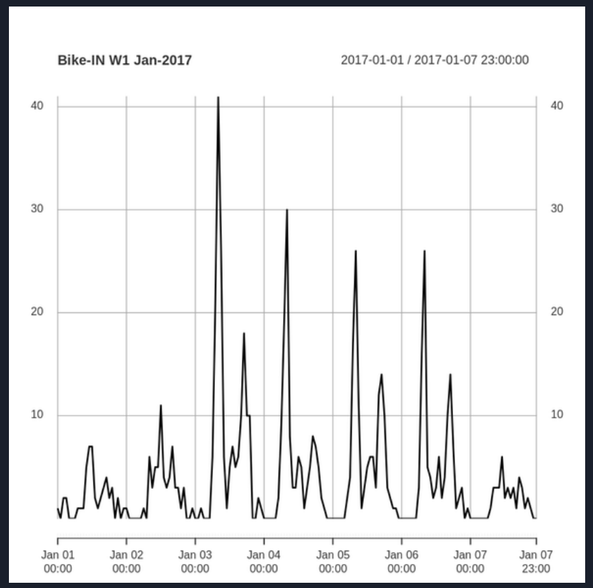
	freq	spec	period
	<dbl>	<dbl>	<dbl>
1125	0.0416666667	4169904.7	24.0
3	0.0001111111	1987636.2	9000.0
2250	0.0833333333	1308505.8	12.0
3375	0.1250000000	1150784.4	8.0
5625	0.2083333333	751739.1	4.8

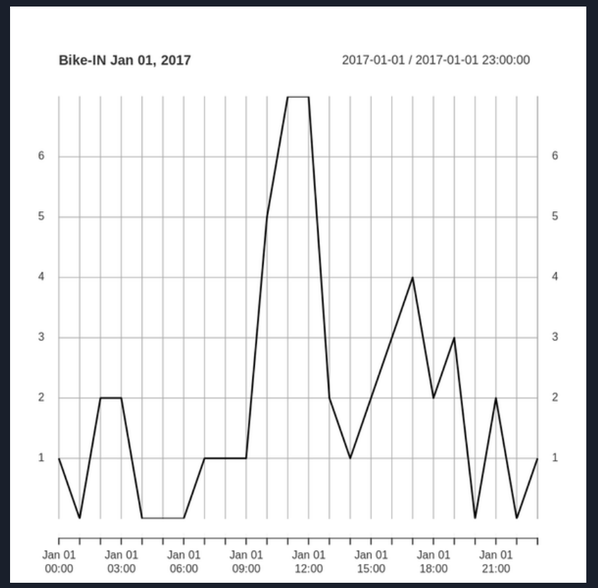
Erratic Data <Neighborhood, Hourly, IN>



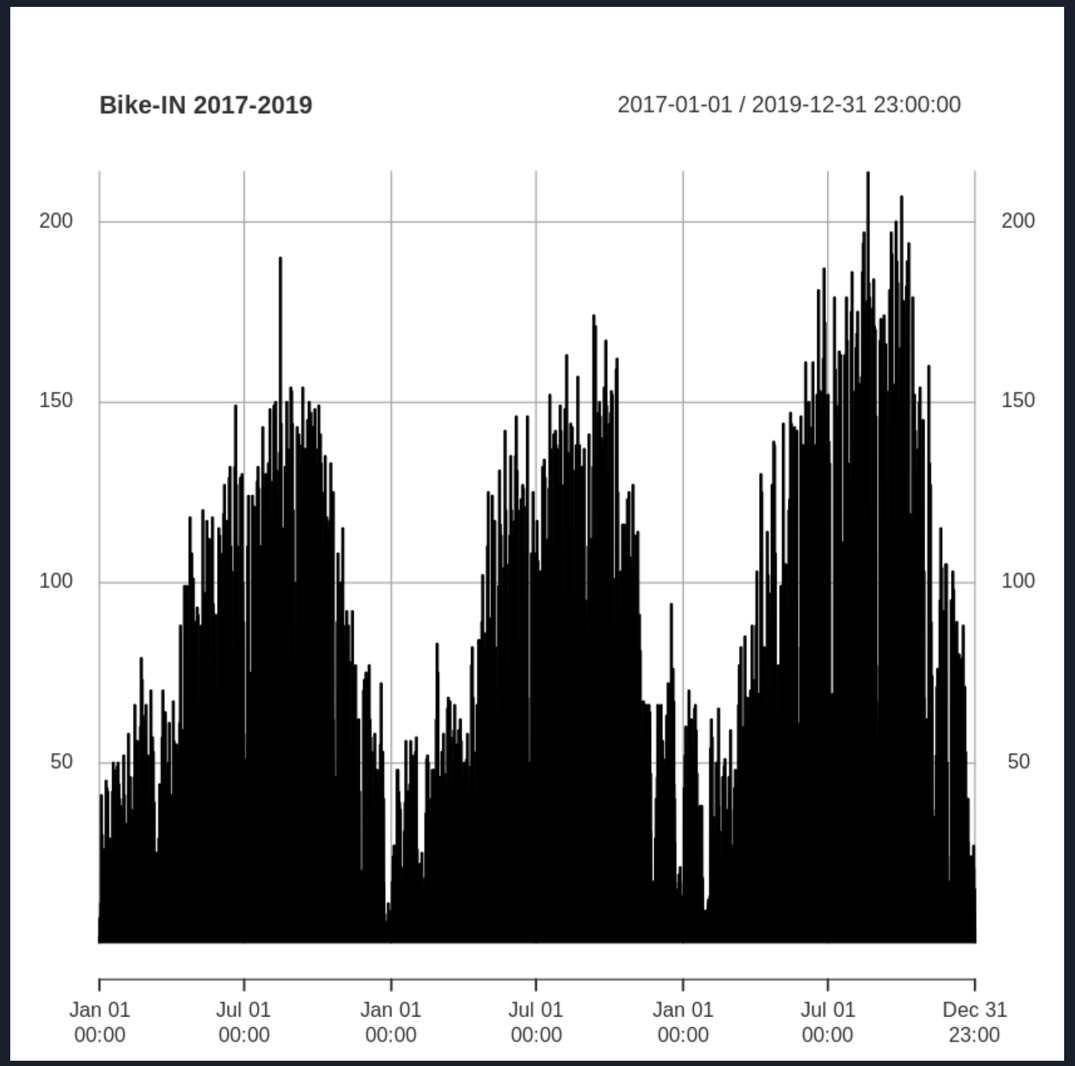


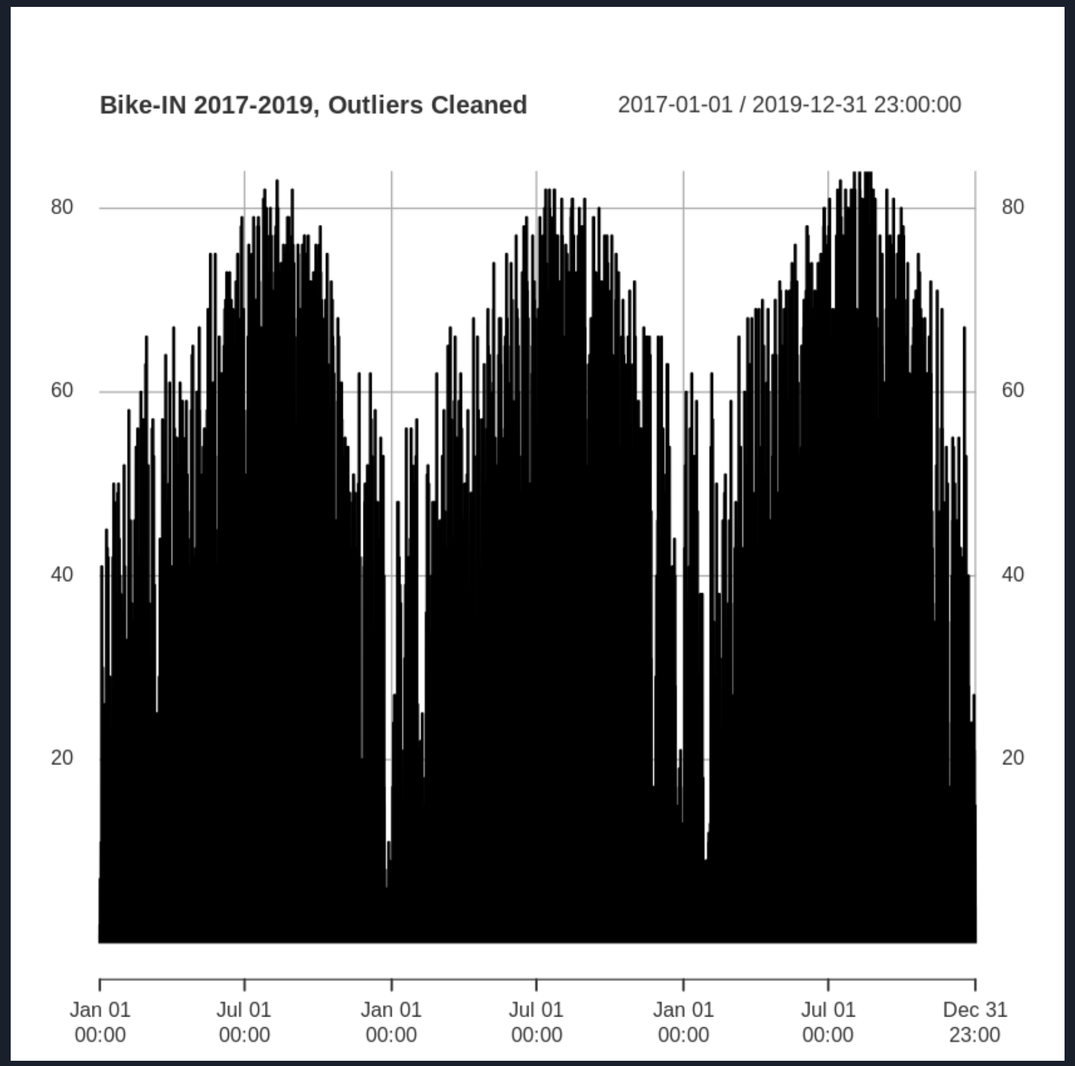




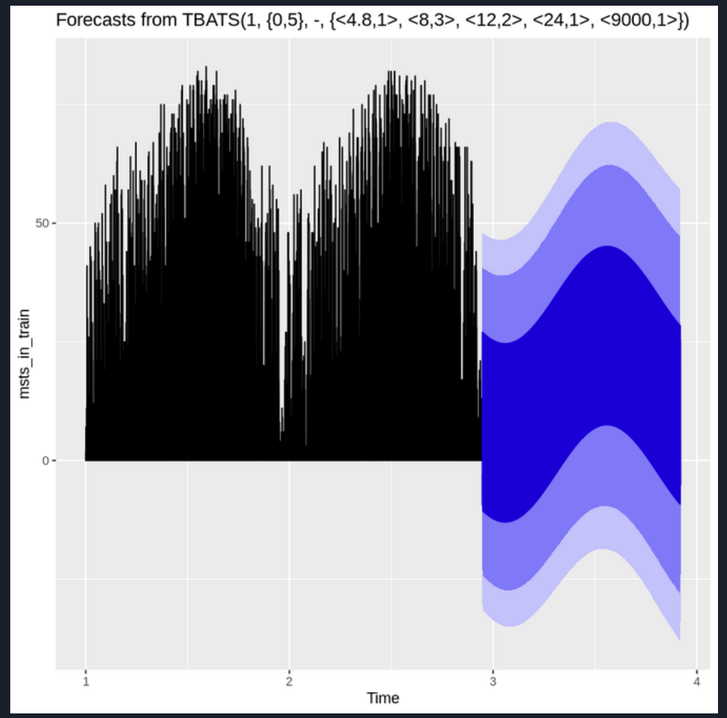


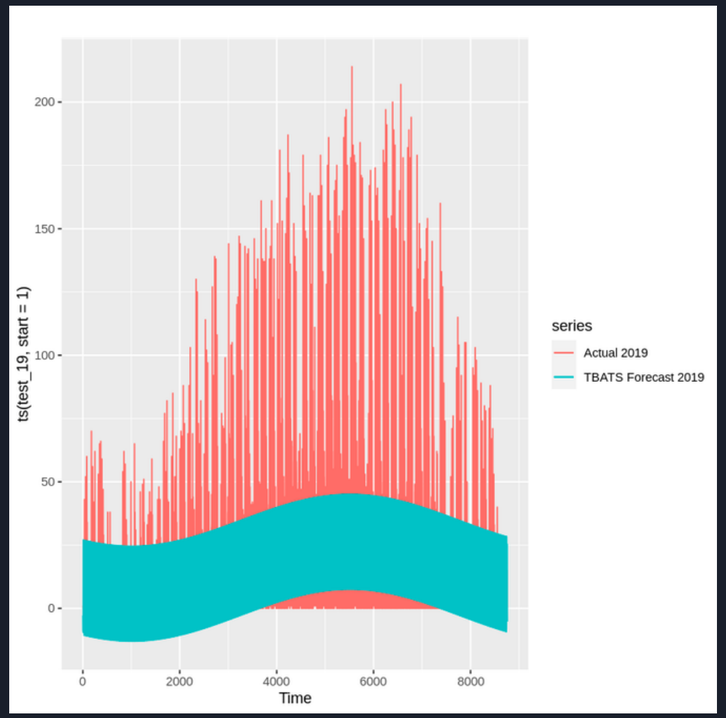
Erratic Data <Neighborhood, Hourly, IN>

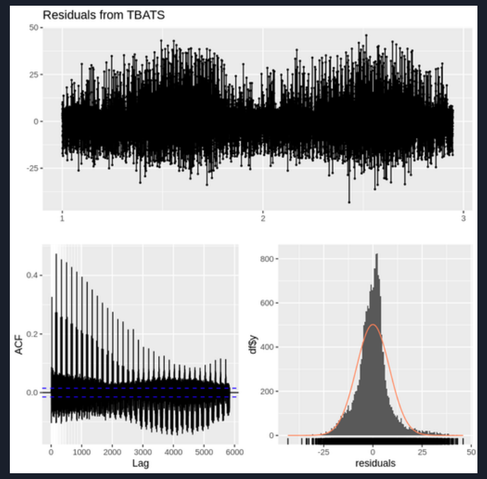


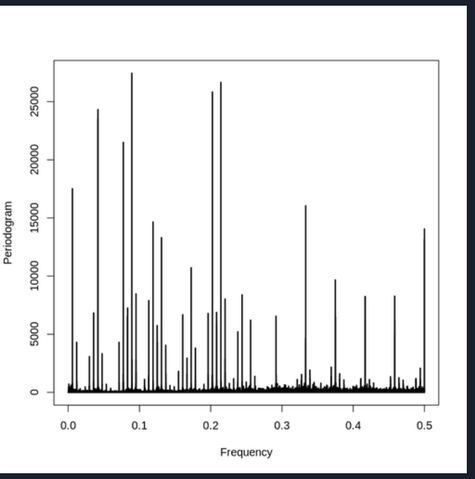


Erratic Data <Neighborhood, Hourly, IN>
Modeling: TBATS

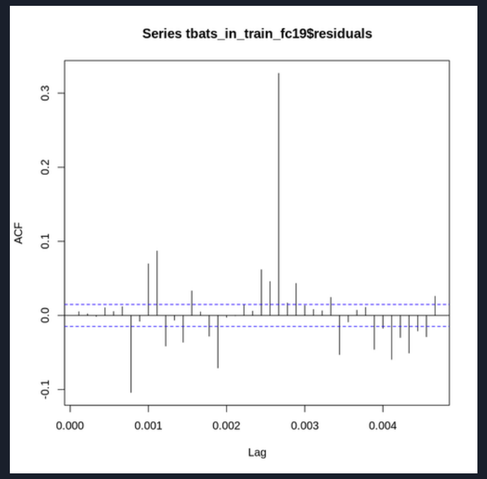


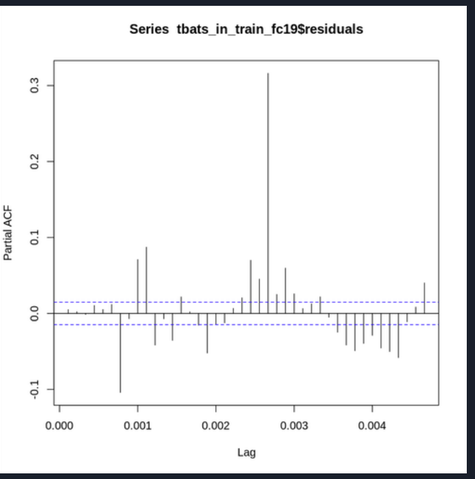






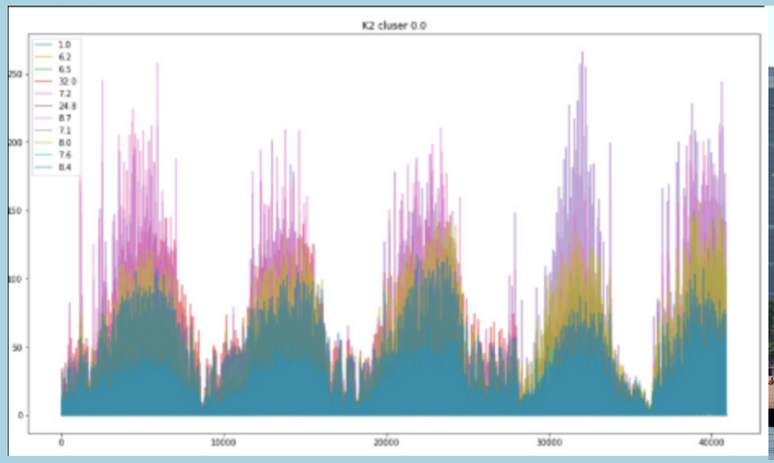
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-5.946371	19.5476	11.20186	-26.44434	172.6922	0.5347444	0.6507094

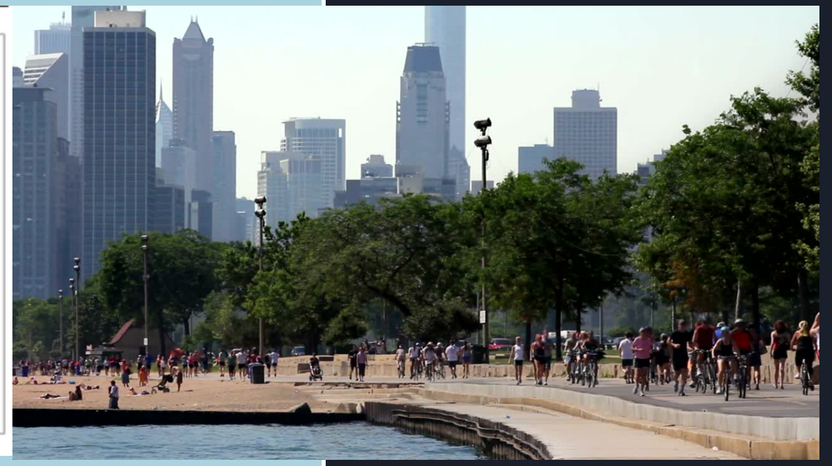


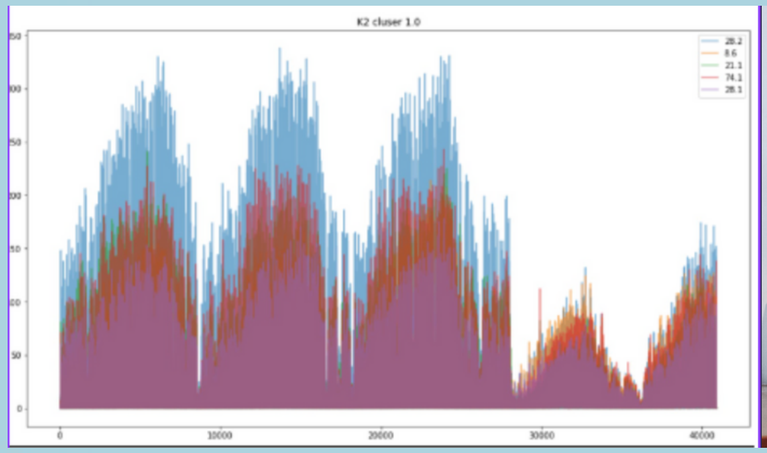



```
[1] "sMAPE - TBATS for Erratic TS:"
92.4730063510853
```

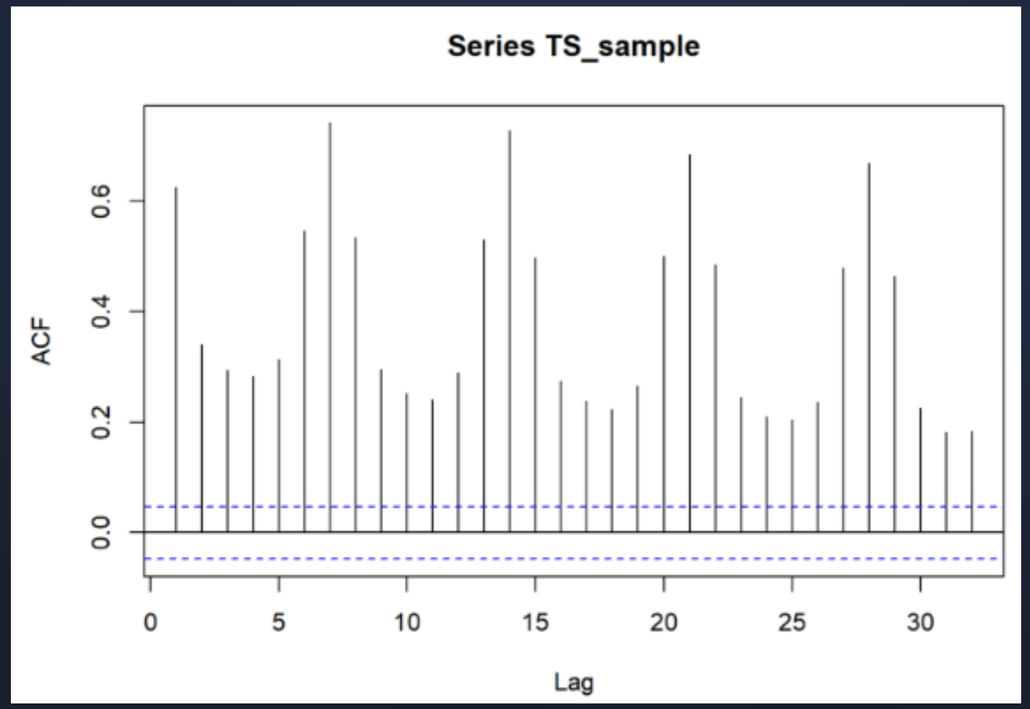
Erratic Cluster k2



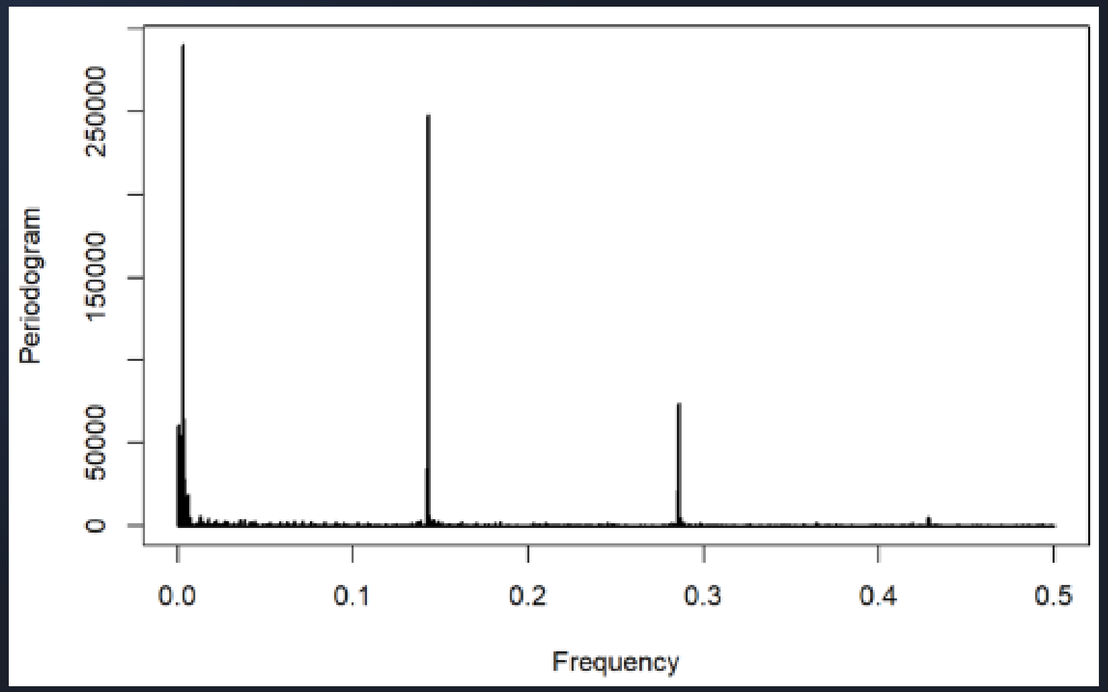




Smooth Data <Station, Daily, IN>
General EDA



There are 3 Dominant frequencies in the Time Series:



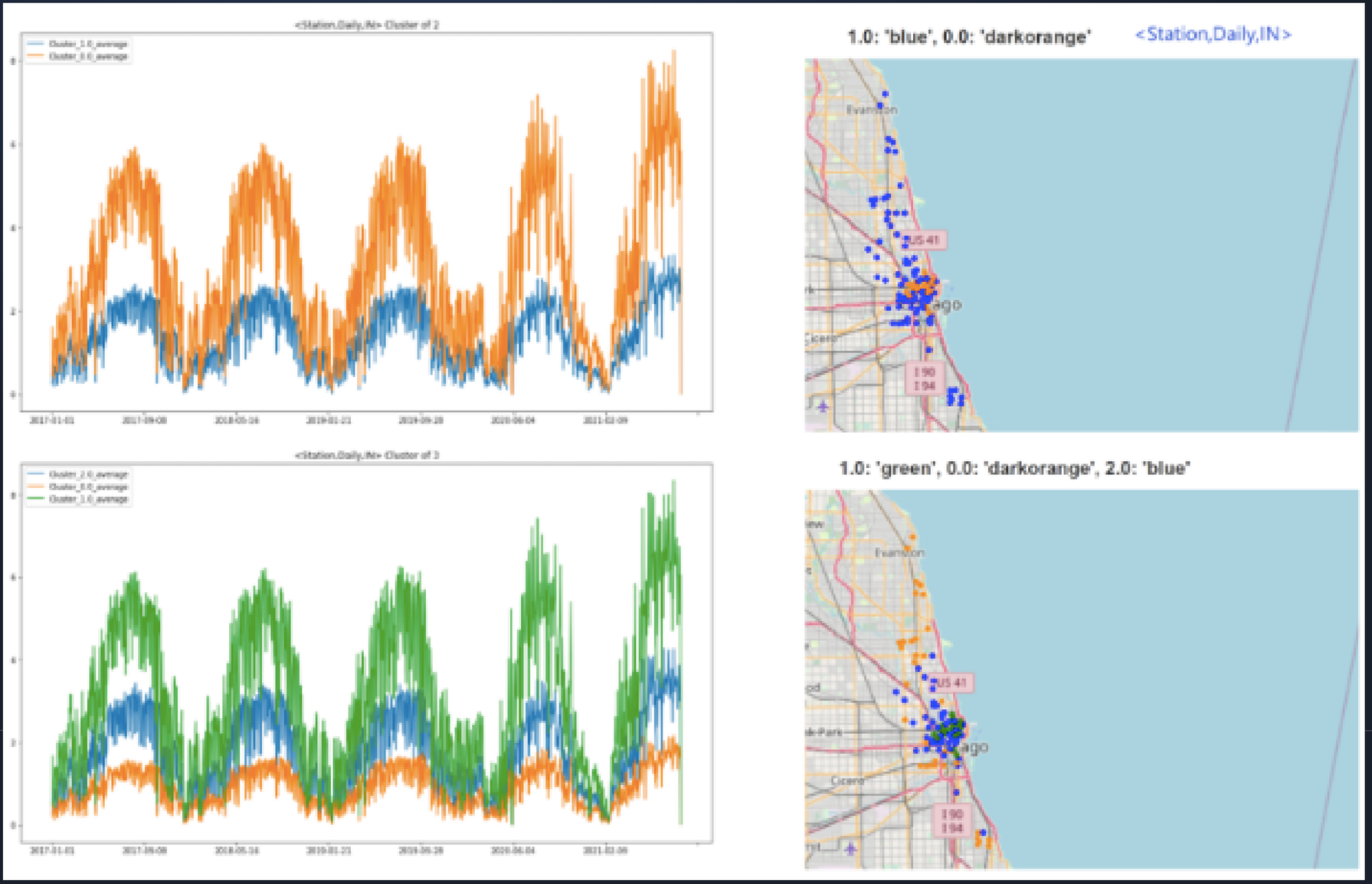
Display the 3 highest "power" frequencies

	freq <dbl>	spec <dbl>	period <dbl>
5	0.002893519	290018.39	345.600000
247	0.142939815	247677.48	6.995951
494	0.285879630	73404.07	3.497976

3 rows

CLUSTERING (Standardized) ON <Station,Daily,IN>

Before clustering, traffic was standardized based on station dock capacity.

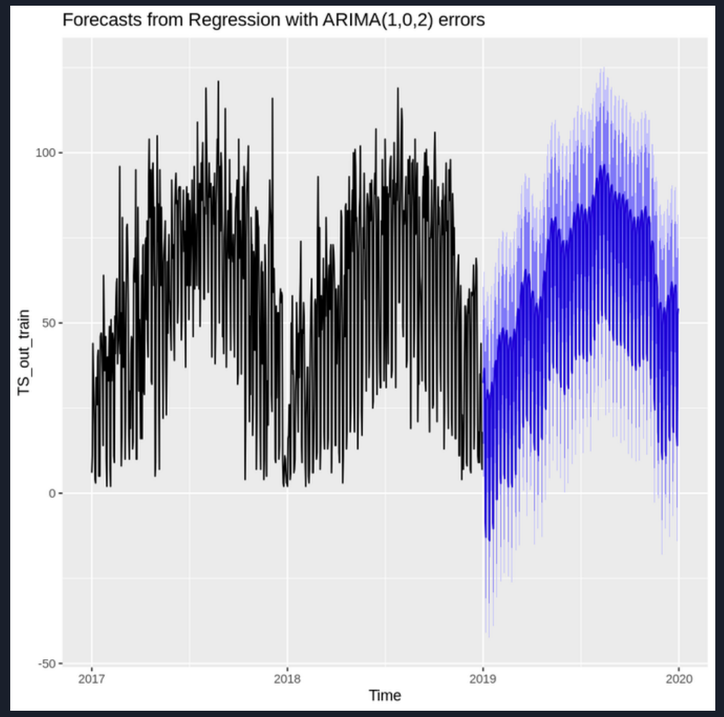


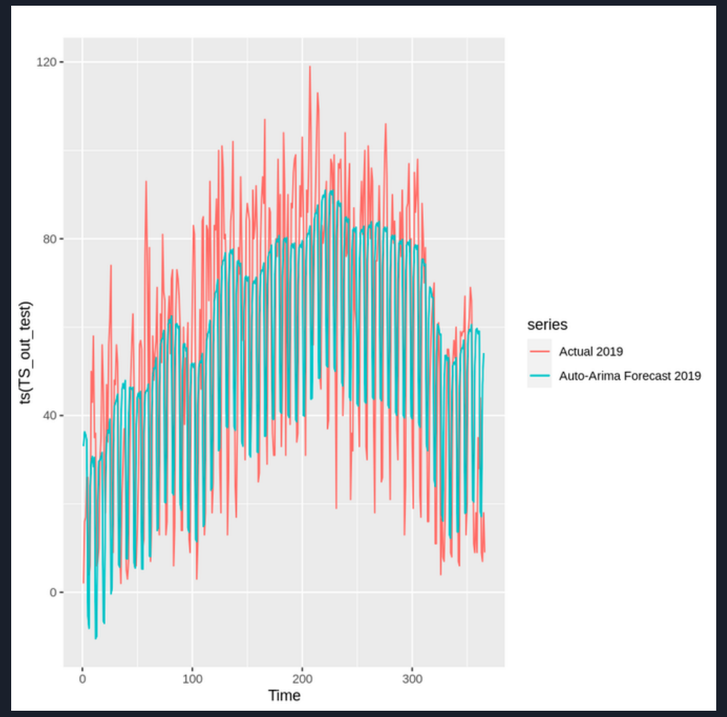
- After standardization & smooth_filtering, k-means nicely separates the stations into multiple parallel layers with different level of traffic intensity.
- The covid intervention was no longer captured by the clustering. We suspect this is due to the smaller geo-granularity. Based on previous observation, COVID hit the city center the hardest, and when aggregated into Neighborhood level, such effect is multiplied and therefore easily visualizable. (But not in the case of station)

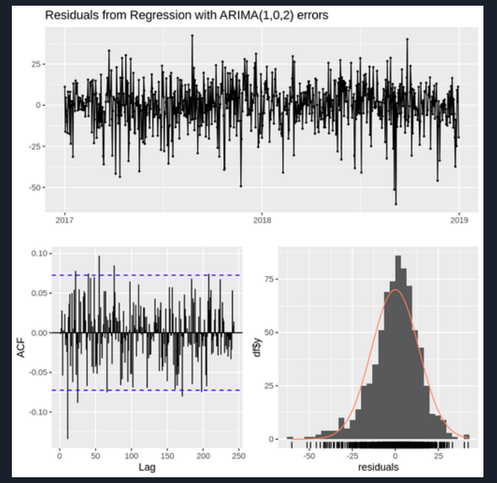
5. Models

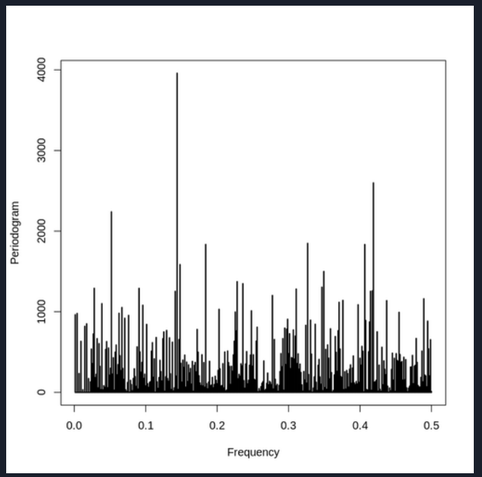
- Auto-Arima (with Fourier Terms)
- Double-Seasonal Holt-Winters
- Prophet
- TBATS
- Dynamic Regression
- Intervention Model

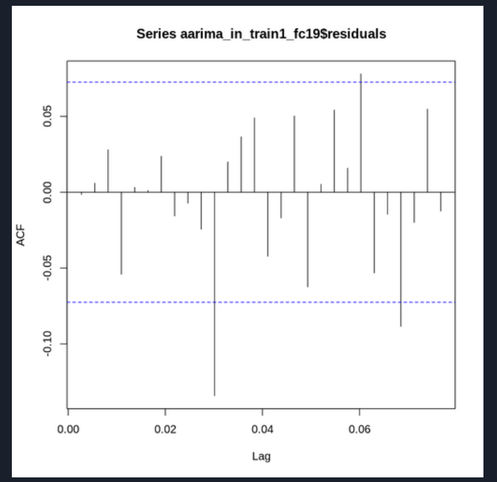
Smooth Data <Station, Daily, IN>
Modeling: Auto-Arima with Fourier Terms

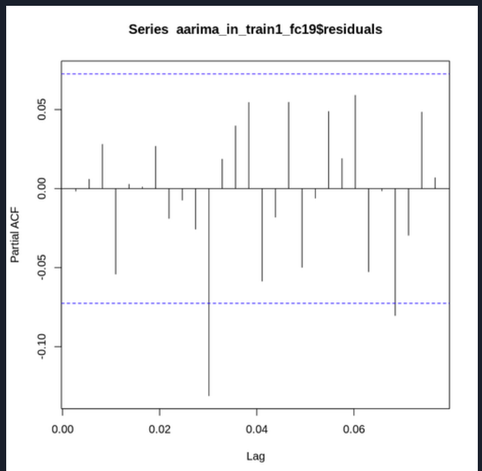








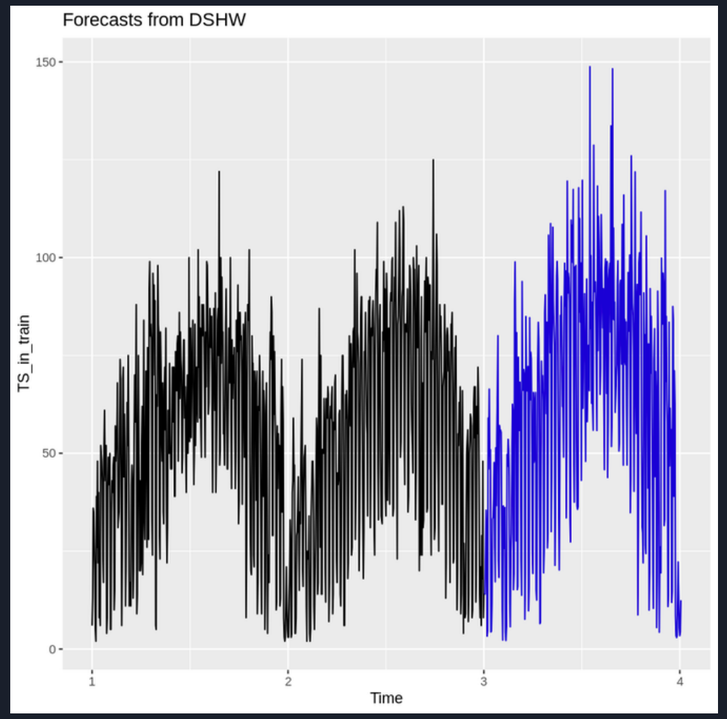


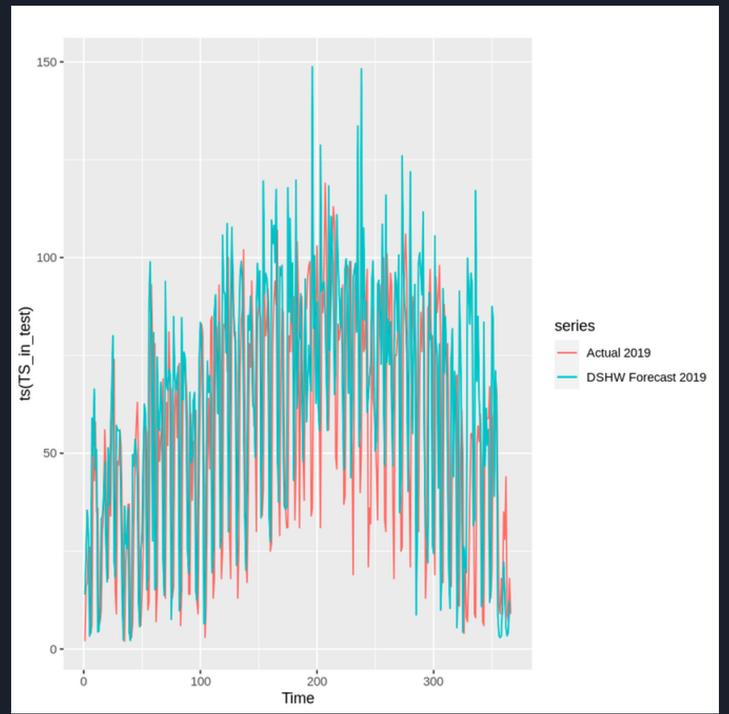


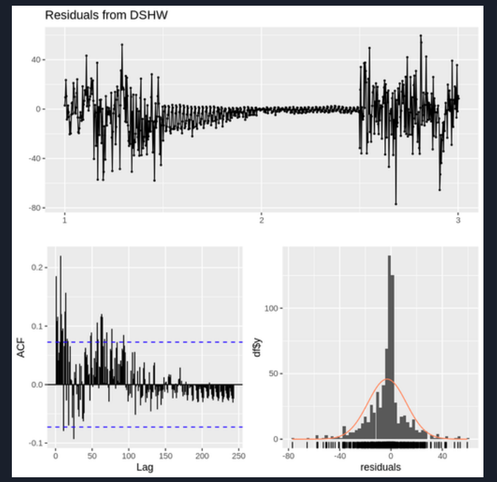
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-2.355181	25.98899	19.79982	44.49055	121.9293	0.2133417	1.346638

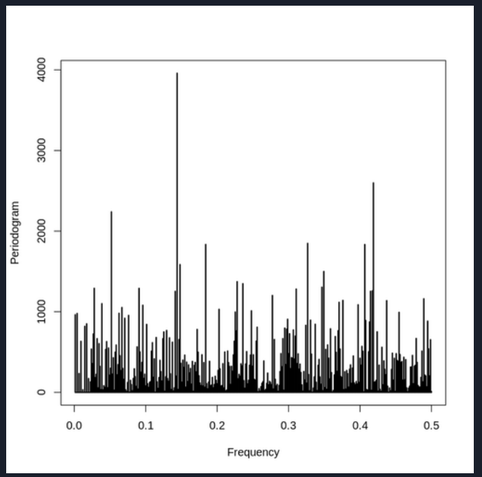
```
[1] "Auto-Arima (Yearly K=8, Weekly K=3) SMAPE"
47.5809707124306
```

Smooth Data <Station, Daily, IN>
Modeling: Double-Seasonal Holt-Winters

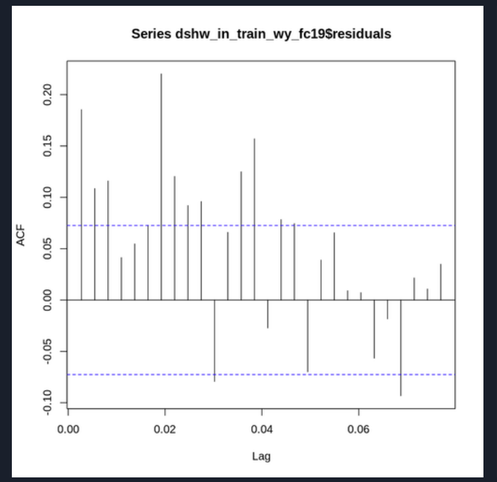


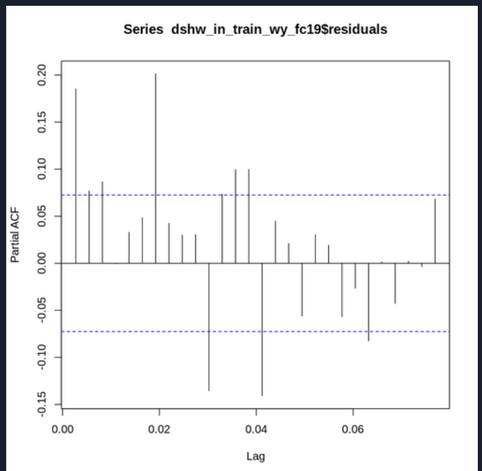






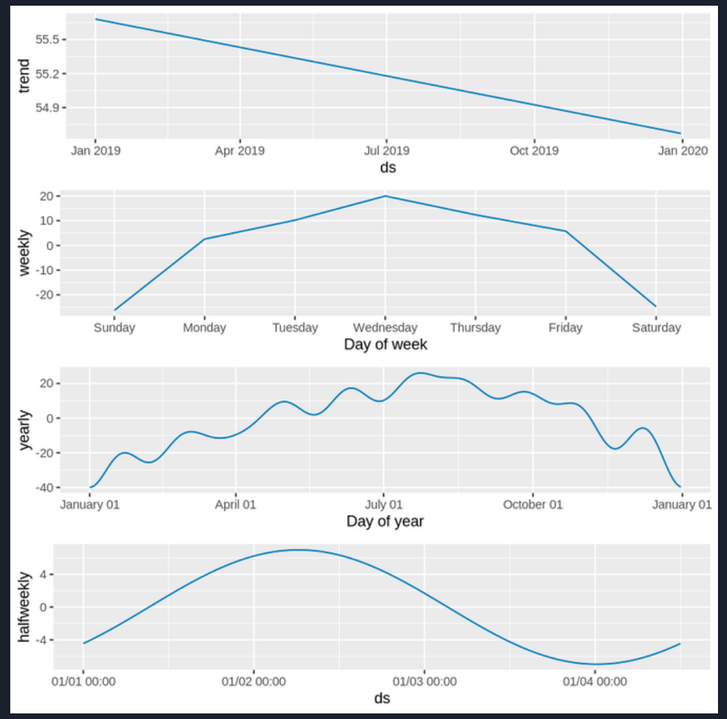
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	6.04653	34.03725	24.47779	-31.25721	72.61358	0.09281662	1.021465

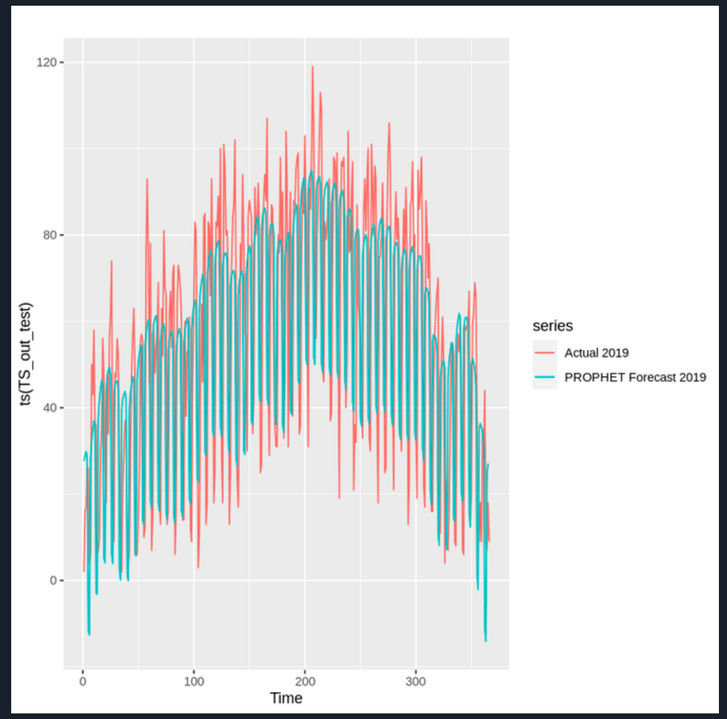


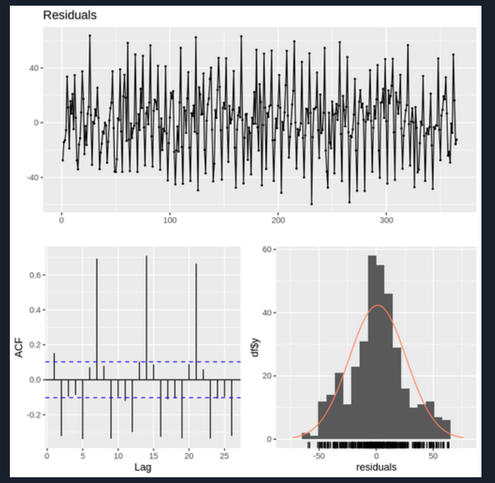


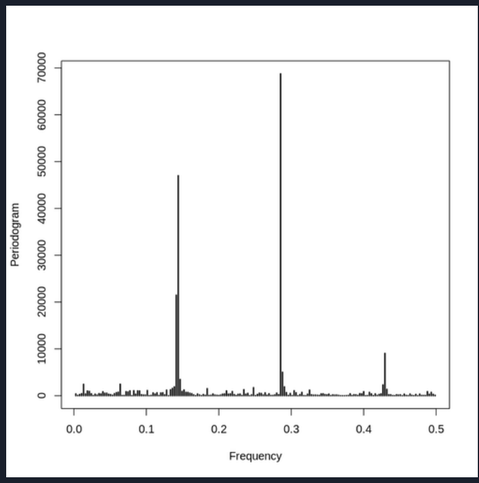

```
[1] "DSHW (Weekly, Yearly) SMAPE"
50.4631770315658
```

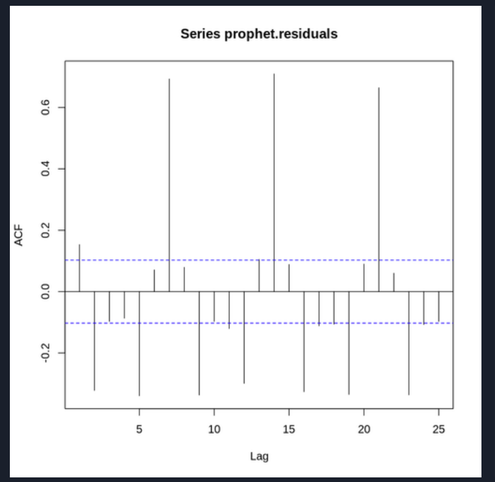
Smooth Data <Station, Daily, IN>
Modeling: Prophet

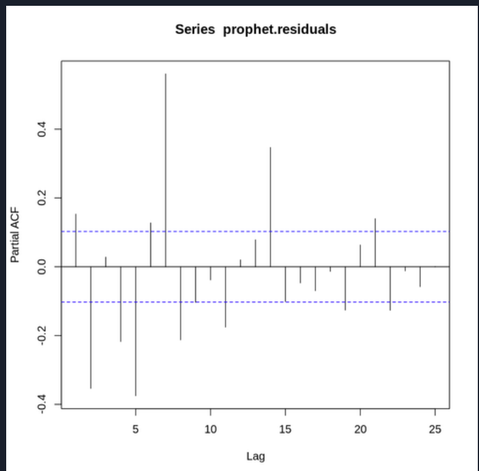








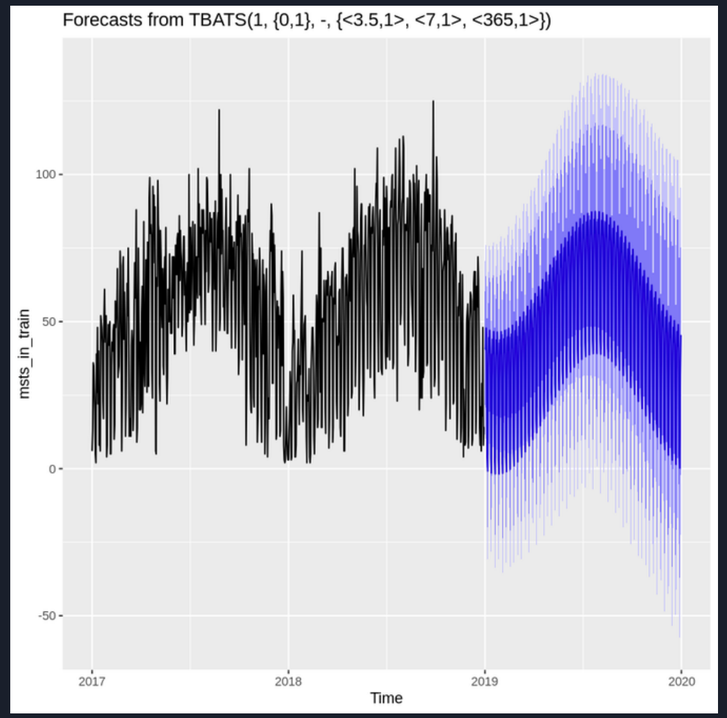


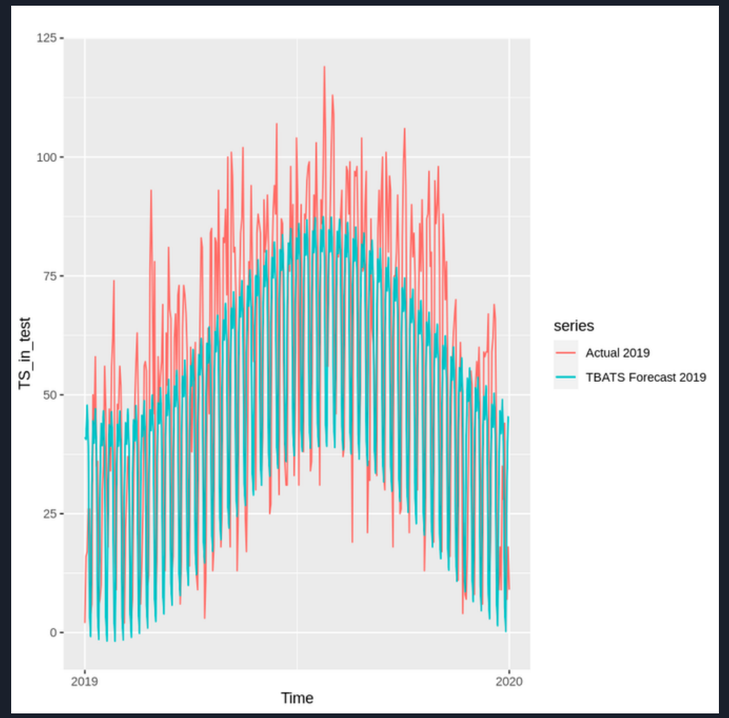


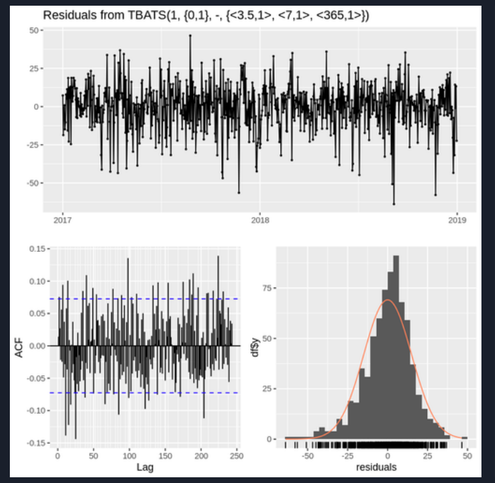
	ME	RMSE	MAE	MPE	MAPE
Test set	-1.335039	24.92865	18.99812	-17.05138	61.15166

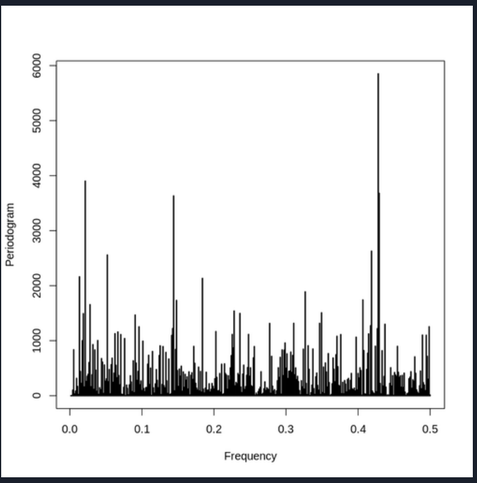
```
[1] "Prophet (Half-weekly, Weekly, Yearly) SMAPE"
45.7530172238639
```

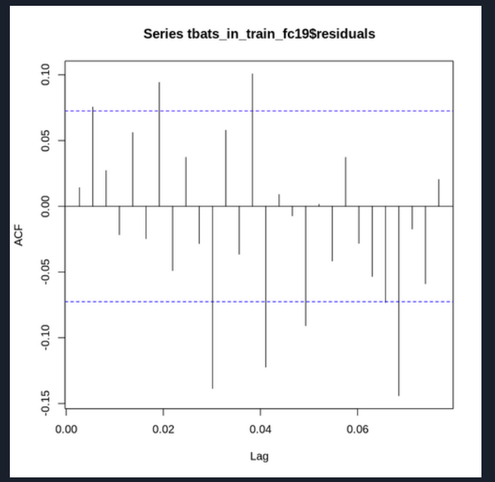
Smooth Data <Station, Daily, IN>
Modeling: TBATS

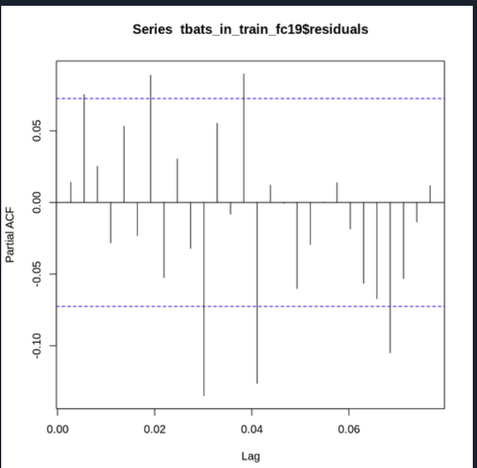








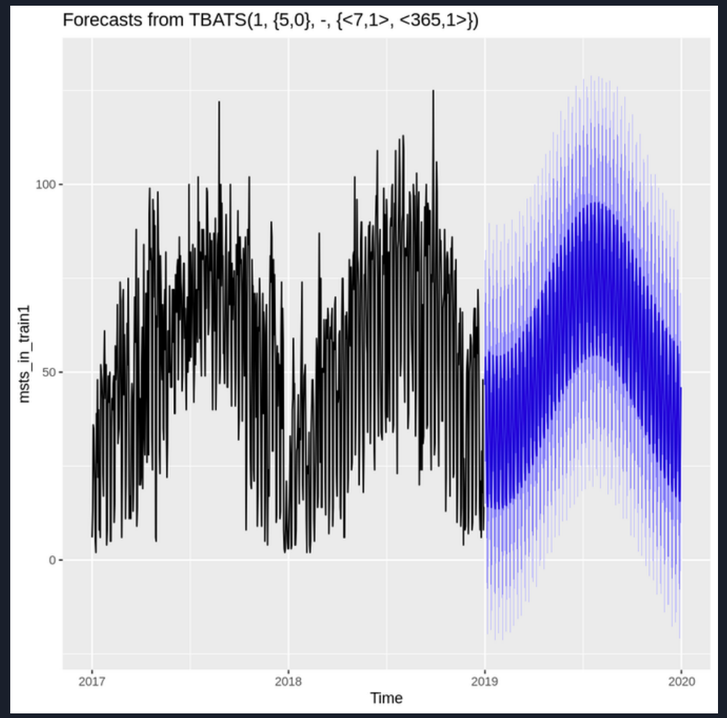


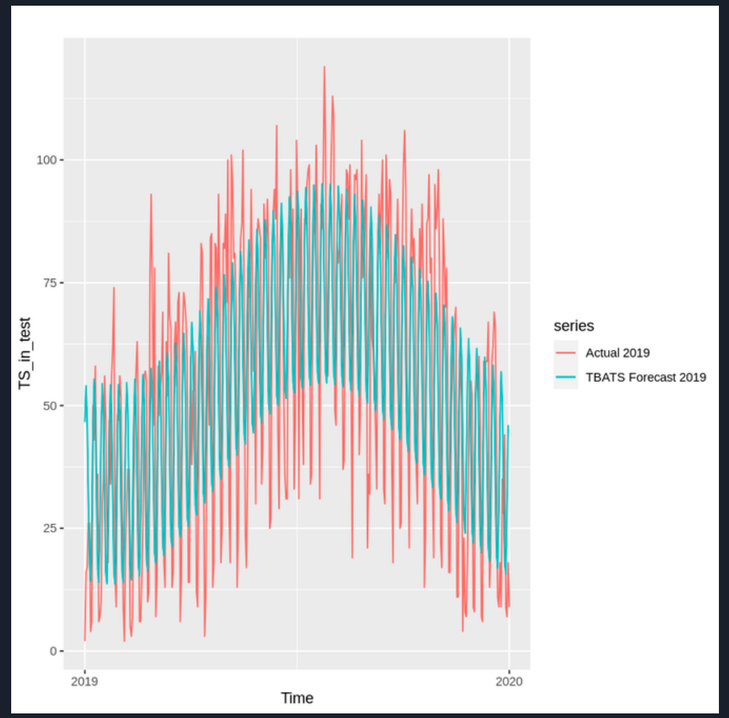


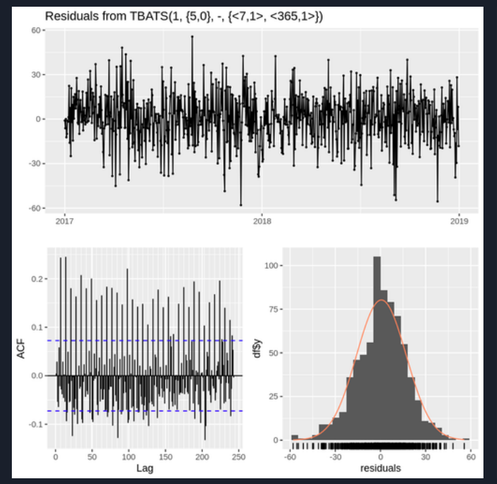
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-7.207978	26.14934	20.63001	-45.91271	119.5336	0.2475894	0.7958925

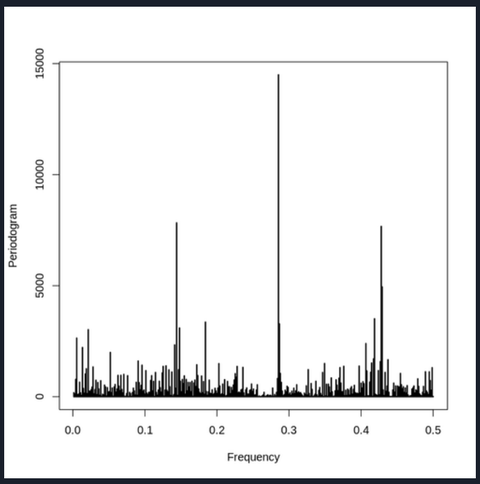
```
[1] "TBATS (Half-weekly, Weekly, Yearly) SMAPE"
52.4580758209308
```

Smooth Data <Station, Daily, IN>
Modeling: TBATS

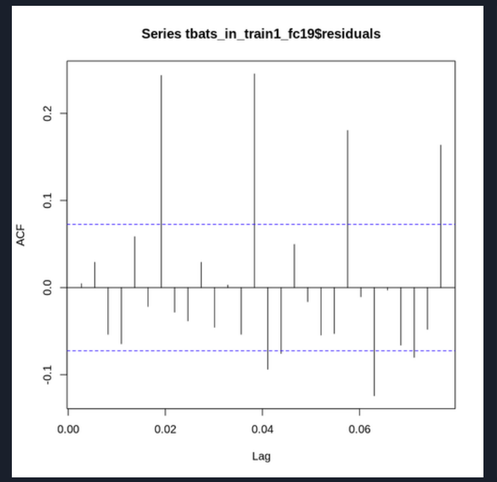


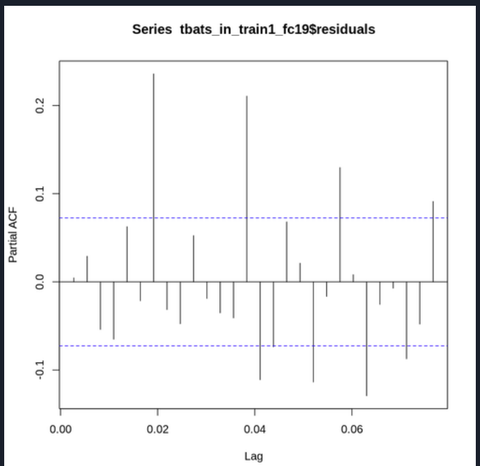






	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-2.142445	23.28578	18.97309	-6.932426	44.72032	0.3266599	1.755569

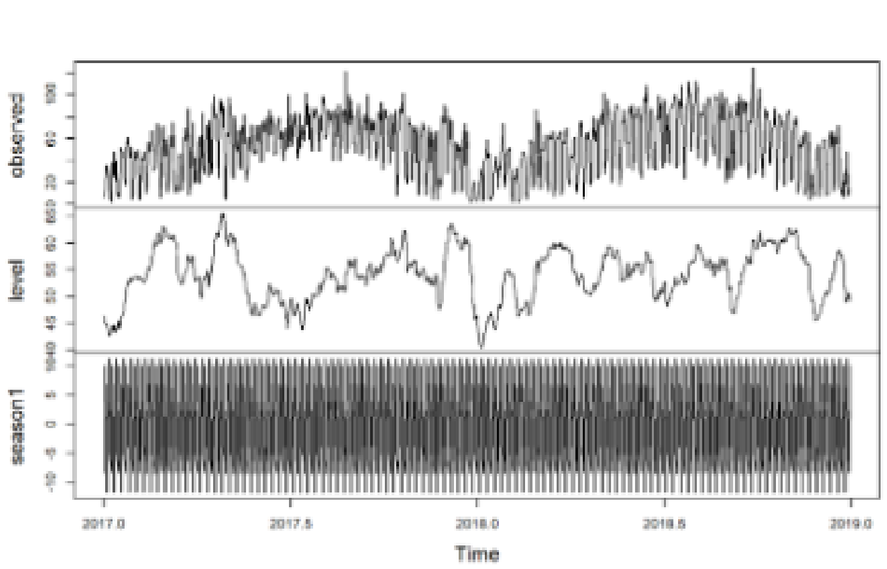


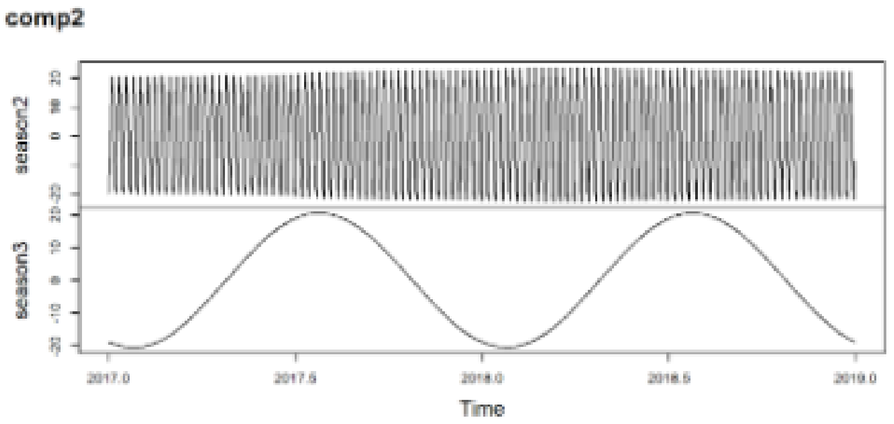



```
[1] "TBATS (Weekly, Yearly) SMAPE"
44.3874843481476
```

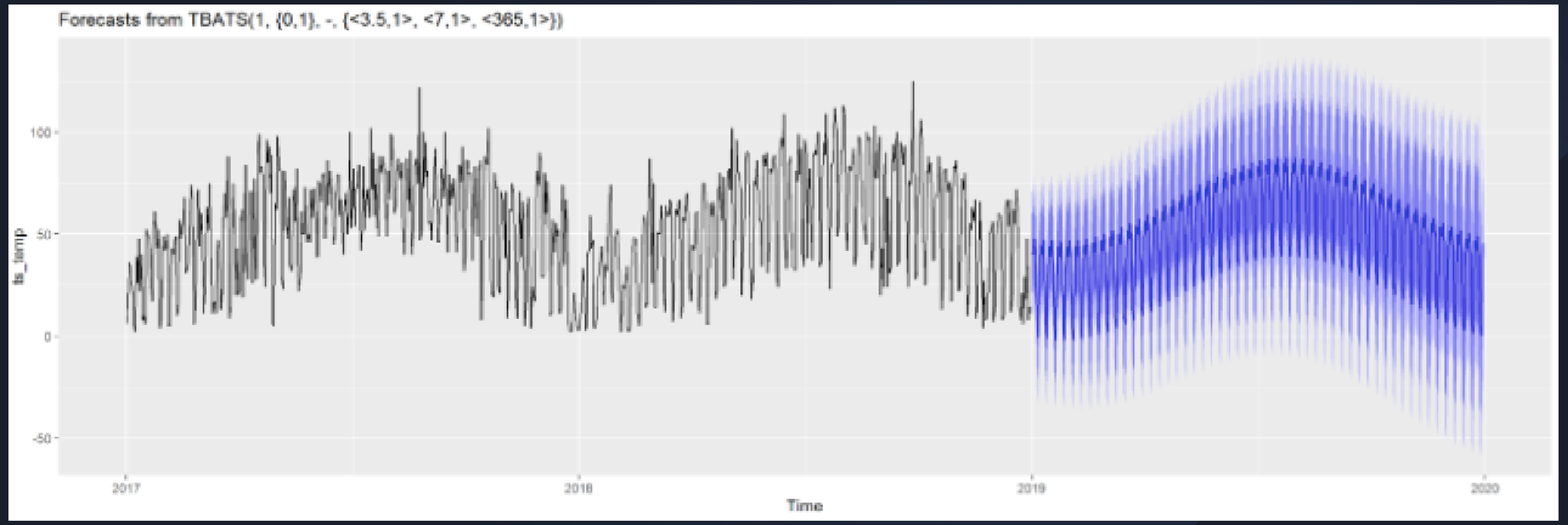
TBATS

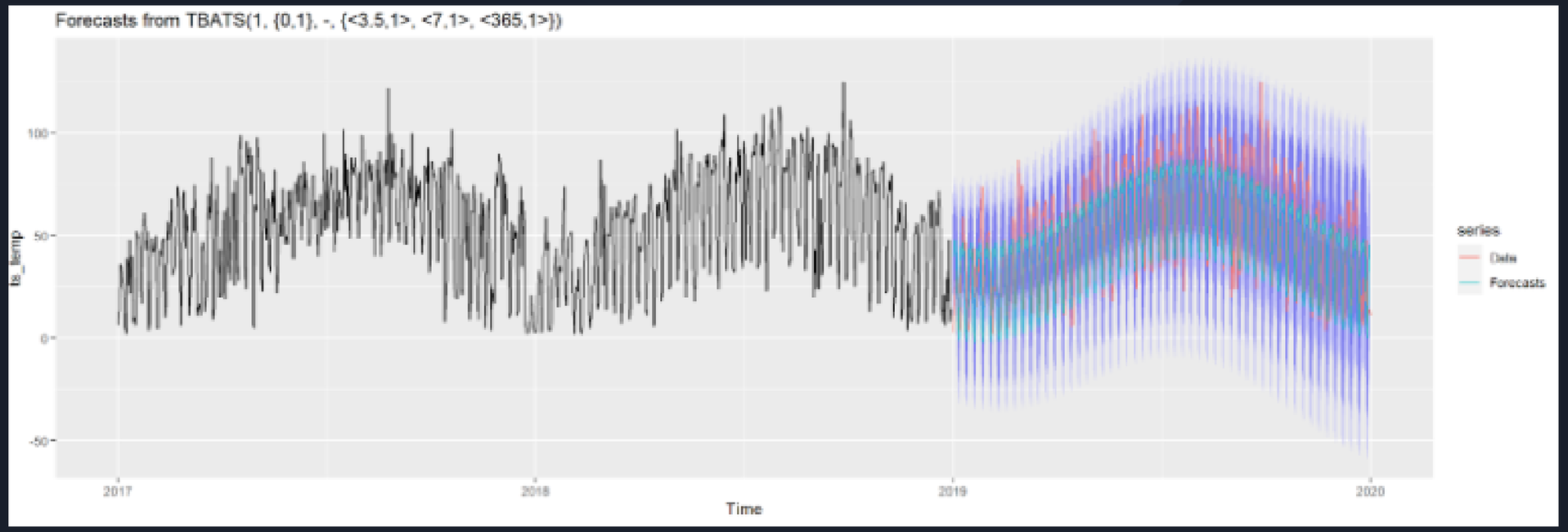
```
#3.5, 7 ,365
ts_temp <- msts(Ts_train, seasonal.periods=c(3.5, 7, 365))
model2 <- tbats(ts_temp)
comp2 <- tbats.components(model2)
plot(comp2)
```

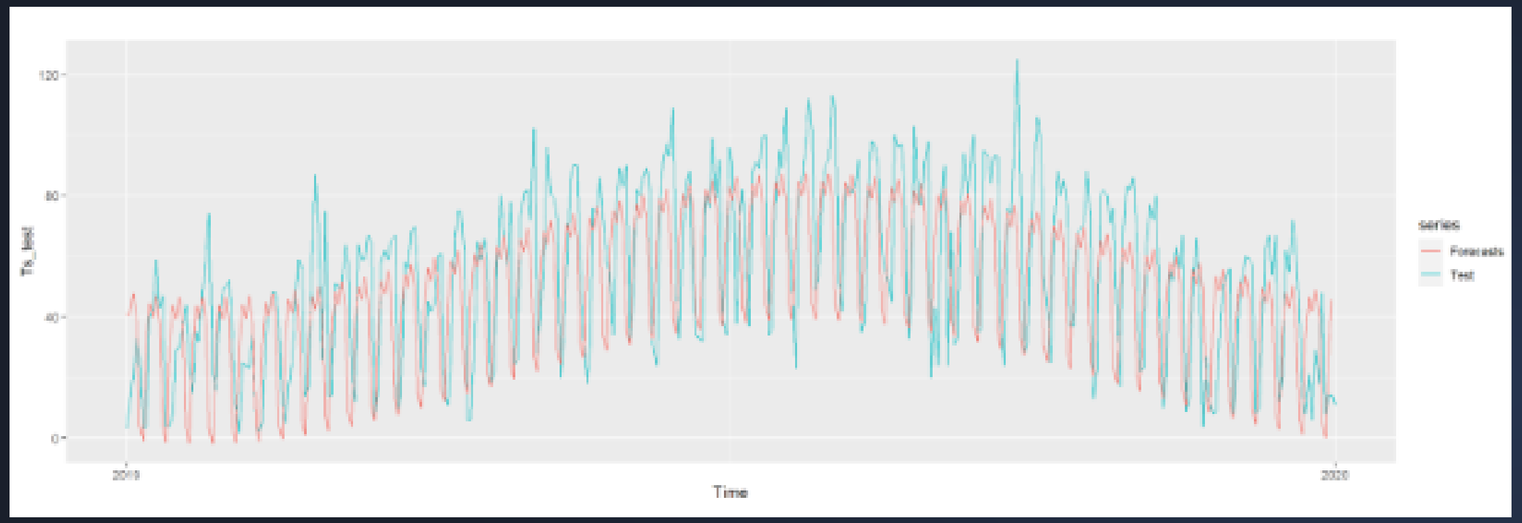


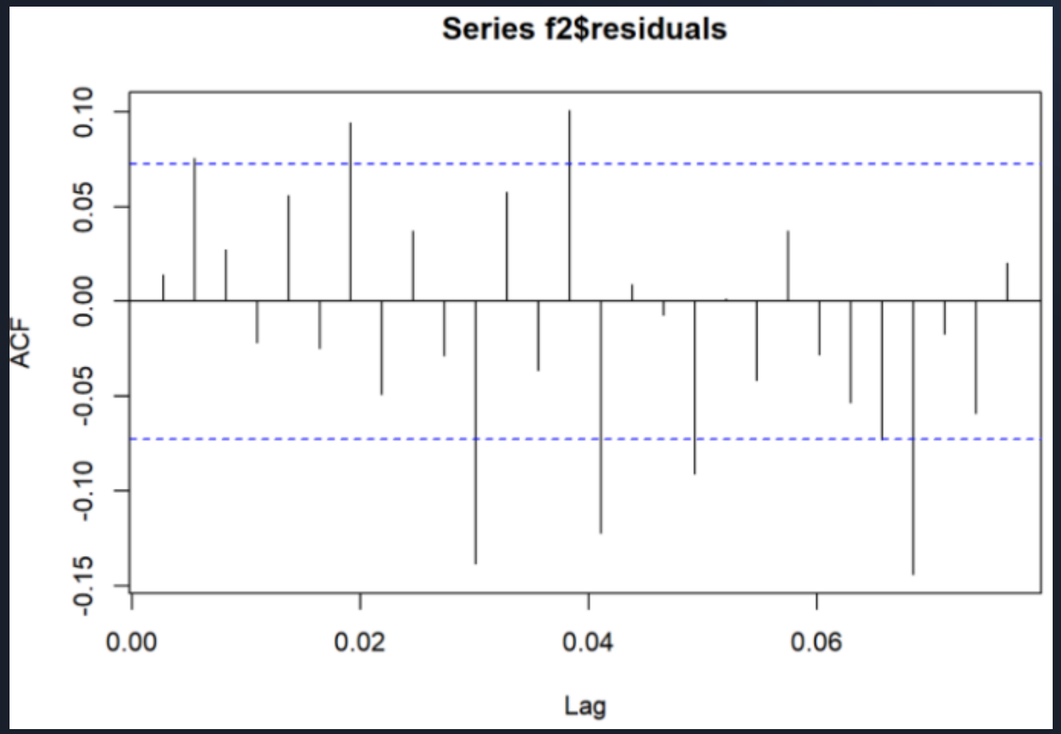


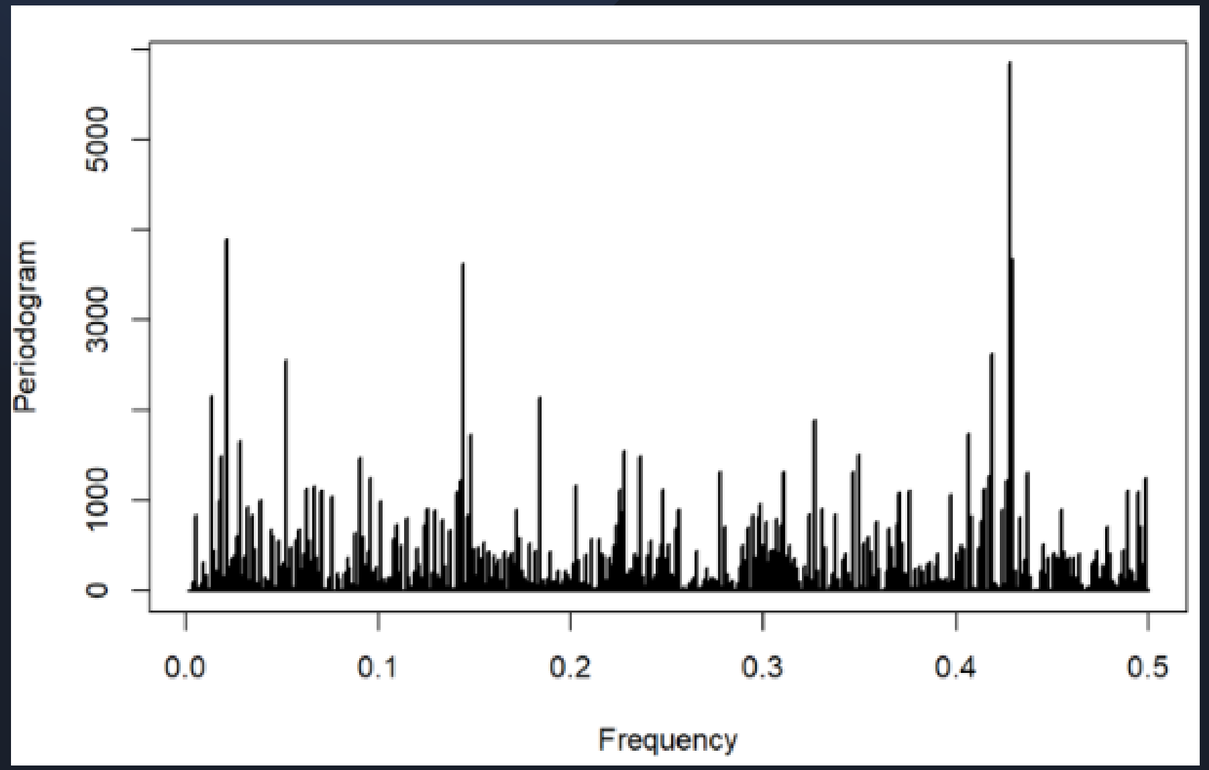
For complex seasonal patterns, we need exponential smoothing


Forecasts from TBATS(1, {0,1}, -, {<3.5,1>, <7,1>, <365,1>})
ts_temp
100
50
0
-50
2017
2018
2019
2020
Time


Forecasts from TBATS(1, {0,1}, -, {<3.5,1>, <7,1>, <365,1>})
ts_temp
100
50
0
-50
2017
2018
2019
2020
Time
series
Data
Forecasts







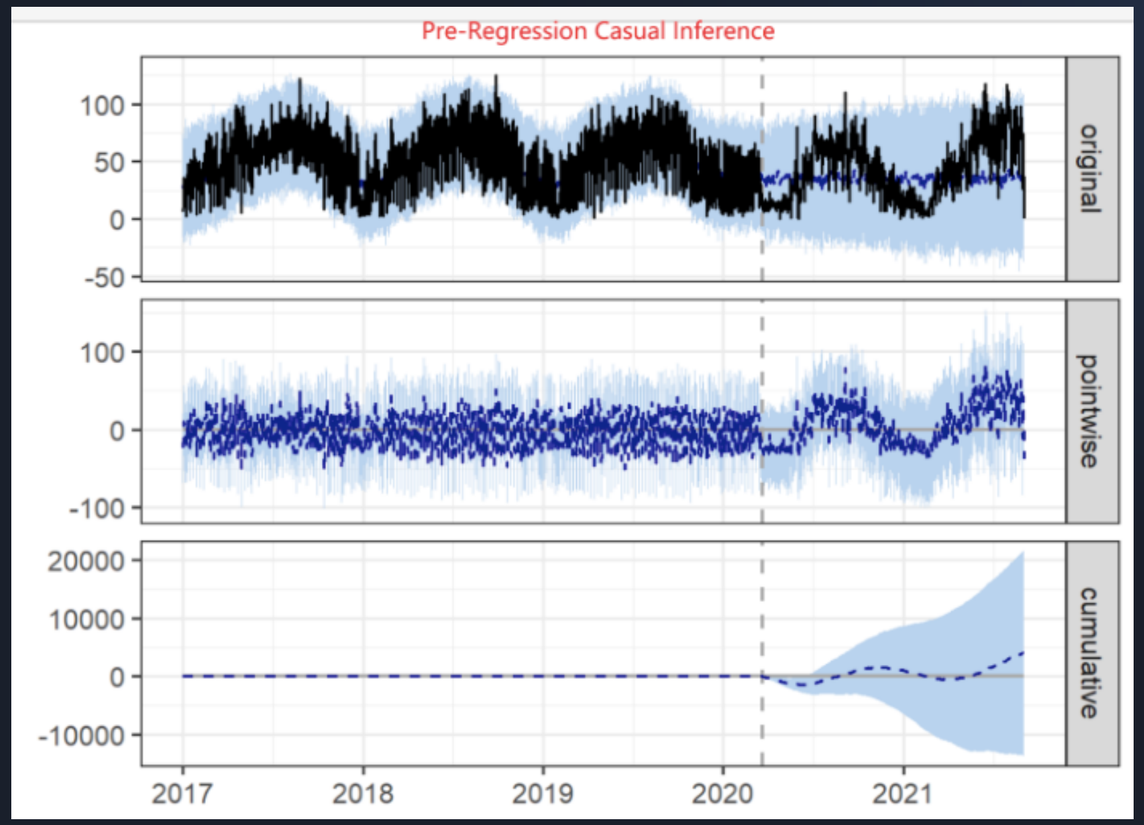
Model residual is NOT auto-correlative .

Periodogram shows there is no dominant seasonality.

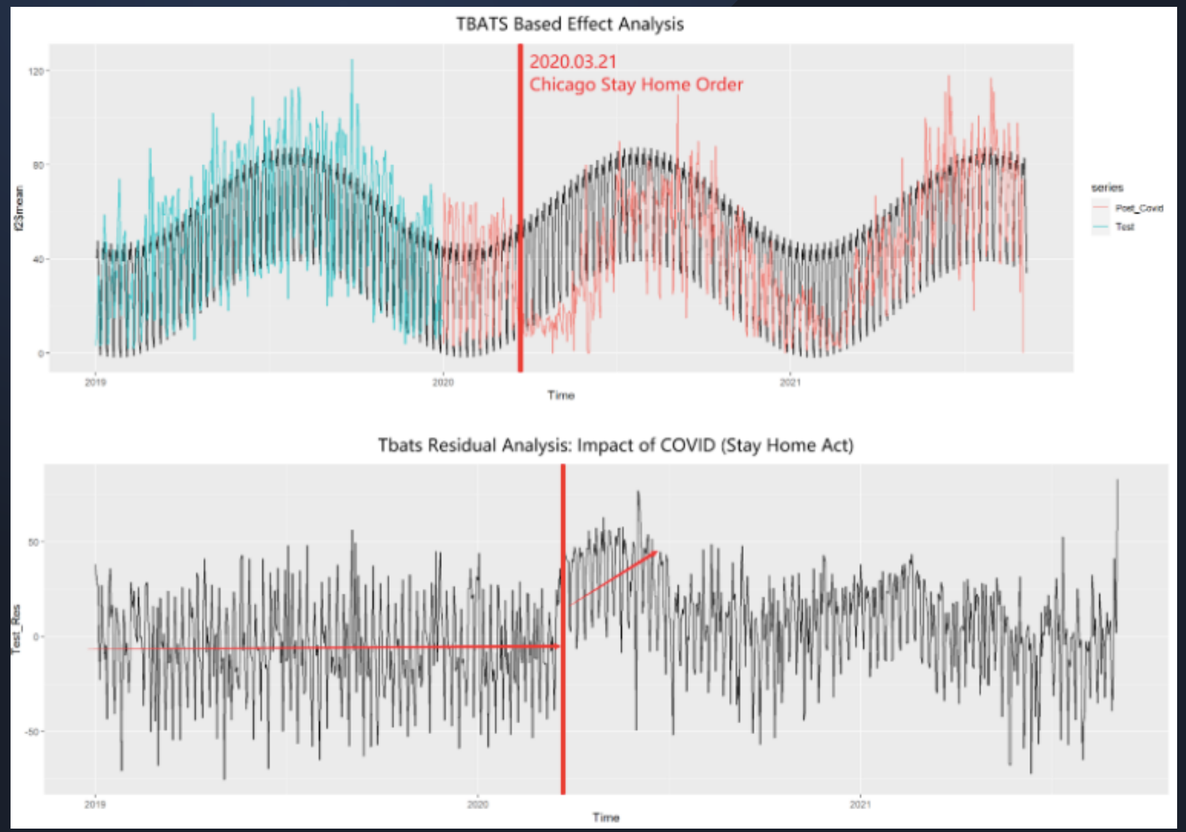
Evaluation

Model	RMSE	sMAPE
TBATS (H, W, Y) TBATS (W, Y)	26.15 23.29	52.46 44.39
Auto-Arima (+ Fourier)	25.99	47.58
Double-Seasonal Holt-Winters	34.04	50.46
Prophet	24.93	45.75

Dynamic Regression



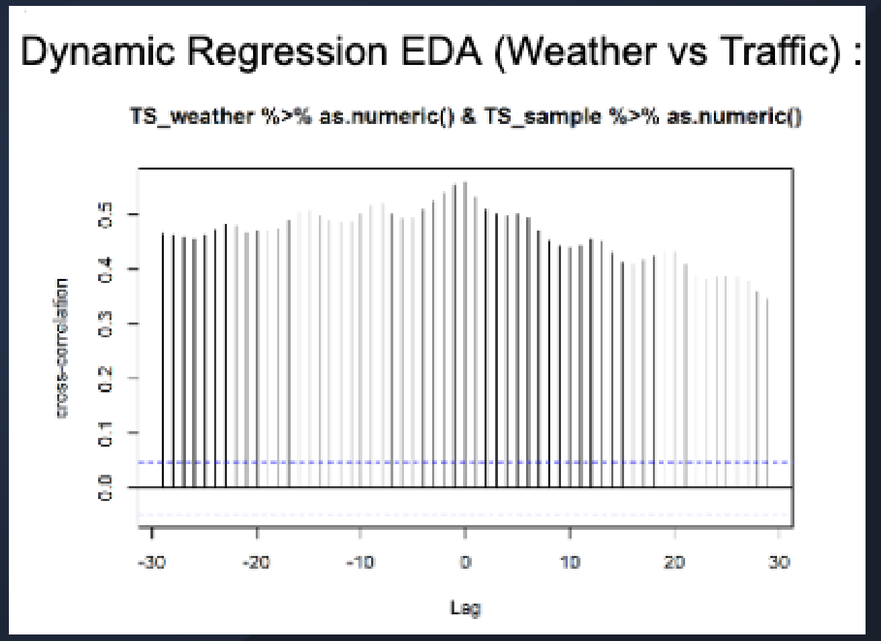
TBATS: Effect and Residual Analysis

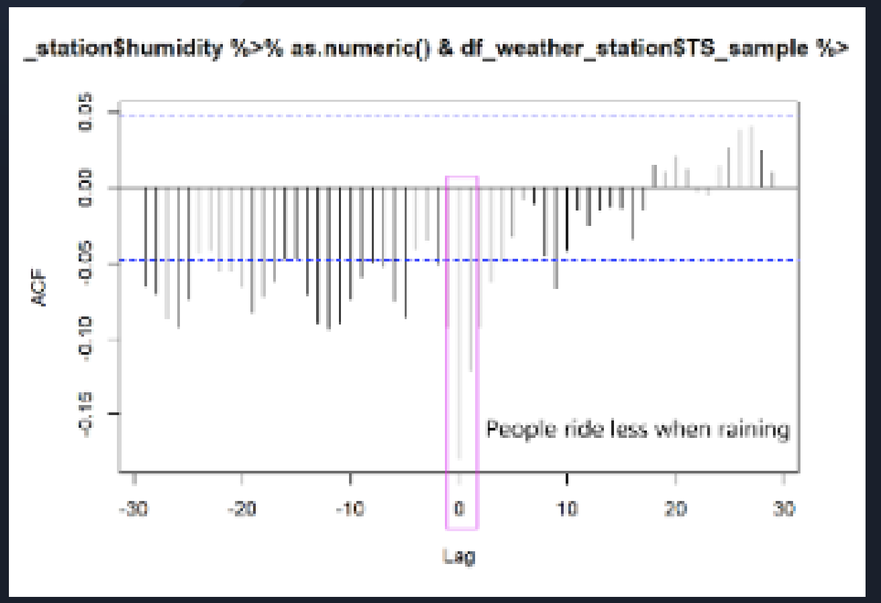


Dynamic Regression EDA

```
## Posterior inference {CausalImpact}
##
##                          Average        Cumulative
## Actual                   43             22750
## Prediction (s.d.)        35 (17)        18763 (8865)
## 95% CI                   [2.3, 69]      [1236.6, 36442]
##
## Absolute effect (s.d.)   7.5 (17)       3987.0 (8865)
## 95% CI                   [-26, 41]      [-13692, 21513]
##
## Relative effect (s.d.)   21% (47%)      21% (47%)
## 95% CI                   [-73%, 115%]   [-73%, 115%]
##
## Posterior tail-area probability p:   0.3253
## Posterior prob. of a causal effect:  67%
##
## For more details, type: summary(impact, "report")
```

Observation: Weather, and Traffic are Cointegrated. Since they have the same level of non-stationarity, their linear combination cancels out the stochastic trends.





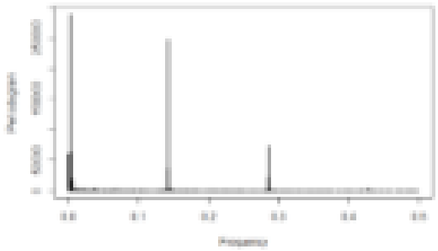
Compare & Contrast Dominant Seasonality After Regression

Description: df [3 x 3]

TS Seasonality Analysis

	freq	spec	period
	<dbl>	<dbl>	<dbl>
5	0.002893519	290018.39	345.600000
247	0.142939815	247677.48	6.995951
494	0.285879630	73404.07	3.497976

3 rows

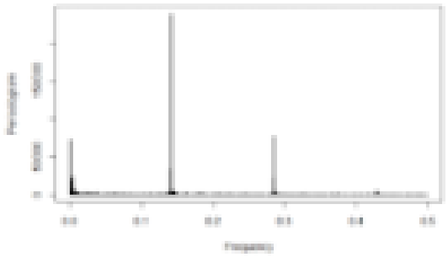


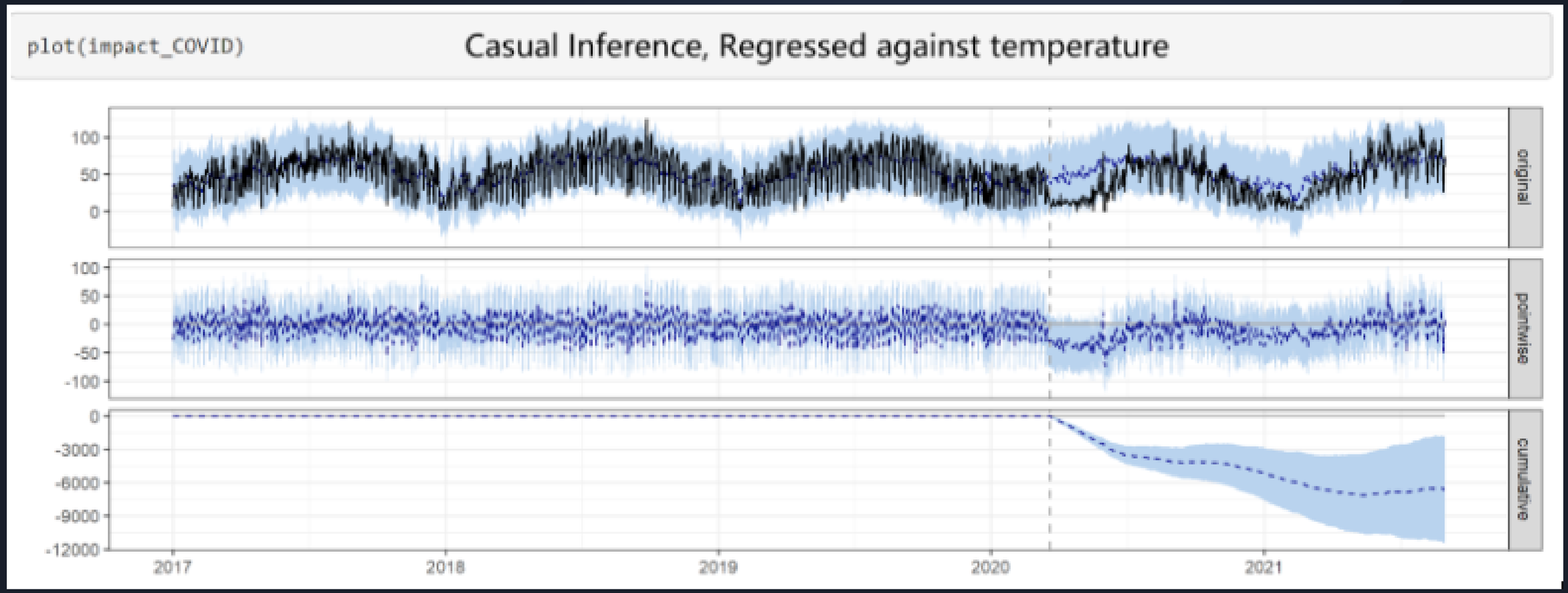
Description: df [3 x 3]

Residual Seasonality Analysis: Post Weather Regressed

	freq	spec	period
	<dbl>	<dbl>	<dbl>
247	0.1429398148	239289.13	6.995951
494	0.2858796296	76912.47	3.497976
1	0.0005787037	72660.63	1728.000000

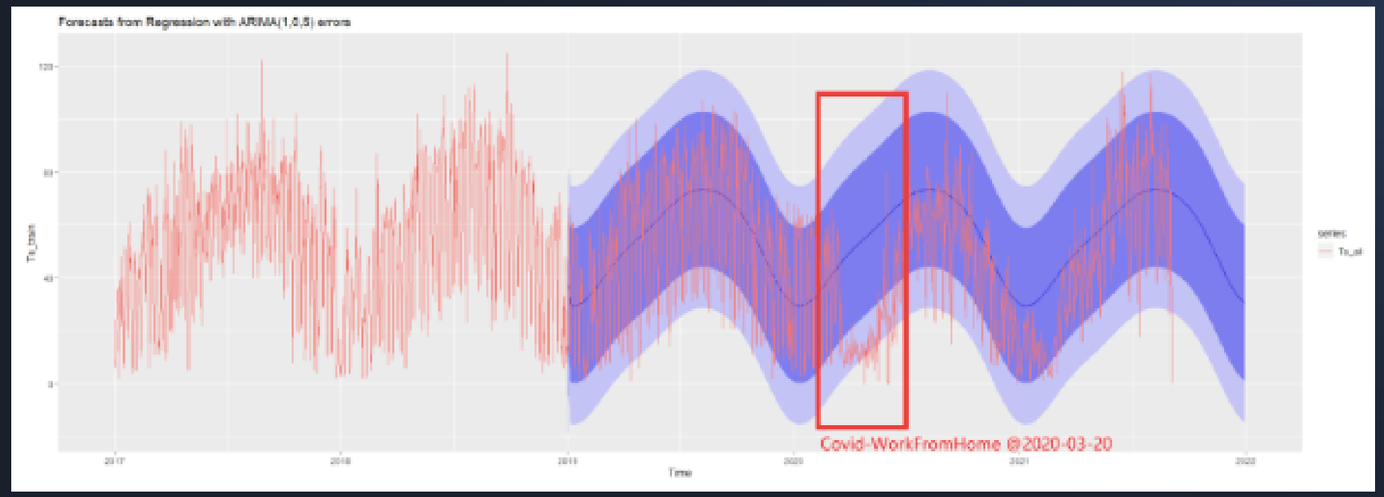
3 rows

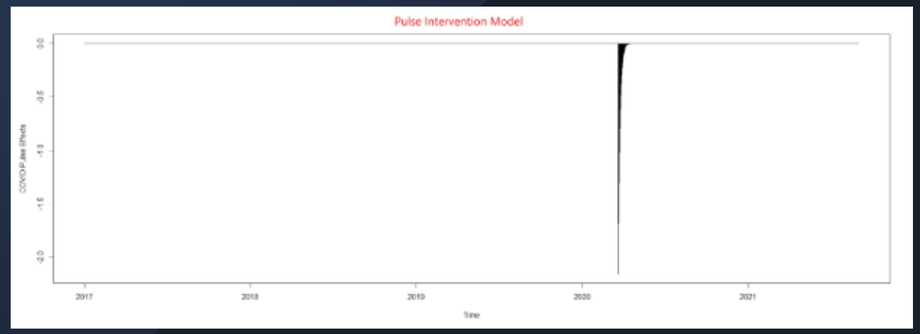


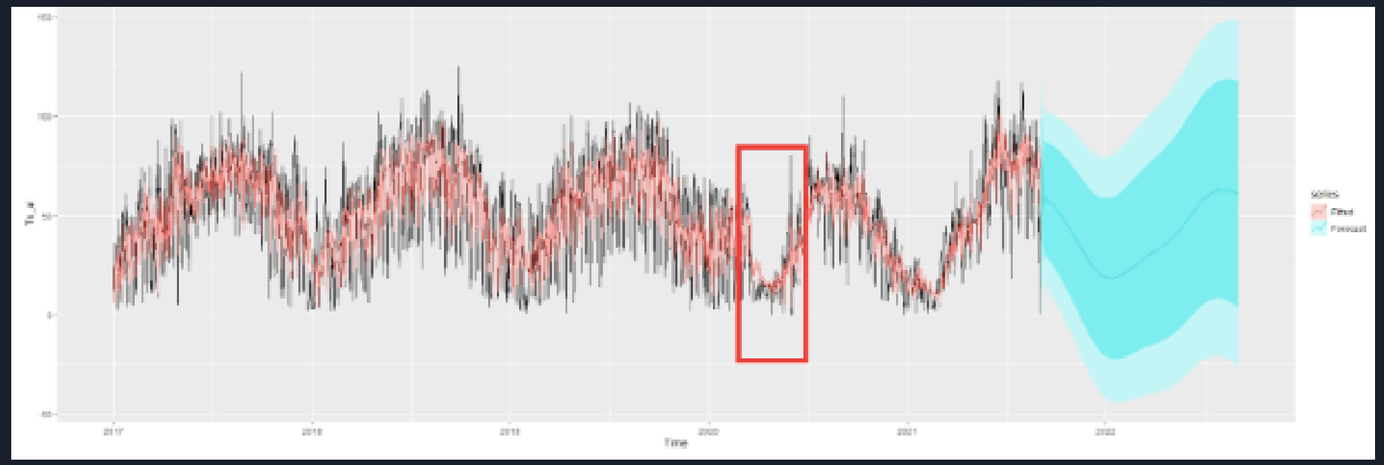


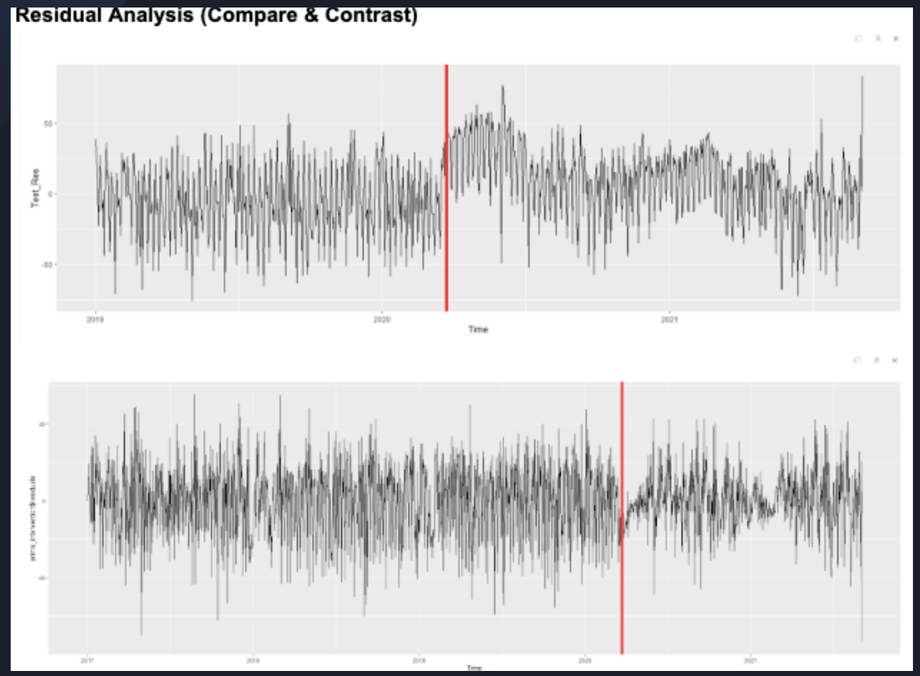
After regression against temperature, which is highly seasonal, the yearly seasonality is removed. However, weekly and half-weekly persist strongly.

Intervention Model









Conclusion

- The best model is TBATS with the lowest error, suitable for complex seasonality
- Erratic data produces high variance forecast
- Aggregation helps a more accurate forecast, which can be used as a base for more granular estimation
- Different region has different pattern
- Weather variable proves to be an important influencing variable as it is cointegrated
- Covid intervention is temporary and has different impact on different region

Future work

1. Casual Inference on temporary stations.
2. Intervention modeling to predict impact of adding or removing temporary stations on dock balance.
3. NPV analysis on weather to implement temporary stations to cut cost and when to do so.
4. Compile In + Out forecast to calculate forecasted balance
5. Develop web-service API prototype that forecasts staion dock avaliability in real time.