Predicting Dengue Outbreaks in Singapore

Using Time Series Analysis

Presentation Slides

Scope of Presentation

- Introduction
- Project Objectives
- Methodology
- Phase I: Creating a model using only Past Observations
- Exploring Possible Variables
- Data Cleaning
- Phase II: Enhancing the model with Exogenous Variables
- Results & Limitations
- Conclusion

Introduction

Dengue is an infectious disease spread by the ***Aedes aegypti*** mosquitoe

Mode of Transmission

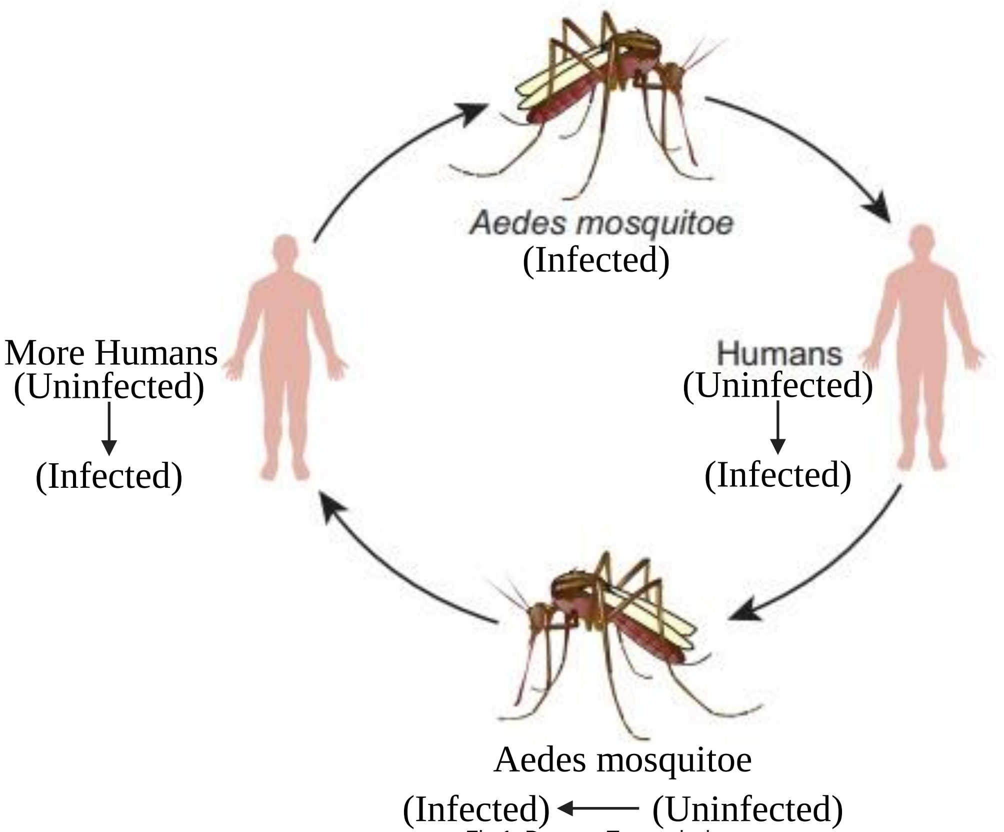


Fig 1. Dengue Transmission

- Healthy person gets infected when bitten via an infected mosquitoe
- Healthy mosquitoe can get infected by biting an infected person

Introduction

Global Impact of Dengue

- 390 million people infected per year
- 500,000 develop into **Dengue Haemorrhagic Fever**
- 25,000 deaths in 2019
- Targets both urban and rural areas. Largely occurring in tropical and sub-tropical areas

- Areas with high rainfall, temperatures & humidity

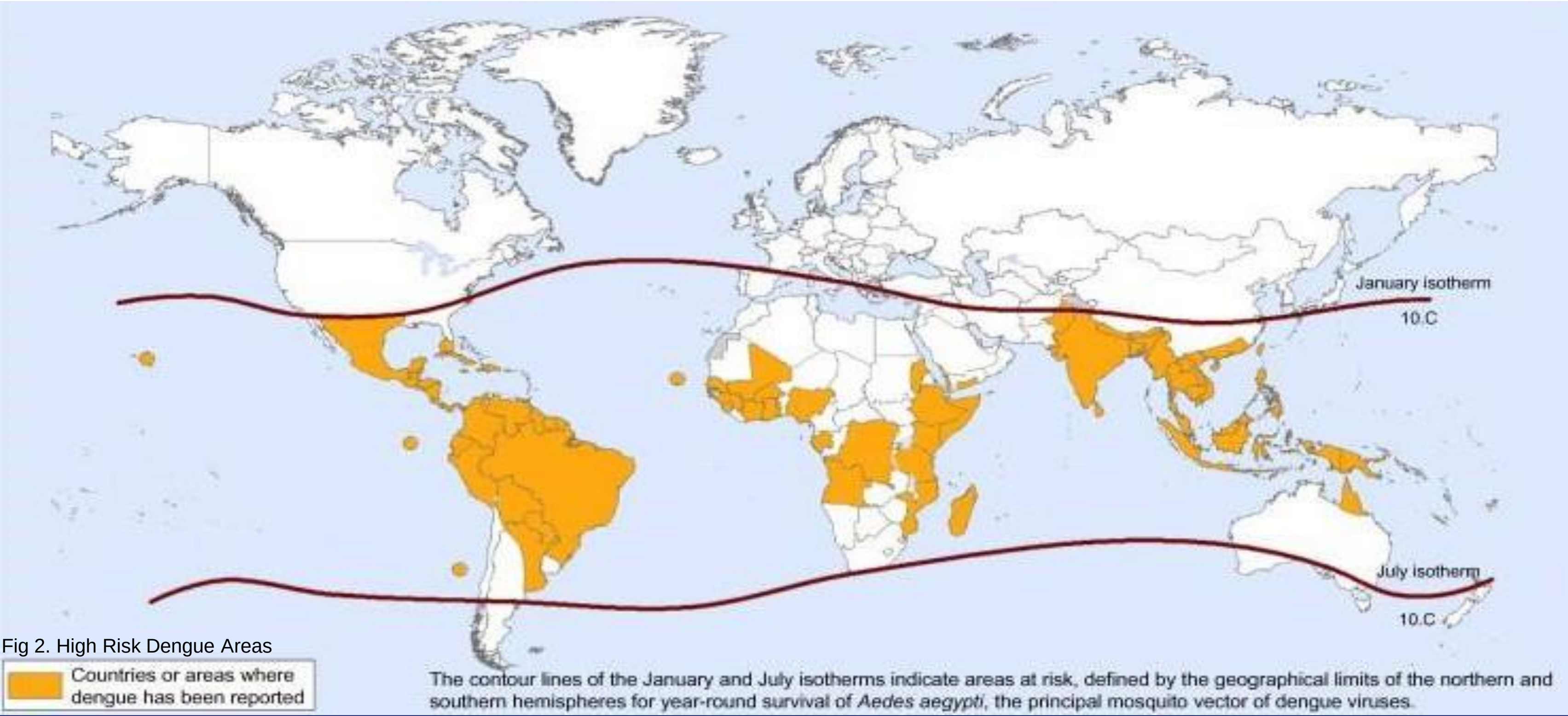


Fig 2. High Risk Dengue Areas

Introduction

Local Impact of Dengue

- Singapore’s location is very suitable for the ***Aedes aegypti*** mosquitoe to breed
- Before January 2020, 2019 had 14, 658 cases. 5.5 times increase from previous year (DENV-2 Strain)
- DENV-3 Strain was detected in January 2020
- This suggests that dengue cases could increase even more, putting more pressure on valuable healthcare resources

Introduction

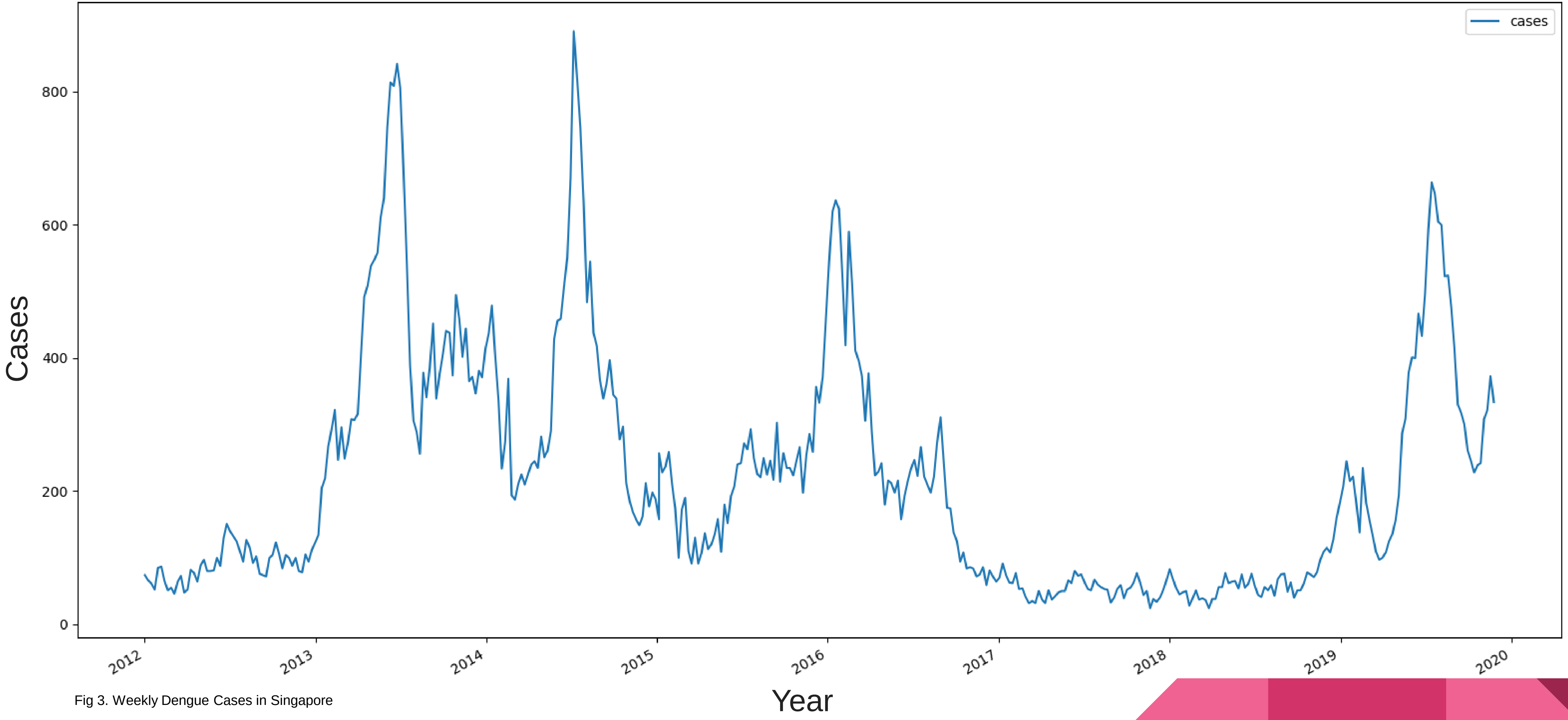
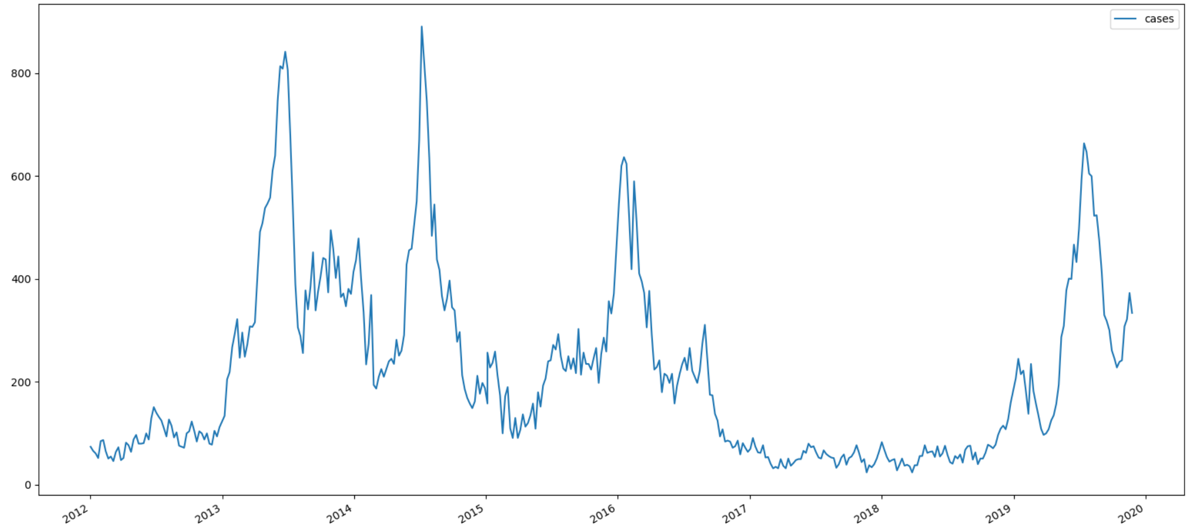


Fig 3. Weekly Dengue Cases in Singapore

Introduction



- Dengue Dataset: 412 observations
- Indexed by Weeks
- Start: January 2012. End: December 2019
- Obtained from NEA via *Data.gov.sg*
- Dengue cases will be the response variable

Project Objectives

- Create a parsimonious Time Series Model that can be used to accurately predict future Dengue Cases

- Future: 2 months look-ahead predictions

- Identify potential exogenous variables that can be used to improve the model

- Identify model limitations and suggest potential ways for improvement

Methodology

Phase I: Creating a model using only Past Observations

- Use **Box-Jenkins Methodology** to select appropriate **model orders**

Phase II: Enhancing the model with **Exogenous Variables**

- Individually insert variables into the best model
- Select best model based on performance on test set

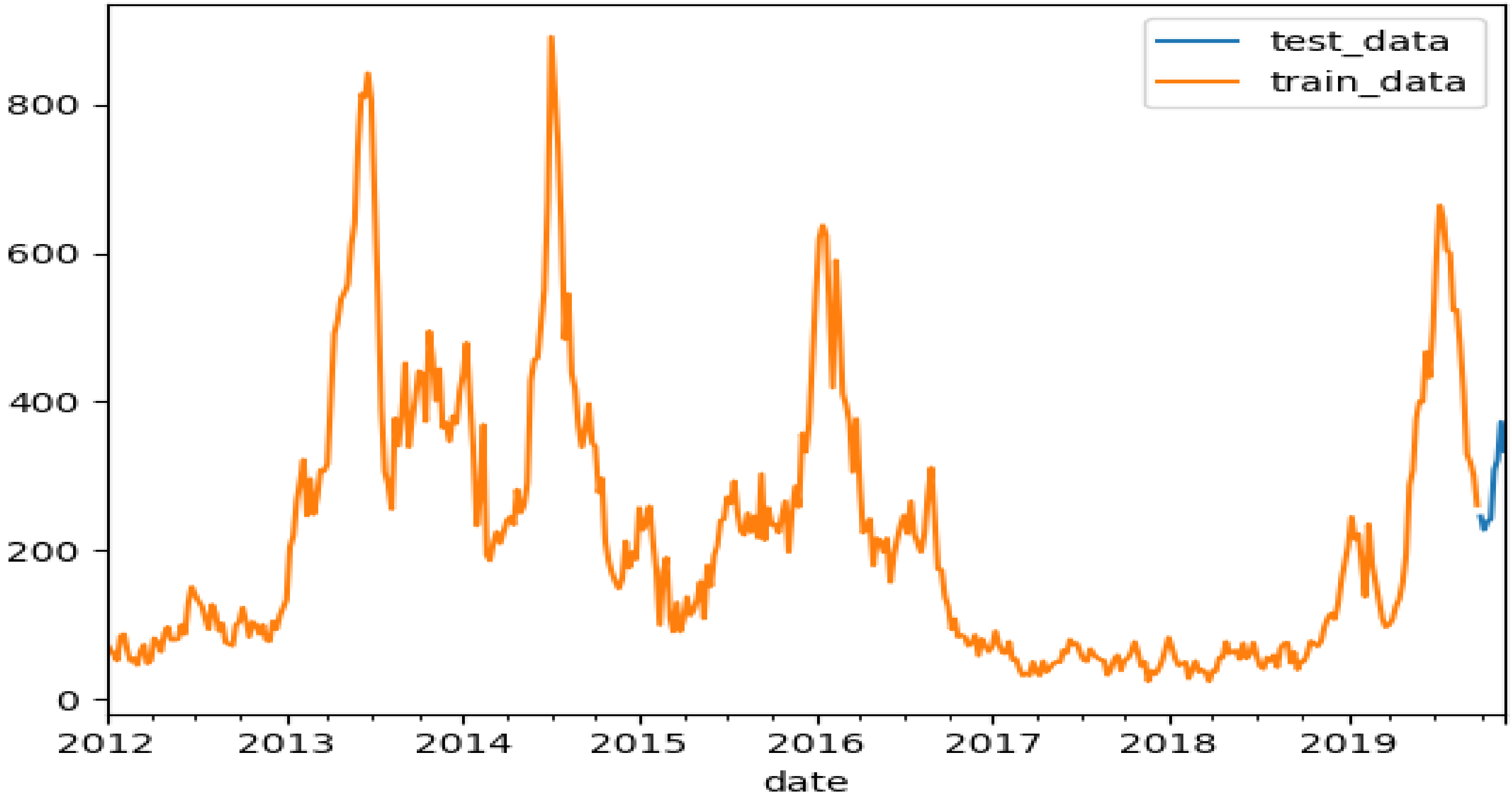
Train-test split

- Performance indicator is the **Mean Squared Error** on the Test Set
- Test Set: **Last 2 months** of the Dengue Dataset
- Avoid traditional randomized train-test split due to temporal structure

Benchmark model: **Persistence Model**

Methodology

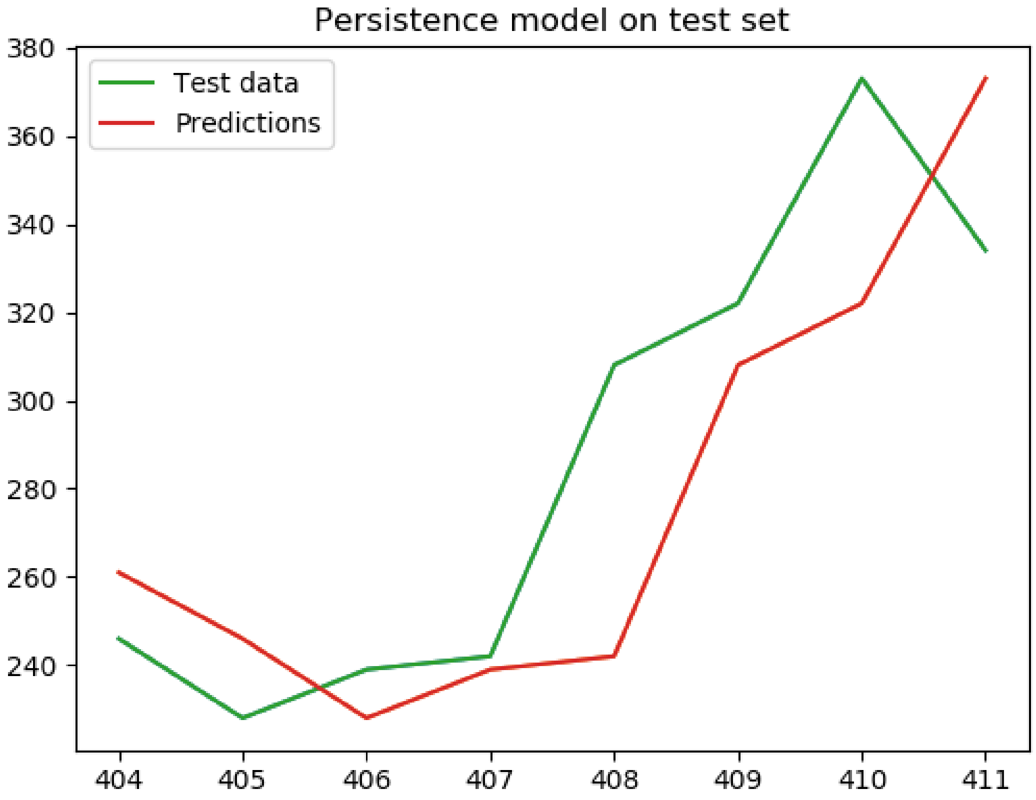
Train-Test Split



Methodology

Persistence Model

- Used as a Benchmark
- “Today’s Predictions” = “Tomorrow’s Predictions”



- Test MSE: 1169.125

ARIMA Terminology

AR: Auto-Regressive terms, lags of the time-series

MA: Moving Average, lags of the error terms

I: Integrated, corresponds to differencing done to make a series more stationary

p = number of AR terms

q = number of moving average terms

d = number of differences

Representation: ARIMA(p, d, q) model

Sample: ARIMA(1, 0, 1) is $y_t = \beta_1 y_{\{t-1\}} + \epsilon_{\{t-1\}} + \epsilon_t$

SARIMA Terminology

S: Seasonal

AR: Auto-Regressive terms, lags of the time-series

MA: Moving Average, lags of the error terms

I: Integrated, corresponds to differencing done to make a series more stationary

<u>ARIMA Component</u>

p = number of AR terms

q = number of moving average terms

d = number of differences

<u>Seasonal Component</u>

P = number of seasonal AR terms

Q = number of seasonal moving average terms

D = number of seasonal differences

m = periodicity of the season

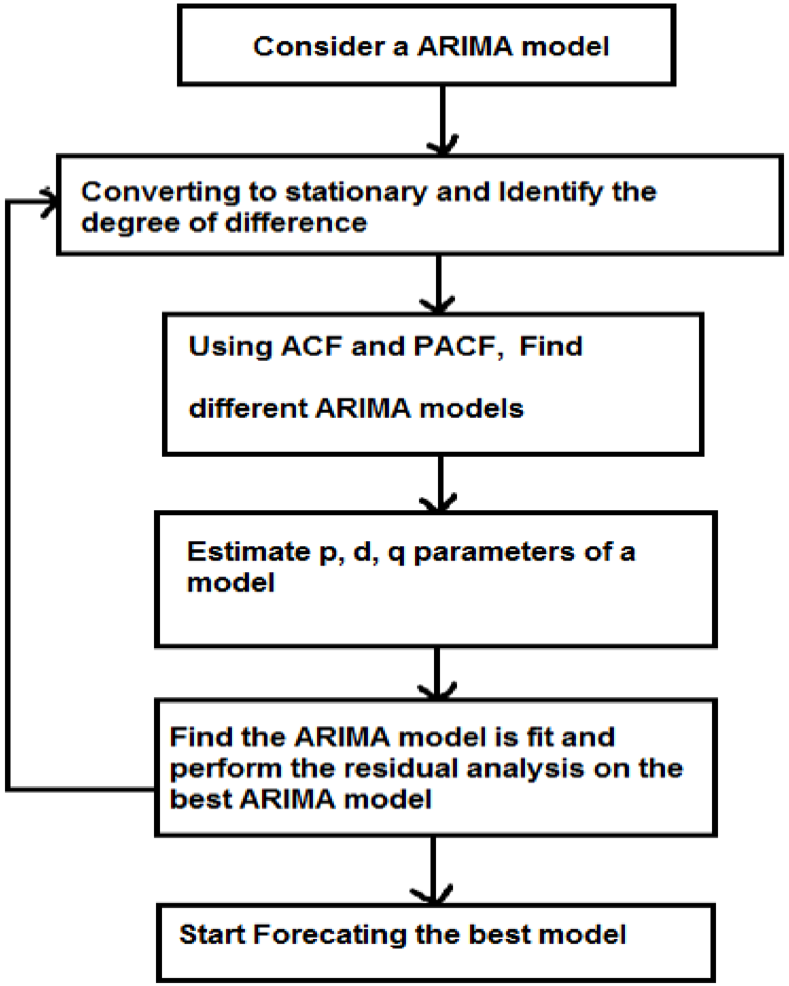
Representation: SARIMA(p, d, q) (P, D, Q, m) model

Sample: SARIMA(0,2,1) (0,0,1,12) is $Y_t - 2Y_{\{t-1\}} + Y_{\{t-2\}} = e_t + \Theta_1 e_{\{t-12\}} + \theta_1 e_{\{t-1\}} + \theta_1 \Theta_1 e_{\{t-13\}}$

Phase I: Creating a model using Past Observations

<u>Box Jenkins Methodology</u>

1. Ensure data is stationary by Augmented Dicky-Fuller Test
   a. If not stationary, difference the time series
1. Plot ACF & PACF to determine possible model order
2. Assess residuals of model
3. Assess MSE on test set



Phase I: Creating a model using Past Observations

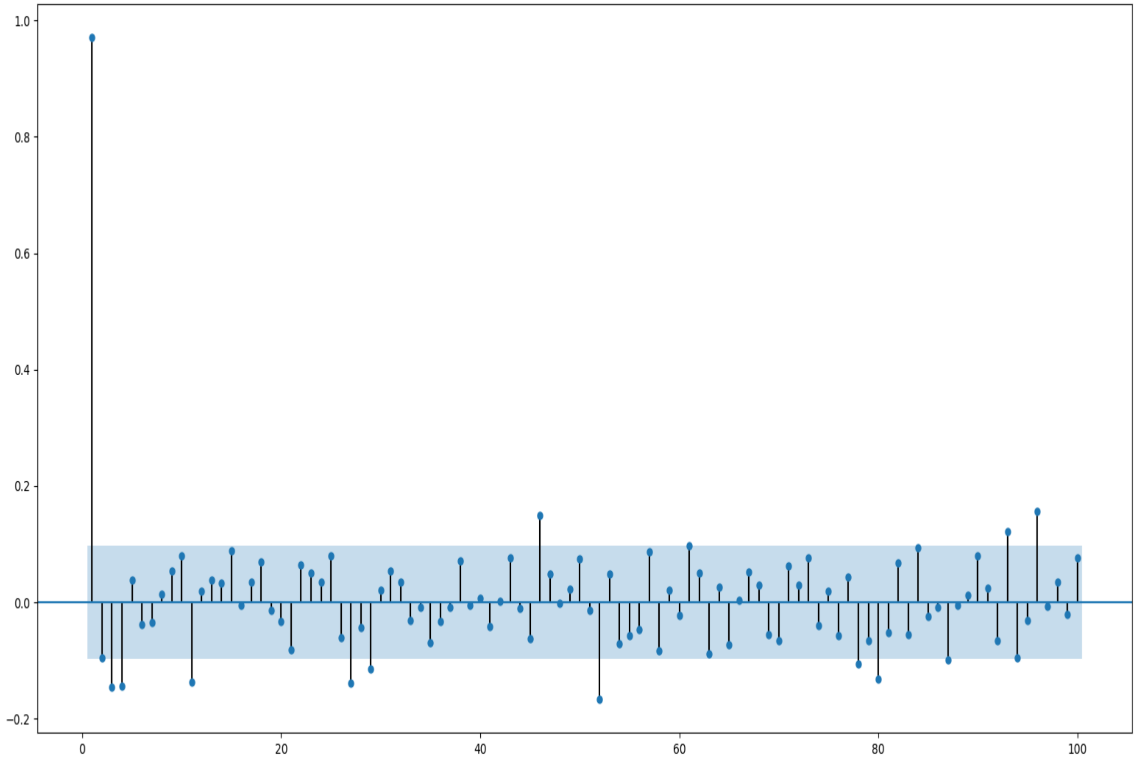
Determining Stationarity

P-value for ADF Test: 0.006948

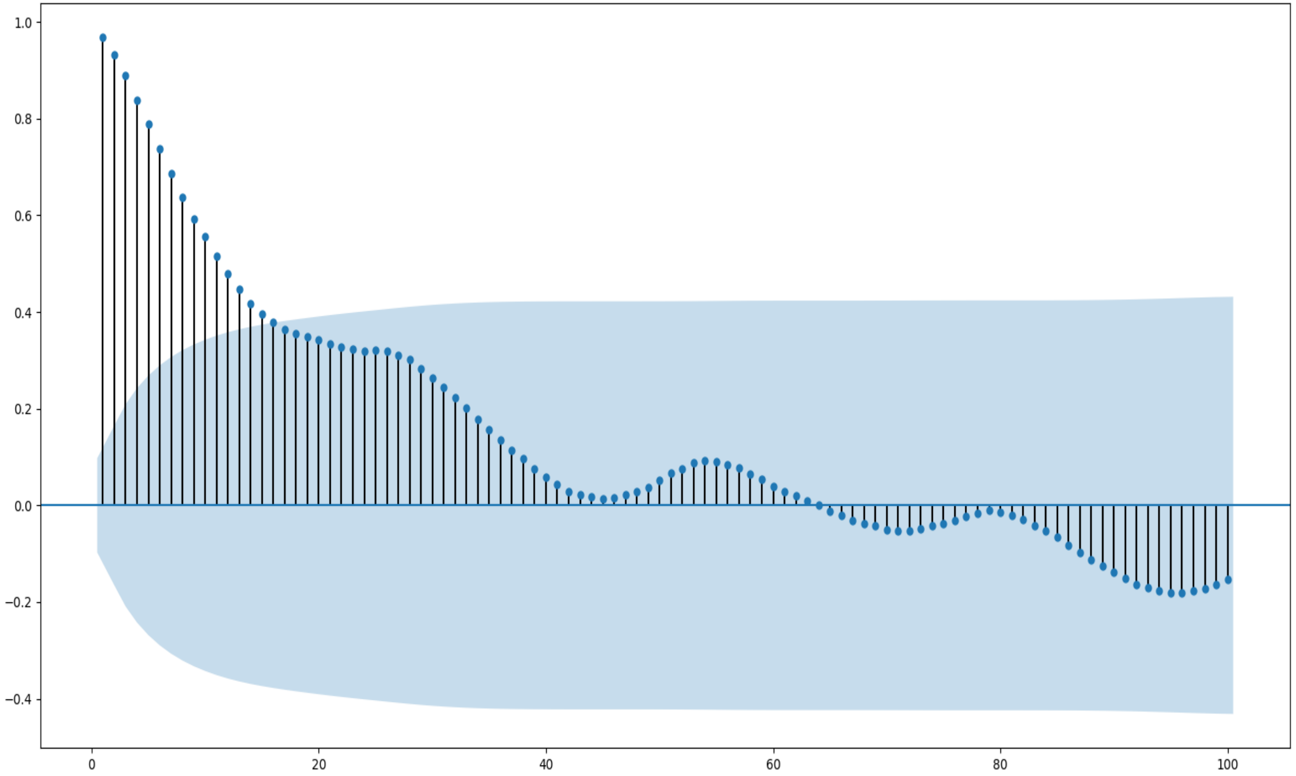
H_0 – Non-stationary time series

H_1 – Stationary

Partial Auto Correlation Plot



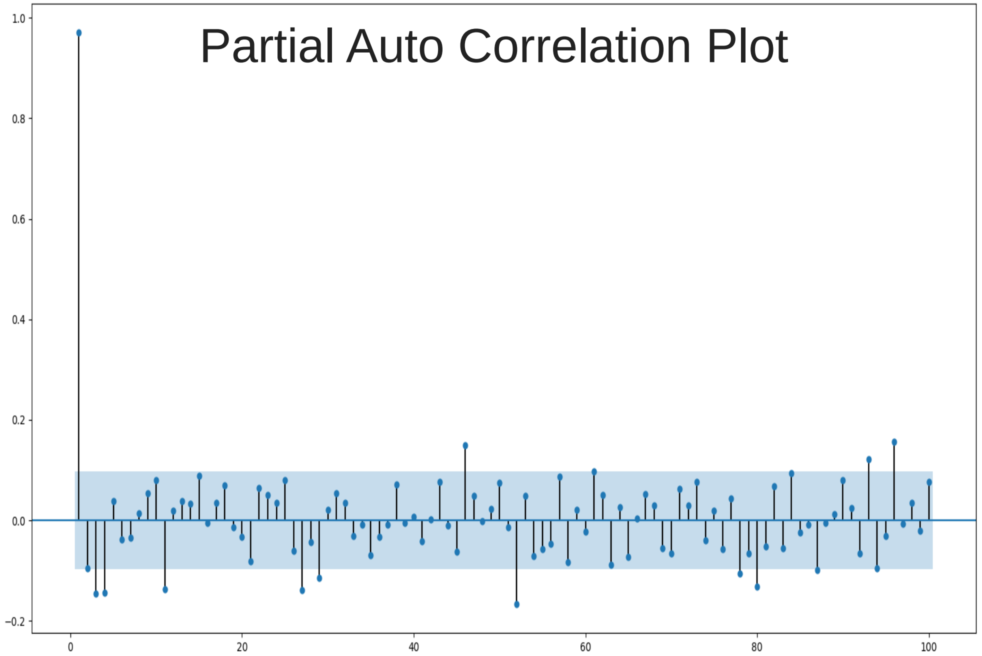
Auto Correlation Plot

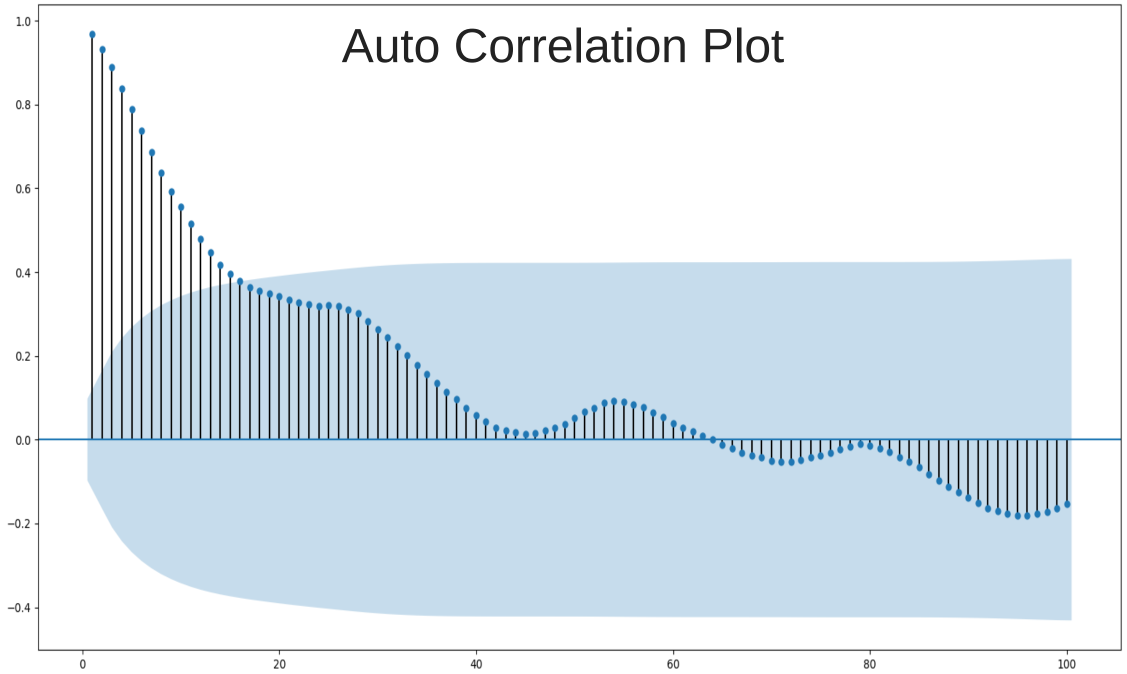


Phase I: Creating a model using Past Observations

Using ACF and PACF to choose model order

	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

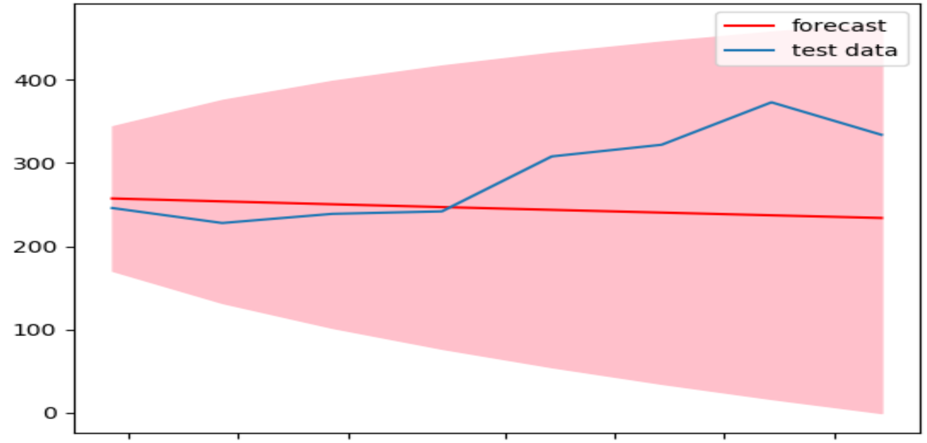




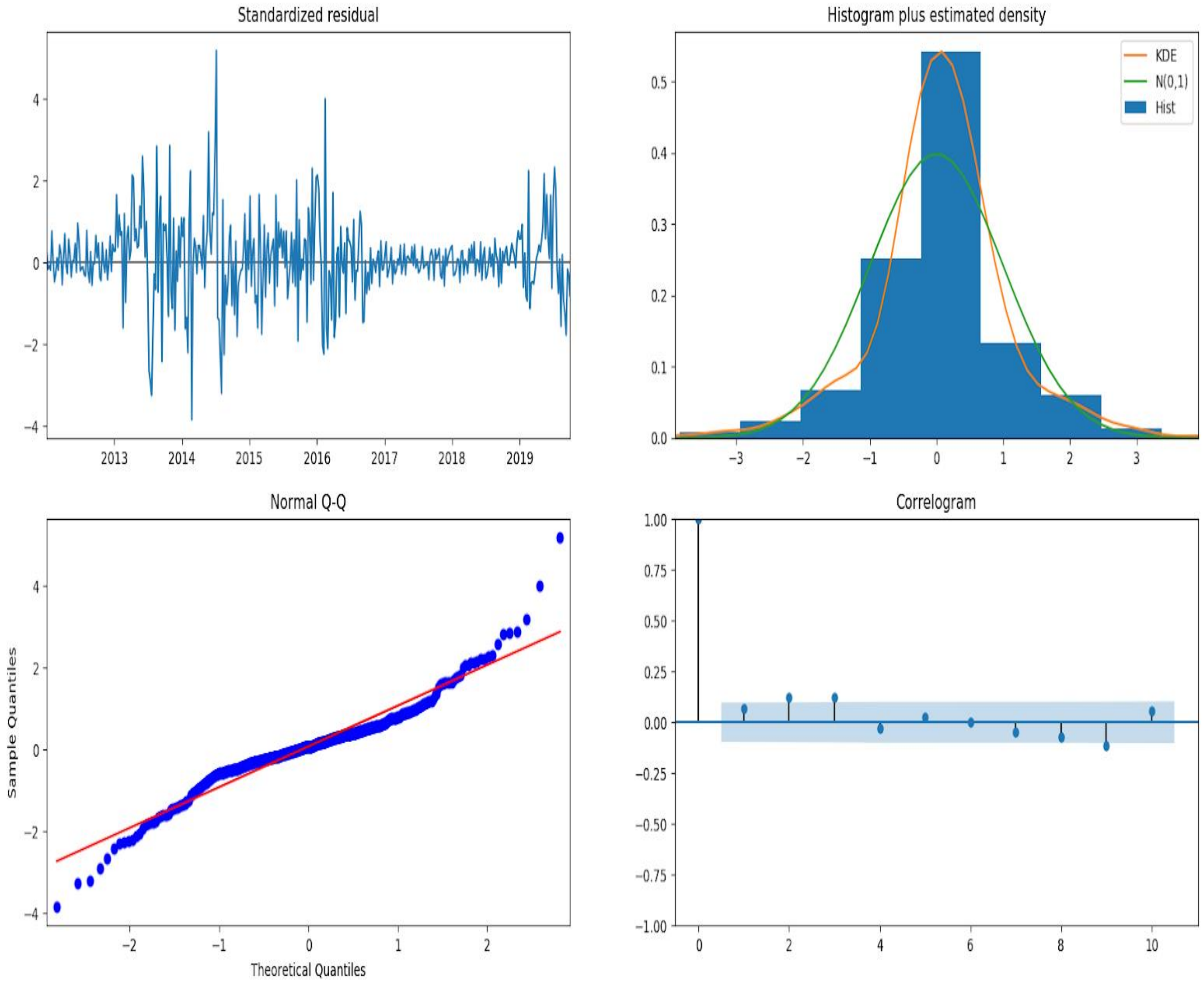
Phase I: Creating a model using Past Observations

<u>Fitting AR(1) Model</u>

Model equation: $y_t = \beta_1 y_{\{t-1\}} + \epsilon_t$



Test MSE: 5017



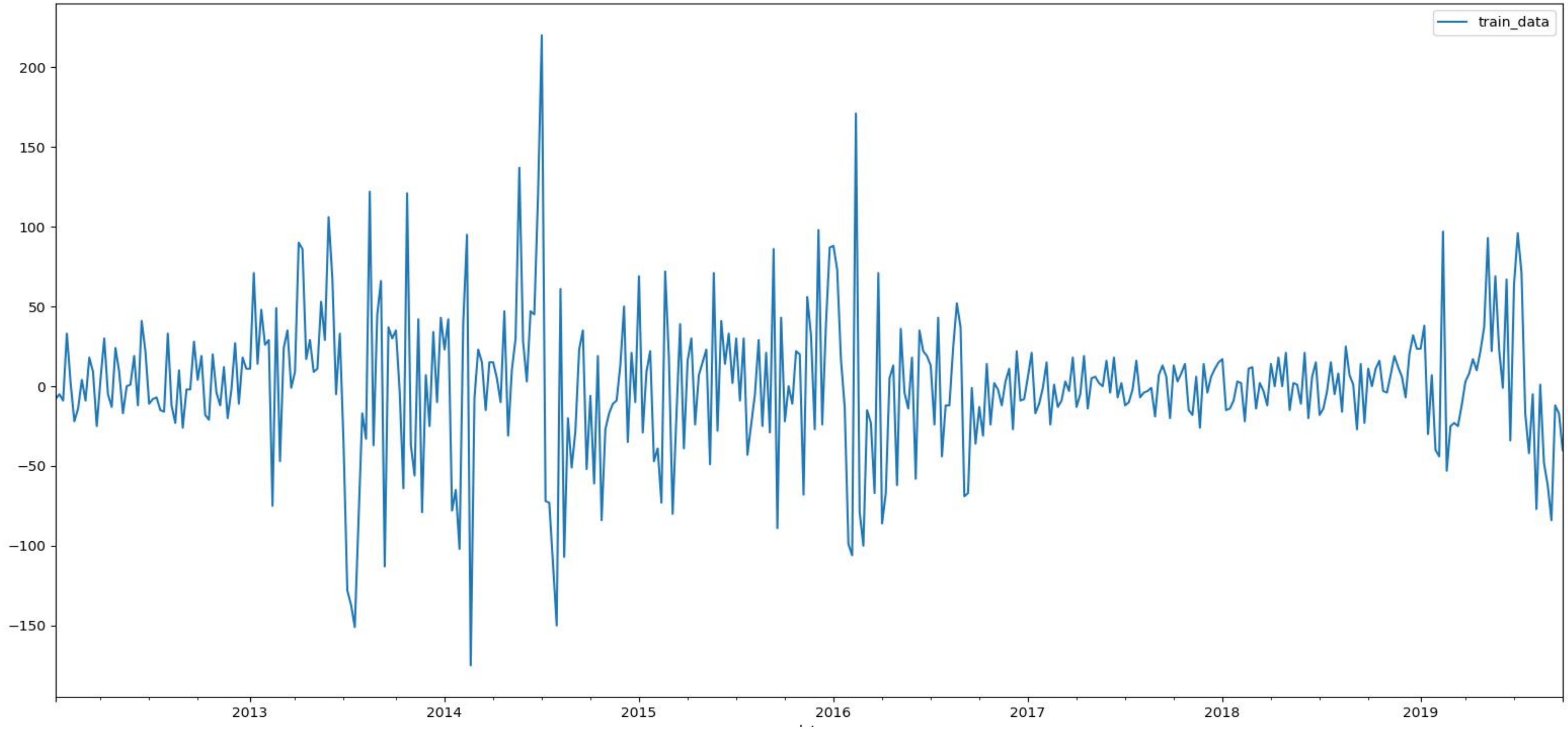
Phase I: Creating a model using Past Observations

First-order Differencing

$y = y_t - y_{\{t-1\}}$

Determining Stationarity

P-value for ADF Test: 1.41×10^{-15}



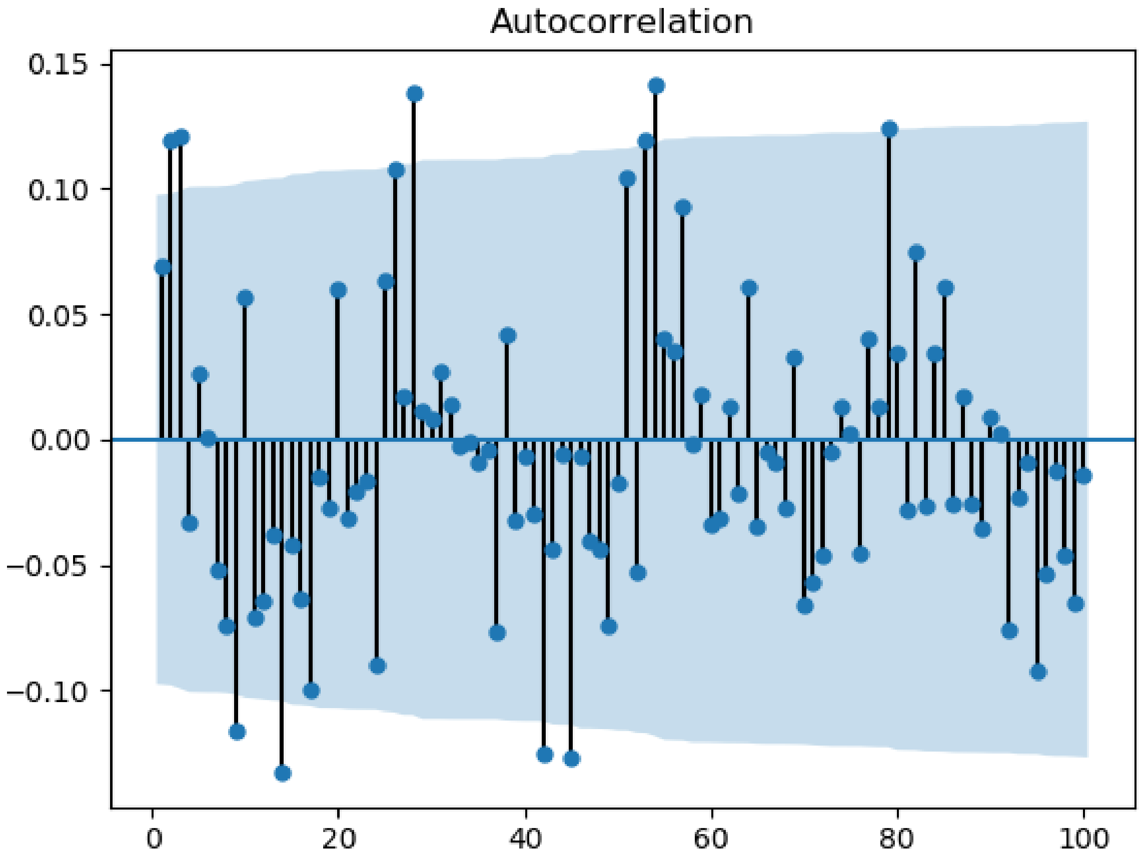
Phase I: Creating a model using Past Observations

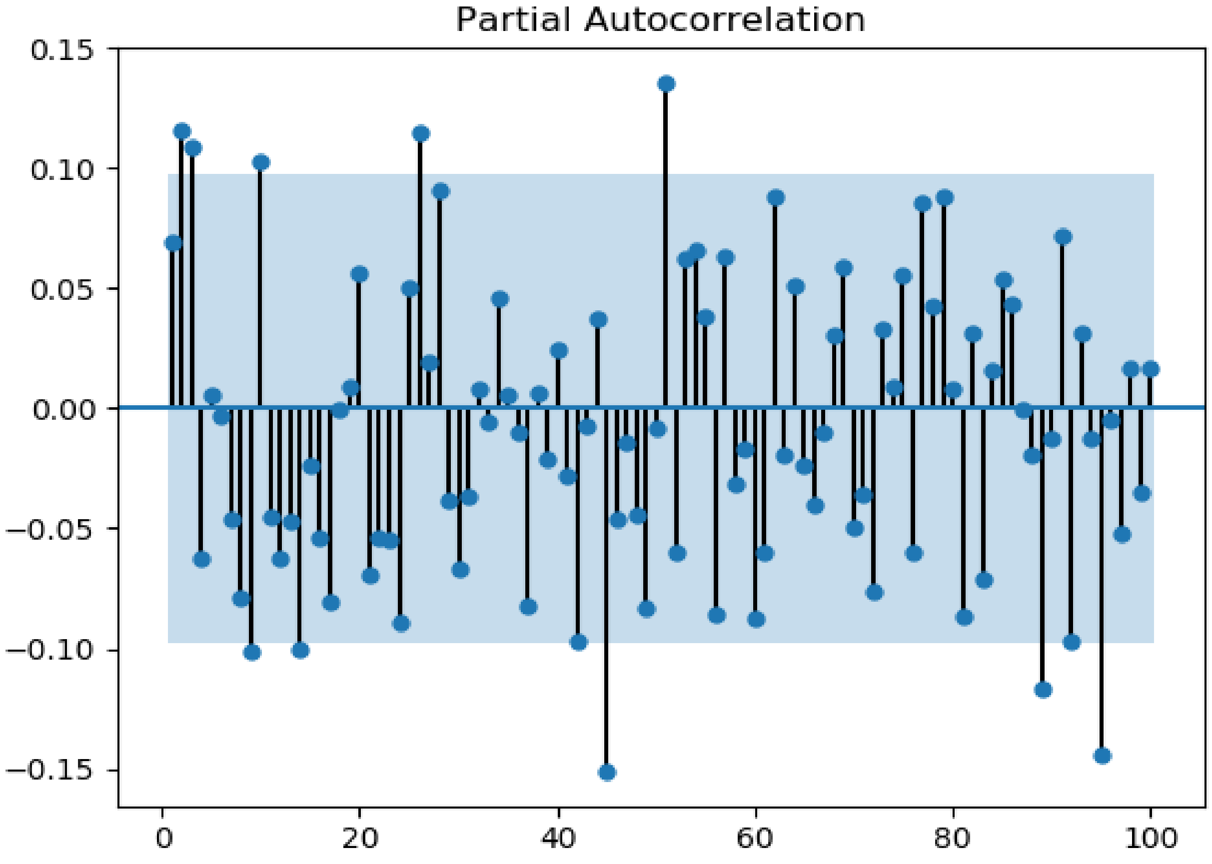
First-order Differencing

$$y = y_t - y_{\{t-1\}}$$

Determining Stationarity

P-value for ADF Test: 1.41×10^{-15}





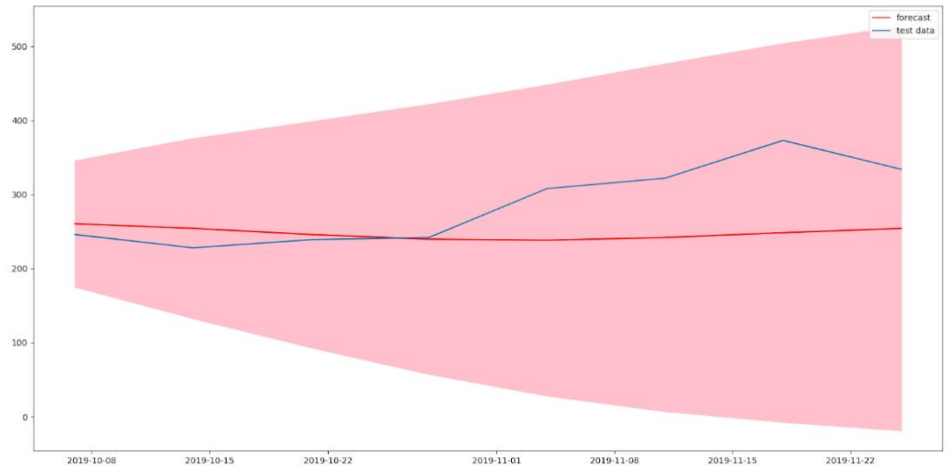
Phase I: Creating a model using Past Observations

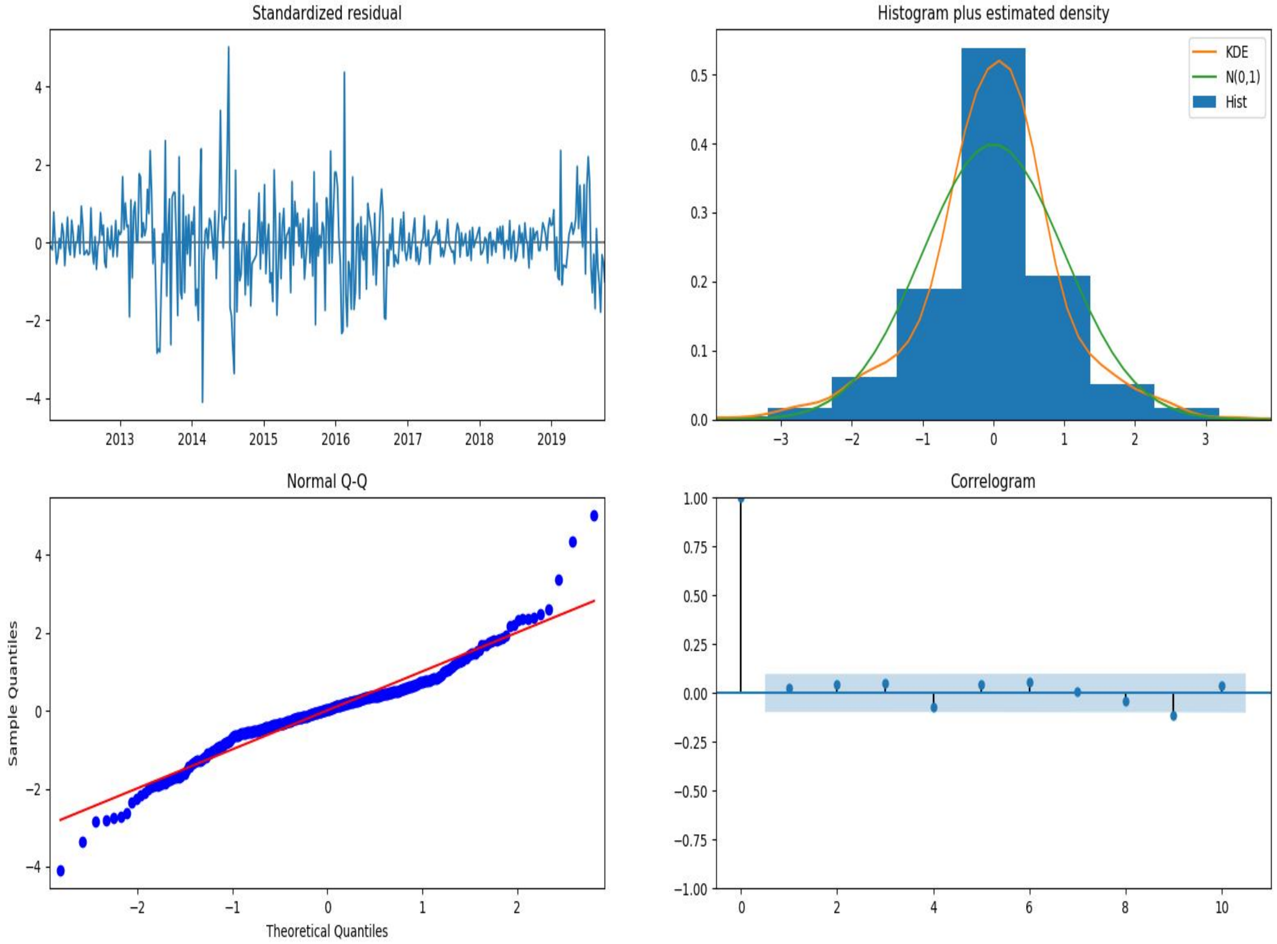
Using PMDARIMA library to search for the best model

Order	AIC	Order	AIC
(2,1,2)	4207.247	(1,1,2)	4211.472
(0,1,0)	4215.667	(2,1,1)	4212.224
(1,1,0)	4215.737	(1,1,1)	4213.104
(0,1,1)	4216.086		

Phase I: Creating a model using Past Observations

Fitting ARIMA(2,1,2) Model





Test MSE: 4263.6

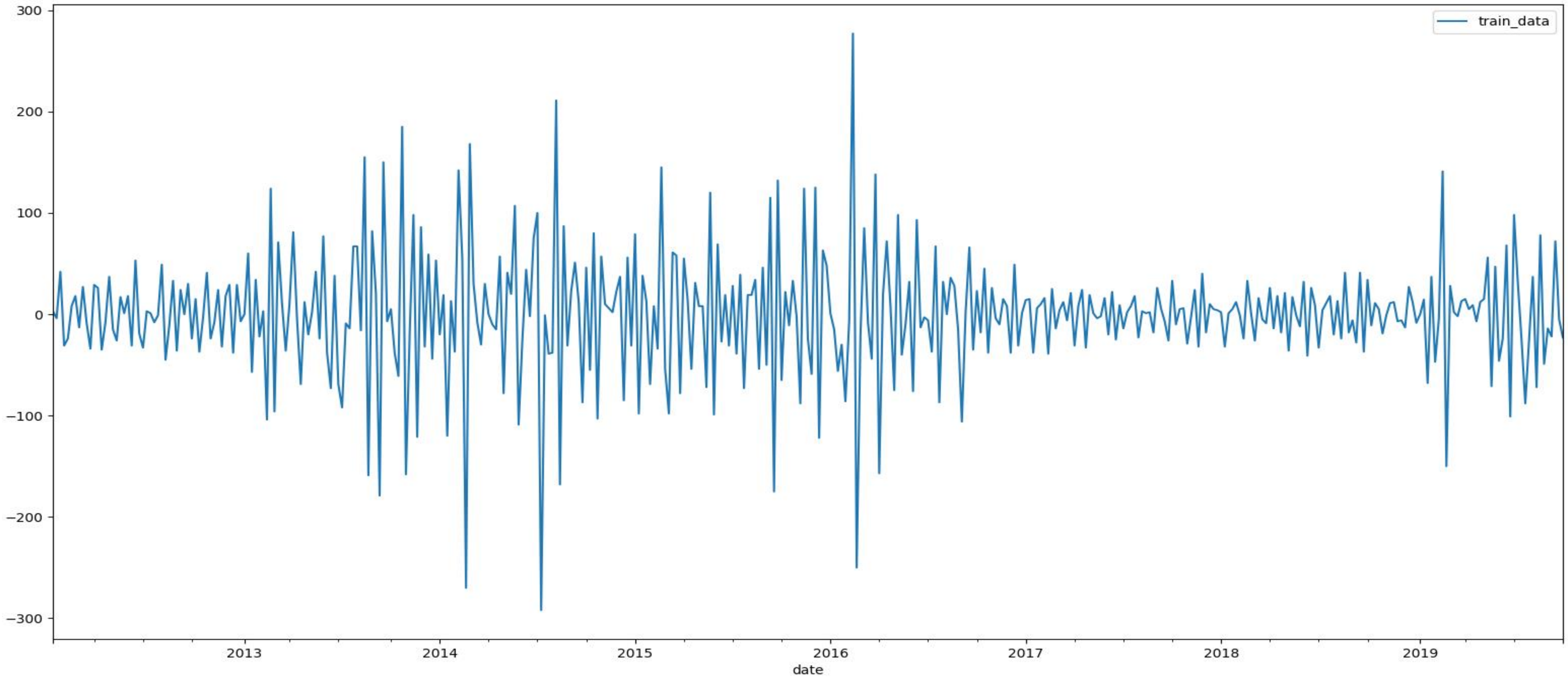
Phase I: Creating a model using Past Observations

Second-order Differencing

$y = y_{t-1} - y_{\{t-2\}}$

Determining Stationarity

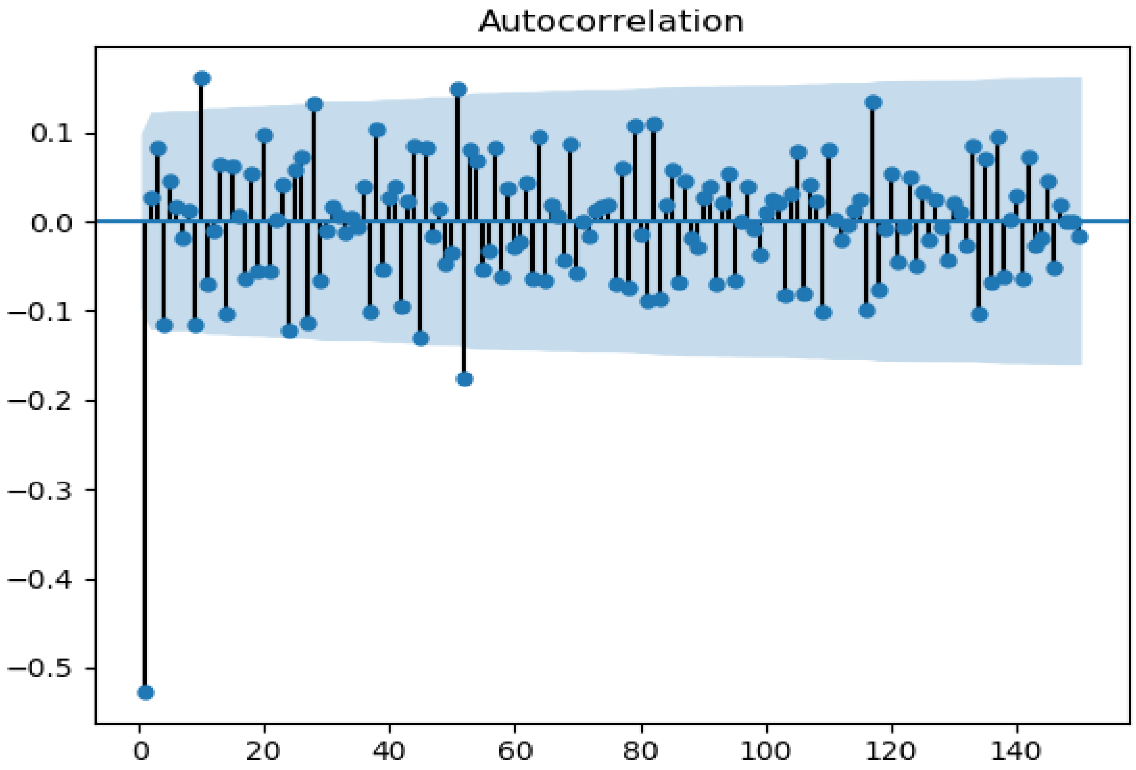
P-value for ADF Test: 9.7274 x 10^{-13}



Phase I: Creating a model using Past Observations

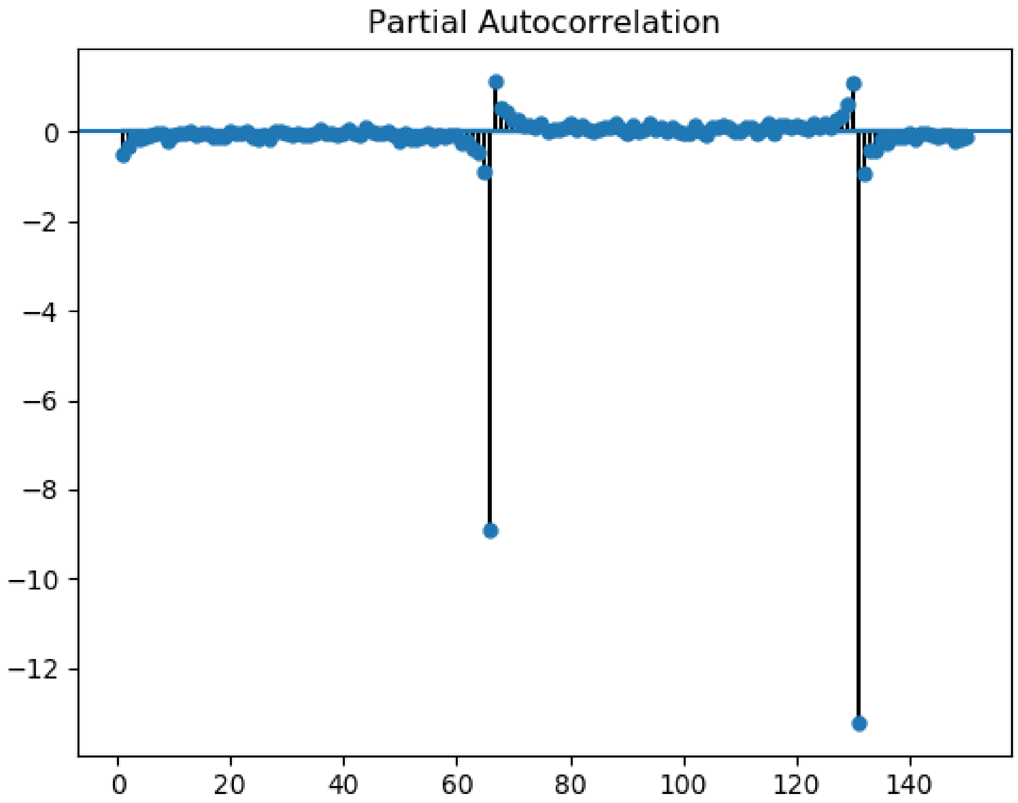
Second-order Differencing

$$y = y_{t-1} - y_{\{t-2\}}$$



Determining Stationarity

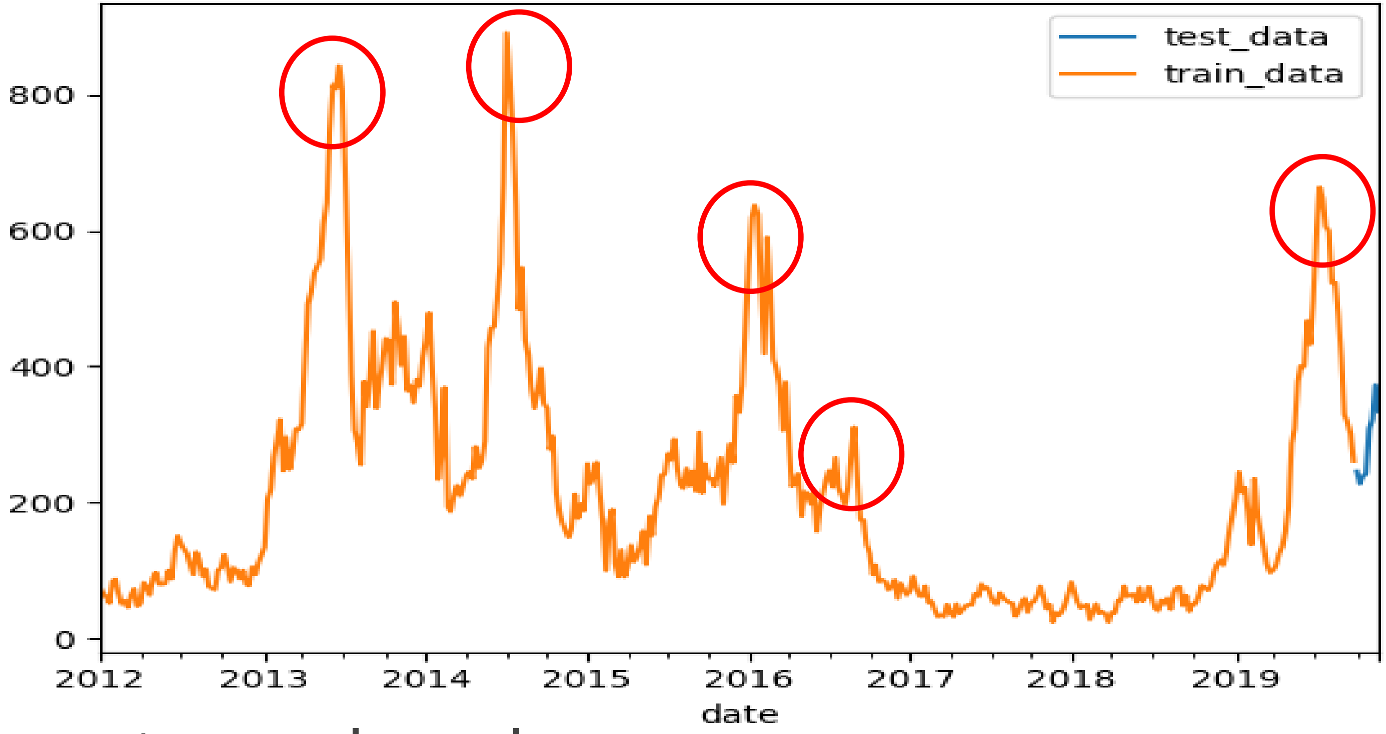
P-value for ADF Test: 9.7274×10^{-13}



Very negative ACF/PACF suggest over-differencing

Phase I: Creating a model using Past Observations

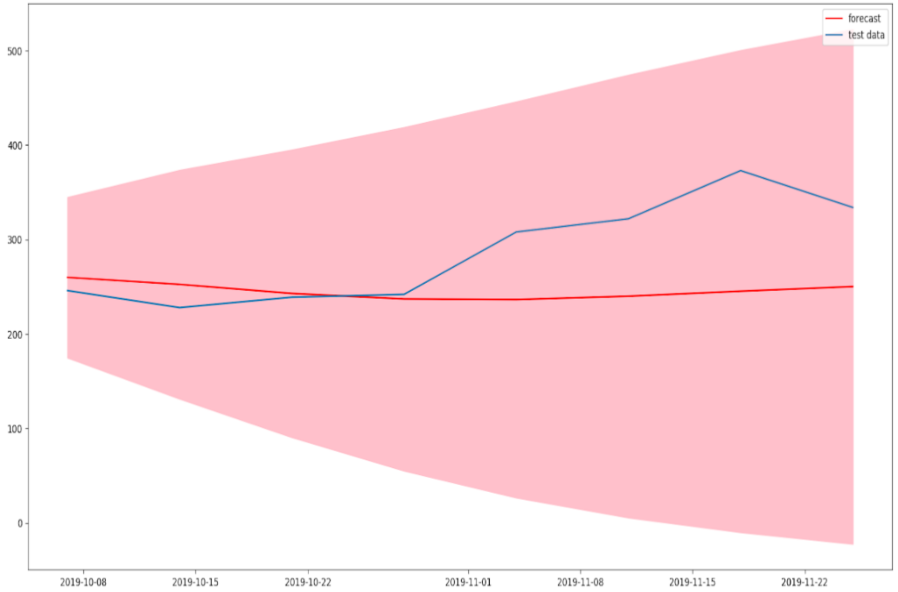
Considering General Seasonality

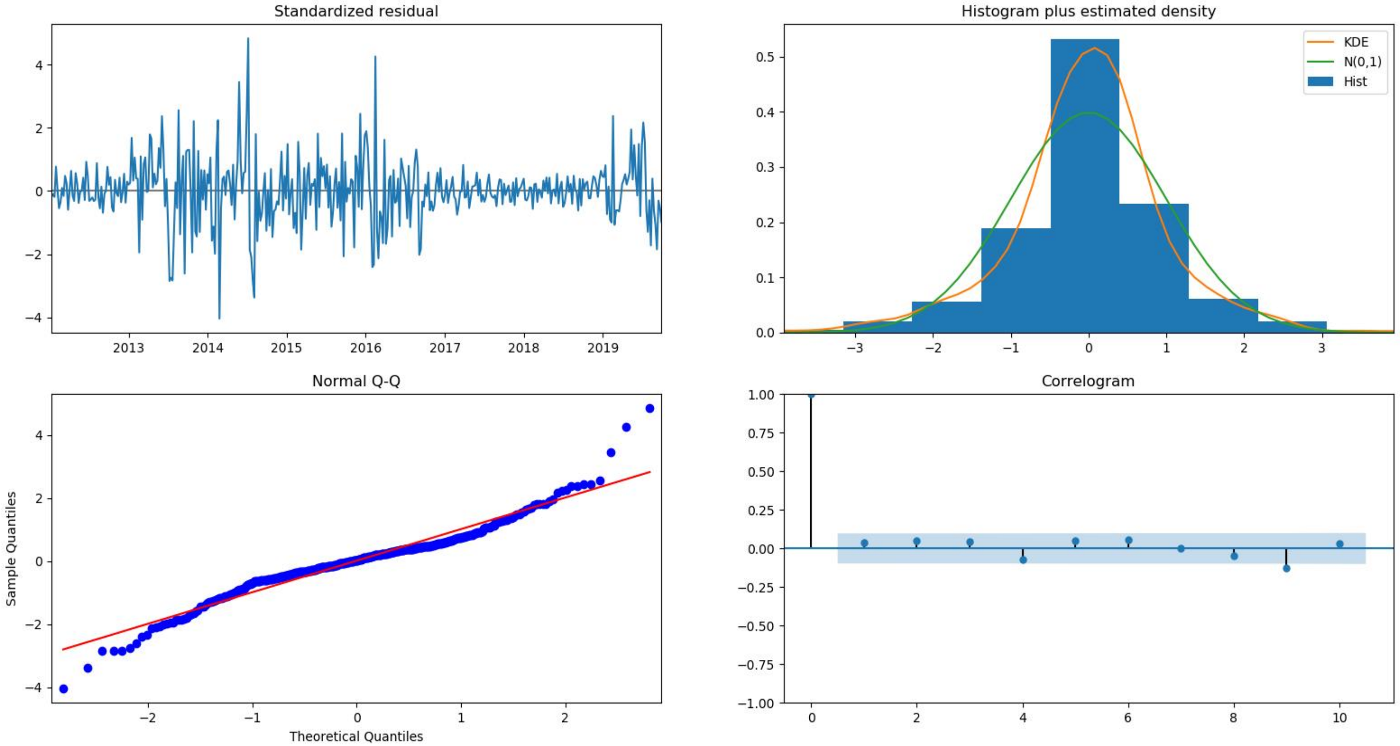


- Intuition suggest a yearly cycle
- Search model space using PMDARIMA with d = 1, m = 52 weeks (1 year)
- PMDARIMA suggests using (2,1,2) (1,0,0,52) SARIMA model

Phase I: Creating a model using Past Observations

Fitting SARIMA(2,1,2) (1,0,0,52) Model

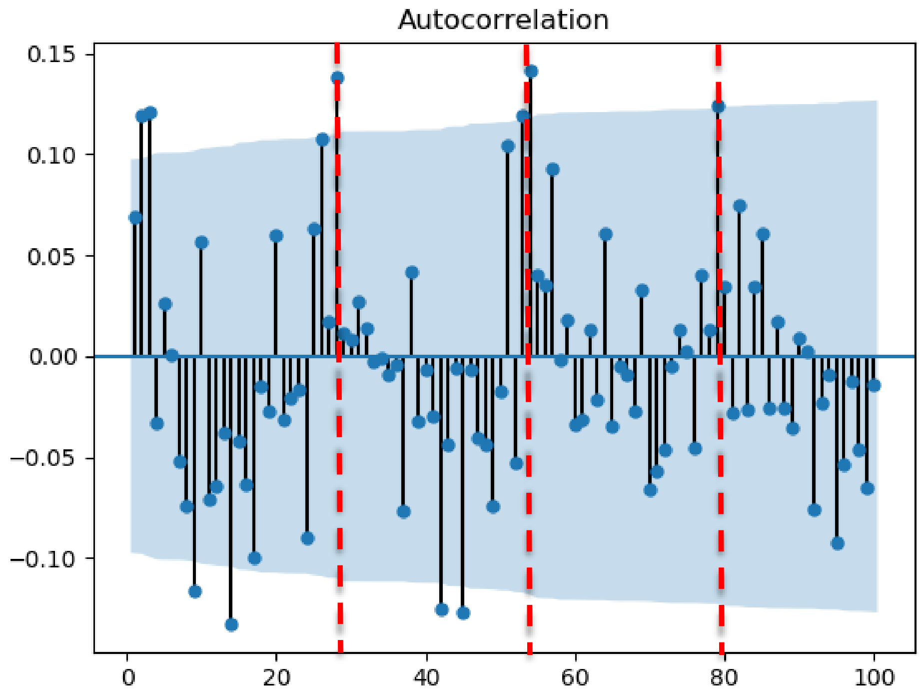


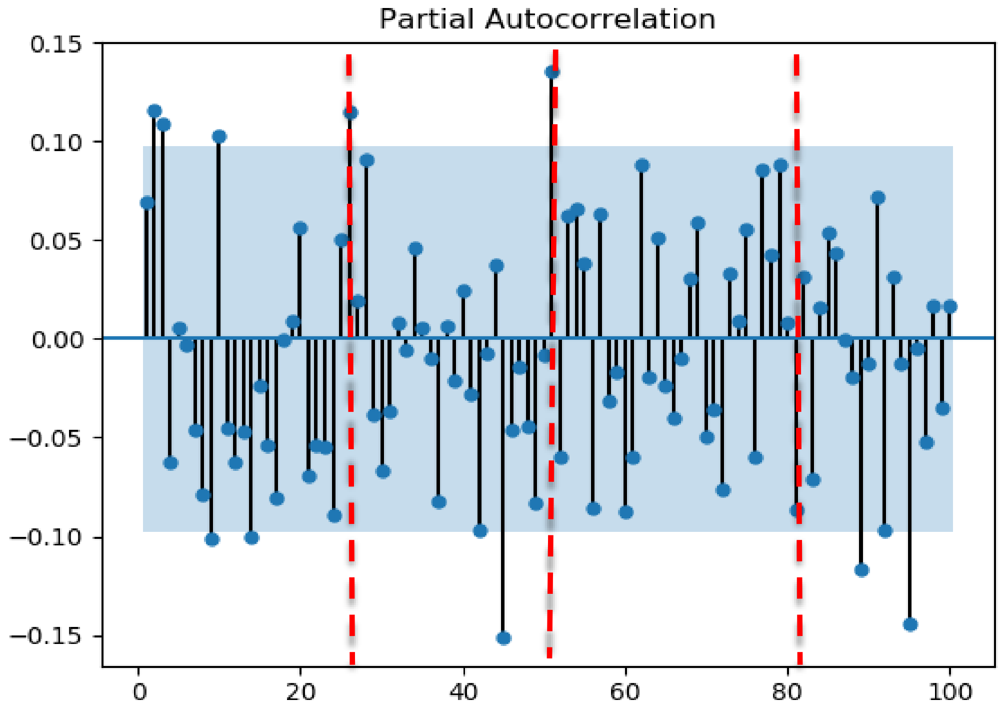


Test MSE: 4497.3

Phase I: Creating a model using Past Observations

Considering First Order Differenced Seasonality

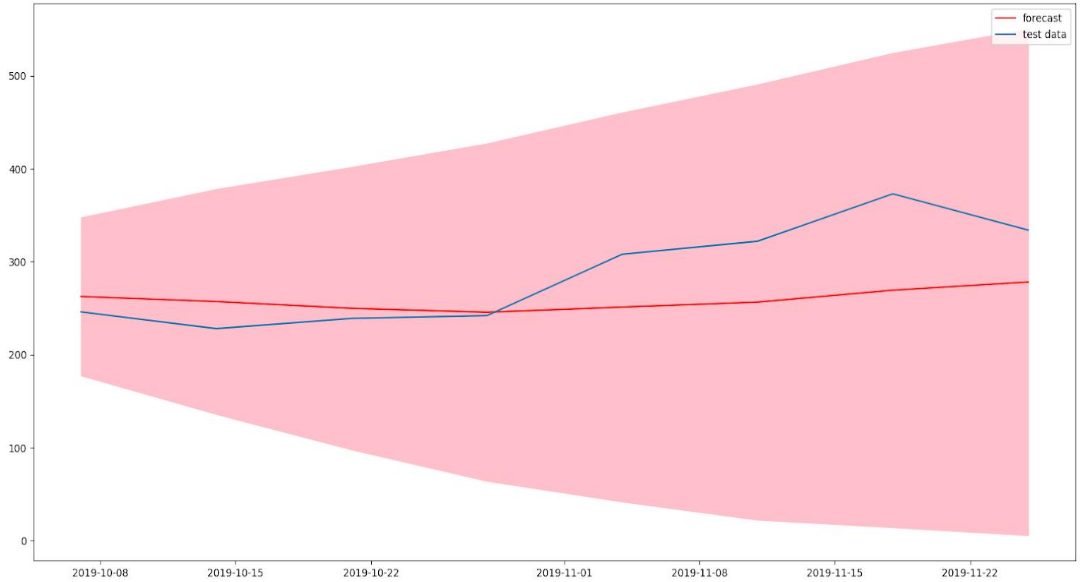


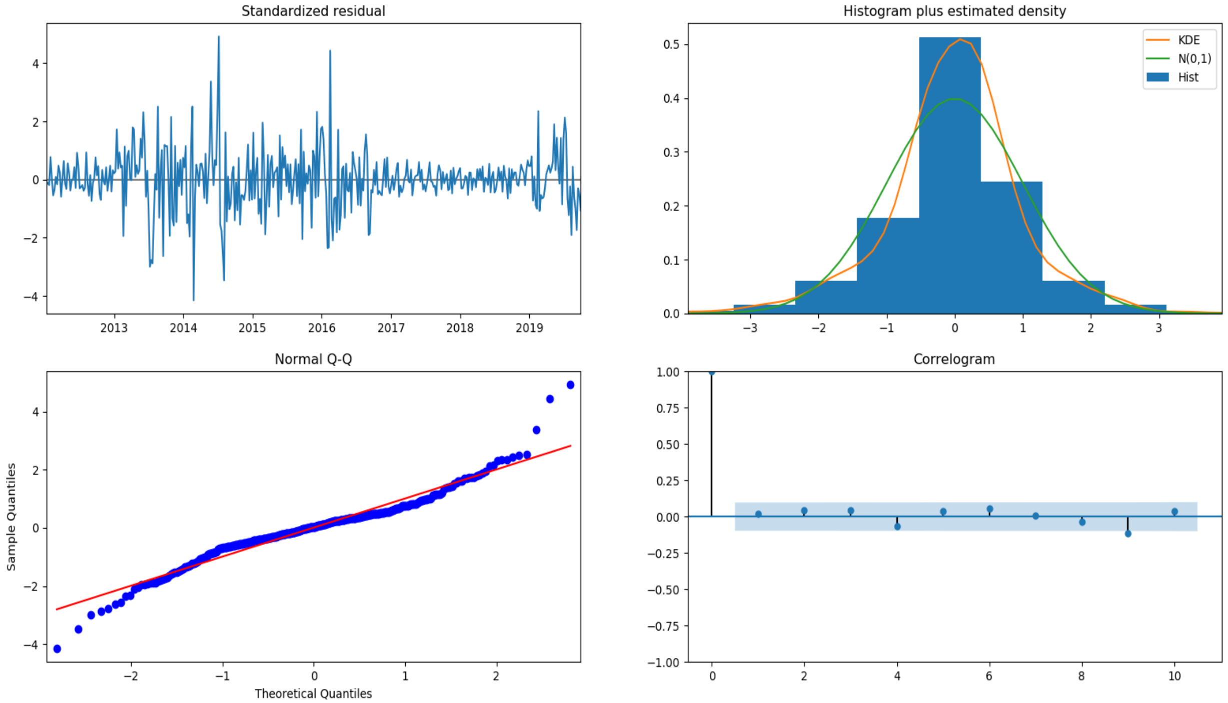


- An approximate cycle occurs at m = 25
- Search model space using PMDARIMA with d = 1, m = 25 weeks
- PMDARIMA suggests using (2,1,2) (0,0,1,25) SARIMA model

Phase I: Creating a model using Past Observations

<u>Fitting SARIMA(2,1,2) (0,0,1,25) Model</u>





Test MSE: 2829.6

Best MSE thus far

Exploring Possible Variables

3 Possible Sources of Data

1. **data.gov.sg Data Portal**

2. Google search terms related to dengue symptoms

3. Scraping 41 weather stations through the Meteorological Service Singapore’s website

Exploring Possible Variables

Summary of Possible Variables

data.gov.sg Portal	Google Search Term	Islandwide Weather Stations
SG Population	Fever	Mean / Max / Minimum Temperature
Changi Total / Max Rainfall	Rash	Max Wind Speed
Changi Daily Minimum / Max Temp	Headache	Avg Highest Rainfall in 30 min
Changi Humidity	Joint Pain	Avg Highest Rainfall in 60 min
Changi No. of Rainy Days	Nausea	Avg Highest Rainfall in 120 min
Changi Mean Temperature	Eye Pain	
	Dengue	

Data Cleaning

Re-aligning time index

- Datasets were interpolated to ensure index is aligned with Dengue's index
- Shape of variables were preserved

Cleaning missing data

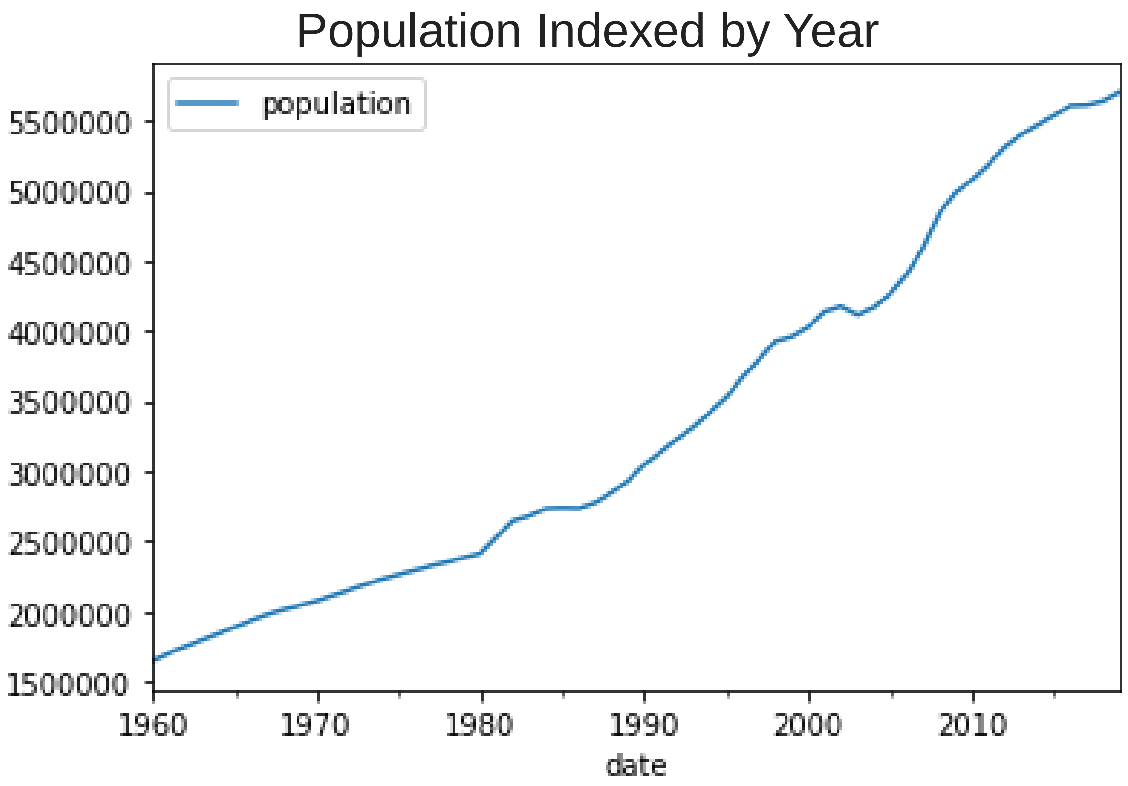
- Scrapped weather data had numerous missing values
- Missing values were ignored
- All 41 stations' data were averaged out

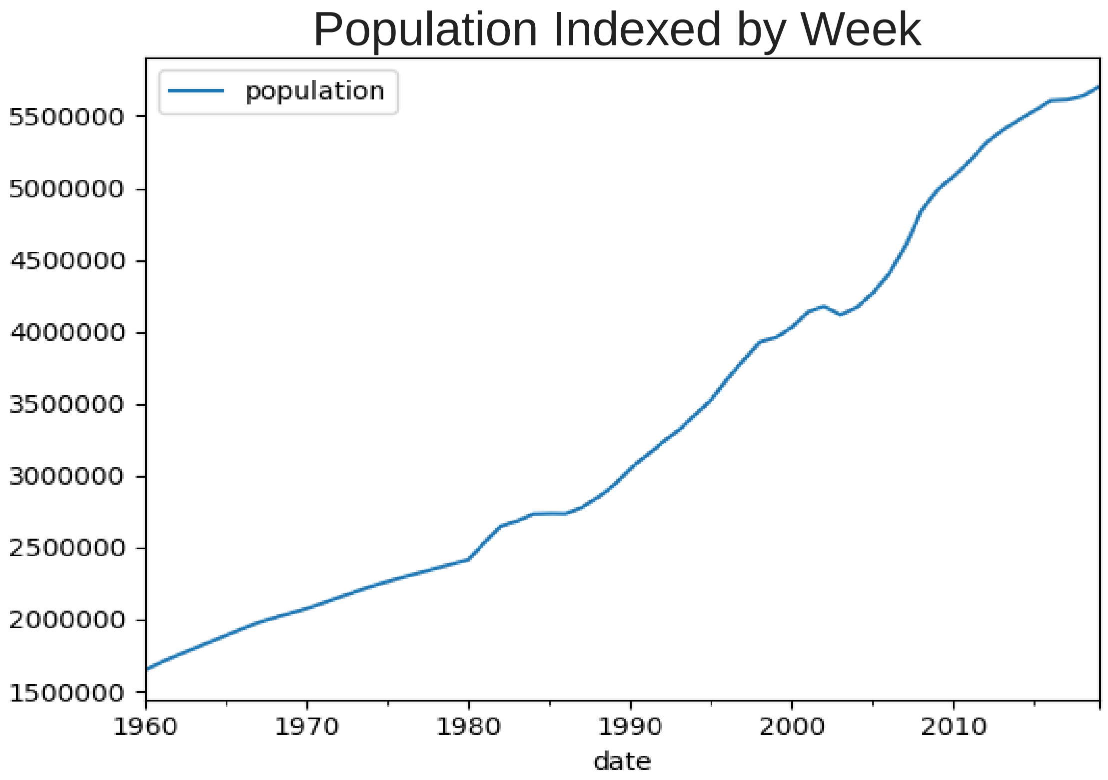
Scaling datasets

- Ensures significance with the Dengue dataset

Re-Alligning Time Index

Population Data originally was indexed by years





Cleaning Missing Data

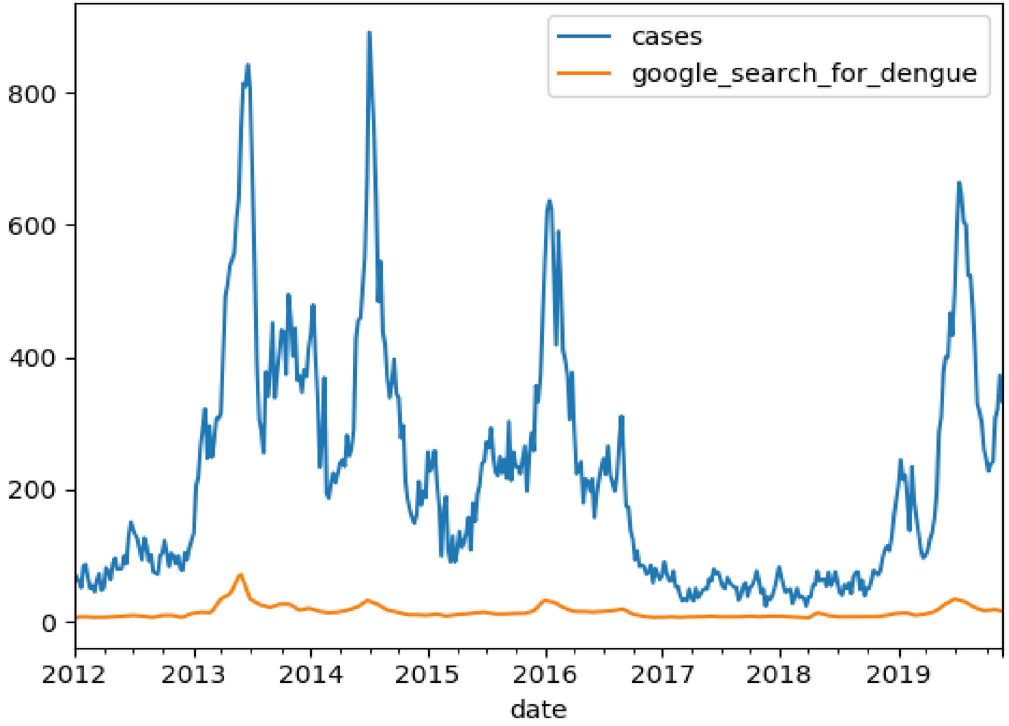
- Scrapped data was daily
- Entries were grouped by weeks
- Entries were averaged out for each week
- Missing entries were not considered during calculations

	Station	Daily Rainfall Total (mm)	Highest 30 Min Rainfall (mm)	Highest 60 Min Rainfall (mm)	Highest 120 Min Rainfall (mm)	Mean Temperature	Maximum Temperature	Minimum Temperature	Mean Wind Speed (km/h)	Max Wind Speed (km/h)	DateTime
0	Kampong Bahru	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-01
1	Kampong Bahru	6.8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-02
2	Kampong Bahru	1.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-03
3	Kampong Bahru	0.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-04
4	Kampong Bahru	30.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-05
5	Kampong Bahru	8.1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-06
6	Kampong Bahru	0.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-07
7	Kampong Bahru	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-08
8	Kampong Bahru	2.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-09
9	Kampong Bahru	5.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-10
10	Kampong Bahru	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-11
11	Kampong Bahru	43.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-12
12	Kampong Bahru	0.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-13
13	Kampong Bahru	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-14
14	Kampong Bahru	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2000-01-15

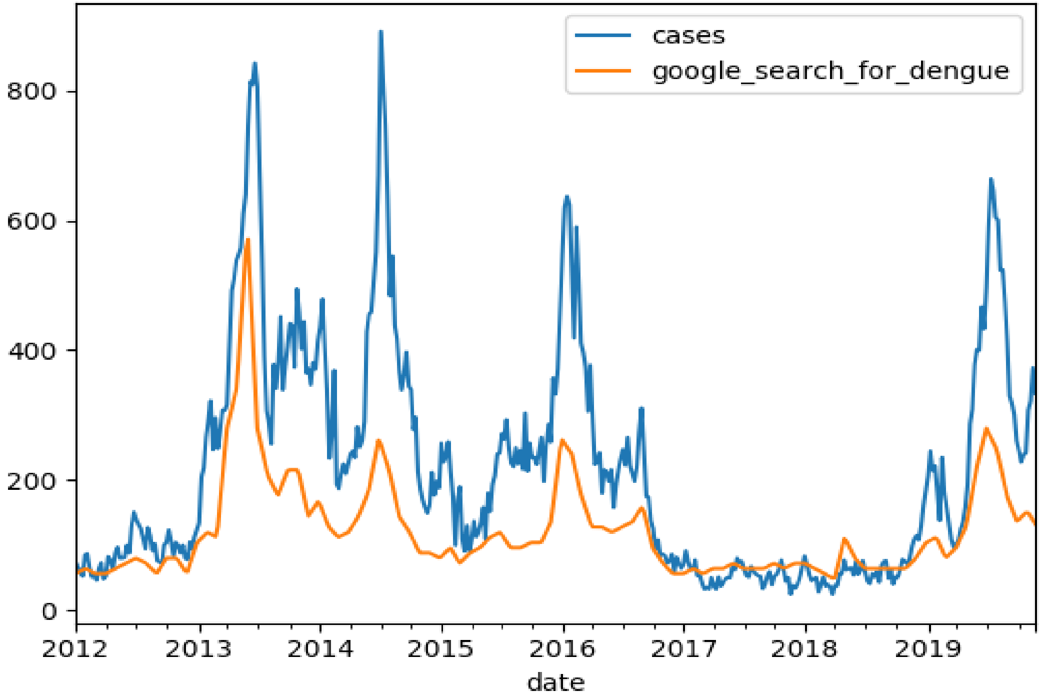
Scaling Datasets

Datasets were scaled to ensure they were significant to the response variable, Dengue Cases

‘Dengue’ Search Term before Scaling



‘Dengue’ Search Term after Scaling



Phase II: Enhancing the model with Exogenous Variables

SARIMAX models of order (2,1,2) (0,0,1,25) were fit with each Exogenous Variable (27 in all)

Each model was evaluated on the test set

Using the top 5 models, a full model was created

Exog Var	Test MSE
fever search	779.079
temp_mean_daily_max changi	1756.088
Max temp changi	1924.362
maximum_rainfall_in_a_day changi	2006.329
Rash search	2095.711

SARIMA Test MSE: 2829

Phase II: Enhancing the model with Exogenous Variables

Backward Elimination was carried out on the full model with 5 exogenous variables

	coef	std err	z	P>|z|	[0.025	0.975]
google_search_for_fever	0.7037	0.184	3.829	0.000	0.344	1.064
google_search_for_rash	-0.0954	0.205	-0.464	0.643	-0.498	0.307
maximum_rainfall_in_a_day	-0.1797	0.101	-1.785	0.074	-0.377	0.018
max_temperature	-1.4254	2.361	-0.604	0.546	-6.054	3.203
temp_mean_daily_max	-0.9199	2.347	-0.392	0.695	-5.519	3.680
ar.L1	0.1066	0.105	1.019	0.308	-0.098	0.312
ar.L2	-0.7772	0.096	-8.055	0.000	-0.966	-0.588
ma.L1	-0.1095	0.088	-1.249	0.212	-0.281	0.062
ma.L2	0.8678	0.080	10.862	0.000	0.711	1.024
ma.S.L25	0.0644	0.051	1.265	0.206	-0.035	0.164
sigma2	1830.5764	83.687	21.874	0.000	1666.553	1994.600

Phase II: Enhancing the model with Exogenous Variables

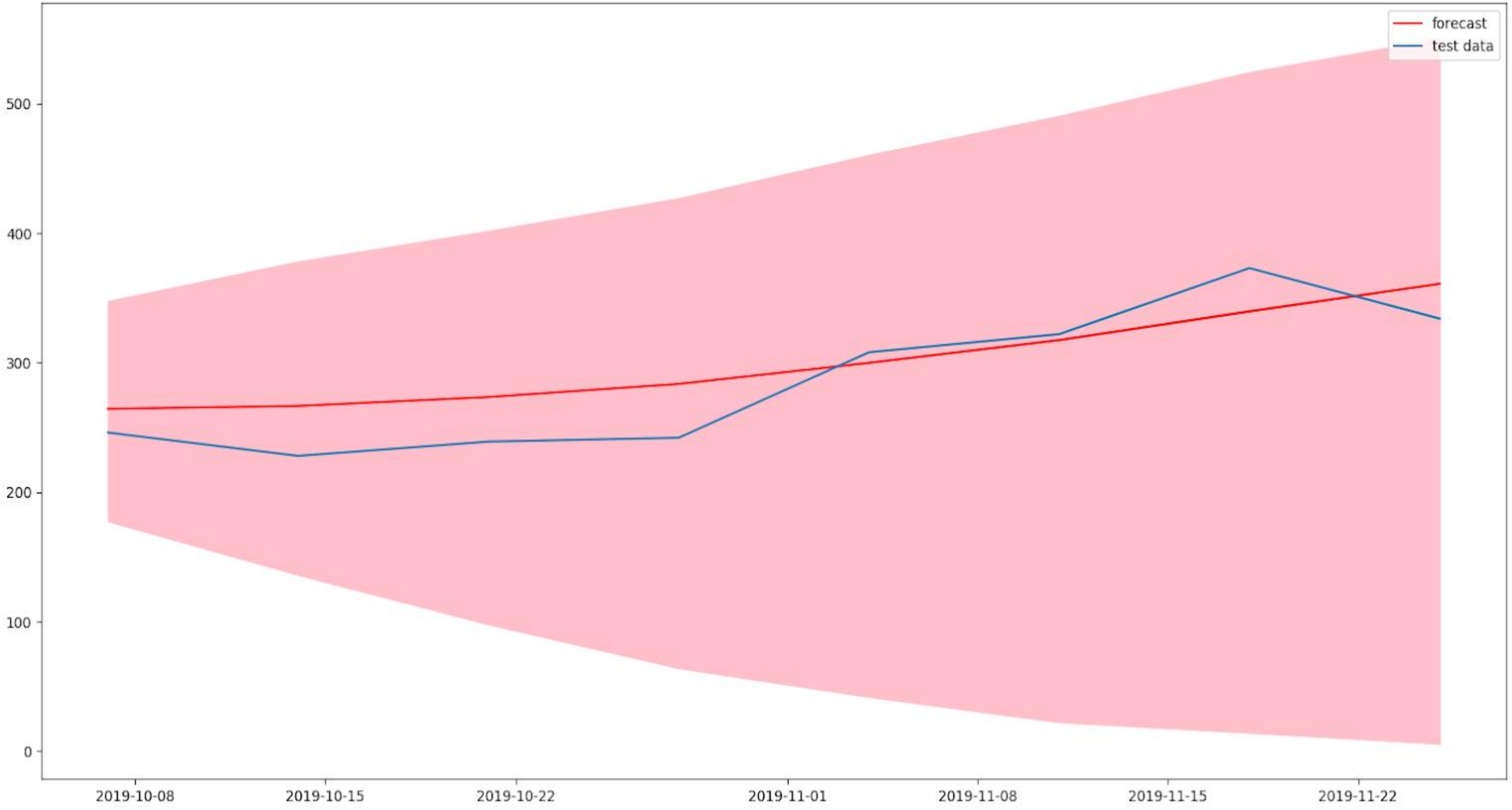
After Backward Elimination, 2 variables are present

	coef	std err	z	P>|z|	[0.025	0.975]
google_search_for_fever	0.7307	0.163	4.489	0.000	0.412	1.050
maximum_rainfall_in_a_day	-0.1499	0.095	-1.585	0.113	-0.335	0.035
ar.L1	0.1150	0.109	1.059	0.290	-0.098	0.328
ar.L2	-0.7621	0.098	-7.787	0.000	-0.954	-0.570
ma.L1	-0.1157	0.092	-1.256	0.209	-0.296	0.065
ma.L2	0.8569	0.082	10.427	0.000	0.696	1.018
ma.S.L25	0.0588	0.051	1.164	0.244	-0.040	0.158
sigma2	1843.8042	82.816	22.264	0.000	1681.487	2006.121

Phase II: Enhancing the model with Exogenous Variables

Test MSE: 833.6

Persistence Test MSE: 1169.125



Results & Limitations

Best model for prediction is a SARIMAX of order (2,1,2) (0,0,1,25),

X = { Fever Search Term, Max Rainfall Changi }

Results & Limitations

Best model for prediction is a SARIMAX of order (2,1,2) (0,0,1,25),
X = { Fever Search Term, Max Rainfall Changi }

Limitations

1. Unrealistic to assume that during prediction, we will have data regarding exogenous variables
2. Better to use lagged exogenous variables instead during training
3. Other transformations, such as log, were not explored
4. A longer time step would allow the model to predict further into the future
5. Instead of combining the 41 stations data into a single entity, each station could had been treated separately

Conclusions

Phase 1: Creating a model using only Past Observations

Phase II: Enhancing the model with Exogenous Variables

Over time, the relationship of other variables should be explored as the relationship is dependant with time

This model should be used along with other time-series prediction models, such as SVMs or ANNs

Thank you!

Q&A

References

- https://www.brainkart.com/media/extra3/86KMsSv.jpg
- https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=3753061_clep-5-299Fig1.jpg

<Template Slide>

- <Placeholder Points>