Rainfall Forecasting SARIMA vs ETS Smoothing Method. Which one is more accurate?

Final Project by Nur Annisa A

MIND MAP

Key Steps

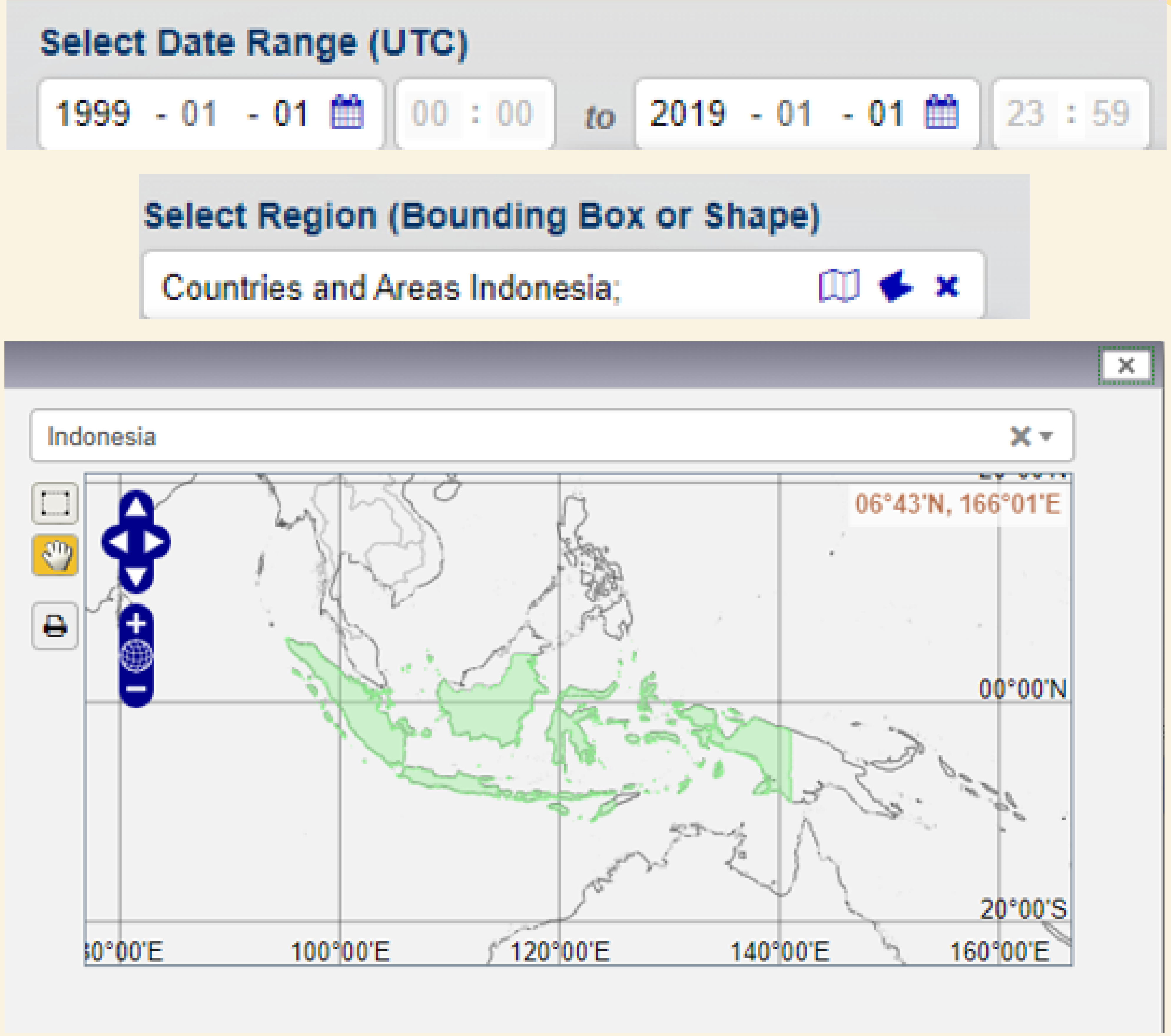
1. Define the Problem
2. Data Collection
3. Data Understanding
4. Data Cleaning
5. Exploratory Data Analysis
6. Modelling
7. Model Evaluation
8. Conclusion
9. Recommendation

Problem Definition

As a country that is quite dependent on agriculture, prediction of future rainfall will be very helpful in the pre-planning water resources. Unexpected amount of rainfall due to climate change can affect agricultural productivity. In this project, we will use ARIMA and ETS model for the rainfall forecasting analysis. The purpose of the analysis is to determine the range of the amount of rain that will occur and see the effects of climate change that is currently happening.

Data Collection

The data used in this analysis were obtained from Giovanni. Giovanni is a Web interface that allows users to analyze NASA's gridded data from various satellite and surface observations.



Data Understanding

Data obtained has 7313 rows and two columns. The head of data contains information about the downloaded data.

- First column : date time
- Second column : amount of rainfall in mm/day

	Title:	Time Series, Area-Averaged of Precipitation Rate daily 0.25 deg. [TRMM ()
0	User Start Date:	1999-01-01T00:00:00Z
1	User End Date:	2019-01-01T23:59:59Z
2	User Bounding Box:	NaN
3	Data Bounding Box:	NaN
4	URL to Reproduce Results:	https://giovanni.gsfc.nasa.gov/giovanni/#servi...
...	...	...
7308	2018-12-28	7.19602013
7309	2018-12-29	6.30557919
7310	2018-12-30	7.00499153
7311	2018-12-31	4.82450294
7312	2019-01-01	13.1867971

7313 rows × 2 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7313 entries, 0 to 7312
Data columns (total 2 columns):
 #   Column                                                                       Non-Null Count  Dtype
---  ------                                                                       --------------  -----
 0   Title:                                                                       7313 non-null   object
 1   Time Series, Area-Averaged of Precipitation Rate daily 0.25 deg. [TRMM ()   7311 non-null   object
dtypes: object(2)
memory usage: 114.4+ KB
```

Data Cleaning

The data used contains :
- Two missing value
- None duplicated value

```
## Check missing value
df.isna().sum()
```

```
Title:                                                                          0
Time Series, Area-Averaged of Precipitation Rate daily 0.25 deg. [TRMM ()       2
dtype: int64
```

```
## Check duplicated value
df.duplicated().sum()
```

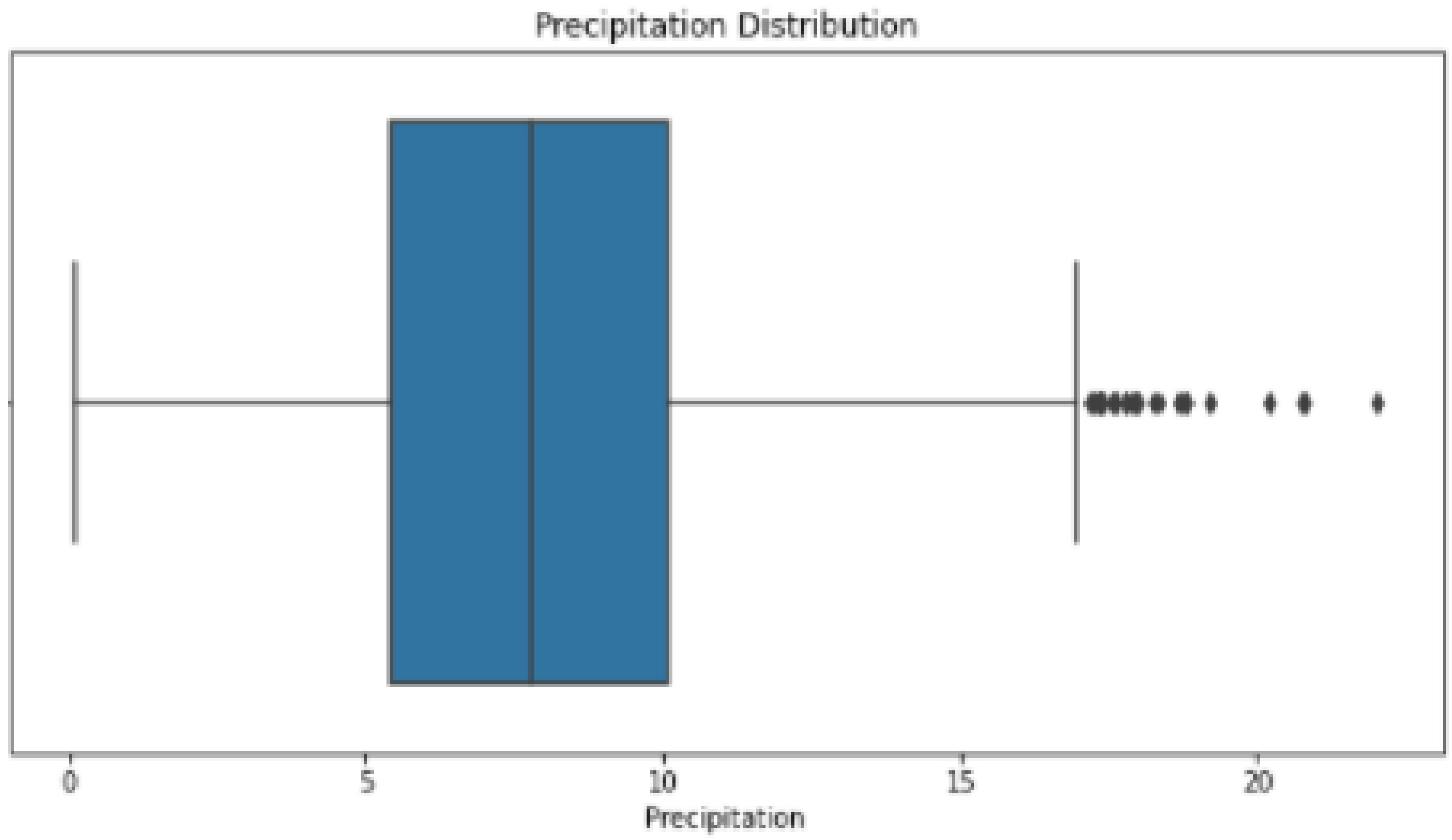
```
0
```

Date	Precipitation
1999-01-01	14.903526
1999-01-02	11.425750
1999-01-03	7.948120
1999-01-04	12.740754
1999-01-05	8.806804
...	...
2018-12-28	7.196020
2018-12-29	6.305579
2018-12-30	7.004992
2018-12-31	4.824503
2019-01-01	13.186797

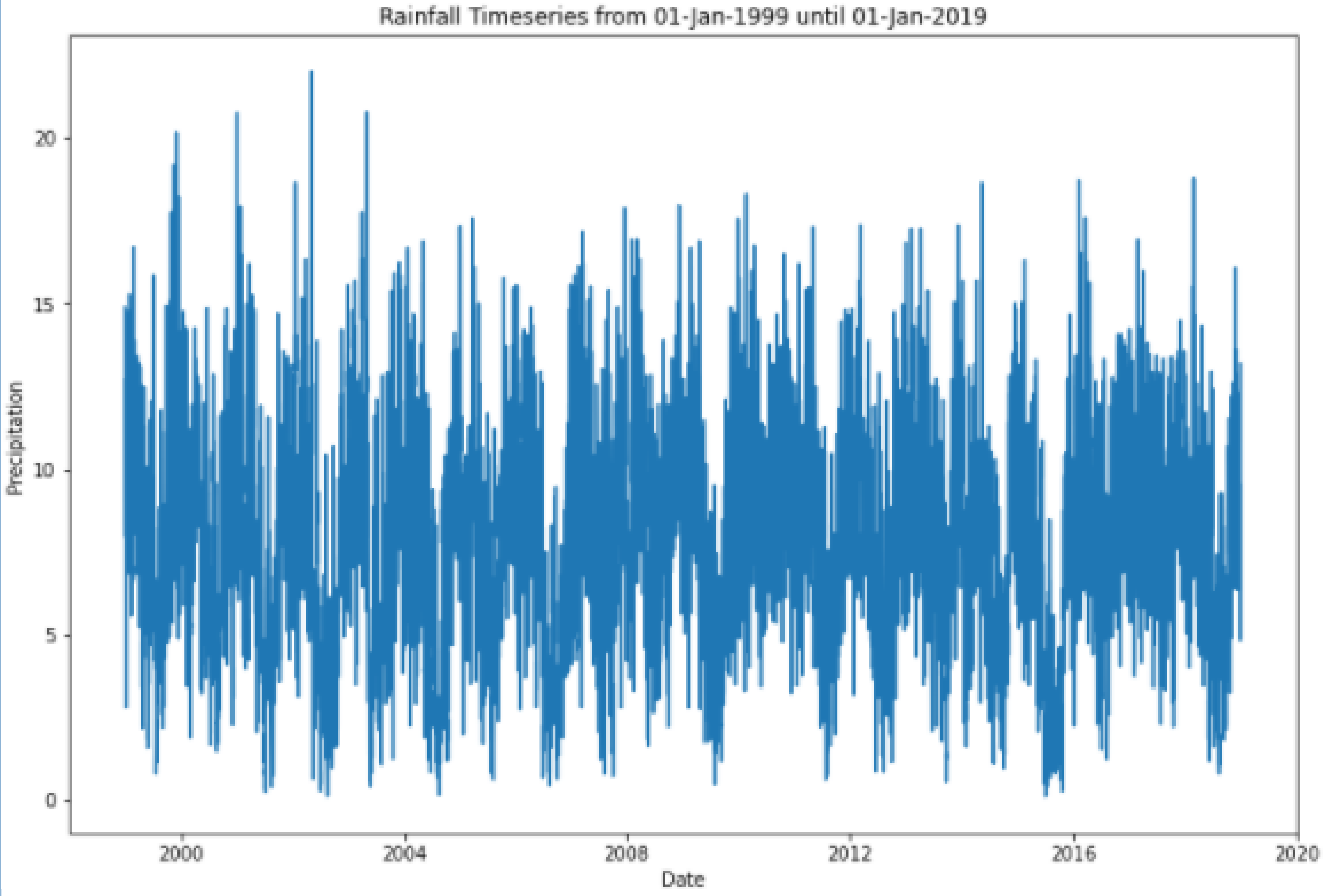
Date	Precipitation	month	year
1999-01	10.000398	1	1999
1999-02	9.691330	2	1999
1999-03	10.648398	3	1999
1999-04	8.398030	4	1999
1999-05	7.350341	5	1999
...	...	...	...
2018-09	5.048592	9	2018
2018-10	6.844830	10	2018
2018-11	9.226878	11	2018
2018-12	9.107919	12	2018
2019-01	13.186797	1	2019

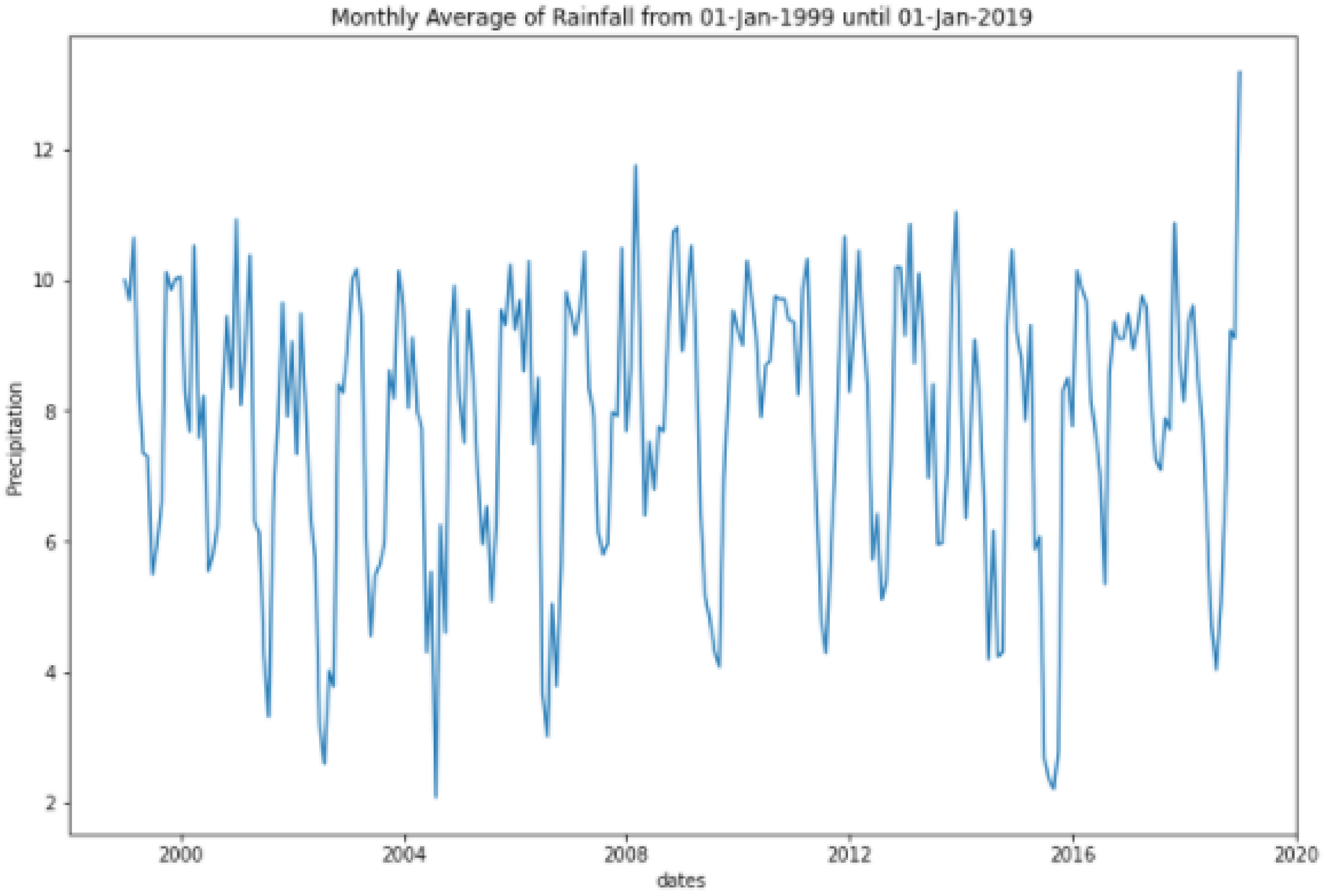
Exploratory Data Analysis (1)

The boxplot on the right shows the data contains outliers. But actually this value is not an outlier because the value still makes sense as the amount of rainfall.



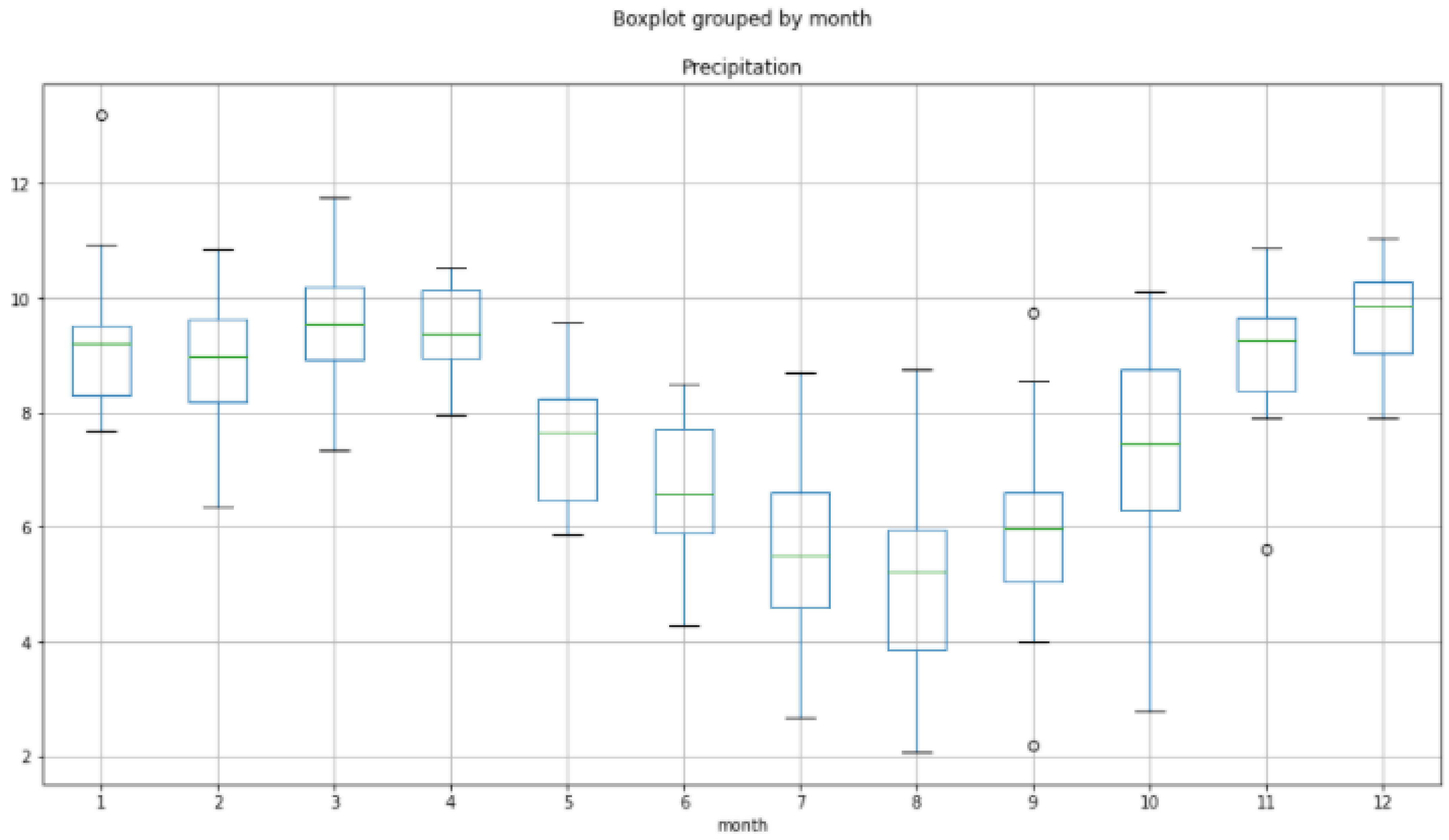
Exploratory Data Analysis (2)





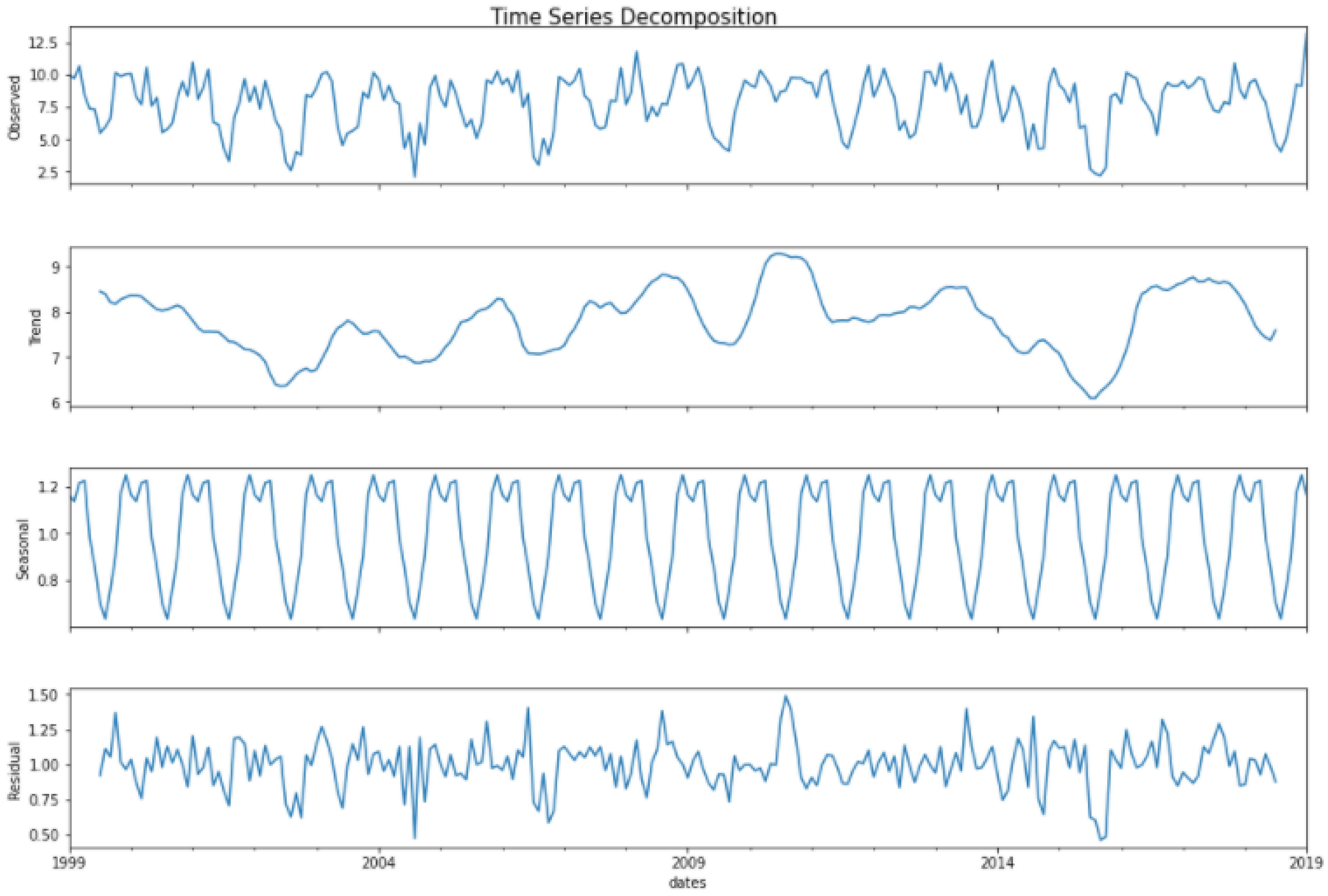
Exploratory Data Analysis (3)

Boxplot grouped by month to check the seasonal pattern in every month from 1999 until 2019



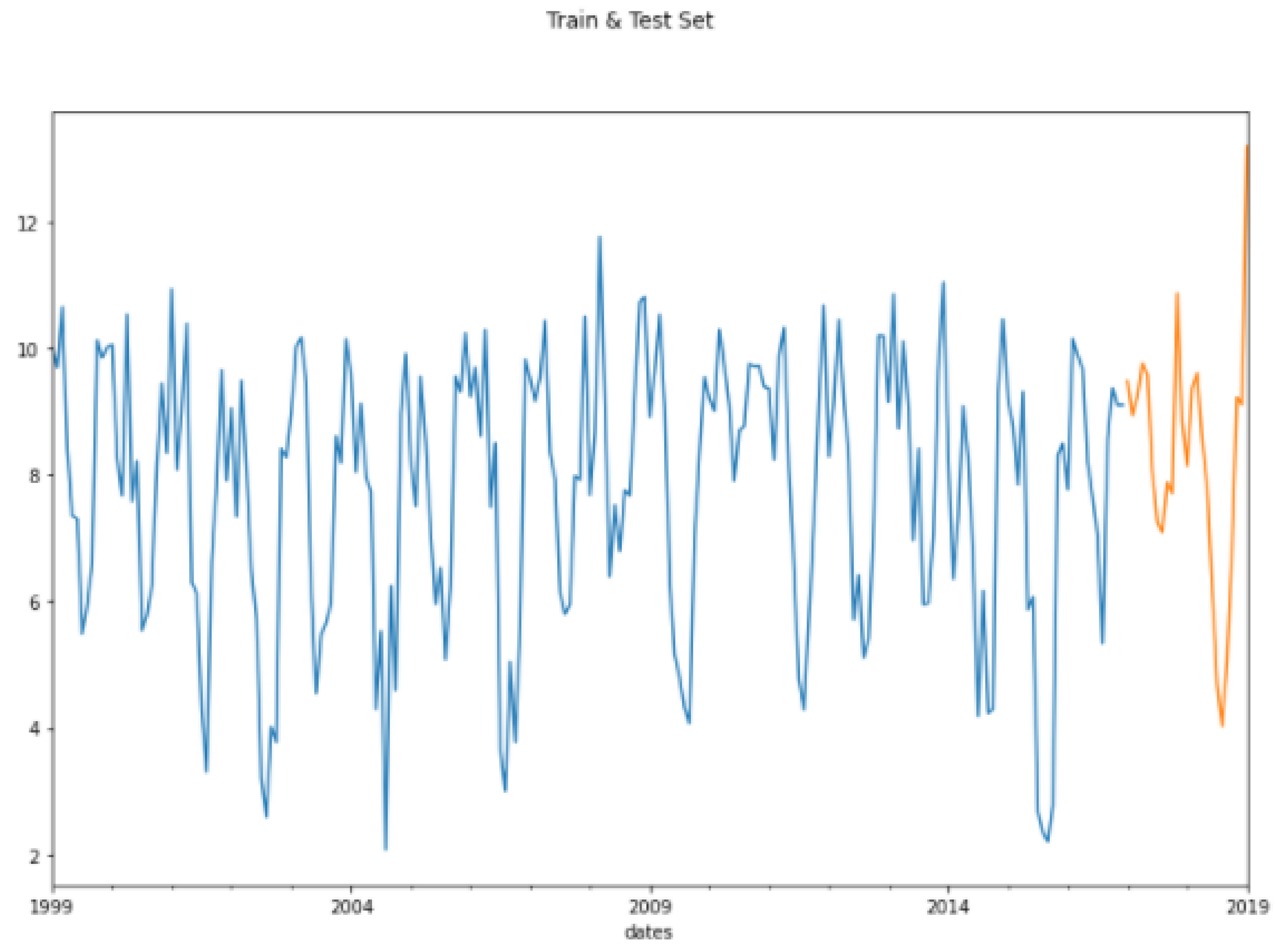
Exploratory Data Analysis (4)

Time Series Decomposition to check the trend and seasonality



Splitting Dataset into Train and Test set

Split into train and test set
Train set : data from 1999 - 2017
Test set : data from 2017 - 2019



Stationarity Check

Dickey-Fuller Hypothesis testing

- Null Hypothesis: The series is not stationary.
- Alternate Hypothesis: The series is stationary.

```
ADF Statistic: -3.698323
p-value: 0.004137
Critical Values:
        1%: -3.463
        5%: -2.876
        10%: -2.574
Result: The series is stationary
```

KPSS testing

- Null Hypothesis : The serie is stationary.
- Alternate Hypothesis : The serie is not stationary.

```
KPSS Statistic: 0.09393575165633575
p-value: 0.1
num lags: 15
Critial Values:
   10% : 0.347
   5% : 0.463
   2.5% : 0.574
   1% : 0.739
Result: The series is stationary
```

The data is stationary!

MODELLING

Seasonal ARIMA

Exponential Smoothing (ETS)

ARIMA

ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. An ARIMA model is characterized by 3 terms:

- p is the order of the AR term
- q is the order of the MA term
- d is the number of differencing required to make the time series stationary

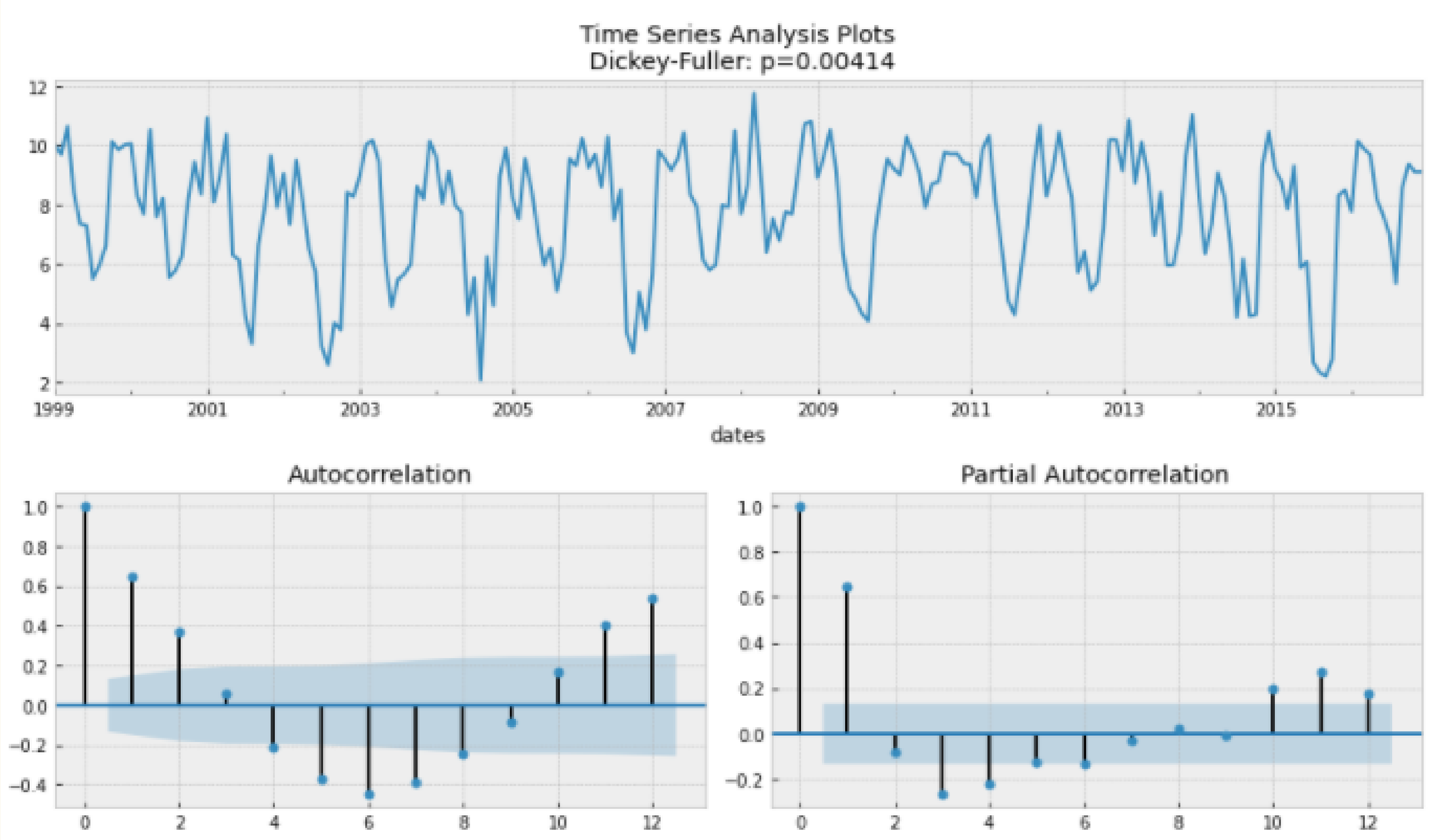
If a time series has seasonal patterns, we need to add seasonal terms and it becomes SARIMA, short for ‘Seasonal ARIMA’.

There are four seasonal elements that are not part of ARIMA that must be configured; they are:

- P: Seasonal autoregressive order.
- D: Seasonal difference order.
- Q: Seasonal moving average order.
- m: The number of time steps for a single seasonal period.

The Modelling Steps

1. Split data into train and test set
2. Find the best parameter to build ARIMA Model
3. Testing on the test set
4. Model Evaluation




```
ARIMA(0, 0, 0)x(0, 0, 0, 12)12 - AIC:1693.5322474199647
ARIMA(0, 0, 0)x(0, 0, 1, 12)12 - AIC:1474.8896216079195
ARIMA(0, 0, 0)x(1, 0, 0, 12)12 - AIC:1048.2690376173314
ARIMA(0, 0, 0)x(1, 0, 1, 12)12 - AIC:923.6884273852249
ARIMA(0, 0, 1)x(0, 0, 0, 12)12 - AIC:1442.1517048732567
ARIMA(0, 0, 1)x(0, 0, 1, 12)12 - AIC:1292.4313411697226
ARIMA(0, 0, 1)x(1, 0, 0, 12)12 - AIC:997.5529625906856
ARIMA(0, 0, 1)x(1, 0, 1, 12)12 - AIC:888.5052980271437
ARIMA(1, 0, 0)x(0, 0, 0, 12)12 - AIC:957.4854282115832
ARIMA(1, 0, 0)x(0, 0, 1, 12)12 - AIC:934.5751923332359
ARIMA(1, 0, 0)x(1, 0, 0, 12)12 - AIC:921.0015107705485
ARIMA(1, 0, 0)x(1, 0, 1, 12)12 - AIC:859.6570712121486
ARIMA(1, 0, 1)x(0, 0, 0, 12)12 - AIC:957.9038053930856
ARIMA(1, 0, 1)x(0, 0, 1, 12)12 - AIC:929.513737092298
ARIMA(1, 0, 1)x(1, 0, 0, 12)12 - AIC:905.936099088098
ARIMA(1, 0, 1)x(1, 0, 1, 12)12 - AIC:838.5062577583068
```

Time Series Analysis Plots

ACF and PACF plots help us to determine the p and q values.

PACF & ACF suggested that range for p value = (0,2) and q = (0,2). P and Q are the same as p and q.

Value of d will be 0 since data already stationary.

By using 'for looping', we can find the best value of each parameter.

The results show that the best parameter values for seasonal ARIMA are : ARIMA(1,0,1)x(1,0,1)12 - with AIC = 838.5062577583068.

Fitting Model

Result Summary Table

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9095	0.037	24.716	0.000	0.837	0.982
ma.L1	-0.4789	0.082	-5.811	0.000	-0.640	-0.317
ar.S.L12	0.9990	0.002	434.177	0.000	0.995	1.004
ma.S.L12	-0.9338	0.077	-12.166	0.000	-1.084	-0.783
sigma2	1.5711	0.183	8.592	0.000	1.213	1.929

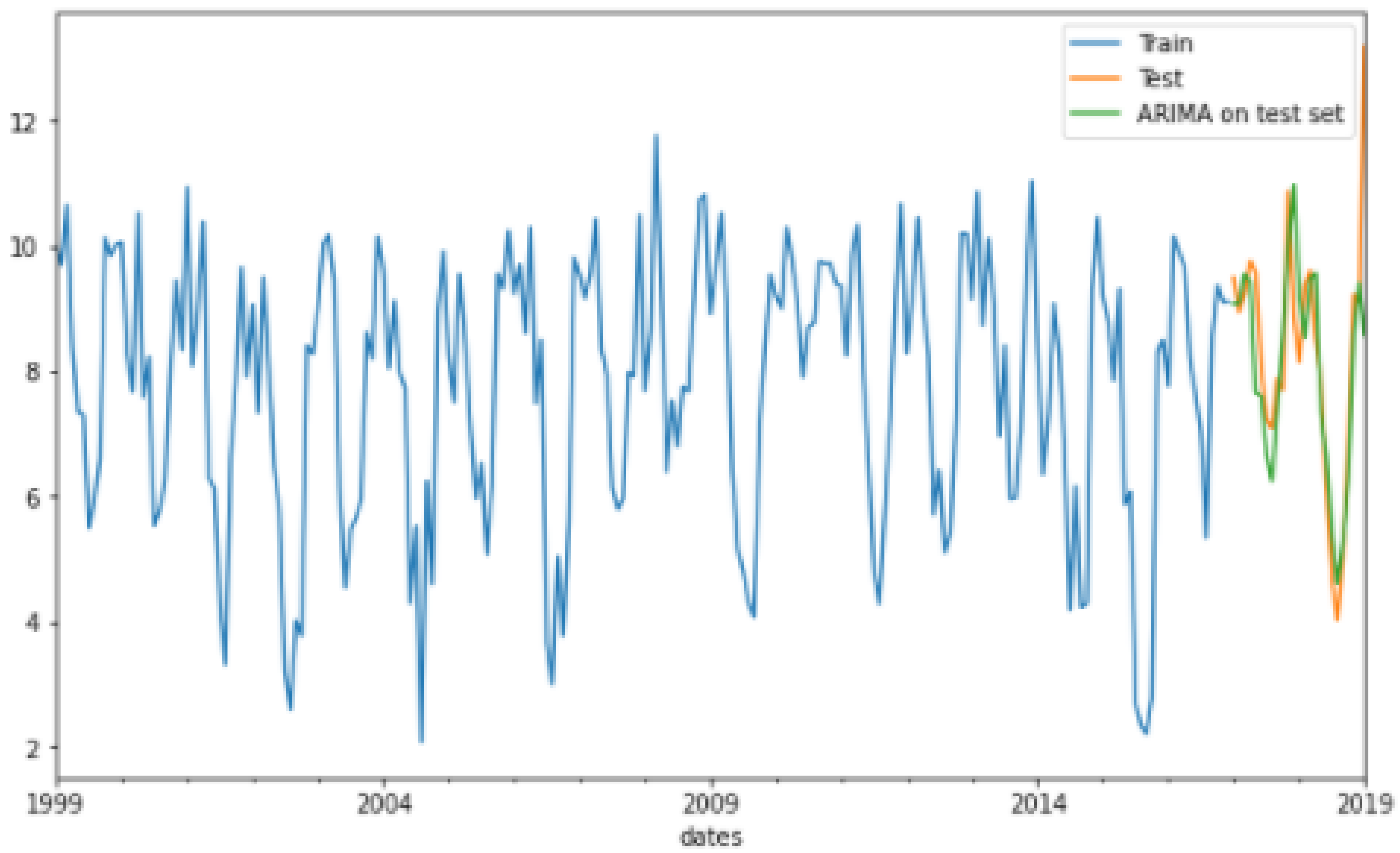
The P>|z| column informs us of the significance of each feature weight. Here, each weight has a p-value lower than 0.05, so it is reasonable to retain all of them in the model.

Model Evaluation : Testing on Test Set

The evaluation metrics used to evaluate the performance of this model :

- MAE
- MAPE
- MSE
- RMSE
- R-square

```
Results of sklearn.metrics for SARIMA Model:
MAE: 0.8429969767982257
MAPE: 0.1321218364390769
MSE: 1.5209238853067064
RMSE: 1.2332574286444442
R-Squared: 0.5926934118384852
```



Exponential Smoothing ETS

Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. There is three type of exponential smoothing methods :

- Simple (single) exponential smoothing uses a weighted moving average with exponentially decreasing weights.
- Holt’s trend-corrected double exponential smoothing is usually more reliable for handling data that shows trends, compared to the single procedure.
- Triple exponential smoothing (also called the Multiplicative Holt-Winters) is usually more reliable for parabolic trends or data that shows trends and seasonality.

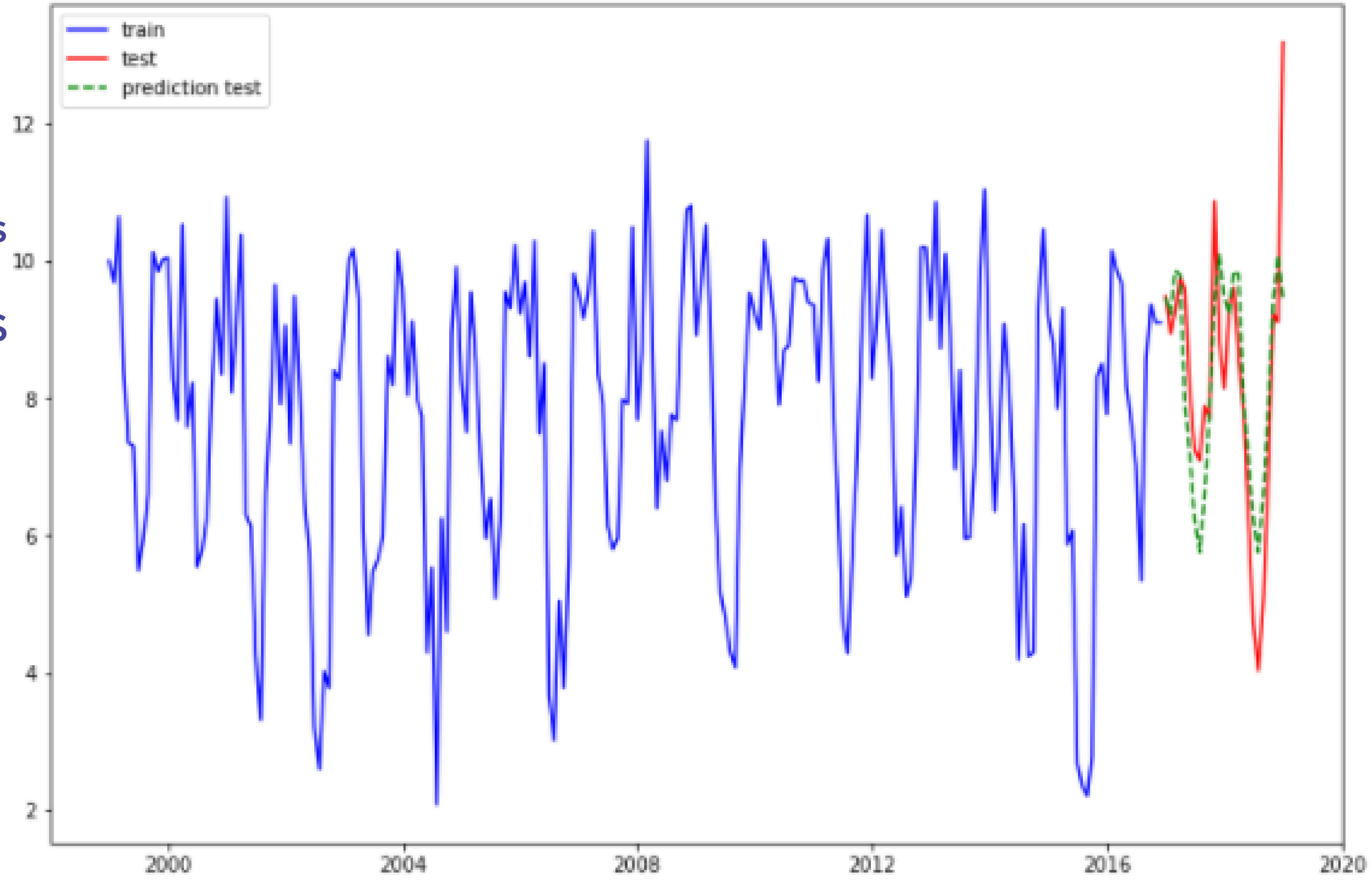
The Modelling Steps

1. Split data into train and test set
2. Define method that will be used
3. Testing on the test set
4. Model Evaluation

Exponential Smoothing ETS

The data used has a seasonality and does not have a significant trend.
The ETS method used in the model is ETS (A, N, A).

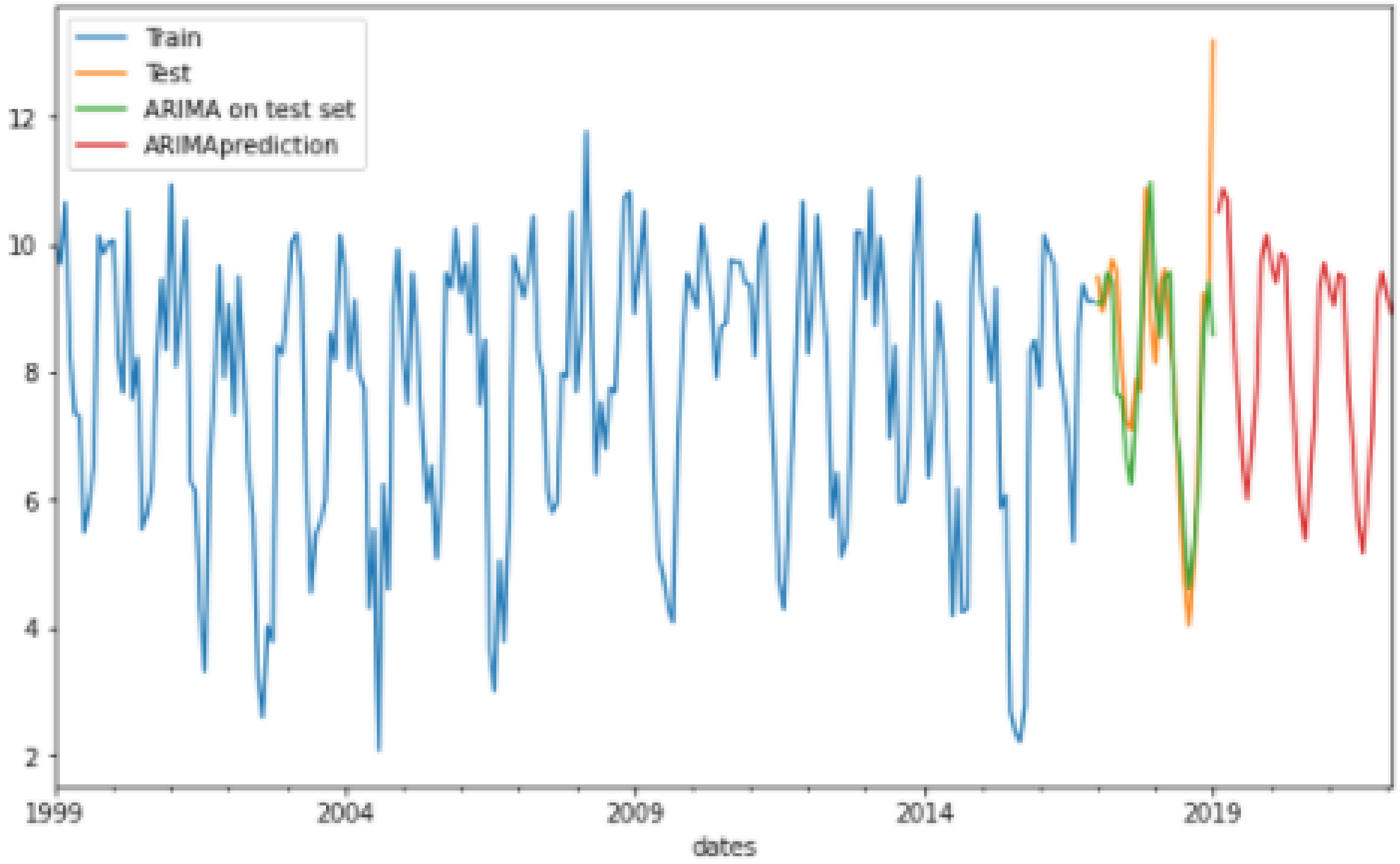
- Error : Additive
- Trend : None
- Seasonality : Additive



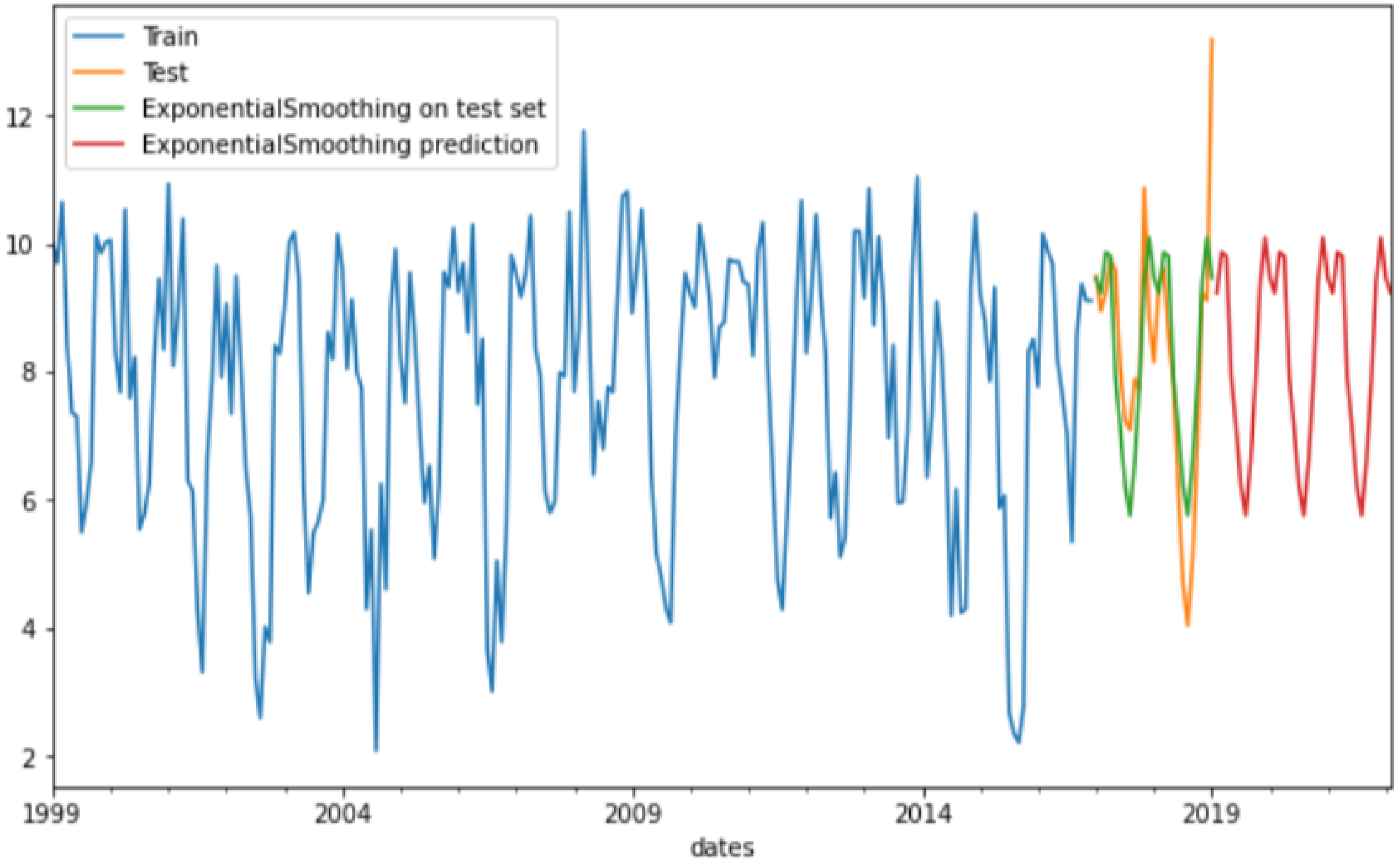
Model Testing on Test Set

```
Results of sklearn.metrics for ETS Model:
MAE: 0.9877797632279066
MAPE: 0.1321218364390769
MSE: 1.604351571158726
RMSE: 1.2666300056286073
R-Squared: 0.5703513035904159
```

Model Prediction & Model Selection

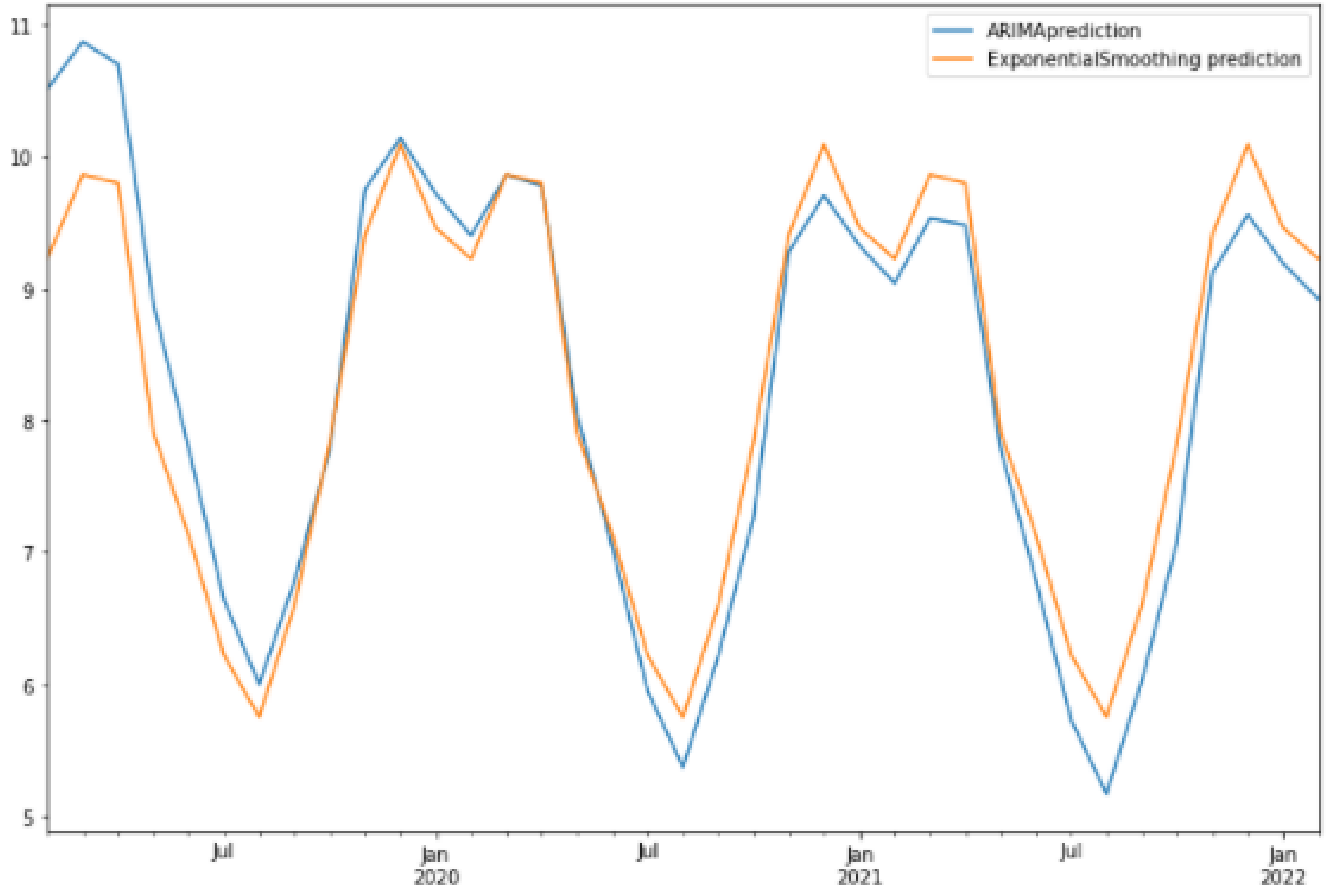


Results of sklearn.metrics for SARIMA Model:
MAE: 0.8429969767982257
MAPE: 0.1321218364390769
MSE: 1.5209238853067064
RMSE: 1.2332574286444442
R-Squared: 0.5926934118384852



Results of sklearn.metrics for ETS Model:
MAE: 0.9877797632279066
MAPE: 0.1321218364390769
MSE: 1.604351571158726
RMSE: 1.2666300056286073
R-Squared: 0.5703513035904159

Model Prediction Visualization



Conclusion

- Seen from the metric evaluation and visualization of the model results, SARIMA model has a better performance in forecasting more than exponential smoothing.
- The model results do not show any seasonal shifts in the next few years and an increase or decrease in trends so that the impact of climate change cannot be seen from the results of the model analysis.
- Since the results of the analysis do not show any seasonal shifts, rainfall in the future would not have an impact on agricultural cropping patterns.
- Two models used in the analysis may provide a range of rain that is likely to occur, but it is still very far from accurate and basically rainfall can be unpredictable due to climate change.

Recommendation

- The more data used in the train set, the more accurate the model can perform.
- Try more combinations of parameters in order to improve the model.