4/13/2022

Time Series – Sunspots

Olivier FONTAINE

TABLE OF CONTENTS

ANALYSIS OBJECTIVE

➔ *Main objective of the analysis that also specifies whether your model will be focused on a specific type of Time Series, Survival Analysis, or Deep Learning and the benefits that your analysis brings to the business or stakeholders of this data.*

The objective of this analysis is to use a set of time series models to predict accurately sunspots numbers in the following years.

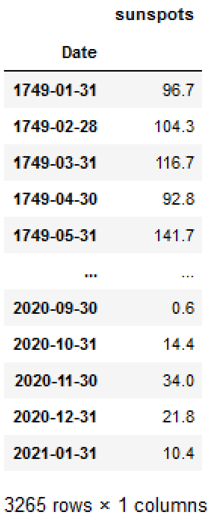
DATA SET DESCRIPTION

➔ *Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.*

The data set name is **Sunspots** Dataset.

The data set is available on KAGGLE: https://www.kaggle.com/datasets/robervalt/sunspots

The data set contain **3,265** Monthly Mean Total Sunspot Number, from 31/01/1749 to 31/01/2021.



	sunspots
Date	
1749-01-31	96.7
1749-02-28	104.3
1749-03-31	116.7
1749-04-30	92.8
1749-05-31	141.7
...	...
2020-09-30	0.6
2020-10-31	14.4
2020-11-30	34.0
2020-12-31	21.8
2021-01-31	10.4

3265 rows × 1 columns

So, in this project, we will try different time series model to predict accurately sunspots numbers in the following years.

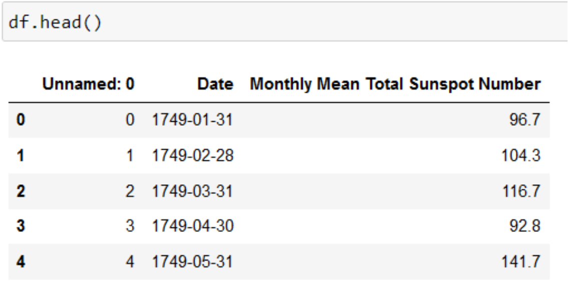
DATA EXPLORATION, DATA CLEANING & FEATURE ENGINEERING

➔ *Brief summary of data exploration and actions taken for data cleaning and feature engineering.*

PREPROCESSING

The dataset needs to be converted in a time series format.

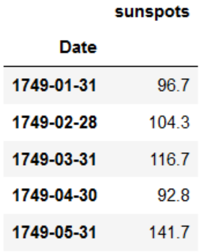
Dataset first values after dataset creation:



```
df.head()
```

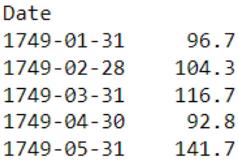
	Unnamed: 0	Date	Monthly Mean Total Sunspot Number
0	0	1749-01-31	96.7
1	1	1749-02-28	104.3
2	2	1749-03-31	116.7
3	3	1749-04-30	92.8
4	4	1749-05-31	141.7

Dataset transformed with date in index:



	sunspots
Date	
1749-01-31	96.7
1749-02-28	104.3
1749-03-31	116.7
1749-04-30	92.8
1749-05-31	141.7

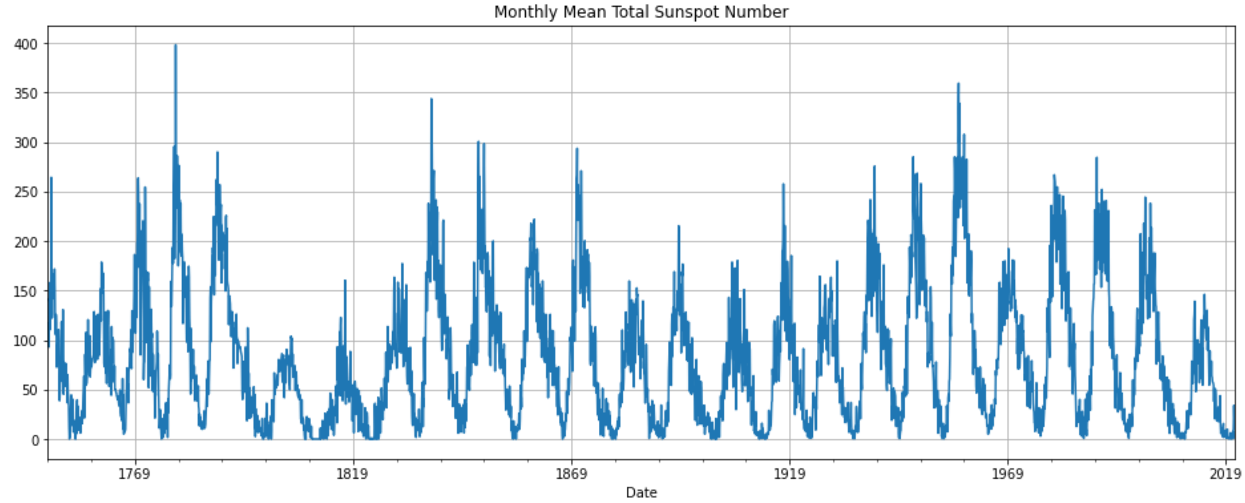
And converted in an array:



```
Date
1749-01-31     96.7
1749-02-28    104.3
1749-03-31    116.7
1749-04-30     92.8
1749-05-31    141.7
```

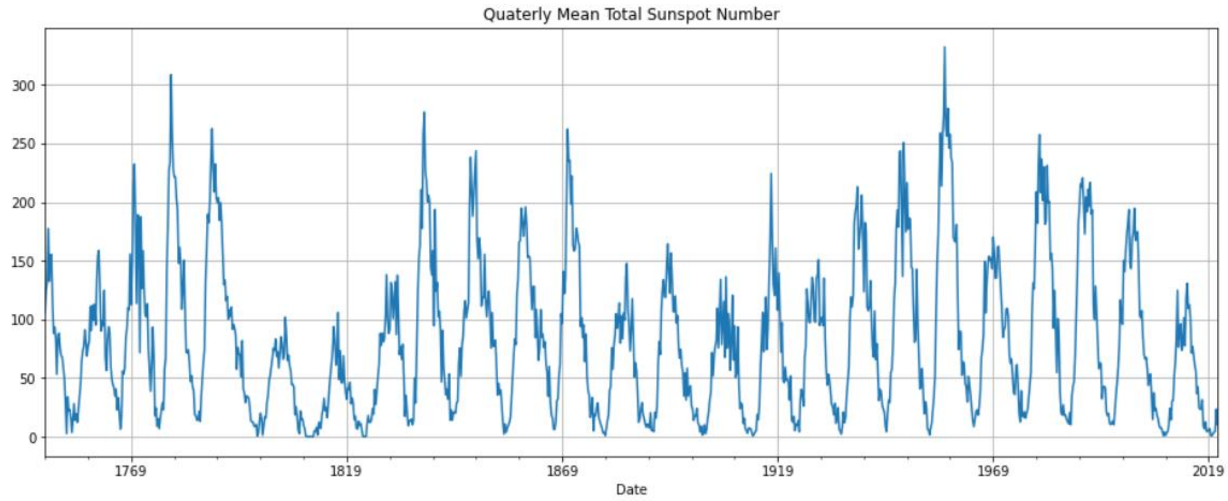
DATA EXPLORATION

Below, you can see the evolution of monthly mean total sunspot number:



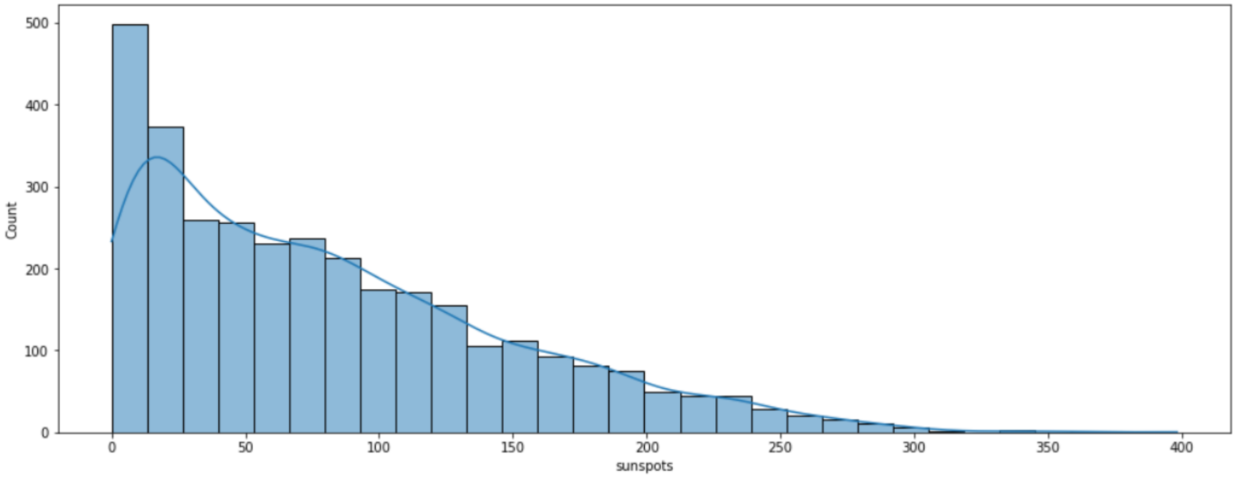
Cycles appears clearly (seasonality). And we can also see that there are cycles within cycles: there are higher peaks at certain times.

Here is now the same data but resampled on quarter period (3 months) instead of month (1 month):



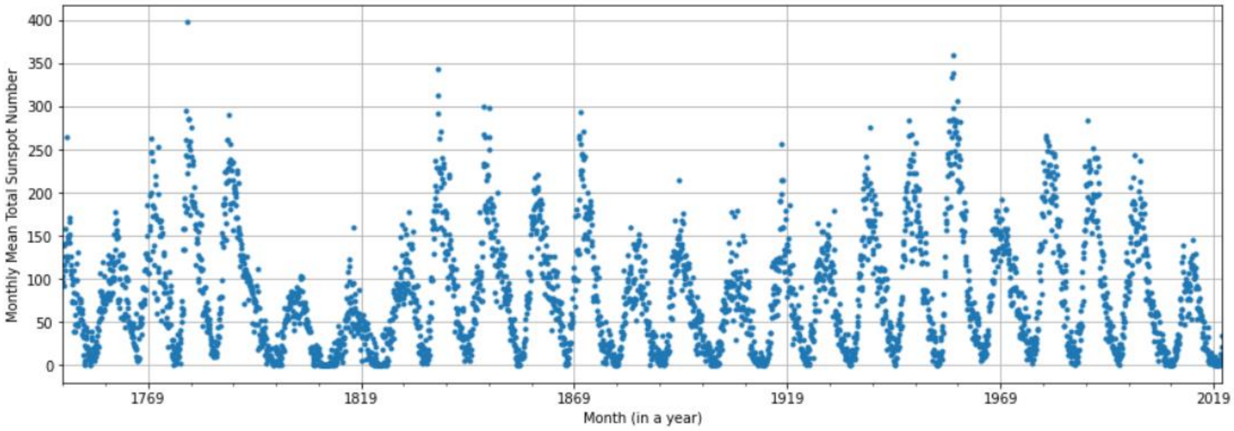
Note: this quarter dataset is easier to use with SARIMA model (less time consuming).

Here is now the histogram plot for the month dataset:

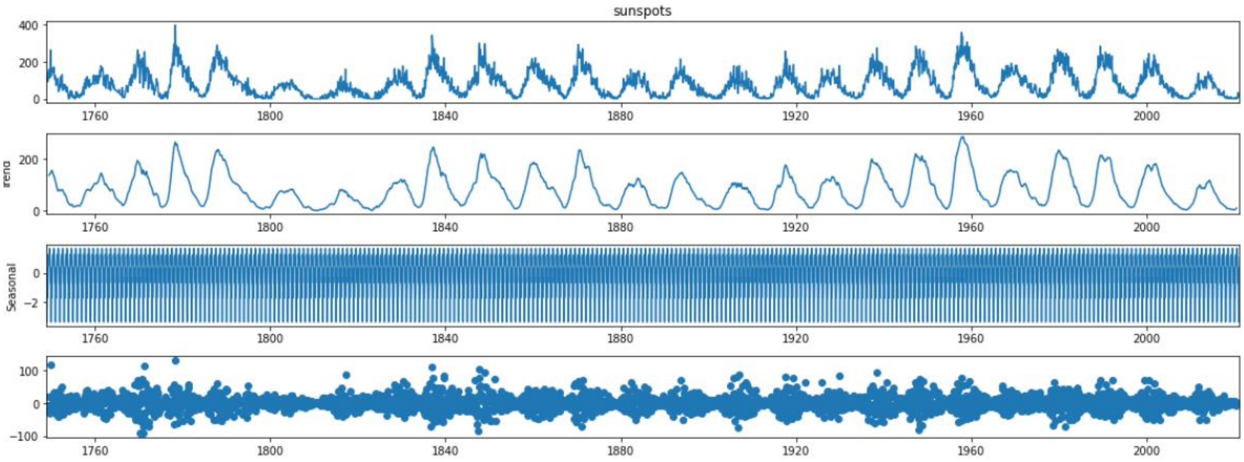


The distribution is right skewed: there are few peaks.

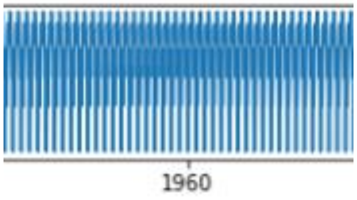
This can also be seen with the following chart where values are dots:



Here is now the decomposition of the time serie:



We can notice in the seasonal plot that we do have the famous 11 years cycles but we also notice 6 dense areas. I.e; the one centered on year 1960.



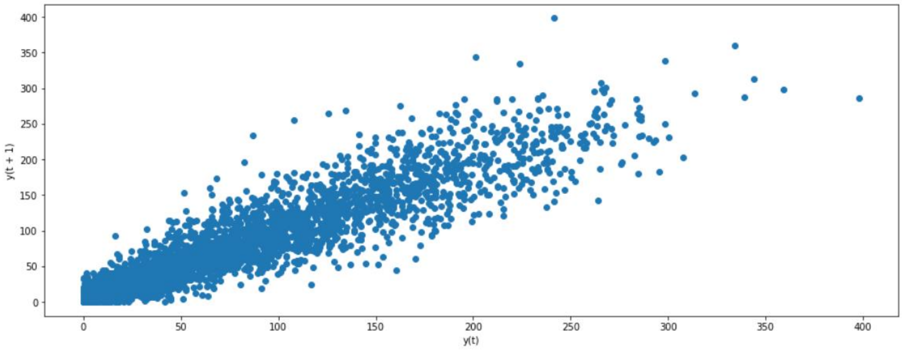
AFD TEST

In order to check the stationarity of the time series, we performed the ADF Test (Augmented Dicker Fuller Test).

- ➔ The P-value is very low (<<0.05) so it means that the Time series is stationary

AUTOCORELLATION & PARTIAL AUTOCORELLATION

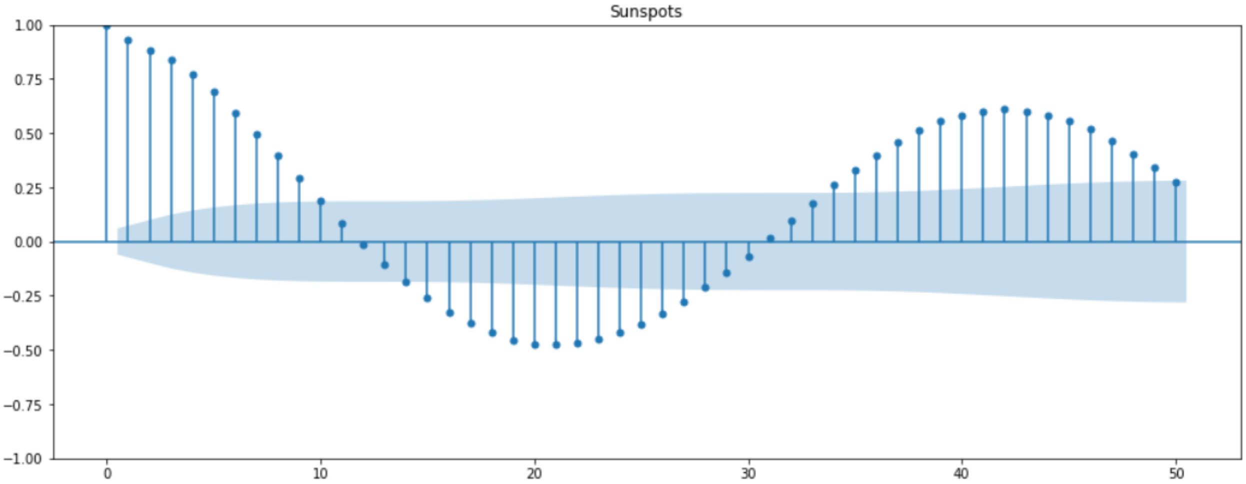
We use lag plot to see the correlation of values between 2 months:



There is a positive correlation between a given month and the corresponding previous month.

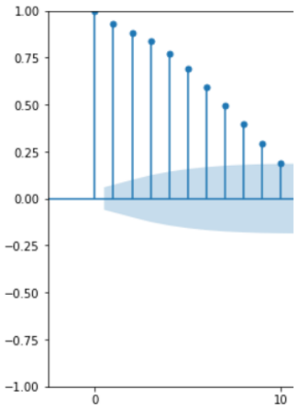
Note: as of now, we will only use the quarter time series.

Here is the autocorrelation plot (ACF) with 50 lags:



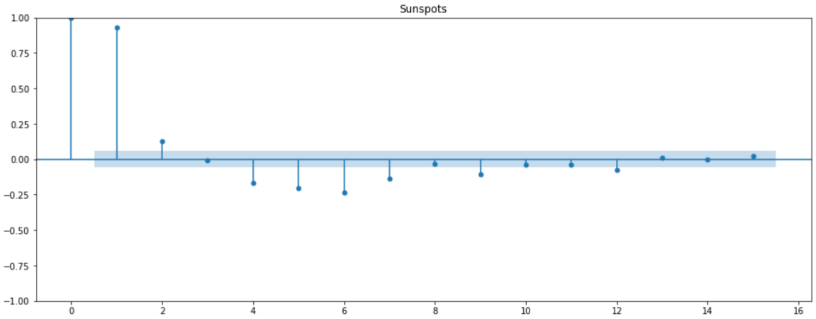
This Autocorrelation plot is commonly used to detect dependence on prior observations.

➔ The last 10 observations (10 quarters = 2 years and a half) have correlations that are statistically significant.



➔ Aaa

Here is the Partial Autocorrelation plot (ACF) with 16 lags:



The partial autocorrelation plot summarizes dependence on past observations. It measures partial results, so it takes into account other lags and remove the effects of other lags and allow you to look at the correlations independently.

➔ Here, when we include all past values, only the 3 most recent values appear significant.

TIME SERIES MODELS

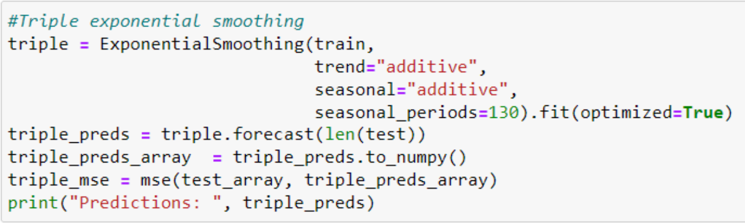
➔ *Summary of training at least three variations of the Time Series, Survival Analysis, or Deep Learning model you selected. For example, you can use different models or different hyperparameters.*

3 time series are evaluated:

- Exponential smoothing
- SARIMA
- LSTM

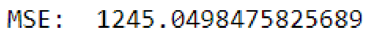
EXPONENTIAL SMOOTHING

We use a smooth model (triple exponential smoothing)



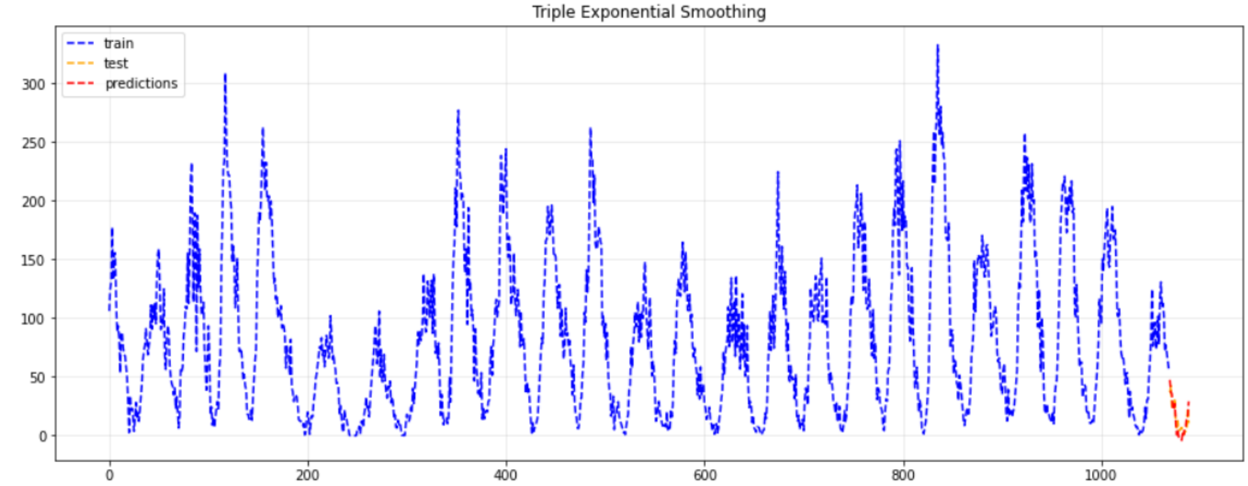
```
#Triple exponential smoothing
triple = ExponentialSmoothing(train,
                              trend="additive",
                              seasonal="additive",
                              seasonal_periods=130).fit(optimized=True)
triple_preds = triple.forecast(len(test))
triple_preds_array  = triple_preds.to_numpy()
triple_mse = mse(test_array, triple_preds_array)
print("Predictions: ", triple_preds)
```

The Mean Squared Error (MSE) is equal to 1245:



```
MSE:  1245.0498475825689
```

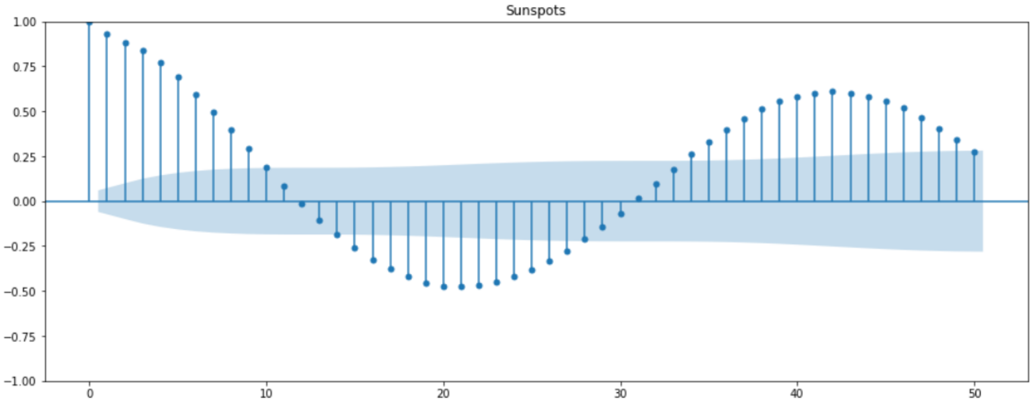
Let's plot the prediction:



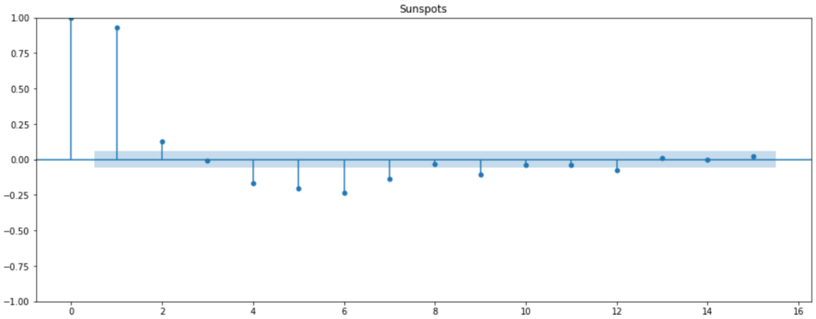
SARIMA

The 2nd model is SARIMA: Seasonal ARIMA.

Based on the ACF plot:



And the PACF plot:



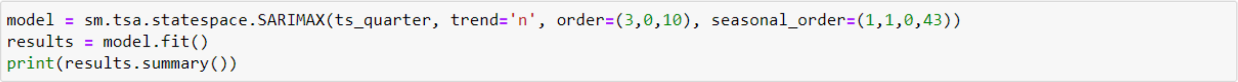
We can evaluate the parameter of the SARIMA model (p, d, q)(P, D, Q):

- p: PACF plot : p = first lag where the value is above the significance level. p=3
- d: The test of stationarity (ADF Test) is significative with no differencing, so d=0
- q: ACF plot : q = first lag where the value is above the significance level. q=10

Seasonal:

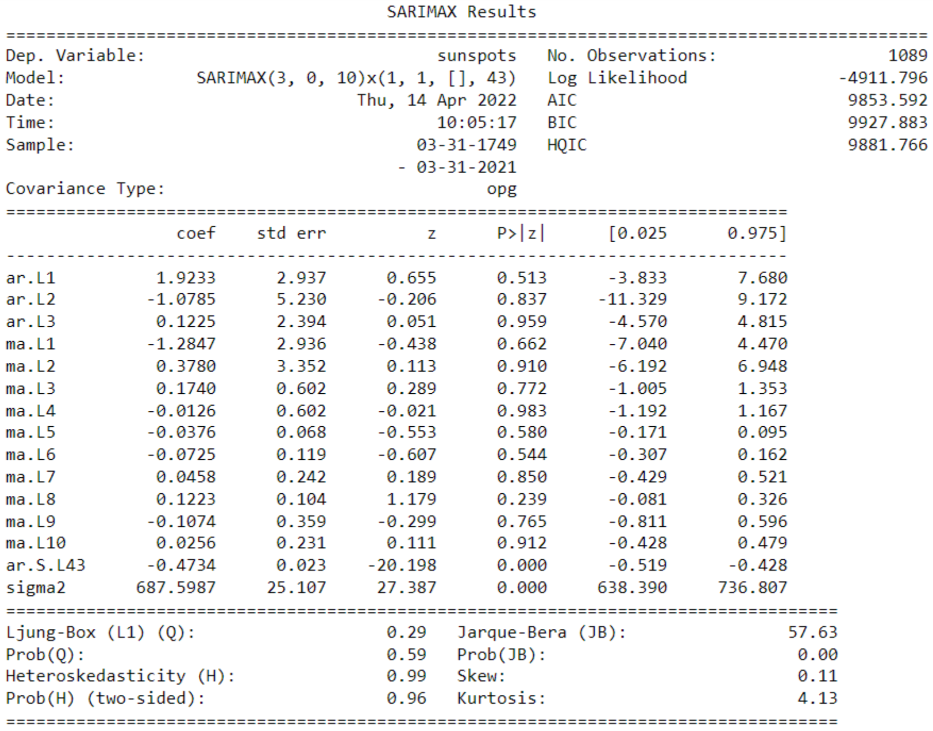
- P: ACF plot : P=1, ACF is positive at lag 43 AND P+Q≤2
- D: D=1, the series has a stable seasonal pattern over time.
- Q: ACF plot : Q=0 ACF is negative at lag 43 AND P+Q≤2

So, we run SARIMA model based on these parameters:



```
model = sm.tsa.statespace.SARIMAX(ts_quarter, trend='n', order=(3,0,10), seasonal_order=(1,1,0,43))
results = model.fit()
print(results.summary())
```

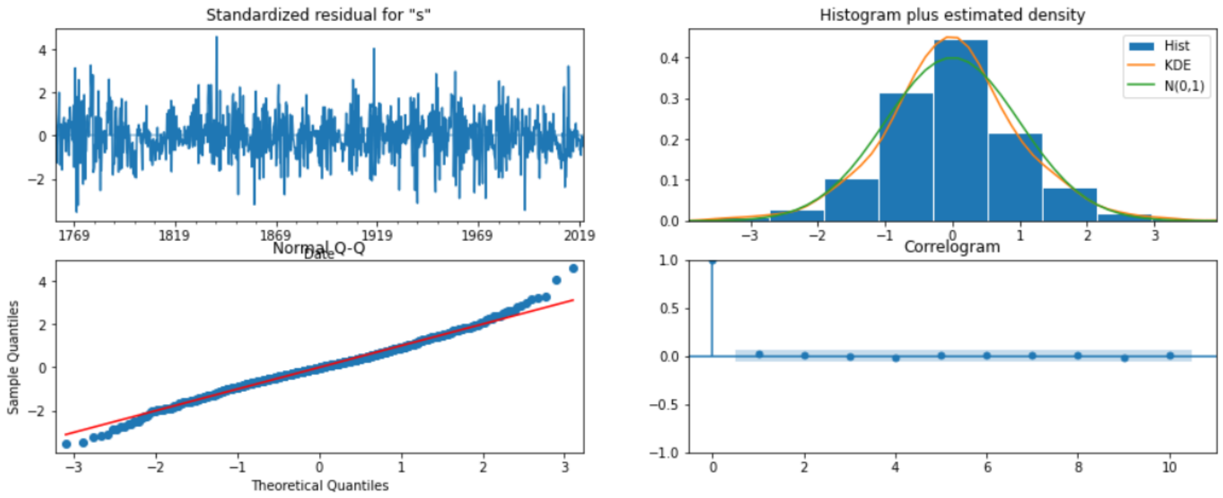
The summary gives the following:



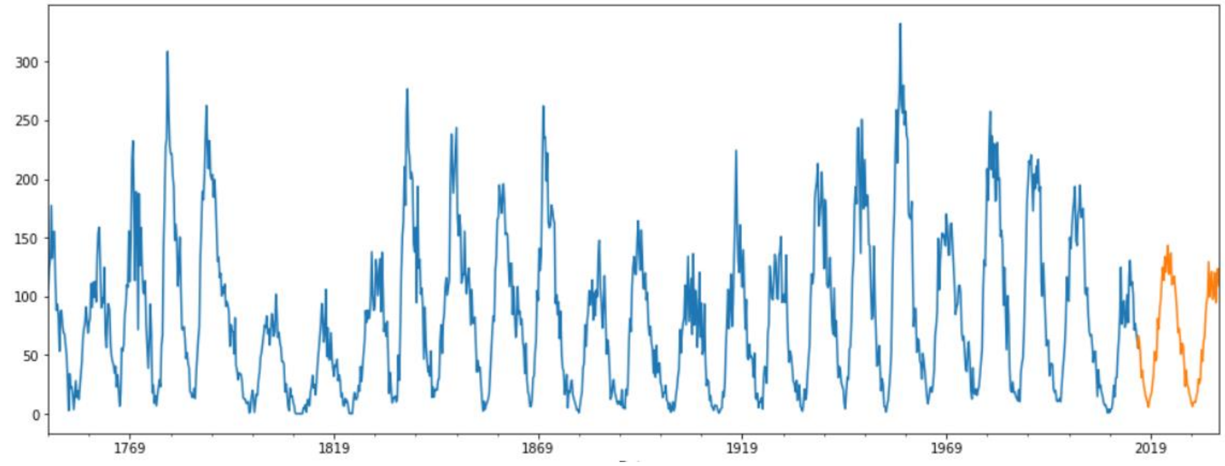
```
                                      SARIMAX Results
==============================================================================================
Dep. Variable:                                sunspots   No. Observations:                1089
Model:             SARIMAX(3, 0, 10)x(1, 1, [], 43)   Log Likelihood              -4911.796
Date:                             Thu, 14 Apr 2022   AIC                          9853.592
Time:                                     10:05:17   BIC                          9927.883
Sample:                                 03-31-1749   HQIC                         9881.766
                                      - 03-31-2021
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.9233      2.937      0.655      0.513      -3.833       7.680
ar.L2         -1.0785      5.230     -0.206      0.837     -11.329       9.172
ar.L3          0.1225      2.394      0.051      0.959      -4.570       4.815
ma.L1         -1.2847      2.936     -0.438      0.662      -7.040       4.470
ma.L2          0.3780      3.352      0.113      0.910      -6.192       6.948
ma.L3          0.1740      0.602      0.289      0.772      -1.005       1.353
ma.L4         -0.0126      0.602     -0.021      0.983      -1.192       1.167
ma.L5         -0.0376      0.068     -0.553      0.580      -0.171       0.095
ma.L6         -0.0725      0.119     -0.607      0.544      -0.307       0.162
ma.L7          0.0458      0.242      0.189      0.850      -0.429       0.521
ma.L8          0.1223      0.104      1.179      0.239      -0.081       0.326
ma.L9         -0.1074      0.359     -0.299      0.765      -0.811       0.596
ma.L10         0.0256      0.231      0.111      0.912      -0.428       0.479
ar.S.L43      -0.4734      0.023    -20.198      0.000      -0.519      -0.428
sigma2       687.5987     25.107     27.387      0.000     638.390     736.807
===================================================================================
Ljung-Box (L1) (Q):                   0.29   Jarque-Bera (JB):                57.63
Prob(Q):                              0.59   Prob(JB):                         0.00
Heteroskedasticity (H):               0.99   Skew:                             0.11
Prob(H) (two-sided):                  0.96   Kurtosis:                         4.13
===================================================================================
```

AIC = 9,853

The following diagnostic gives information about the residuals. They are normally distributed.



Finally, we plot the prediction:



➔ The results looks good.

LSTM

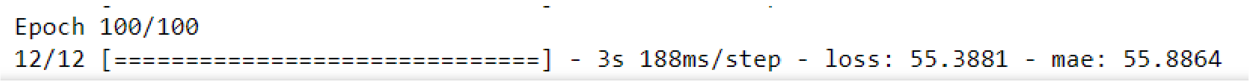
The 3rd model is LSTM (Deep Learning). This is mainly inspired from one of the project on KAGGLE.

The model is the following:

Layers
Conv1D layer

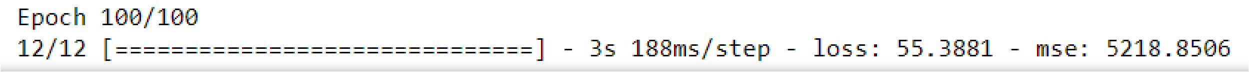
LSTM
Batch normalization
LSTM
Batch normalization
Dense
Batch normalization
Dense
Dense
Lambda

- ➔ After 100 epochs, MAE is equal to around 56.



```
Epoch 100/100
12/12 [==============================] - 3s 188ms/step - loss: 55.3881 - mae: 55.8864
```

- ➔ After 100 epochs, MSE is equal to around 5218.



```
Epoch 100/100
12/12 [==============================] - 3s 188ms/step - loss: 55.3881 - mse: 5218.8506
```

TIME SERIES MODEL RECOMMENDATION

- ➔ *A paragraph explaining which of your models you recommend as a final model that best fits your needs in terms of accuracy or explainability.*

The smooth model was very basic and on the small period. He will not evaluated here.

The LSTM got a better score on MSE than the SARIMA model so LSTM is the best model.

KEY FINDINGS AND INSIGHTS

- ➔ *Summary Key Findings and Insights, which walks your reader through the main findings of your modeling exercise.*

The key finding is that we confirm the 11 years cycle of the sunspots. We are able to predict future periods based on our last model.

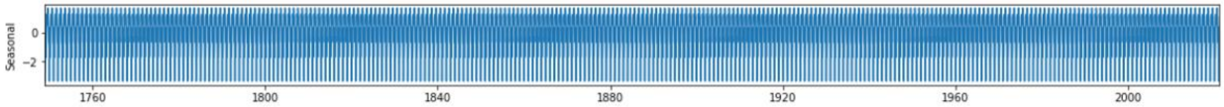
SUGGESTIONS FOR NEXT STEPS IN ANALYZING THIS DATA

- ➔ *Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model or adding specific data features to achieve a better model.*

On the SARIMA model, I could have used gridsearch to find the best hyperparameters of this model.

On the LSTM model, I could have fine-tuned the model by selecting different values of those neural network hyper parameters.

It could be interesting to dive into the other cycle type discovered in data exploration:



APPENDIX: CODE

The code for this project is available on GITHUB:

https://github.com/Olivier-FONTAINE/IBM-Machine-Learning-professional-certificate/blob/main/06-TSS-Sunspot%20analysis.ipynb