Final Project 29 – Final Report

Team Members – Abhik Choudhury, Ganapathy Raaman Balaji, Vamsi Perumallu Yalamanchili

Summary:

As we all know, one of the most challenging problems these days is to understand and maintain the web traffic and also keep ourselves ready for the future. As part of this project our team collected data for approximately 145,000 Wikipedia webpages. The goal of this project is to create models that can accurately forecast the future web traffic. Since Python does not have enough packages like R for time series forecasting, to solve the problem using Python is a challenge that our team has decided to solve.

Introduction:

For the last few years, more people are getting access to the internet all around the world. Those companies that manages the increase in website traffic most efficiently would be the ones to succeed.[1] There have been much research done to forecast web traffic. A few of them include Linear models such as Holt Winters Model, AR Model, and MA Model. Nonlinear models are focused on forecasting with Recurring Neural Networks. ARMA and ARIMA models are advantageous in univariate time series forecasting.

ARMA (Auto Regressive Moving Average) process combines the Autoregressive and moving average processes. It states that the present value is linearly dependent on its own previous values and a constant (like the AR process), and on the mean of the time series, the current error term and previous error terms (like the MA process).[2] This is represented by ARMA(p, q) where p is the order of autoregressive process, and q is the order of Moving Average process. The ARMA process equation is given by:

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

ARMA(0, q) is the equivalent to MA(q), and ARMA(p,0) is equivalent to AR(p) process.

ARIMA describes trend and seasonality in time series as a function of lagged values.[2] In both ARIMA model involves differencing (also called Integrating) the original time series. This helps remove seasonality, trends, and variance in the data. The components of ARIMA is Auto-Regressive (AR) and Moving Average (MA), but in terms of differenced time series data. ARIMA process is denoted by ARIMA(p,d,q), where p is the order of the AR(p) process and q is the order of MA(q) process. d is the order of integration, which is the number of times the series has been differenced to make the series stationary. In simple terms, ARIMA model is ARMA model that is applied on non-stationary time series. While the series has to be rendered stationary before performing ARMA model, ARIMA can be applied on non-stationary time series (given the value of d).[2] The ARIMA equation is given by:[5]

$$y'_t = C + \varphi_1 y'_{t-1} + \cdots + \varphi_p y'_{t-p} + \theta_1 \epsilon'_{t-1} + \cdots + \theta_q \epsilon'_{t-q} + \epsilon_t$$

Final Project 29 – Final Report

Team Members – Abhik Choudhury, Ganapathy Raaman Balaji, Vamsi Perumallu Yalamanchili

Dataset Description:

The dataset was downloaded from Kaggle. [4] Fig 1 shows the first 5 rows of the dataset. It has a total of 145063 pages data, and the number of pages containing at least one missing value were 27786. Time period is between 01-Jul-2015 to 31-Dec-2016

	Page	2015-07-01	2015-07-02	2015-07-03	2015-07-04	2015-07-05	2015-07-06	2015-07-07	2015-07-08	2015-07-09	...	2016-12-22
0	2NE1_zh.wikipedia.org_all-access_spider	18.0	11.0	5.0	13.0	14.0	9.0	9.0	22.0	26.0	...	32.0
1	2PM_zh.wikipedia.org_all-access_spider	11.0	14.0	15.0	18.0	11.0	13.0	22.0	11.0	10.0	...	17.0
2	3C_zh.wikipedia.org_all-access_spider	1.0	0.0	1.0	1.0	0.0	4.0	0.0	3.0	4.0	...	3.0
3	4minute_zh.wikipedia.org_all-access_spider	35.0	13.0	10.0	94.0	4.0	26.0	14.0	9.0	11.0	...	32.0
4	52_Hz_I_Love_You_zh.wikipedia.org_all-access_s...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	48.0

Fig 1: First 5 rows of the dataset.

Analysis:

1. Data Cleaning and Exploration:

After loading the dataset, we find that there are 145063 total number of records, and 27786 records having more than 1 Null values. Since the original dataset needs to be pivoted, we have used the melt() function to pivot the dataframe from wide to long. We then converted date column to a pandas datetime object, and extracted Year, Month and Day from the dataframe (Fig 2).

	Page	Date	Visits	Year	Month	Day
0	2NE1_zh.wikipedia.org_all-access_spider	2015-07-01	18.0	2015	7	1
1	2PM_zh.wikipedia.org_all-access_spider	2015-07-01	11.0	2015	7	1
2	3C_zh.wikipedia.org_all-access_spider	2015-07-01	1.0	2015	7	1
3	4minute_zh.wikipedia.org_all-access_spider	2015-07-01	35.0	2015	7	1
4	52_Hz_I_Love_You_zh.wikipedia.org_all-access_s...	2015-07-01	NaN	2015	7	1

Fig 2: Pivoted dataframe

Looking at the number of Null records for the pivoted dataframe, all columns except ‘Visits’ columns have no null records. 7.7% of total records are Nulls. So, we dropped those rows containing Null values. We plotted the mean of page visits across all the pages to understand the trend and seasonality (Fig 3). We observed that there is a slight trend and cyclicity.

Final Project 29 – Final Report

Team Members – Abhik Choudhury, Ganapathy Raaman Balaji, Vamsi Perumallu Yalamanchili

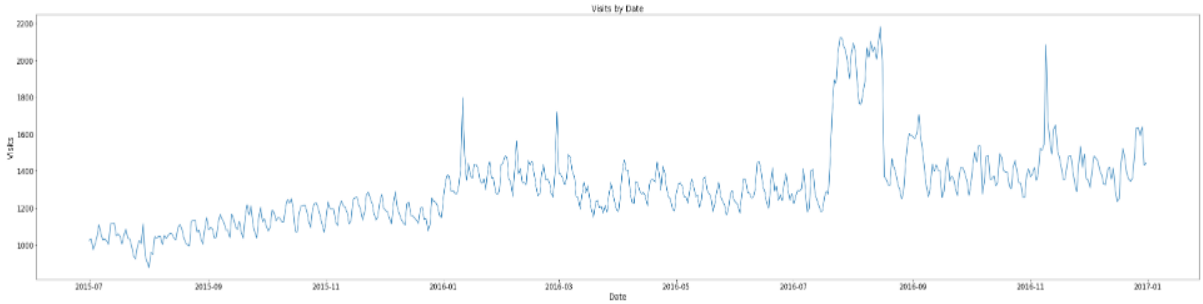

Fig 3: Daily mean of page visits for all pages

Next, Fig 4 shows the top 10 visited pages in the dataset across all languages visited by users.

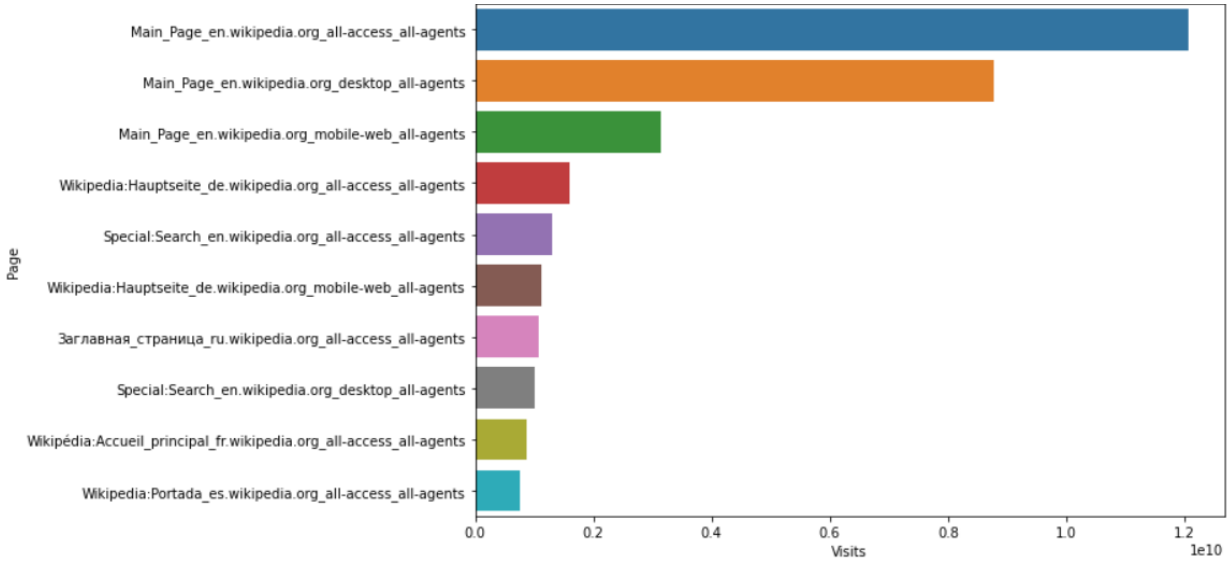

Fig 4: Top 10 visited pages across all languages

Fig 5 shows Language distribution of all the pages and Fig 3 shows the top 10 pages visited by users.

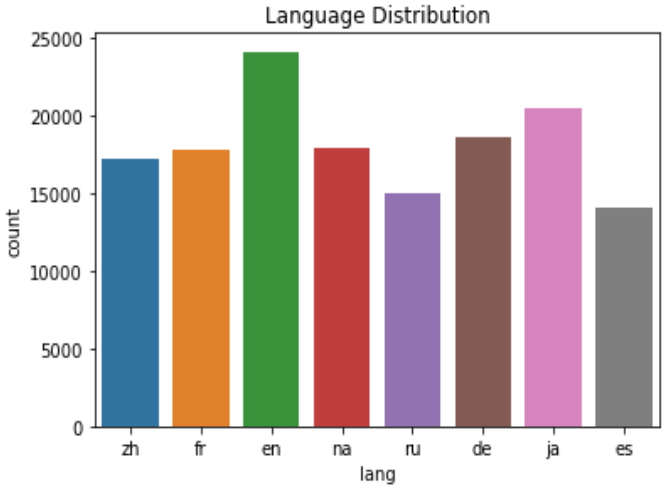

Fig 5: Languages distribution

Final Project 29 – Final Report

Team Members – Abhik Choudhury, Ganapathy Raaman Balaji, Vamsi Perumallu Yalamanchili

Fig 6 shows the daily mean of page visits segmented by language. We observed an increasing trend and cyclicity across all the languages. The order of languages in the plots in Fig 6 are (from top to down and then across) is as follows: 'ja', 'na', 'fr', 'en', 'es', 'de', 'zh', and 'ru'.

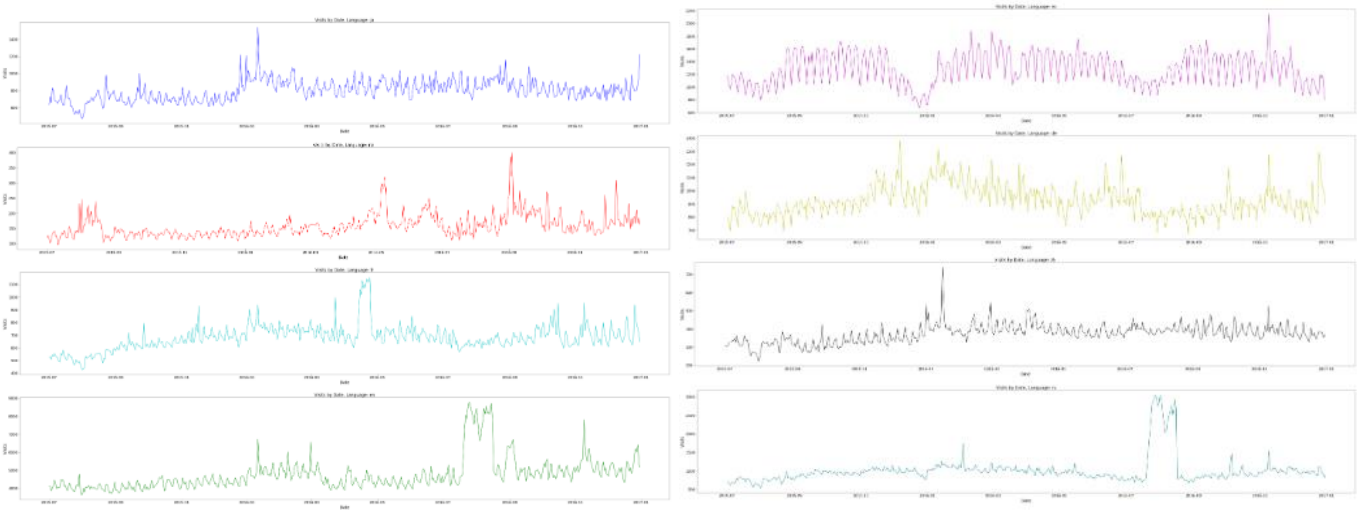

Fig 6: Mean page visits for pages published (all languages)

Since the pivoted dataframe (with only the Spanish webpages) has over 7 million rows, we decided to choose one webpage for our initial analysis. Fig 7 shows the mean page visits for pages published in Spanish.

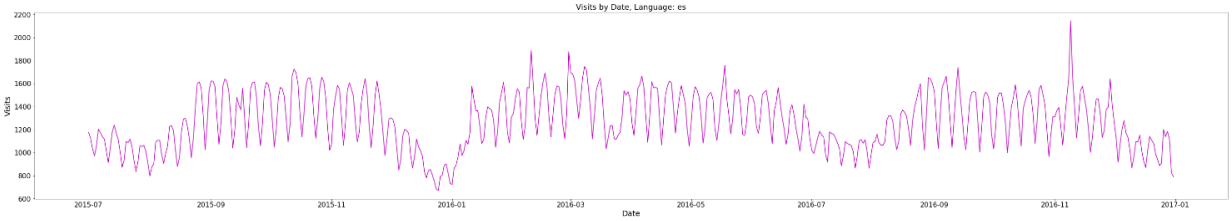

Fig 7: Mean page visits for pages published in Spanish

2. <u>Time Series Modeling:</u>

We performed an Augmented Dickey Fuller test (ADF test) for one of the Spanish webpages ('100_metros_es.wikipedia.org_desktop_all-agents'). The ADF statistic is -5.586, and the p-value of 1.36e-06. This suggests that the series is stationary. We can now take the 1st order difference of the series and perform ADF Test on the differenced data to check for stationarity. We plotted the ACF and PACF plots as shown below (Fig 8).

Final Project 29 – Final Report

Team Members – Abhik Choudhury, Ganapathy Raaman Balaji, Vamsi Perumallu Yalamanchili

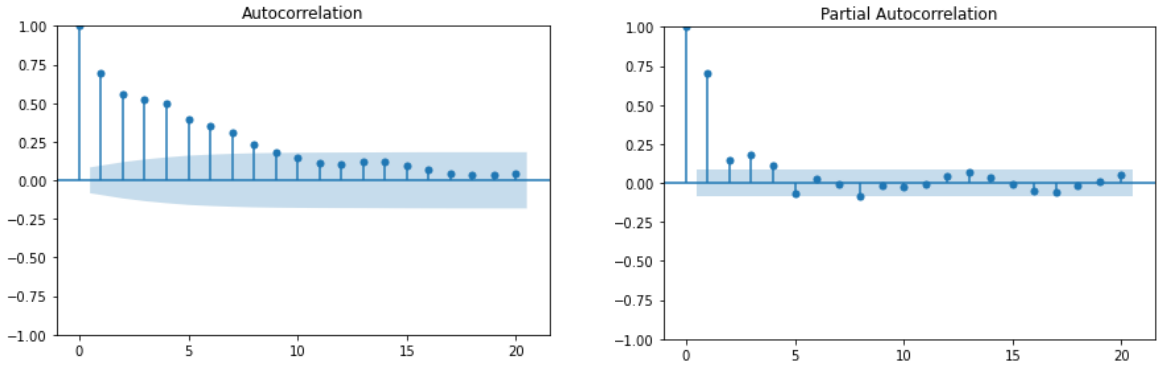

Fig 8: ACF Plot (On the left) of the actual number of visits for the Spanish webpages. Notice how the plot is slowly decaying. This indicates an autoregressive process. From the PACF plot, we can see that after lag 4, the partial autocorrelation coefficients are not significantly different from 0.

Next, the ADF Statistic of the differenced time series is -9.45 and the p-value is 4.67e-16. This shows the data is stationary. Fig 9 shows the ACF and PACF plots of the differenced time series with a max lag of 20.

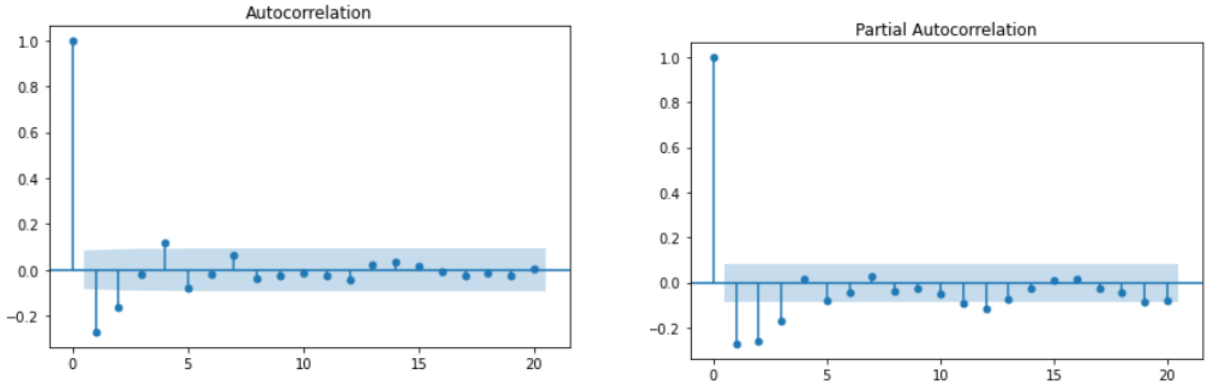

Fig 9: ACF Plot (On the left) of the differenced number of visits for the Spanish webpages. We can see from both the plots that this is an ARMA process with max p = 3, and max q = 2.

From the ACF and PACF plots, it is clear that the process is an ARMA (3, 2) process. Using SARIMAX function in Python, we created an ARMA model, iterated over p and q values. The best model is evaluated using Akaike Information Criterion (AIC). AIC is an estimation of the model quality relative to other models. Since some information will be lost when a model is fitted to the data, AIC quantifies relative amount of information lost by the model. Lower AIC means better model. AIC is a function of the number of parameters k in a model and the maximum value of the likelihood function ($\hat{L}$). AIC is given by:

$$AIC = 2k - 2ln(\hat{L})$$

For the different combinations of p and q values, Fig 10 shows the values of AIC.

Final Project 29 – Final Report

Team Members – Abhik Choudhury, Ganapathy Raaman Balaji, Vamsi Perumallu Yalamanchili

	(p,q)	AIC
0	(0, 0)	6644.347367
1	(1, 0)	6645.964747
2	(0, 1)	6645.983116
3	(1, 1)	6647.815484
4	(2, 0)	6647.836935
5	(2, 1)	6649.811375

Fig 10: AIC values for the different combinations of p and q

Looking at the different values, we selected ARMA (0, 1) model. The next step was to evaluate goodness of fit of the model (as shown in Fig 11).

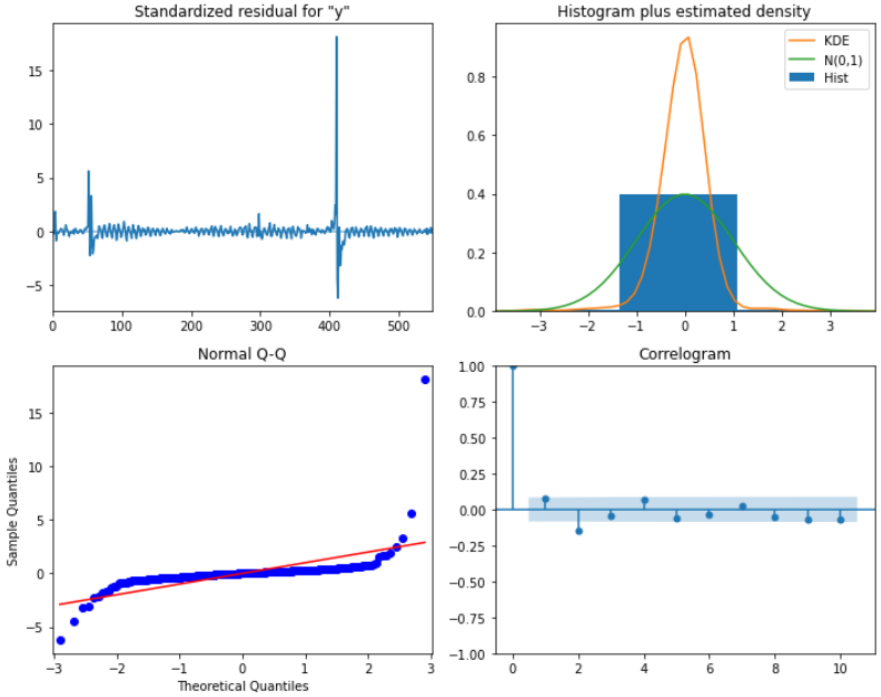

Fig 11: Goodness of fit for ARMA (0, 1) From the top left (clockwise) – Residual plot, residual histogram, ACF plot of the residuals and Q-Q plot.

The goodness of fit (Q-Q Plot) shows that the residuals are fat tailed, and the distribution is different from normal distribution. The p – values of each lag is shown in fig 12. Each p-value is less than 0.05. Thus, at each lag we can reject the null hypothesis, and the residuals are not independently distributed and are correlated.

```
from statsmodels.stats.diagnostic import acorr_ljungbox

residuals = model_fit.resid
lbvalue, pvalue = acorr_ljungbox(residuals, np.arange(1, 11, 1))
print(pvalue)
```

```
[0.0778887  0.00064512 0.00120788 0.00099649 0.00096322 0.00171958
 0.00308207 0.0032213  0.00212781 0.00132524]
```

Fig 12: p – values of each lag

Final Project 29 – Final Report

Team Members – Abhik Choudhury, Ganapathy Raaman Balaji, Vamsi Perumallu Yalamanchili

3. Forecasting:

Next, we forecasted the visits using the ARMA model. For this, we split the data into training and testing with a 75/25 split. Fig 13 shows the actual vs. predicted values.

visits_diff	pred_ARMA
39.0	-62.921537
32.0	-36.257580
-122.0	-3.870071
100.0	8.147730
-111.0	108.332666

Fig 13: Actual vs. Predicted values

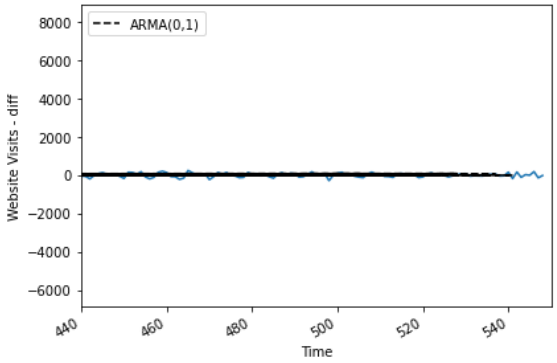

Fig 14: Actual vs. Predicted plot

We can observe that the prediction is not very accurate using ARMA model. Mean Squared error is 370958.28. This high value of error is also indicative of the fact that ARMA might not be the best model. Fig 14 plots the predicted vs actual values for ARMA(0,1) model.

Conclusion:

The model that we decided for this webpage forecast is ARMA as this data has an Auto Regressive component and Moving Average component. Not all the webpages might be stationary. So, we might need to do differencing to get them as stationary. The data was differenced to make it stationary. But the results are not accurate, and might need more work done.

Future Work:

For future, we would need to perform ARIMA and RNN modeling to forecast the webpages to get better results. Since the data is non-stationary, we propose ARIMA as a solution to get accurate results.

References:

[1] *Analytics/AQS/pageviews*. Wikitech. (n.d.). Retrieved December 4, 2022, from https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews

[2] Peixeiro, Marco. "6." *Time Series Forecasting in Python*, Manning Publications Co, Shelter Island, NY, 2022.

[3] Shelatkar, T., Tondale, S., Yadav, S., & Ahir, S. (2020). Web traffic time series forecasting using Arima and LSTM RNN. *ITM Web of Conferences*, *32*, 03017. https://doi.org/10.1051/itmconf/20203203017

[4] *Web traffic time series forecasting*. Kaggle. (n.d.). Retrieved December 4, 2022, from https://www.kaggle.com/competitions/web-traffic-time-series-forecasting/data

[5] Peixeiro, Marco. "7." *Time Series Forecasting in Python*, Manning Publications Co, Shelter Island, NY, 2022.