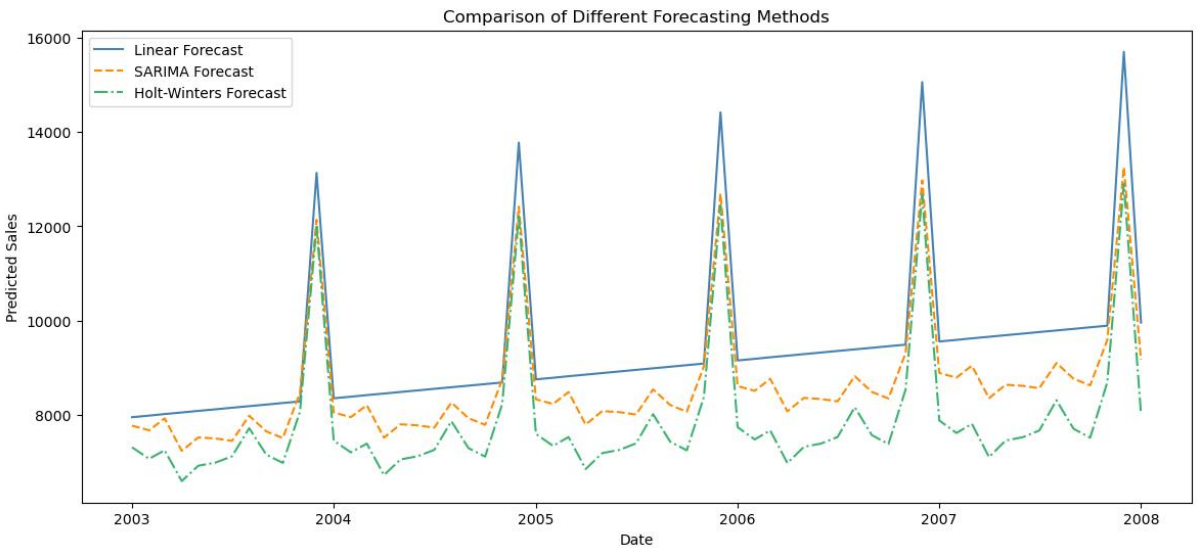
Time Series Analysis and Forecasting: Unveiling Insights from Diverse Datasets

Benjamin Nicholson
Seton Hill University



Introduction:

The use of time series has become increasingly apparent with the rise in computing power and data collection. Time series is the continuous series of data with an equal interval between each datapoint. The opportunity that arises with using a time series is the opportunity to break it down into individual components which are responsible for the movement of the data. The goal of this project is to build a comprehensive understanding of the time series datasets using the historical data which allows for future projections. These projections will be based on particular forecasting techniques such as

- Exponential Smoothing (SES, Holt Winter's Seasonal Method)
- ARIMA Modeling (AR, MA, ARMA, ARIMA, SARIMA, SARIMAX)
- Linear Regression

Utilising a variety of forecasting techniques with their own advantages offers a variety of forecasts to be made which can be evaluated using a variety of metrics. These techniques will be used on four different datasets.

- Sunspots Data : Real historical data on number of sunspots per month
- Website Traffic Data : Average monthly visits to an online website
- Electricity Consumption Data : Hourly electricity consumption
- Dinosaur Park : Monthly sales of dinosaurs

Methodology:

1 **Exponential Smoothing (SES)**:

Exponential smoothing is a simple and commonly used forecasting technique for time series data. It involves recursively updating a weighted average of past observations to generate forecasts. The weighting decreases exponentially as observations move further into the past, giving more recent data greater importance in the forecast. SES is particularly useful for smoothing out irregular fluctuations in data and generating short-term forecasts.

2 **Holt Winter's Seasonal Method**:

Holt-Winters' seasonal method, also known as triple exponential smoothing, extends exponential smoothing to handle seasonality in time series data. It comprises three components: level (average value of the series), trend (rate of change in the series), and seasonality (repeating patterns within the series). Holt-Winters' method includes parameters for smoothing these components, allowing the model to capture both trend and seasonality in the data. This technique is especially effective for forecasting data with both trend and seasonal patterns.

3 **ARIMA Modeling (AR, MA, ARMA, ARIMA, SARIMA, SARIMAX)**:

ARIMA stands for Autoregressive Integrated Moving Average, which is a popular and powerful time series forecasting method. ARIMA models are capable of capturing a wide range of temporal dependencies in data. The acronym ARIMA denotes three main components:

- Autoregression (AR): This component models the relationship between an observation and a number of lagged observations (autoregressive terms). It captures the linear dependence between an observation and its lagged values.
- Moving Average (MA): This component models the relationship between an observation and a residual error from a moving average model applied to lagged observations. It captures the effects of random shocks or noise in the data.
- Integration (I): This component accounts for differencing, which transforms a non-stationary time series into a stationary one by taking differences between consecutive observations.

- SARIMA and SARIMAX (Seasonal ARIMA and Seasonal ARIMAX) extend the ARIMA model to handle seasonal patterns in the data. They include additional seasonal autoregressive and moving average terms to capture seasonal variations.

4 **Linear Regression**:
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In the context of time series forecasting, linear regression can be applied when there is a linear relationship between the predictor variables (e.g., time) and the response variable (e.g., sales). It estimates the coefficients of a linear equation that best fits the observed data, allowing for the prediction of future values based on the observed trend. While linear regression is simple and interpretable, it may not capture complex temporal patterns or nonlinear relationships in the data as effectively as other methods.

With the knowledge of each of these methods they can applied appropriately to the different datasets. All of the data was in the format where it had a date column and then a value column so there was no need for feature engineering or modification of the data to improve forecasting.

Method

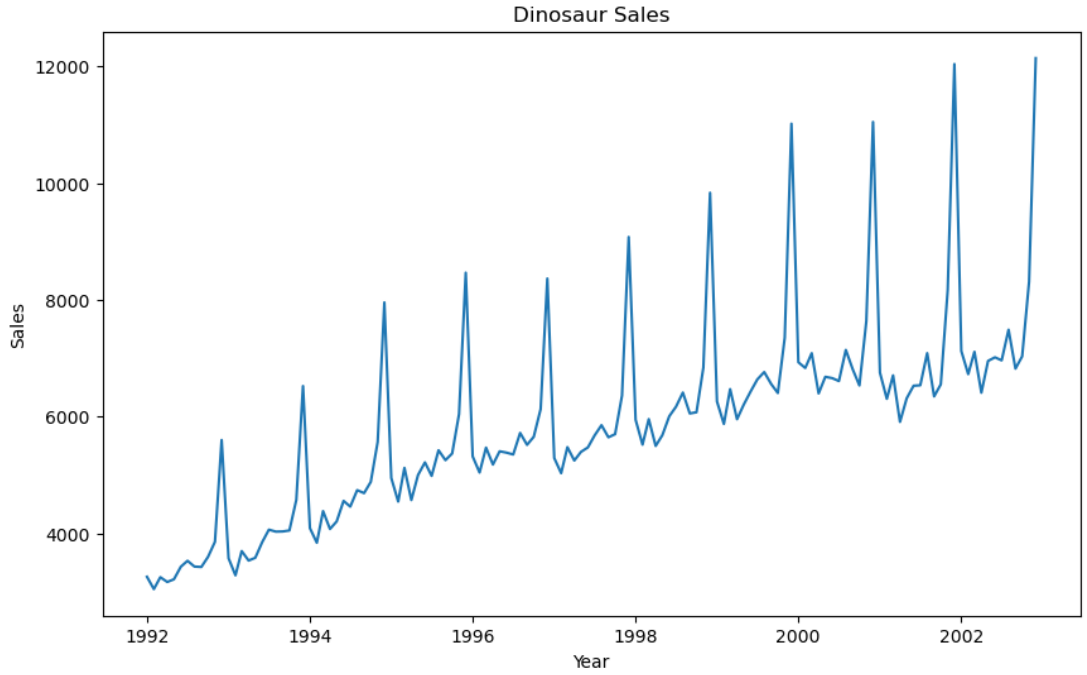
1. Preprocess data: Ensure that the data has datetime columns and output data with no missing data. You can also check for anomalies looking the z-score of all different data points. If it is greater than 3 then it can be considered an anomaly and the decision to remove these datapoints will be considered, taking into account their impact on analysis. Aggregation is necessary if the data does not clearly display movement in data.
2. Exploratory Data Analysis: This section is about understanding the type of data that is present and what patterns can be picked up. Patterns can be observed both visually and with using techniques such as ACF and PACF plots. There use is explained in the code files, these plots can help describe correlations with different time intervals to the current data. It is essential to break down the data using seasonal decomposition and understanding trend, seasonality, cyclical nature and irregularities.
3. Model Building : As discussed above ARIMA is a very powerful forecasting tool. However, this section is also going to include SES and Holt Winter's Seasonal Method as they offer their own benefits to explaining the data. When looking to create an ARIMA model you have to test different p,d and q values which are reflective of AR, I and MA respectively. This is discussed above. With the range of values for the ARIMA model they are evaluated using AIC and BIC values. The best few models will then be used in the next section
4. Model Evaluation : Incorporating the models with the lowest AIC and BIC, these models with be tested to see their accuracy. The data is splt into training and test data to understand how their performance measures up using a variety of evaluation metrics such as MAE and MSE. Whichever had the lowest would be used to forecast duture data.
5. Forecasting : Using the best model from before forecast future data for an appropriate time period.

Dinosaur Sales

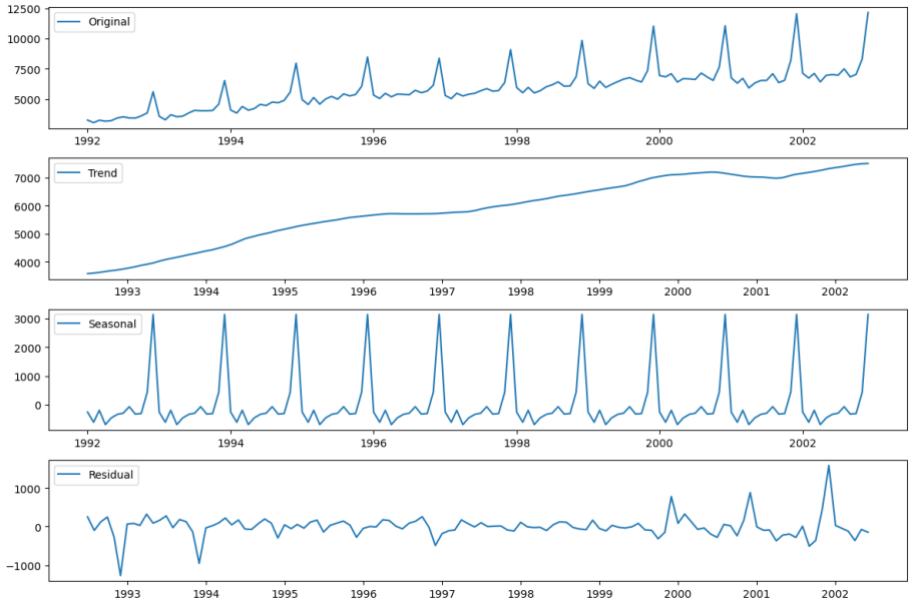
Preprocess:

The dinosaur dataset had a year and a month column alongside a sales column. This meant that in order to use the dataframe the year and month columns would need to be converted to datetime. Besides that the integrity of the data was intact and there was no need for removal of anomalies or checking for missing values

EDA



The data spans from 2002 to 2004 where it shows a strong linear trend with a seasonal component which grows with time. This suggests that it might be a multiplicative Holt Winter's Seasonal model. It was found that the seasonality took place every 12 values during the last month of the year. This suggested that during December the sales for dinosaurs went way up. The image below is the seasonal decomposition with seasonal period of 12



The seasonal decomposition shows a very good model where the residual is low but it is clear the peaks are going from small to high which confirms the suspicion that it has a multiplicative seasonal component. Analysis showed that there was two rates of growth, for months 1-11 and for the 12th month. Combining these two growth rates could be a good basis for a linear regression forecasting model.

Model Building

Through visual inspection of ACF, PACF and ADF tests the values for pdq were found. As well as the seasonally adjusted PDQ values were also found by visual inspection of new ACF, PACF and ADF tests, using that seasonally adjusted data. The results are shown to the right. Using the ARIMA function found in the code file it returned this as the lowest AIC and BIC. As discussed in the comments, using the higher order for ARIMA models is not its purpose so for future ARIMA modelling I will keep it at less than 5. As values greater than this are likely going to belong to those plots that include seasonality.
This was observed where the SARIMA model with (0,1,2,0,1,0,12) had an AIC of 1618 and BIC of 1927 which is far lower than the other ARIMA models.

	Model	Order	AIC	BIC
19	ARIMA	(12, 1, 1)	1864.071460	1904.324222
21	ARIMA	(12, 2, 1)	1866.141284	1906.286766
10	ARMA	(12, 0, 12)	1885.061289	1960.014139
9	ARMA	(12, 0, 1)	1993.021178	2036.263207
20	ARIMA	(12, 1, 12)	2039.007024	2110.886957
22	ARIMA	(12, 2, 12)	2048.009196	2119.697557
2	AR	(12, 0, 0)	2052.195855	2092.555082
16	ARIMA	(3, 1, 12)	2152.279864	2198.283021
12	ARIMA	(1, 1, 12)	2153.848926	2194.101688
4	MA	(0, 0, 12)	2170.377692	2210.736919
6	ARMA	(1, 0, 12)	2199.331267	2242.573296
8	ARMA	(3, 0, 12)	2207.348803	2256.356436
18	ARIMA	(3, 2, 12)	2210.727113	2256.607664
11	ARIMA	(1, 1, 1)	2223.493659	2232.119251
15	ARIMA	(3, 1, 1)	2224.857452	2239.233438
17	ARIMA	(3, 2, 1)	2227.173235	2241.510907
13	ARIMA	(1, 2, 1)	2242.229042	2250.831645
7	ARMA	(3, 0, 1)	2244.469200	2261.766012
5	ARMA	(1, 0, 1)	2247.723088	2259.254296
14	ARIMA	(1, 2, 12)	2254.008272	2294.153754
1	AR	(3, 0, 0)	2254.530064	2268.944074
0	AR	(1, 0, 0)	2264.718990	2273.367396
3	MA	(0, 0, 1)	2290.344040	2298.992446

Model Evaluation

The following were the errors for the different ARIMA models
ARIMA (12,1,1) MAE: 378.1747824813723
ARIMA (1,1,1) MAE: 917.583286119831
SARIMA (0,1,2) (0,1,0,12) MAE: 401.85417090161354

It is clear that the SARIMA model once again performs best which follows with what was found in the AIC and BIC.

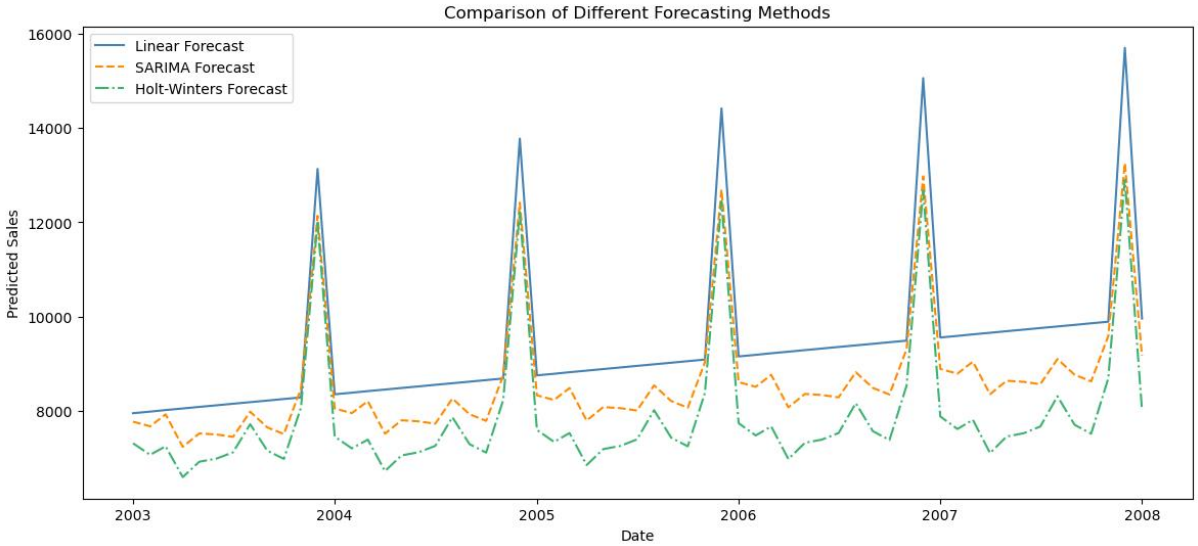
To compare how these did with other methods, I used the Holt Winter's Seasonal Method as well as Regression Forecasting.

Holt Winters MAE : 238.75
Linear Forecasting MAE: ~317

The best training test model was in fact the Holt Winter's Model.

Forecasting

I forecasted each of these on a graph which can be observed below and it compares how each different model will give slightly different predictions offering their own benefit.



Linear forecast obviously does not have small adjustments from month to month and is going to be a good basis by using least square lines to make predictions.

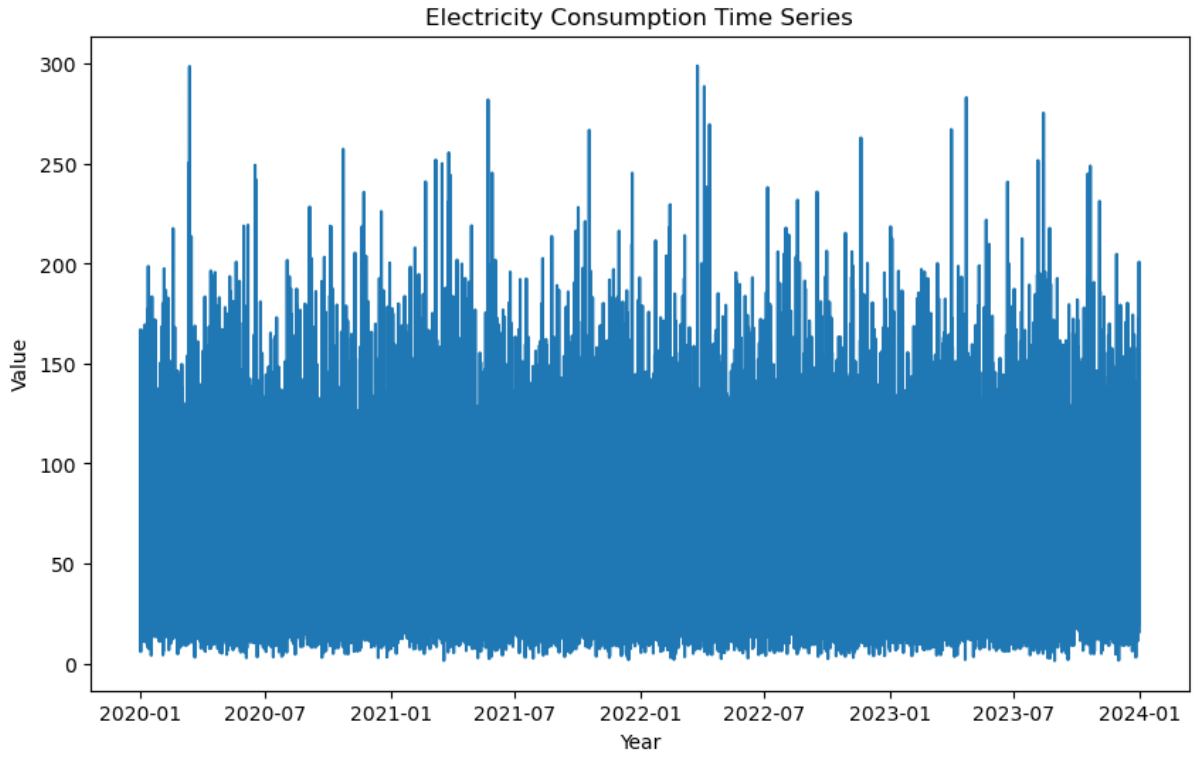
Holt Winter's had the most accurate model which is likely because it is not as optimistic in the month to month data but still matches the peaks of the SARIMA forecast.

Electricity Consumption

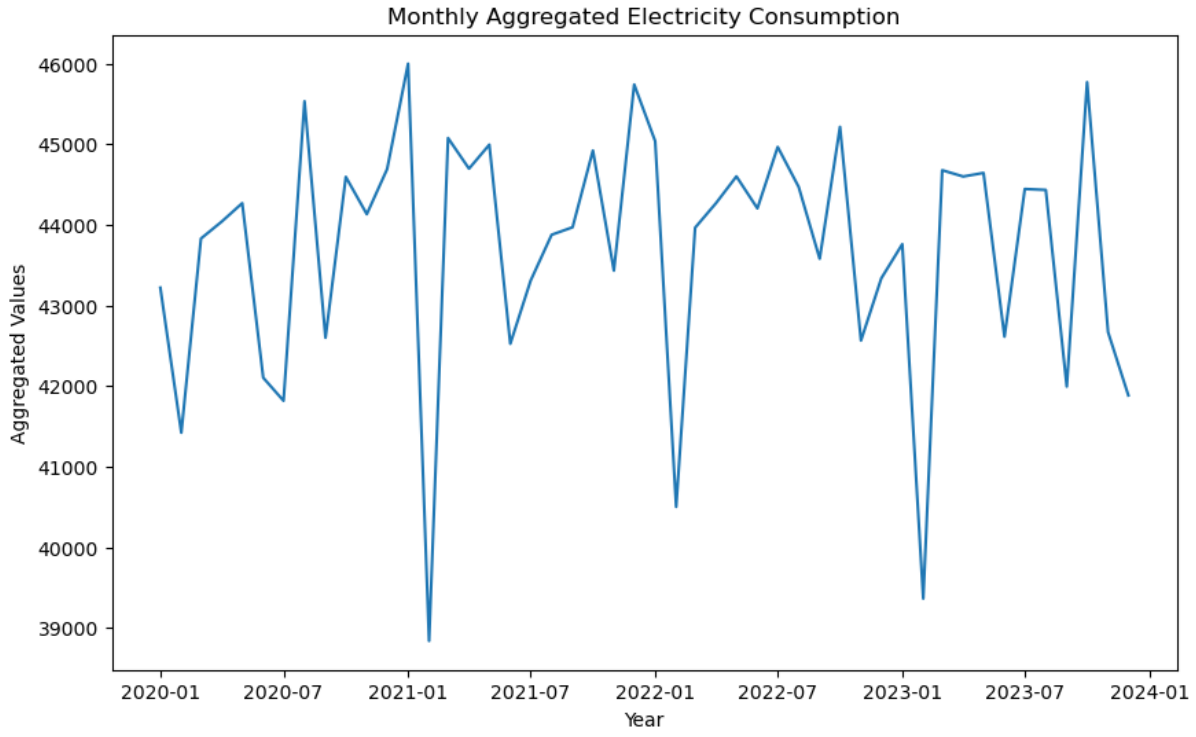
Preprocess

Electricity consumption had a date column in form YYYY-MM-DD HH:MM:SS where there was 35,041 entries. This resulted in the original graph of the data to be very hard to digest. Following checking for missing values and data anomalies it was found that there was 430 rows where the z score was above 3 standard deviations away from the mean. Due to the high amount of anomalies, there is going to be a dataframe with and without anomalies so that depending on the demands of the dataframe the appropriate one can be used. Removing data anomalies allows for the overall trend of the data to be understood as outliers are removed. However due to the high volume of them it is likely that these are actually essential to understanding the data. One further analysis that could have happened is checked the frequency of these occurring to understand if it took place on a reoccurring time period.

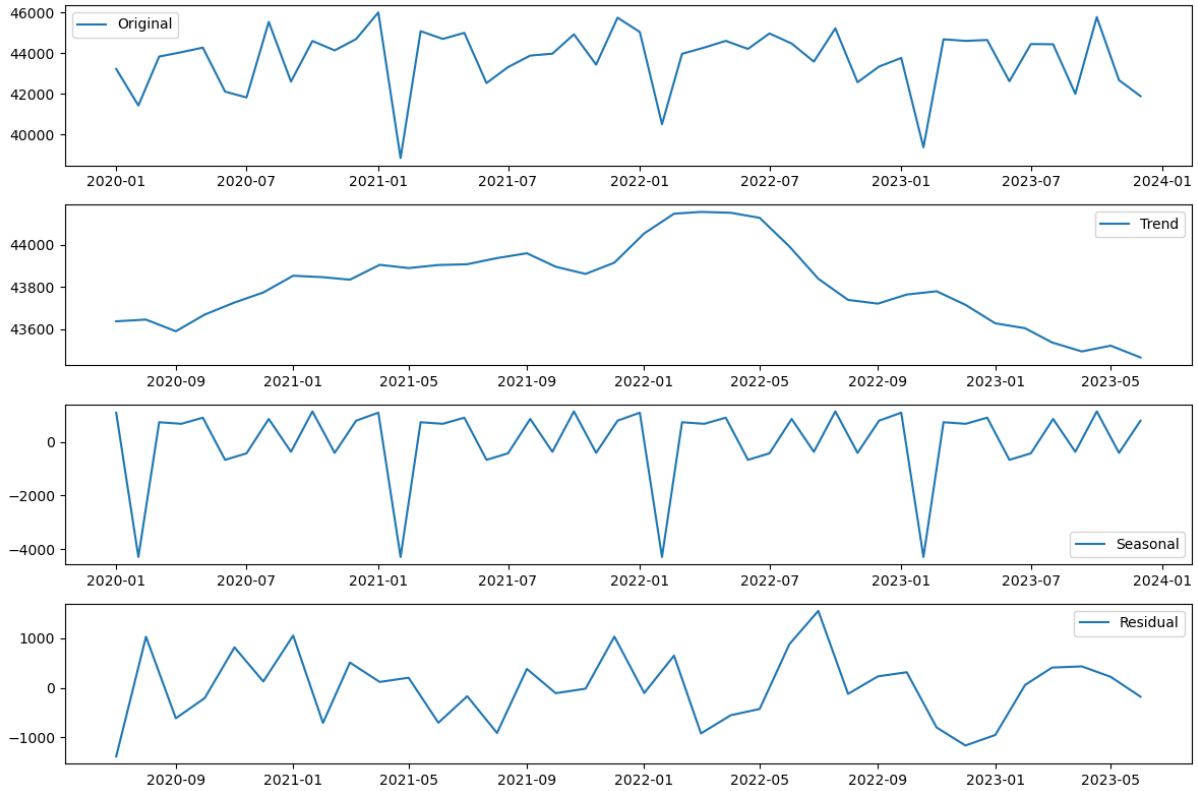
EDA



This plot shows the deeply concentrated the data is, this means that there needs to be aggregation applied to the data. By converting from hourly data to daily or weekly data it is going to be easier to visually digest patterns in the data. It was observed that using a monthly aggregation of the data would effectively summate the data in a way that can represent the movement of the data.



As the data was aggregated based off of month then it is likely that a seasonality would take place on a yearly period. It showed that the second month had a large decline which can be observed in the 'Monthly Aggregated Electricity Consumption' graph where the beginning of 2021, 2022 and 2023 all showed this large decline. Suggesting that electricity during this month is far lower.



The residual is equally spread amongst a mean of 0 which suggests that the overall pattern of the data has been understood. There is no real trend in the data but consumption in 2021-2022 is higher than prior and after to these time periods.

Model Building

Looking at the ACF and PACF plots did suggest that this data did not have that high of an autoregressive (p) and a moving average (q) stature as well as already being a very stationary time series. This explains why the most accurate models were MA of (0,0,1). However the SARIMA model (1,0,0,0,0,2,12) which included a seasonal period of 12 months did improve to a lower AIC than the non seasonal ARIMA models.

	Model	Order	AIC	BIC
2	MA	(0, 0, 1)	842.497007	848.110610
4	ARMA	(1, 0, 1)	843.587484	851.072288
8	ARIMA	(1, 0, 1)	843.587484	851.072288
0	AR	(1, 0, 0)	844.077841	849.691444
1	AR	(6, 0, 0)	848.217660	863.187268
3	MA	(0, 0, 6)	848.497862	863.467470
6	ARMA	(6, 0, 1)	849.169454	866.010263
10	ARIMA	(6, 0, 1)	849.169454	866.010263
7	ARMA	(6, 0, 6)	849.764093	875.960907
11	ARIMA	(6, 0, 6)	849.764093	875.960907
5	ARMA	(1, 0, 6)	850.961889	867.802698
9	ARIMA	(1, 0, 6)	850.961889	867.802698

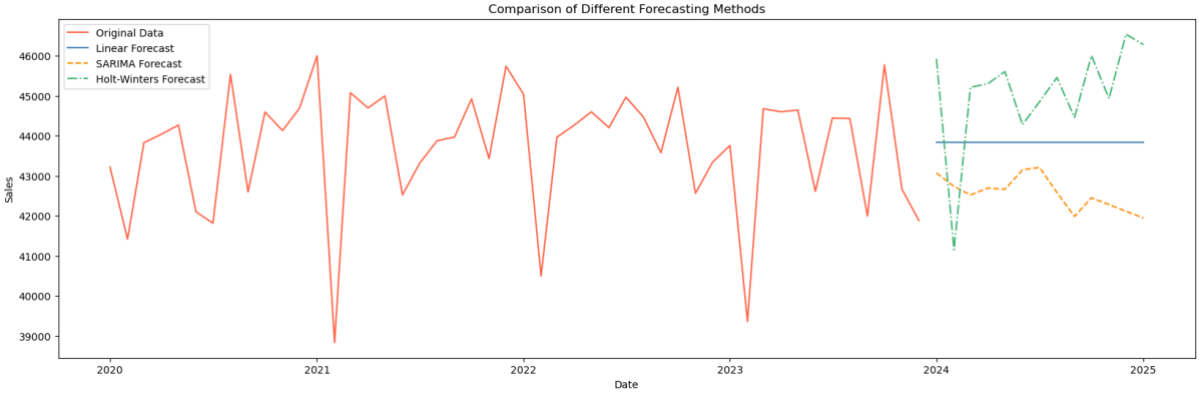
Model Evaluation

As the values for this dataset are very high using a percentage difference makes more sense when digesting the accuracy of the model.

```
mae ARIMA (0,0,1): 2.965495680661785
mae SARIMA (1,0,0) (0,0,2,12): 2.929138823372951
mae Holt Winters: 2.9183637270689826
```

The percentage values are all very similar where the Holt Winter's model once again outperforms the ARIMA and SARIMA models to get a more accurate train/test mae percentage. However as they are all very close in error rates each will be forecasted and understood.

Forecasting



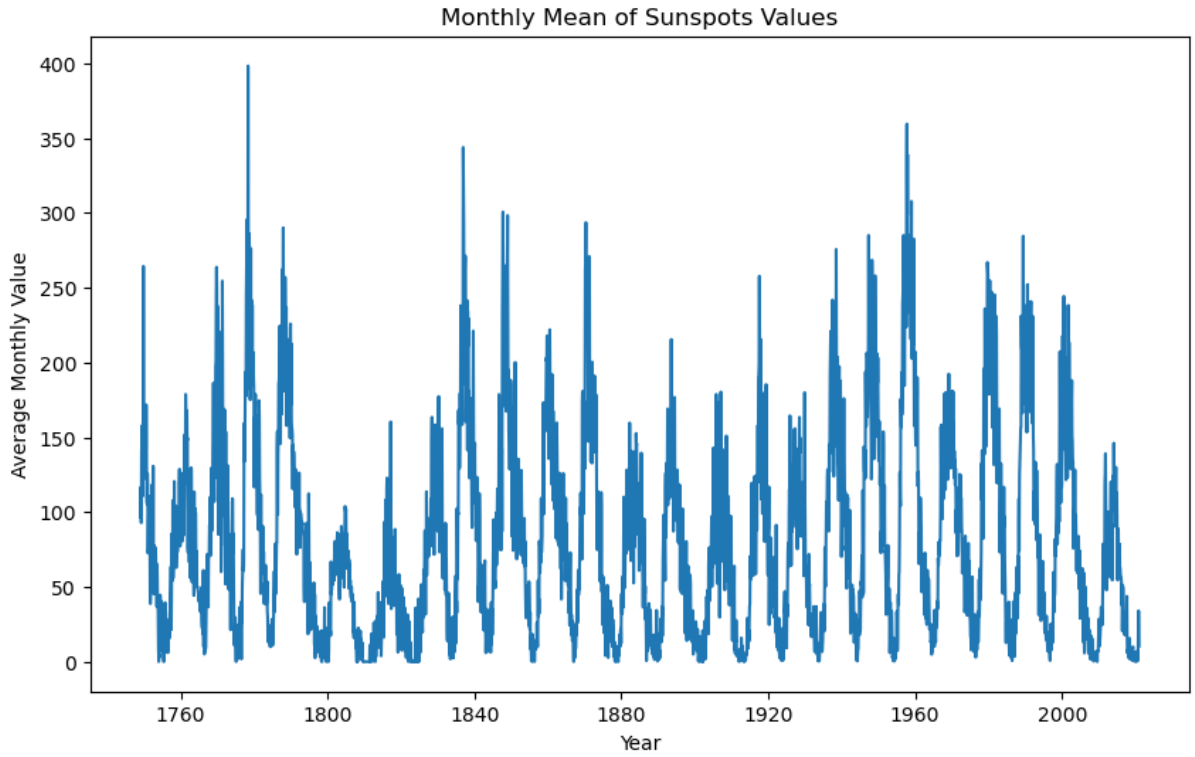
The forecast shows how each one of these took a different approach despite being trained on the same training and test data. The linear forecast likely makes most sense in continuing the stationarity of the graph where as the Holt-Winters forecast and SARIMA model forecast are going in opposite directions. The Holt Winters seasonal pick up the large drop in values at the start of 2024 where as the SARIMA looks to continue on the slight downwards trend that is shown following the high period of 2021-2022. It is likely that the linear forecast / ARIMA (0,0,1) is going to be best overall prediction to not overfit the data and understand the overall trend.

Sunspots Data

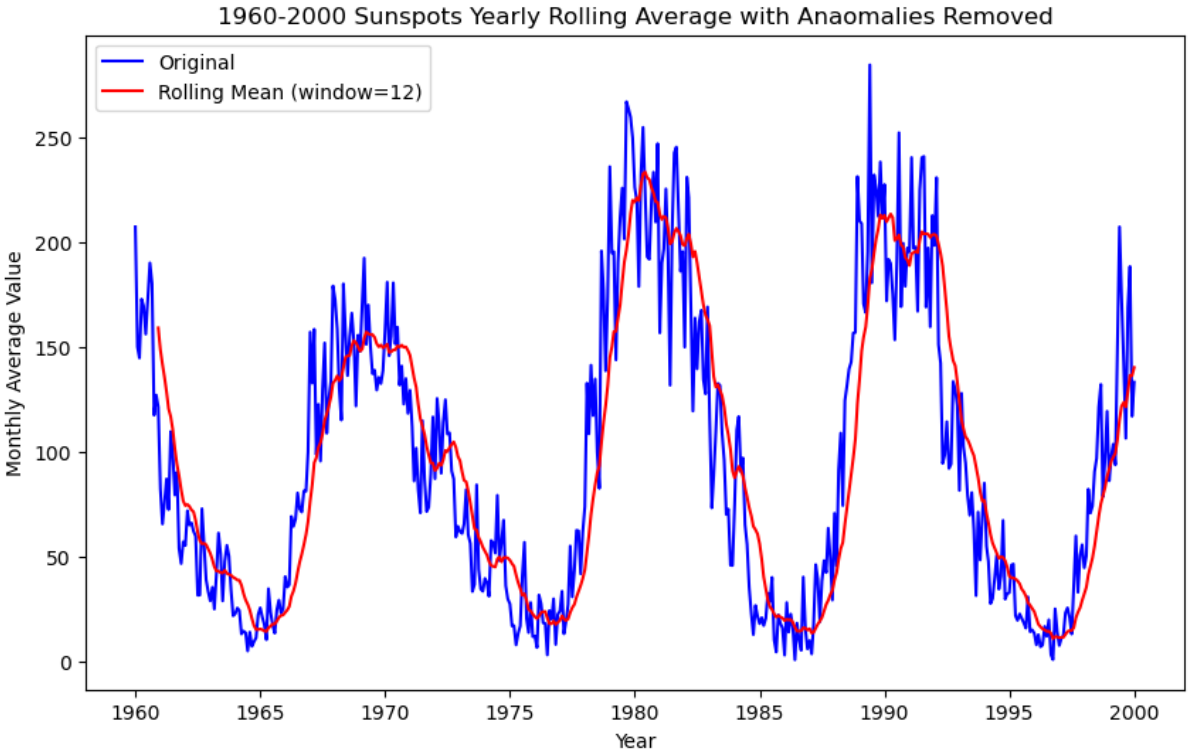
Preprocess

Sunspots is real data that has been taken from 1749 to 2021. It has the monthly mean total sunspots number with a column of date which is in the form YYYY-MM-DD. Converting it into a datetime where the day is at the start of the month rather than the end is going to help with interpretability so that the days are not jumping around from 31 in January to 28 in February. After checking for null values which was none, the number of anomalies was quite low but it was found that they took place during certain clumps. This might have been due to certain cosmological events causing the sun to have more sunspots during this period. They were in 1778, 1836-1837 and 1957-1959 which meant their values were three standard deviations away from the mean.

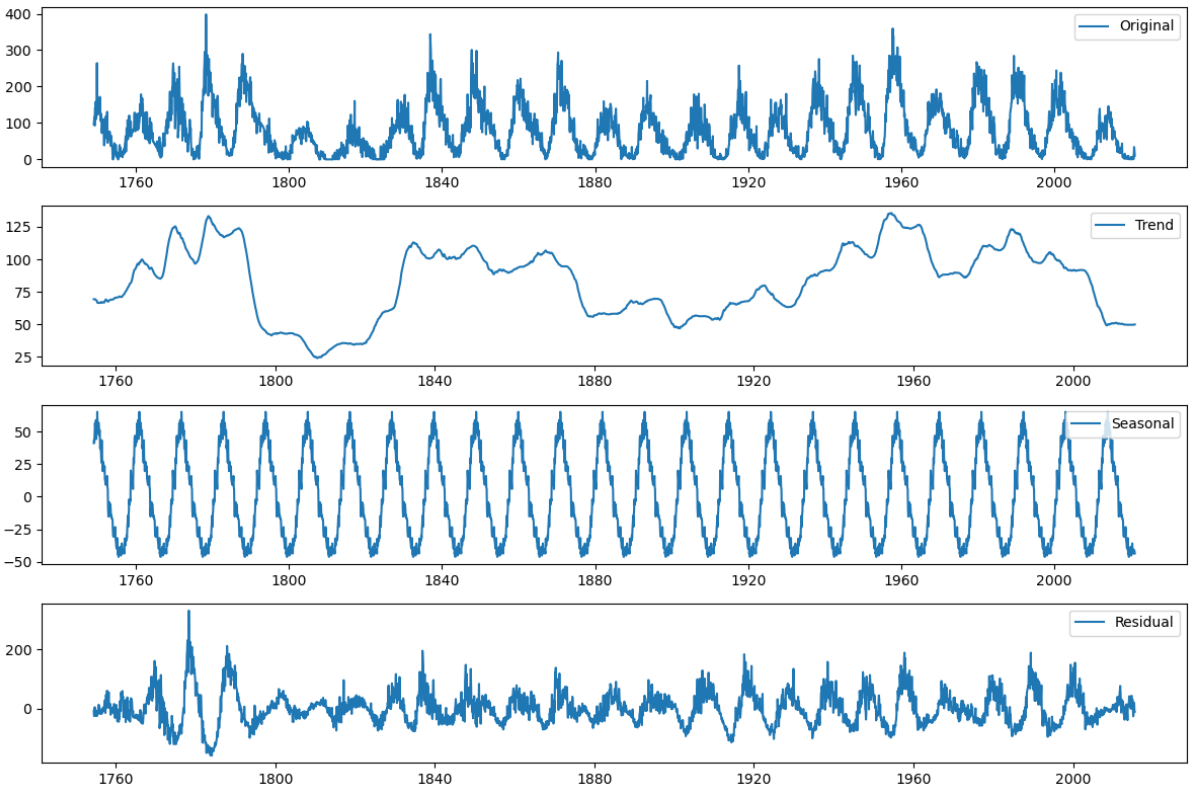
EDA



The data clearly shows a strong pattern where values will return close to 0 and to some average maximum of about 150-200 monthly sunspots. It is hard to visually observe the pattern as there are so many of them. The ACF and PACF plots also do not show any conclusive data for the seasonality. The graph below shows a zoomed in data selection from 1960 to 2020.



It was found that there was a seasonal period of 11 years. Further investigation into this dataset suggests that due to the poles of the sun swapping every 11 years has an influence on the number of sunspots measured per month.



The decomposition displays the strong seasonal pattern and decent residual. However there does seem to be a pattern to the residual, further improvements could consist of exploring this

pattern. There is no overall trend and it does suggest a stationary dataset with an ADF statistic of -10.5 when only needed a -3.43 for a 99% confidentiality that the data will be stationary.

Model Building

Due to the complexity of the data fitting a seasonality of 132 intervals (11 years of 12 values per year) it was very difficult to find a seasonal ARIMA model. Further improvements could be to take a sample of the data to try to build a less complex model. The ARIMA models were all very low pdq values as for the complexity of the ACF and PACF plots which did not display any real pattern in the autocorrelation and partial autocorrelation for q and p values respectively. It was already discussed how the stationarity of the graph was proven with the ADF test so the d value can be left at 0.

	Model	Order	AIC	BIC
15	ARIMA	(2, 1, 2)	30250.957980	30281.411524
5	ARMA	(1, 0, 2)	30316.781285	30347.236360
9	ARIMA	(1, 0, 2)	30316.781285	30347.236360
7	ARMA	(2, 0, 2)	30318.779675	30355.325766
13	ARIMA	(2, 0, 2)	30318.779675	30355.325766
6	ARMA	(2, 0, 1)	30321.223487	30351.678562
12	ARIMA	(2, 0, 1)	30321.223487	30351.678562
14	ARIMA	(2, 1, 1)	30329.219256	30353.582091
11	ARIMA	(1, 1, 2)	30331.183799	30355.546633
10	ARIMA	(1, 1, 1)	30333.403911	30351.676037
4	ARMA	(1, 0, 1)	30360.870706	30385.234767
8	ARIMA	(1, 0, 1)	30360.870706	30385.234767
1	AR	(2, 0, 0)	30515.008081	30539.372141
0	AR	(1, 0, 0)	30766.404603	30784.677648
3	MA	(0, 0, 2)	32960.383449	32984.747509
2	MA	(0, 0, 1)	34223.100649	34241.373694

Model Evaluation

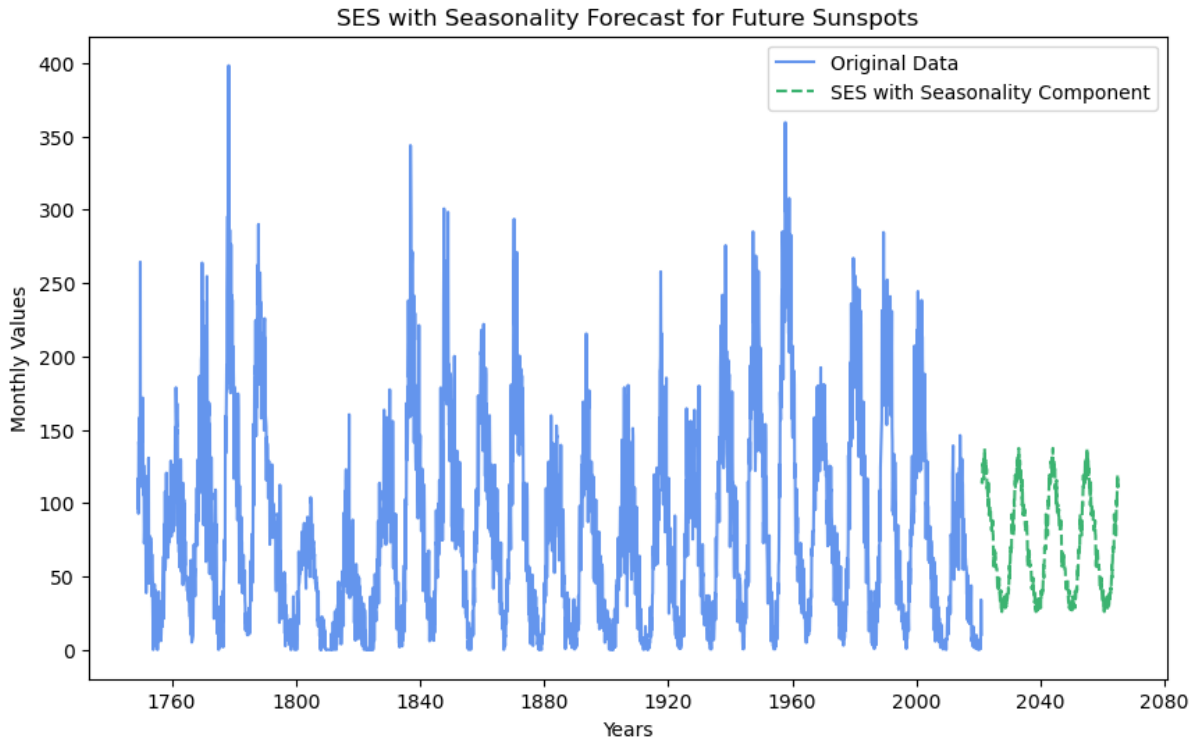
As the ARIMA model creation was not that accurate due to the complexity of the data. Involving an SES model which is simply taking the smoothing of recent data values as well as using the Holt Winter's Seasonal Method to include the 11 year period of the season is likely to give an accurate model.

```
SES mae: 57.97045909511802
SES with Seasonal Components mae: 40.83485855936685
ARIMA mae: 63.86739600092766
Holt Winters mae: 49.145153591694005
```

The list of mae displays that's using an SES which takes the recent mean going through the data in combination with the seasonality as stated before gives the most accurate model. Using the SES without a seasonal component still gives a more accurate model than MAE while the Holt Winter's Seasonal Method once again displays a good model however goes not find the mean as well as the SES.

Forecasting

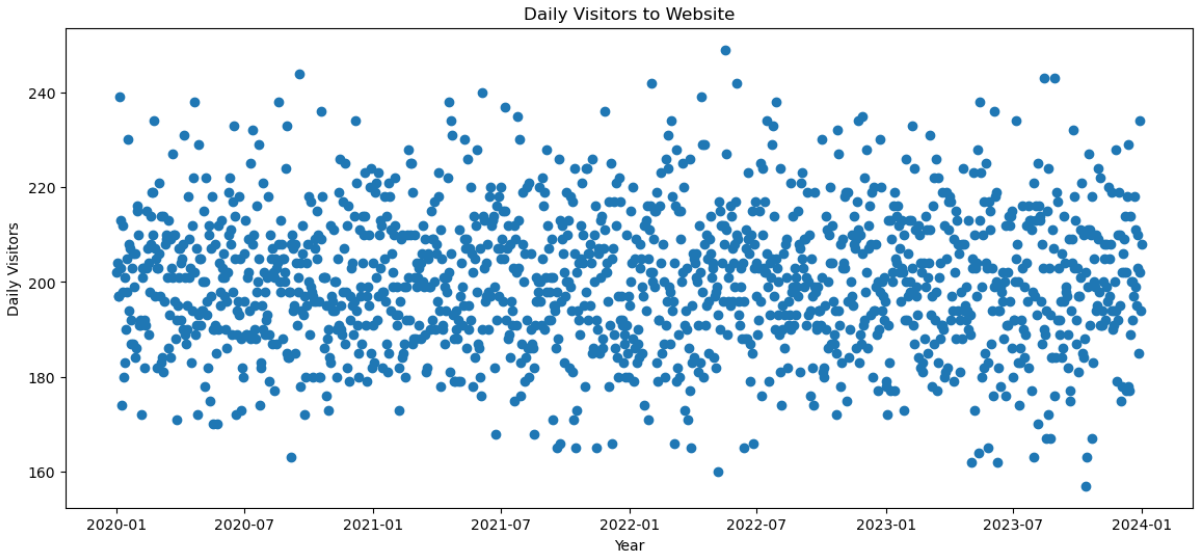
Using the SES with seasonal component the forecast until 2070 is going to be listed. This suggests that the monthly sunspots will continue to remain stationary with a seasonal fluctuation based on the time interval in the 11 year cycle. The irregularities as observed in previous values will likely return at some point in the future however this is quite an accurate basis for the prediction of sunspots over the next 50 years.



Website Traffic

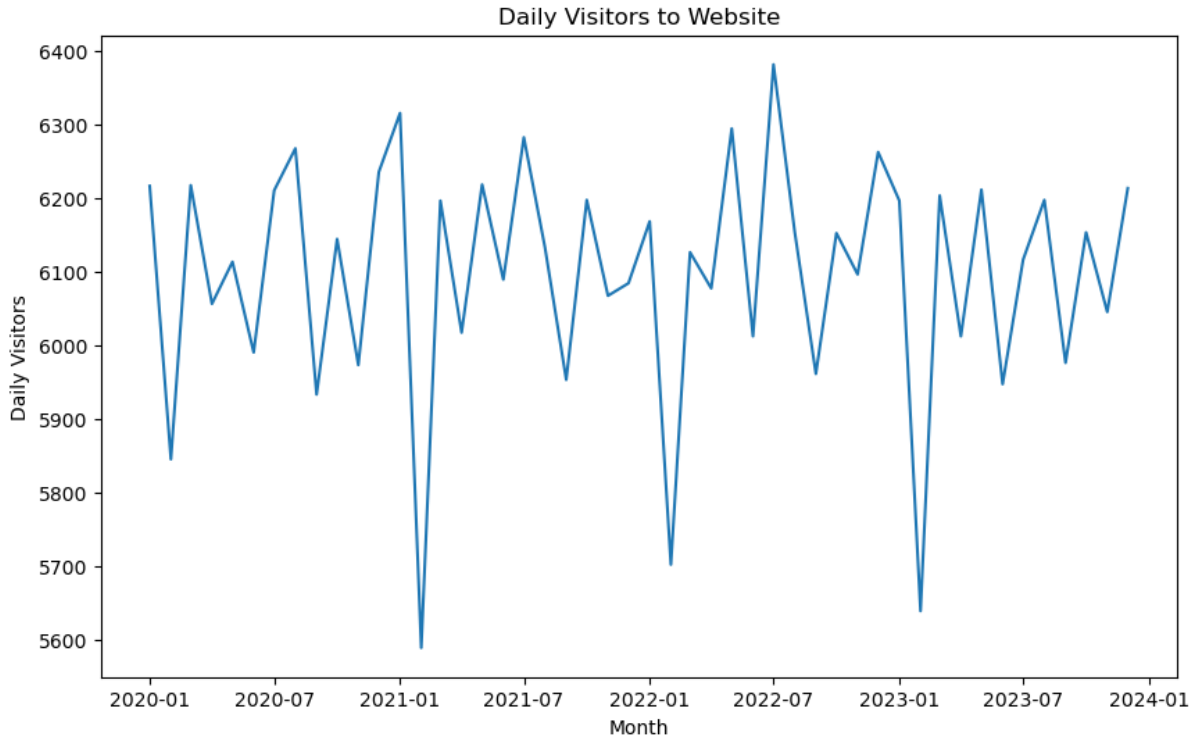
Preprocess

Website traffic returned a datetime in form of YYYY-MM-DD with a value column of daily visitors. There were no missing values and there was just one value that was considered anomalies. However it was found that the data was very concentrated and in order to understand the data, there needed to be some form of aggregation. Using month as an aggregation it shows a very similar pattern to the data that was found in Electricity Consumption data.

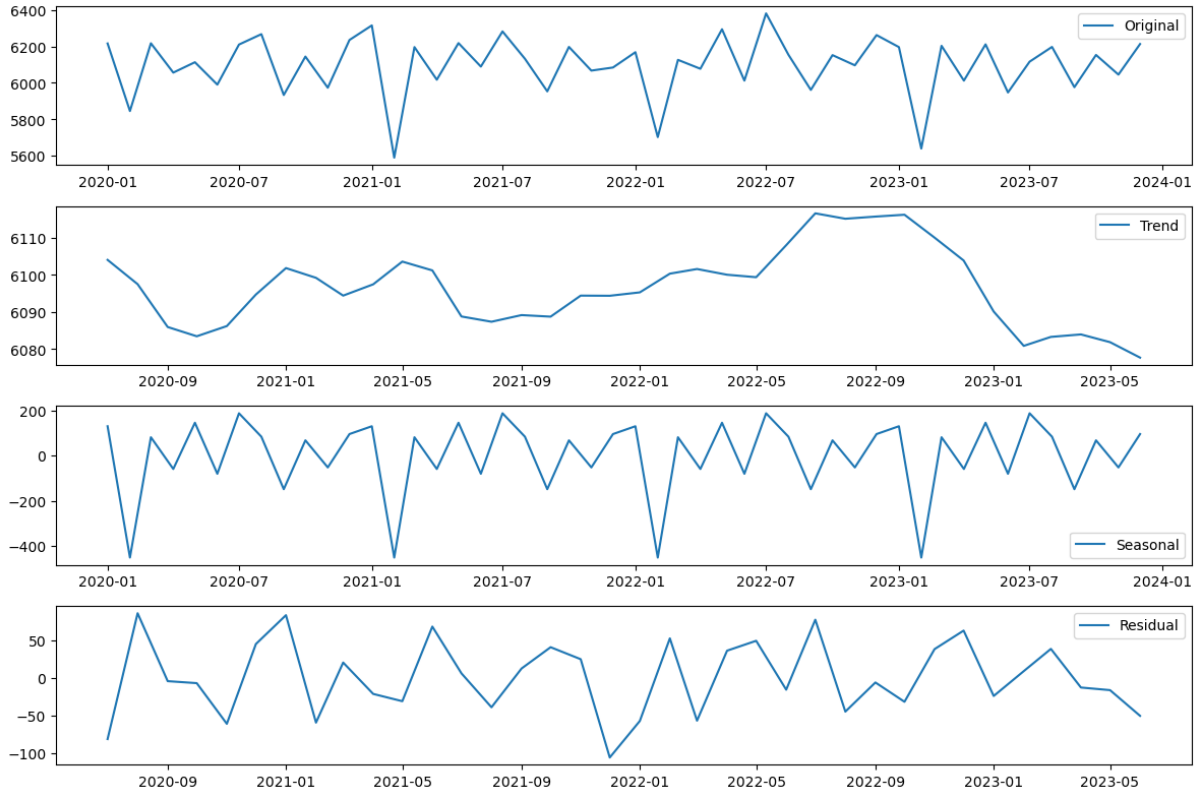


EDA

Once the data has been aggregated on a monthly basis it creates the following line series



The ACF and PACF plots both suggest that there is higher correlation at lags 12 so this is going to be assumed as the seasonal period.



The trend of the data is quite stationary with a spike taking place in 2022 to 2023. The residual shows a altering values around a mean of 0 which suggests that the pattern of the data has been decomposed. The seasonal period shows a value that spikes downwards far greater than other time periods of the 12 month time interval.

Model Building

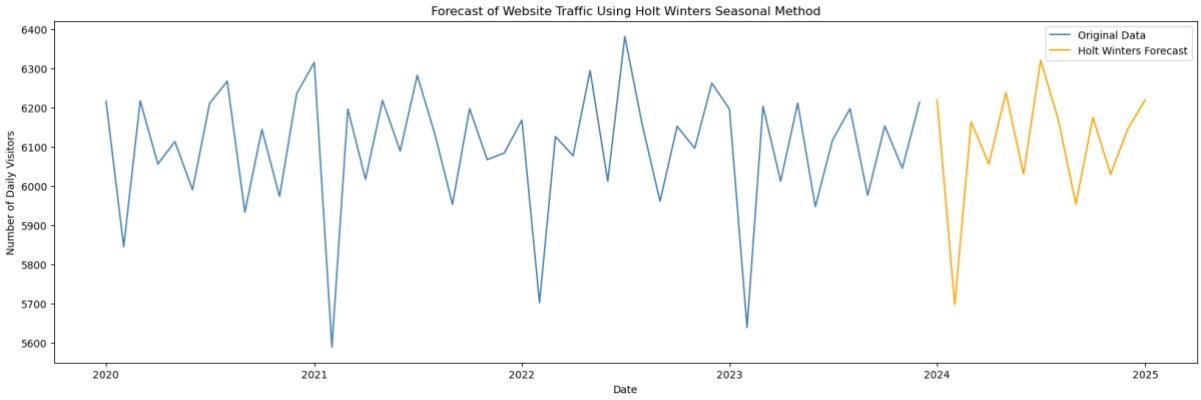
Once again the SARIMA model (0,0,1,0,0,1,12) outperforms the ARIMA models. This is because incorporating a seasonal component when there is any type of relationship that happens over a cyclical interval is going to prefer a SARIMA model which in this case had an AIC of 602 compared to the next closest of 611, from ARIMA (2,1,5). The SARIMA model has a moving average of 1 with everything else kept the same.

	Model	Order	AIC	BIC
15	ARIMA	(2, 1, 5)	611.801164	626.602345
10	ARIMA	(1, 1, 1)	614.966322	620.516765
7	ARMA	(2, 0, 5)	615.061804	631.902613
13	ARIMA	(2, 0, 5)	615.061804	631.902613
4	ARMA	(1, 0, 1)	616.132897	623.617701
8	ARIMA	(1, 0, 1)	616.132897	623.617701
14	ARIMA	(2, 1, 1)	616.611260	624.011851
11	ARIMA	(1, 1, 5)	617.468058	630.419092
6	ARMA	(2, 0, 1)	617.561590	626.917596
12	ARIMA	(2, 0, 1)	617.561590	626.917596
2	MA	(0, 0, 1)	619.949810	625.563413
3	MA	(0, 0, 5)	621.026591	634.124998
0	AR	(1, 0, 0)	622.604891	628.218494
5	ARMA	(1, 0, 5)	622.679063	637.648671
9	ARIMA	(1, 0, 5)	622.679063	637.648671
1	AR	(2, 0, 0)	623.757966	631.242770

Model Evaluation

Investigating the best performing ARIMA model (without seasonal components), the SARIMA model and Holt Winter's Seasonal Method. The most accurate model was Holt Winter's Seasonal Method with a percentage error of just under 1% for the test/training data. This will be the only model that will be forecasted as the SARIMA (0,0,1,0,0,1,12) model failed to understand the movement of the data and the ARIMA (2,1,5) had an error of about 3%.

Forecasting



The model follows the stationarity nature of the data while capturing the large drop that takes place in the 2nd value of the seasonal period It also correctly highlights that the rest of the months usually take on some random value varying with a certain range.

Future Improvements

1. Utilising forecasting uncertainty is a big component that will be beneficial to improving readability and accuracy of the model as there is always uncertainty involved in the future.
2. Improving on the understanding of parameters of Exponential Smoothing techniques for Holt Winters method. With a greater explanation and testing of different parameters a more accurate model can be made.

Conclusion

Each of the investigated forecasting techniques offer their own benefit to different types of datasets. It does seem that Holt Winter's Seasonal Method thrives on seasonal factors and can sometimes outperform the SARIMA model dependent on the accuracy of the parameters which influence Holt winter's Seasonal Method. Where as ARIMA models perform better in capturing temporal dependencies and linear regression as observed. The linear prediction only works when the least squares values is very small and oscillate around that least squares with no specific pattern. For the future using uncertainty parameters will help the readability and accuracy of the forecasts. As well as utilising some type of parameter optimisation for the SES model.