Time Series, Assignment 1

Petersen Trentin (PTRTRE004)

September 15, 2024

Abstract

Within this report are the findings and conclusions made for a project conducted for Eskom, with a strong focus on the forecasting of peak daily electricity demand and total energy lost due to unplanned outages (UCLF+OCLF). Using historical data from April 1, 2019 to September 13, 2023, a multitude of forecasting techniques and models were employed to produce valuable insights into the energy supply issues faced by South Africa. Thorough analysis of these forecasts by Eskom might enable the mitigation of load shedding and unplanned outages which will ultimately result in more stability in the national grid.

Introduction

Over the past decade, the demand for electricity in South Africa has only increased which has proven itself to be a tough challenge for Eskom, South Africa's primary supplier of power. Unplanned outages caused by ageing infrastructure and poor management combined with the sharp fluctuations in the demand for electricity result in hindrances in meeting the energy demands of South Africa as well as causing alarming rates of instability in the national power grid. This has led to an increase of the amount of load shedding hours in South Africa, even reaching a new record of consecutive days of load shedding in 2023 which had a detrimental impact on the economy of South Africa and the lives of many people. The aim of this project is to produce robust forecasts for the peak daily electricity demand as well as the proportion of energy lost due to all unplanned outages.The robust forecasting of these metrics is vital to Eskom's success in reducing the frequency of load shedding and improving the stability and reliability of the national grid.

The data which will be analysed and then used to create forecasts is named 'ESK6816.csv' and is provided by Eskom. The dataset contains hourly observations beginning $01/04/2019 \quad - \quad 00:00$ and ending $31/03/2024 \quad - \quad 23:00$. For the sake of the analysis and forecasting there are two primary variables of interest, $RSA.Contracted.Demand$ and $Total.UCLF.OCLF$. $RSA.Contracted.Demand$ represents the hourly average demand that Eskom must fulfill. $Total.UCLF.OCLF$ represents the total proportion of Eskom's plant capacity that unavailable due to unplanned outages.

Questions/Hypotheses

Eskom states that there are two peak electricity demand periods within a day: 6am to 9am and 5pm to 9pm, the focus of this report will be on the latter. During this period of the day is when most households in South Africa use appliances such as electric stoves, heaters and entertainment devices; at the same time there may also still be some commercial or industrial activities which overlap with this increased household demand. This spike in electricity demand, puts a lot of strain on the grid and the ageing infrastructure.

It is hypothesized that the seasons of the year as well as the economic activity will have a large impact on the daily peak demand, causing an annual pattern to emerge, with demand peaking in winter and reaching a minimum in summer. A weekly pattern is also expected as during the work week, Monday to Friday, the demand will go up and on the weekends the demand will greatly reduce. These predictable seasonal patterns should assist with producing robust forecasts.

Producing robust forecasts for the energy loss due to the unplanned outages will be more of a challenge as it depends on factors such as the availability of funds and skilled personnel to maintain the ageing infrastructure and water availability for the cooling of the large number of coal power plants. A slight upward trend in the energy loss due to unplanned outages is expected due to ageing infrastructure.

Using the forecasts, some critical questions might be answered. For instance, how can load shedding as well as unplanned outage frequency be minimized during the peak demand period 5pm - 9pm? Additionally how can electricity pricing and contract negotiation be optimized to best mimic the patterns in the peak daily demand which will help reduce operational costs.

The forecasts produced will be a blend of short-term and medium-term forecasts which will be used to support operational decisions, grid management and resource allocation for maintanence and power generation.

Section 1 - EDA

1.1 - Data and variable names/classes

The data is read in from the ESK6816.csv dataset. The raw data consists of 22 variables and 43848 observations. Of the 22 variables we are only interested in the *Date*, *RSA.Contracted.Demand*, *RSA.Contracted.Forecast* and $Total.UCLF.OCLF$.

A POSIXct, *Date.Time*, variable is created and then filtered to only include the hourly observations from 5pm-9pm, the peak demand period. The table below presents a some summary statistics of the data for the two variables of interest - *RSA.Contracted.Demand* and $Total.UCLF.OCLF$.

1.2 - Summary statisitcs and Missing Observations

Table 1: Summary statistics - RSA.Contracted.Demand & Total.UCLF.OCLF

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Demand	18783.38	26702.1	28447.4	28360.85	30023.18	35004.75	1000
Energy Loss	5982.00	10532.0	12962.0	12891.27	15142.00	21308.00	1000

Table X shows that there are 1000 NA observations, on inspection the data shows that from the 14th of September onwards the datasets only contains values for the *Date* and *RSA.Contracted.Forecast* variables. The choice was made to truncate the data such that the final observation in the dataset had the Date and time: $2023/09/13 - 21:00$. The dataset contained no other missing data.

Table 2: The first and last two observations with missing data.

Date	RSA.Contracted.Forecast	RSA.Contracted.Demand
2023-09-14	30836.54	NA
2023-09-14	32437.46	NA
2024-03-31	24694.54	NA
2024-03-31	22910.12	NA

1.3 - Plotting the raw data

We can now plot our data for both the *RSA.Contracted.Demand* and $Total.UCLF.OCLF$ variables using time plots as shown below. MA-720 (Monthly) smoothing was applied to expose the trend within the raw data.

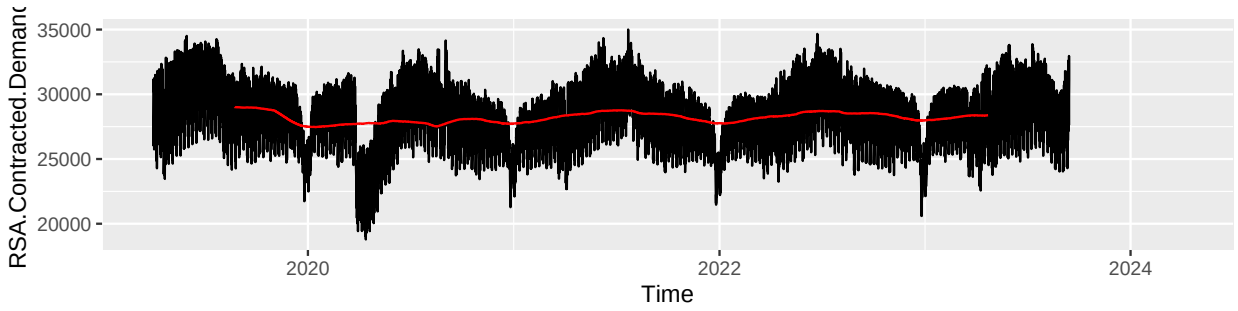


Figure 1: Time plot of the raw demand data.

The figure shows that there is some annual seasonality/cyclicality in the demand data. This can be logically explained with the seasons of the year and their correlation with electricity demand. In summer the demand is lower, then it increases until it peaks in mid-winter when the days are shorter and colder, requiring more electricity for lighting and heating. Every year there is also a sharp decline in December/January, we will investigate this further in Section 1.5. The demand time plot also shows the dramatic effect that Covid had on the demand in 2020 making it an outlier amongst the other 4 years (2019,2021,2022,2023).

The demand time plot has no clear trend and the variance appears to be fairly constant (although quite high) with the exception of the Covid era and the December/Christmas period.

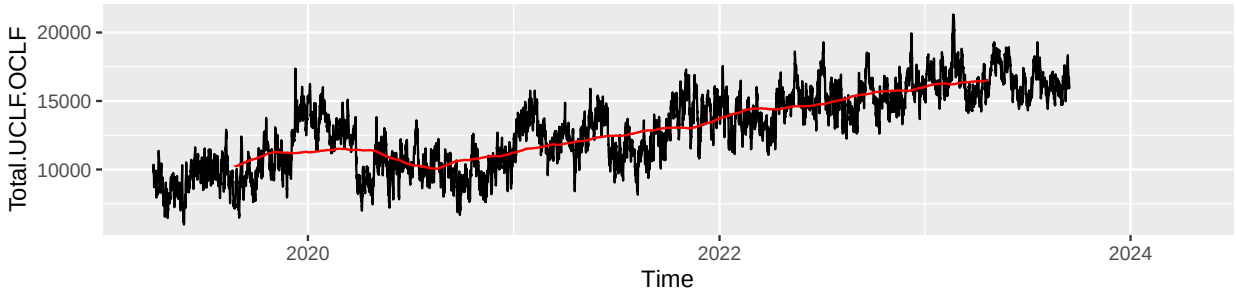


Figure 2: Time plot of the raw energy loss data.

The time plot for the total energy lost due to unplanned outages presents quite a clear positive linear trend but lacks any noticeable seasonality. The variance also appears to be constant for the majority of the time. We notice that there doesn't seem to be much correlation between the demand and the energy lost due to unplanned outages with the exception of the Covid era. The increasing energy loss can likely be attributed to the lack of adequate infrastructure maintenance.

1.4 - Examining major decreases in December

It was mentioned in section 1.4 that there appeared to be a drastic decrease in the demand around December/January every year. The time plot in the Appendix shows that the sharp decrease is due to Christmas and New years celebrations/holidays. During these times businesses will often close or scale down as people take leave. Another contributor is the people that leave the country to go on holiday, however the impact of this might be cancelled out by tourists visiting South Africa during that period.

1.5 - Data Cleaning

Figure X shows how the demand in 2020 differed from the years 2019 and 2021-2023. Events such as the Covid pandemic are irregular occurrences and thus it is sensible to truncate the data such as to not include demand data which was greatly affected by the Covid pandemic. By comparing 2020 to 2021, it appears that by July 2020 the seasonal pattern seemed to have stabilized , therefore a choice was to truncate the data such that it begins in September of 2020 which will give us 4 full years of data. Thus observations in the final cleaned demand data will begin $2020/09/14 - 17:00$ and run until $2023/09/13 - 21:00$.

Much like for the demand data the effect of the Covid pandemic is reflected in the energy loss data. However unlike for demand data, there is no clear annual pattern or predictable trend for energy loss due to unplanned generation outages. This is most likely because it depends on current circumstances such as infrastructure failures, maintenance issues or other events that might affect generation capacity. Thus for forecasting unplanned outages, the data that should be used should be recent.

Therefore the data is truncated to include only a single year and will run from 2022/09/14 to 2023/09/13.

1.6 Aggregating data

(*UNFINISHED*) In order to convert the hourly demand data into daily data such that the peak daily demand can be forecasted, the hourly data will be aggregated by taking the maximum demand of the 5pm-9pm period each day.

1.7 - Decomposition of data

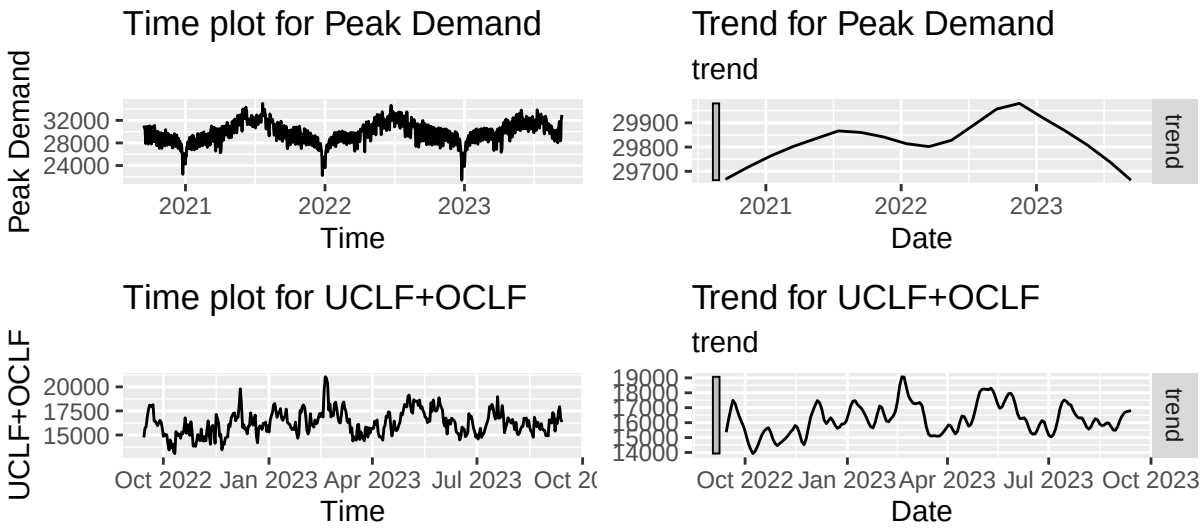


Figure 3: Time plots of cleaned/transformed data with trend plots

The figure above shows that there is some non-linear trend in the peak demand data but no clear trend in the data for energy loss due to unplanned outages.

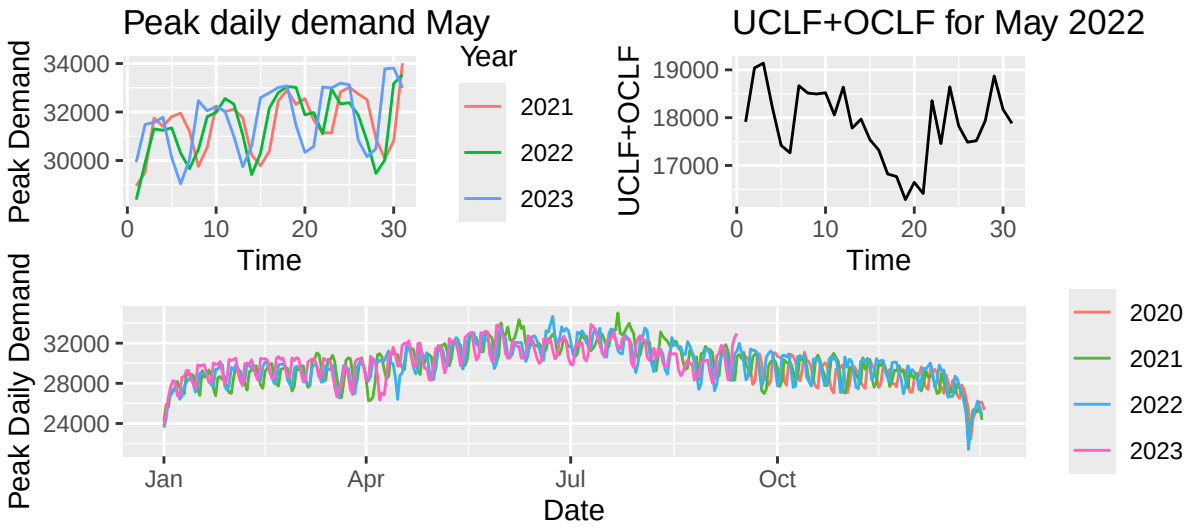


Figure 4: Weekly and annual seasonality of cleaned/transformed data.

The figure shows that there is strong weekly seasonality in the peak daily demand data which is consistent

across 2021-2023. There does not appear to be any weekly seasonality for the energy loss data.

There is clear annual seasonality pattern for the daily peak demand data. Because we are only using a single year of data for the energy loss, it is not sensible to plot the annual seasonality.

(*UNFINISHED*)

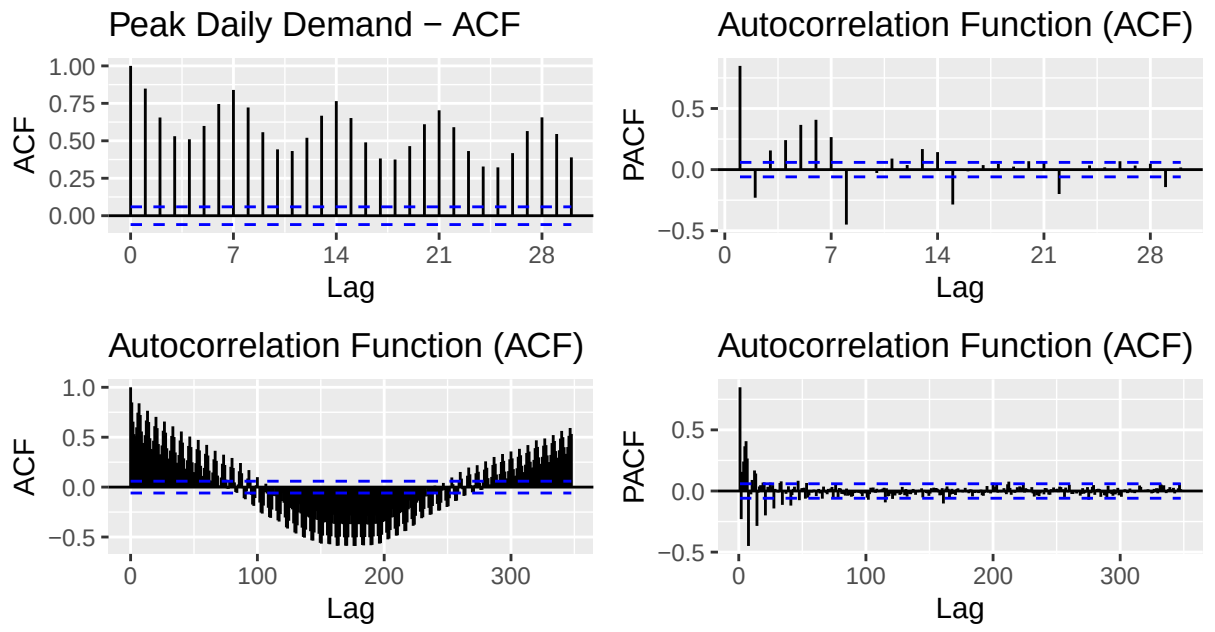


Figure 5: ACF and PACF plots of Peak Daily Demand Data.

Figure X shows significant autocorrelation in the series up to relatively high lags in both the PACF and ACF. There are clear spikes in autocorrelation every 7 lags in both the ACF and PACF which represent the strong weekly seasonality discussed in Section 1.8.2. The figure also suggests that there is some trend in the data through the decaying pattern of the autocorrelation in the ACF, this happens as if there is trend then observations that are close to each other in time are close in size.

The cyclicality in the ACF is representative of the annual seasonality/cyclicality in the peak daily demand data. From the ACF plot as well as the supporting time plots we can infer that the annual period is roughly 365 days. Using Figure X and the figure above we it is clear due due to the presence of trend and strong seasonality that the Peak Demand data is non-stationary.

Section 2 - Peak Daily Demand Model Fitting

For the purpose of the model fitting, forecasting and comparison, the data will be split up into training/test sets. The forecast horizon was set to 92 days (3 months), a decision which was based on the aims of the project which include producing short-term forecasts which will be used in handling load shedding and medium-term forecasts which will be used in grid management decisions. Therefore, the training set will consist of all observations from 2020/09/14 to 2023/06/13 and the test set will be made up of the last 3 months from 2023/06/14 to 2023/09/14.

2.0 Model Preparation

2.0.1 Differencing for Non-seasonal Models

Figure X shows that the peak daily demand data is clearly non-stationary as the series does not fluctuate around a mean of zero and there is a clear presence of seasonality which is supported by the ACF plot in Figure X. Therefore for the non-seasonal models (AR, MA, ARMA and ARIMA) we can take first differences until the data appears stationary.

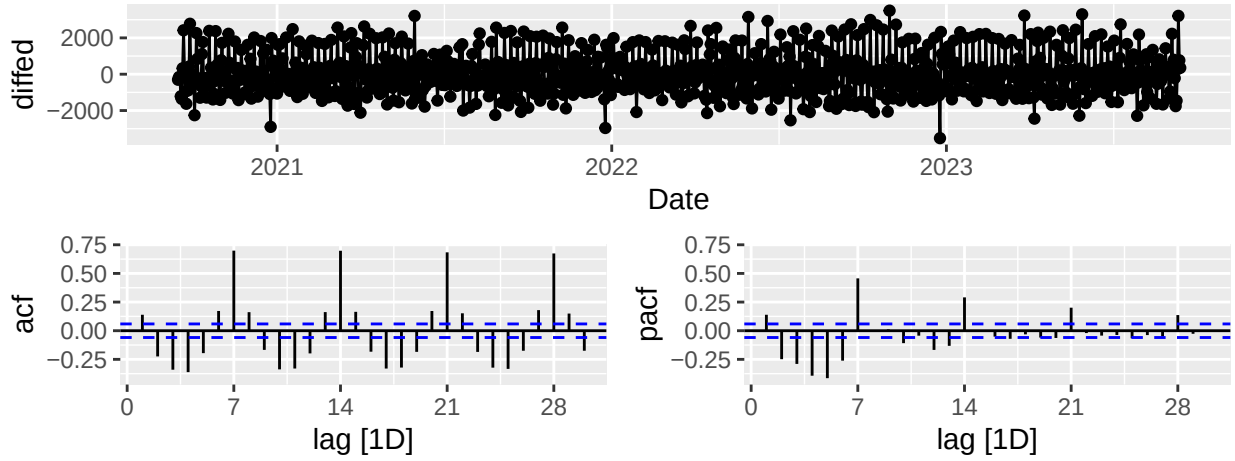


Figure 6: Differenced data and ACF/PACF plots (Non-seasonal).

The figure above shows data that is representative of stationary series as it fluctuates around a zero-mean and displays no heteroskedasticity.

We can then perform an Augmented Dickey-Fuller test, to test for the presence of a unit root in the series. The null hypothesis of the test is as follows H_0 : There is a unit root present in the series, i.e., the series is non-stationary. And the alternative hypothesis is H_1 : There is no unit root present in the series, i.e., the series is stationary.

Table 3: ADF Test Results

Statistic	P.Value	Alternative
-12.26219	0.01	stationary

As shown in the table above, the p-value of the ADF test is $p = 0.01$, this means that there is sufficient evidence to reject the null hypothesis. Thus we can conclude that the series is stationary.

2.0.2 Differencing for Seasonal Models

As mentioned in Section 1.9.1, Figure X shows that the peak daily demand data is clearly non-stationary. From the investigation into the seasonality present in the data in Section 1.8.3, it is evident that there is strong weekly seasonality in the data which suggests the use of a model such as a SARIMA which takes into account the seasonal components of a series. Seasonal differencing ($m = 7$) was applied.

The plot below shows that the data still somewhat exhibits the annual seasonal pattern, therefore a first difference will also be applied to the series. The seasonally and first differenced is shown in the figure below. The series appears to have been stationarized.

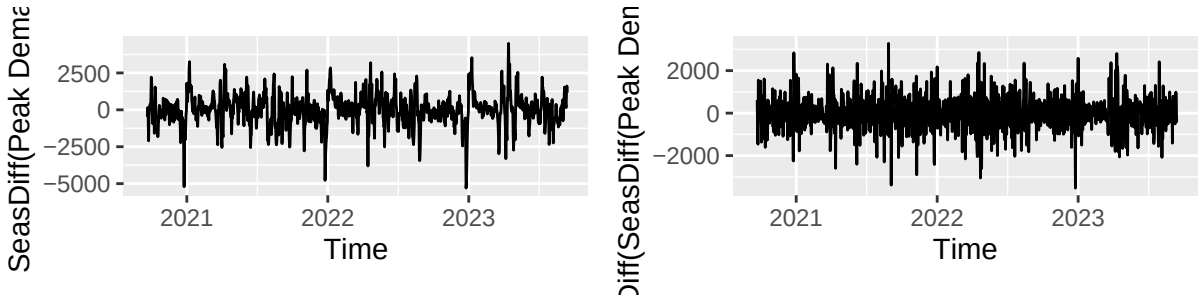


Figure 7: Differenced data (Seasonal)

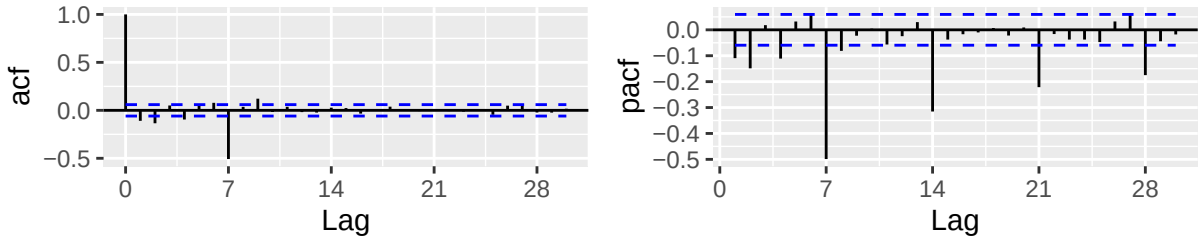


Figure 8: ACF/PACF of Differenced data (Seasonal)

Once again an ADF test can be performed to determine if the differenced series has reached stationarity. H_0 : There is a unit root present in the series, i.e., the series is non-stationary. H_1 : There is no unit root present in the series, i.e., the series is stationary.

Table 4: ADF Test Results

Statistic	P.Value	Alternative
-14.59346	0.01	stationary

The p-value of the ADF test is $p = 0.01$, this means that there is sufficient evidence to reject the null hypothesis. Thus we can conclude that the series is stationary.

2.1 AR/MA/ARMA/ARIMA (Non-seasonal) Models

To begin, simple model autoregressive, moving-average, ARMA and ARIMA models will be developed, these are relatively straight forward to fit however, given the seasonality present in the data as shown in Section 1.8.2, we expect that AR, MA, ARMA and ARIMA models will not fit the data adequately. This is because they do not take the seasonal components of the data into consideration, this will result in poor quality forecasts and residuals which indicate that there is information in the series which has not been captured by the model. For the sake of comprehensiveness and providing a baseline, non-seasonal models are included in the analysis.

Because of the theoretical limitations of the non-seasonal models, the model fitting and selection process was drastically simplified, leaving the focus on the seasonal models/methods. The non-seasonal models were fit using the same simple methodology:

1. Inspect the ACF and PACF plots of the stationarized data in Section 1.9.1.
2. Determine the relevant orders for p, d and q for a candidate model.

3. Fit multiple models by varying the relevant orders such as $p = 5, 6, 7, 8$ for an $AR(p)$ model.
4. Choose the best model by comparing AICc values. (Appendix)

Table 5: Best model specifications chosen using AICc.

AR(8)	MA(7)	ARMA(5,6)	ARIMA(5,1,4)

A residual analysis was conducted on the models listed above. This was comprised of interpreting the plots of the residuals (Appendix) and the results of the Ljung-Box test for autocorrelation of the residuals.

The hypotheses for the Ljung-Box test are as follows:

H_0 : There is no autocorrelation in the residuals for a fixed number of lags (REF).

H_1 : There is autocorrelation present in the residuals at some lag/s l in the residuals.

Table 6: Ljung-Box Test Results

Model	Test Statistic	P Value	AC Present (Y/N)	First Significant AC at Lag
AR(8)	106.2382	0	Y	7
MA(7)	398.1555	0	Y	6
ARMA(5,6)	143.3960	0	Y	2
ARIMA(5,1,4)	272.8049	0	Y	5

The Ljung-Box test results above and the plots in the Appendix confirm the expectation that AR, MA, ARMA and ARIMA models will not be appropriate for modelling the given series as all models display significant autocorrelation in the residuals which indicates that there is information that the models are failing to capture. All of the non-seasonal models showed significant autocorrelation at lag 7 which indicates that the strong weekly seasonality was not captured by the models and was still present in the series. Hence the forecast plots are not closely analyzed as the non-seasonal models are inadequate for modelling this series, however, the plots can be found in the Appendix.

2.2 SARIMA Models

Section 2.1 presented the shortcomings of applying non-seasonal models to the series as they failed to provide a good fit to the series and produced inaccurate forecasts. The focus is now on $SARIMA(p, d, q)(P, D, Q)[M]$ seasonal models. By adding seasonal components $(P, D, Q)[M]$ to the already existing $ARIMA(p, d, q)$ model framework, we aim to capture the seasonality present in the data which should enable the development of more robust/accurate forecasts. The orders of p, d, q, P, D, Q were selected using the ACF and PACF plots in Section 2.0.2.

2.2.1 SARIMA(4,1,4)(2,1,1)[7]

The spikes at lag 7 and lag 14 in the PACF plot in Figure 8 suggest that if we choose $m = 7$ then we start with $P = 2$ for the seasonal AR component as the initial candidate model. The ACF plot in Figure 8 suggests $Q = 1$ for the seasonal MA component due to the significant spike at lag 7. The significant spikes at lag 4 in the ACF/PACF of Figure 8 suggest that $p = 4$ and $q = 4$ be used as the non-seasonal AR and MA components.

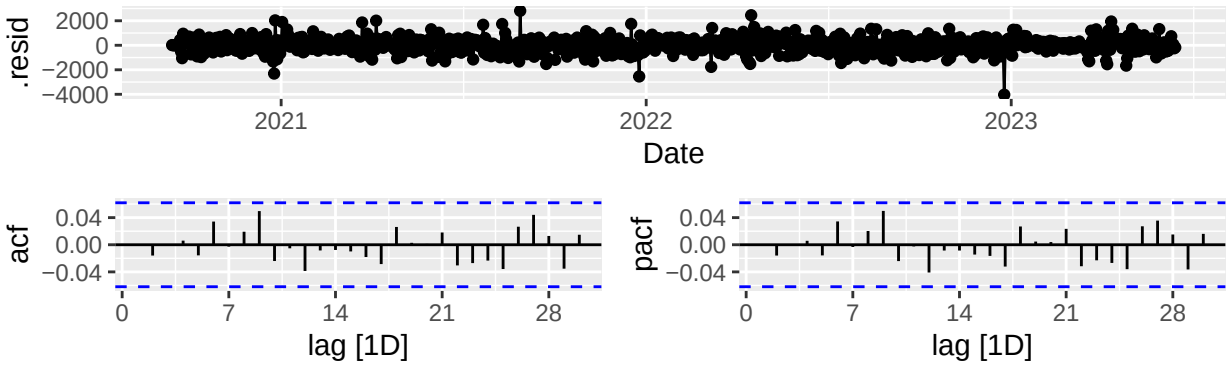
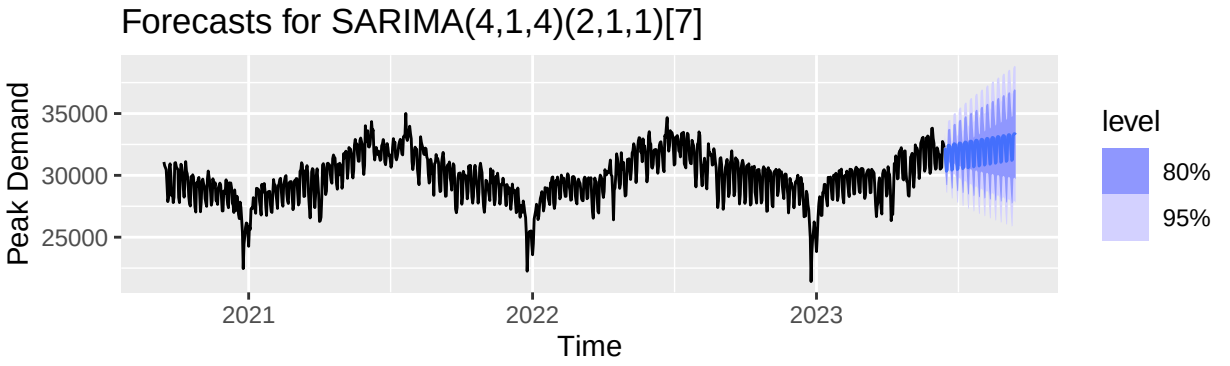


Figure 9: SARIMA(4,1,4)(2,1,1)[7] residual plot.

Table 7: Ljung-Box Test Results

LB Statistic	P-value	Max Lag
6.87	0.009	365

The residuals of the $SARIMA(4,1,4)(2,1,1)[7]$ model suggest that the model is a good fit to the data as there does not appear to be any significant autocorrelation in the ACF/PACF plots. This is supported by the results of the Ljung-Box test which show $p-value >> 0.05$ indicating that we fail to reject H_0, therefore we can conclude there is no significant autocorrelation in the residuals.



Whilst the $SARIMA(4,1,4)(2,1,1)[7]$ model incorporates seasonal differencing as well as autoregressive and moving average components the forecasts suggest that the model is not a good fit. While the model appears to capture the weekly seasonality, it falls short in capturing the annual seasonality that is present in the data, thus the forecasts fail to align with what we'd expect of the forecasted values given the historical data. This suggests the need for an additional component which will model the annual pattern.

2.2.2 SARIMA(3,0,3)(0,1,0)[365]

A $SARIMA(3,0,3)(0,1,0)[365]$ model is then fit to the data as an alternative approach in order to try and capture the annual seasonality that the $SARIMA(4,1,4)(2,1,1)[7]$ model failed to capture, this is done by setting the seasonal period $m = 365$. It is noted that this will very likely affect the model's ability to capture the weekly seasonality.

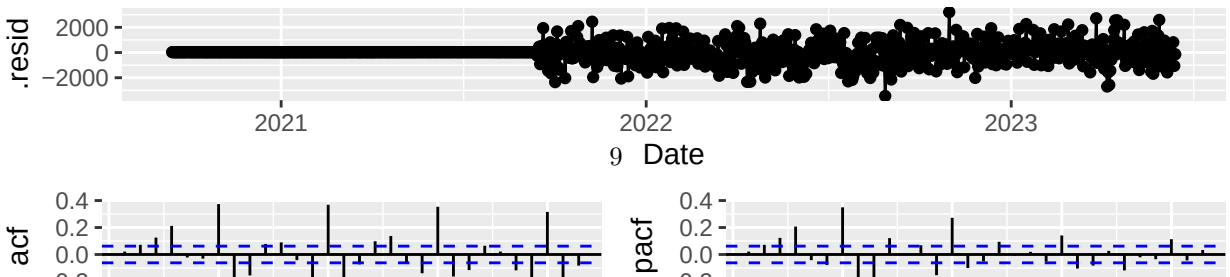
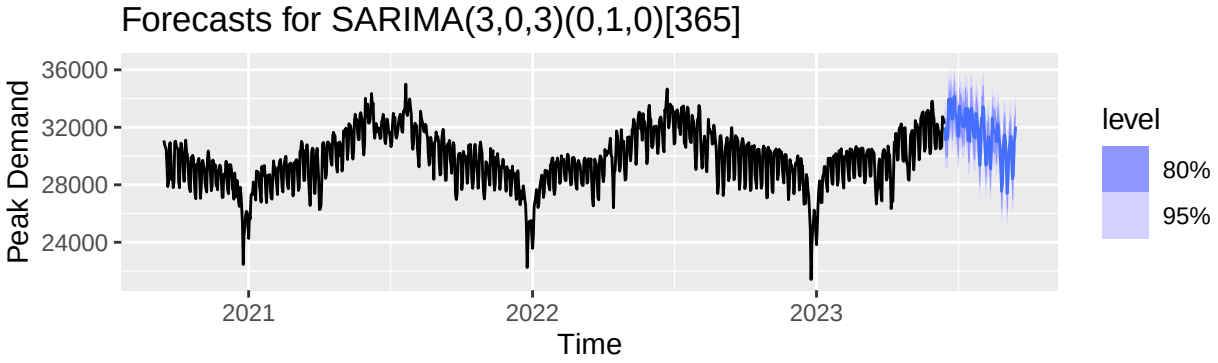


Table 8: Ljung-Box Test Results

LB Statistic	P-value	Max Lag
933.398	0	365

The significant autocorrelation at lags 2,3,4,7 in the ACF and PACF plots of the residuals suggest that the model is an inadequate fit to the data and that there is information in the series that the model has failed to capture. It is clear that by neglecting the weekly seasonality by choosing $m = 365$ the model has failed to capture the weekly seasonality which can be seen by the spikes in autocorrelation in the ACF and PACF at lags that are a multiple of 7.



The forecasts for this represent the annual pattern in the series fairly well as they don't just fluctuate around a constant mean but seem to follow the patterns seen in the historical data, however, the strong autocorrelation present in the residuals at lag 7 make it clear that this model is not appropriate for the series. So whilst the forecasts appear more promising than the $SARIMA(4,1,4)(2,1,1)[7]$ model, further model fitting and exploration is required to find a model that can account for multiple/complex seasonalities.

2.2.3 SARIMA(4,1,4)(2,1,1)[7]+Fourier(365)

A possible solution to handling a series with complex/multiple seasonality is fitting a dynamic harmonic regression model with an ARIMA error structure (*REF*). Put simply, Fourier terms are added to a SARIMA model to account for some one of more of the seasonalities in the data.Therefore, the $SARIMA(4,1,4)(2,1,2)[7] + Fourier(p = 365, k = 7)$ model is proposed (where p is the period of the Fourier term and k is the order of fourier terms). This specification uses the Fourier terms to try capture the annual seasonality in the data as the period of the annual seasonality is estimated to be 365 days, k is chosen to be 7 as this allows for a more flexible fit which will help avoid underfitting.

Table 9: Ljung-Box Test Results

LB Statistic	P-value	Max Lag
12.798	0.119	365

The residuals for the $SARIMA(4,1,4)(2,1,1)[7] + Fourier(p = 365, k = 7)$ model suggest that the model is a good fit to the data as there is no significant autocorrelation shown in the ACF or PACF plots. This is supported by the results of the Ljung-Box test for which the $p - value = 0.119$ which as per the hypotheses stated in Section 2.1 means that there is insignificant evidence to reject H_0, therefore, we can conclude there is insignificant autocorrelation in the residuals and that the model is capturing the available information in the series well.

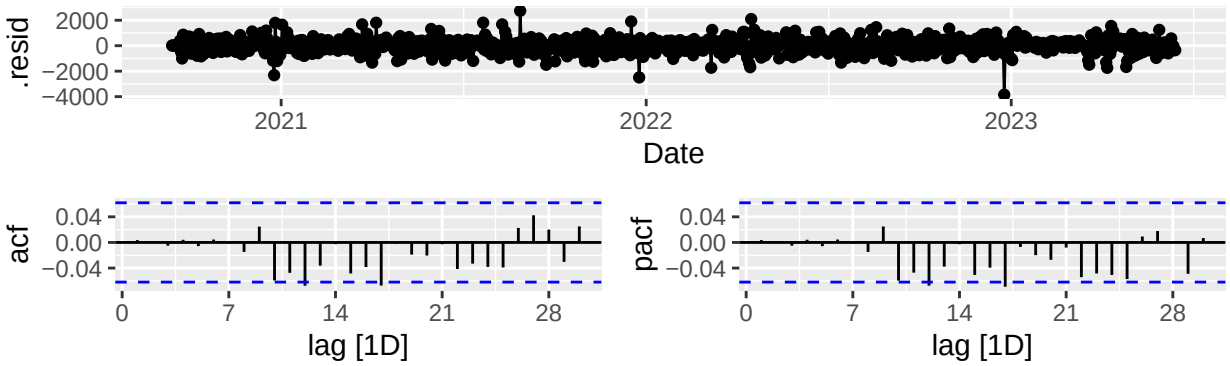
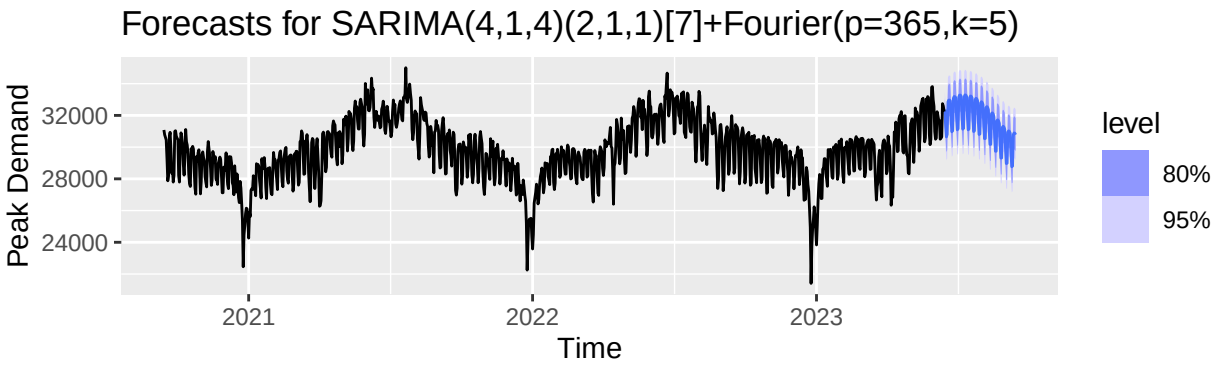


Figure 11: SARIMA(4,1,4)(2,1,1)[7]+Fourier(365) residual plot.

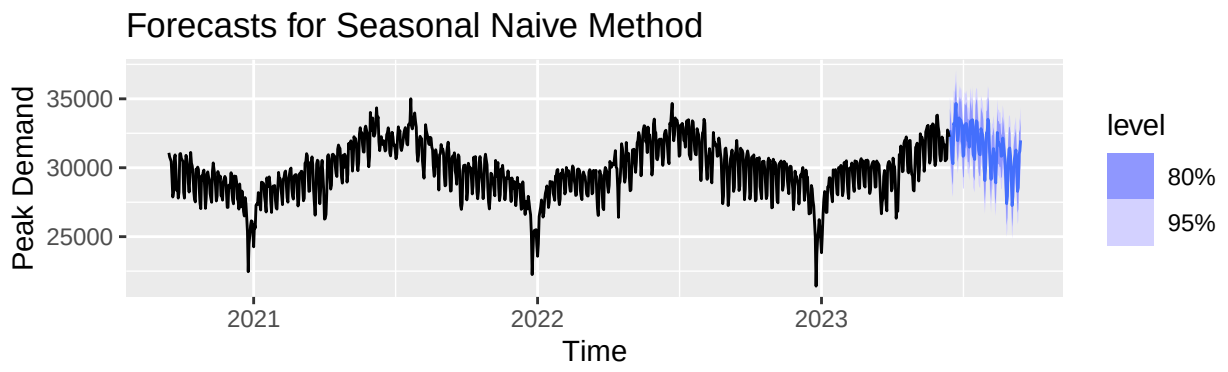


The forecasts support the claim that the model is a good fit to the data as the model appears to capture both the weekly and annual seasonality and produce forecasts similar to what we'd expect given the historical data. The addition of the Fourier term seems to have helped capture the annual seasonal component, allowing the SARIMA component to focus on the weekly seasonality. Although it must be noted that the forecast do appear quite smooth which might indicate that the model is underfitting.

2.3 Benchmark Model - Seasonal Naive

Models that we build for forecasting should always be compared to a simple benchmark model as we need to be able to justify the additional complexity introduced when using a model such as an ARIMA or SARIMA model. The more complex model should significantly outperform and present improvements over the simpler forecasting model/method.

In the context of this problem, the seasonal naive forecasting method presents itself as the most appropriate benchmark method for forecasting the series due to the presence of strong seasonality in the series.



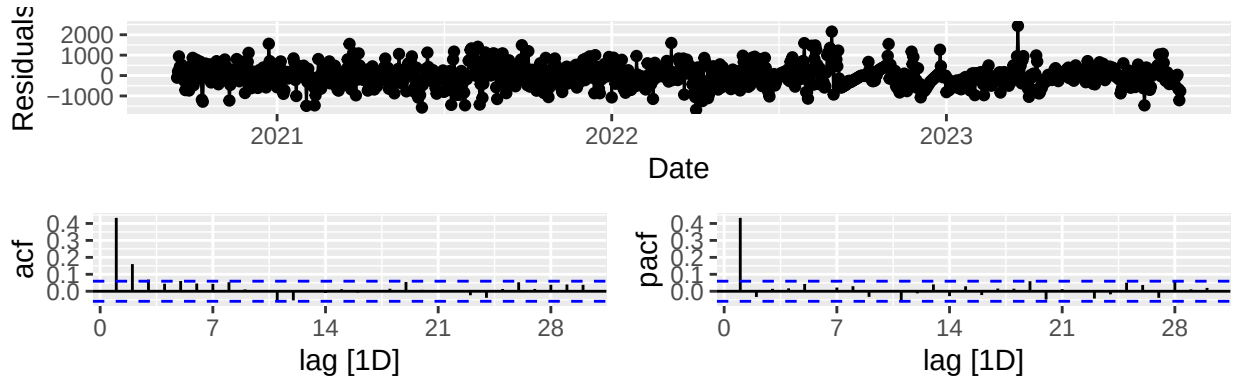
The forecasts produced by the Seasonal Naive model appear to be accurate and represent both the weekly and annual seasonalities present in the series. The accuracy of the forecasts will be compared to the other models in Section 2.7 in which the quality of the Seasonal Naive forecasts will then be determined.

2.4 ARCH

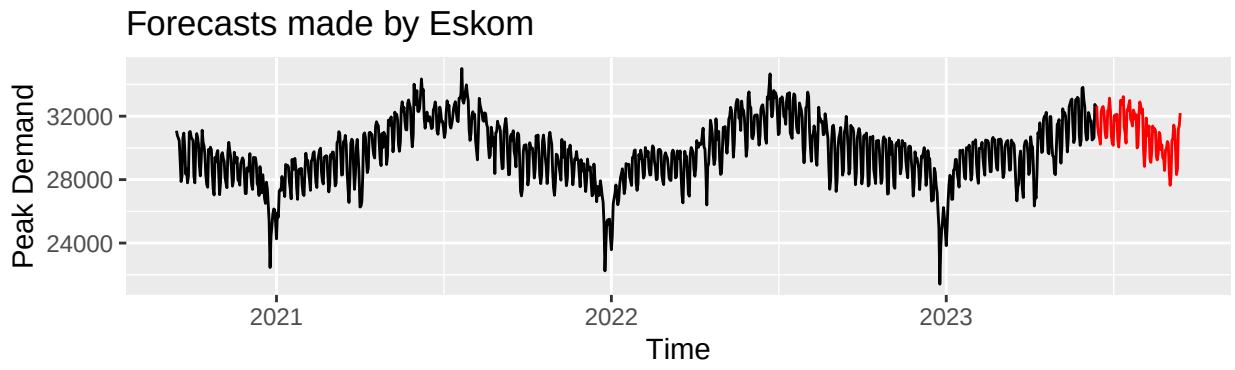
2.5 GARCH

2.6 Eskom's Forecasts

The residuals for the forecasts made by Eskom appear to be stationary, however, they show that there is strong autocorrelation at lag 1 which suggests that the model that they fit/method they use to produce the forecasts has failed to capture all of the information available in the data and that observations 1 time period (day in this case) apart are highly correlated. We cannot perform a Ljung-Box test as the model which Eskom used to produce these forecasts is not specified, however, the presence of autocorrelation in the residuals is clear.



The forecasts made by Eskom are in line with what we'd expect from the series.



2.7 Comparing Forecast Accuracies

Table 10: Forecasting accuracies of models/methods

	ME	RMSE	MAE	MPE	MAPE
AR(8)	-708.111	1490.474	1178.205	-2.464	3.902
MA(7)	-511.157	1428.732	1163.137	-1.837	3.831
ARMA(5,6)	-488.536	1430.187	1123.635	-1.716	3.687
ARIMA(5,1,4)	-383.299	1393.841	1114.426	-1.381	3.647
SARIMA(4,1,4)(2,1,1)[7]	-1039.273	1783.741	1376.094	-3.510	4.549
SARIMA(3,0,3)(0,1,0)[365]	-444.946	1061.571	865.834	-1.461	2.804
SARIMA(4,1,4)(2,1,1)[7]+Fourier(365)	-575.307	1215.011	966.023	-1.930	3.147
Seasonal Naive	-364.706	1019.768	857.335	-1.188	2.769
Eskom's Forecasts	37.207	435.362	332.929	0.105	1.074

The forecasted values for the peak demand of electricity were then used to calculate the accuracy of the forecasts by comparing them to the observed demand values in the test set. The table above shows various accuracy metrics for the forecasts produced using AR, MA, ARMA, ARIMA and SARIMA models as well as the Seasonal Naive forecasting method and the forecasts provided by Eskom.

2.8 Discussion and selection of forecasting models/methods.

The findings in Section 2.7 and 2.1 allow for the non-seasonal models to be ruled out of contention, due to poor model fit and forecast quality/accuracy.

The residuals for the $SARIMA(4,1,4)(2,1,1)[7]$ showed good model fit based on the residual analysis in Section 2.2.1, however, the forecast accuracy was lacking as the model struggled to capture the annual seasonality.

The $SARIMA(3,0,3)(0,1,0)[365]$ model had the best forecast accuracy on the test set,however, there was significant autocorrelation present in the residuals as found in Section 2.2.2 make it unreliable as there is clearly information that was not captured by the model. The performance on a single test set does not determine that the model generalizes well to all cases.

The most promising model was the $SARIMA(4,1,4)(2,1,1)[7]+Fourier(365)$ model. The residual analysis in Section 2.2.3 suggested that the model was a good fit and the forecasts aligned well with what we'd expect to see based on the historical data from the series. The forecast accuracy was also relatively good and could be improved through further tweaking of the model parameters.

The Seasonal Naive forecasting method performed very well on the test set, however, the simplicity of this method could result in poor generalization to other test sets and more rigorous testing such as Cross-Validation would be necessary to determine the general performance of this method.

Eskom's forecasts showed excellent accuracy, however, the presence of significant autocorrelation at lag 1 as discovered in the residual analysis in 2.6, suggests that whatever model/method was used to produce these forecasts failed to capture all of the available information. This might prove to be a problem in the long term for providing accurate and robust forecasts.

To produce robust forecasts, the $SARIMA(4,1,4)(2,1,1)[7] + Fourier(365)$ model should be fine-tuned to improve its forecast accuracy. If it turns out that the seasonal naive method consistently outperforms the optimized $SARIMA(4,1,4)(2,1,1)[7] + Fourier(365)$ model, even after rigorous testing, then the seasonal naive method could be preffered due to its simplicity, interpretability and it's relatively low computational complexity.

Conclusion

*Could investigate if water supply correlates with unplanned outages