

Business Statistics and Modelling

Gas turbine and Time Series data

Baldip Singh | Quarter 1 – 2021

Contents

[Modelling 2](#_Toc69423208)

[Abstract 2](#_Toc69423209)

[Data Preparation 2](#_Toc69423210)

[The Data 3](#_Toc69423211)

[Data Analysis 4](#_Toc69423212)

[Model Analysis 6](#_Toc69423213)

[Testing the Model 7](#_Toc69423214)

[Conclusion 8](#_Toc69423215)

[Time Series 9](#_Toc69423216)

[Abstract 9](#_Toc69423217)

[The Data 9](#_Toc69423218)

[Data Preparation 9](#_Toc69423219)

[Trend Analysis 10](#_Toc69423220)

[Seasonality Analysis 11](#_Toc69423221)

[Forecasting 12](#_Toc69423222)

[Conclusion 14](#_Toc69423223)

[Appendices 15](#_Toc69423224)

[Bibliography 15](#_Toc69423225)

[R-Code-Modelling 17](#_Toc69423226)

[R-Code Timeseries 18](#_Toc69423227)

# Modelling

## Abstract

Gas Holes© is a medium sized drilling company which use turbine powered drilling systems as a part of their boring process. A new tunnelling job requires the drills to be operating at higher RPM’s than usual. The previous drill set up required 130mwh of power supply whereas the new set up is estimated to require 160mwh.

Operators have noticed the gas turbine system is not providing adequate power to the drills to operate at the required levels. Staff do not know how to control power output from the turbines accurately, however, have a collection of sensory readings from the gas turbines from the past five years.

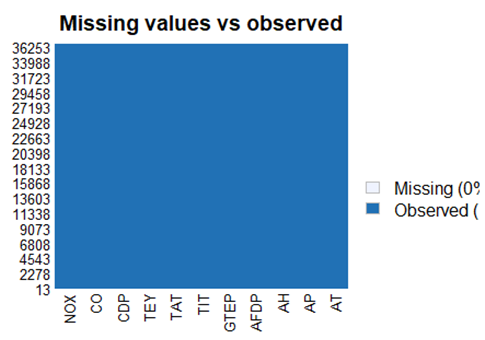
Gas Holes© needs you to investigate 1. which of the turbine readings is the most statistically useful attribute(s) for controlling the energy yield? 2. Propose a model to predict future energy returns and 3. Suggest a simple solution to controlling energy yield to meet new power demand.

## Data Preparation

The data is supplied in five .csv files, each having over 7,000 observations. Each file corresponds to a particular year’s data readings, ranging from 2011-2015. I intend on creating a model based on regression analysis, to predict the Total Energy Yield – TEY.

We must create two sets of data; one set of data will be the training set the second a test set. The accuracy of the model will be tested by using the model created from the training set and applying it to the test set (data that the model has not yet seen). This is to measure the accuracy of the model and will subsequently tell us if the model should be practically applied.

All years of data are merged and at random each observation will be allocated, to either the test or training data set. The split selected is 3 : 1, 75% of observations will be randomly assigned to the training set.

The data did not have any missing/NA values and was complete, shown in the missingness map below.

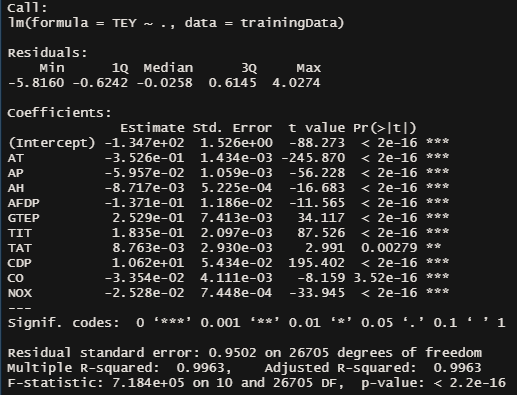
## The Data

The data consists of eleven different sensor measurements from a gas turbine. A gas turbine in short is a piece of machinery that creates energy through the burning of compressed air and fuel.

These different sensory measurements can be classified into three different categories Environmental (atmospheric readings), Mechanical (measurements taken from the machinery itself) and By-Product (the outcome of the process). Shown in the table below.

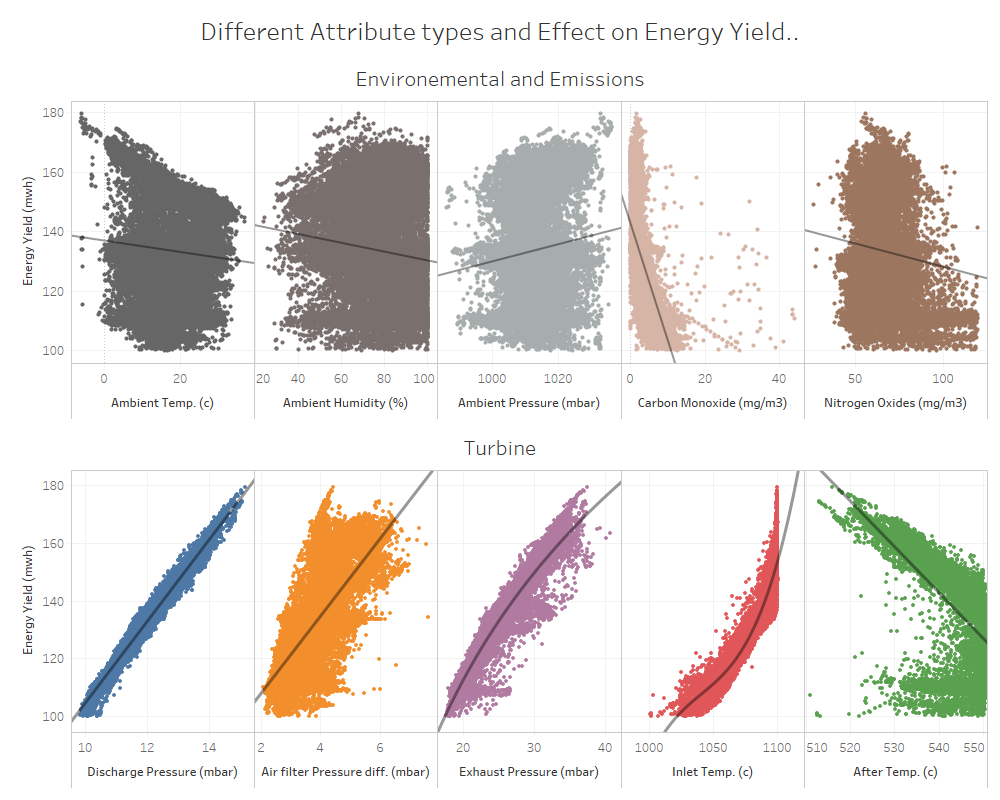
|  |  |  |
| --- | --- | --- |
| Environment | Mechanical | By-Product |
| Ambient temperature (AT)  Units: “C” | Air filter difference pressure (AFDP)  Units: “mbar” | Turbine energy yield (TEY)  Units: “mwh” |
| Ambient pressure (AP)  Units: “mbar” | Gas turbine exhaust pressure (GTEP)  Units: “mbar” | Carbon monoxide emission (CO)  Units: “mg/m3” |
| Ambient humidity (AH)  Units: “%” | Turbine inlet temperature (TIT)  Units: “C” | Nitrogen oxides emission(NOx)  Units: “mg/m3” |
|  | Turbine after temperature (TAT)  Units: “C” |  |
|  | Compressor discharge pressure (CDP)  Units: “mbar” |  |

## Data Analysis

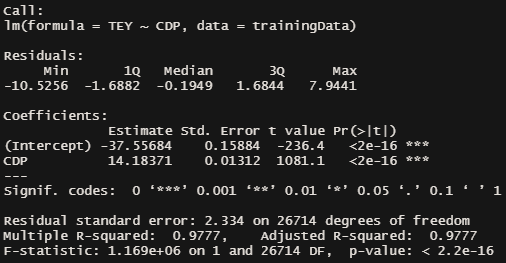


Initial regression output showed all variables to be statistically significant in modelling energy yield. This however is costly and requires too much time and resources for our business requirement and we require a much simpler method to calculate/predict the future total energy yield.

To narrow down the number of attributes contributing to the energy output the next step involved creating plots of each attribute against total energy yield in the hope of indicating each attributes contribution to total energy yield. Illustrated on the next page.



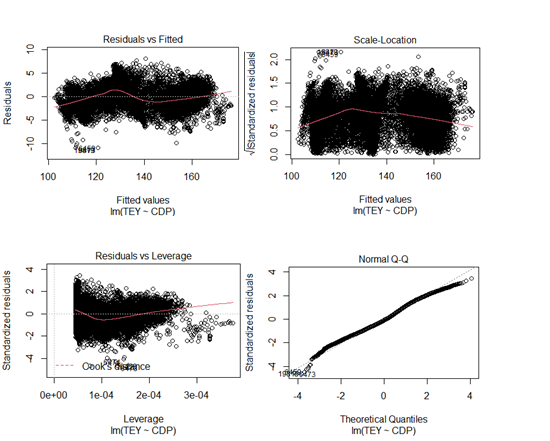
Plotting environmental and emission type attributes showed none to minimal statistical relationship in predicting energy yield. However, ALL turbine readings show some to strong relationships with the discharge pressure (in blue, bottom left) illustrating the strongest relationship, displayed by how tightly the data follows the trend line.

Creating a model based on CDP – compressor discharge pressure was the single attribute with the highest R-Squared value, 0.9777.

Compared to the R-Squared value of having all variables (0.9963) modelling on the CDP alone is a powerful predictor of the total energy yield very nearly as powerful as having all attributes.

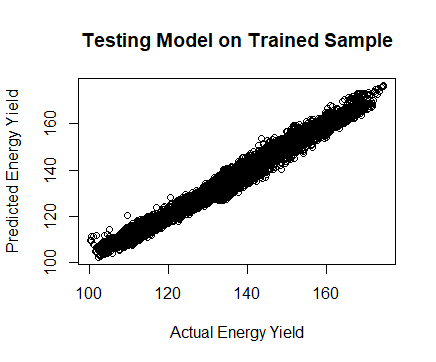
Therefore, the final model to predict energy yield will be based on CDP alone.

## Model Analysis



Model plots show all regression modelling assumptions are fulfilled, and the modelling method used is appropriate.

1. Residual vs Fitted: No non-linear relationships visible.
2. Scale-Location: Residuals are generally spread evenly; line is horizontal enough.
3. Residuals vs Leverage: Observations are independent of each other, although some outliers are present the model is good enough.
4. Normal Quantile-Quantile plot: Values follow the dashed line, negligible deviations from straight line.

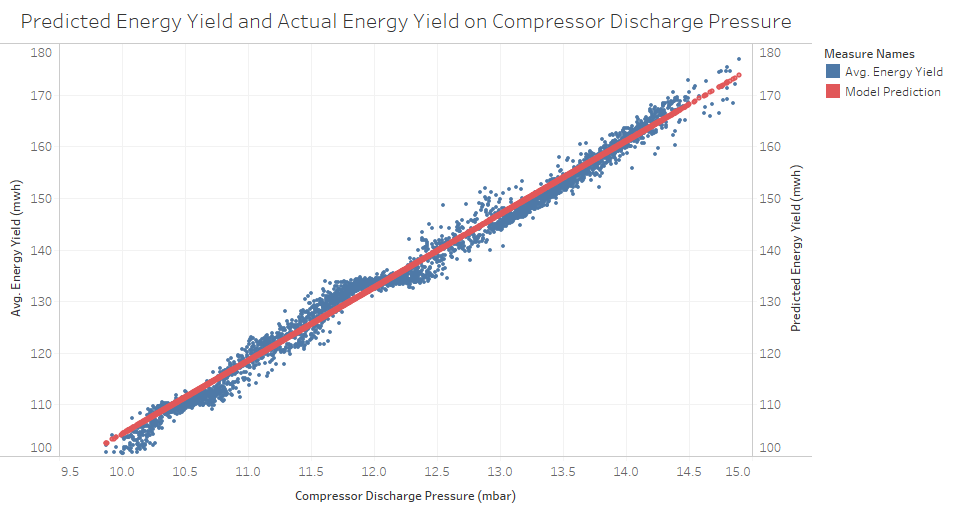


Above plot shows the model is good at predicting energy yield within the training sample and is ready to be applied to test set.

## Testing the Model

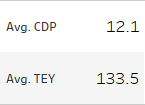
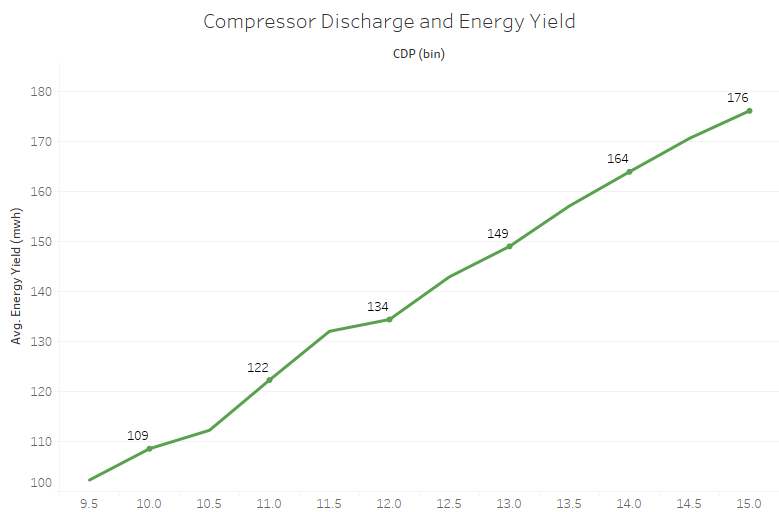
Below we see the average energy yields for different readings of discharge pressure. We have the model predictions against the test sample. We can clearly see the prediction is accurate for the most part, we do see slight deviation between 11.5-12mbar discharge pressure however for the most part the model is predicting well enough.

The model was just as good at predicting the energy yield in the test set as it was in the training set. This is supported by the similar Root Mean Squared Error (RMSE) values, for the training data set we had a RMSE of 2.334 vs the test data set RMSE of 2.332.



## Conclusion

Compressor discharge pressure (mbar)



Compressor discharge pressure (mbar)

Compressor discharge pressure is the best attribute to predict future energy yields with great accuracy. Shown above, for each point we increase the compressor discharge pressure we get a ~10% increase in energy yield alternatively each point we decrease the discharge pressure we lose ~10% energy yield.

To meet new energy demands I recommend the gas turbine compressor discharge pressure be kept at 14mbar, meeting the 160mwh power output required by the new drill set up.

Tuning CDP alone can help increase energy output from the gas turbines, when we find our drills are being under powered. The model below can accurately forecast future energy yields. On average we will see energy yield predictions to be 2-3 mwh above or below the actual energy yield.



# Time Series

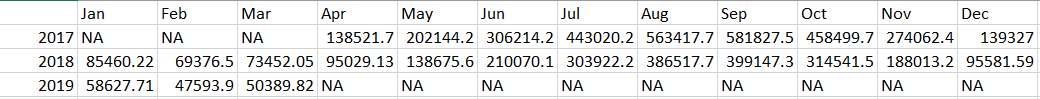
## Abstract

IceBreaker© is a large producer and retailer of thermal wear with multiple stores across New Zealand, specializing in extra thick thermals.

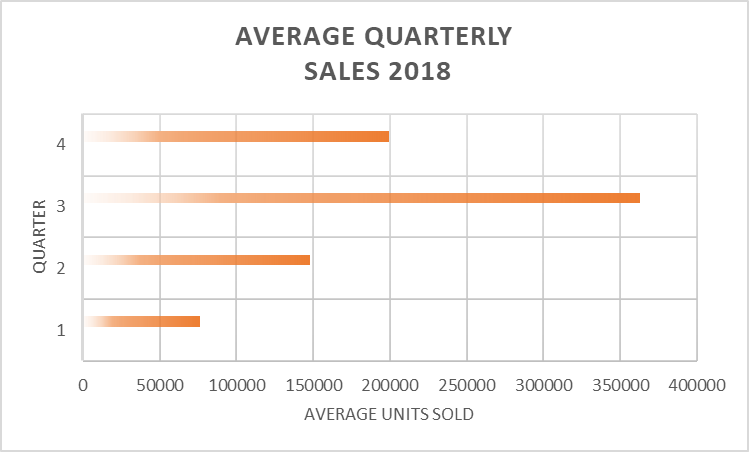
IceBreaker©’s needs you to forecast the next three months’ worth of extra thick thermal sales so the required production material can be on hand. Also identify and provide a possible explanation to any observable trends in units sold, and comment on whether this product should be continued.

## The Data

We are supplied univariate timeseries data which ranges from April 2017 to now (March 2019) . The data consists of units sold each month in table format.



Only 2018 had a full years’ worth of data showing an average of 196648.9 units sold per month. The Quarter with the highest number of units sold was quarter three and the quarter with the lowest being the first.

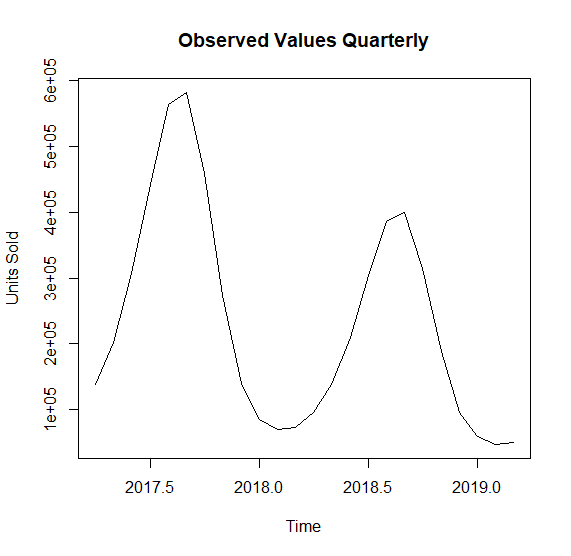
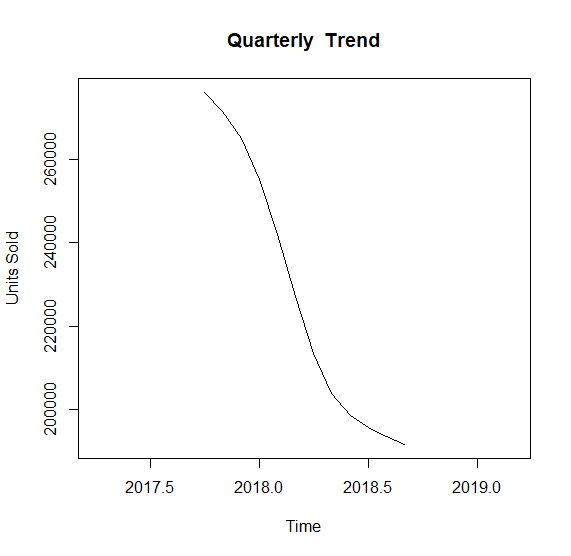


## Data Preparation

For the timeseries function to interpret this data, the data must be of type numeric vector only. Only the single variable, units sold was kept without years, and months and further converted into vertical format. This was accepted as “readable” by the timeseries function. The years and months were later assigned using the arguments available in the timeseries function.

Other than the change mentioned above, the data provided was exactly what was required to conduct the timeseries analysis and no further changes were required.

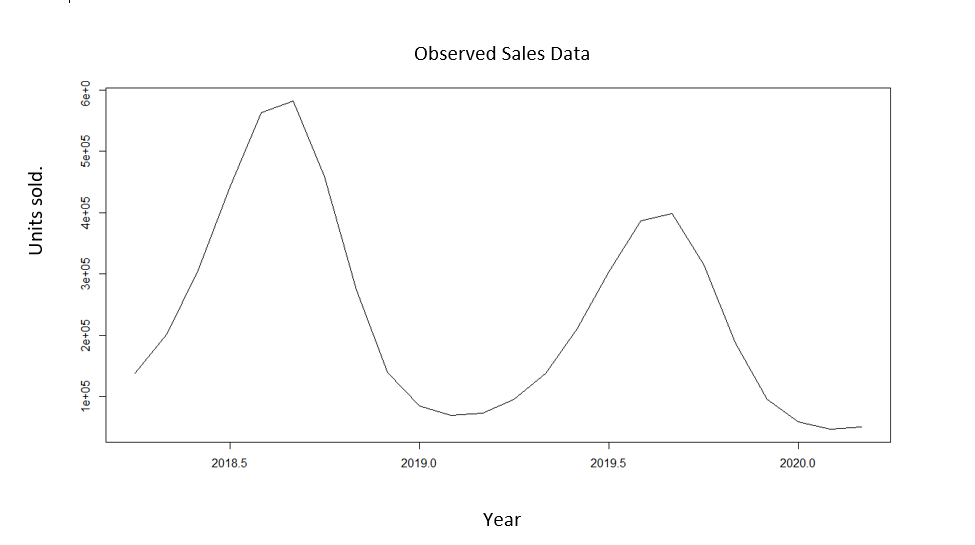
## Trend Analysis

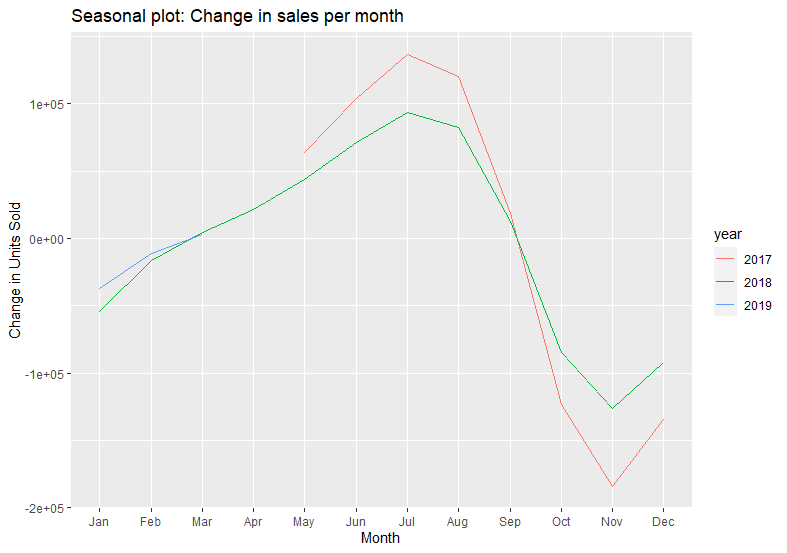
 

Preliminary decomposition of the observed values shows clear modulation of units sold over time (Observed Values Quarterly plot), indicating the time of year and attributes associated with the period affect the number units of extra thick thermals sold.

We also view a negative trend exists for the sales meaning the units sold at any point annually are decreasing as time passes (Quarterly Trend plot).

## Seasonality Analysis



We see strong seasonality trends present, all though the number of units being sold may differ, the peaks and troughs in sales occur around the same time each year (shown in blue dashed lines above). We see that each year we will see the highest number of sales(shown in red) around the beginning of the third quarter, and the lowest number of units sold during the first quarter. 

On the previous page is a monthly breakdown of the difference in sales compared to the previous months sales, this way we get a better idea of the demand for the thermals. We see that we always have a sharp drop in sales from the month of August through till November. From November we see sales begin to gradually grow.

It can be noted we see clear seasonal patterns for the months of May through till August. May being the final month of Autumn temperatures begin to drop, June through August being winter. With winter being the coldest time of the year, it is not surprising to see the sales of extra thick thermals to be increasing during these times. July being the coldest month of the year shows a clear correlation with extra thick thermal sales in that we see the greatest spike.

To note although the rate of sales for September are dropping rapidly, September is found to be the month with the most units sold.

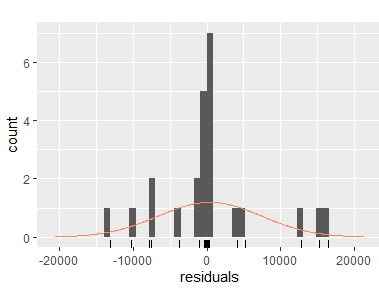
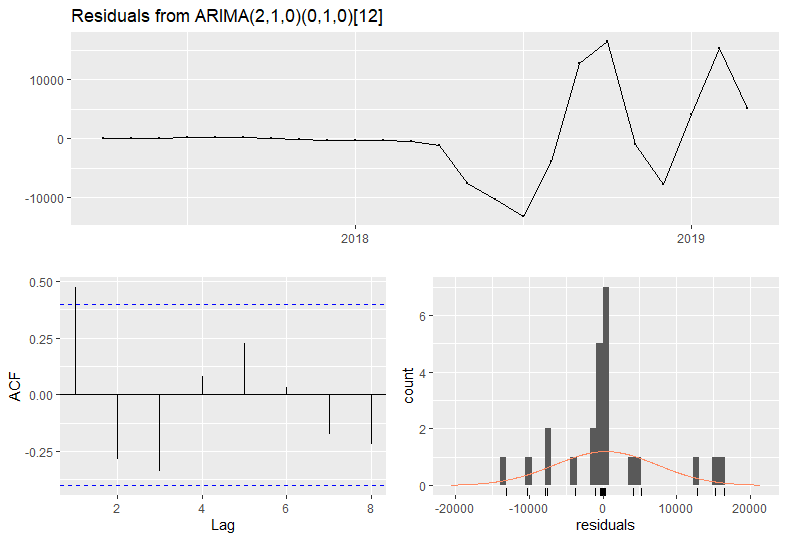
Both seasonality and trends exist within the sales of extra thick thermals. Therefore, forecasting future sales is viable.

## Forecasting

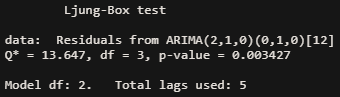
Different models were implemented (naïve and ETS) before finalizing with the ARIMA model to be used for forecasting future sales numbers as it was deemed the best fit.

Briefly, for an ARIMA model to be deemed fit for use, two assumptions must be fulfilled. There must only be a single variable (units sold in this case) and second the residuals of the output must be randomly distributed proving no external variables are warping the model.

Using auto.arima function in R, we get the following output when applying the best model calculated by ARIMA.

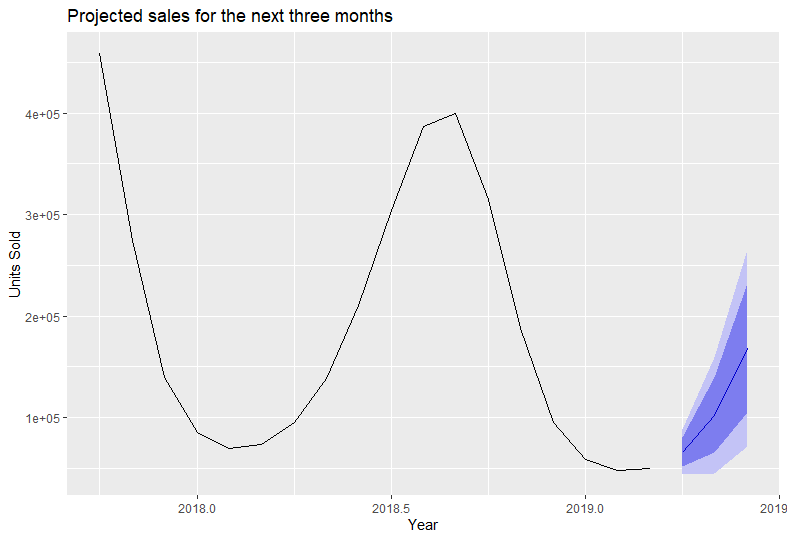


The autocorrelation function (above-left) shows the lag is white noise i.e., the residuals do not show significant autocorrelations. Again, the residuals are normally distributed(above-right), solidifying the residuals are random.



The Ljung-box test(above) shows a p-value of less than 0.05 therefore we cannot reject the null hypothesis and can say the residuals of the ARIMA model are random.

All the above plots and output prove the residuals to be random and the model CAN be used to project the next three months’ worth of data.



Above is the predicted number of units sold for the next three months. The darker blue Projecting dark blue band being 80% confidence range vs light blue 95% confidence.

**Point Forecast Lo 80 Hi 80 Lo 95 Hi 95**

**Apr 2020** 80450.66 70051.98 90849.35 64547.25 96354.08

**May 2020** 113285.47 95529.87 131041.07 86130.62 140440.31

**Jun 2020** 172663.75 141555.09 203772.41 125087.16 220240.34

Above are the number of units sold with relevant confidence intervals.

## Conclusion

With the limited amount of data available we see obvious seasonal trends; we should avoid over stocking on production material as we observe sharp drop in sales for the months of August through till November but should have enough stock available for September sales.

Sales will begin to pick up from November however we see the greatest jump in sales for the winter months May, June, and July. Have production material ready for sale during these months. This is critical.

Below are the projected units required for the next three months along with a 95 % accuracy range.

|  |  |  |
| --- | --- | --- |
| Month | Projected Units Required | 95% Accuracy Range |
| April | 66,123 | 44,209 to 88,036 |
| May | 102,605 | 45,599 to 159,612 |
| June | 168,646 | 71,955 to 265,336 |

During the last twenty years we see the average temperatures for winter are gradually increasing. 2017 had above average temperatures on record, 2018 the 6th warmest winter on record and 2019 is projected to be the 7th warmest temperature on record. Not only this but 2020 is(was) expected to be THE warmest winter on record. This may be a contributing factor to why our extra thick thermals are not being purchased as often and show an overall negative trend.

Although we do see sales trending negatively, we are still managing to sell hundred thousand of units during our good months. We should continue the sale of this product however purchase conservatively and review sales statistics quarterly. In my opinion we should purchase for units towards the bottom end of the 95% accuracy range to minimize risk.

# Appendices

## Bibliography

Alam, M. (2019). *Time series forecasting: from naive to ARIMA and beyond*. Retrieved from towards Data Science: https://towardsdatascience.com/time-series-forecasting-from-naive-to-arima-and-beyond-ef133c485f94

Brownlee, J. (2020). *How to Identify and Remove Seasonality from Time Series Data with Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/time-series-seasonality-with-python/

Brownlee, J. (n.d.). *White Noise Time Series with Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/white-noise-time-series-python/

Chaterjee, S. (2018 ). *Time Series Analysis Using ARIMA Model In R*. Retrieved from Programming in R: https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/

Coghlan, A. (2010). *Using R for Time Series Analysis*. Retrieved from Time Series Documentation: https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html

Duke. (2016). *Linear Regression models*. Retrieved from Duke Education: http://people.duke.edu/~rnau/testing.htm#:~:text=If%20any%20of%20these%20assumptions,model%20may%20be%20(at%20best)

Editor, M. B. (2012). *Why You Need to Check Your Residual Plots for Regression Analysis: Or, To Err is Human, To Err Randomly is Statistically Divine*. Retrieved from Minitab: https://blog.minitab.com/en/adventures-in-statistics-2/why-you-need-to-check-your-residual-plots-for-regression-analysis#:~:text=The%20non%2Drandom%20pattern%20in,explaining%20all%20that%20is%20possible.

Guy, S. D. (2018). *Introduction to time series in R*. Retrieved from Youtube: https://www.youtube.com/watch?v=VrJYOItZYLQ&ab\_channel=SeattleDataGuy

KASSAMBRA. (11/03/2018). *Regression Model Diagnostics*. Retrieved from STHDA: http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

Kim, B. (2015, September 21). *Understanding Diagnostic Plots for Linear Regression Analysis*. Retrieved from Research Data Sciences: https://data.library.virginia.edu/diagnostic-plots/

Magazine, P. (n.d.). *Gas Turbine Image.* Retrieved from https://www.google.com/imgres?imgurl=https%3A%2F%2Fwww.powermag.com%2Fwp-content%2Fuploads%2F2017%2F12%2F9ha-gas-turbine\_ge-power.jpg&imgrefurl=https%3A%2F%2Fwww.powermag.com%2Ftest-your-knowledge-gas-turbine-failure-modes%2F&tbnid=3KwKUZcGAlNCqM&vet=12ah

NIWA. (n.d.). Retrieved from Winter 2020: https://niwa.co.nz/climate/summaries/seasonal/winter-2020

NIWA. (n.d.). *Winter 2017*. Retrieved from https://niwa.co.nz/climate/summaries/annual-climate-summary-2017

NIWA. (n.d.). *Winter 2019*. Retrieved from https://niwa.co.nz/climate/summaries/seasonal/winter-2019

Team, T. R. (2021). *What is a Gas Turbine?* Retrieved from Realpars.com: https://realpars.com/gas-turbine/

University, T. A. (2019, September 12). *Simple Linear Regression | MSE RMSE & MAE | Model Evaluation Techniques - Part 2*. Retrieved from Youtube: https://www.youtube.com/watch?v=YSB7FtzeicA&ab\_channel=TheAIUniversity

Wikipedia. (2021, March 25). *Mean absolute percentage error*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Mean\_absolute\_percentage\_error

## R-Code-Modelling

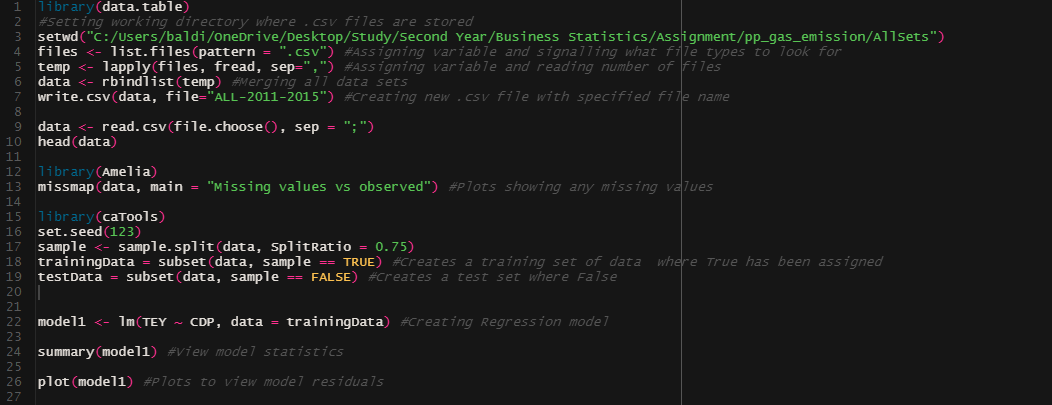
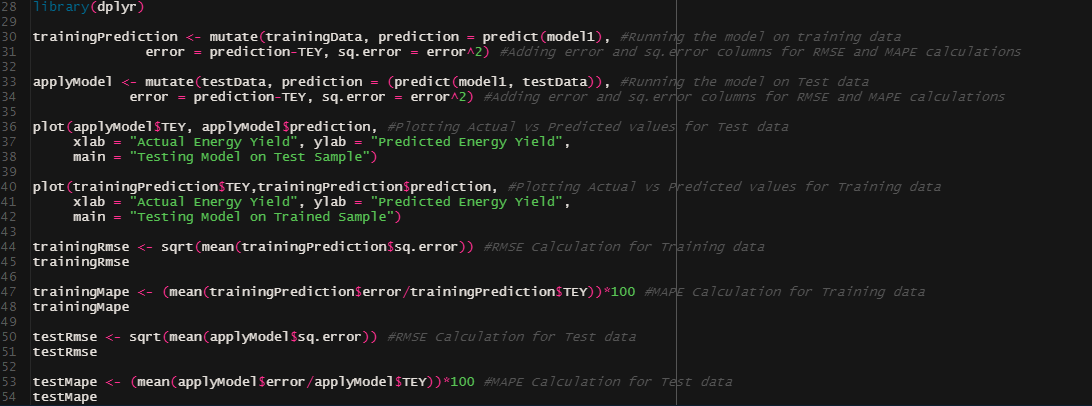


Figure - Final full code. Merges files, testing for completeness, creates training and test samples. Applying and testing model accuracy, refer to annotations.

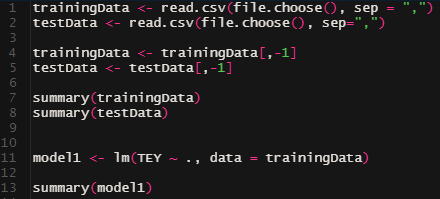


Figure 2 - Initial regression model all variables.

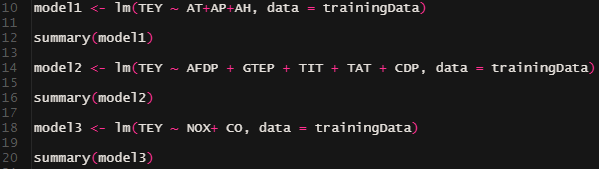


Figure 3 – Initially Modelling different attribute types.

## R-Code Timeseries

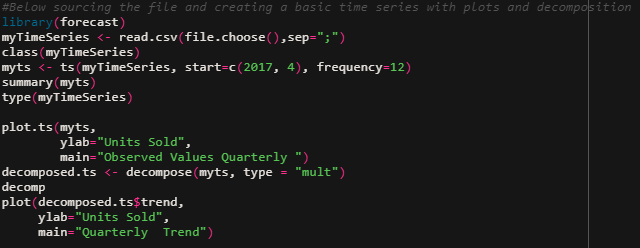


Figure 1-Below sourcing the file and creating a basic time series with plots and decomposition.

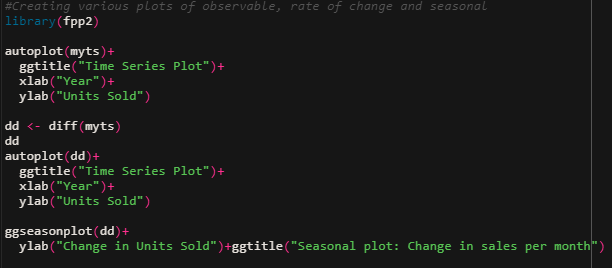


Figure 2-Creating various plots of observable, rate of change and seasonal.

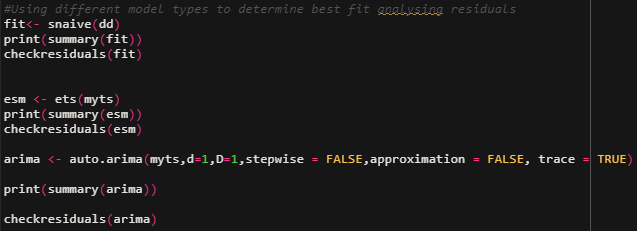


Figure 3-Using different model types to determine best fit analyzing residuals.

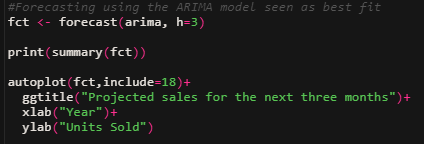


Figure 4-Forecasting using the ARIMA model seen as best fit.