**Time series analysis and forecasting using data from the National Office of Statistics regarding online sales versus retail sales from 2006 to 2023.**

Contents

[Executive Summary 3](#_Toc144636995)

[Data details 4](#_Toc144636996)

[Regression Analysis 4](#_Toc144636997)

[Time Series Analysis 7](#_Toc144636998)

[Forecasting model 9](#_Toc144636999)

[Recommendations and improvements 12](#_Toc144637000)

[References 13](#_Toc144637001)

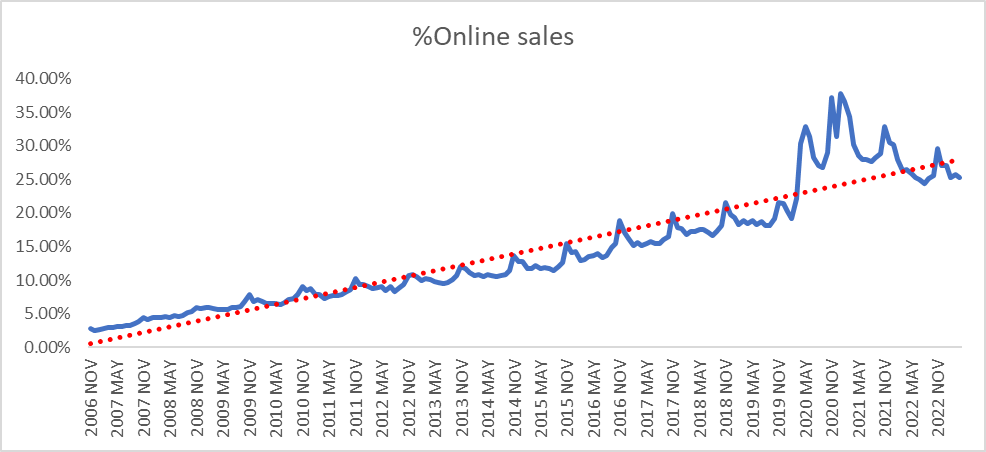
# Executive Summary

The aim of this project is to create a forecasting model using time series analysis that will be able to predict the percentage of sales done online versus in retail stores in future years.

Many businesses are seeing a shift into digital channels and this insight will be useful in helping them understand where their digital infrastructure needs to be in the future. It can also be used as guide to determine how their business is performing compared to the average. In turn, this analysis could help drive decisions needed to expand business’s digital capabilities to keep in line with evolving customer behaviour.

Analysis of the data has shown a strong linear correlation with monthly trends and has proved to be a statistically significant data model through regression analysis using month and year as the independent variables and percentage of sales as the dependent variable. (Image 1)

The prediction model output shows that sales will increase from the current 25.2% to 30.72% in 2025. The model showed a confidence level of 8% and a Root Mean Squared Error of 0.215 (based on 2023 actuals v predicted)

Image 1.

# Data details

Source – Office of National Statistics (ons.gov.uk)

[Internet sales as a percentage of total retail sales (ratio) (%) - Office for National Statistics (ons.gov.uk)](https://www.ons.gov.uk/businessindustryandtrade/retailindustry/timeseries/j4mc/drsi)

As the data is already structured and complete, it only needed a small amount of processing. (Image 2)

I split the data into groups of year, month, and quarter. I then added each to a line graph to help me decide the best set to work with. The month was decided on because there was a clear seasonal trend in the data that could be used to give a more accurate forecast. It is also the largest data set which should improve the statistical significance of the data.

A screenshot of a data

Description automatically generated with low confidenceImage 2.

To prepare the data for analysis, the date format needed amending into a number format rather than text to be able to run the regression analysis. I used text to columns to split the month from the year and formatted the months onto a categorical number.

# Regression Analysis

While not always suitable for time series data, (www.linkedin.com, n.d.) I have chosen to run a regression analysis model so that I can see whether there is a strong correlation between the month (independent variable) and the online sales %’s (dependant variable). I can then use this information towards the forecasting model and confidently take into account the seasonality changes within the year. When we look at the initial line graph, it does show an increase in certain months and the regression analysis will show me how strong the correlation is across those months.

To run the regression using months as the independent variable, I split the months out using dummy variables to be able to run the regression analysis and analyse the month trends. (Image 3)

Image 3.A screenshot of a computer

Description automatically generated with low confidence

Plotting the data into a line graph and adding a trendline helps visualise the trends in the data. (Image 4)

It is clear that there is an upwards trend year on year and that there are seasonality trends with an increase seen each November. It is also clear that 2019 and 2020 saw some large changes in customer behaviour with online sales % increasing far above the trend and this can be attributed to the Covid pandemic and retail stores being closed for much of this period.

What is interesting, is that once we enter 2022, the overall linear trend seems to come back with a reduction in online sales seen from 2021 to 2022. This could be useful information for companies if having to react to another pandemic.

A graph with a line and a red dotted line

Description automatically generatedImage 4.

I ran the regression analysis only using the data up to 2019. This removes the impact seen from covid lockdown as the trend changes substantially over this period but then during 2022 the online sales % seems to come back in line with trend as can be seen from the line graph. (Image 5)

Image 5.

A graph with a line and a red dotted line

Description automatically generated

Image 6.

The regression analysis output shows (Image 6) the data set has a Multiple R of 99% and an R Square of 98% which tells us that there is a strong relationship between the month and the % of sales made online.

It also has a very high F Stat (655.425) and the associated P Value is 0 meaning that the model is significant and I can reject the Null hypothesis. (The month of year has no impact to online sales %’s)

Finally, the standard of error for the final model is very low, suggesting further that the model is very precise.

# Time Series Analysis

As well as the regression model, I will also run a time series analysis model in excel. This is because, as mentioned previously, linear regression is not always suitable for time series data. While it was useful to prove the correlation between months and sales, the trends seen in this instance are clearly non-stationary and a time series approach would allow for better insights.

To complete this, I used the analysis function of exponential smoothing using the online sales %’s as the input range. Excel then calculates the standard errors and plots the data with a forecast into a graph. (Image 7)

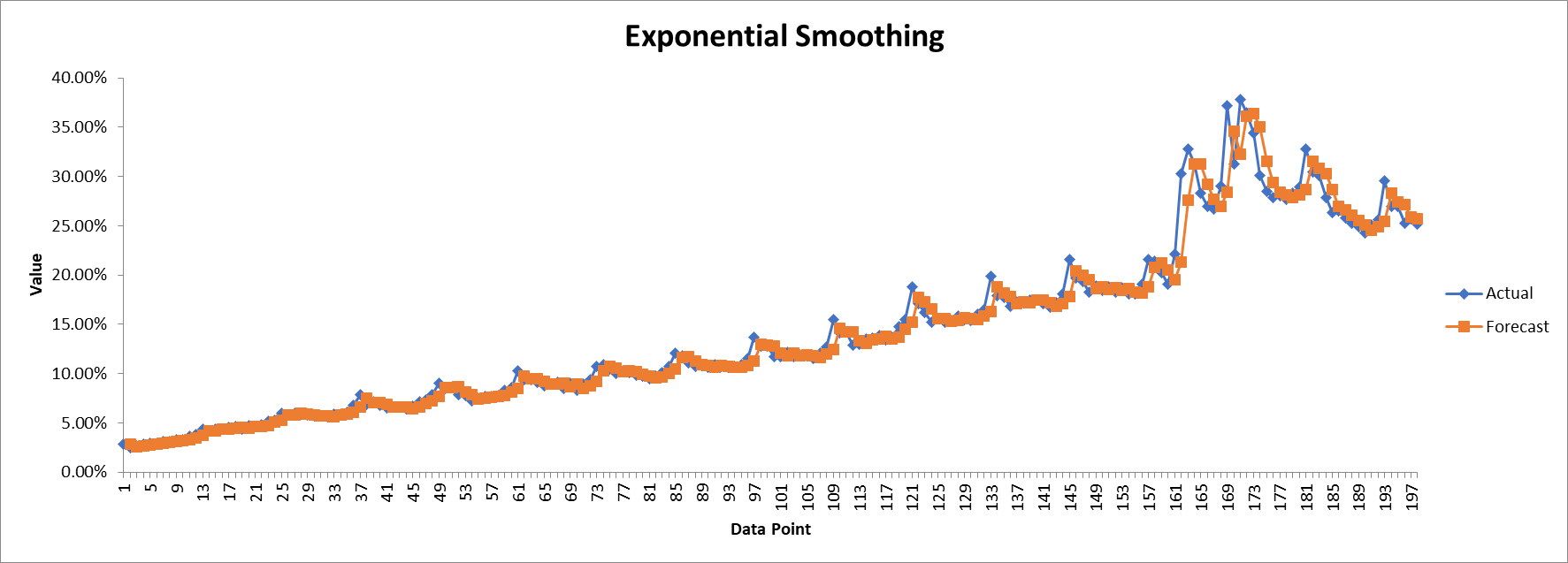
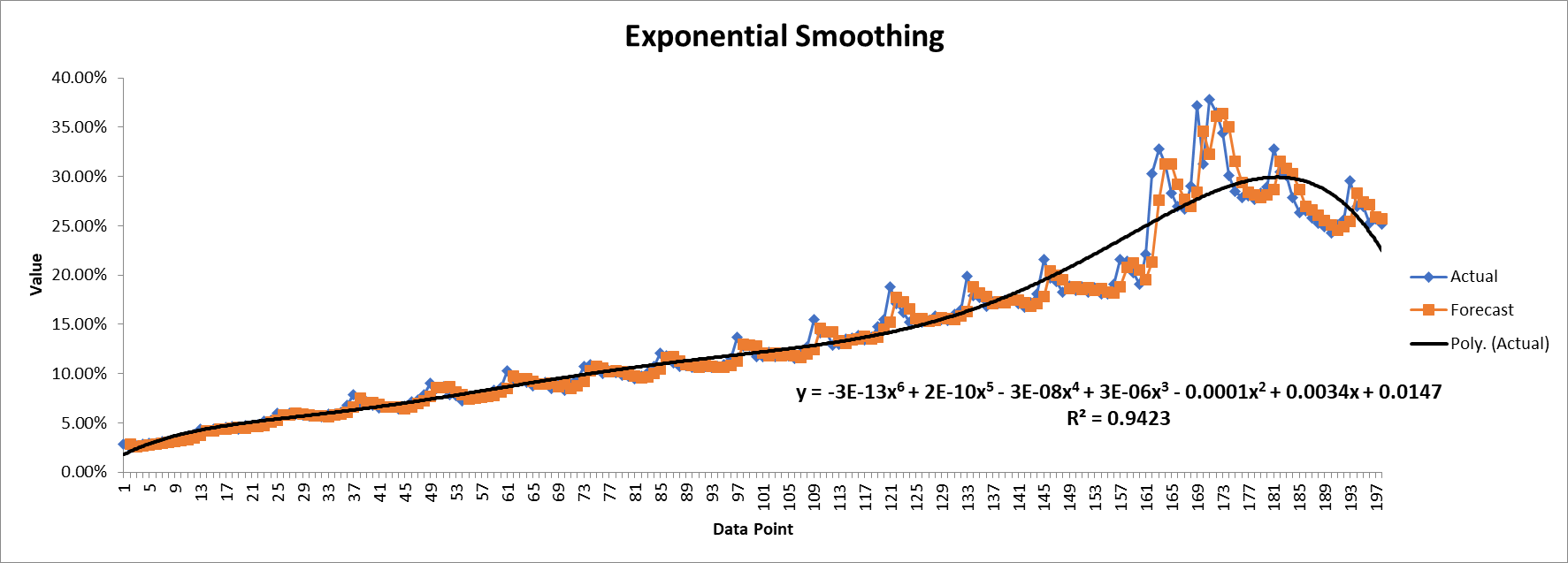
Image 7.

Image 8.

A screenshot of a computer

Description automatically generated

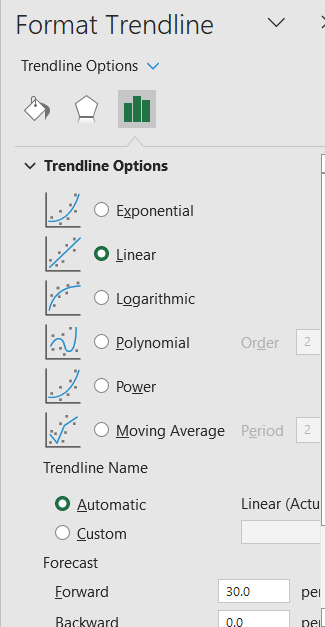
I then plot a polynomial trend line to this data to minimise the error of the forecasting model and add the R-squared values and the equations on the chart. (Image 8)

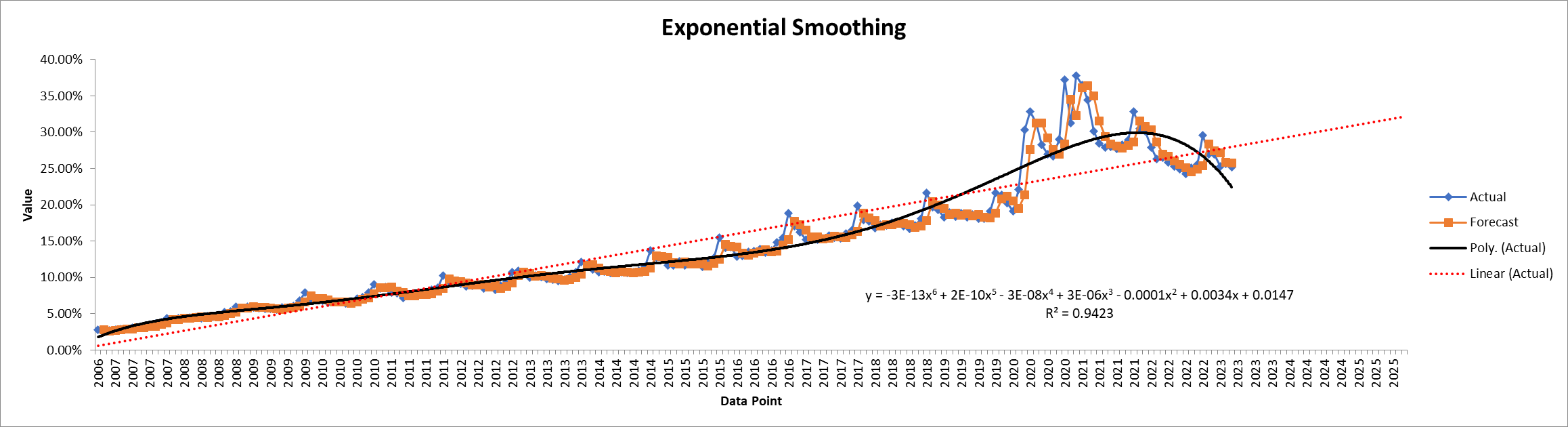
Image 9.

The output shows (Image 9) that the R squared equals 0.9423 which means that this ratio explains 94.23% of changes in sales in process of time.

We can now add a linear trend line which is recommended by most authors to do here (exceltable.com, n.d.) This allows for forecasting the sales in future months by specifying the number of periods to project the trendline forwards. I used 30 as this takes me up to the end of 2025. (Image 10)

Image 10.



Image 11.

The above (Image 11) shows that by the end of 2025 the online sales % will be at 32.5%

This method of forecasting, however, has a few limitations. Firstly, does not show seasonality or trends in the forecast. Also, while the high R2 value suggests that a significant portion of the variance is explained by the model, we must consider the possibility that the model has overfitted the data and this could be problematic as it could lead to poor generalisation to the future data points. (Liu, 2000)

# Forecasting model

To be able to take into account trends and seasonality in the data, I have looked to use the forecasting function in excel to compare the results and determine whether this is a more accurate way to forecast future online sales.

I have excluded 2021 and 2022 from the data used for the forecast model as I do not want it to impact the forecast. This is important to do because the shopping behaviour during COVID-19 mainly changed due to enforced regulations rather than inherent shifts in preferences. We can see in the line graph that as lockdowns lifted, shopping behaviour reverted to the usual trends, indicating that the changes were a result of external factors rather than lasting alterations in consumer habits.

The forecasting model has been built in excel using the forecasting function available. The reason for this is that it is widely used and makes the model easy for others to replicate. It is also the main tool in my workplace making it accessible to everyone.

A new table has been created with the data and a new column added to turn the year/month into a format that excel will recognise as a date. As shown below in image 12.

Image 12.

A screenshot of a spreadsheet

Description automatically generated

I have chosen to use the Exponential Triple Smoothing (ETS) method in excel as it captures trends and seasonality but also allows you to specify the period of the seasonal trends. It also gives more weight to the more recent observations to help with accuracy. (Time series Forecasting using Holt-Winters Exponential Smoothing, 2004) As it is clear in the graph that some months see a higher % of online sales, it is important to capture this in the forecast.

I have used the forecast formula:

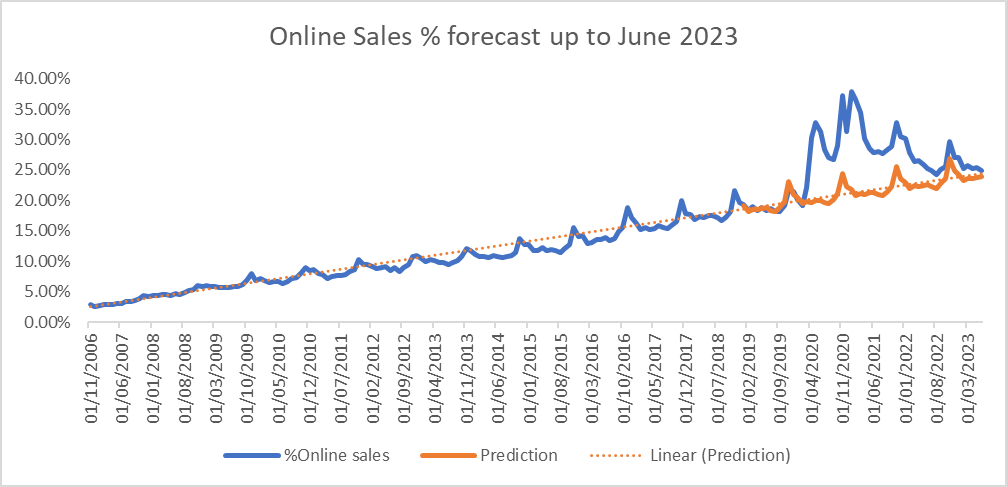
=FORECAST.ETS(B148,$C$4:$C$147,$B$4:$B$147,12))

This covers the data from 2006 up to and including 2018. As mentioned before, I am excluding 2020 and 2021 due to the changes in trends over the pandemic, but I am also excluding 2019 and 2022 as I can use these periods to test the accuracy of the forecast.

The output shows a confidence interval of 0.7% for 2019 which increases to 3.5% for 2023. It has a Root Mean Squared Error (RMSE) of 0.215 which implies that the model has relatively good predictive accuracy, with predictions being, on average, within approximately 0.215% for the 2023 period.

The graph below (Image 13) shows the prediction for 2019 to 2023 versus the actual online sales percentage and the seasonal trends can be easily seen. It essentially shows a relatively good prediction model if we exclude the impact of the pandemic in 2020 and 2021.

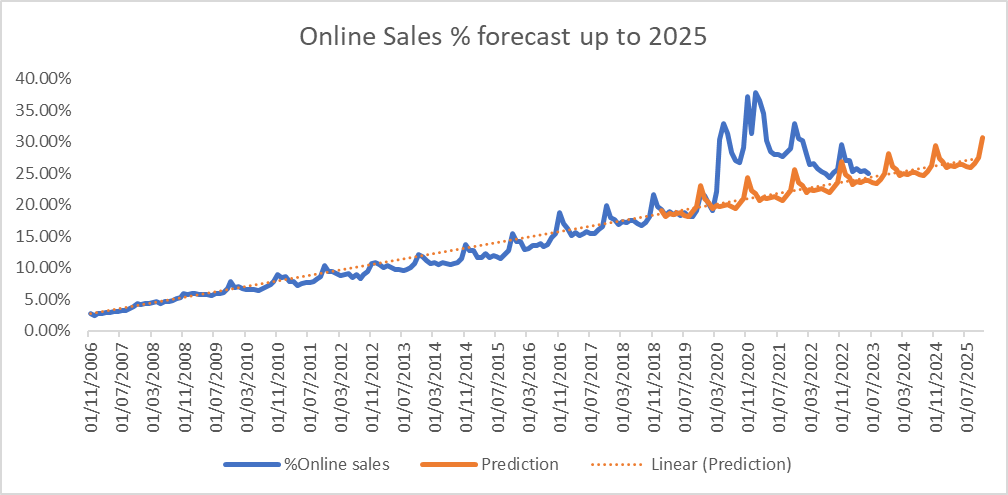
Image 13.



Using this forecasting model to look further ahead, the confidence interval reduces 7% for 2025, which would be expected as forecasts become for uncertain over a long period.

This model (Image 14) shows that, by the November 2025, the online sales % will have increased to 31% from the 25% in July 2023.

Image 14.



# Recommendations and improvements

The recommendations here would be to test different forecasting models to find the one with the highest confidence interval and the lowest RMSE. Options to consider would be:

* ARIMA (AutoRegressive Integrated Moving Average): ARIMA is a powerful and widely used time series forecasting method that combines autoregressive and moving average components to model the data,
* Seasonal Decomposition of Time Series (STL): STL decomposes a time series into its seasonal, trend, and residual components, making it easier to forecast each component separately.

(Wen et al., 2019)

The data set could also be used to build a forecasting model in other tools such as R or Python. Both are widely used and provides a range of packages for data analysis, visualization, exploratory data analysis and time series forecasting. (Sahoo et al., 2019)

# References

Exceltable.com. (n.d.). Time series analysis and forecasting in Excel with examples. [online] Available at: <https://exceltable.com/en/analyses-reports/time-series-analysis>.

Linkedin.com. (n.d.). What are some applications and examples of time series and regression in your industry? [online] Available at: <https://www.linkedin.com/advice/0/what-some-applications-examples-time-series#:~:text=Time%20series%20assumes%20that%20the>

Liu, Y. (2000). Overfitting and Forecasting: Linear versus non -linear time series models.

Sahoo, K., Samal, A.K., Pramanik, J. and Pani, S.K. (2019). Exploratory Data Analysis using Python. International Journal of Innovative Technology and Exploring Engineering, 8(12), pp.4727–4735. doi: <https://doi.org/10.35940/ijitee.l3591.1081219>.

Time series Forecasting using Holt-Winters Exponential Smoothing. (2004). Available at: <https://c.mql5.com/forextsd/forum/69/exponentialsmoothing.pdf>.

Wen, Q., Gao, J., Song, X., Sun, L., Xu, H. and Zhu, S. (2019). RobustSTL: A Robust Seasonal-Trend Decomposition Algorithm for Long Time Series. Proceedings of the AAAI Conference on Artificial Intelligence, 33, pp.5409–5416. doi: <https://doi.org/10.1609/aaai.v33i01.33015409>.