Using Time Series Analysis for Sales and Demand Forecasting

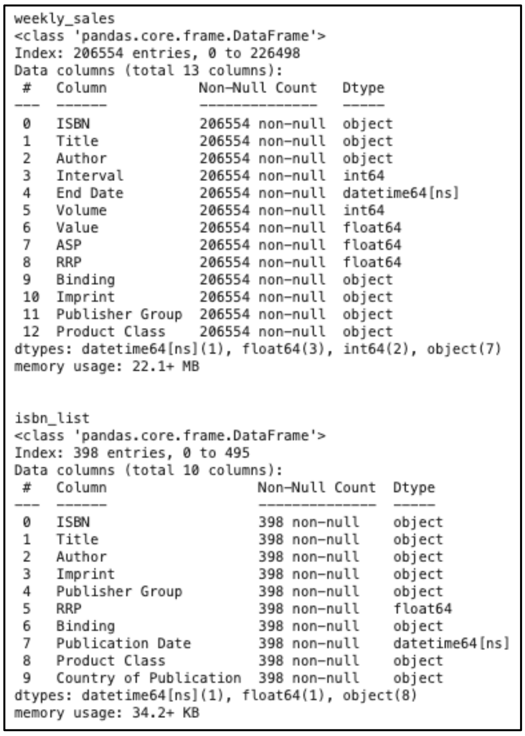
Diptarko Chowdhury
7TH FEBRUARY 2025

Problem Statement

Demand for books can fluctuate drastically, causing difficulties for small to medium-sized independent publishers in investment, stocking and distribution. This project aims to create time series models that can follow seasonal patterns in sales trends for certain books that can assist such publishers.

The Data Set & Exploratory Data Analysis

The datasets (Figure 1) used in this project were twofold. `weekly_sales` contained information around how many sales for each book occurred in weeks from 01/06/2001 to 07/20/2024 (with the Column End Date) referring to the last day of the sales week. `isbn_lis`t contained information only pertaining to the actual book, much of this information was shared in weekly_sales, but it included two more variables, `Publication Date` and `Country of Publication`.
Initial steps included joining the tables by `ISBN`, ensuring all columns had correct data types, and creating extra rows to include weeks where books had zero sales, as this was not present in `weekly_sales`. For further analysis, only books with ISBNs which had sales data beyond 01/07/2024 were kept, of which there were 61.



```
weekly_sales
<class 'pandas.core.frame.DataFrame'>
Index: 206554 entries, 0 to 226498
Data columns (total 13 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   ISBN             206554 non-null  object
 1   Title            206554 non-null  object
 2   Author           206554 non-null  object
 3   Interval         206554 non-null  int64
 4   End Date         206554 non-null  datetime64[ns]
 5   Volume           206554 non-null  int64
 6   Value            206554 non-null  float64
 7   ASP              206554 non-null  float64
 8   RRP              206554 non-null  float64
 9   Binding          206554 non-null  object
 10  Imprint          206554 non-null  object
 11  Publisher Group  206554 non-null  object
 12  Product Class    206554 non-null  object
dtypes: datetime64[ns](1), float64(3), int64(2), object(7)
memory usage: 22.1+ MB


isbn_list
<class 'pandas.core.frame.DataFrame'>
Index: 398 entries, 0 to 495
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   ISBN                    398 non-null    object
 1   Title                   398 non-null    object
 2   Author                  398 non-null    object
 3   Imprint                 398 non-null    object
 4   Publisher Group         398 non-null    object
 5   RRP                     398 non-null    float64
 6   Binding                 398 non-null    object
 7   Publication Date        398 non-null    datetime64[ns]
 8   Product Class           398 non-null    object
 9   Country of Publication  398 non-null    object
dtypes: datetime64[ns](1), float64(1), object(8)
memory usage: 34.2+ KB
```

Figure 1: Raw data sets used for analysis

There was a large variance in sales trend across books, as can be seen in Figure 2. The vast majority of books though, had far fewer sales between 2013-2024, than they did in 2001-2012, as witnessed in Figure 3. The sales patterns could be due to a multitude of reasons.

- Books look to undergo a high volume of sales around release, this could be due to marketing compaigns and anticipation from fans in the release build-up
- Seasonal peaks may arise around gift-giving periods such as Christmas
- The advent of e-books such as kindles, audiobooks, as well as alternate forms of entertainment becoming more popular (TV, streaming services) may have led to the decline of people buying books in print

For further analysis in this project, two books are selected: ‘The Alchemist’ and ‘The Very Hungry Caterpillar’.

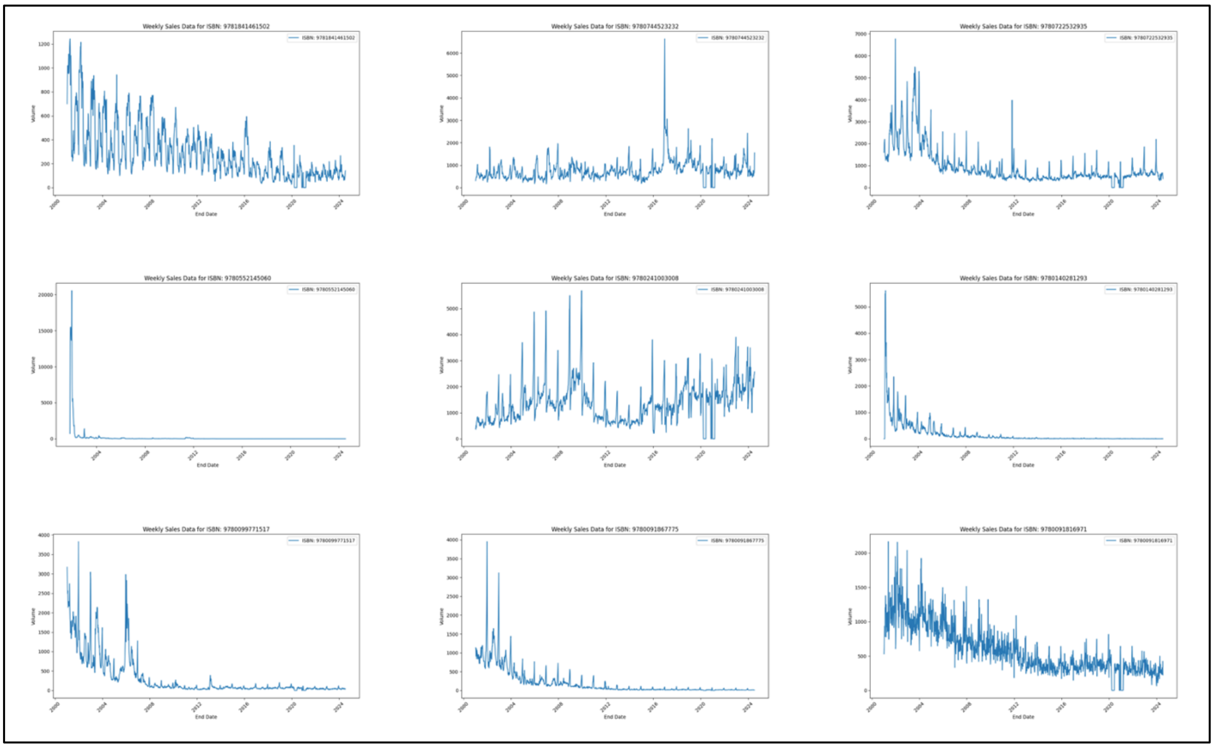


Figure 3: Weekly sales data for a random selection of books

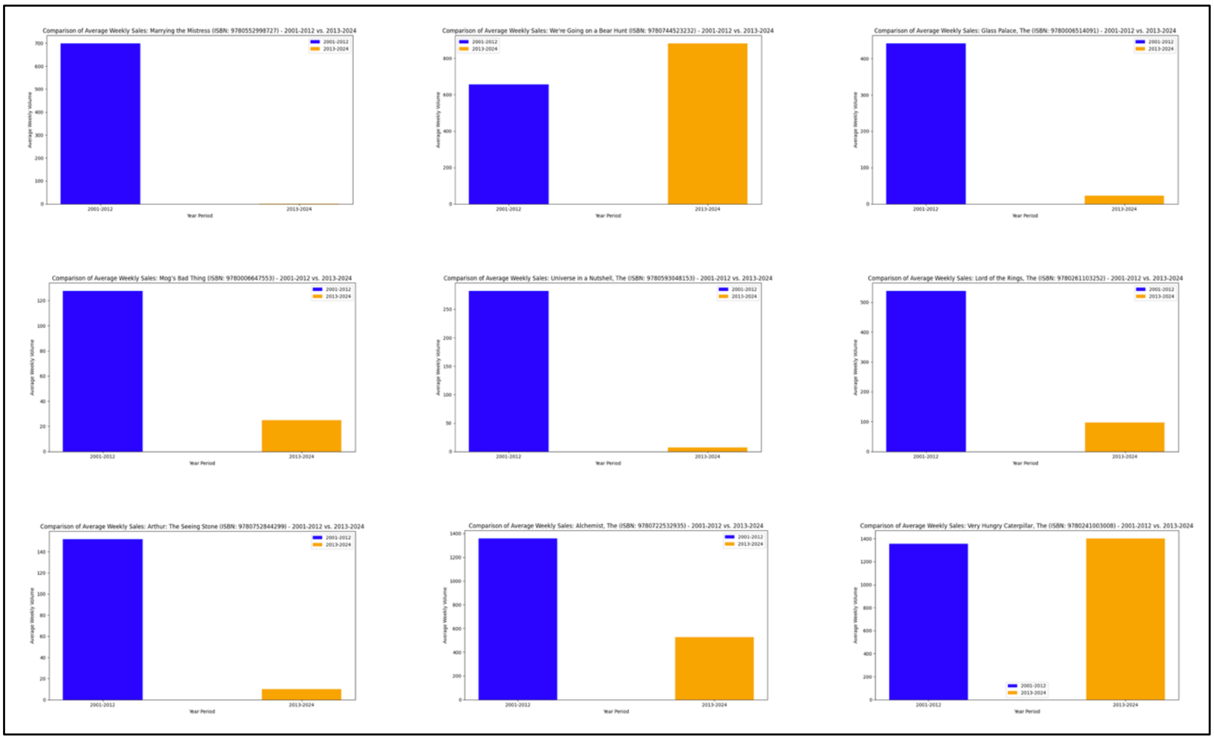


Figure 2: Notable drop off in sales for most books

Classical Time Series Analysis

The sales volume trends of voth ‘The Alchemist’ and ‘The Very Hungry Caterpillar’ can be seen in Figure 4. Since there are zero values for some weeks (perhaps coinciding with closures due to the Covid-19 pandemic), multiplicative decomposition is not possible, so additive decomposition is done for the time series as shown in Figure 5.

Both books exhibit strong seasonal patterns with increasing seasonal peaks. This is crucial information for publishers and book stockers.

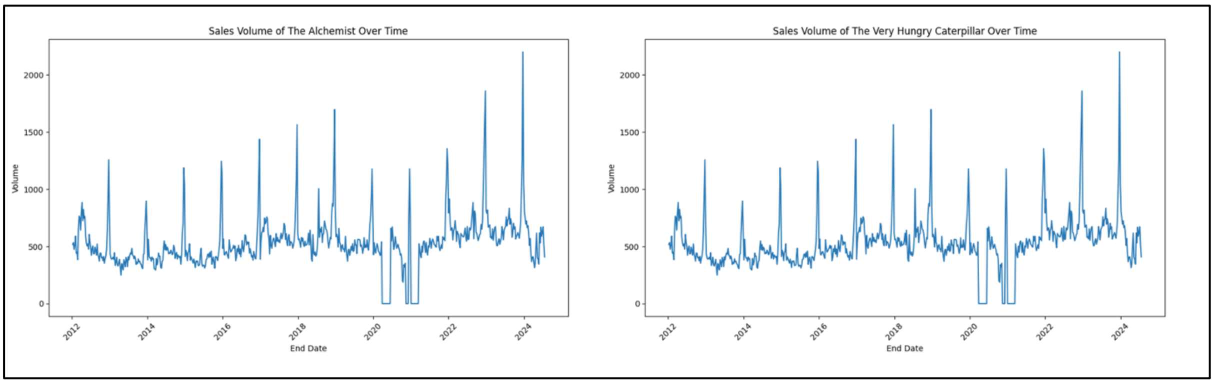


Figure 4: Sales volume of both books over time

The autocorrelation function (ACF)) result (Figure 5) for the alchemist show positive correlations with the first few lags, with then the 52nd lag also showing significance showing a yearly seasonality. Partial autocorrelation (PACF) shows a strong peak at lag 1; a first-order autogressive (AR(1)) process may be a good fit. The ACF for the Very Hungry Caterpillar shows even more long-term dependeencies, with fluctuating peaks at roughly twelve week intervals. Simlarly, AR(1) looks to be a good fit.

After confirming that both series were stationary using the Augmented Dickey-Fuller test, uising a forecast horizon of 32 weeks, Auto Arima was used to find the optimal model parameters for a forecast model, results shown in Figure 6.

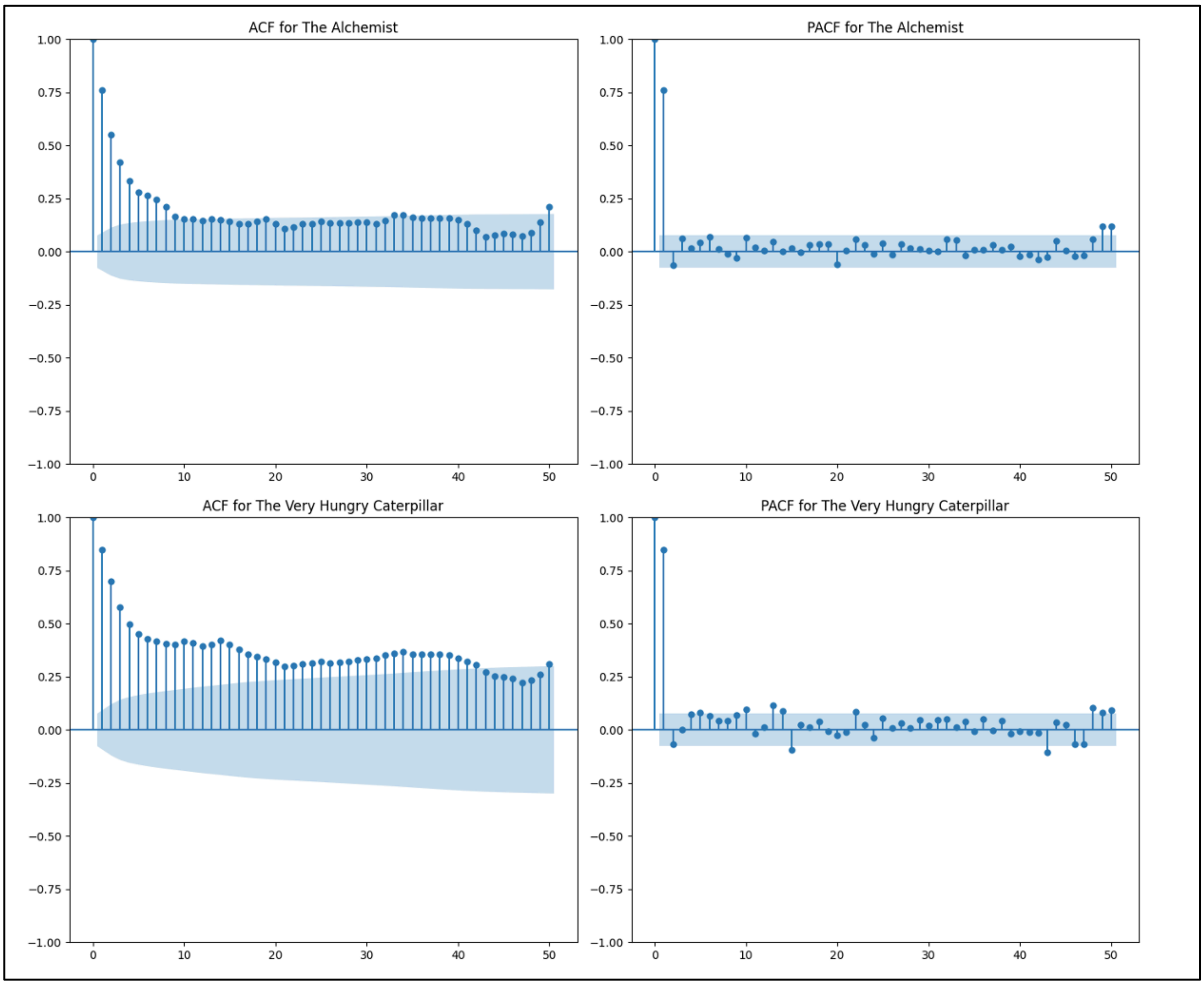
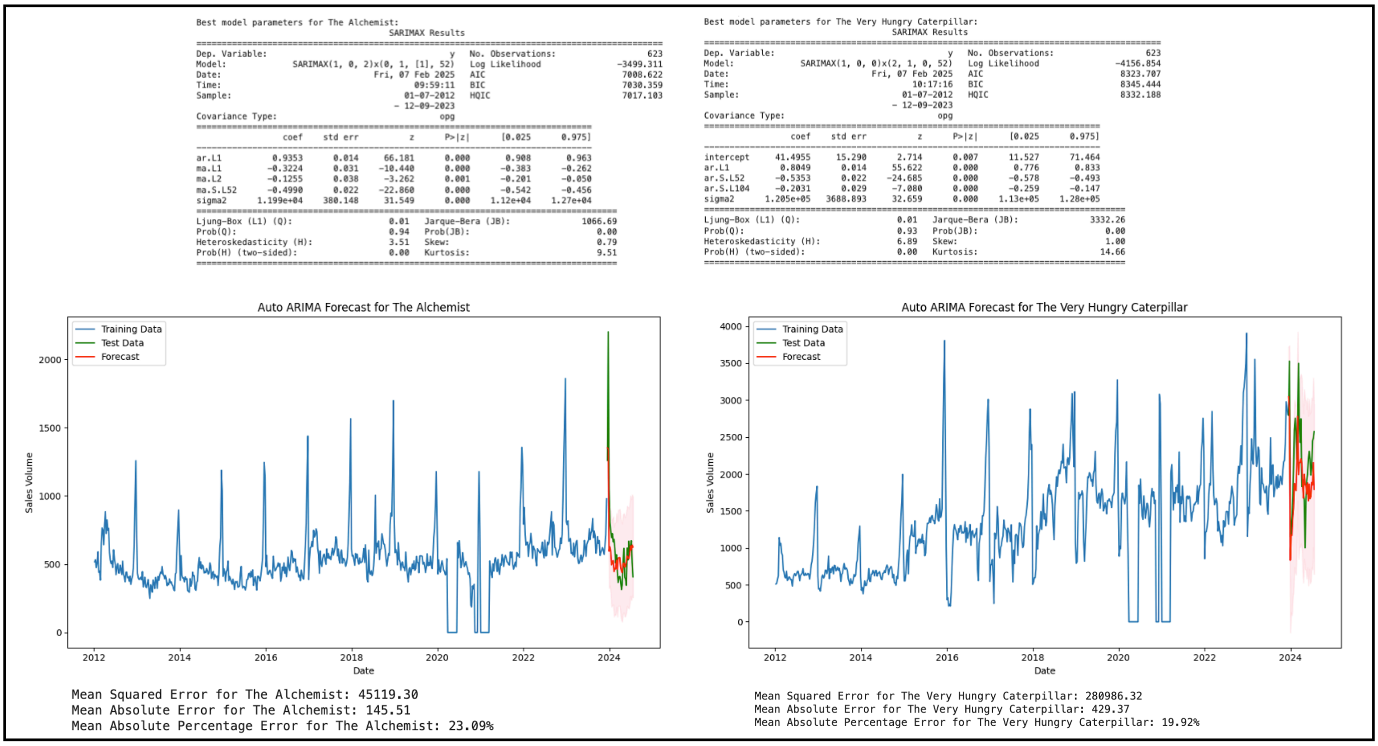


Figure 5: ACF and PACF plots for both books



```
Best model parameters for The Alchemist:
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                  623
Model:             SARIMAX(1, 0, 2)x(0, 1, [1], 52)   Log Likelihood               -3499.311
Date:                          Fri, 07 Feb 2025   AIC                           7008.622
Time:                                  09:59:11   BIC                           7030.359
Sample:                              01-07-2012   HQIC                          7017.103
                                   - 12-09-2023
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9353      0.014     66.181      0.000       0.908       0.963
ma.L1         -0.3224      0.031    -10.440      0.000      -0.383      -0.262
ma.L2         -0.1255      0.038     -3.262      0.001      -0.201      -0.050
ma.S.L52      -0.4990      0.022    -22.860      0.000      -0.542      -0.456
sigma2      1.199e+04    380.148     31.549      0.000    1.12e+04    1.27e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):              1066.69
Prob(Q):                              0.94   Prob(JB):                         0.00
Heteroskedasticity (H):               3.51   Skew:                             0.79
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.51
===================================================================================
```

```
Best model parameters for The Very Hungry Caterpillar:
                                     SARIMAX Results
=========================================================================================
Dep. Variable:                               y   No. Observations:                  623
Model:             SARIMAX(1, 0, 0)x(2, 1, 0, 52)   Log Likelihood               -4156.854
Date:                         Fri, 07 Feb 2025   AIC                           8323.707
Time:                                 10:17:16   BIC                           8345.444
Sample:                             01-07-2012   HQIC                          8332.188
                                  - 12-09-2023
Covariance Type:                           opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     41.4955     15.290      2.714      0.007      11.527      71.464
ar.L1          0.8049      0.014     55.622      0.000       0.776       0.833
ar.S.L52      -0.5353      0.022    -24.685      0.000      -0.578      -0.493
ar.S.L104     -0.2031      0.029     -7.080      0.000      -0.259      -0.147
sigma2      1.205e+05   3688.893     32.659      0.000    1.13e+05    1.28e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):              3332.26
Prob(Q):                              0.93   Prob(JB):                         0.00
Heteroskedasticity (H):               6.89   Skew:                             1.00
Prob(H) (two-sided):                  0.00   Kurtosis:                        14.66
===================================================================================
```

```
Mean Squared Error for The Alchemist: 45119.30
Mean Absolute Error for The Alchemist: 145.51
Mean Absolute Percentage Error for The Alchemist: 23.09%
```

```
Mean Squared Error for The Very Hungry Caterpillar: 280986.32
Mean Absolute Error for The Very Hungry Caterpillar: 429.37
Mean Absolute Percentage Error for The Very Hungry Caterpillar: 19.92%
```

Figure 6: Optimal ARIMA model results for both books

The Alchemist: SARIMAX(1, 0, 2)x(0, 1, [1], 52)

- AR(1) coefficient: 0.9353: the first lag of the series has a strong positive relationship with the current value
- MA(1) coefficient: -0.3224: there is some negative short-term correction in the model's error terms.
- MA(2) coefficient: -0.1255: another negative correction term for the second lag.
- Seasonal MA(S)(52) coefficient: -0.4990, indicating strong seasonal effects on sales on an annual basis, with a negative relationship between this lag and the sales volume.
- Sigma^2 (Variance): 1.199e+04, which provides a measure of the model's residual variance.

The Very Hungry Caterpillar: SARIMAX(1, 0, 0)x(2, 1, 0, 52)

- AR(1) coefficient: 0.8049: first lage of the series has a strong positive relationship with the current value, just slightly less than in the case of The Alchemist
- Seasonal AR(52) coefficient: -0.5353: negative seasonal effect that peaks at yearly intervals
- Seasonal AR(104) coefficient: -0.2031: another negative (but weaker) seasonal effect every two years
- Sigma^2 (Variance): 1.205e+05: order of ten higher than The Alchemist model, meaning that there's greater variance in this data (which matches with the decomposition that we did earlier)

From the plot of residuals (Figure 7), we can ascertain that the distribution of residuals in both cases is fairly normal, with the centre near zero. There are outliers present in both cases, with the spread of residuals for The Very Hungry Caterpillar being notably larger, so there may be some seasonal fluctions or trends not fully accounted for tby the models. The residuals’ ACF plots in both cases indicate a largeaurocorrelation at lag = 0 but then minimal significance beyond that, suggesting that the residuals are infact random and uncorrelated (after trends and seasonality have been extracted).

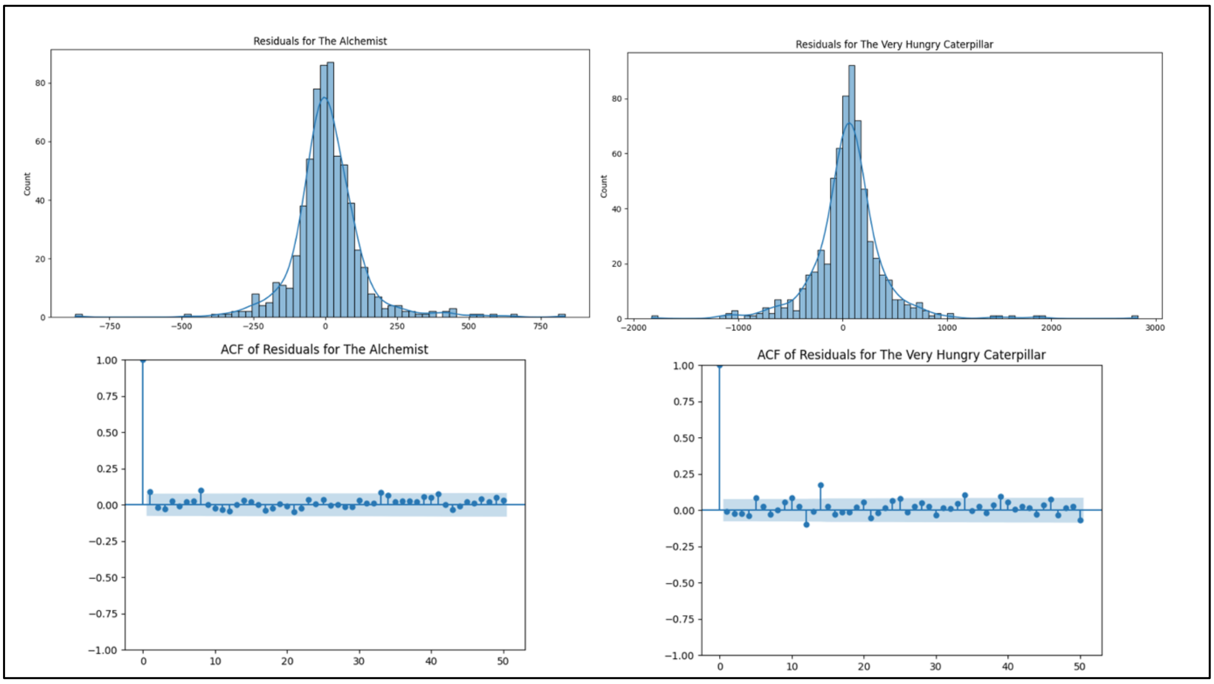
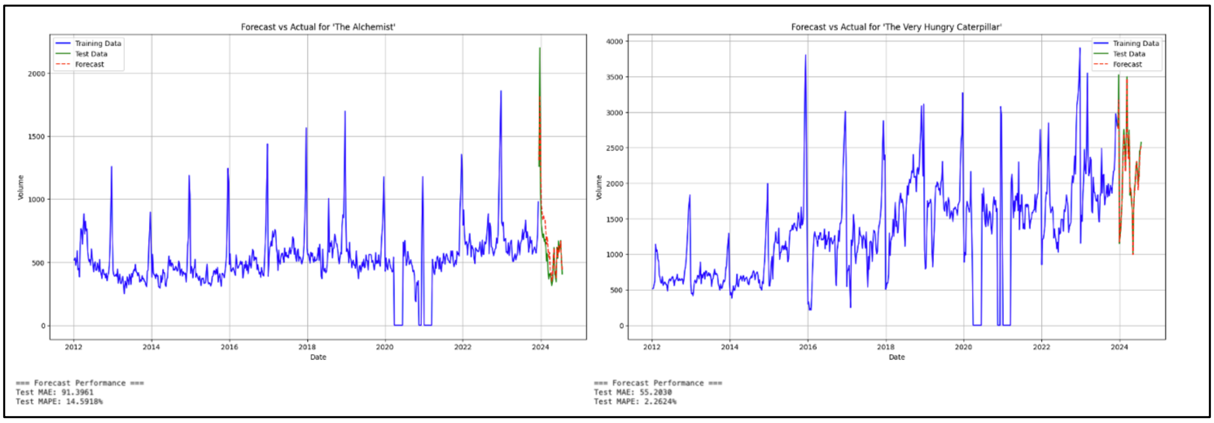


Figure 7: Residual plots for both books

Time Series Analysis with Machine Learning

The first ML technique used was XGBoost. After conducting hyperparameter tuning, the optimal XGBoost model was fitted with results as in Figure 8.

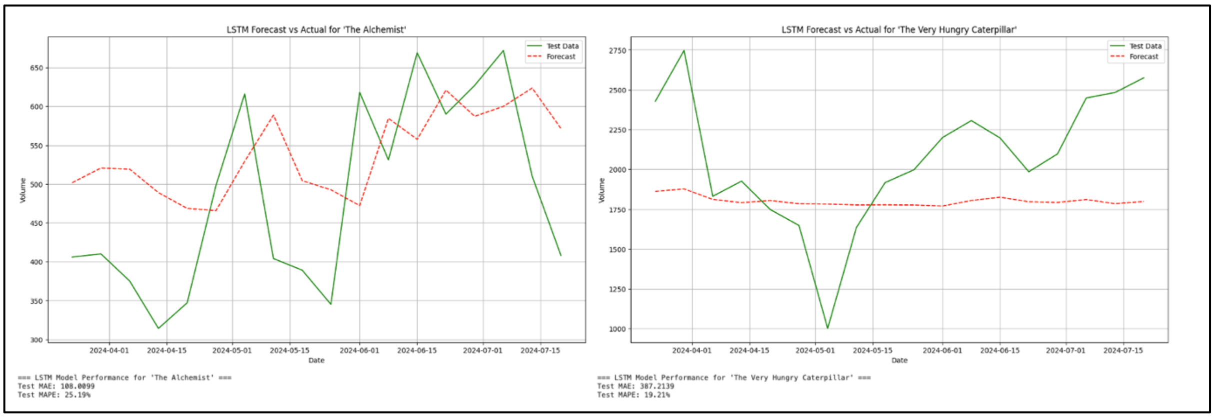


```
=== Forecast Performance ===
Test MAE: 91.3961
Test MAPE: 14.5918%
```

```
=== Forecast Performance ===
Test MAE: 55.2030
Test MAPE: 2.2624%
```

Figure 8: XGBoost forecast for both books

With MAPEs of 14.6% and 2.3% these were a considerable improvement on the Auto ARIMA. The second ML technique used was LSTM with results as in Figure 9; a poorer fit with higher MAPEs, the Very Hungry Caterpillar fit in particular looks very underfitted with short term volatility almost entirely omitted from the model. Interestingly there seems to be a lag in The Alchemist fit by one data point, although the model visually looks to follow the test data better, it has a higher MAPE.



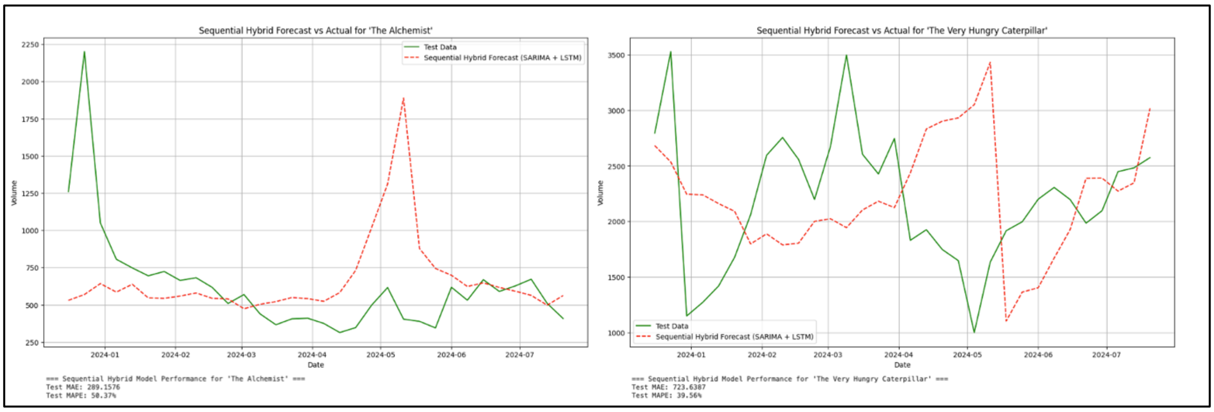
```
=== LSTM Model Performance for 'The Alchemist' ===
Test MAE: 108.0099
Test MAPE: 25.19%
```

```
=== LSTM Model Performance for 'The Very Hungry Caterpillar' ===
Test MAE: 387.2139
Test MAPE: 19.21%
```

Figure 9: LSTM forecast for both books

Hybrid Models

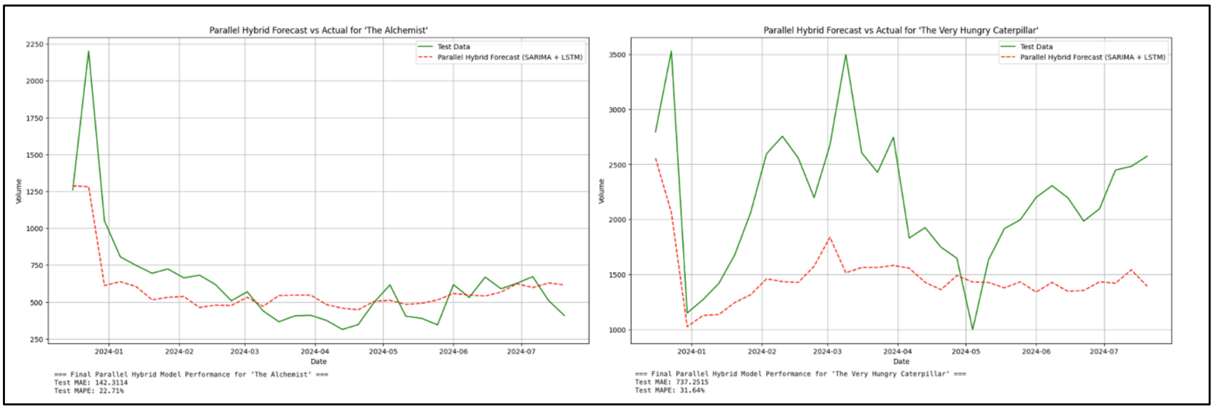
Sequential and Parallel Hybrid models (SARIMA + LSTM) were also tested (Figures 9, 10).



```
=== Sequential Hybrid Model Performance for 'The Alchemist' ===
Test MAE: 289.1576
Test MAPE: 50.37%
```

```
=== Sequential Hybrid Model Performance for 'The Very Hungry Caterpillar' ===
Test MAE: 723.6387
Test MAPE: 39.56%
```

Figure 10: Sequential hybrid model for both books



```
=== Final Parallel Hybrid Model Performance for 'The Alchemist' ===
Test MAE: 142.3114
Test MAPE: 22.71%
```

```
=== Final Parallel Hybrid Model Performance for 'The Very Hungry Caterpillar' ===
Test MAE: 737.2515
Test MAPE: 31.64%
```

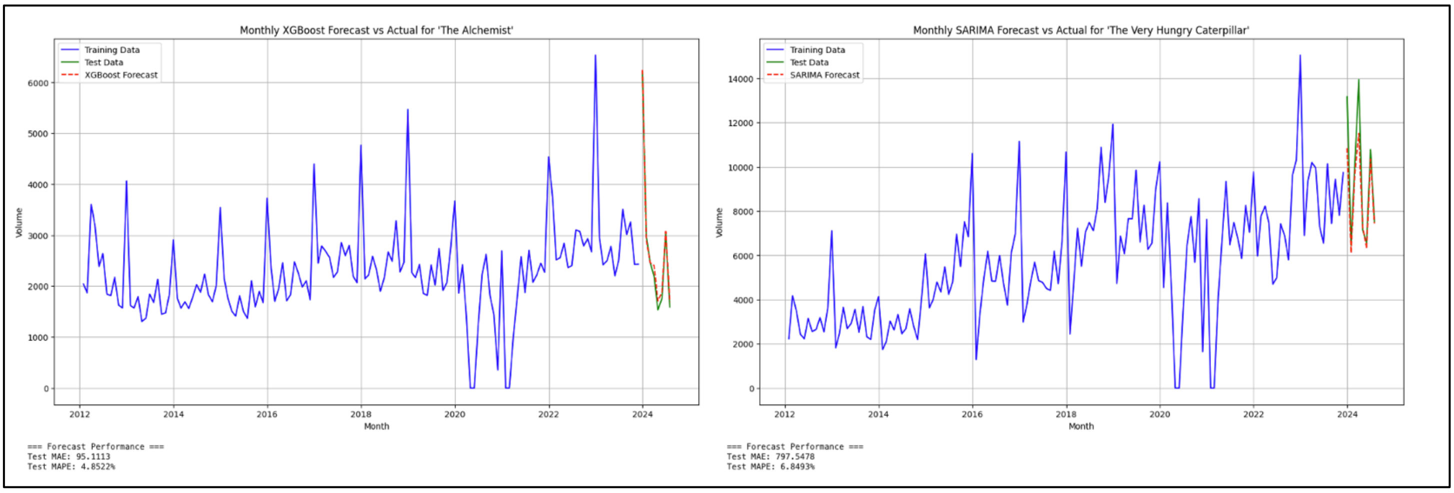
Figure 11: Parallel hybrid model selection for both books

In both instances the parallel hybrid model was an improvement (in the case of the Alchemist, a marked improvement) on the sequential model. However these still fell short of the XGBoost model's performance. Interestingly, the optimal weighting of the LSTM in the parallel combinations was low (0.2, 0.0), displaying the LSTM model's ineffectiveness in effectively modelling these time series (without further tuning).

Monthly Prediction

As a final exercise, the dataframes for both bookes were re-aggregated so that sales data was logged monthly. The ensuing monthly data was then put through two models, an XGBoost and a SARIMA model (Figure 12).

In both instances, treating the sales volume in a monthly manner resulted in a much better model performance. This is to be expected, as by aggregating sales volume over a longer time period smoothens out short-term fluctuations. If the residuals were to be calculated, it is likely they'd be smaller. It is intuitively obvious that weekly data would be skewed more by the effect of holidays and other events, making monthly aggregation overall preferrable for analysis.



```
=== Forecast Performance ===
Test MAE: 95.1113
Test MAPE: 4.8522%
```

```
=== Forecast Performance ===
Test MAE: 797.5478
Test MAPE: 6.8493%
```

Figure 12: Monthly forecast models

Conclusion

Further exploration and tuning is obviously required, but in the current state, using an XGBoost model in particular (especially across monthly data) to forecast the sale of both The Alchemist and The Very Hungry Caterpillar proves to be an effective tool for publishers. Next steps should be to try hybrid models using XGBoost, as opposed to LSTM, as this is yet unexplored and could provide further accurate models.