BAX 372T/ OM 372T/ STA 372T

Group Project #2

Oliver Wu, Olivia Pan

Executive Summary:

In this project, our goal was to predict the house sales volume in each month in the UK so that the general public and real estate developers will have more knowledge on the housing market. This will reduce the real estate market loss due to excessive construction or irrational housing market investment. We analyzed historical data from the UK, seasonally adjusted the data, and built an ARIMA model on both seasonally adjusted transactions and seasonalities to forecast future sales volumes. Our approach accounted for major disruptions such as COVID (pandemic) and utilized time series decomposition to separate long term trends from seasonal effects. The final forecasts provide point predictions along 95% prediction intervals. This gives a practical outlook for the near future of the UK housing market.

Report

Introduction:

Living expenses have always been a significant part of people's daily expenditures, typically accounting for around 50% of take-home pay. Numerous factors influence house transaction prices, including overall market conditions, property features, and amenities associated with the properties.

Investigating the long-term trends, seasonality patterns, and forecasting the growth of house transaction prices is crucial because it provides valuable insights for the general public to make informed investment decisions. It also enables real estate developers to better understand market dynamics and strategize accordingly. In order to eliminate confusing factors such as wealthy level differences and culture differences, we only selected house sales history in the UK. Therefore, the central question we aim to address is: **How can we accurately forecast and predict house transaction volumes in the UK?**

Dataset Overview:

The UK property prices with sales history data set comes from Kaggle, available at: UK Property Prices with Sale History. The dataset contains columns property_id which is the unique id that links each transaction to a property; id which is the transaction id; display price, deedDate which represents the transaction date; newBuild which means whether it is a new construction; tenure which stands for ownership type at the time of sale; and percentageChange which is the percent increase or decrease

from the last sale. We only selected deedDate because we only need the transaction time. We sorted the column so it is in chronological order.

Methodology:

We started by processing UK house transaction data, focusing on transaction dates (deedDate) to measure monthly sales volume. After aggregating the number of sales per month from January 2010 through October 2024, we noticed a strong seasonality and a pandemic related disruption around 2020-2023, shown in figure 1. To stabilize variance we applied a logarithmic transformation (ln) to the monthly sales figures, the line plot of log sales versus time is shown in figure 2.

To address the pandemic effect, sales data from March 2020 to May 2023 were set to missing, shown in figure 3. Then we decomposed the log-transformed series using STL with a window size of 31, chosen after multiple rounds of experimentation on the ACF (autocorrelation functions) plots. The seasonally adjusted series was denoted as LogA, the line plot is shown in figure 4.

Stationarity was assessed through plots, the ACF, and the KPSS test. The ACF plot has only positive spikes, and most of the spikes are outside the confidence band, meaning there is positive correlation between the current seasonally adjusted housing market transaction volumes and lagged seasonally adjusted transaction volumes. The KPSS test gives a p value of 0.01. Using the significance value of 0.05, we reject the null hypothesis that there is stationarity in the data. Differencing was applied where necessary to achieve stationarity, with a KPSS test p value of 0.1.The line plot of first differences of log A is shown in figure 5. The ACF plot of first differences of seasonally

adjusted house transaction volumes, shown in figure 6, still has some spikes outside the confidence band: the spike around lag 1 has the biggest negative spike, and there is the biggest positive spike at lag 6. There are roughly even numbers of positive spikes and negative spikes.

An ARIMA (0,1,2) model was selected as the best model with the least AICc value by R to fit the LogA series, shown in figure 7. The variance estimated in the model is 0.018. An ARIMA (0,0,0) (0,1,0) model was fitted to the seasonal component, shown in figure 9, with estimated variance of 0.00011.

To check normal distribution residuals, we applied the gg_residual method on the LogA ARIMA model and the ad test on the residuals. According to the histogram of residuals, there is a roughly normal distribution with a few outliers to the left. The p value of 0.143 in the ad test is bigger than the significance level of 0.05. As a result, we failed to reject the null hypothesis that the residuals are normally distributed (shown in figure 10).

Using these ARIMA models we forecasted the next four months for both the seasonal and adjusted parts, combined the forecasts, and exponentiated the results to return to the original scale. 95% prediction intervals were calculated by considering the variances and covariances between the two forecast components. The forecasts and prediction intervals are visualized alongside historical sales volumes for easier interpretation, shown in figure 11 and 12, with figure 12 providing more emphasis on the forecast visualization.

Note: all modeling, analysis,etc. were performed using R, through fable, fpp3, and forecast packages.

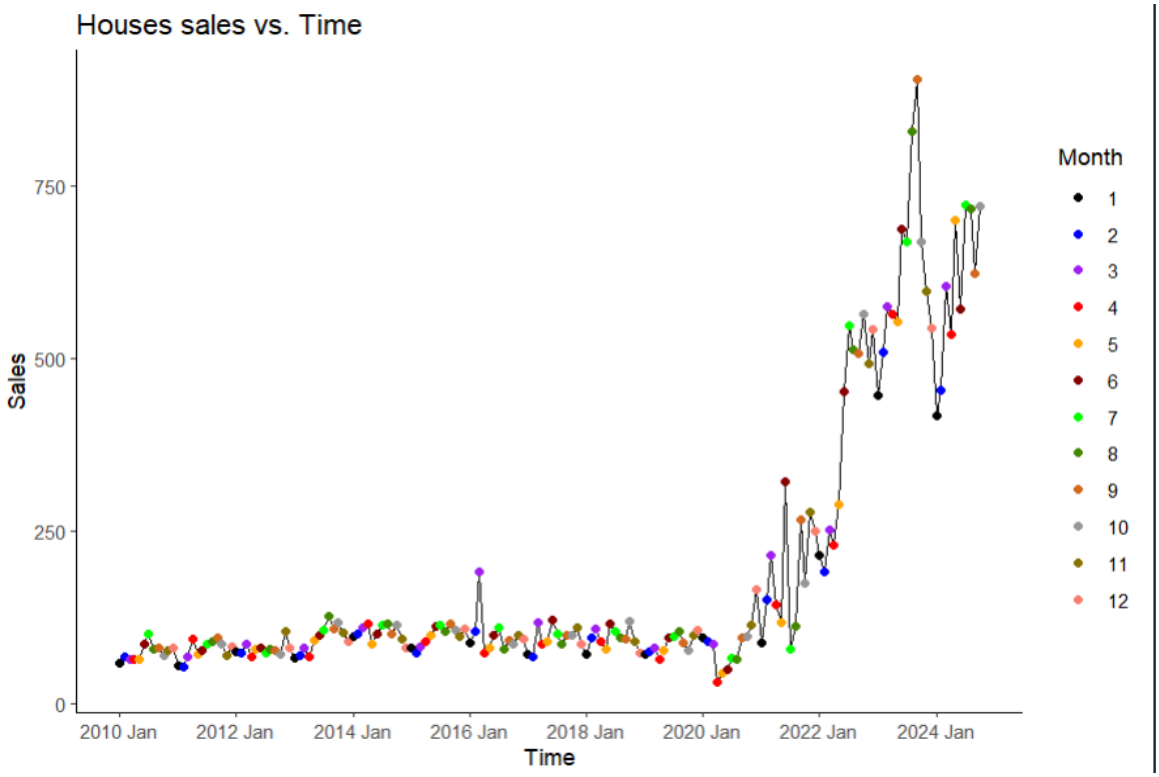
Discussion:

Our analysis shows that UK house sales volume exhibits strong seasonal behavior with clear annual cycles. The seasonal adjustment and decomposition methods effectively isolate the seasonality. This allows us to model the underlying trends separately from recurring seasonal effects.

The fitted ARIMA models displayed white noise residuals and passed normality and stationarity tests, confirming our model choices. Forecasts for the next four months suggest that the house sales volume will remain relatively stable, with slight seasonal fluctuations. The width of the prediction intervals reflect both the variability inherent in the housing market and the uncertainty introduced by recent disruptions such as the pandemic.

Our results provide a valuable tool for individuals considering buying or selling property, as well as for developers planning construction projects. However, we are not able to clearly answer why the house transaction volume peaked after the pandemic period. Due to brief research, the peak in transaction volume after the pandemic period might be resulted by higher demand for houses in the UK, overall more flexible working mode (more companies offer remote working options) offered to workers. While our models performed well on the available data, we must acknowledge that there are many limitations, and external factors can make these models unreliable.

Appendix: Figure 1, Figure 2:



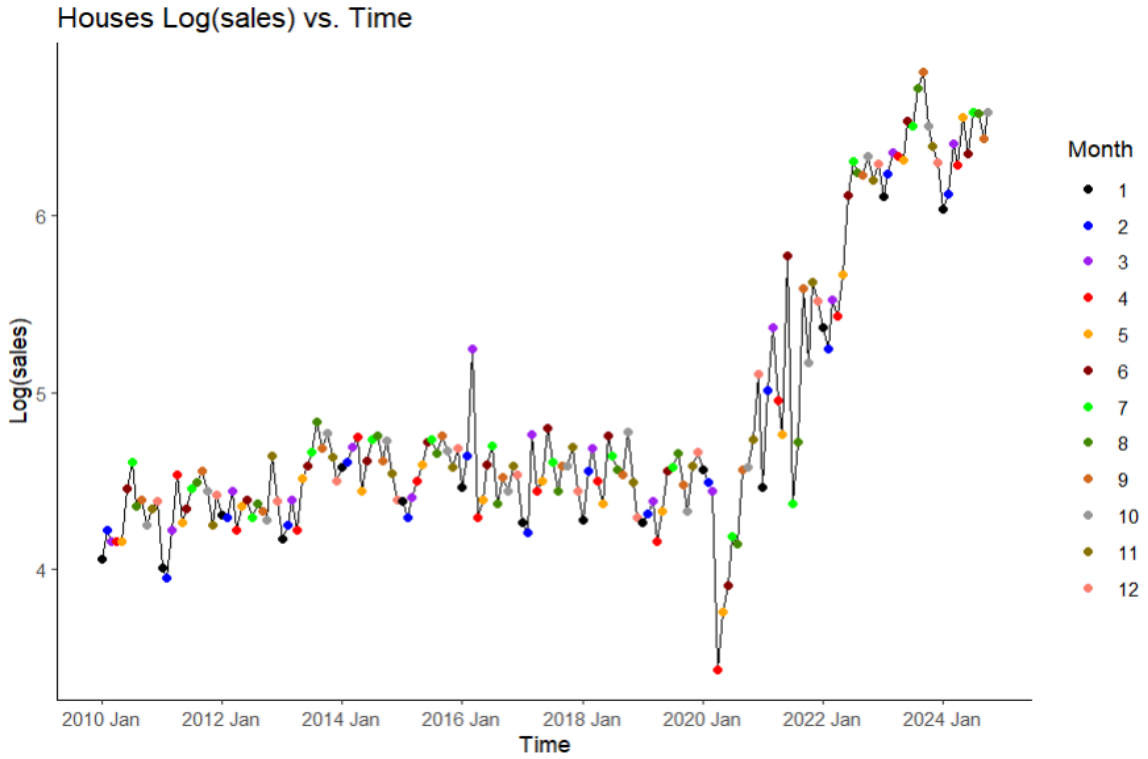
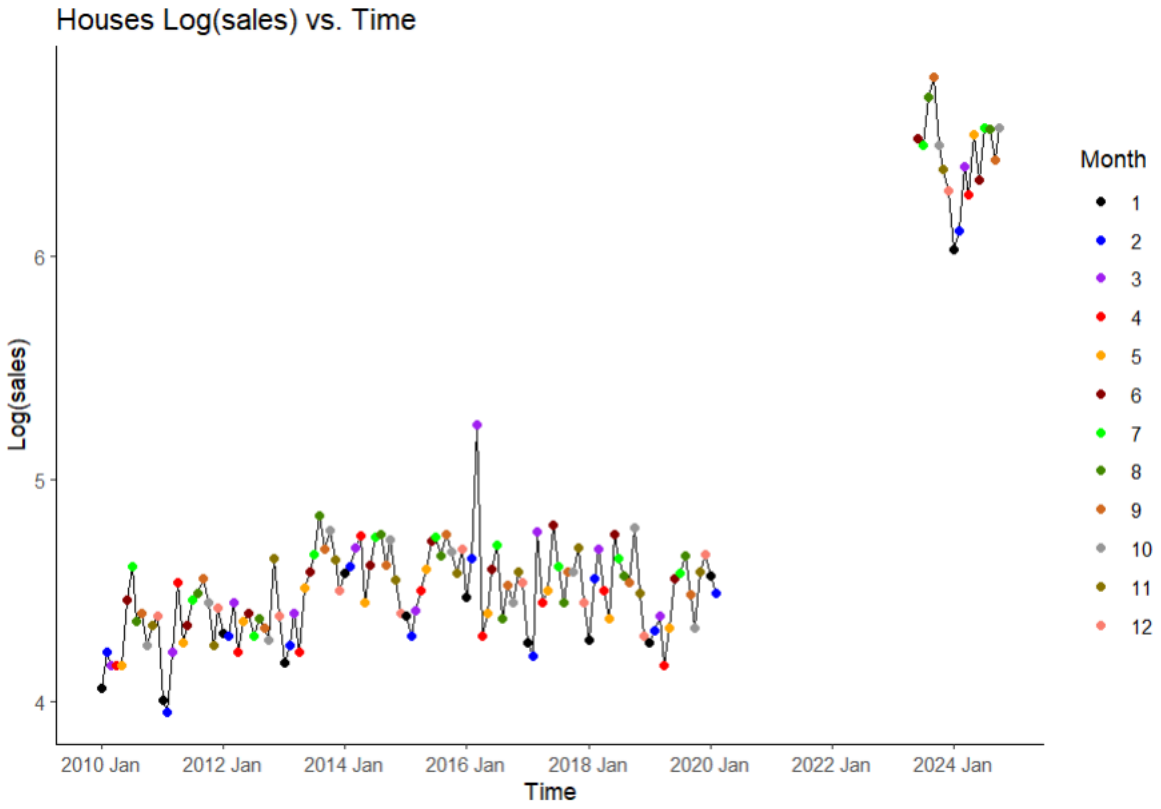


Figure 3, Figure 4:



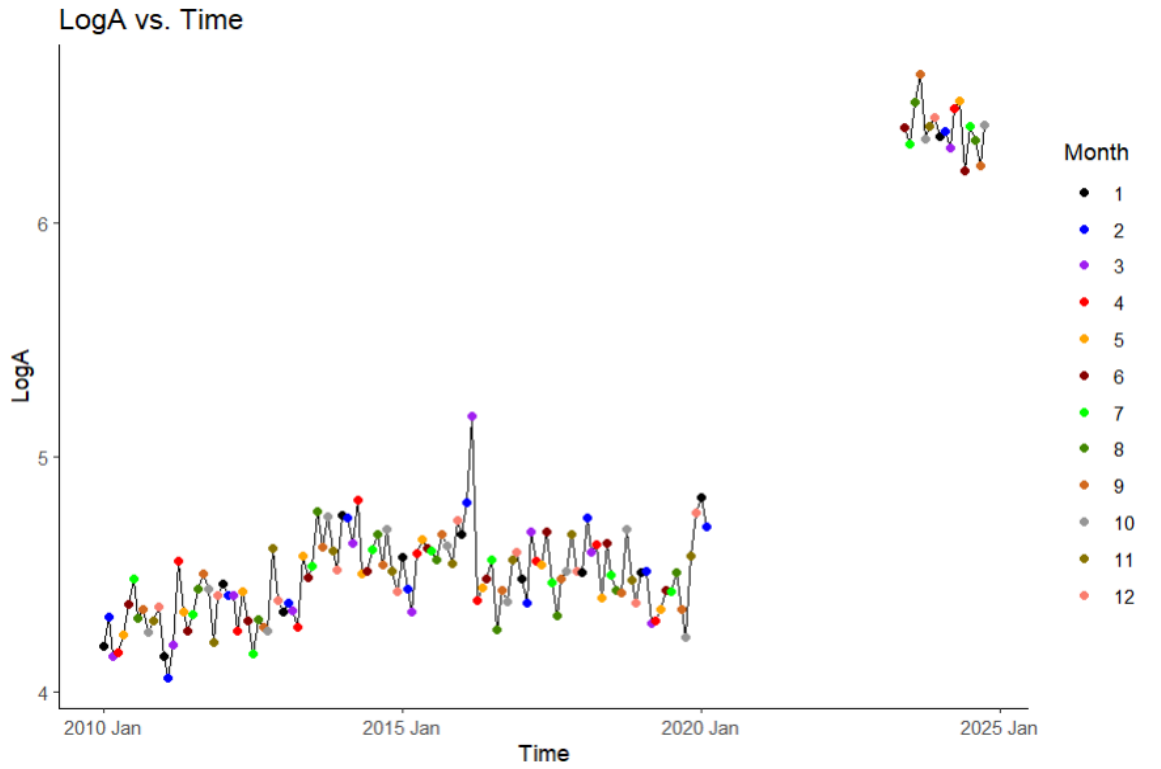
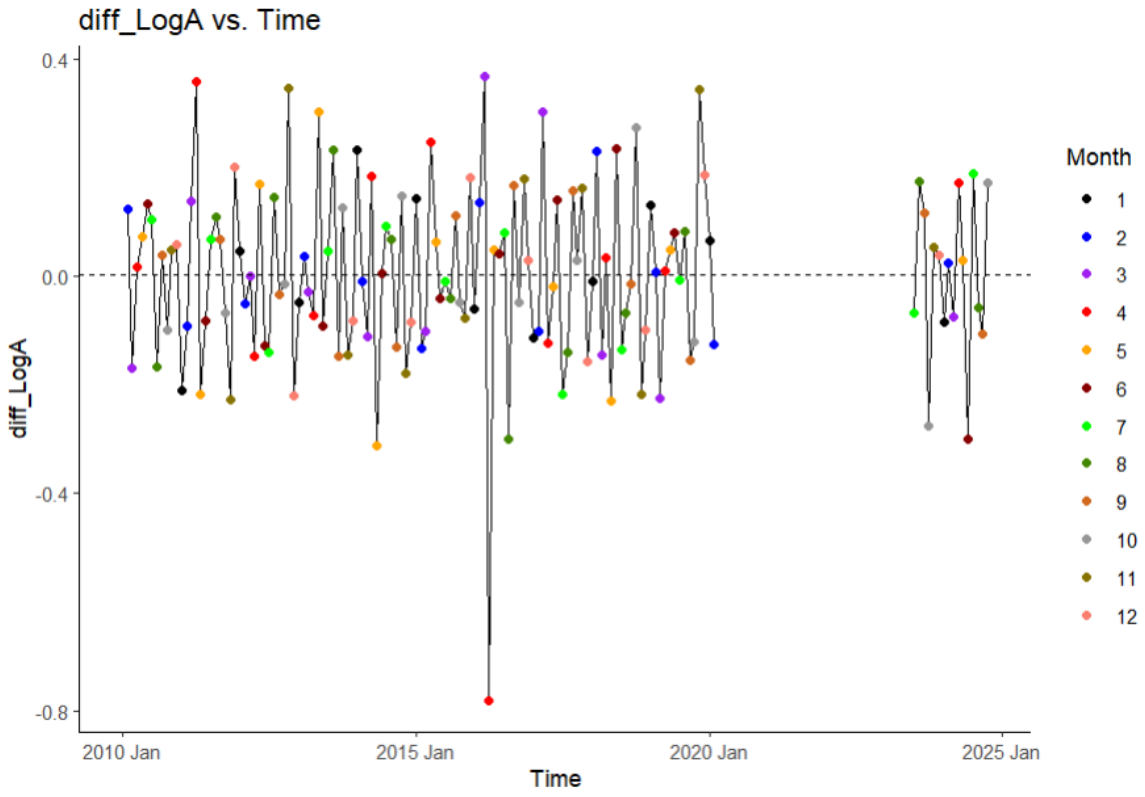
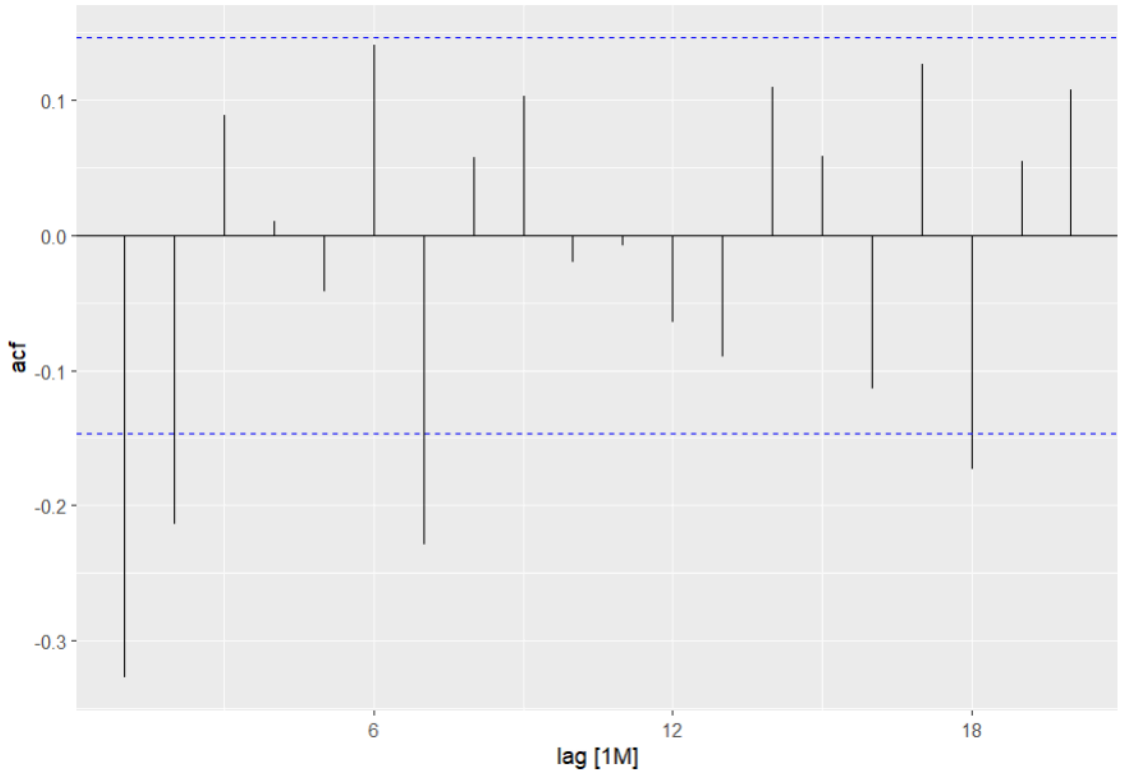


Figure 5, 6


diff_LogA vs. Time
0.4
0.0
-0.4
-0.8
diff_LogA
2010 Jan
2015 Jan
2020 Jan
2025 Jan
Time
Month
1
2
3
4
5
6
7
8
9
10
11
12


0.1
0.0
-0.1
-0.2
-0.3
acf
6
12
18
lag [1M]

```
> report(result_ARIMA_LogA)
Series: LogA
Model: ARIMA(0,1,2)

Coefficients:
          ma1      ma2
      -0.4235  -0.0683
s.e.   0.1012   0.0954

sigma^2 estimated as 0.01771:  log likelihood=64.68
AIC=-123.35   AICc=-123.21   BIC=-113.82
```

Figure 7,8:

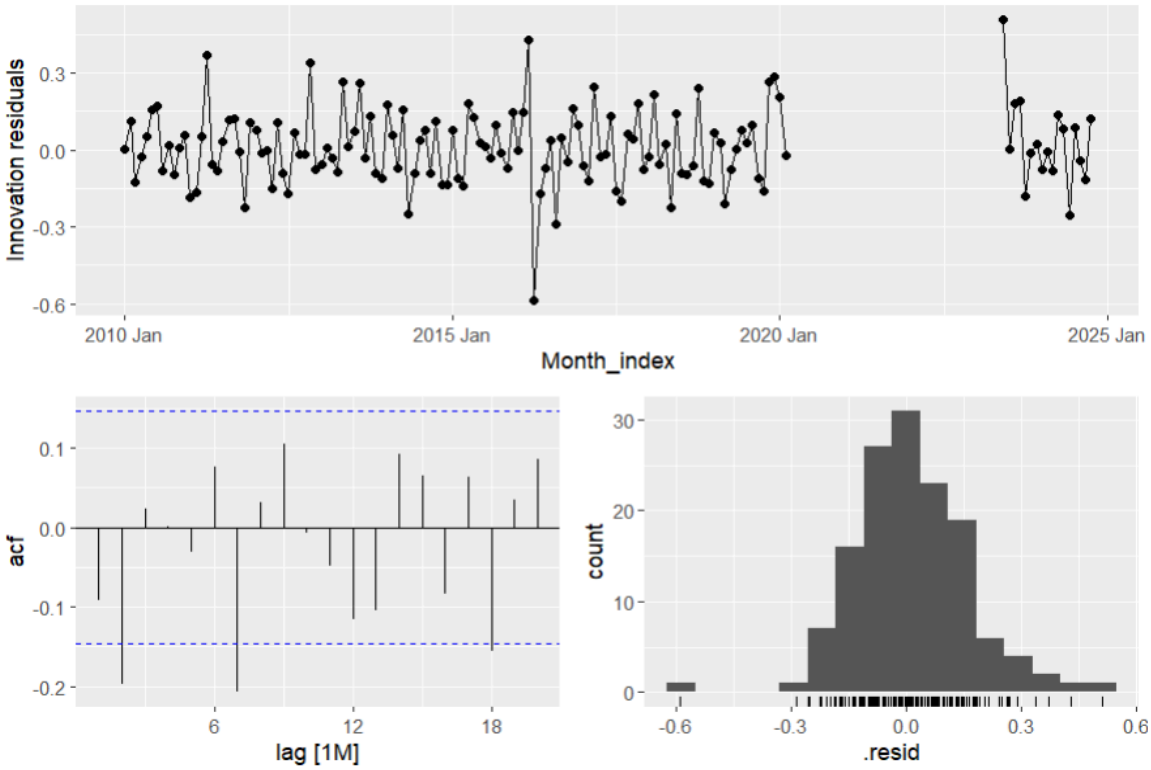
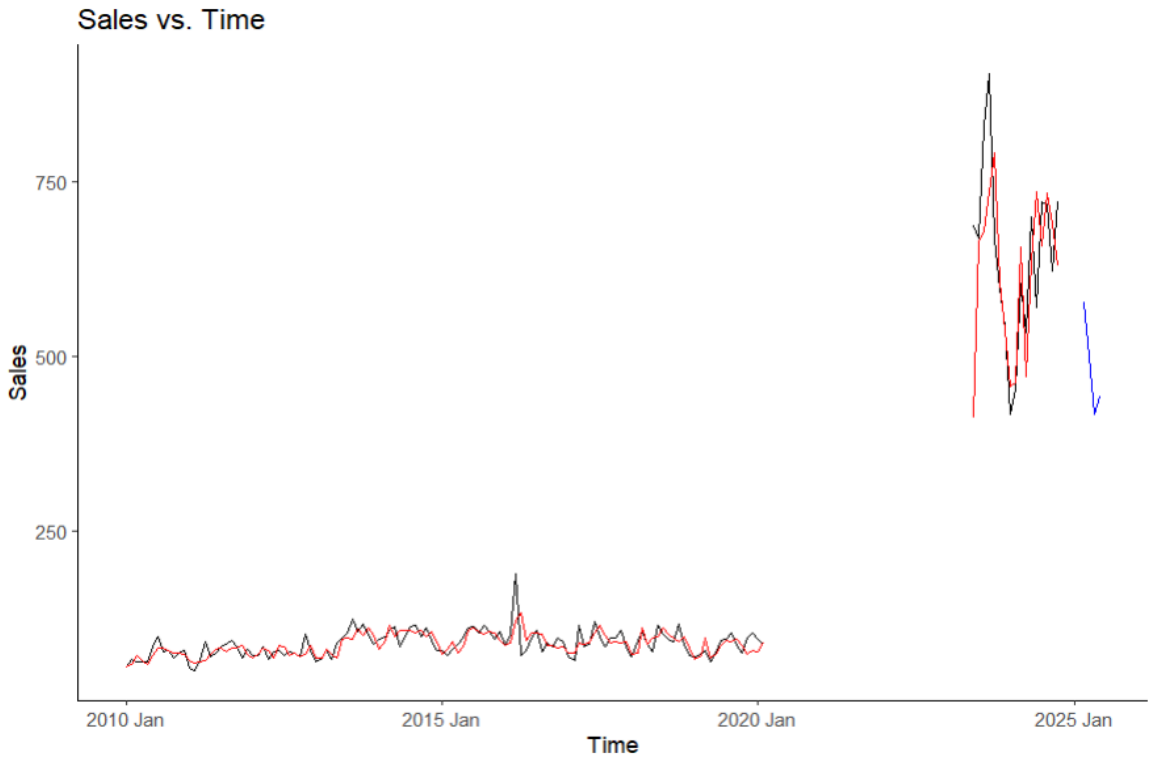


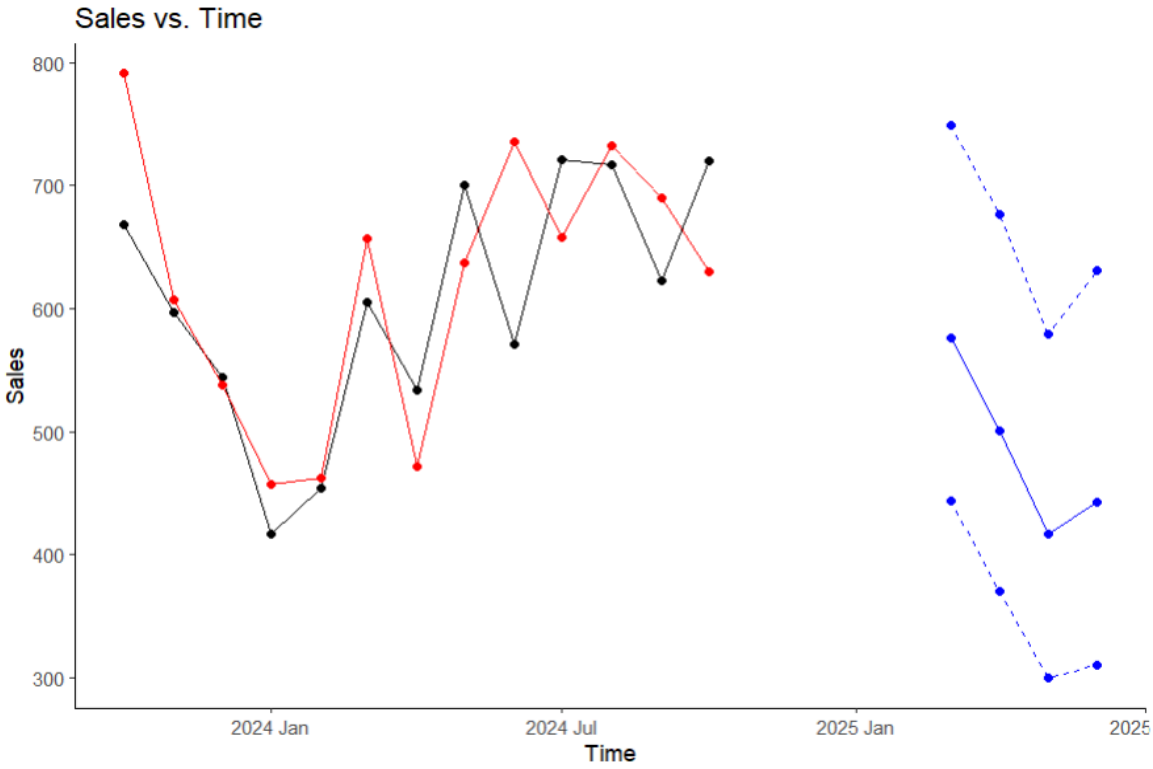
Figure 9,10:

```
> report(result_ARIMA_season)
Series: season
Model: ARIMA(0,0,0)(0,1,0)[12]

sigma^2 estimated as 0.000114:  log likelihood=518.05
AIC=-1034.11   AICc=-1034.09   BIC=-1031
```

```
> ad.test(result_ARIMA_LogA_augment$.resid)

        Anderson-Darling normality test

data:  result_ARIMA_LogA_augment$.resid
A = 0.56289, p-value = 0.1428
```

Figure 11, 12:


Sales vs. Time
750
500
250
Sales
2010 Jan
2015 Jan
2020 Jan
2025 Jan
Time


Sales vs. Time
800
700
600
500
400
300
Sales
2024 Jan
2024 Jul
2025 Jan
2025
Time