Time Series and Business Data

Time Series Group Project (Group 4)

Group members: Radu Miu, Yeuk Lan Ng

Dataset: U.S. Air Traffic Data (2003-2023)
https://www.kaggle.com/datasets/yyxian/u-s-airline-traffic-data

Data Summary:
This dataset contains monthly airline traffic data for the United States from 2003 to 2023, including Total Air Travel Passengers(Pax), flights, revenue passenger miles (RPM), available seat miles (ASM), and load factor.
Time series analysis is a statistical method for analyzing data points gathered or recorded at predetermined time intervals. This strategy can reveal trends, patterns, and other important information about the data as it changes over time.
In the context of airline traffic statistics, time series analysis can be used to forecast future air passenger numbers using previous data, assisting with capacity planning, pricing strategies, and other commercial choices.

Our Approaches:

1. Data cleaning and preparation
2. Description of the trend
    - Time series plot for total passengers over time
    - Seasonal plot with different month over time
3. Stationarity Transformation
    - First transformation: log
    - Second transformation: 1st order difference
    - Third transformation: difference of order 12
4. Box-Jenkins methodology and ARMA
    - Model building and analysis
    - Residual diagnostic
5. Model Validation and Forecast

Data cleaning and preparation

After loading the dataset(air_traffic), we checked for N.A. values and duplicates, this dataset does not contain any of them. Then, we created a new data column(date) and got rid of the comma in the 'Pax' column for a smoother data processing later. Beside the usual practice, we designed a series of functions to integrate the coding process.

Description of the trend

- Time series plot for total passengers over time

We created the plot to observe the overall trend of total number of passengers from 2003 to 2023.

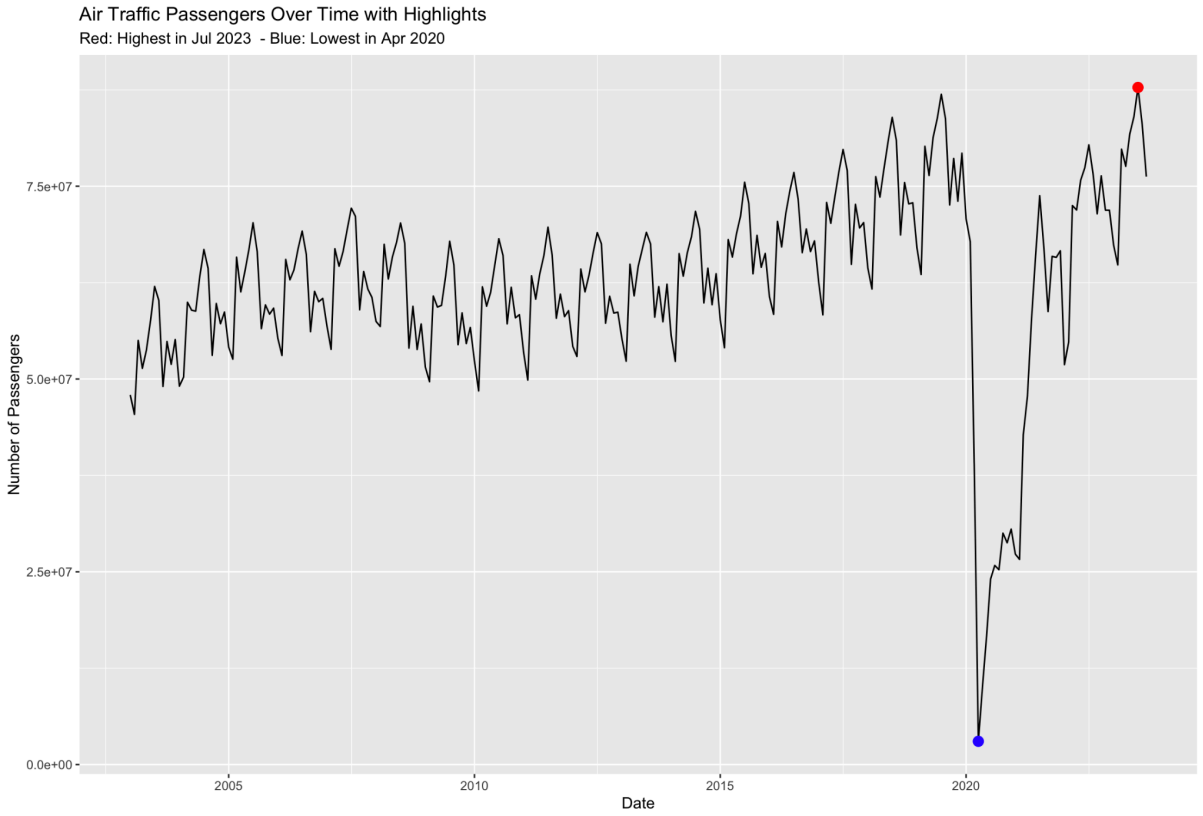

Figure 1: Air Traffic Passengers Over Time with Highlights

We could find that the trend shows a seasonality trend with similar patterns observed over time. Following a dramatic decrease in 04/2020 (total traffic 30,138,99) which gradually reached to the highest in 07/2023 with 878,107,72 total traffic.

- Seasonal plot with different month over time

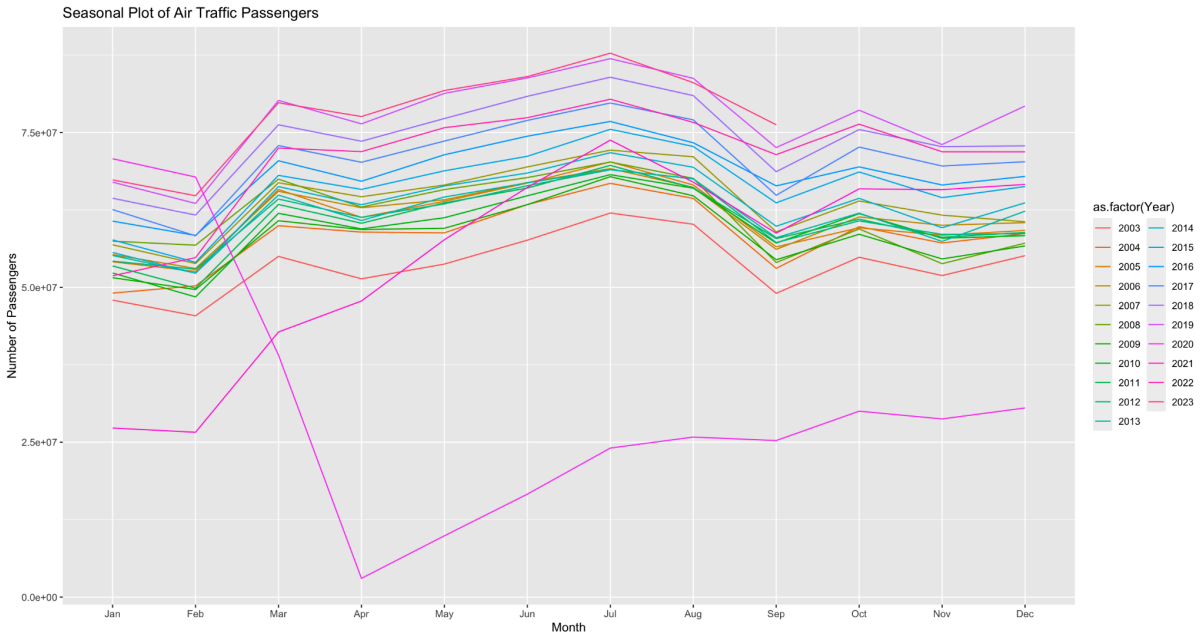

Figure 2: Air Traffic Passengers Seasonal plot

We detected the seasonality among months in different years. In March, June to August, October, and December are the peak traffic months. Except 2020 when there was a sharp decline from February. We suppose the seasonality is due to the impact of COVID-19 on the aviation industry during 2020. And for the months previously mentioned, we assumed that these are the periods of holiday so a lot of traffic was recorded.

Stationarity Transformation

- Original data shows non-stationarity and seasonality, especially in 2020, due to the Covid-19 influence. Significance occurred on lag 1, 7, 13 on PACF plot while there's ACF plot shows no significant decay to 0.

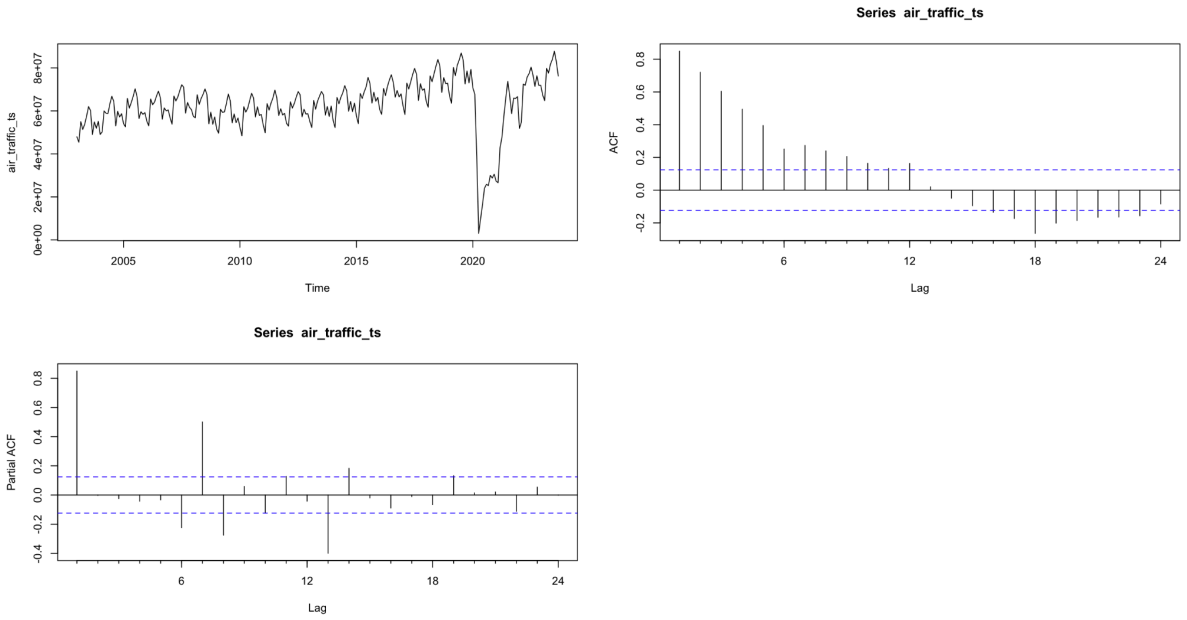

Figure 3: Air Traffic Passengers Time Series plot with ACF and PCF

- First transformation: log

  The transformed time series with log is not stationary still, with no constant trend following ACF not decaying to 0.

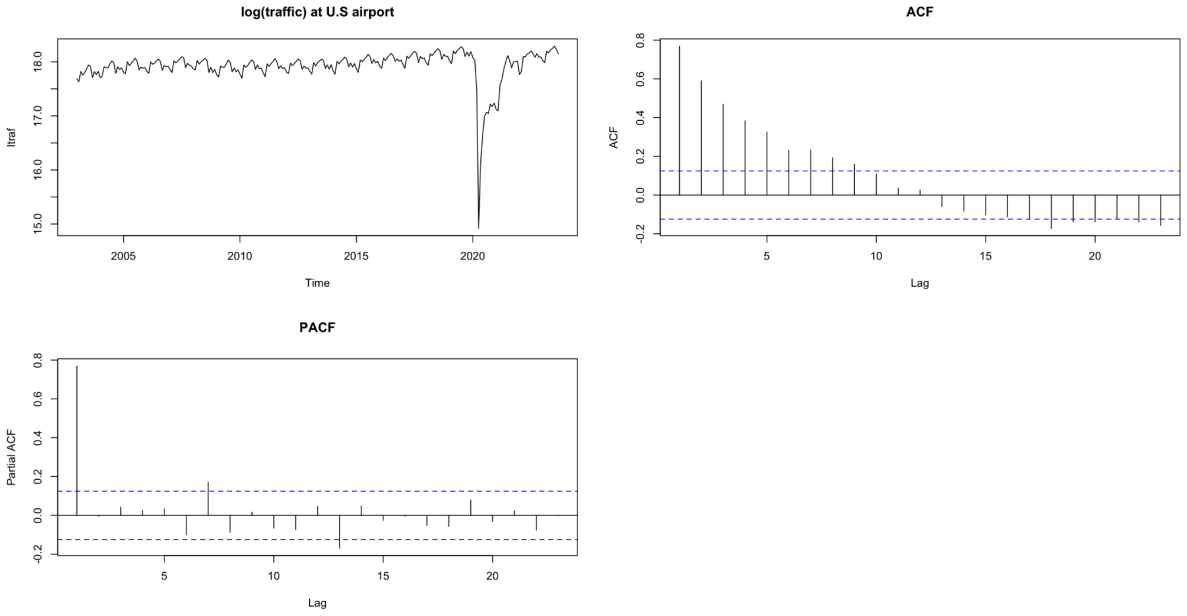

Figure 4: log(traffic) Air Traffic Passengers Time Series plot with ACF and PCF

- Second transformation: 1st order difference to remove trend, there's significant coefficient at lag 6 of PACF plot and ACF has no sign to decay to 0. The overall trend is not constant.

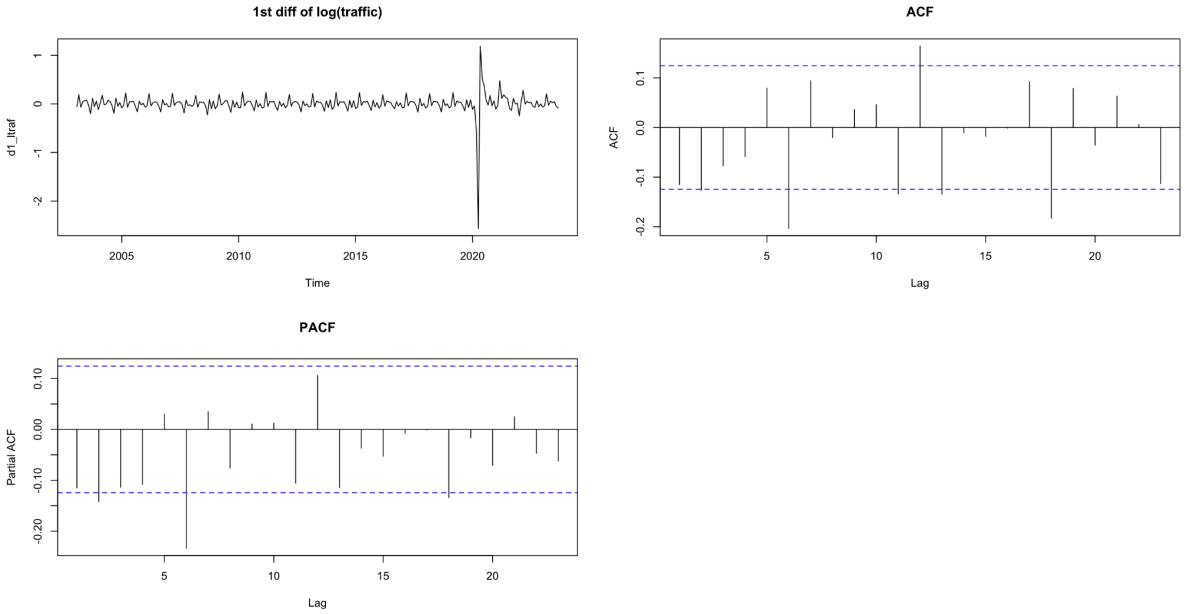

Figure 5: log1(traffic) Air Traffic Passengers Time Series plot with ACF and PCF

- Third transformation: difference of order 12 to remove seasonality. The seasonality is removed before 2020. ACF and PACF both have significant coefficients at lag 12.

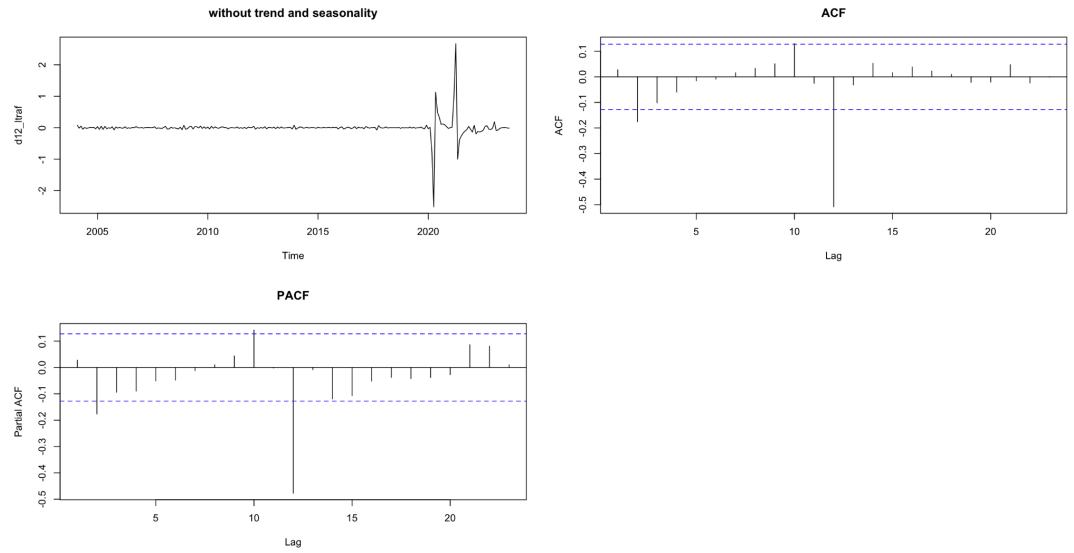

Figure 6: log12(traffic) Air Traffic Passengers Time Series plot with ACF and PCF

Box-Jenkins methodology and ARMA

Model building and analysis: based on the total 6 models we built, here we presented model 5 and 6 that outperformed the rest of them.

- Model 5

  Model 5 was built using the original time series with the following parameters: SARIMA (1, 1, 2) (0, 1, 1).

```
Call:
stats::arima(x = air_traffic_ts, order = c(0, 1, 1), seasonal = list(order = c(2,
    1, 3), period = 12), method = "ML")

Coefficients:
         ma1     sar1    sar2     sma1     sma2    sma3
      0.3737  -0.4382  0.4614  -0.5731  -0.7973  0.4985
s.e.  0.0591   1.4004  1.1243   1.4591   0.4327  1.0976

sigma^2 estimated as 1.413e+13:  log likelihood = -3918.04,  aic = 7850.07

SBC value is :
7866.462

Pvalues are :
         ma1          sar1         sar2         sma1         sma2         sma3
2.486069e-10 7.543306e-01 6.814966e-01 6.944984e-01 6.534452e-02 6.496892e-01
```

Figure 7: Model analysis

Even though all the p values are significant, and the model seems to fit the data well and we don't have useless parameters, the AIC and the SBC are high, about 7800 for both. We think this is due to the sharp drop in the timeseries caused by the pandemic. The AIC and SBC are better than others for model 5. Even so, we can still feel the impact of the drop on those indicators. (cf figure7)

```
> residual_diagnostic(mod5)

        Ljung-Box test

data:  Residuals from ARIMA(0,1,1)(2,1,3)[12]
Q* = 30.498, df = 18, p-value = 0.03288

Model df: 6.   Total lags used: 24


        Shapiro-Wilk normality test

data:  res
W = 0.57782, p-value < 2.2e-16
```

Continuing with the residual analysis, they follow a normal distribution and are not significant. However, the p value of the Ljung-Box test is slightly under 0.05 which means that the non-autocorrelation hypothesis can be rejected.

As for the homoscedasticity of the residuals we used the McLeod-Li test. For lag 12 the p value is bigger than 0.05, thus we can accept the homoscedasticity hypothesis. (cf figure 8)

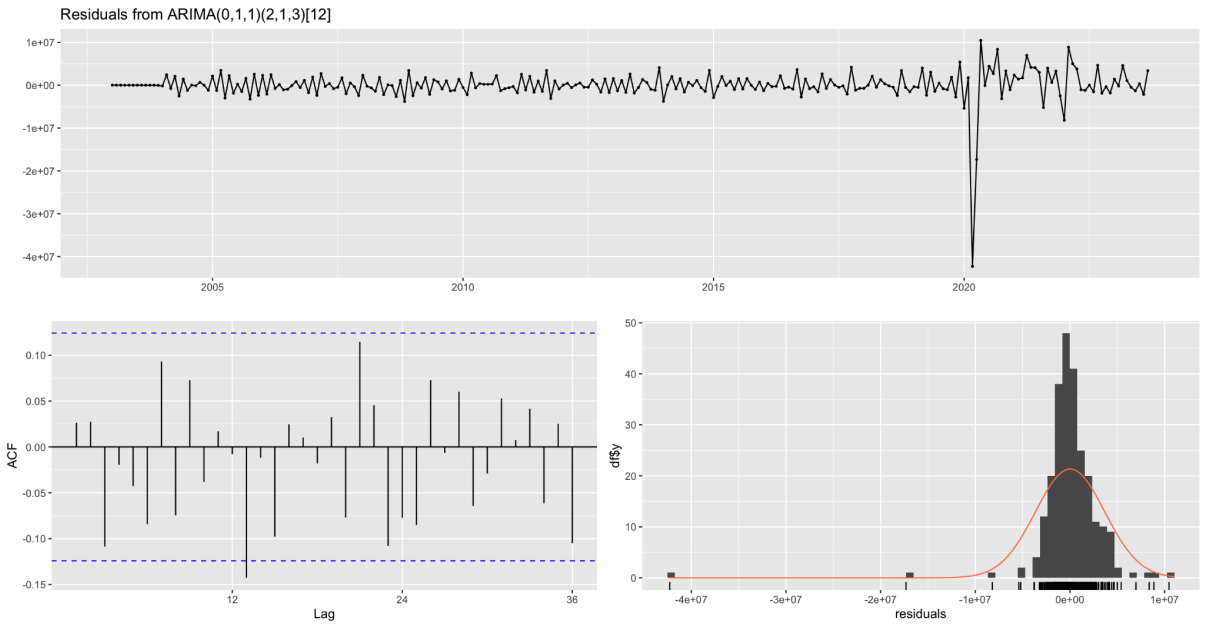

Figure 8: Model 5 plot

McLeod-Li test plot for model 5:

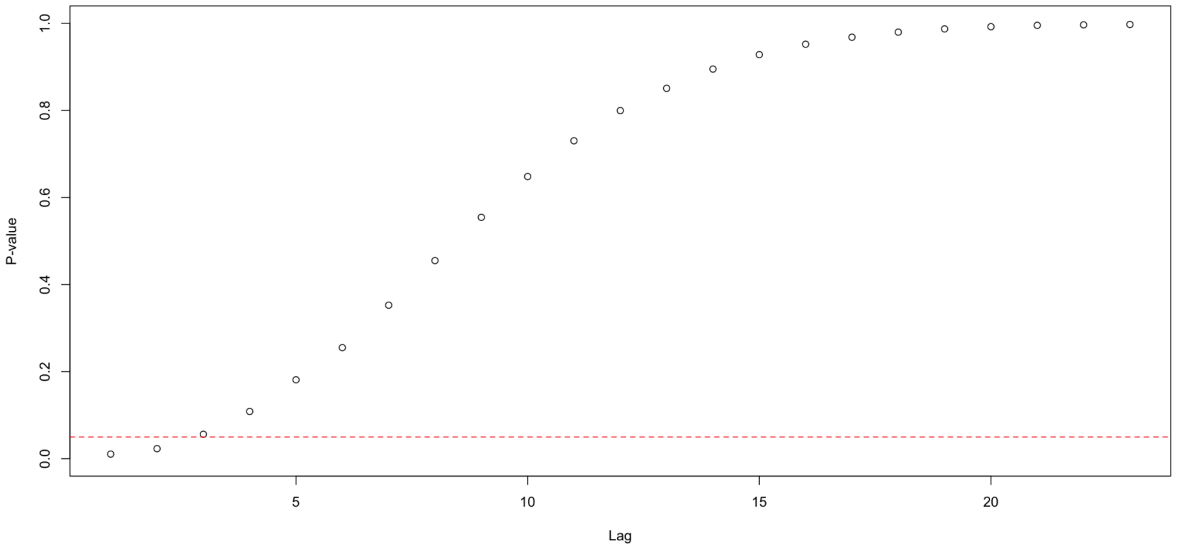

Figure 9: Model 5 McLeod-Li test plot

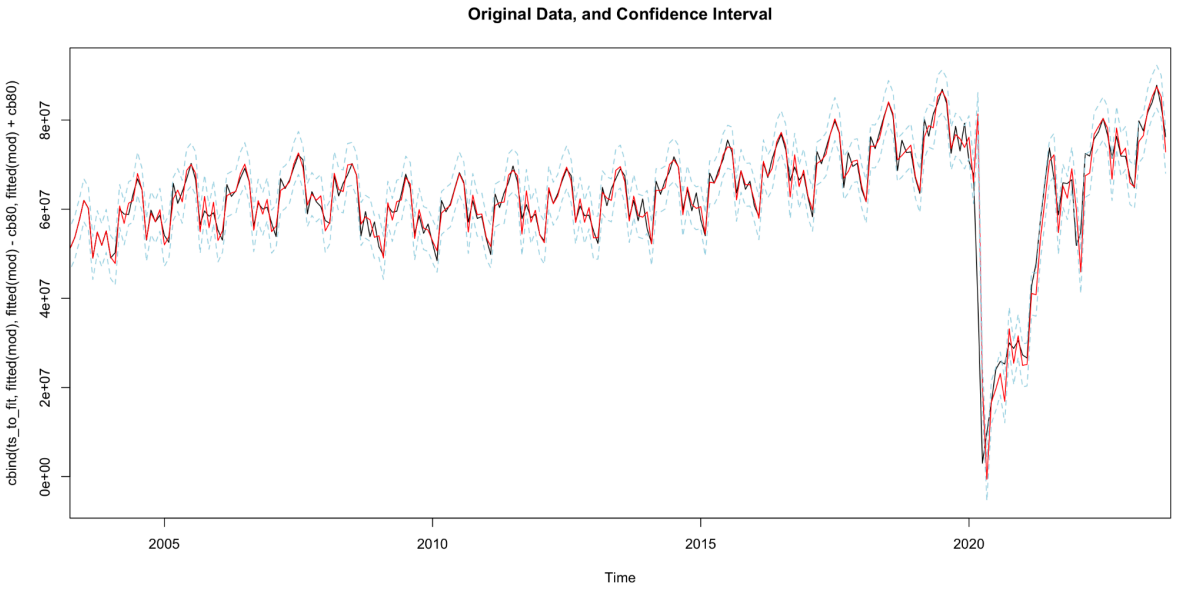

Figure 10: Model 5 confidence interval compared to the time series

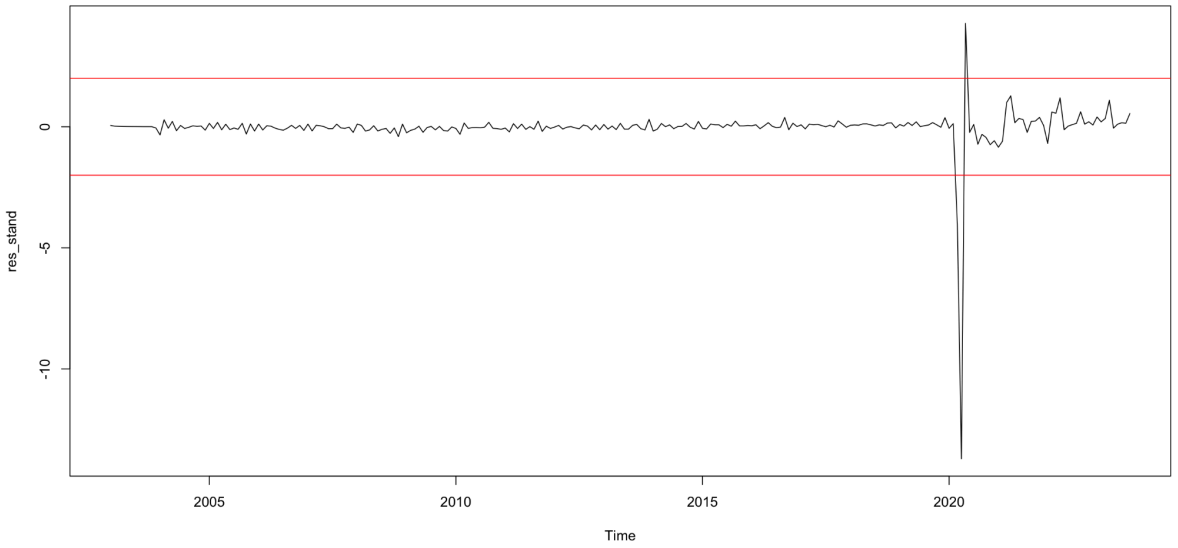

Figure 11: Model 5 residual plot

- Model 6 analysis

On the other hand, model 6 was built using the log of the original time series. This transformation is not without importance, as this will diminish the impact of the drop.

```
Call:
stats::arima(x = ltraf, order = c(1, 1, 2), seasonal = list(order = c(0, 1,
    1), period = 12), method = "ML")

Coefficients:
         ar1      ma1      ma2     sma1
      0.7772  -0.8753  -0.1247  -0.9448
s.e.  0.0529   0.0887   0.0849   0.0771

sigma^2 estimated as 0.03394:  log likelihood = 48.11,  aic = -86.22

SBC value is :
-75.28982

Pvalues are :
     ar1      ma1      ma2     sma1
0.000000 0.000000 0.142113 0.000000
```

Figure 12: Model 6 analysis

Indeed, as we can see the AIC and SBC for model 6 are smaller, -86 for AIC and -75 for SBC.

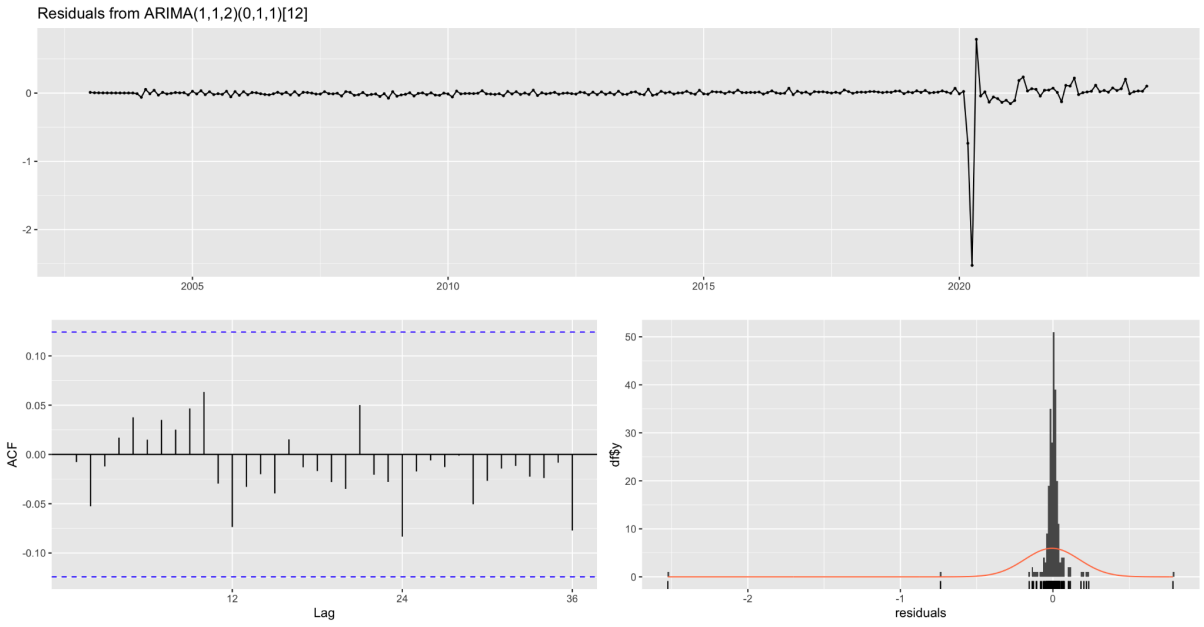

Figure 13: Model 6 plot

Moreover as we can see in the residual analysis plot, they follow a normal distribution and are not significant.

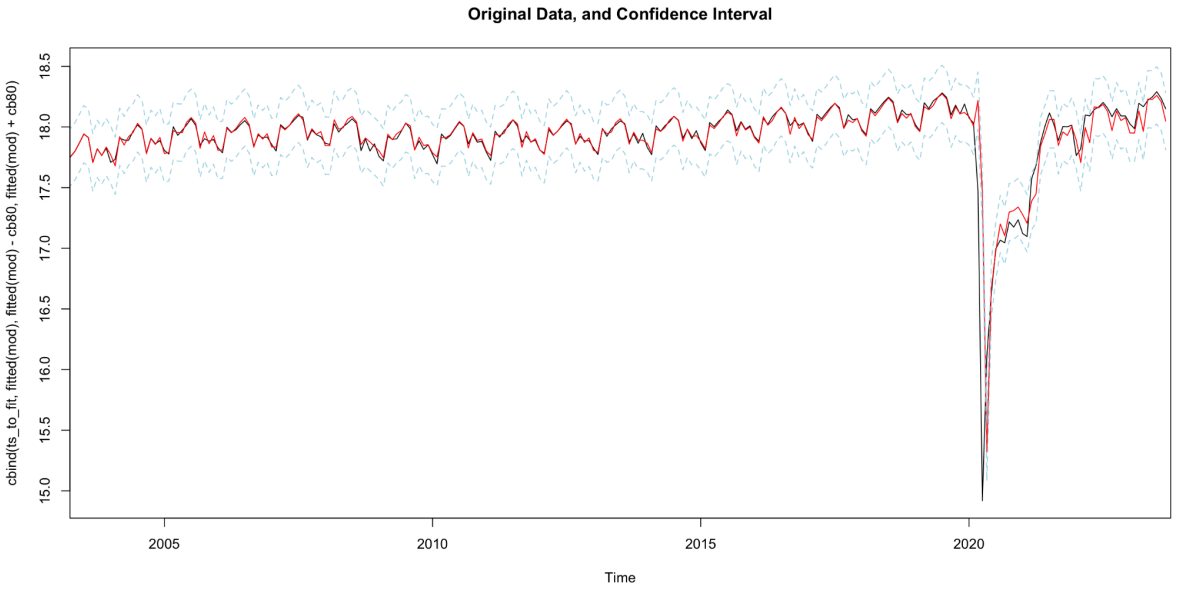

Figure 14: Model 6 confidence interval compared to the time series

The residuals are following a normal distribution and are not significant (cf plots and Shapiro-test).

```
        Ljung-Box test

data:  Residuals from ARIMA(1,1,2)(0,1,1)[12]
Q* = 9.4955, df = 20, p-value = 0.9764

Model df: 4.   Total lags used: 24


        Shapiro-Wilk normality test

data:  res
W = 0.22337, p-value < 2.2e-16
```

The p value of the Ljung-Box test is close to one thus we can accept the non-autocorrelation hypothesis.

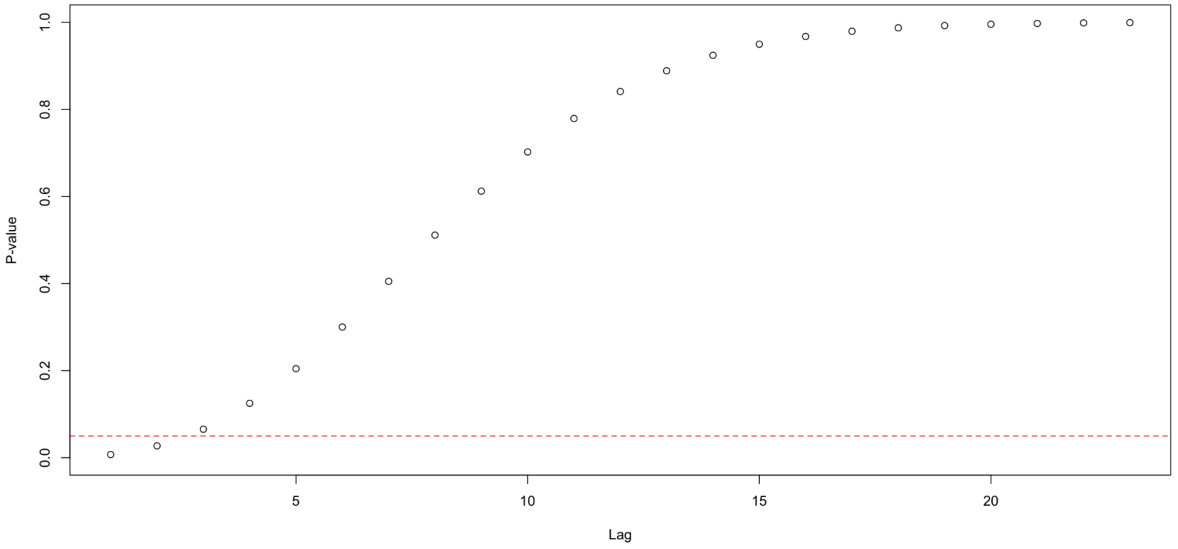

Figure 15: Model 6 McLeod-Li test plot

As for the homoscedasticity of the residuals we used the McLeod-Li test. For lag 12 the p value is bigger than 0.05, thus we can accept the homoscedasticity hypothesis.

In short, the two models' plots are very similar, with model 5 having a tighter confidence interval. Thus, we can theorize that the metrics of model 5 are so bad because model 5 wasn't built using the log of the dataset, thus errors are amplified compared to model 6 who was built using the log of the dataset.

Model Validation and Forecast

In total we built 6 models however we only presented the last 2 which achieved the best results so we trained our forecast based on them.

We split the data in the following way: the first 200 observations are in the training set and the remaining observations are in the test set. These are further validated with an in-sample and out-of-sample analysis. Based solely on the numeric indicators model 6 is a lot better than model 5. However, the plot of the fitted values yields interesting insights.

- Models prediction shows model 6 has higher accuracy compared to model 5 with lower RMSE and MAE.

Figure 16: Model 5 prediction plot

Figure 17: Model 6 prediction plot

- Model 5 & 6 forecast for the next three periods

Model 5 and 6 forecasts for the next 3 years are very different, in comparison to the model validation plots. As we used models that were trained on the whole data for this forecasting, we can theorize that the divergence is due to the different way they process the and learn from the sharp drop in 2020. Indeed model 5 predicts a faster recovery compared to model 6.

In order to improve our predictive power and have models that converge towards more similar predictions we recommend diminishing the impact of the drop as it is due to an external impact (pandemic) that can't be predicted. Other metrics retaining information about the amount of risk we want to take in the future could be integrated in order to have more exhaustive predictions.

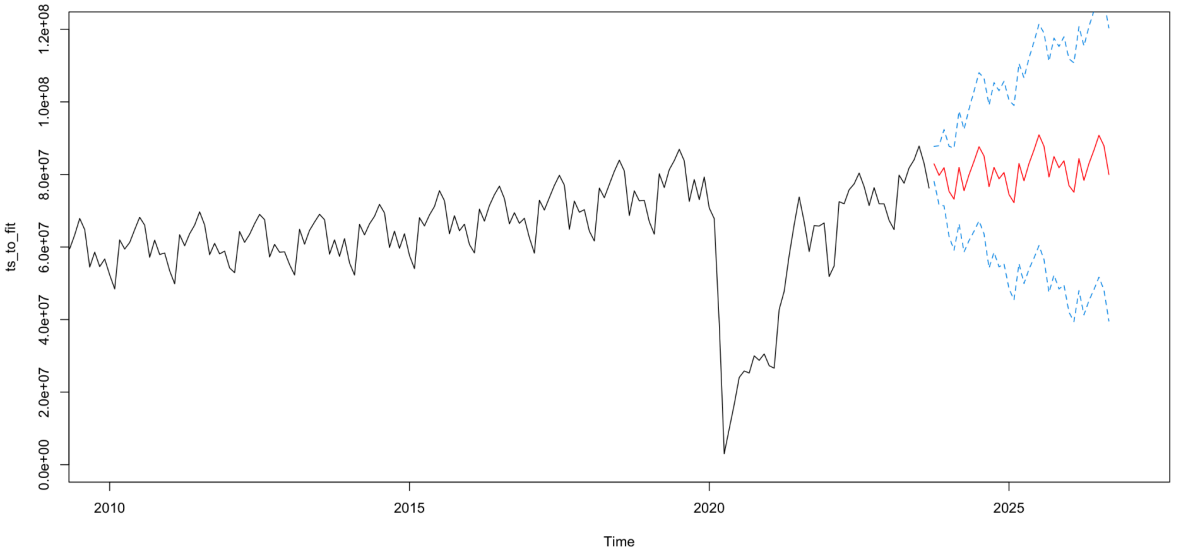

Figure 18: Model 5 forecast plot

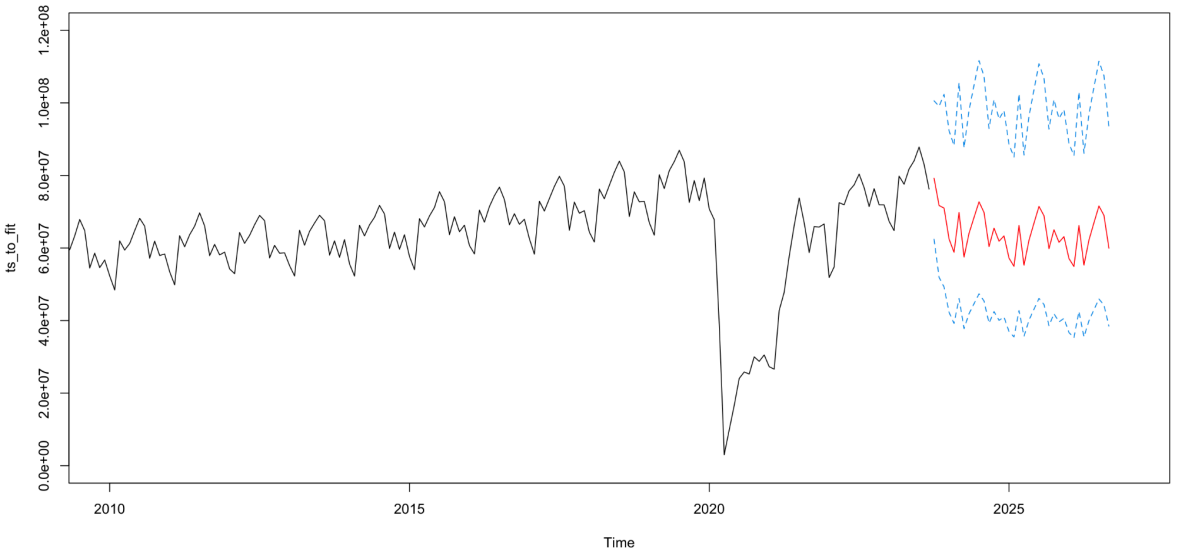

Figure 19: Model 6 forecast plot