Assignment Forecasting

By: Arunkkumar Karthikeyan

30th April, 2023

Introduction

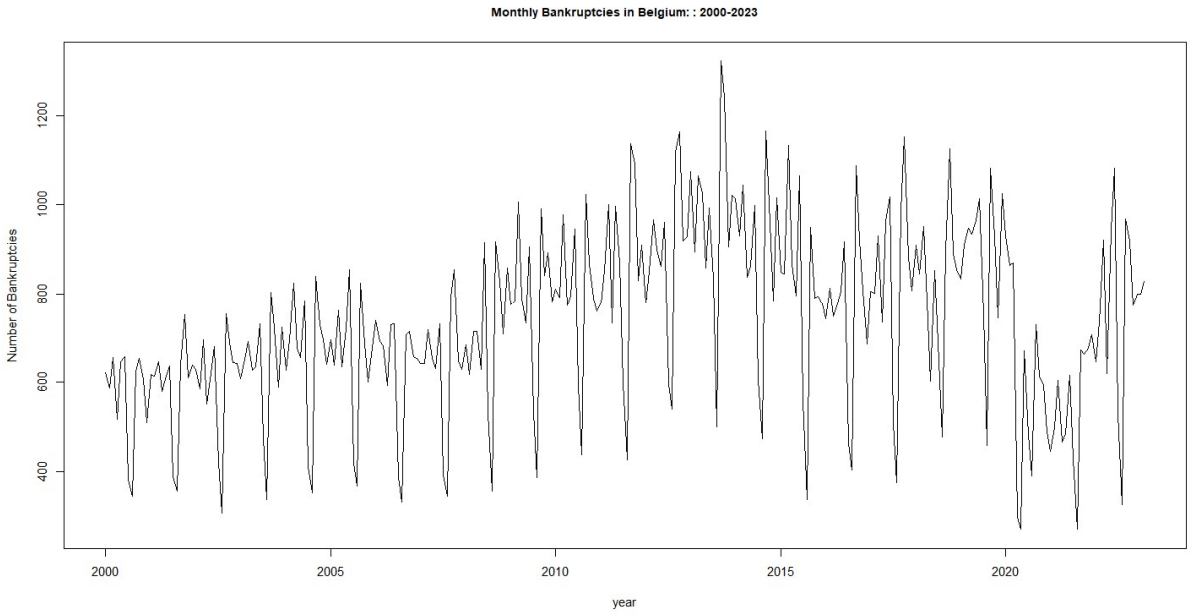
The objective of this paper is to apply the different concepts of forecasting in real datasets. For this purpose, this paper is dived in two parts, one using the dataset Bankruptcy and another one using the dataset unemployment rate in Colombia.

EXERCISE 01 : Time Series: Monthly Bankruptcy in Belgium : 2000-2023

The dataset contains monthly information about the bankruptcy in Belgium for all economic activities from 2020 to 2023. The objective of the monthly bankruptcy data for all economic activities in Belgium is to track the evolution of bankruptcy rates over time and provide insight into the overall health of the economy.

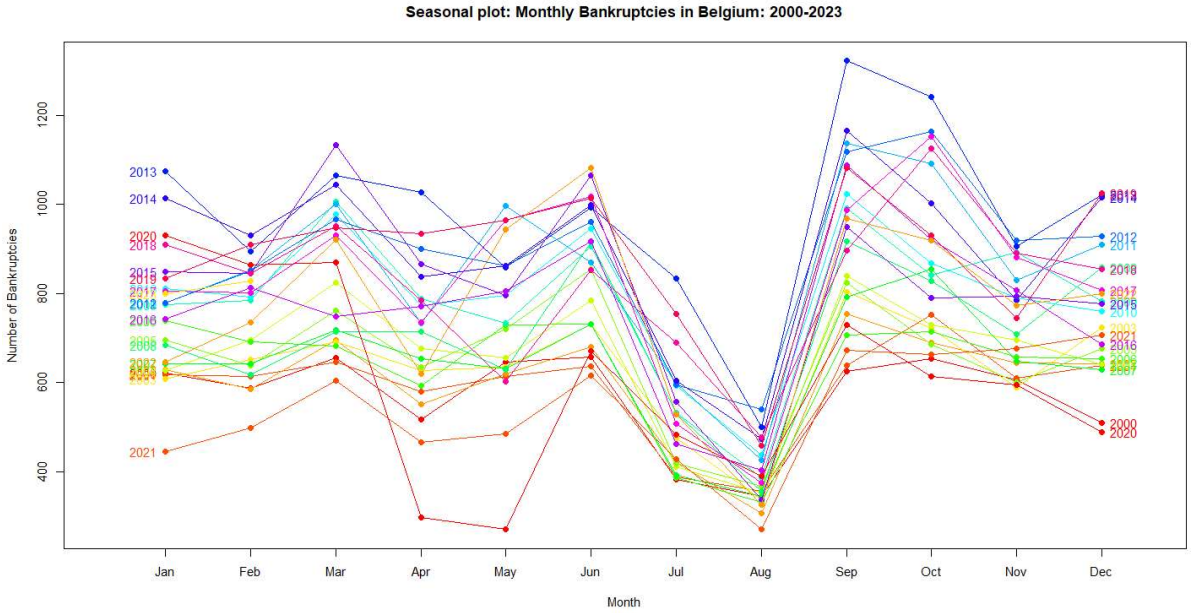
1. Data Exploration

The dataset contains monthly information of the bankruptcy from January 2000 to February 2023, the next figure shows the monthly bankruptcies in Belgium over the time:

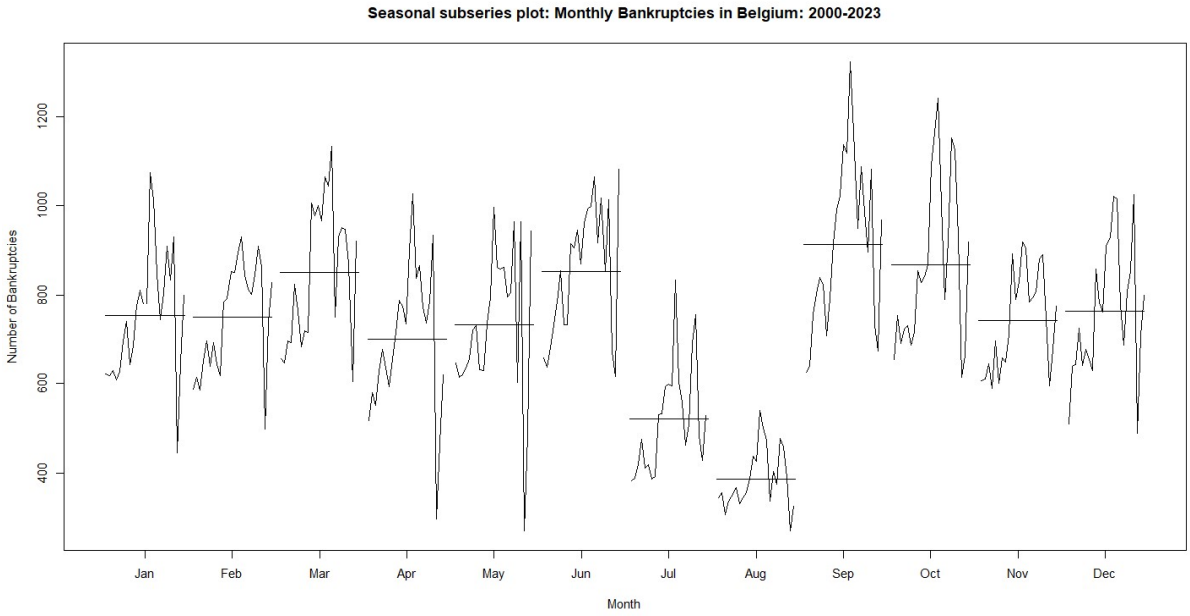


We can observe that the series has a seasonal pattern, which can be seen by the regularity of the data. The series also has a positive trend pattern, suggesting that the series is increasing over time. The slight variation in the series over time suggests that the series is composed of multiplicative components. Also, the data shows that the number of bankruptcies has been increasing over time, with a slight dip during the 2008 financial crisis. The highest number of bankruptcies occurred in September 2011, with 1138 bankruptcies, while the lowest number occurred in August 2013, with only 500 bankruptcies.

The following seasonal plots will be used to confirm the assumptions of seasonal and trend patterns:

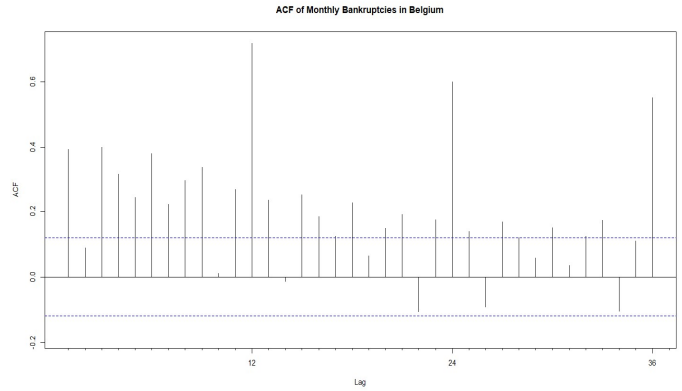


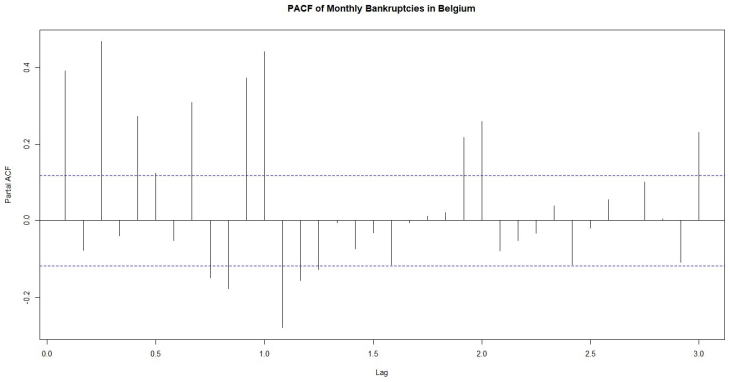
In this first seasonal plot, we get the information about the different years on top of the other. We can observe that the sub-series in some years are a little overlapped. However, it can be confirmed that there is a general increasing trend over time, the rate of increase is not consistent, and there are periods of fluctuation and even slight decreases. For example, The plot shows a clear pattern of seasonality with peaks occurring in the months of September and October, followed by a decline in the month of August. The number of bankruptcies in the summer months is generally lower than the other months. We can also see that the amplitude of the seasonal component is more or less constant except in 2000, 2012, and 2020.



Using the month seasonal plot, we can confirm the presence of an increasing trend. However, from the above plot we can see that there is no clear increasing trend visible in the plot, as the number of bankruptcies in each month appears to fluctuate around an average level, without showing a consistent upward or downward trend over time.

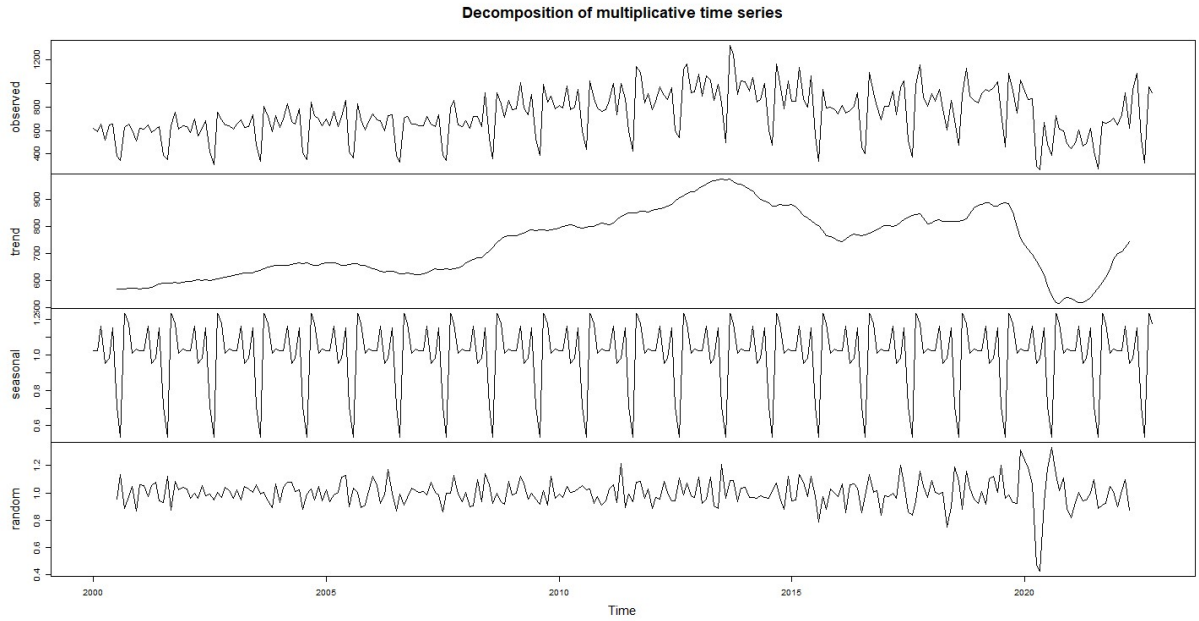
For supporting the assumptions of trend and seasonal pattern, we will look at the autocorrelation function of the time series:





The ACF plot shows that there is a significant positive autocorrelation lags 12, 24, 36, and so on, indicating a monthly seasonal effect and there is an increasing trend as well. Overall, the ACF plot suggests the presence of both trend and seasonal components in the Belgium bankruptcy data.

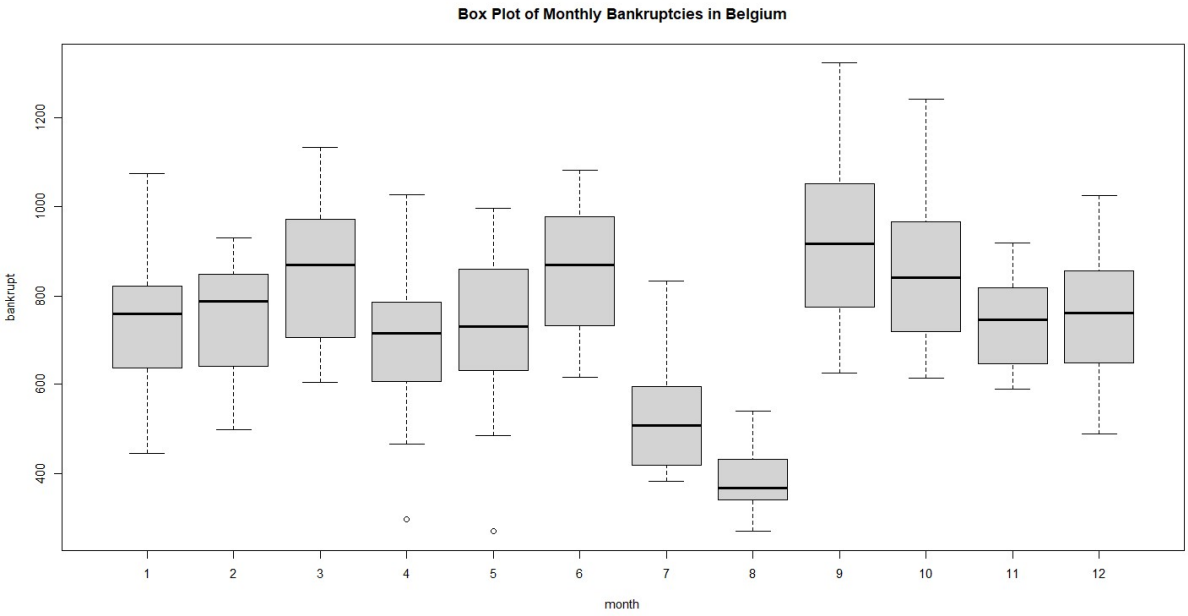
In the next graph, we can observe the seasonal and trend component of the time series, the trend component seems to increase slightly over the years till 2014 and it seems to decrease after that.



Finally, in order to describe the data we will get some summary statistics and the boxplot by month.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   270.0   617.0   732.5   735.9   868.8  1322.0
```

Looking at the boxplot we can observe that mean value each month is different and does not present remarkable differences. September is the month with the highest bankruptcy value



2. Transformations

I have performed the Augmented Dickey-Fuller (ADF) test for testing the null hypothesis of whether a time series is stationary or not. The ADF test is a statistical test that is commonly used to check for the presence of a unit root in a time series.

```
Augmented Dickey-Fuller Test
```

```
## data:  bankrupt
```

```
## Dickey-Fuller = -3.555, Lag order = 6, p-value = 0.03788
```

```
## alternative hypothesis: stationary
```

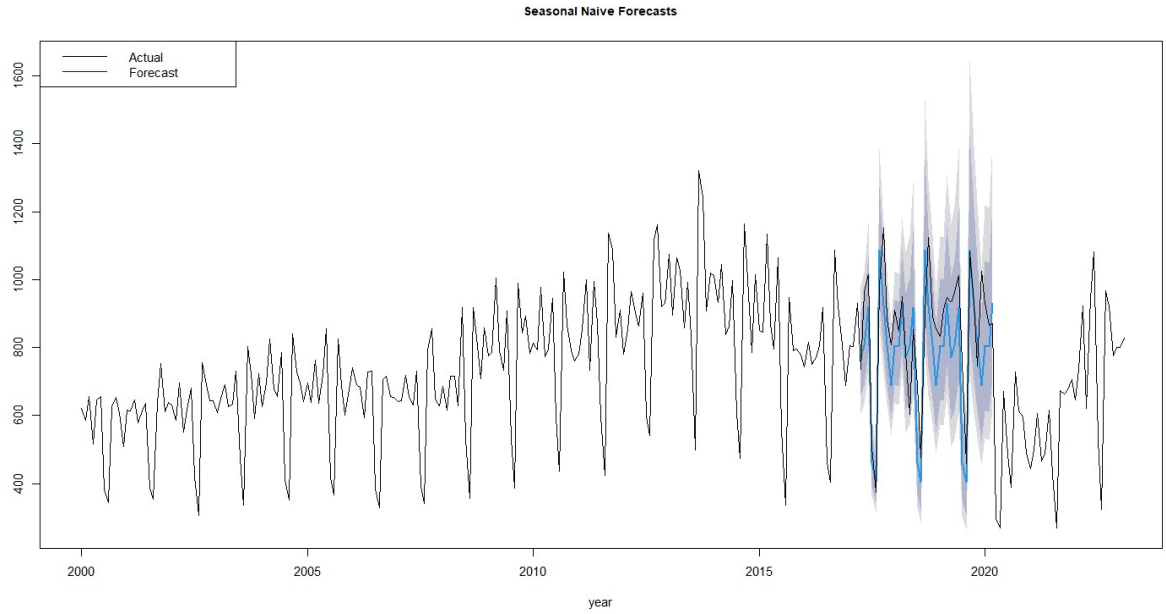
Since the p-value of the test is less than the significance level of 0.05, we can reject the null hypothesis of non-stationarity and conclude that the bankrupt time series is likely stationary. Therefore, there is no need to apply a transformation or adjustment to the time series.

Also, there is no need to find out the optimal lambda value using Box Cox method since there is no transformation is required.

3. Forecast : Seasonal Naïve Method

In this section, we will forecast the Belgium Bankruptcy data using different models , in order to evaluate the performance of the different forecasts I will divide the data in a training and test set. The training dataset contains information from January 2000 to March 2017 and the test dataset from April 2017 to March 2020.

The next plot shows the forecasts made by the seasonal naive method



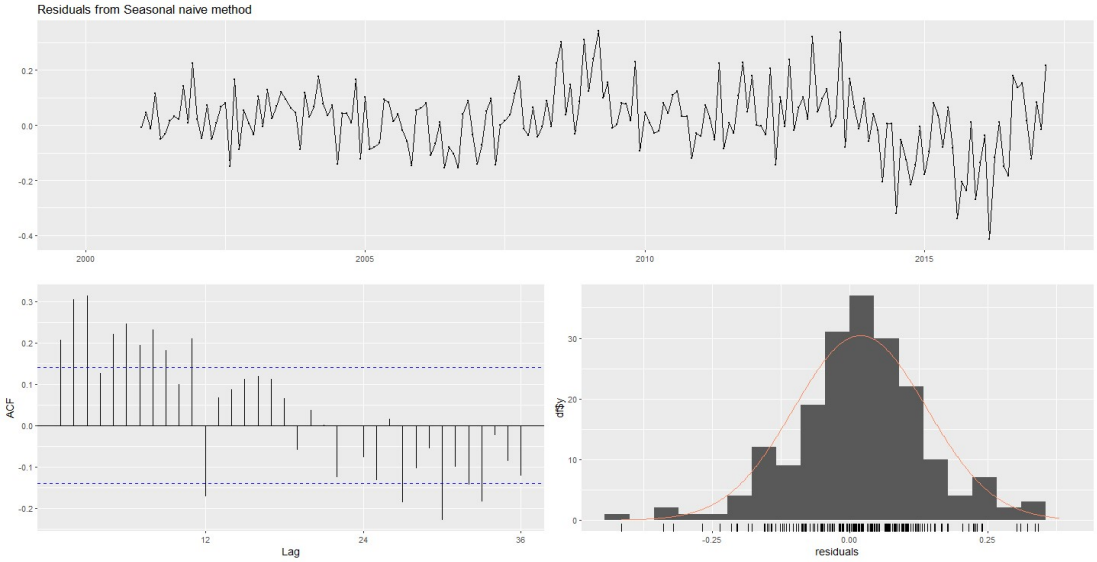
The resulting plot displays the actual values until February 2023 in black and the forecasted values until February 2022 in blue. It can be observed that the forecasted values exhibit a similar seasonal pattern to the actual values and are reasonably consistent with the trend.

Checking the forecast accuracy we have the following results:

```
##                      ME      RMSE       MAE      MPE     MAPE     MASE
## Training set  13.32308  95.37069  68.73846 1.164853  9.086461 1.000000
## Test set      65.58333 134.72740 107.36111 7.181171 12.786016 1.561878
##                   ACF1 Theil's U
## Training set 0.22608292        NA
## Test set    -0.03370048 0.3951562
```

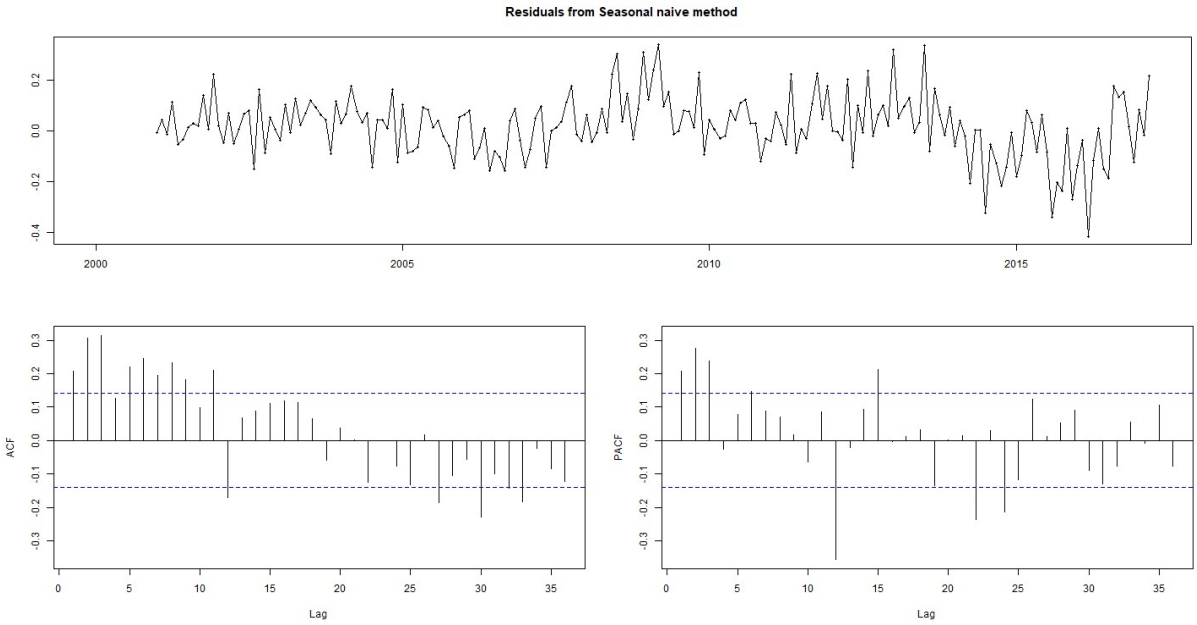
The model shows a better performance(in terms of RMSE, MAE, MAPE and MASE) in the training set than in the test set. Additionally, the ACF1 value is positive for the training set, indicating a positive autocorrelation, whereas the ACF1 value is negative for the test set, indicating a negative autocorrelation. Later, we will use these metrics to compare this model with other models.

Checking the residuals we have the following results:




```
##
##  Ljung-Box test
##
## data:  Residuals from Seasonal naive method
## Q* = 133.03, df = 24, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 24
```

In the previous plot, one can suggest that the residuals are around zero and they do not follow completly a normal distribution. There is a high peak in the distribution.

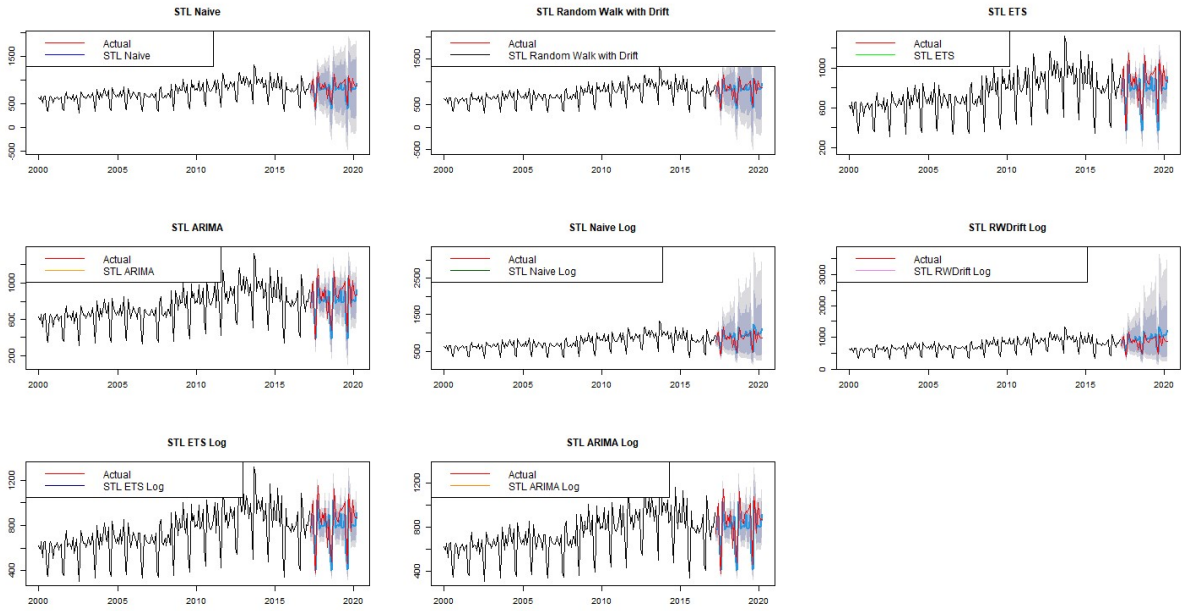


Looking at the ACF, we can observe that we have significant spikes in the lags 2, 3, 6, and 12, what it means that we still have information in the residuals that we are not considering in the seasonal naive model and it would not be reflected in the forecast.

The Ljung Box test suggests that we can reject the null hypothesis of white noise in the residuals since we have small p-values in all lags. This confirms that we still have information in the residuals and we are not capturing completely the data generation process. So in order to determine if these are good forecasts, we need to run other models.

4. Forecast: STLF Decomposition

The next plot shows the forecasts made by the STLF decomposition method. This model combine forecasts of the seasonally adjusted data using a naïve, random walk with drift method, ets, arima methods. The red line represents the actual test dataset and the other colored plots denotes the forecasts of different methods



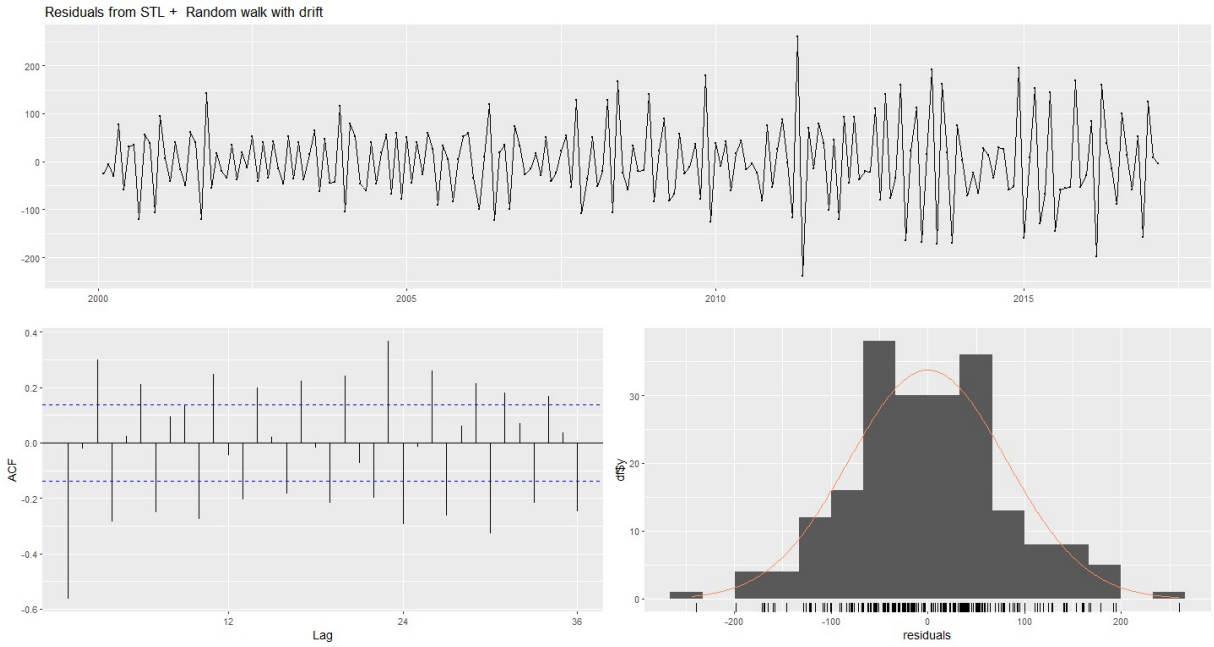
Checking the forecast accuracy and residuals we have the following results:

	Model	Accuracy_Train.RMSE	Accuracy_Train.MAE	Accuracy_Train.MAPE	Accuracy_Train.MASE	Accuracy_Test.RMSE
1 STL Seasonal Naive	stlf	81.18879	64.27866	9.003926	0.9351193	110.4517
2 STL Random Walk with Drift	stlf	81.18277	64.29786	9.010373	0.9353986	102.3765
3 STL ETS	stlf	60.76960	46.46273	6.395061	0.6759349	119.0959
4 STL ARIMA	stlf	58.07320	44.77898	6.174608	0.6514399	110.5275
5 STL NAIVE LOG	stlf	79.77829	61.78330	8.426455	0.8988170	113.4301
6 STL RWDRIFT LOG	stlf	79.83006	61.84286	8.441720	0.8996835	140.9635
7 STL ETS LOG	stlf	58.86785	44.81192	6.025132	0.6519192	115.5934
8 STL ARIMA LOG	stlf	56.76196	42.86390	5.795552	0.6235796	112.1173

	Accuracy_Test.MAE	Accuracy_Test.MAPE	Accuracy_Test.MASE	Residuals.Q.	Residuals.df	Residuals.p.value
1 STL Seasonal Naive	91.09621	10.845480	1.325258	286.50209	24	0.000000e+00
2 STL Random Walk with Drift	83.52577	9.976188	1.215124	286.50209	24	0.000000e+00
3 STL ETS	97.65418	11.527148	1.420663	129.93667	24	1.110223e-16
4 STL ARIMA	91.12241	10.846234	1.325639	51.95891	22	3.167208e-04
5 STL NAIVE LOG	90.89345	11.032378	1.322308	208.96300	24	0.000000e+00
6 STL RWDRIFT LOG	110.50833	13.345307	1.607664	208.96300	24	0.000000e+00
7 STL ETS LOG	94.94758	11.138317	1.381288	91.32515	24	8.689854e-10
8 STL ARIMA LOG	91.69230	10.813152	1.333930	53.15555	20	7.710019e-05

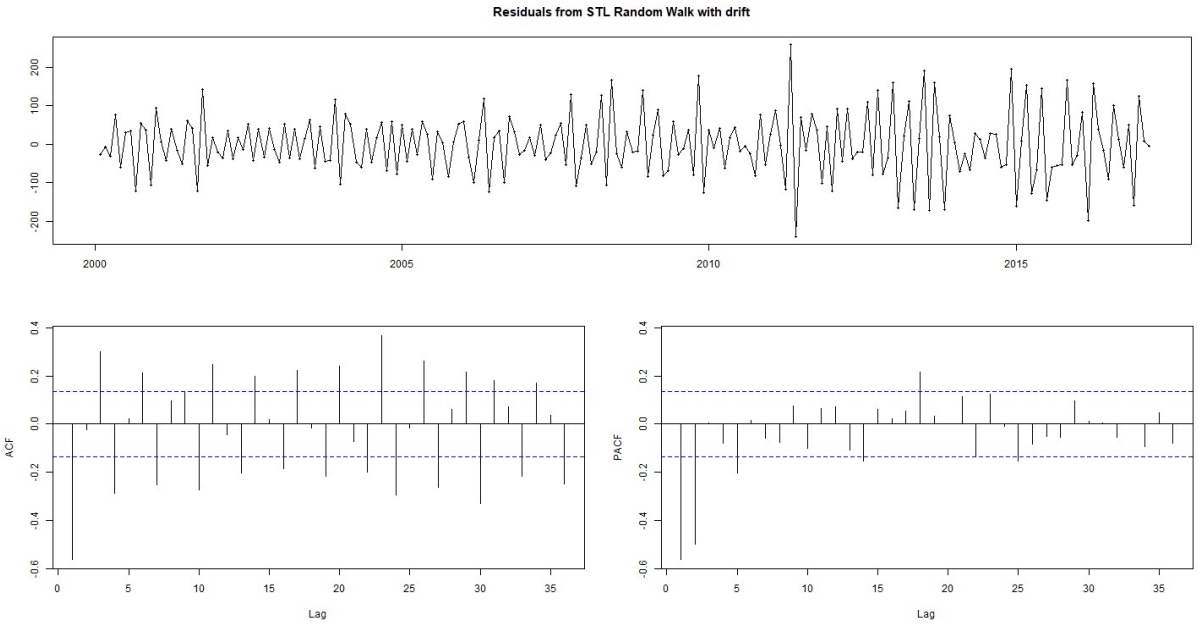
We can say from the above results that **STL random walk with drift** model is the best performing, as it has the lowest test residual error (MASE, RMSE, MAE, MPE) compared to the others, which indicates that it fits better on unseen data compared to other models. However, we can also see that the prediction interval of random walk with drift is large. Hence, we need to check some other models as well to get the final forecast predictions.

Checking the residuals we have the following results :




```
##
##  Ljung-Box test
##
## data:  Residuals from STL +  Random walk with drift
## Q* = 286.5, df = 24, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 24
```

In the previous plot, we can suggest that the residuals are not around zero and they follow approximately a normal distribution.



Looking at the ACF, we can observe that we have significant spikes in many lags (3, 4, 6, 7..) what it means that we still have information in the residuals that we are not considering in the STL Random walk with Drift model and it would not be reflected in the forecasts.

The Ljung Box test suggests that we can reject the null hypothesis of white noise in the residuals since we have p-values very close to zero in the all lags. This confirms that we still have information in the residuals of the STL Random walk with Drift model and we are not capturing completely the data generation process.

5. Forecast: ETS

For building an ETS model, I select and fit various models with different combinations of error, trend, and seasonality components. The chosen models are: AAN, ANN, AAA, MMM, and MAM.

Then, we can compare the models using information criteria such as AIC and BIC:

```
##            AIC      BIC
##     AAN 3228.222 3248.219
##     ANN 3227.745 3237.743
##     AAA 2890.330 2950.319
##     MMM 2825.244 2885.233
##     MAM 2828.380 2885.036
```

Based on the obtained results, the MMM model shows a lower AIC, indicating a better fit, while the MAM model has a lower BIC, which is indicative of better performance when accounting for complexity. However, it is important to note that neither AIC nor BIC alone can determine the absolute "best" model. Therefore, to make a more informed decision, we will further analyze the accuracy metrics and residual diagnostics.

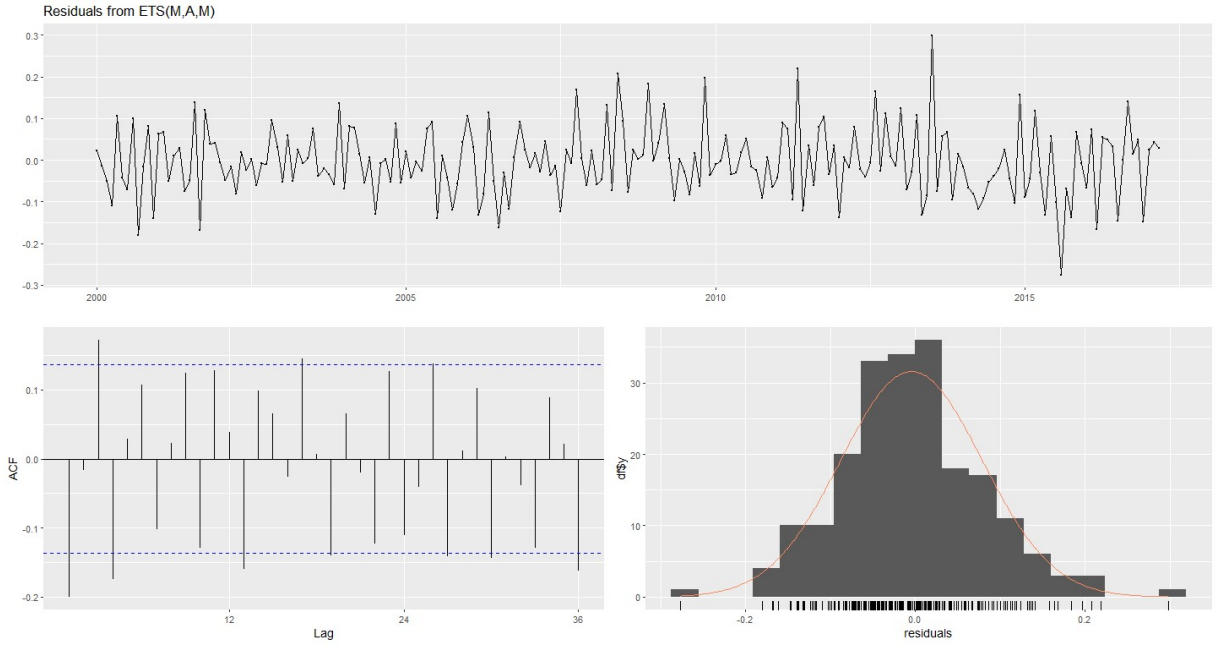
Checking the forecast accuracy and residuals we have the following results

	Model	Accuracy_Train.RMSE	Accuracy_Train.MAE	Accuracy_Train.MAPE	Accuracy_Train.MASE	Accuracy_Test.RMSE	Accuracy_Test.MAE
1 AAN	ETS	164.39852	120.69572	20.026318	1.7558688	191.1906	159.67537
2 ANN	ETS	166.60619	123.80715	20.155812	1.8011336	185.3147	151.02506
3 AAA	ETS	68.59077	53.72790	7.713328	0.7816279	107.9847	86.34934
4 MMM	ETS	62.25914	47.39461	6.433476	0.6894918	120.8976	99.14956
5 MAM	ETS	62.19617	47.51199	6.560347	0.6911995	100.4588	80.02241

	Accuracy_Test.MAPE	Accuracy_Test.MASE	Residuals.Q.	Residuals.df	Residuals.p.value
1 AAN	21.233286	2.322941	407.21791	24	0.000000e+00
2 ANN	20.825152	2.197097	418.52163	24	0.000000e+00
3 AAA	10.118600	1.256201	79.85815	24	6.409098e-08
4 MMM	11.577688	1.442417	60.16498	24	6.053941e-05
5 MAM	9.574288	1.164158	66.97489	24	6.219158e-06

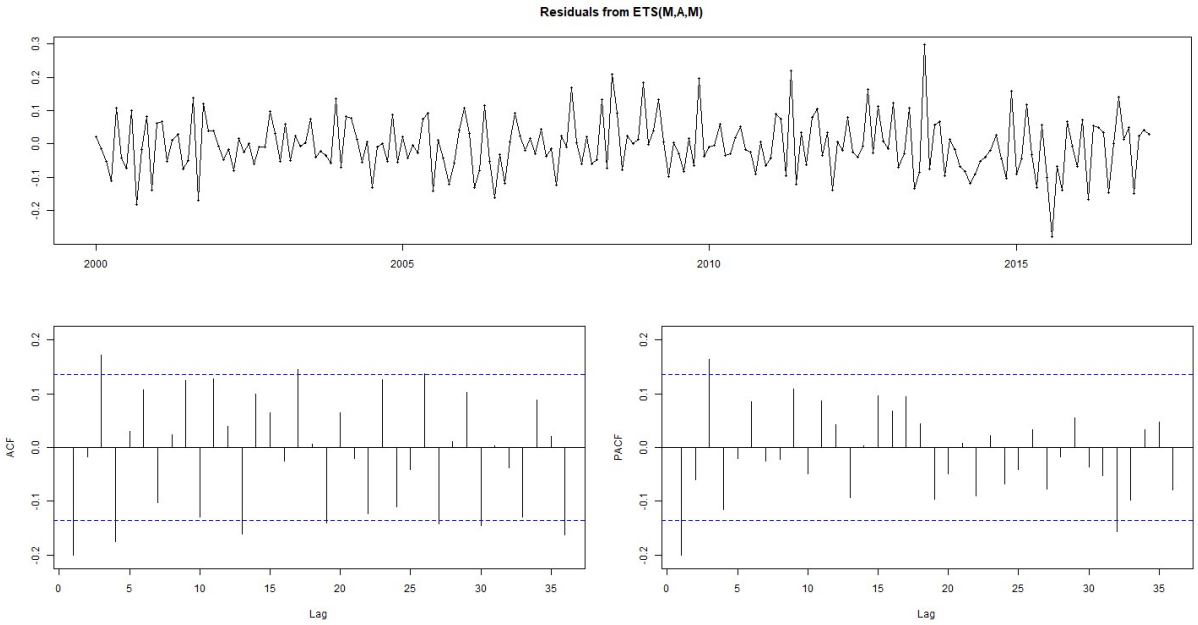
We can say from the above results that **MAM** (Multiplicative error, Additive trend, Multiplicative seasonality) model is the best performing, as it has the lowest test residual error (MASE, RMSE, MAE, MPE) compared to the others, which indicates that it fits better on unseen data compared to other models.

Checking the residuals we have the following results :




```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(M,A,M)
## Q* = 66.975, df = 24, p-value = 6.219e-06
##
## Model df: 0.   Total lags used: 24
```

In the previous plot, we can suggest that still the residuals are not completely around zero and they follow approximately a normal distribution.



Looking at the ACF, we can still observe that we have some spikes in lags 3, 4 what it means that we still have information in the residuals that we are not considering in the STL Random walk with Drift model and it would not be reflected in the forecasts.

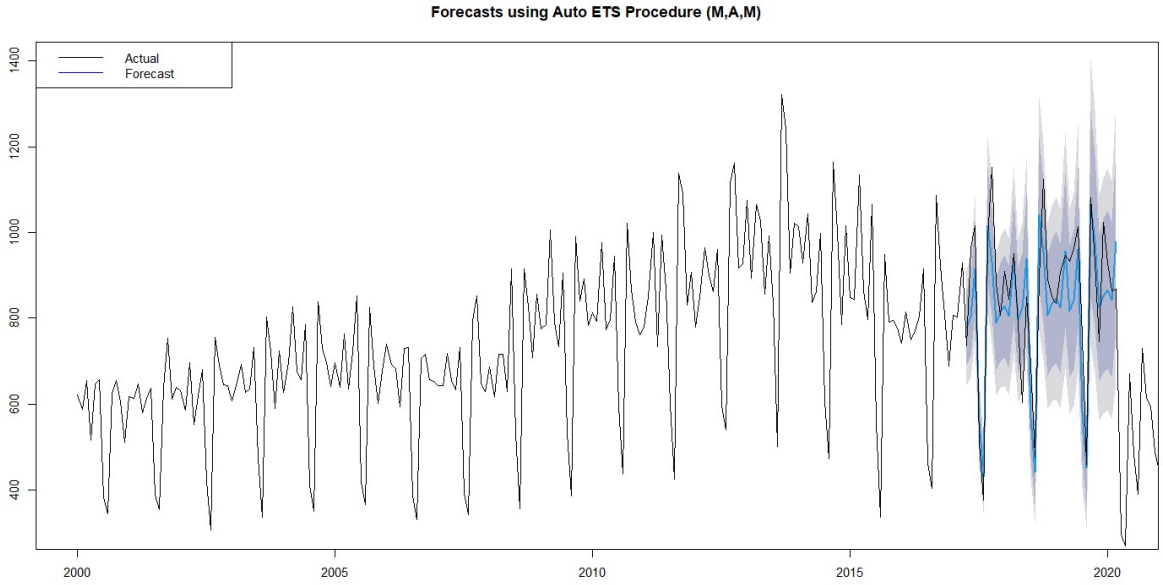
The Ljung Box test suggests that we can reject the null hypothesis of white noise in the residuals since we have p-values very close to zero in the all lags. This confirms that we still have information in the residuals of the ETS(M,A,M) and we are not capturing completely the data generation process.

The parameters associated to the model are:

```
##         alpha          beta         gamma             l             b
##        0.2782          8e-04          1e-04      589.0881        2.3357
##            s0            s1            s2            s3            s4
##        1.0113        0.9815        1.1649        1.2735        0.5402
##            s5            s6            s7            s8            s9
##        0.6976        1.1545        1.0147        0.9827        1.1553
##           s10           s11
##        0.9963        1.0274
```

The alpha value indicates that the model is not taking into account only the last observations, also it is taking older information to create the data generation process. The value of the beta parameter suggests that there is a slight change in the slope over time and we are modeling a constant and linear trend.

Let's now run an automated ETS procedure and we can see that the auto ETS result is MAM.



6. Forecast: ARIMA

In this section, we will first fit auto ARIMA model and later we will fit four seasonal ARIMA models. First, I will use the auto.arima function from R, this function returns the best model base on the AIC criteria. In the second model, we will fit all models by defining the parameters taking into account the ACF and PACF plots associated with the data.

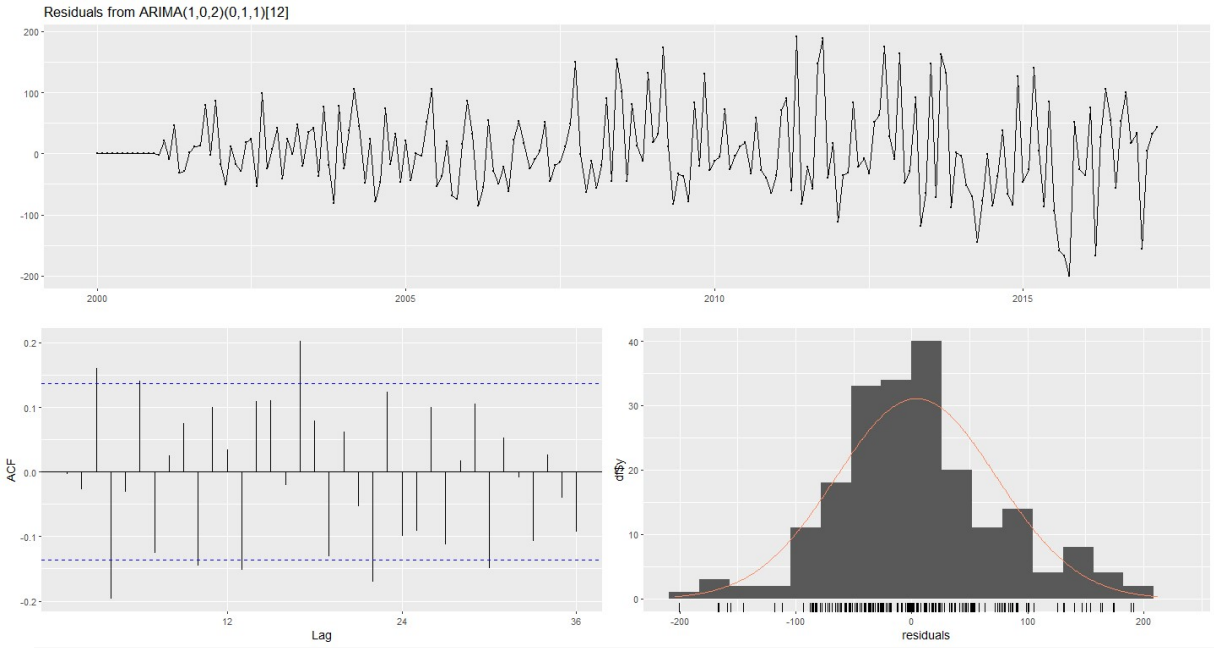
Auto ARIMA

Now, we will evaluate the results provided by the function auto.arima from R. We can see from the below result that the AUTO ARIMA model is the seasonal Arima model with a seasonal difference of 1 (denoted by the 'd' parameter in the seasonal order), a non-seasonal difference of 0 (denoted by the 'D' parameter in the seasonal order), an autoregressive (AR) term of 1, and a moving average (MA) term of 2 in the non-seasonal order. The seasonal order indicates that there is no seasonal autoregressive term (SAR=0) and a seasonal moving average term of 1 (SMA=1), with a seasonal period of 12 (denoted by the [12] parameter).

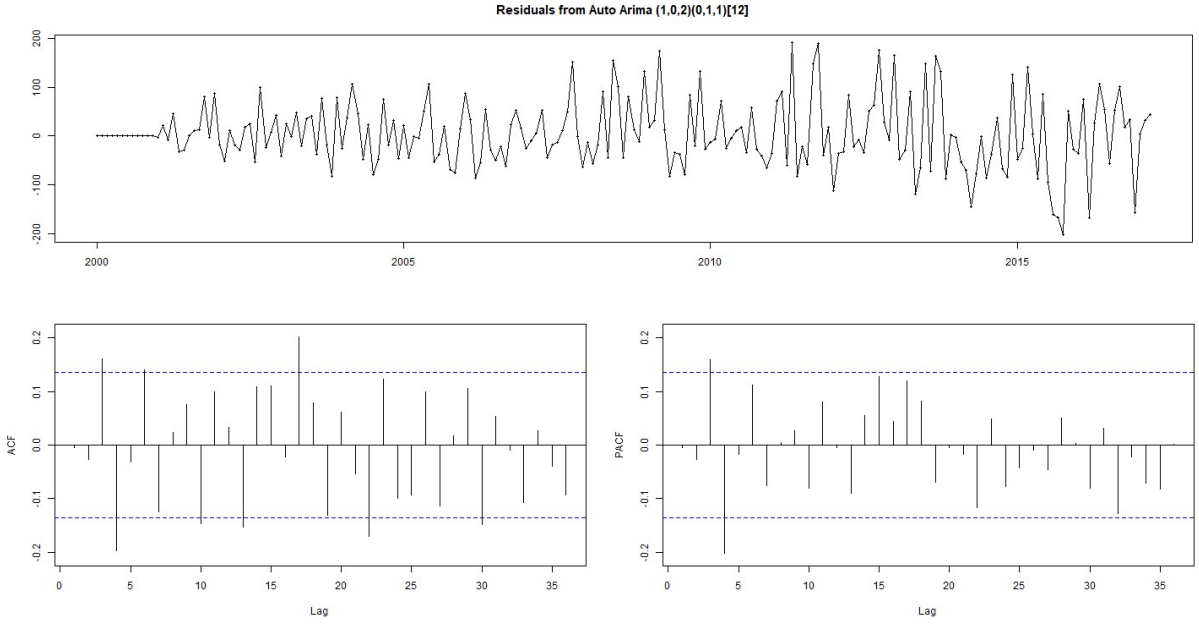
The summary result of AUTO Arima Model is as follows :

```
ARIMA(1,0,2)(0,1,1)[12]

Coefficients:
          ar1      ma1     ma2     sma1
      0.9767  -0.8944  0.1757  -0.7417
s.e.  0.0180   0.0706  0.0708   0.0564

sigma^2 = 5214:  log likelihood = -1113.93
AIC=2237.87   AICc=2238.18   BIC=2254.23

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE         ACF1
Training set 4.028675 69.35942 52.15576 -0.4910388 7.143469 0.7587566 -0.004027925
```

Checking the residuals we have the following results :




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,2)(0,1,1)[12]
## Q* = 69.942, df = 20, p-value = 1.861e-07
##
## Model df: 4.   Total lags used: 24
```



Looking at the ACF, we can still observe that we have some spikes in lags 3, 4,6,10 what it means that we still have information in the residuals that we are not considering in the Auto Arima model and it would not be reflected in the forecasts.

The Ljung Box test suggests that we can reject the null hypothesis of white noise in the residuals since we have p-values very close to zero in the all lags and our previous ETS(M,A,M) model has less Q* value compared to Auto Arima model. This confirms that we still have information in the residuals of the Auto Arima and we are not capturing completely the data generation process.

The parameters associated to the Auto Arima model are:

```
##           ar1          ma1          ma2          sma1
##     0.9767486   -0.8944041   0.1756626      -0.7416716
```

The ARIMA(1,0,2)(0,1,1)[12] equation in terms of the backward shift operator B can be written as:

(1 - ar1B)(1 - ma1B - ma2B^2)(1 - sma1B^12)yt = et

(1 - 0.9767B)(1 - B)(1 - 0.8944B + 0.1757B^2)(1 - 0.7417B^12)yt = et

This can be further simplified as,

yt - 0.9767yt-1 + 0.8944et-1 - 0.1757et-2 - 0.7417yt-12 = et/B

Checking the forecast accuracy we have the following results:

```
##                       ME      RMSE       MAE       MPE     MAPE     MASE
## Training set  4.028675  69.35942 52.15576 -0.4910388  7.143469 0.7587566
## Test set     65.600805 118.65068 96.83687  7.1298161 11.383273 1.4087728
##                    ACF1 Theil's U
## Training set -0.004027925      NA
## Test set      0.034562256 0.3327778
```

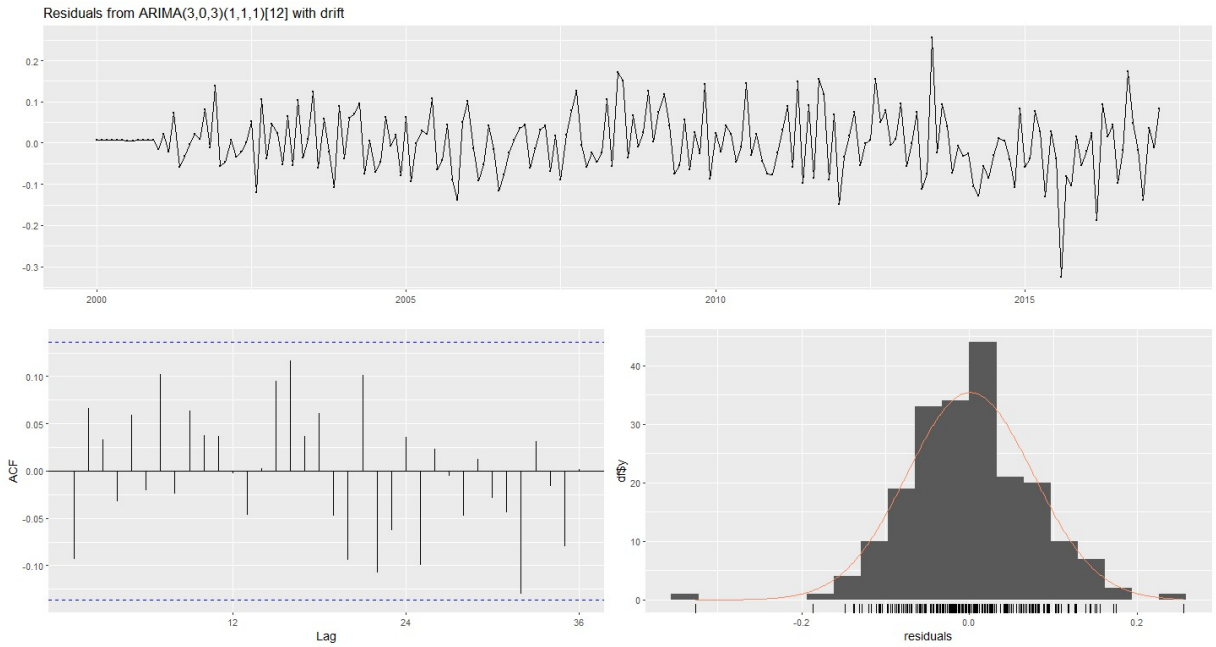
Manual ARIMA

For starting to fit the model, I will look at the ACF and PACF plots of the transformed data taking into account one seasonal difference. The function ndiffs suggests that after this difference we do not need more differences.

Looking at the ACF and PACF functions I will suggest implementing an ARIMA(3,0,3)(1,1,1)[12] model. This because we can observe that there are significant spikes since the lag 3. Regarding the seasonal part in the PACF, we can observe significant spikes in the lags 12,24 and 36, and in the ACF in the lag 12. This suggests that I could model the seasonal part as (1,1,1) taking into account that the data has one seasonal difference.

The summary result of ARIMA(3,0,3)(1,1,1)[12] Model is as follows :

```
Forecast method: ARIMA(3,0,3)(1,1,1)[12] with drift

Model Information:
Series: bank_train
ARIMA(3,0,3)(1,1,1)[12] with drift
Box Cox transformation: lambda= 0

Coefficients:
          ar1     ar2     ar3     ma1     ma2      ma3    sar1     sma1   drift
      -0.1617  0.1609  0.9991  0.3950  0.1268  -0.7570  0.0333  -0.9997  0.0015
s.e.   0.0045  0.0047  0.0011  0.1084  0.1336   0.0417  0.0772   0.0284  0.0013

sigma^2 = 0.006282:  log likelihood = 203.75
AIC=-387.5   AICc=-386.3   BIC=-354.77

Error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set 2.925488 56.75498 42.77172 -0.1974671 5.775402 0.6222386 -0.05993036
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,0,3)(1,1,1)[12] with drift
## Q* = 23.569, df = 16, p-value = 0.09934
##
## Model df: 8.   Total lags used: 24
```

The total parameters associated with the model are 9 including the drift, analyzing their p-values we observe that only 2 are not significant.

Checking the forecast accuracy we have the following results:

	Model	Accuracy_Train.RMSE	Accuracy_Train.MAE	Accuracy_Train.MAPE	Accuracy_Train.MASE	Accuracy_Test.RMSE
1 ARIMA_102_111_DRIFT	Arima	61.27284	46.10614	6.127690	0.6707473	97.18198
2 ARIMA_102_111	Arima	62.25452	46.84388	6.183377	0.6814799	117.07597
3 ARIMA_303_111_DRIFT	Arima	56.75498	42.77172	5.775402	0.6222386	98.83350
4 ARIMA_303_211_DRIFT	Arima	56.76593	42.79416	5.778973	0.6225651	100.19299

	Accuracy_Test.MAE	Accuracy_Test.MAPE	Accuracy_Test.MASE	Residuals.Q.	Residuals.df	Residuals.p.value
1 ARIMA_102_111_DRIFT	74.62747	9.026006	1.085673	43.84742	19	0.0009913786
2 ARIMA_102_111	94.90246	11.103674	1.380631	43.83448	19	0.0009954694
3 ARIMA_303_111_DRIFT	82.50755	9.963399	1.200311	23.56944	16	0.0993392203
4 ARIMA_303_211_DRIFT	83.98466	10.109505	1.221800	23.19401	15	0.0801065601

We notice that for this ARIMA_303_111_DRIFT model, the value of Q* is 23.56, and the p-value is greater than 0.05, providing insufficient evidence to reject the null hypothesis of white noise. This indicates that our model may be capturing all the necessary information in the time series data.

7. Benchmark Forecasting Models

After creating the previous models, now we are going to compare them in terms of residuals and accuracy. To compare the different models in terms of residual diagnostics and forecast accuracy, we can calculate the root mean squared error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE) and the mean absolute scaled error (MASE) for each model on the train and test sets, as well as examine their residual diagnostics.

```
> summary_table
                        Model Accuracy_Train.RMSE Accuracy_Train.MAE Accuracy_Train.MAPE Accuracy_Train.MASE Accuracy_Test.RMSE
1 STL Seasonal Naive     stlf           81.18879          64.27866           9.003926          0.9351193          110.45171
2 STL Random Walk with Drift  stlf      81.18277          64.29786           9.010373          0.9353986          102.37652
3 STL ETS                stlf           60.76960          46.46273           6.395061          0.6759349          119.09590
4 STL ARIMA              stlf           58.07320          44.77898           6.174608          0.6514399          110.52750
5 STL NAIVE LOG          stlf           79.77829          61.78330           8.426455          0.8988170          113.43015
6 STL RWDRIFT LOG        stlf           79.83006          61.84286           8.441720          0.8996835          140.96352
7 STL ETS LOG            stlf           58.86785          44.81192           6.025132          0.6519192          115.59345
8 STL ARIMA LOG          stlf           56.76196          42.86390           5.795552          0.6235796          112.11728
1 AAN                     ETS          164.39852         120.69572          20.026318          1.7558688          191.19057
2 ANN                     ETS          166.60619         123.80715          20.155812          1.8011336          185.31471
3 AAA                     ETS           68.59077          53.72790           7.713328          0.7816279          107.98465
4 MMM                     ETS           62.25914          47.39461           6.433476          0.6894918          120.89761
5 MAM                     ETS           62.19617          47.51199           6.560347          0.6911995          100.45882
1 ARIMA_102_111_DRIFT   Arima           61.27284          46.10614           6.127690          0.6707473           97.18198
2 ARIMA_102_111         Arima           62.25452          46.84388           6.183377          0.6814799          117.07597
3 ARIMA_303_111_DRIFT   Arima           56.75498          42.77172           5.775402          0.6222386           98.83350
4 ARIMA_303_211_DRIFT   Arima           56.76593          42.79416           5.778973          0.6225651          100.19299
                        Accuracy_Test.MAE Accuracy_Test.MAPE Accuracy_Test.MASE Residuals.Q. Residuals.df Residuals.p.value
1 STL Seasonal Naive             91.09621          10.845480          1.325258    286.50209           24      0.000000e+00
2 STL Random Walk with Drift     83.52577           9.976188          1.215124    286.50209           24      0.000000e+00
3 STL ETS                        97.65418          11.527148          1.420663    129.93667           24      1.110223e-16
4 STL ARIMA                      91.12241          10.846234          1.325639     51.95891           22      3.167208e-04
5 STL NAIVE LOG                  90.89345          11.032378          1.322308    208.96300           24      0.000000e+00
6 STL RWDRIFT LOG               110.50833          13.345307          1.607664    208.96300           24      0.000000e+00
7 STL ETS LOG                    94.94758          11.138317          1.381288     91.32515           24      8.689854e-10
8 STL ARIMA LOG                  91.69230          10.813152          1.333930     53.15555           20      7.710019e-05
1 AAN                           159.67537          21.233286          2.322941    407.21791           24      0.000000e+00
2 ANN                           151.02506          20.825152          2.197097    418.52163           24      0.000000e+00
3 AAA                            86.34934          10.118600          1.256201     79.85815           24      6.409098e-08
4 MMM                            99.14956          11.577688          1.442417     60.16498           24      6.053941e-05
5 MAM                            80.02241           9.574288          1.164158     66.97489           24      6.219158e-06
1 ARIMA_102_111_DRIFT            74.62747           9.026006          1.085673     43.84742           19      9.913786e-04
2 ARIMA_102_111                  94.90246          11.103674          1.380631     43.83448           19      9.954694e-04
3 ARIMA_303_111_DRIFT            82.50755           9.963399          1.200311     23.56944           16      9.933922e-02
4 ARIMA_303_211_DRIFT            83.98466          10.109505          1.221800     23.19401           15      8.010656e-02
```

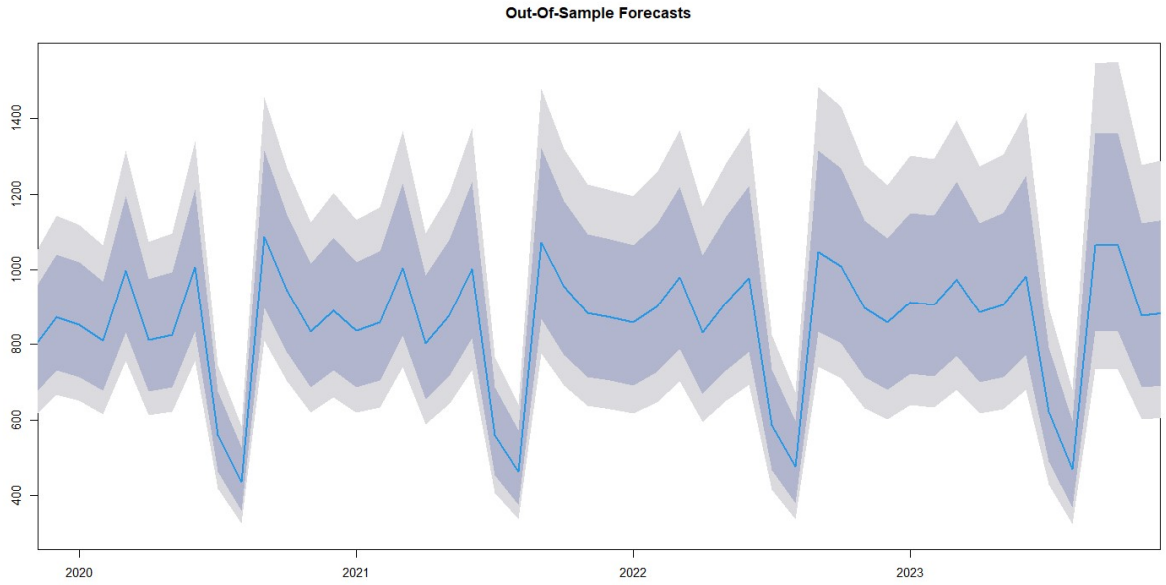
Based on the accuracy measures, the ARIMA_303_111_DRIFT model has the lowest RMSE/MAE/MAPE/MASE on the train set, but the ARIMA_102_111_DRIFT model has the lowest metrics on the test set.

Based on the residuals, ARIMA (3,0,3)(1,1,1) with drift and ARIMA (3,0,3)(2,1,1) with drift have low Q* values of 23.56 and 23.19 respectively and high p values of 0.09 and 0.08 respectively, suggesting the evidence of white noise. While all the remaining models have significant p values thus rejecting the null hypothesis of white noise.

Therefore, we will select the ARIMA model(3,0,3)(1,1,1) as our final model for forecasting the monthly number of corporate bankruptcies in Belgium.

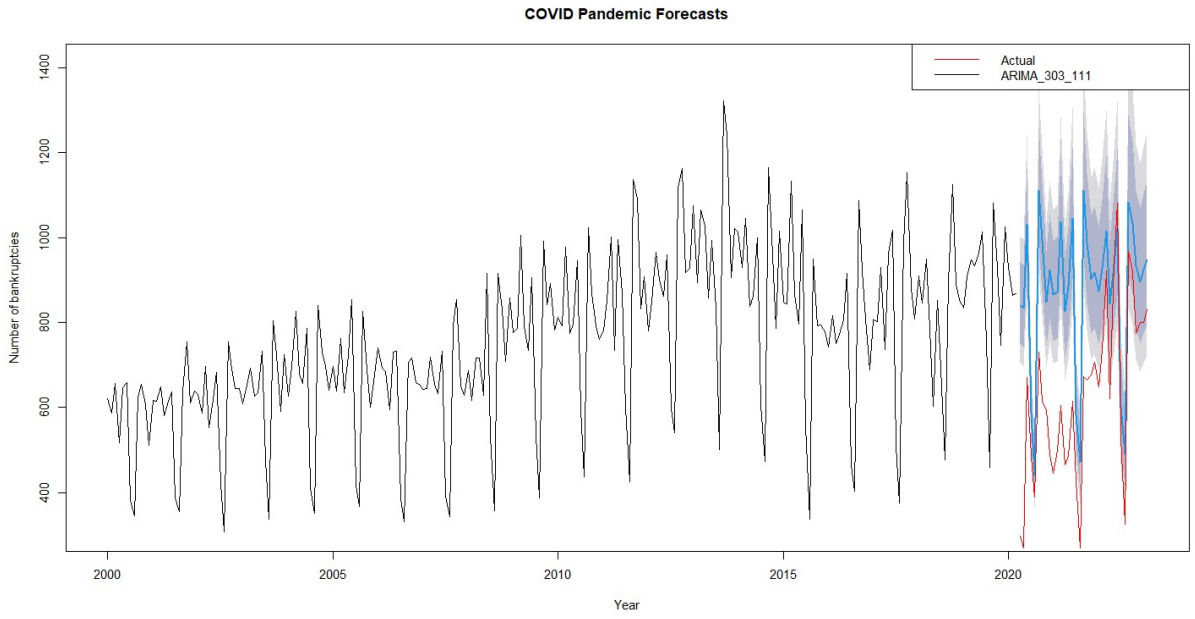
8. Out of Sample Forecasts: Model ARIMA (3,0,3)(1,1,1) with Drift

For generating the forectasts up to December 2023, first we need to fit the model in the whole data from January 2000 to March 2020. After this we can make the predictions.



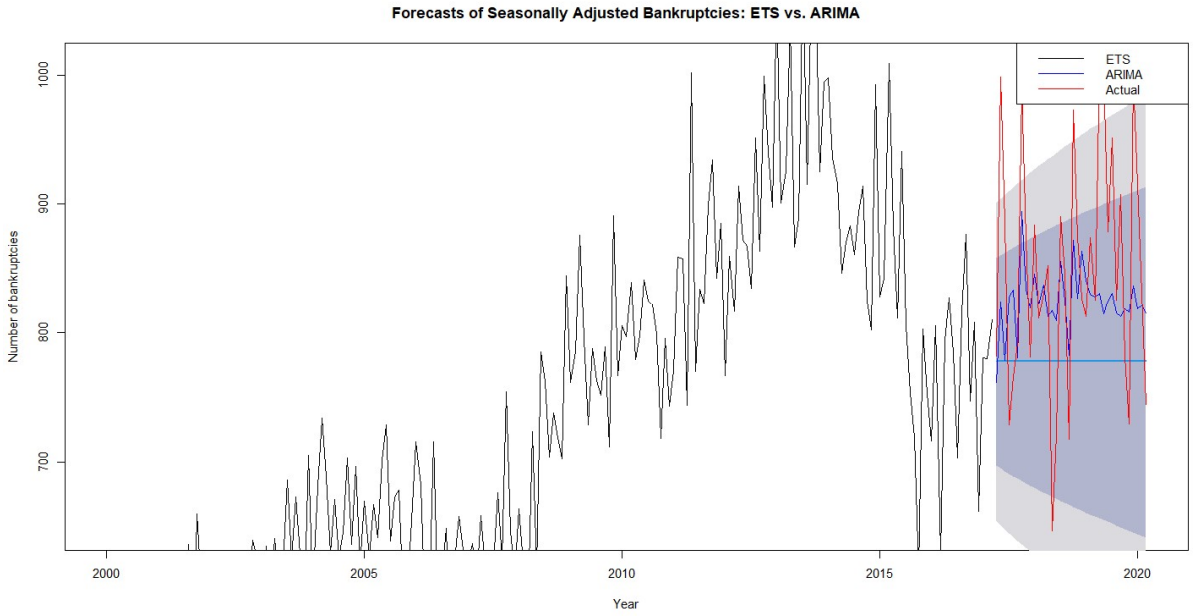
9. COVID Impact

We can clearly see from the below plot results that Covid changed the usual trend of bankruptcies in Belgium. Less bankruptcies were reported during the post pandemic period than we have forecasted based on the historical data. The red line indicates the actual values where we can see that in Aug/Sep 2020 there are only less number of bankruptcies but our forecasted results are higher during these months.



10. Seasonal Adjustment using mstl

Based on the below plot, we can see that the Auto ARIMA model (ARIMA(0,1,2)(2,0,1)[12]) is a better fit for the seasonally adjusted data, as it has lower accuracy measures and smaller difference between train and test sets.




```
> accuracy(pred_ets, test)
                    ME      RMSE      MAE         MPE      MAPE      MASE         ACF1 Theil's U
Training set  4.36137  60.81777 46.43992 0.08964232  6.232179 0.6713215 -0.12228568        NA
Test set     71.17657 115.43449 92.77990 7.31278889 10.405288 1.3411982  0.07945618 0.9254166
>
> # Accuracy of ARIMA model
> accuracy(pred_arima, test)
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1 Theil's U
Training set 10.38671 51.66893 40.03085 0.9425414 5.364596 0.5786739 -0.03399018        NA
Test set     24.41351 87.24458 69.30769 1.8683570 8.059382 1.0018911  0.19601044 0.6937255
```

And the parameters of the Auto Arima model is as follows :

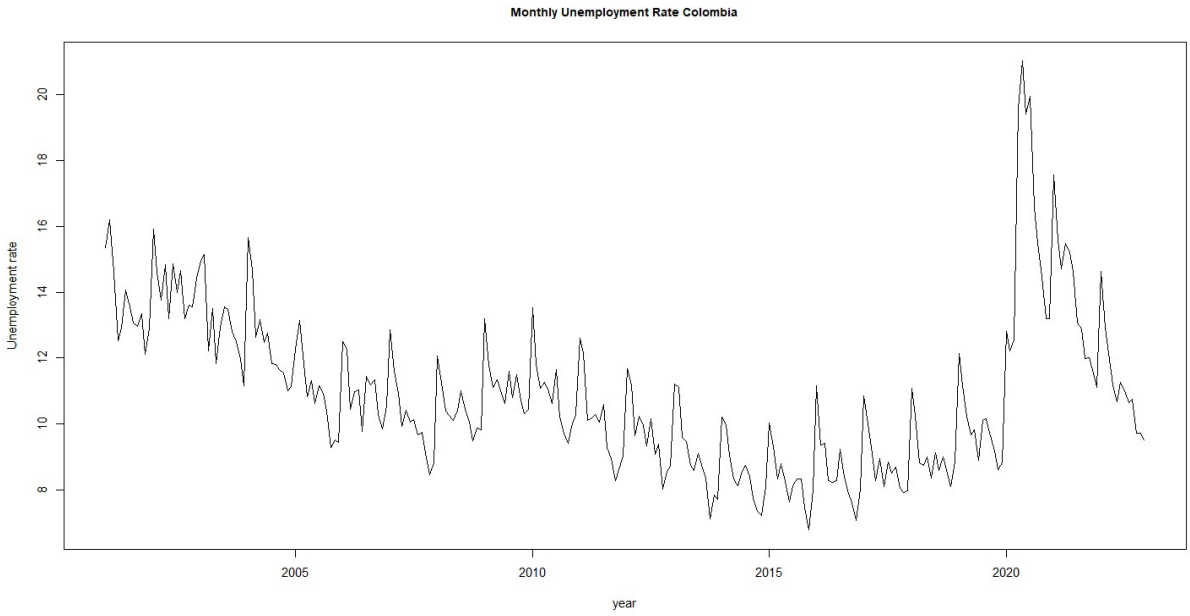
```
> fit_arima
Series: train
ARIMA(0,1,2)(2,0,1)[12]

Coefficients:
          ma1     ma2    sar1     sar2     sma1
      -0.8152  0.2304  0.4464  -0.2976  -0.7891
s.e.   0.0722  0.0664  0.0876   0.0791   0.0837

sigma^2 = 2749:  log likelihood = -1111.31
AIC=2234.62   AICc=2235.04   BIC=2254.58
```

EXERCISE 02 : Time Series: Unemployment Rate Colombia

For this section, I will use a time series that contains the unemployment rate in Colombia that is measured in percentage. The dataset is collected from the source (Banco de la Republica, Colombia)

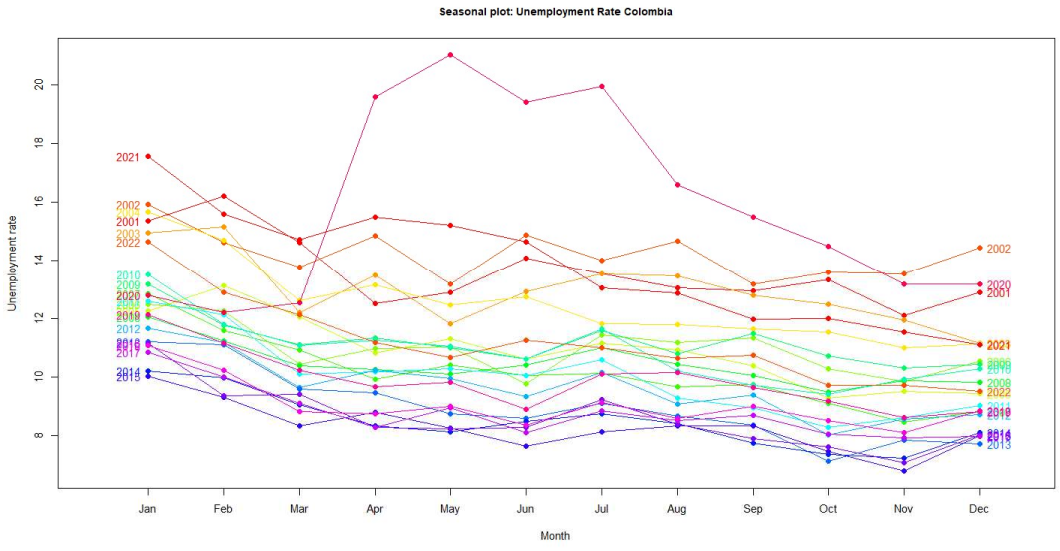
1. Data Exploration

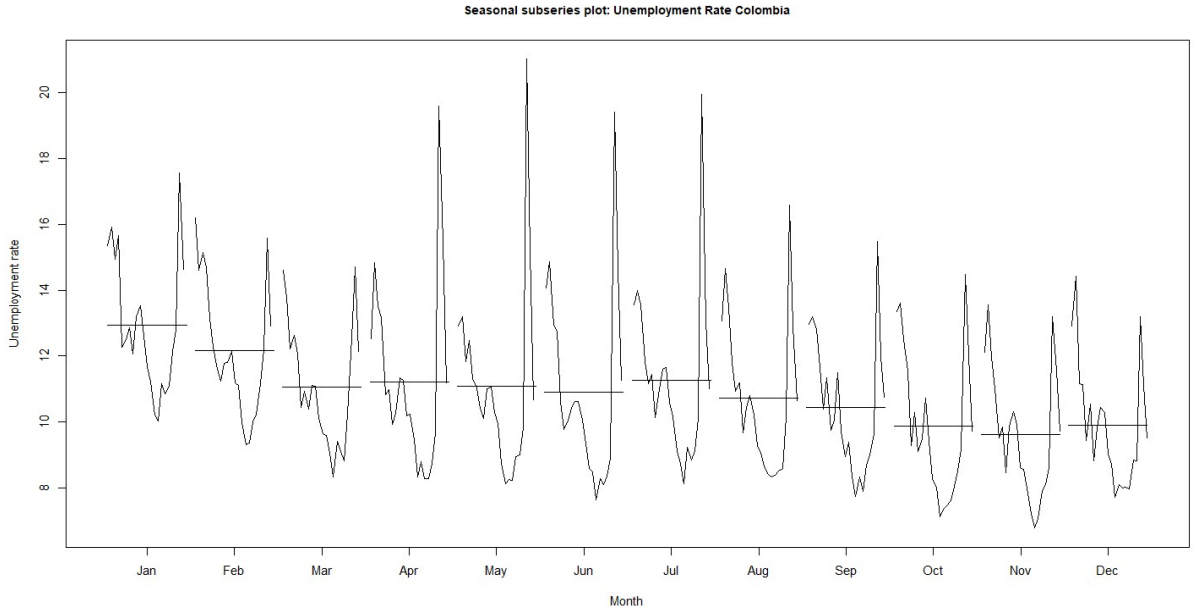
The dataset contains monthly information of the unemployment rate from January 2001 to December 2022, the next figure shows the unemployment rate series over the time:



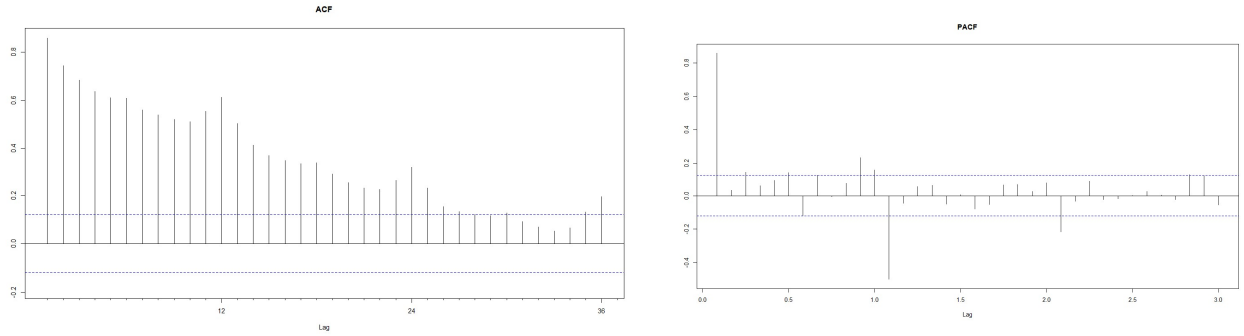
We can observe that the variation of the time series is changing over time what suggests that the data is a multiplicative time series. We will explore some relevant graphs in order to identify the different components of the data. The highest number of unemployment rate occurred in May 2020, with 21%, while the lowest number occurred in November 2015, with only 6.8%.

The following seasonal plots will be used to confirm the assumptions of seasonal and trend patterns:



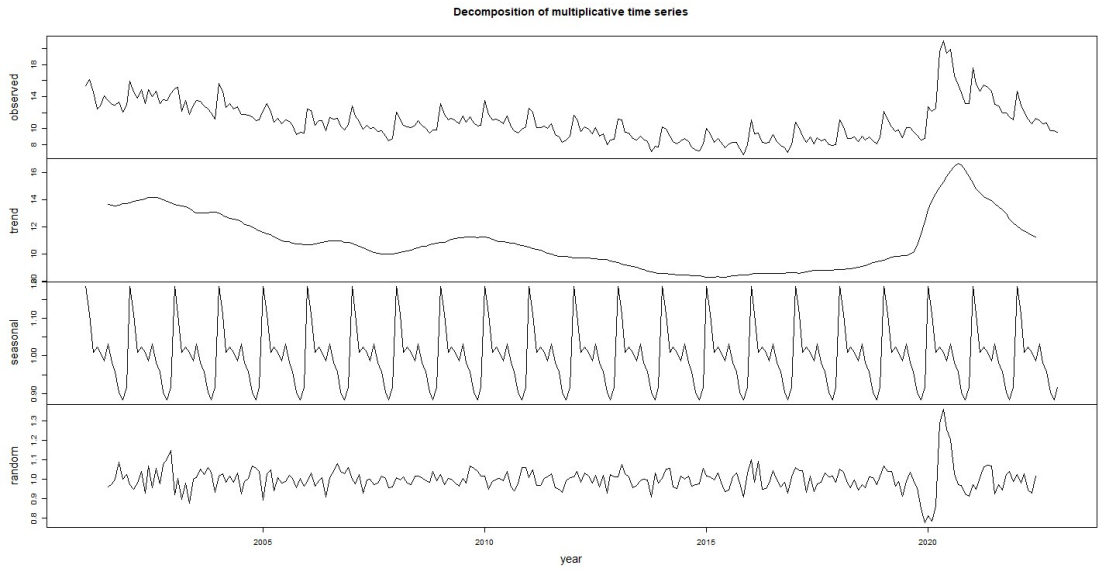


Analyzing the first plot is not clear the trend component of the data because the data is overlapped. Using the month seasonal plot, we can confirm the presence of an increasing trend. However, from the above plot we can see that there is no clear increasing trend visible in the plot, as the unemployment rate in each month appears to fluctuate around an average level, without showing a consistent upward or downward trend over time.



The above ACF graph helps to confirm the presence of a trend and a seasonal component in the data. The ACF related to the time series shows that all spikes are significant, which means that there is a trend pattern in the data.

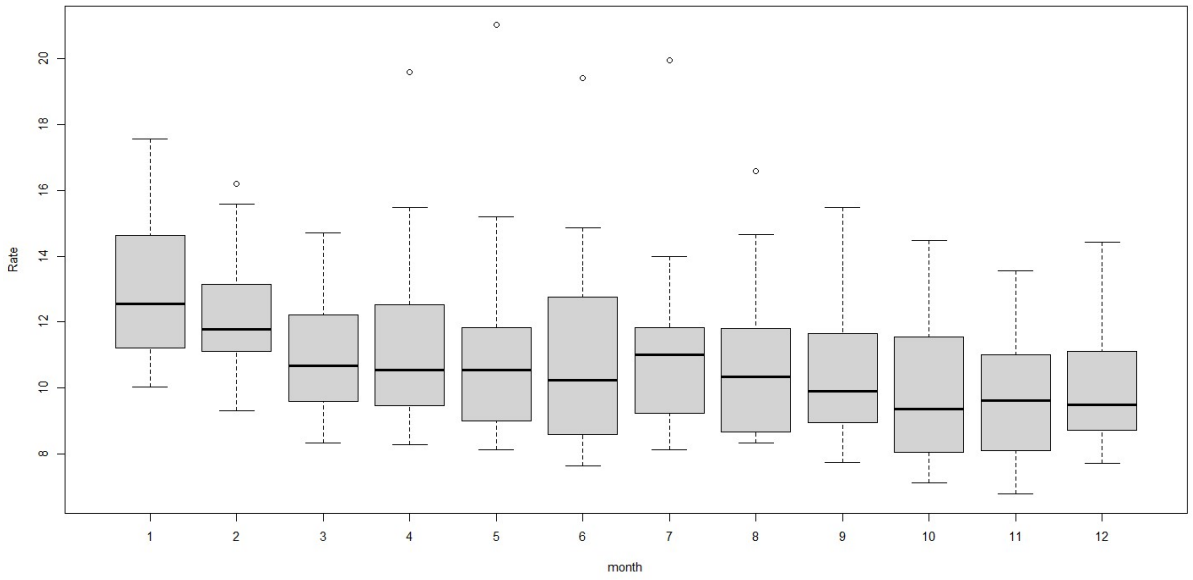
In the next graph, we can observe the seasonal and trend component of the time series, the trend component is decreasing over time till 2020 and increased till 2021 and again the curve started decreasing.



Finally, to describe the data we will get some summary statistics and the boxplot by month.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.774   9.078  10.485  10.919  12.215  21.013
```

Looking at the boxplot we can observe that mean value each month is a little different. October is the month with the lowest average unemployment rate value.



2. Transformations

I have performed the Augmented Dickey-Fuller (ADF) test for testing the null hypothesis of whether a time series is stationary or not. The ADF test is a statistical test that is commonly used to check for the presence of a unit root in a time series.

```
Augmented Dickey-Fuller Test
```

```
## data:  unemployment
```

```
## Dickey-Fuller = -2.7883, Lag order = 6, p-value = 0.2439
```

```
## alternative hypothesis: stationary
```

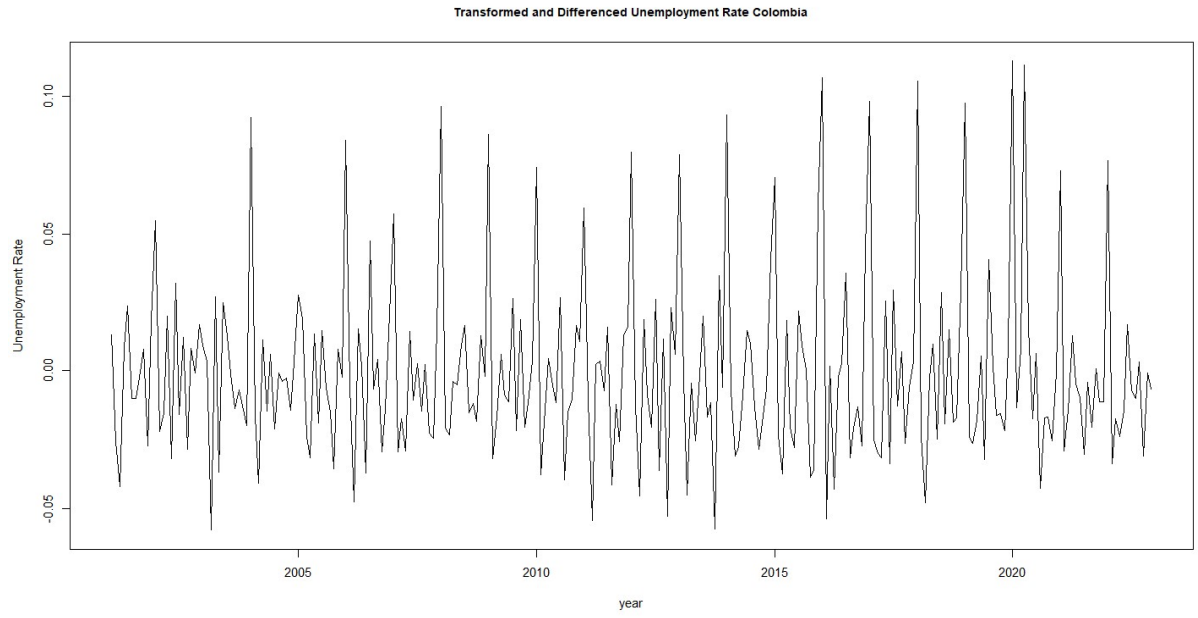
Since the p-value of the test is greater than the significance level of 0.05, we cannot reject the null hypothesis of non-stationarity and conclude that the bankrupt time series is non-stationary. Therefore, we need to apply a transformation or adjustment to the time series.

In order to suggest a transformation to stabilize the variance in the time series, I will use the Bow-Cox transformation to get the optimal value of Lamba:

```
BoxCox.lambda(unemployment)
```

```
## [1] -0.5055
```

The Box-Cox gives a value of Lambda -0.50. The optimal lambda value of -0.50 indicates that a negative power transformation is required to stabilize the variance of the time series data. So, I am using the first-order difference to remove the trend. This can help in making the time series stationary and suitable for further analysis using time series models.



After I applied the transformation I checked once again the adf test to find out whether p value is significant and the below results confirms that the time series is stationary based on the applied transformation.

```
Augmented Dickey-Fuller Test
```

```
## data:  unemp_diff
```

```
## Dickey-Fuller = -7.8928, Lag order = 6, p-value = 0.01
```

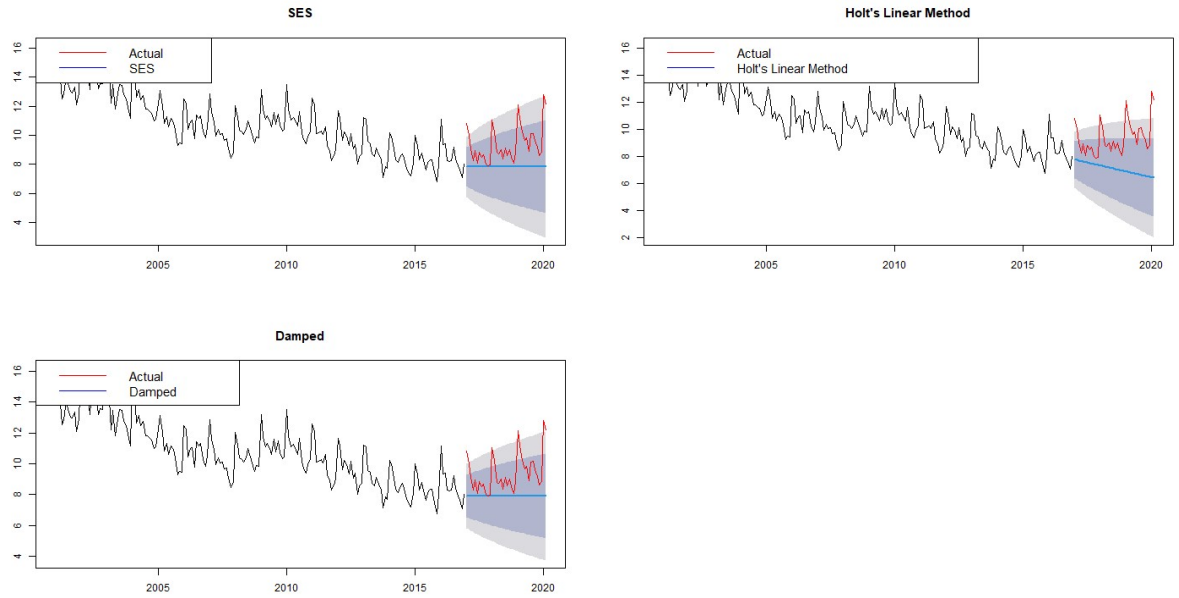
```
## alternative hypothesis: stationary
```

3. Forecast modeling

Now, we will forecast the unemployment rate using two models: ETS and ARIMA methods. In order to evaluate the performance of the different forecasts and compare the models I will divide the data in a training and test dataset. The training dataset contains information from January 2001 to December 2016 and the test dataset from January 2017 to February 2020. I have kept the data from March 2020 to December 2022 for my later analysis (during COVID Pandemic).

Forecast: Exponential Smoothing Methods

The below plot shows the forecasts made by Exponential Smoothing methods such as SES, Holt-Linear and Damped. The red line represents the actual test dataset and the blue colored line denotes the forecasts of different methods.

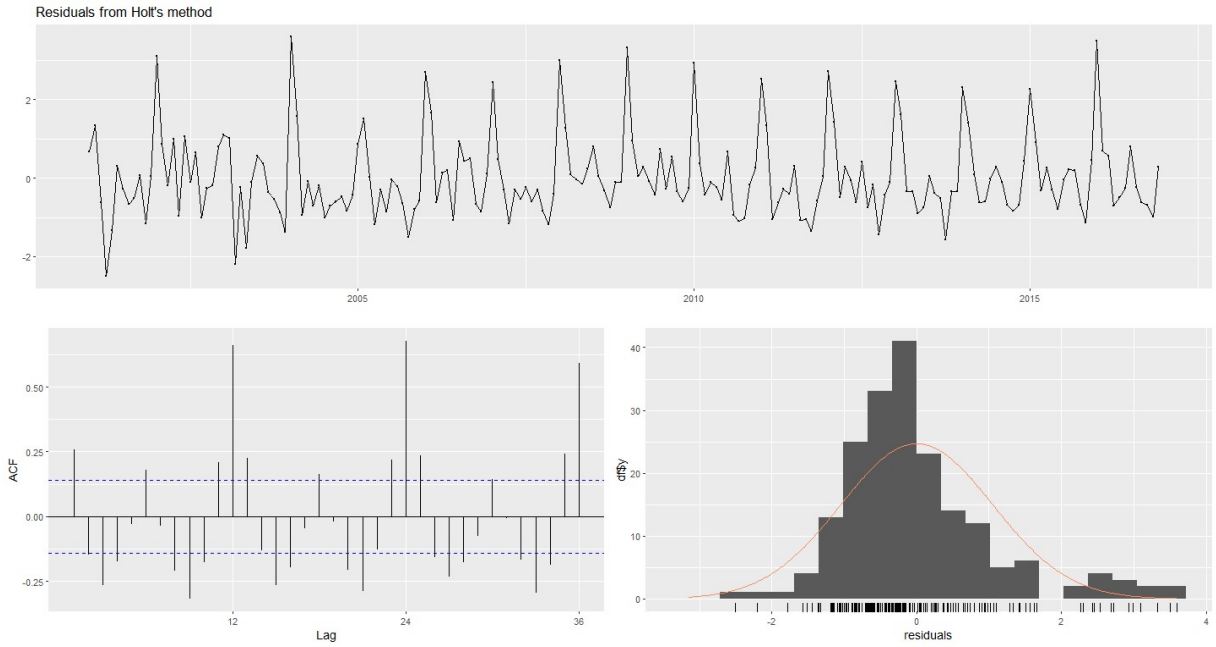


Checking the forecast accuracy and residuals we have the following results :

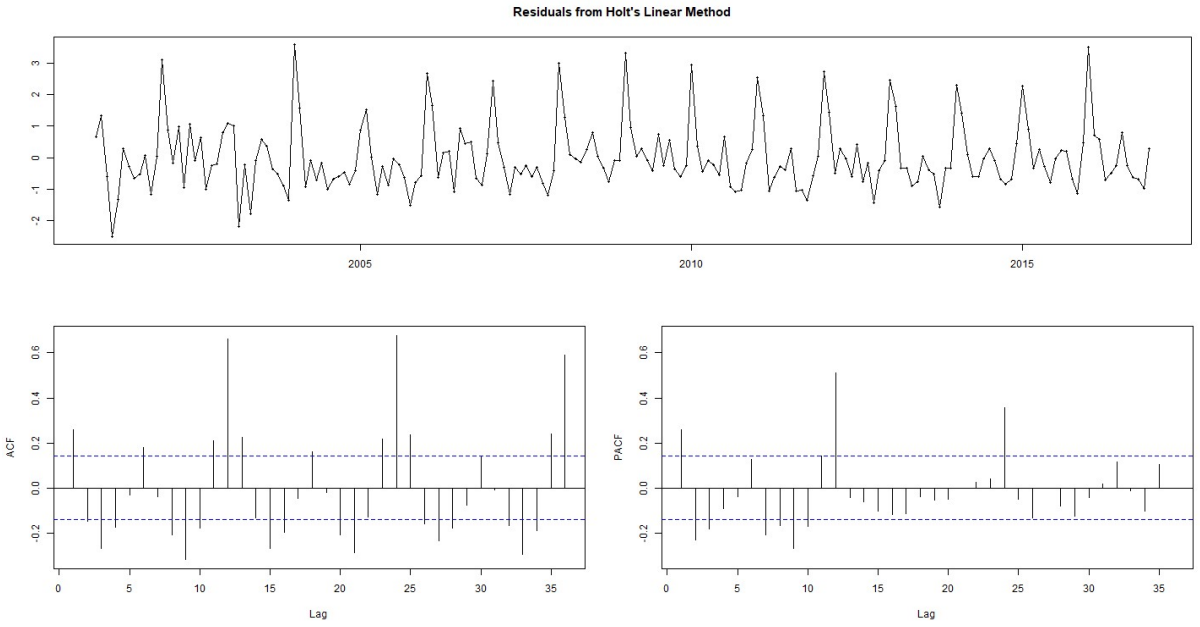
```
> es_df
                          Model Accuracy_Train.RMSE Accuracy_Train.MAE Accuracy_Train.MAPE Accuracy_Train.MASE Accuracy_Test.RMSE
1 SES           Exponential Smoothing          81.18879          64.27866           9.003926          0.9351193          110.4517
2 Holt's Linear Exponential Smoothing          81.18277          64.29786           9.010373          0.9353986          102.3765
3 Damped        Exponential Smoothing          60.76960          46.46273           6.395061          0.6759349          119.0959
                Accuracy_Test.MAE Accuracy_Test.MAPE Accuracy_Test.MASE Residuals.Q. Residuals.df Residuals.p.value
1 SES                    91.09621          10.845480          1.325258     286.5021           24      0.000000e+00
2 Holt's Linear          83.52577           9.976188          1.215124     286.5021           24      0.000000e+00
3 Damped                 97.65418          11.527148          1.420663     129.9367           24      1.110223e-16
>
```

We can say from the above results that **Holt's Linear** model is the best performing, as it has the lowest test residual error (MASE, RMSE, MAE, MPE) compared to the others, which indicates that it fits better on unseen data compared to other models. Also, we need to check some other models as well to get the final forecast predictions.

Checking the residuals we have the following results :




```
##
##  Ljung-Box test
##
## data:  Residuals from Holt's method
## Q* = 366.06, df = 24, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 24
```



Looking at the ACF, we can observe that we have significant spikes in many lags (3, 4, 6, 8..) what it means that we still have information in the residuals that we are not considering in the Holt's Linear model and it would not be reflected in the forecasts.

The Ljung Box test suggests that we can reject the null hypothesis of white noise in the residuals since we have p-values very close to zero in the all lags. This confirms that we still have information in the residuals of the Holt's Linear model and we are not capturing completely the data generation process.

Forecast: ARIMA

I will fit an ARIMA model using the function auto.arima from R, this function returns the best model base on the AIC criteria. We can see from the below result that the AUTO ARIMA model (ARIMA(2,0,2)(2,1,0)[12] can be described as the non-seasonal order with an autoregressive term of 2, no differencing (d=0), and a moving average term of 2 and the seasonal order with an autoregressive term of 2, a first-order differencing (D=1), and no moving average term with a seasonal period of 12 (denoted by the [12] parameter).

The summary result of AUTO Arima Model is as follows :

```
> summary(auto_arima)
Series: unm_train
ARIMA(2,0,2)(2,1,0)[12]

Coefficients:
         ar1      ar2      ma1     ma2     sar1     sar2
      1.5395  -0.5568  -1.3236  0.5508  -0.7513  -0.2999
s.e.  0.1807   0.1780   0.1701  0.1262   0.0784   0.0859

sigma^2 = 0.3717:  log likelihood = -167.41
AIC=348.82   AICc=349.47   BIC=371.17

Training set error measures:
                     ME      RMSE       MAE        MPE     MAPE      MASE        ACF1
Training set -0.04719237 0.5803565 0.4410456 -0.5844044 4.166001 0.5672924 0.007912007
```

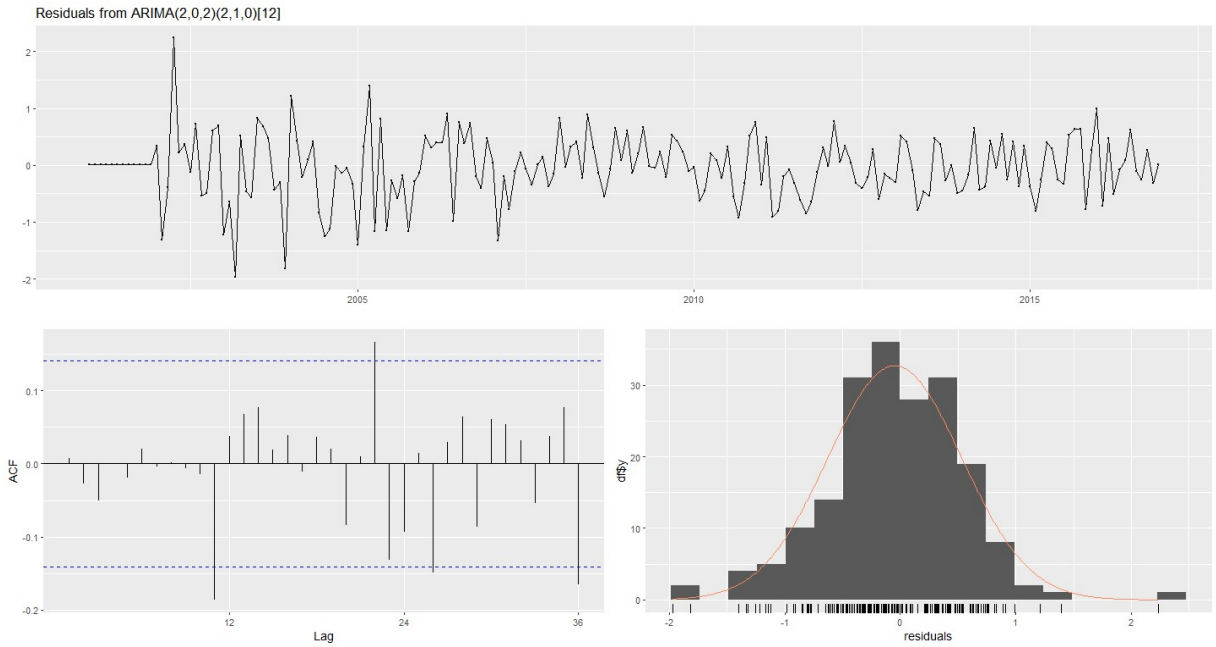
The generated model is an ARIMA(2,0,2)(2,1,0)[12]. This model has 6 estimated parameters and they are significant.

Checking the forecast accuracy we have the following results:

```
##                        ME      RMSE       MAE        MPE      MAPE      MASE
## Training set   -0.04719237 0.5803565 0.4410456 -0.5844044 4.166001 0.5672924
## Test set        0.87363291 1.1010484 0.8940107  8.9191310 9.161808 1.1499163
##                     ACF1 Theil's U
## Training set 0.007912007        NA
## Test set     0.580282688 0.8524172
```

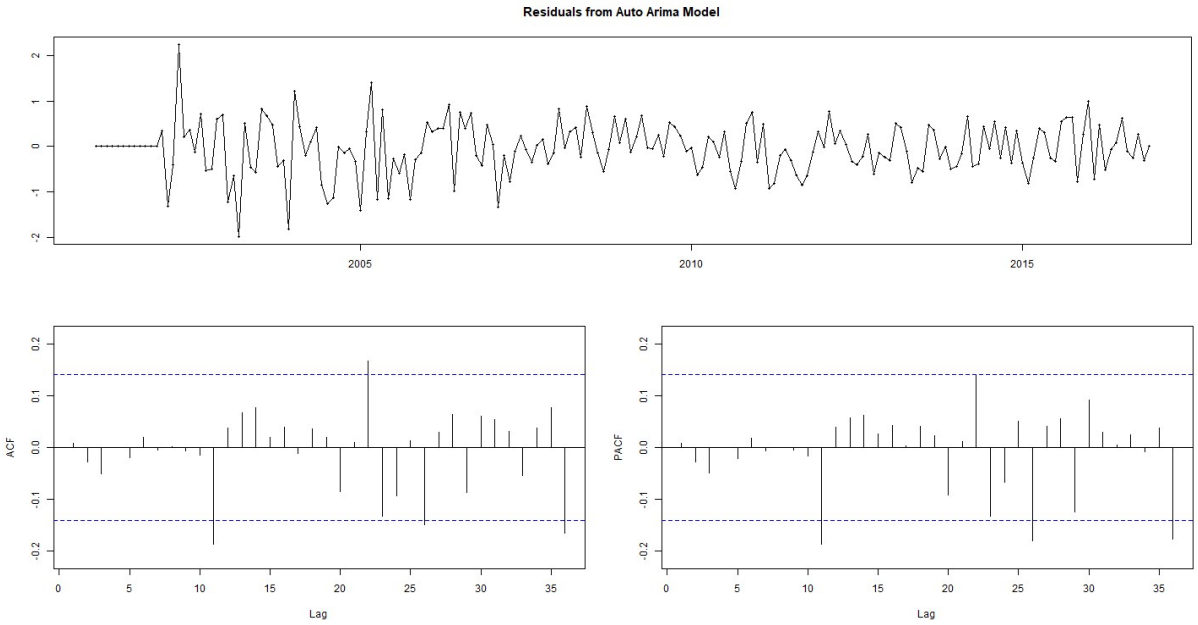
The model shows a better performance in the training set than the test set. Later, we will use these metrics to compare this model with the ETS model.

Checking the residuals we have the following results:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,0,2)(2,1,0)[12]
## Q* = 24.745, df = 18, p-value = 0.1321
##
## Model df: 6.   Total lags used: 24
```

Looking at the previous plot, one can suggest that the residuals are not around zero and they do not follow a normal distribution, there are some peaks and it seems that there is a fat-tailed distribution.

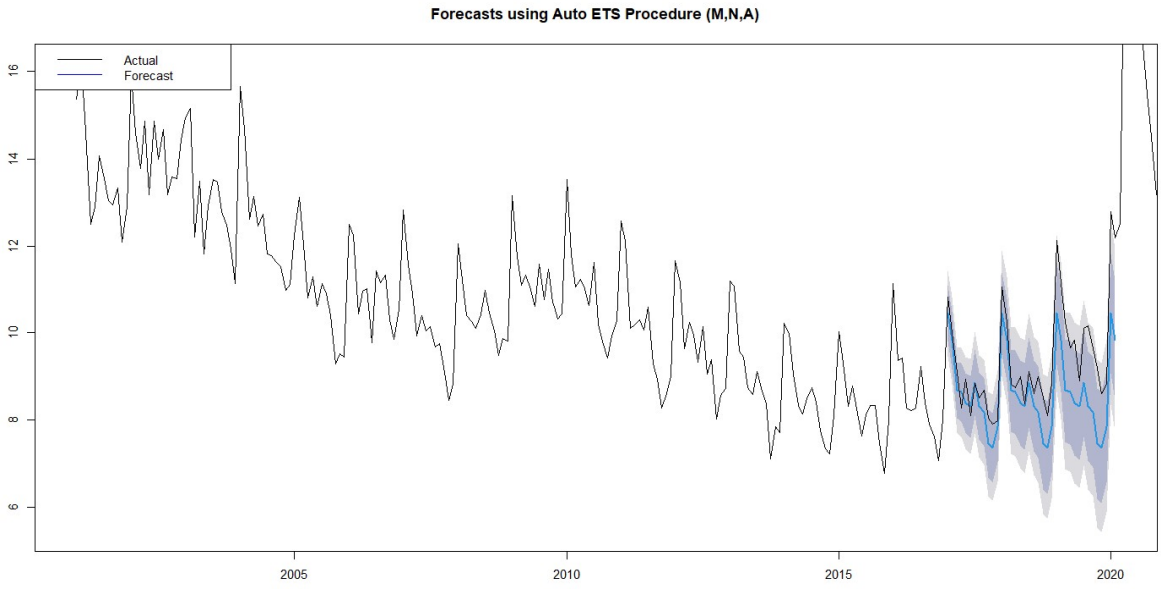


The ACF of the residuals suggests that we still have information that we are not considering in the ARIMA model since there are significant spikes in some lags (11, 22).

The Ljung Box test indicates that we can reject the null hypothesis of white noise in the residuals since we have some p-values smaller than 0.05. This confirms that we still have information in the residuals of the model and we are not capturing completely the data generation process.

Forecast: ETS

For building an ETS model, I will use the function ETS from R. In the next plot we can observe the forecast made by this model.



The result is a method "M,N,A" (Multiplcative error, no trend, and additive seasonal component). This function returns automatically the best ETS method based on some criteria as AIC, MSE, etc.

The parameters associated to the model are

```
##         alpha         gamma             l            s0            s1
##        0.3749         2e-04       13.9533       -0.6709       -1.1571
##            s2            s3            s4            s5            s6
##       -1.070       -0.3489       -0.2082        0.3243       -0.2179
##            s7            s8            s9           s10
##        -0.135        0.1082        0.1502        1.3058
```

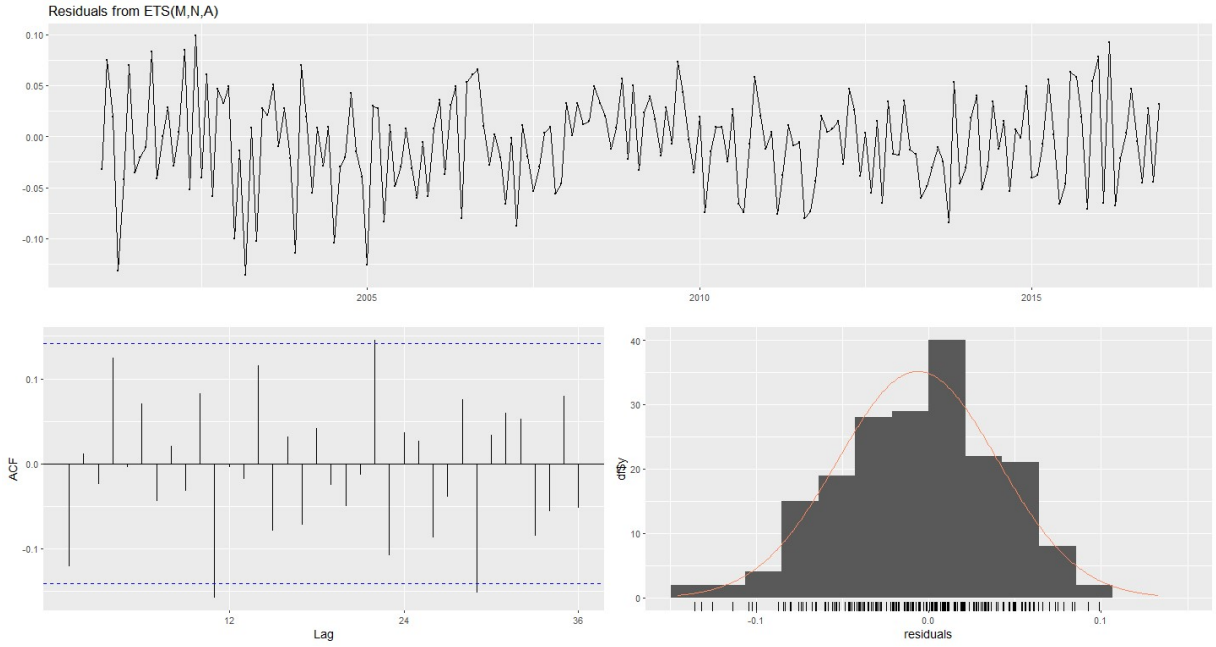
The alpha value indicates that the model is not taking into account only the last observations, also it is taking older information to create the data generation process.

Checking the forecast accuracy we have the following results:

```
##                       ME      RMSE       MAE         MPE     MAPE      MASE
## Training set -0.07539471 0.5467826 0.4170203 -0.8662748 3.851604 0.5363901
## Test set      0.77860633 1.0253108 0.8098249  7.8650511 8.244421 1.0416327
##                    ACF1 Theil's U
## Training set -0.1463966        NA
## Test set      0.6564853 0.8018988
```

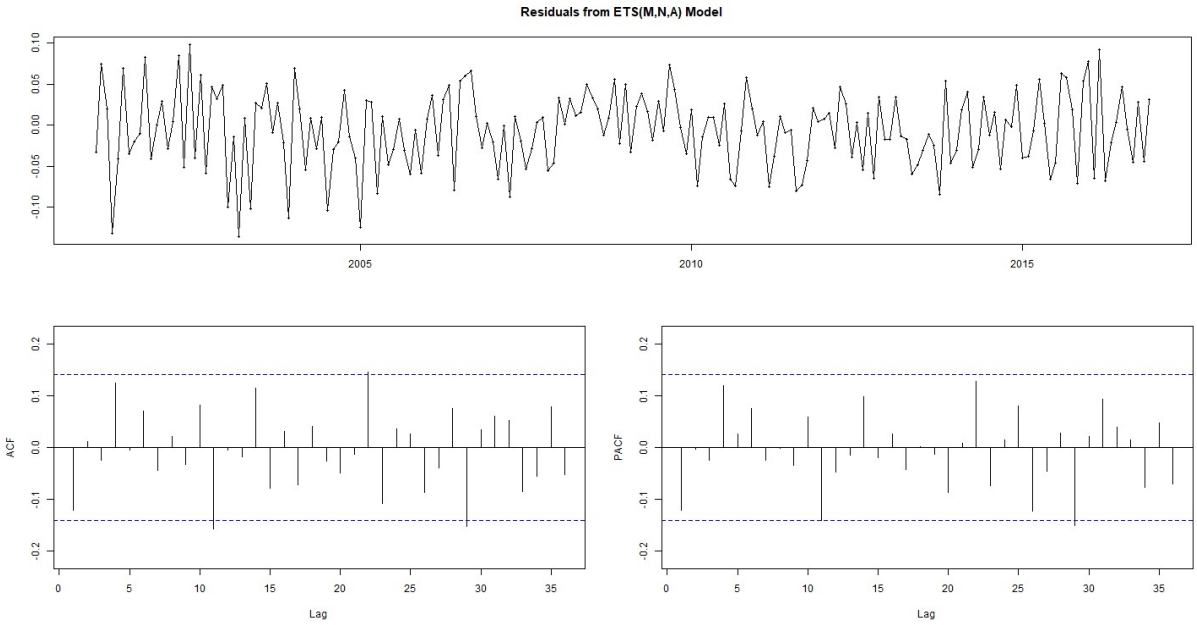
The model shows a better performance in the training set than the test set. If we compare these results with the accuracy metrics from the ARIMA model, we can observe that the performance is better in the ETS model. All metrics are smaller in the ETS models, and in the tess dataset some of them (MPE,MAPE), the difference is notable.

Checking the residuals we have the following results:




```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(M,N,A)
## Q* = 28.413, df = 24, p-value = 0.243
##
## Model df: 0.   Total lags used: 24
```

Looking at the previous plot, we can suggest that the residuals are around zero and they do not follow completely a normal distribution, It seems that the distribution is skewed.

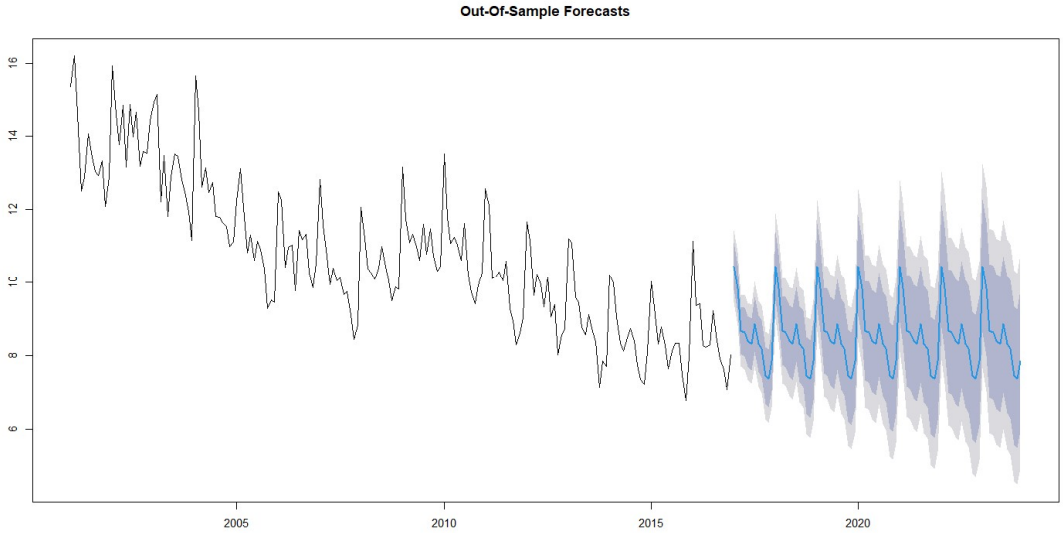


Looking at the ACF of the residuals, we can observe that we have significant spikes in some lags (11, 22, 24) what it means that we still have information in the residuals that we are not considering in the ETS("MNA") model and it would not be reflected in the forecasts.

The Ljung Box test suggests that we can reject the null hypothesis of white noise in the residuals since we have p-values that are almost zero in all lags. This confirms that we still have information in the residuals of the model and we are not capturing completely the data generation process.

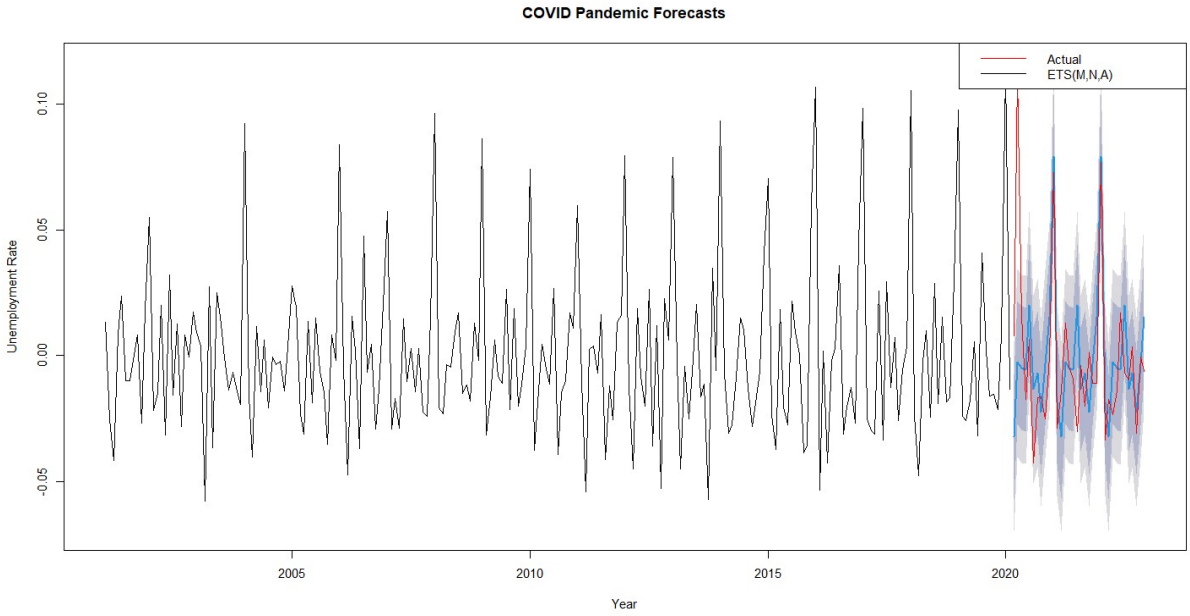
4. Out of Sample Forecasts: Model ETS (M,N,A)

Based on my analysis of the unemployment rate in Colombia dataset using different models, I have chosen the ETS("MNA") model for predicting the behavior of this variable. The residuals diagnostic for Auto Arima model is better compared to ETS but still in both models there may be some underlying information that is not being captured. Hence, I support my conclusion in terms of accuracy of my chosen model based on the results. The forecast plot displays the predicted values for the dataset up to December 2023.



5. Covid Impact :

We can clearly see from the below plot results that Covid changed the usual trend of unemployment rate in Colombia especially in the month of May 2020. High unemployment rates were reported during the post pandemic period than we have forecasted based on the historical data. The red line indicates the actual values where we can see that in May 2020 there are high percentage of unemployment rate but our forecasted results are less during these months.



References

1. BANREP (Banco de la Republica) Colombia - https://totoro.banrep.gov.co/estadisticas-economicas/faces/pages/charts/line.xhtml?facesRedirect=true
2. Colombia Unemployment Rate - https://tradingeconomics.com/colombia/unemployment-rate
3. https://www.analyticsvidhya.com/blog/2021/11/performing-time-series-analysis-using-arima-model-in-r/
4. https://towardsdatascience.com/an-introduction-to-time-series-analysis-with-arima-a8b9c9a961fb