Zambon Pharma

Product sales data analysis and forecasts

Authors:
Cecilia Tomba
Giulio Nebbiai

Course:
Business, Financial and Economical data
Professor:
Mariangela Guidolin

Presentation Outline

1 INTRODUCTION

2 EDA

3 DATA ANALYSIS

4 DATA MODELLING

5 FORECASTS

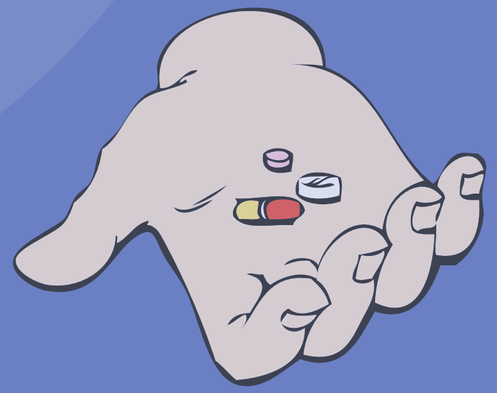
6 CONCLUSIONS

Who is Zambon?

"We are a modern chemical-pharmaceutical multinational built upon the history and values of an Italian family-owned business, with ambitious plans for growth and development."

FOCUS

- Neurological Disorders
- Respiratory System Diseases
- Urological System
- Pain
- Gastrointestinal Area



Data Provided

Dataset 1

2019-01 to 2023-10 monthly observations about:

- Sales
- Sold Units
- Market Sales
- Market Units

of product 1

Dataset 2

2019-01 to 2023-10 monthly observations about:

- Sales
- Sold Units
- Market Sales
- Market Units

of product 2

The problem

For privacy reasons the company didn't provide us the **names** and the **cathegory** of the two products, neither the **countries** where they were sold.

The only provided infos were:

- One of the two products is a "**seasonal**" one, so relatable to flu and respiratory diseases
- Both the products were sold in a **central european country**
- We assumed **Italy** as sample country

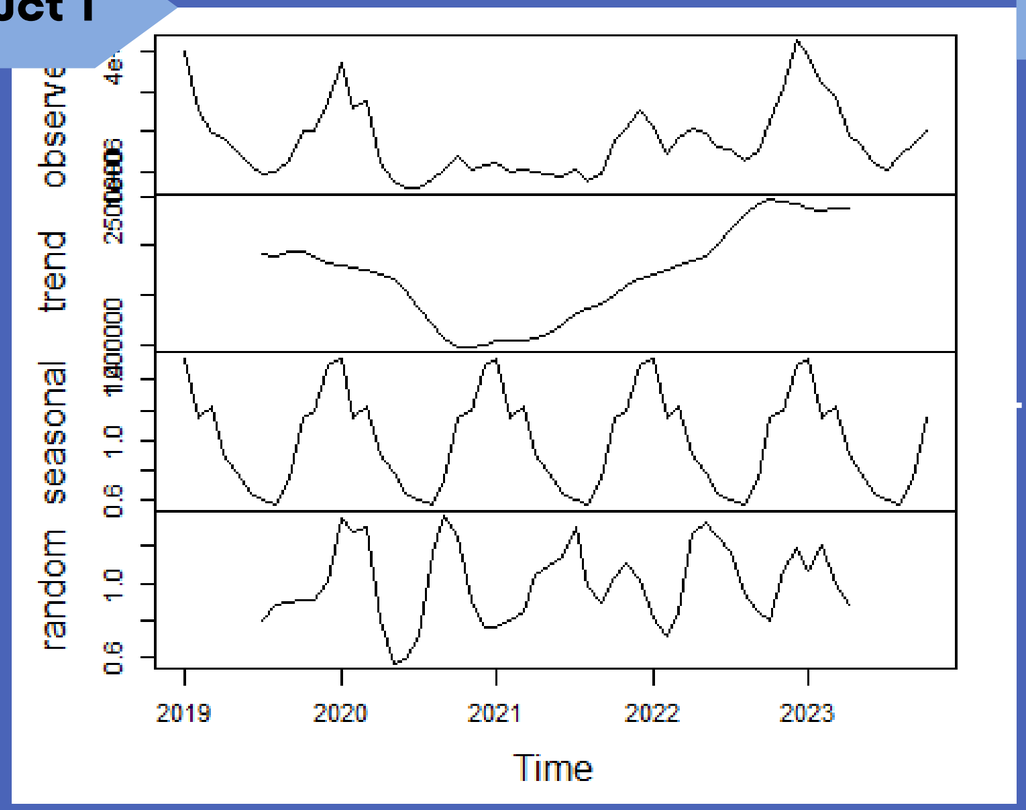
These **limitations** didn't allow us to find all the possibile **covariates** which could affect the phenomena

The Challenges

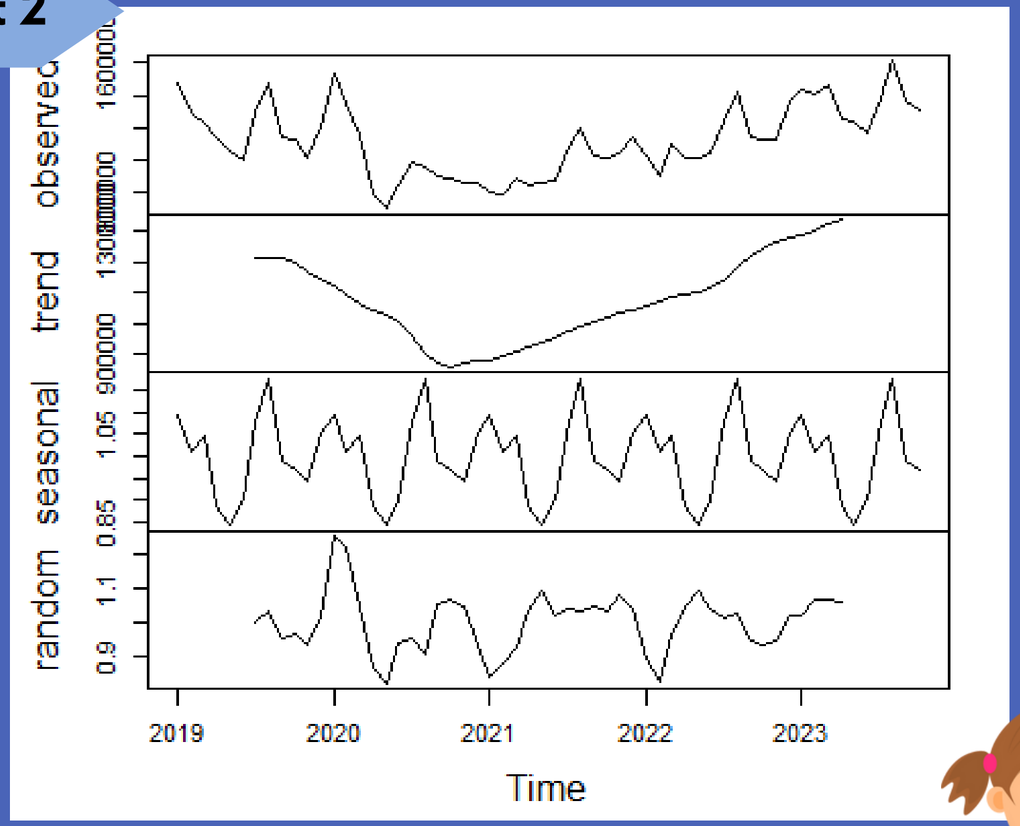
- Finding out the seasonal product
- Looking for the right covariates
- Guessing the product

We take a look to the seasonal habit of the two sales times series

Product 1



Product 2



Product 1 is the one whose seasonality is strongly relatable to yearly seasons and so respiratory diseases. For this reason we decide to focus our analysis on this product.

At this point we can define our OBJECTIVES

- Estimate the model which best describes the phenomenon

- Find the proper exogenous variables that could have affected the phenomenon of sales

- Produce midlle term ex-ante forecasts assuming the dgp doesn't change

By the end of the work we will be able to provide to Zambon predictions (with a level of uncertainty) about **product 1 future sales from November to March** (most consistent months)

Our ultimate dataset

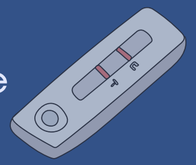
Contains 58 monthly observations of 10 variables

- Sales -> variable of interest
- Units -> number of units sold
- Market Sales -> sales by the target market
- Market Units -> number of units sold by the target market
- Covid positivities -> number of people currently affected by COVID
- Flu incidence -> Indicator related to the percentage of the population with flu
- Temperature -> monthly average temperature
- Zambon Google trends -> to check the popularity of the company
- Unemployment rate -> ratio of the unemployed to the corresponding labour force
- Old-age dependency ratio - > percentage ratio of the population aged 65 years over to the population aged 0-14 years (annual)

SOURCES

Covid positivities

https://lab24.ilsole24ore.com/coronavirus/?refresh_ce



Flu incidence

https://respivirnet.iss.it/pagine/rapportoInflunet.aspx

Temperature

https://www.ilmeteo.it/portale/archivio-meteo?refresh_cens

Old age ratio

https://www.istat.it/

Unemployment rate

https://www.istat.it/

EDA:

Exploratory Data Analysis

Data Pre-Processing

Infer Missing value

1

Data Overview

Turn all variables into Times Series and presentation

2

Correlations

Investigate any possible relantionship between predictors and response

3

Data Pre-Processing

Infer missing values for Flu incidence

Missing Observations:

9 observations

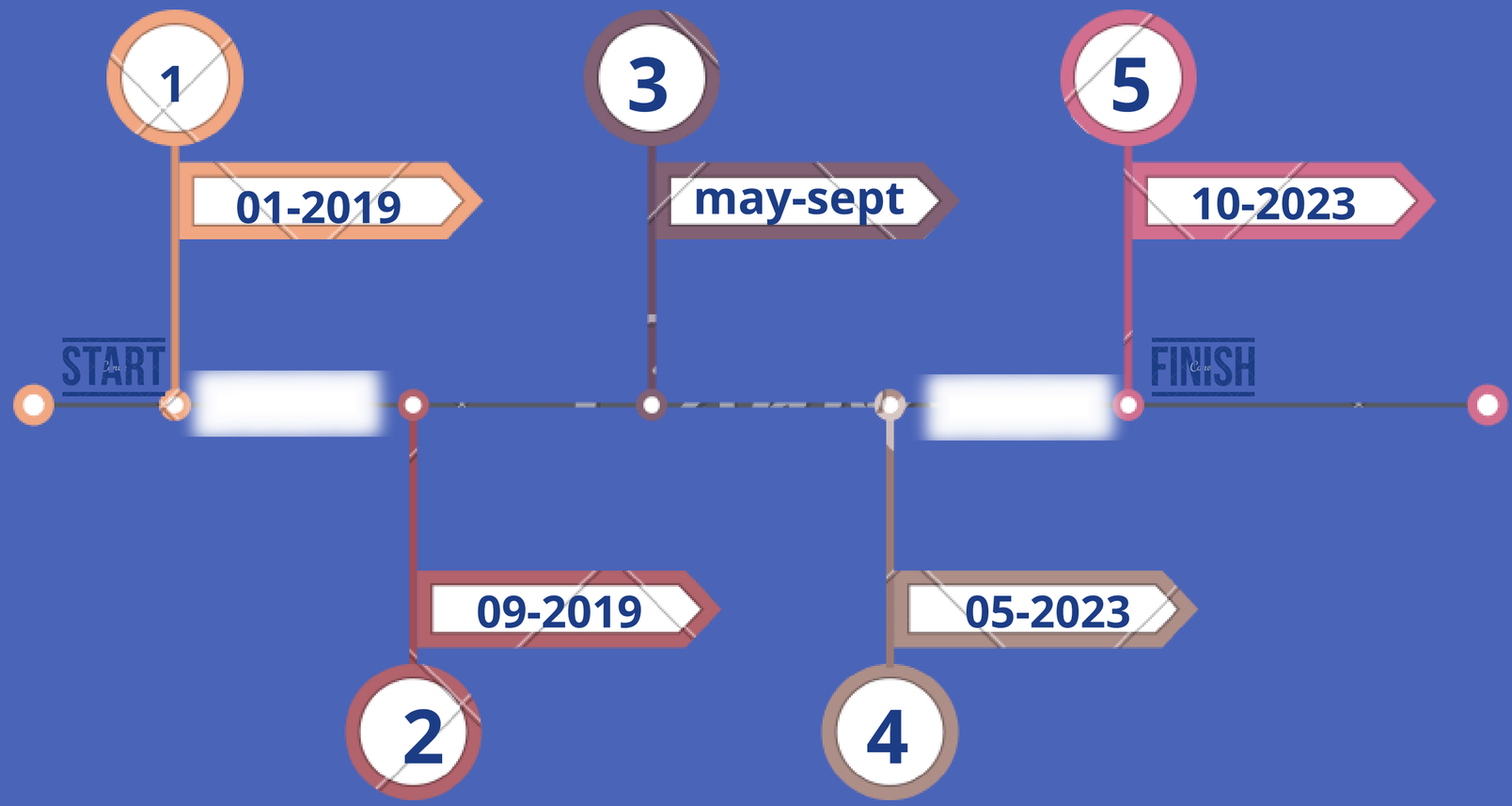
At the beginning of the series: observations missed from January 2019 to September 2019

5 observations

Observations missed from May to September every year

6 observations

At the end of the series: Observations missed from May 2023 to October 2023



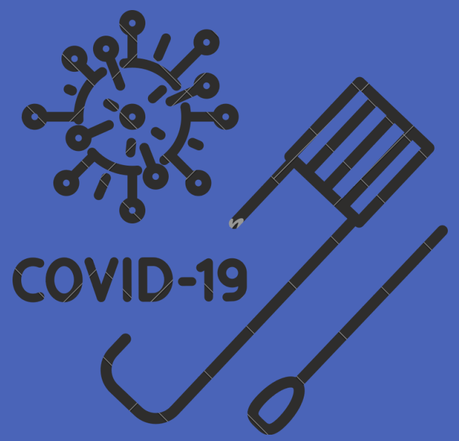
Data Pre-Processing

We provided different methods in order to infer missing values in the 2 tails:

Linear regression + **COVID**, **TEMP**

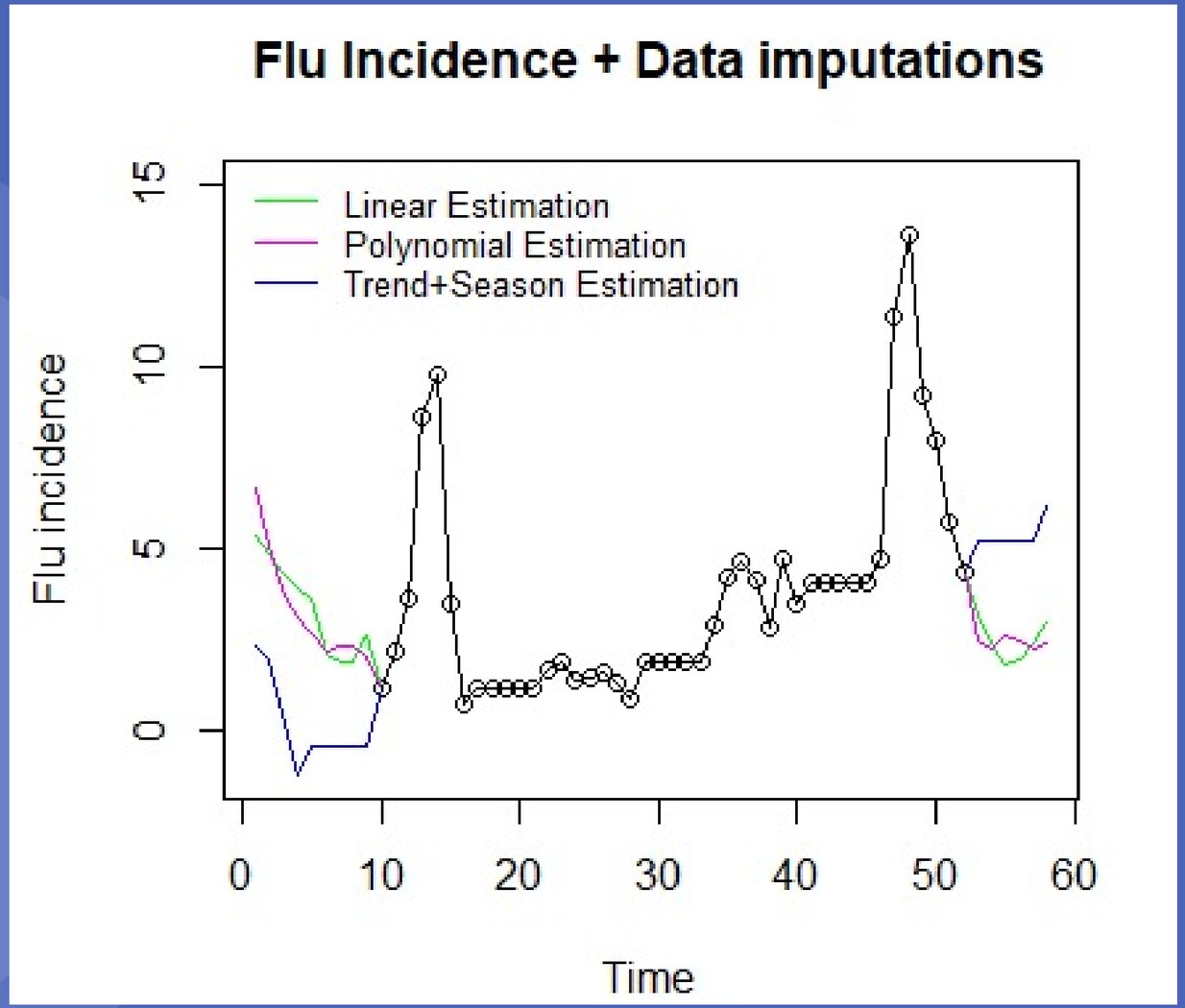
Polynomial regression + **COVID^2**, **TEMP^2**

TS Regression + **TREND**, **SEASON**



Data Pre-Processing

Data comparison and imputation



Multiple R-squared

LR	0.14
PR	0.19
TREND + SEASON	0.5

1 Data Pre-Processing

Data comparison and imputation

Multiple R-squared
Really low in all models

We select:

Predictions 01/2019 - 10/2019								
1	2	3	4	5	6	7	8	9
6.67	5.06	3.71	3.1	2.7	2.16	2.31	2.33	2.04

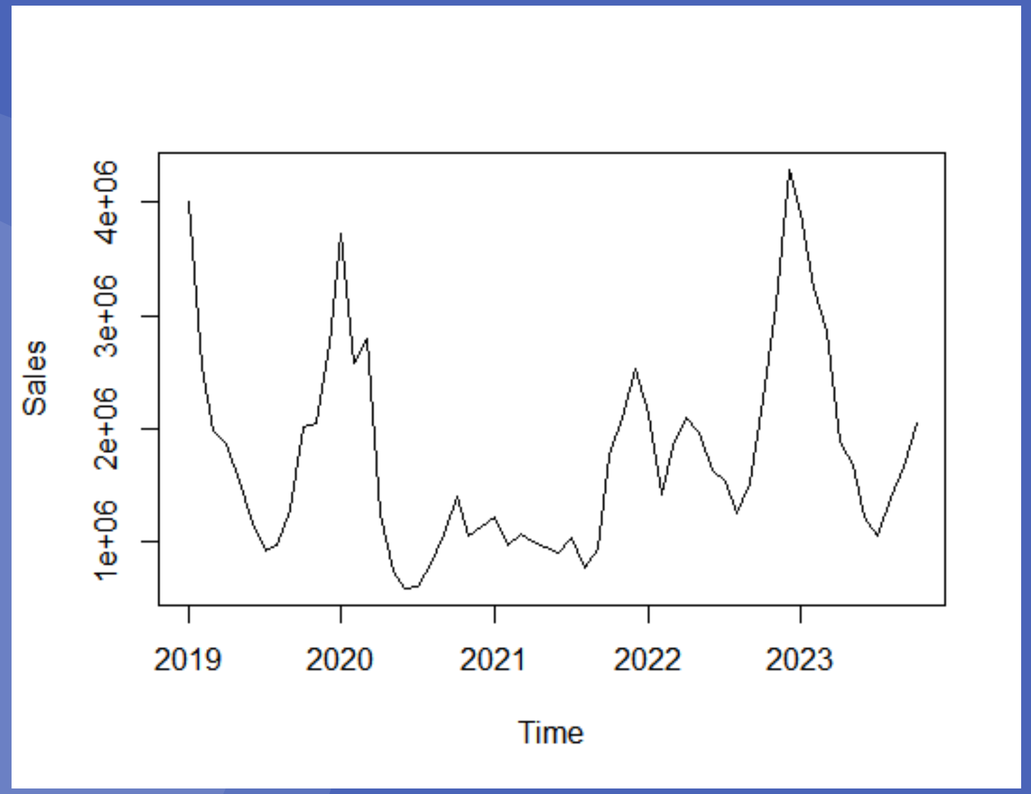
Predictions 5/2023 - 10/2023					
1	2	3	4	5	6
2.47	2.24	2.62	2.46	2.28	2.39

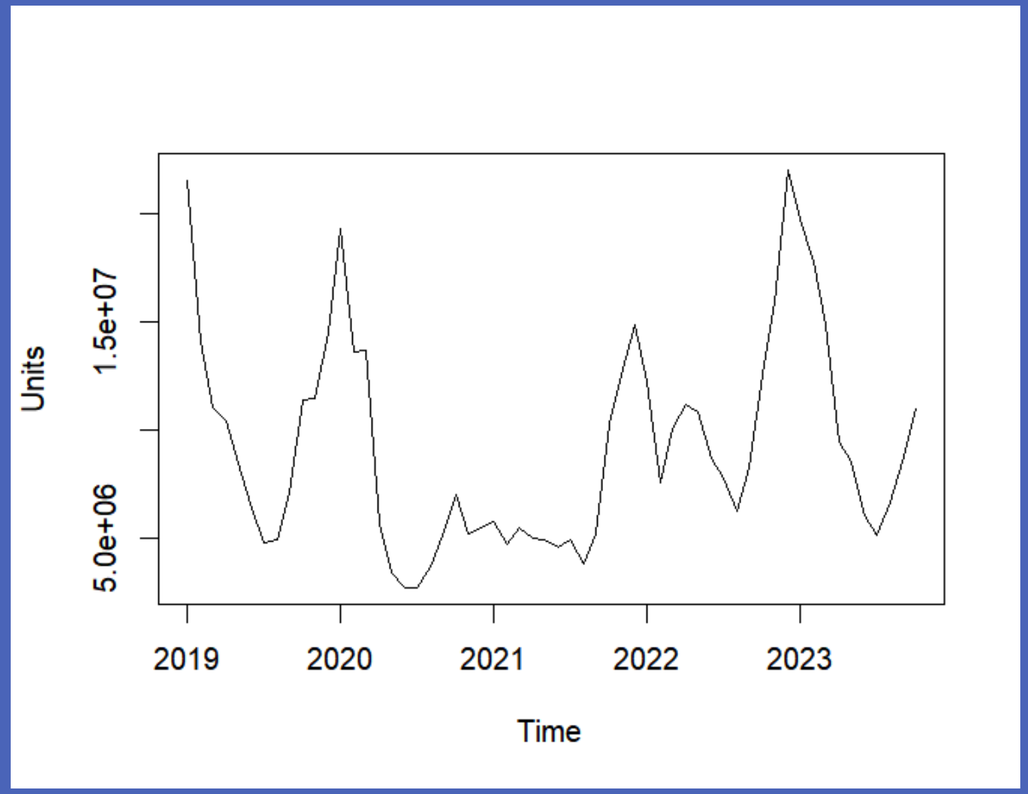
Polynomial Model

- Slightly better R^2 compared to Linear regression
- Looking at the graph, it seems to do not differ so much from original data
- We used this model to estimate value the tail of the series
- Instead we use the mean as method to infer values in the middle of the period and maintain them costant all over the period

Data Overview

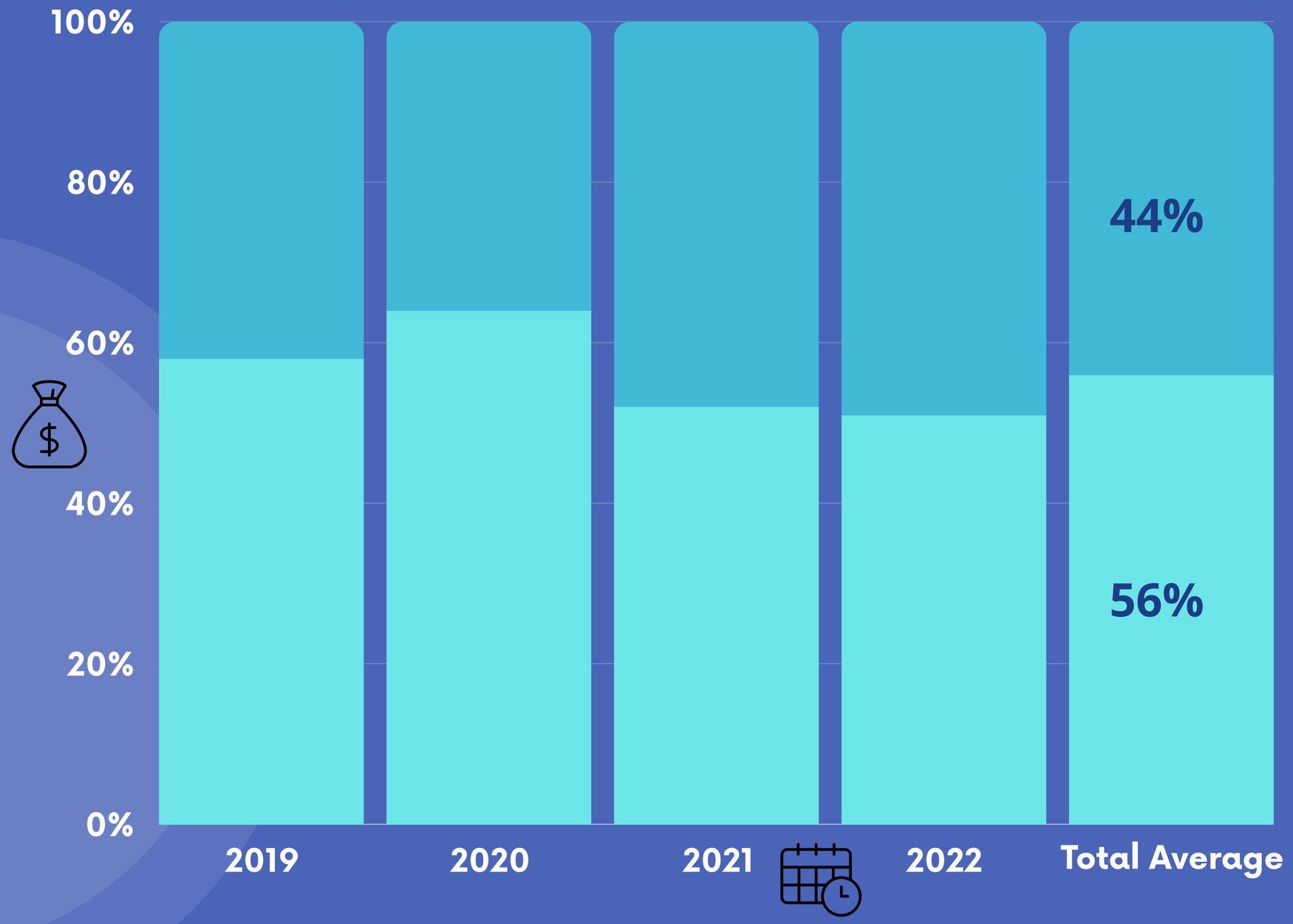
Zambon Times series





Because of the COVID pandemic it's hard to find a strong regularity in both the series

Data Overview



Zambon's annual average sales

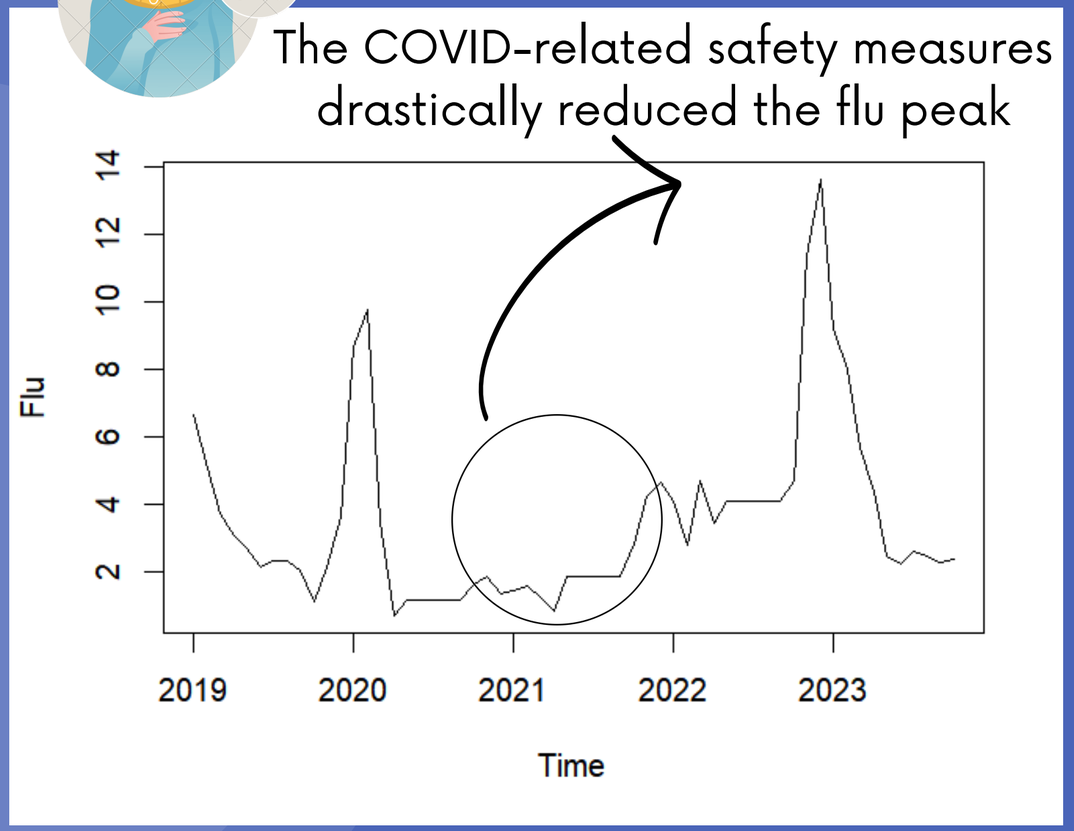
7 less relevant months: From April to October

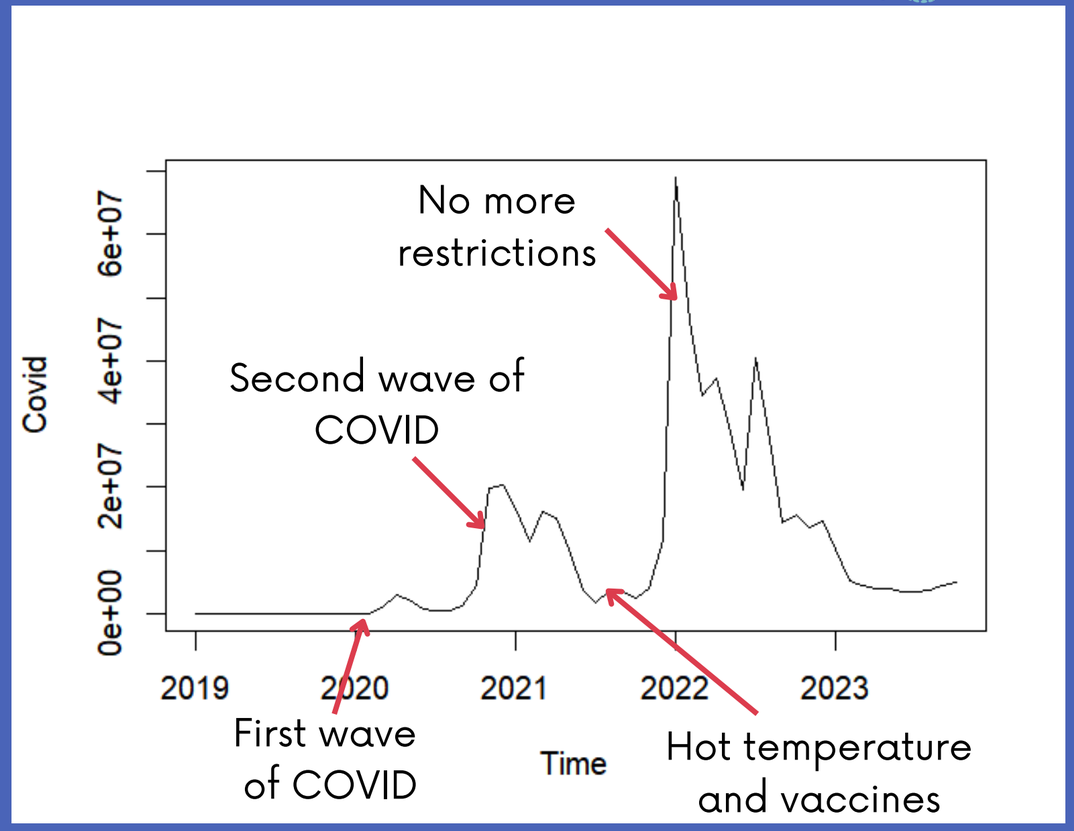
5 most relevant months: From November to March

(Greatest part of the nevenue comes from this period)

2 Data Overview

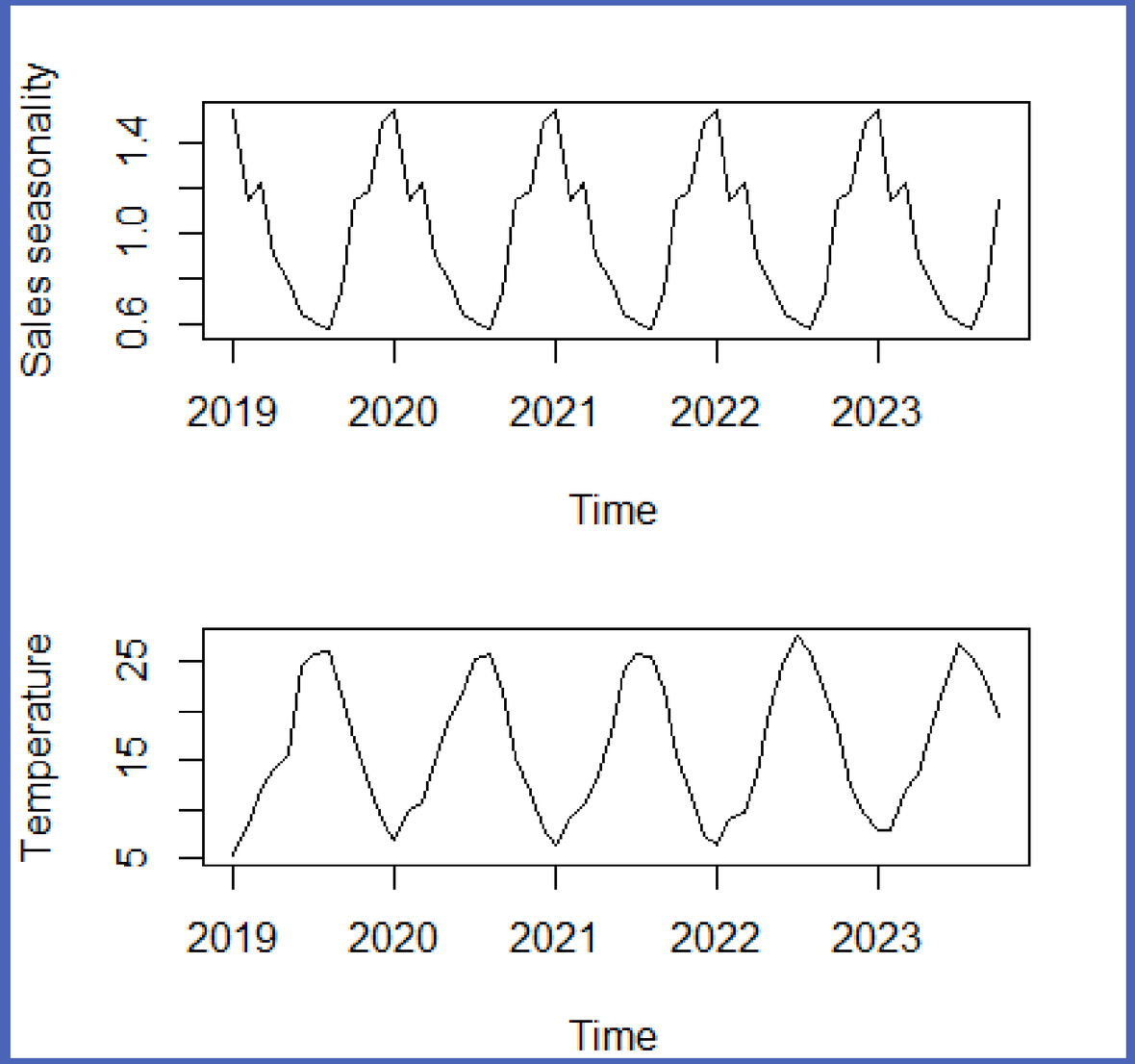
Health Times series





Data Overview

Health Times series



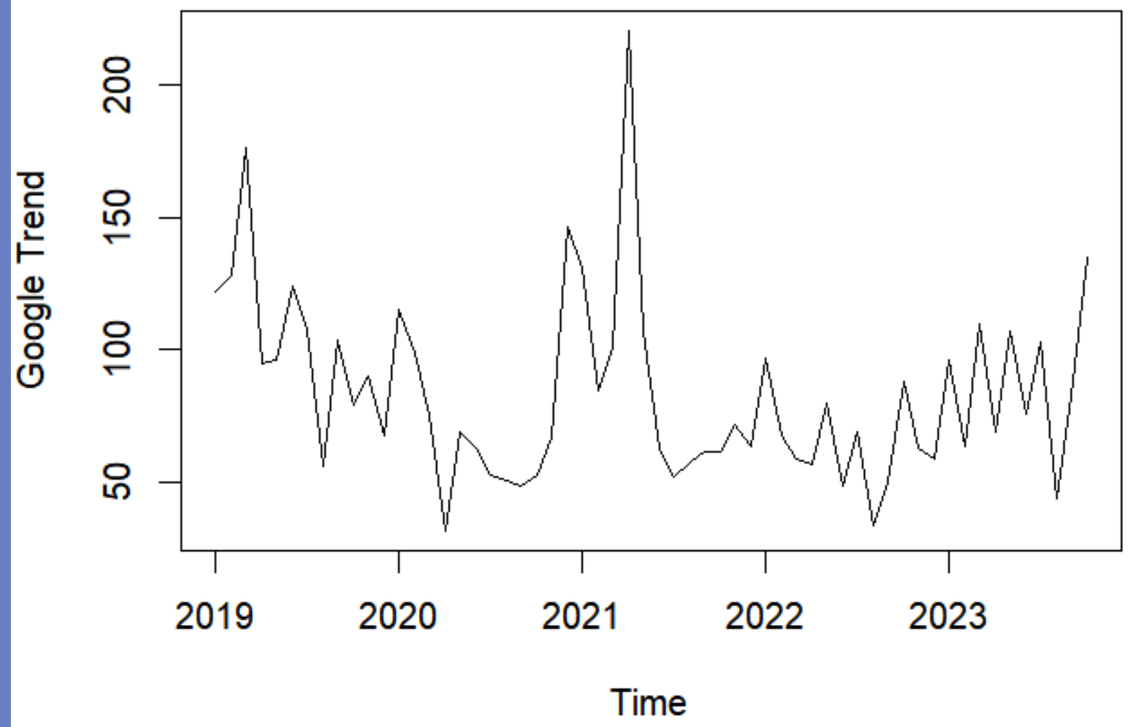
Confirm of the seasonality of the product

2 Data Overview

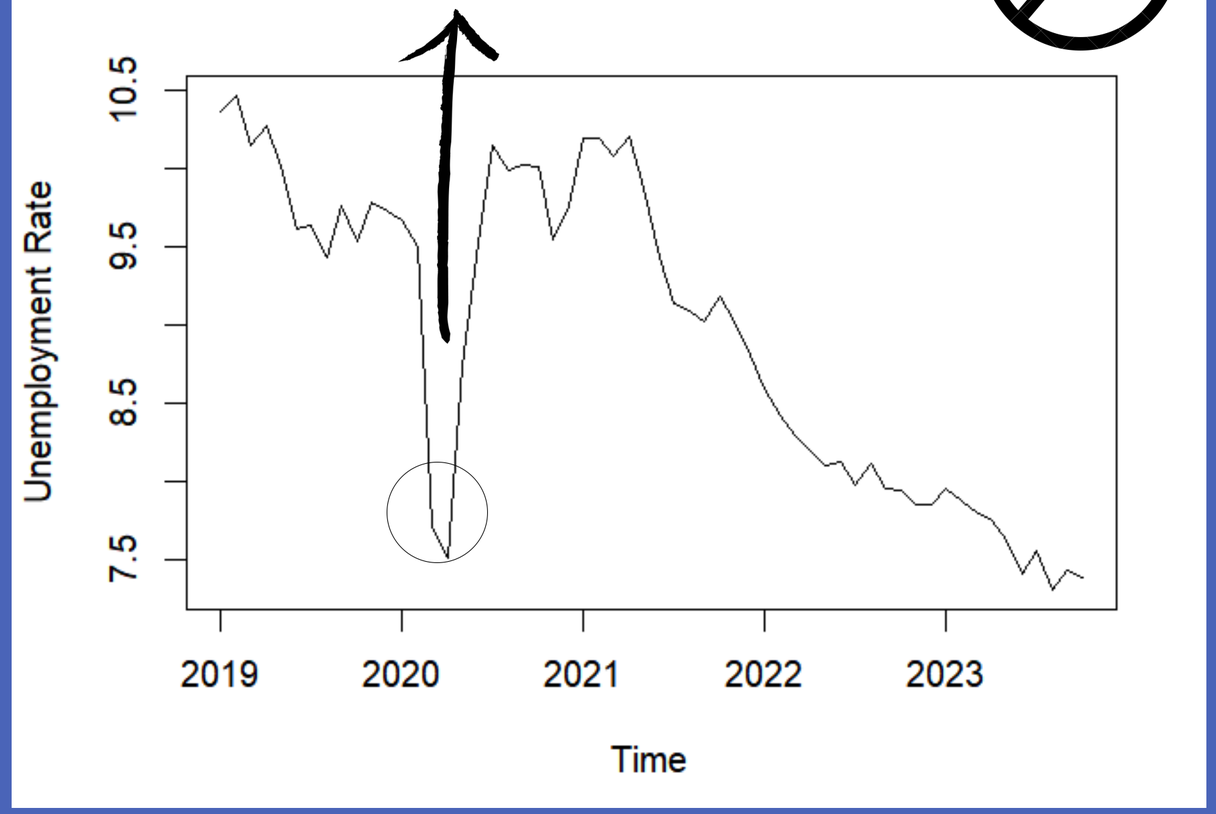
Socio-Economic Times series

Check of the popularity of the brand on the web

No presence of a trend

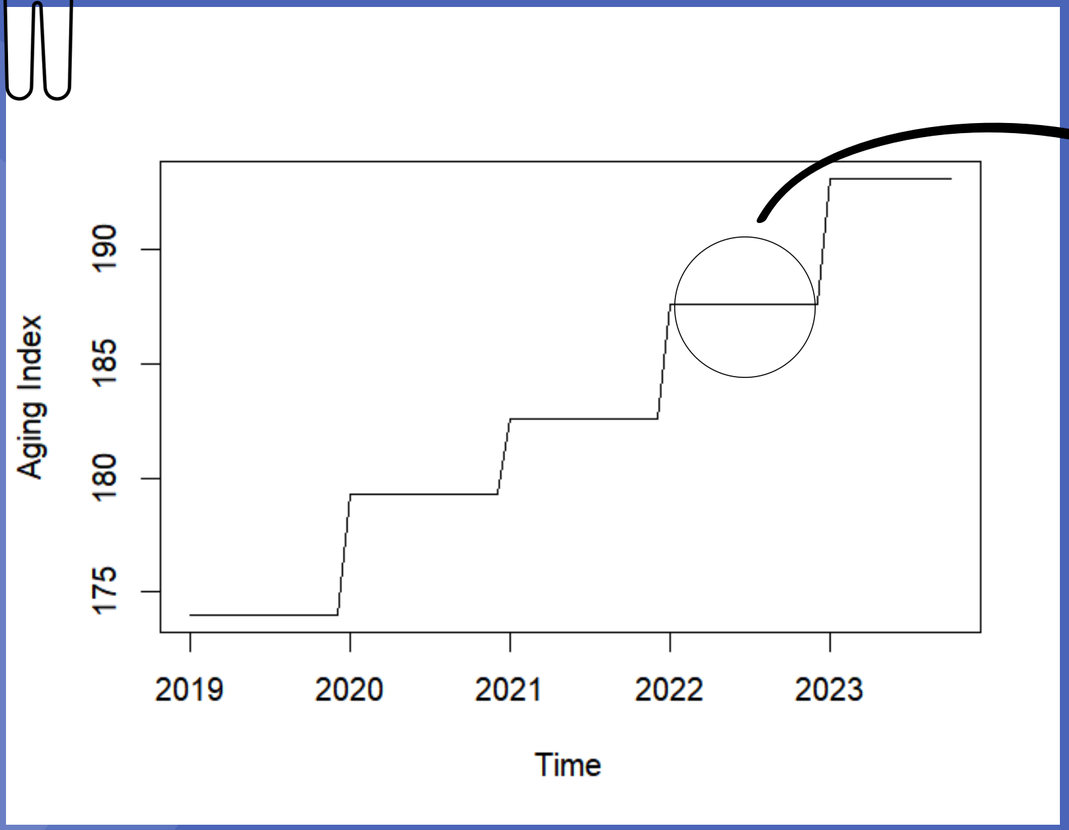


Sudden decrease in the unemployment rate during the initial period of Covid-19



2 Data Overview

Socio-Economic Times series



- Annual Observations, held constant all over the year
- Increasing Trend

Correlations

Compare :
1. Sales
2. Units

	Sales	Units	Market Sales	Market Units
Sales	1	0.998	0.993	0.992
Units	0.998	1	0.993	0.994
Market Sales	0.993	0.993	1	0.999
MarketUnits	0.992	0.994	0.999	1

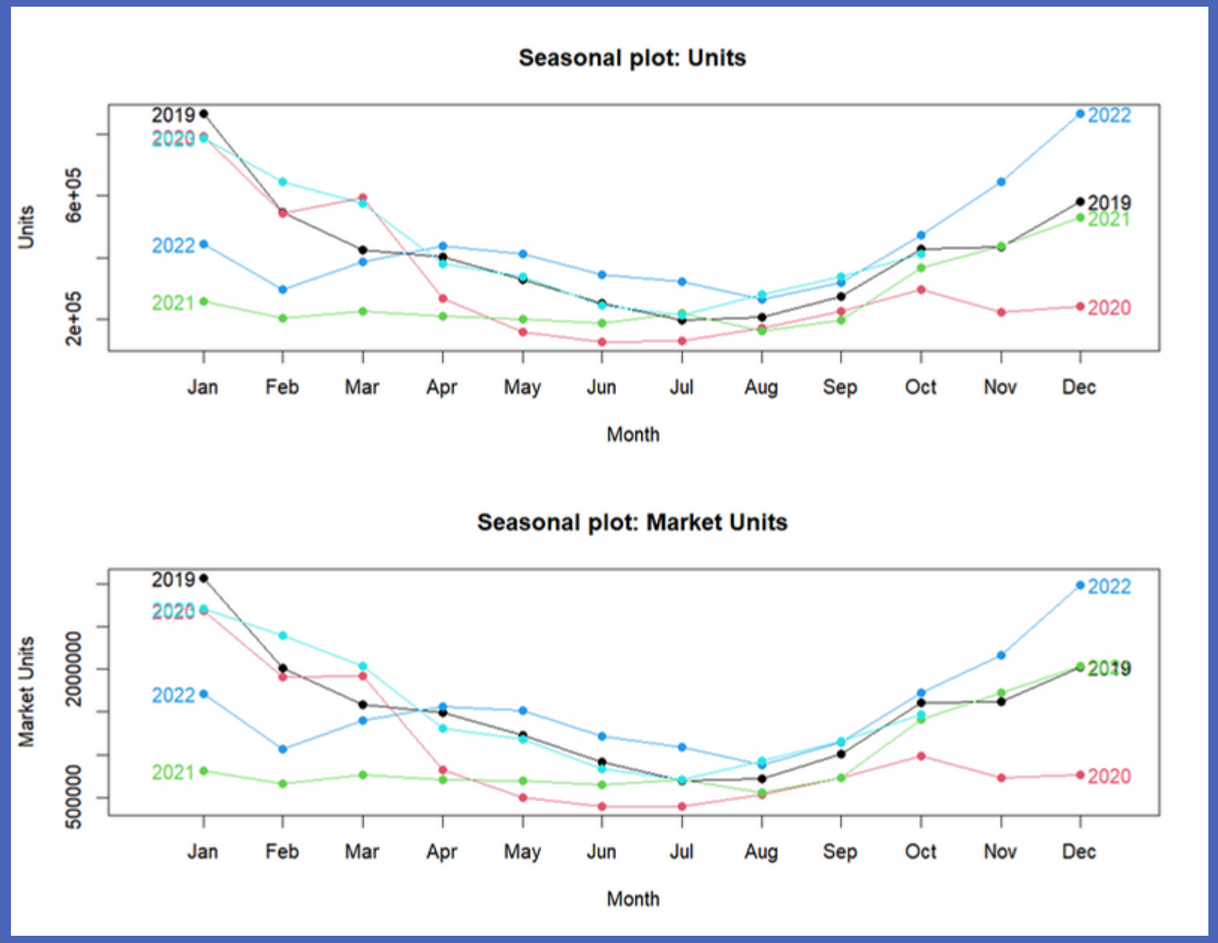
- Sales and units are highly correlated
- Multi-collinearity problem
- Since all the variable explain the phenomenon in the same way we choose to focus on Sales

3 Correlations

The trend of the company follows the one of the market

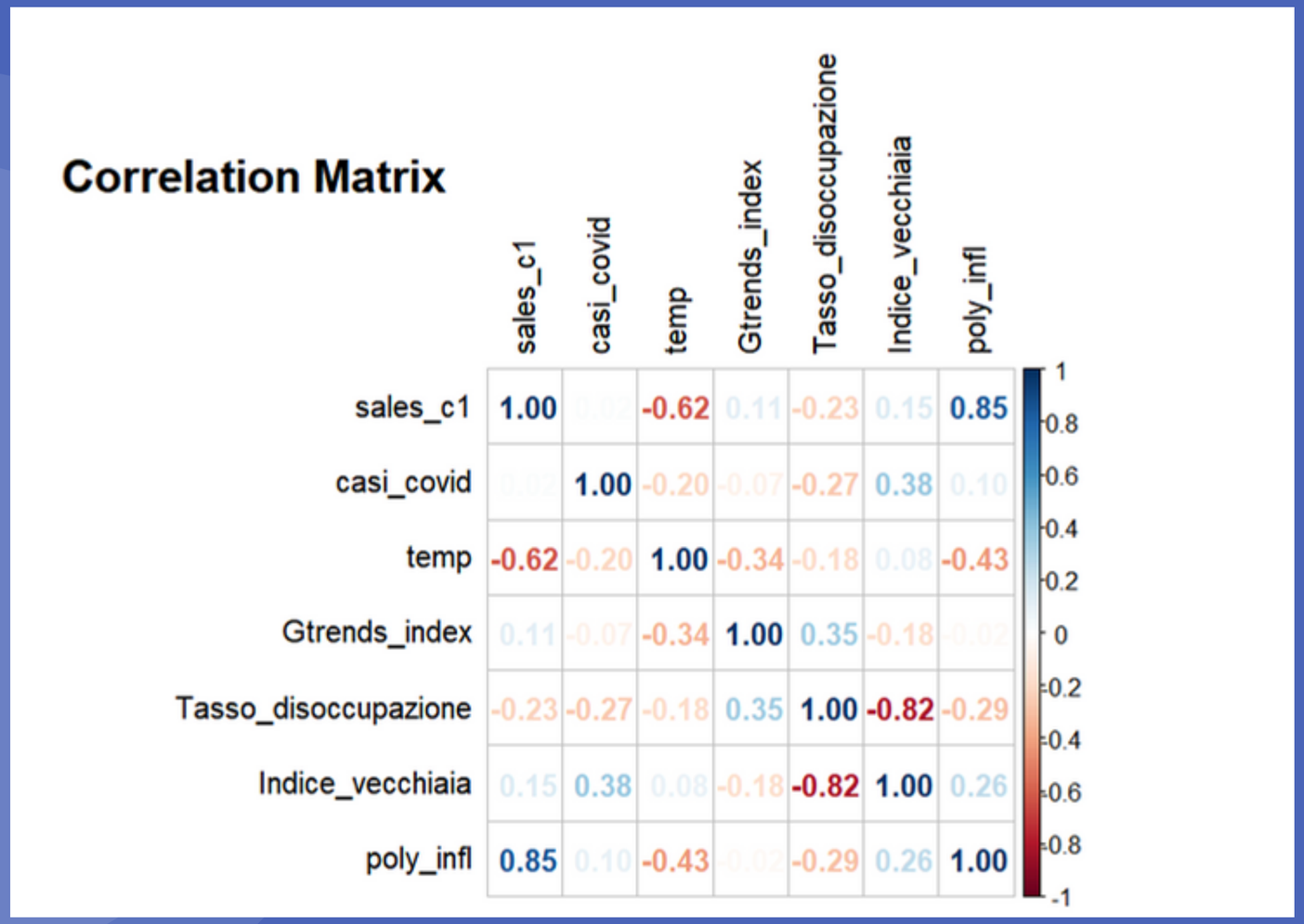
vs Compare:

1. Sales - Market Sales
2. Units - Market Units



Correlations

Between All variables



- High correlation between sales and flu -> Anti-Flu product
- Inverse Correlation between sales and Temp

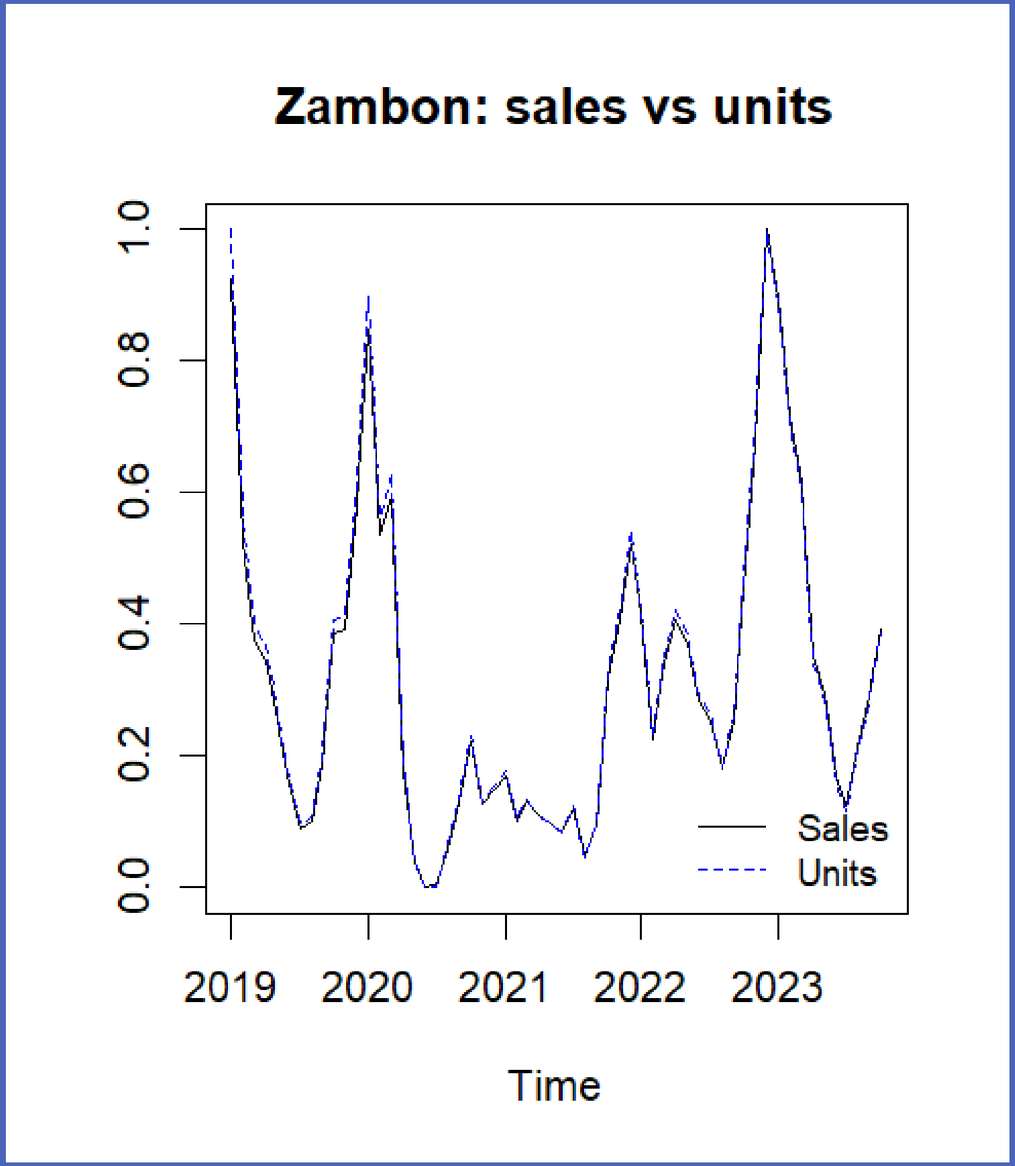
No relevant correlation between: old-age index and sales, Google Trend index and sales, COVID and sales.

DATA ANALYSIS

Comparative analysis: Product 1 sales vs their product 1 units.

Comparative analysis: Zambon's Performance vs Reference Market Trend

Comparative analysis: Product 1 vs Product 2



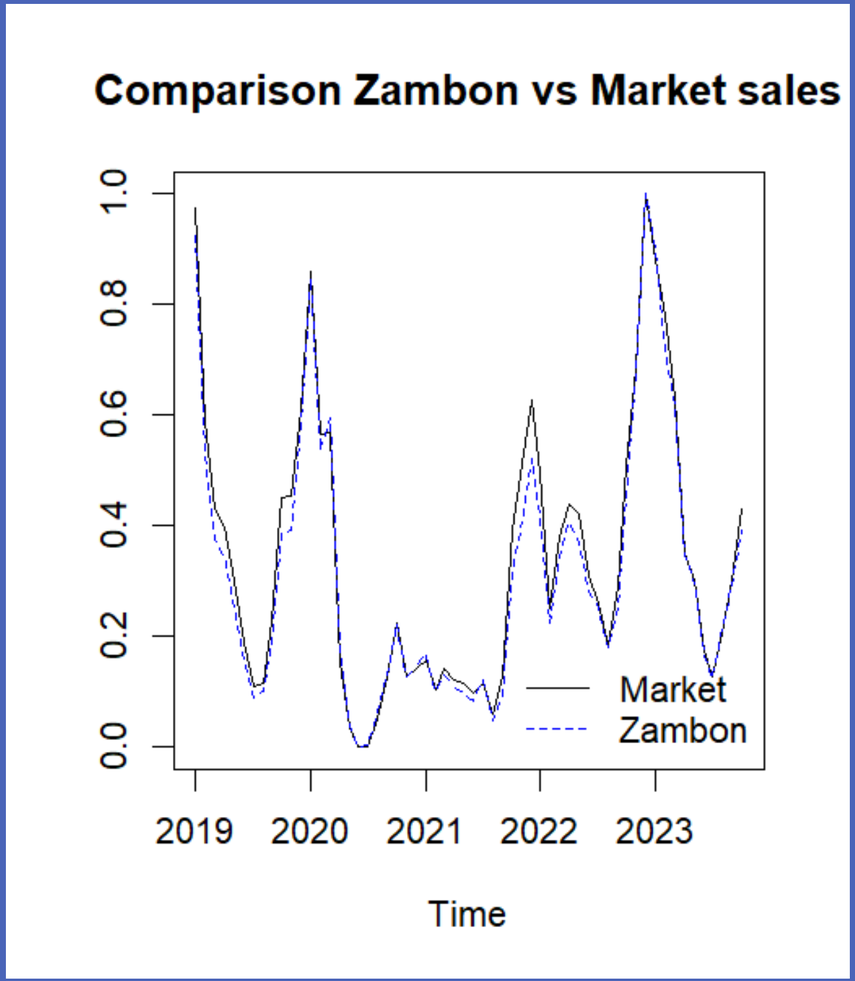
- The chart illustrates the comparison between time series of units produced by Zambon and the sales series

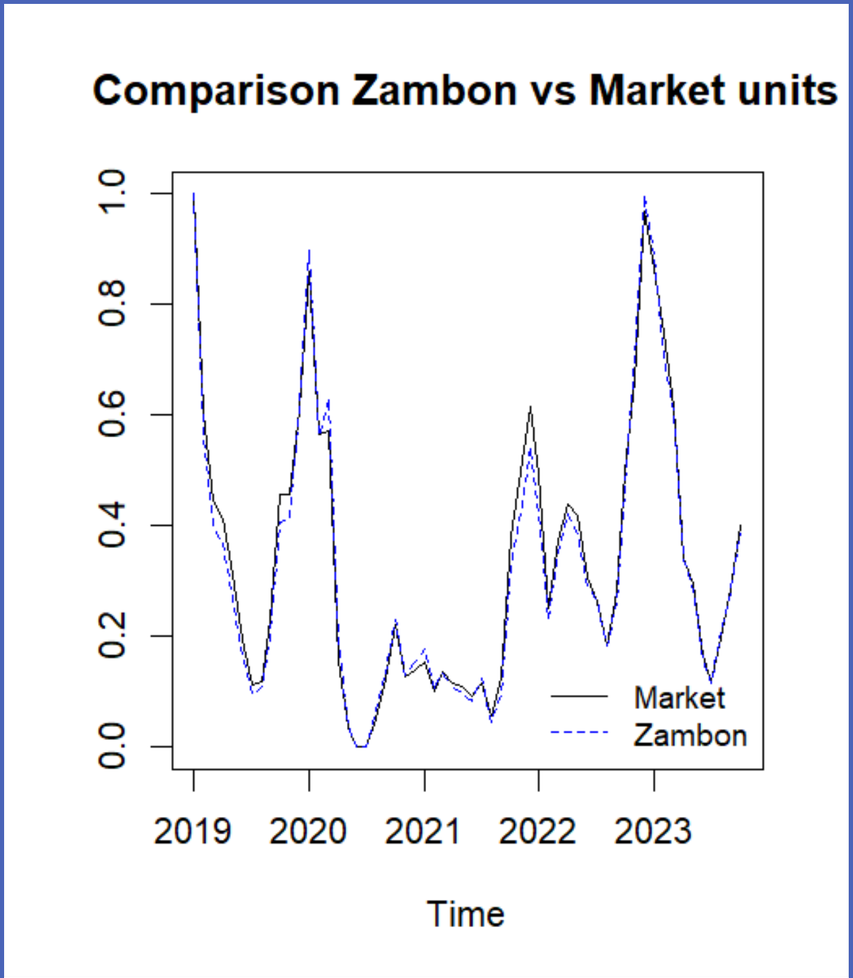
- They exhibit a similar temporal behaviour.

Correlation: 0.9984

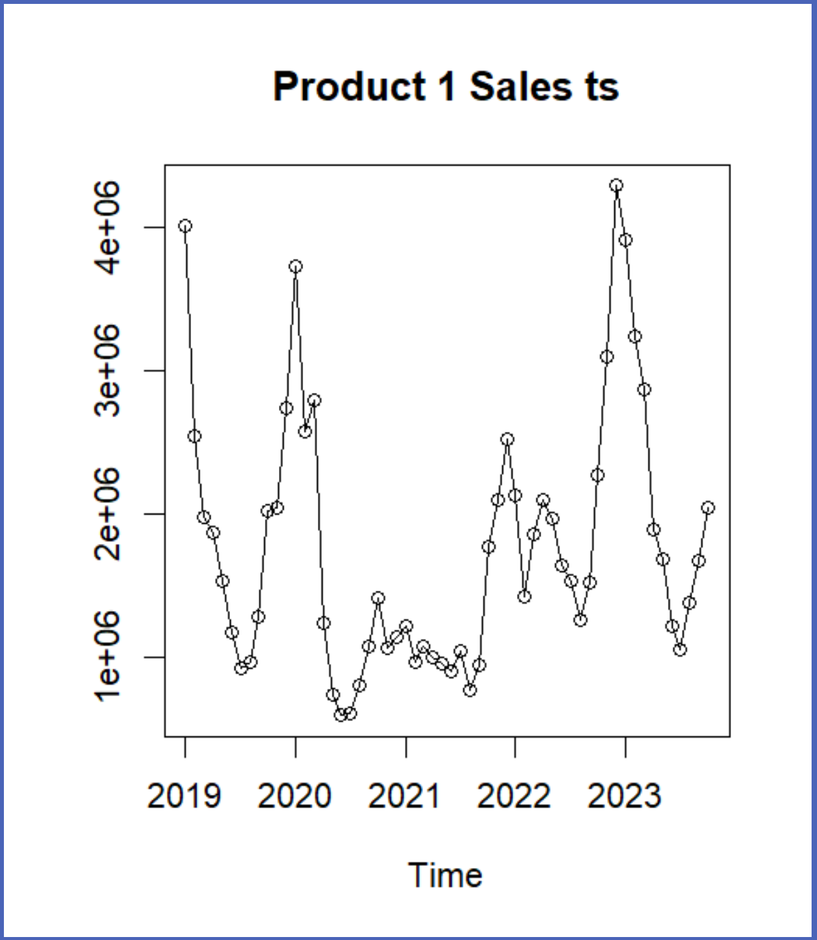
*Scaled with MinMaxScaler

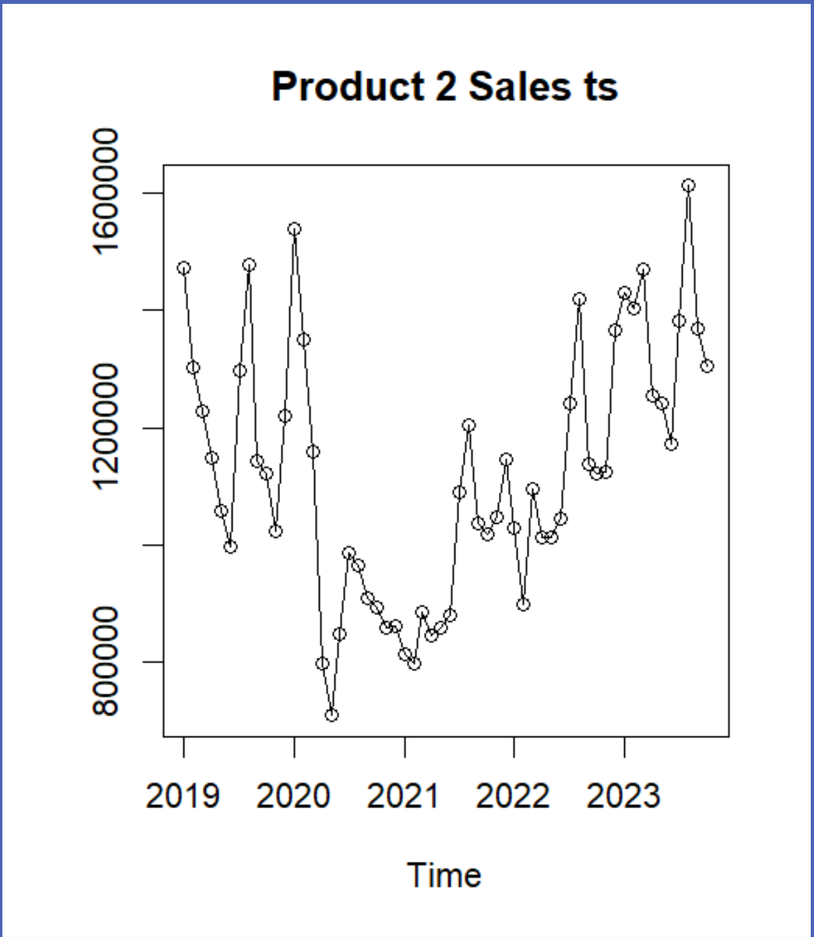
Comparing Zambon's Performance with the Reference Market Trend



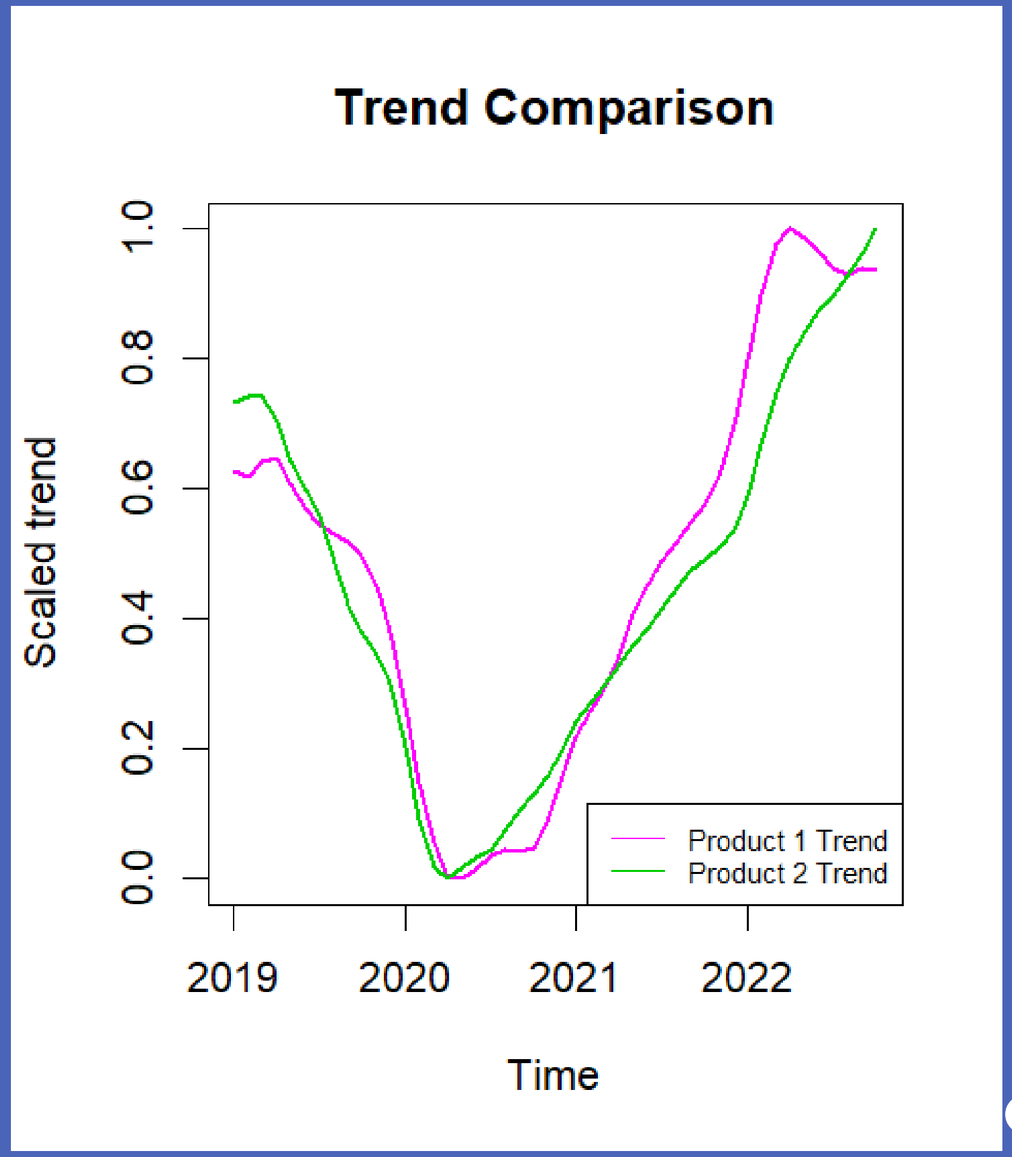


*Scaled with MinMaxScaler





- We compare the **first** and the **second** product
- We get some information about the habit of the products
- We try to understand something about Zambon in general.
- So, purifying from the effect of seasonality, we look some information in the **trend**.



*Scaled with MinMaxScaler

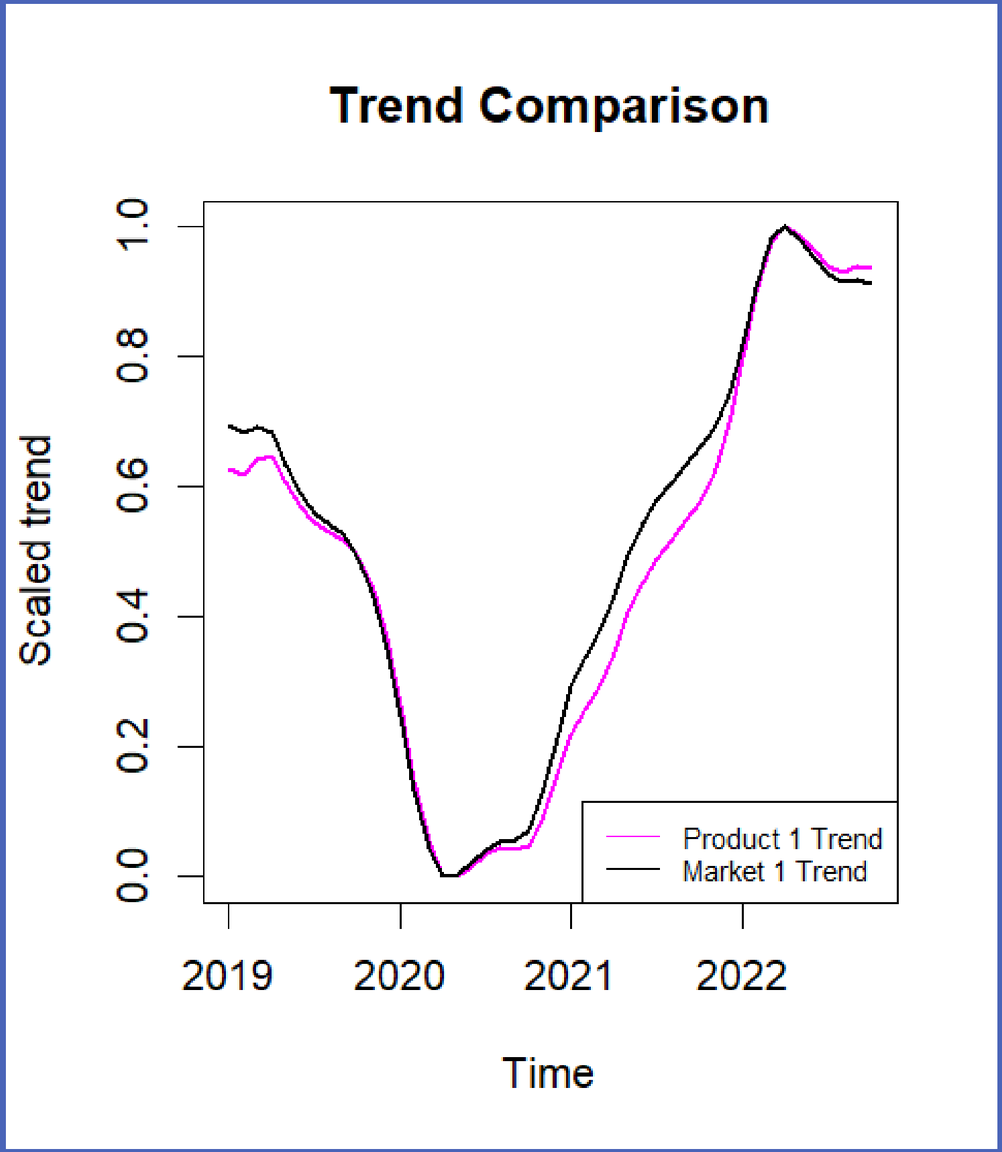
- We decompose the series in its components: Trend, Random and Seasonality, in order to remove the effect of the last one

- We analyze and compare trends: the highly similar trends, confirmed by the calculated correlation, lead us to conclude that both products exhibit a growing trend.

Could this be explained by a growing in popularity of Zambon?

We compare each product's trend with its respective reference market trend.

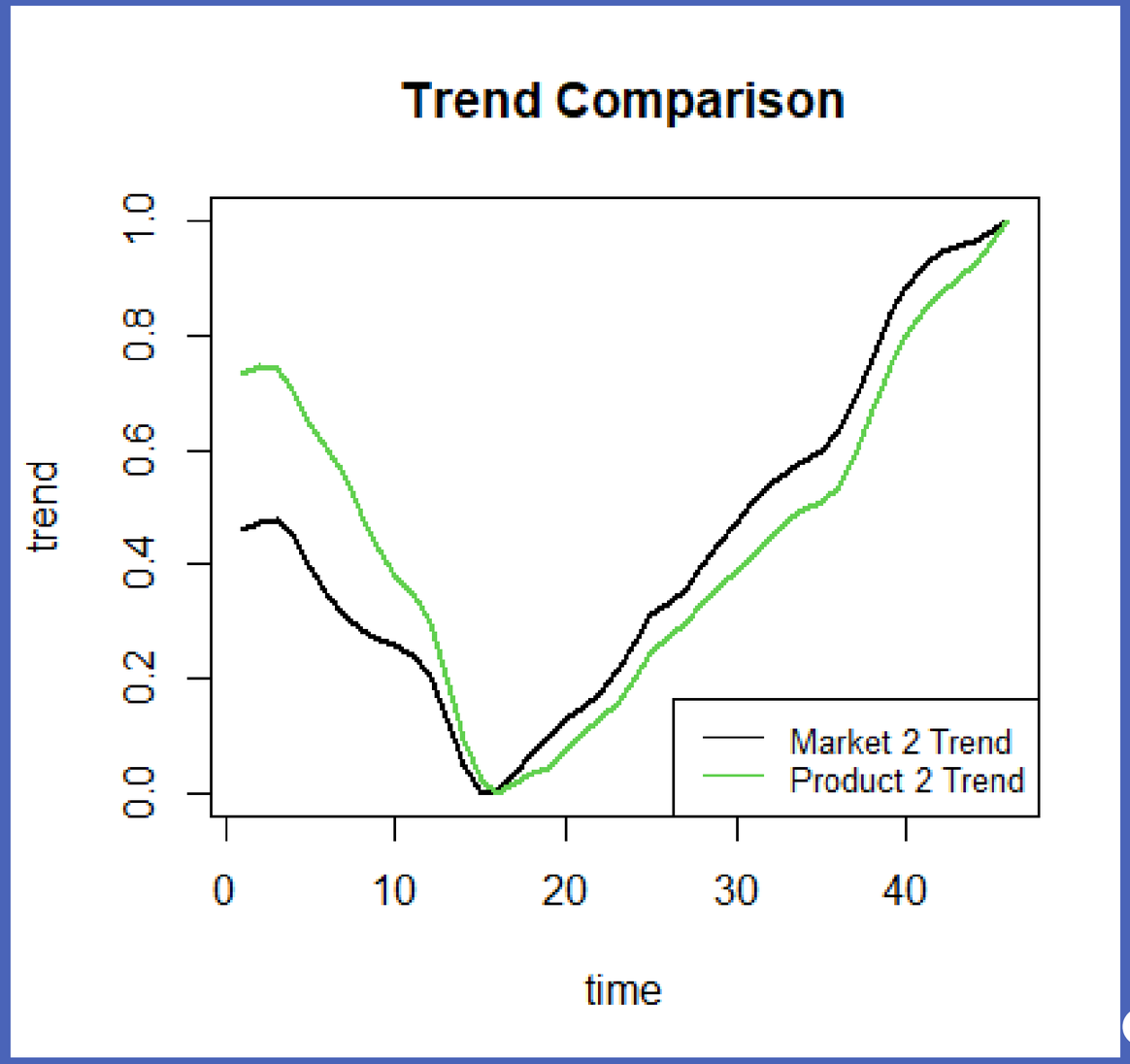
Correlation: 0.9615



*Scaled with MinMaxScaler

- The Reference Market's behavior appears to be followed by the Product 1 trend
- This indicates that the growing trend is not the result of Zambon's increasing popularity, but
- Just a confirmation of the fact that the product is in line with the Reference Market

Correlation:0.9926



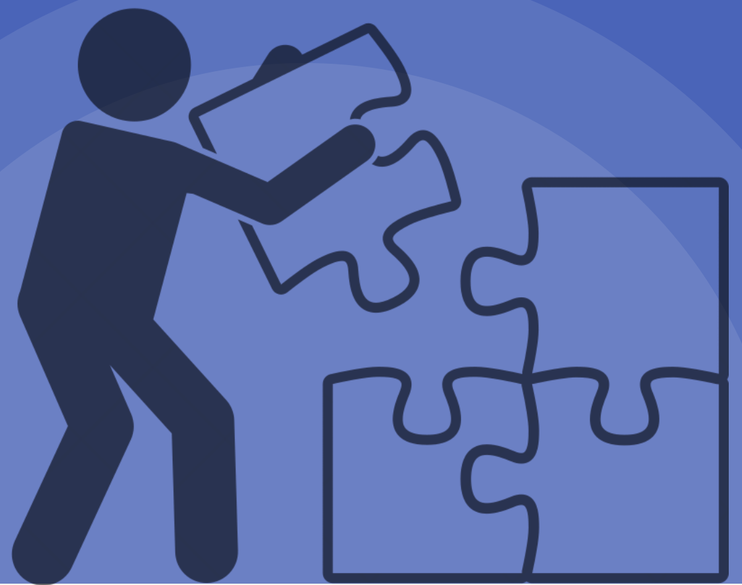
*Scaled with MinMaxScaler

- After the first Covid pandemic, Product Trend 2 reverses in relation to the Reference Market trend.

- Numerous factors (laws, the introduction of new products onto the market, etc.) could explain this, but since we are unaware of the kind of product, we are unable to continue our investigation.

Correlation: 0.9072

DATA MODELING



- LINEAR REGRESSION
- ARMAX
- ARIMA MODELS
- GENERALIZED ADDITIVE MODEL
- TIME SERIES DECOMPOSITION ANALYSIS
- EXPONENTIAL SMOOTHING

LINEAR REGRESSION

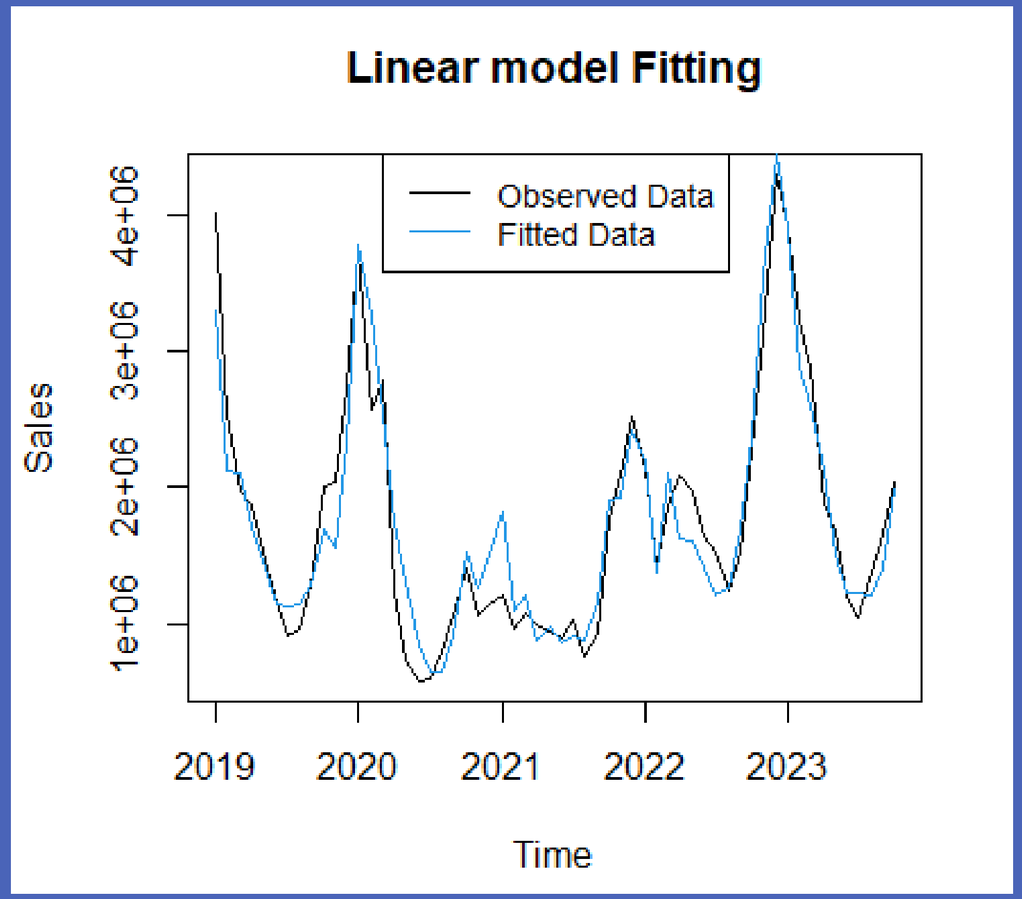
*We get to this results dropping the variables by hand
*Kept variables are significant at 10%

tslm(sales_ts ~ trend + season + covid_ts + disoc_ts + influenza_ts)

Coefficients	Estimate	Pr(>|t|)	Coefficients	Estimate	Pr(>|t|)
Intercept	4.449e+06	2.16e-05	Season 8 (Aug)	-1.374e+06	1.07e-06
Trend	-9.503e+03	0.054315	Season 9 (Sep)	-1.092e+06	5.85e-05
Season 2 (Feb)	-7.982e+05	0.000484	Season 10 (Oct)	-5.329e+05	0.032976
Season 3 (Mar)	-5.983e+05	0.011127	Season 11 (Nov)	-8.152e+05	0.000934
Season 4 (Apr)	-8.093e+05	0.001339	Season 12 (Dic)	-3.900e+05	0.090198
Season 5 (May)	-1.039e+06	7.42e-05	covid_ts	-9.363e-03	0.016259
Season 6 (June)	-1.316e+06	2.68e-06	disoc_ts	-2.482e+05	0.004420
Season 7 (July)	-1.348e+06	1.08e-06	influenza_ts	2.144e+05	4.54e-12

LINEAR REGRESSION

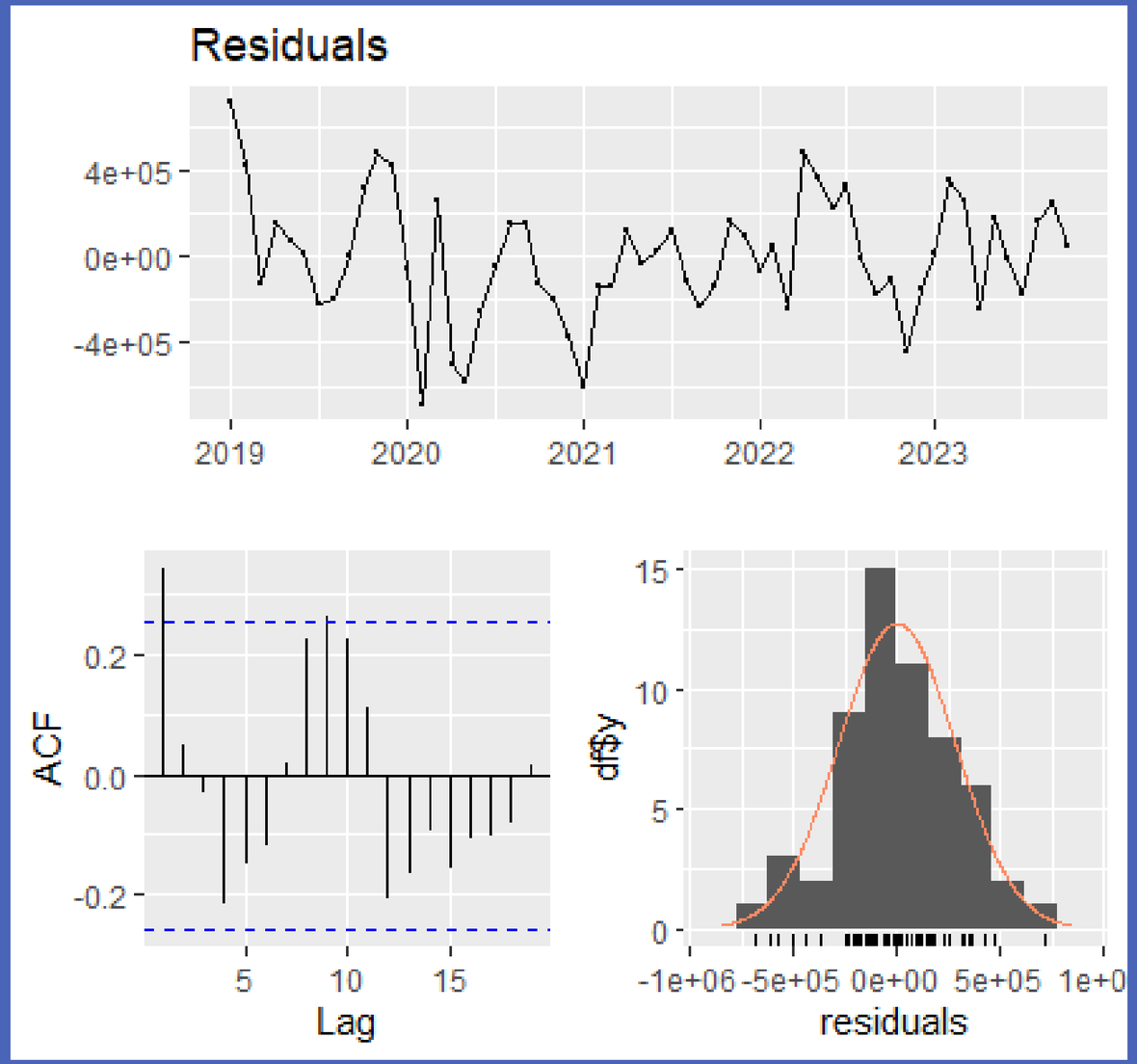
tslm(sales_ts ~ trend + season + covid_ts + disoc_ts + influenza_ts)



R-Squared	AIC	RMSE
0.8993	1653.742	280507

*We get to this results dropping the variables by hand
*Kept variables are significant at 10%

FROM LINEAR REGRESSION TO ARMAX

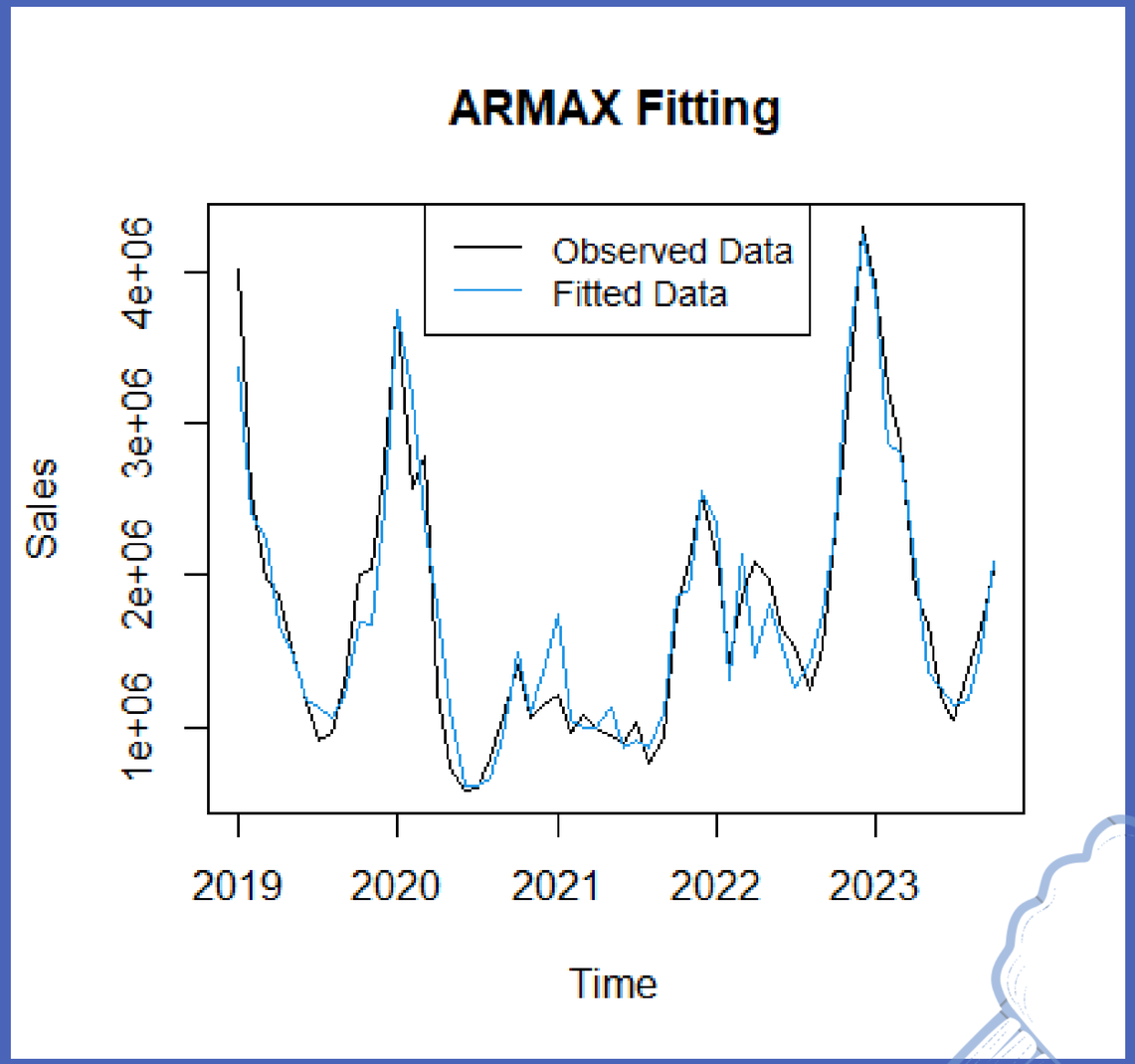


Durbin Watson Test
1.1997

The **graphical representation** and the **Durbin Watson Test** suggest that there's still **autocorrelation** in the residuals

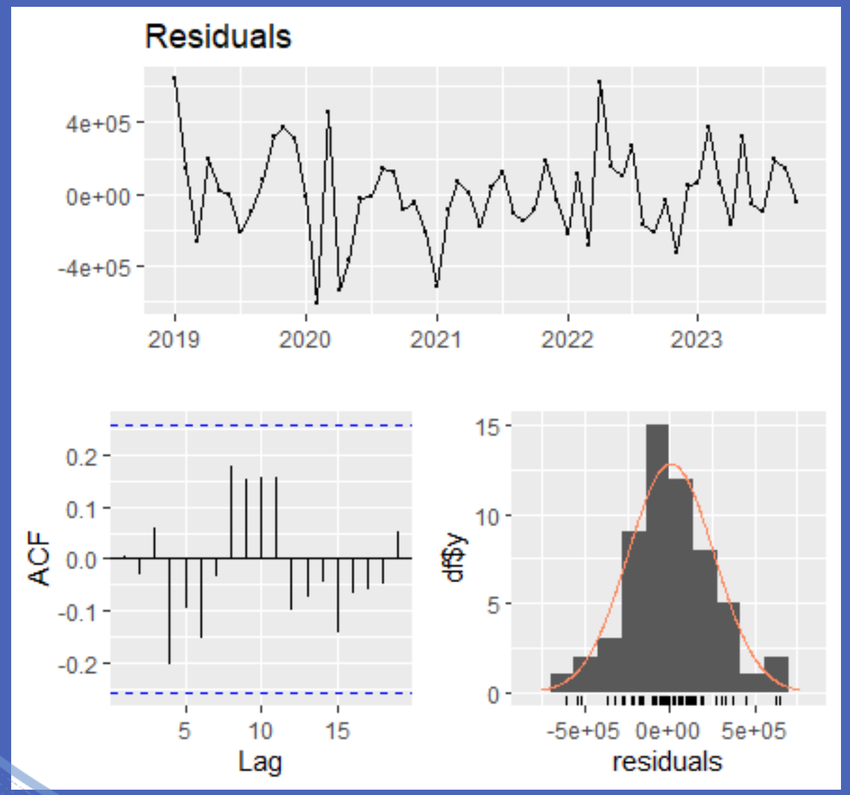
-> we decide to model an ARMAX

ARMAX with auto.arima

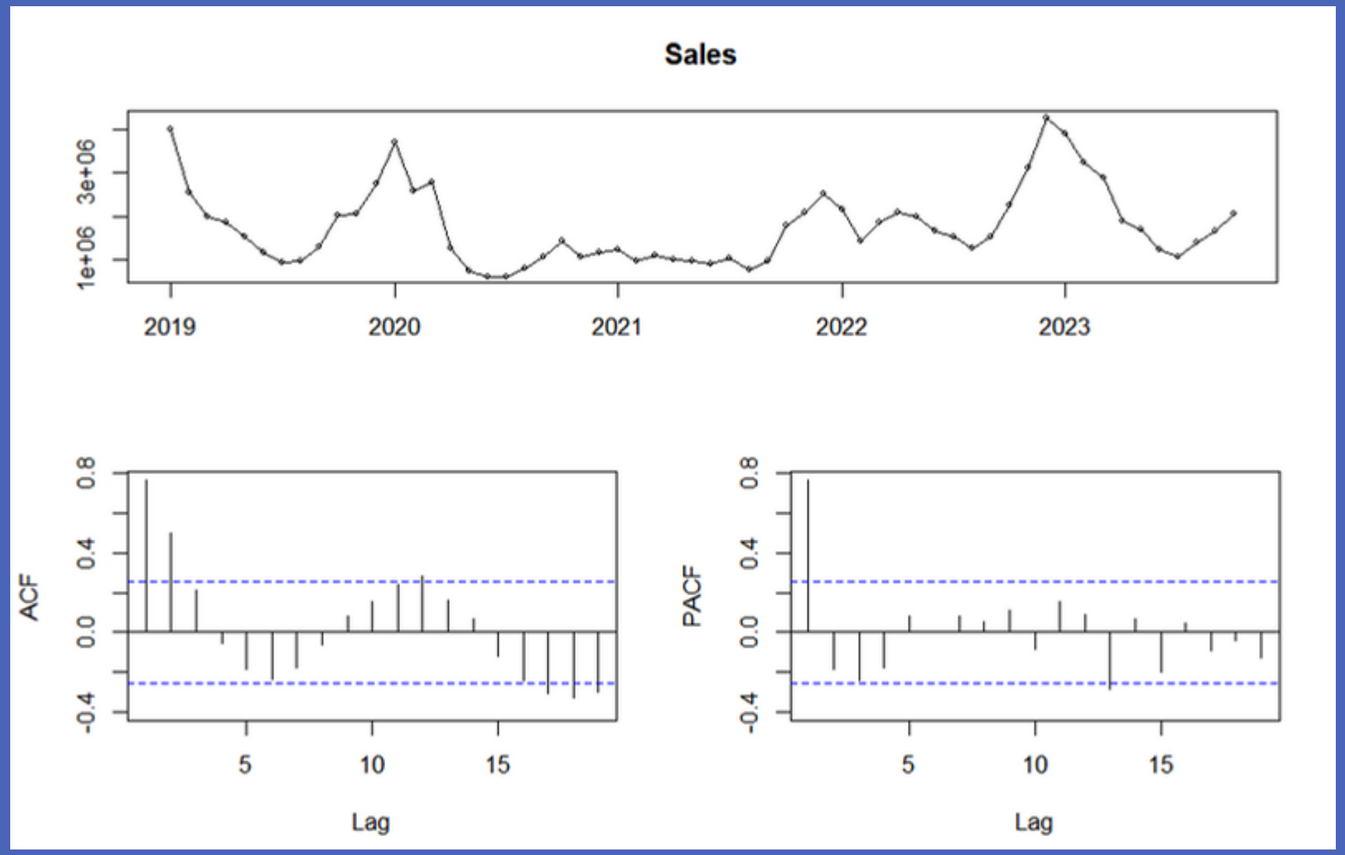


Regression
with ARIMA(1,0,0)(1,0,0)[12] errors

R-Squared	AIC	RMSE
-	1615.81	250580.7



ARIMA



- Sinusoidal oscillation of ACF

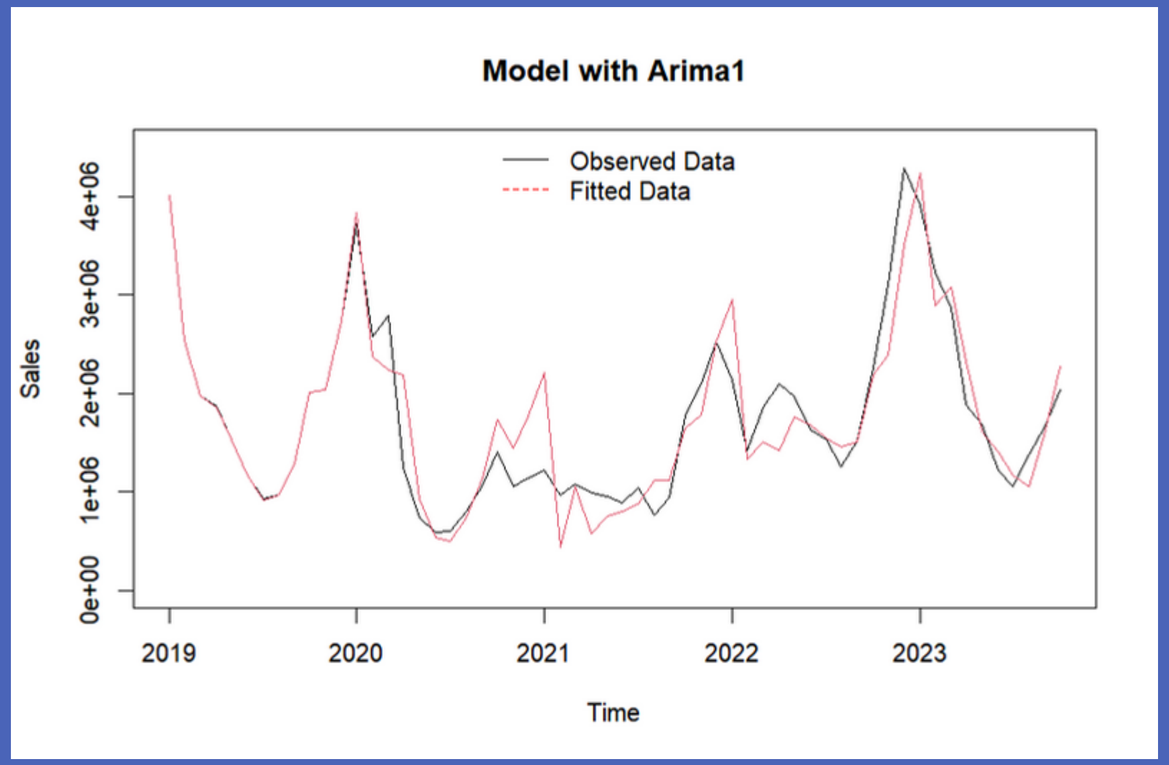
- By looking at PACF fix parameter **p equal to 1** based on the spike at lag 1

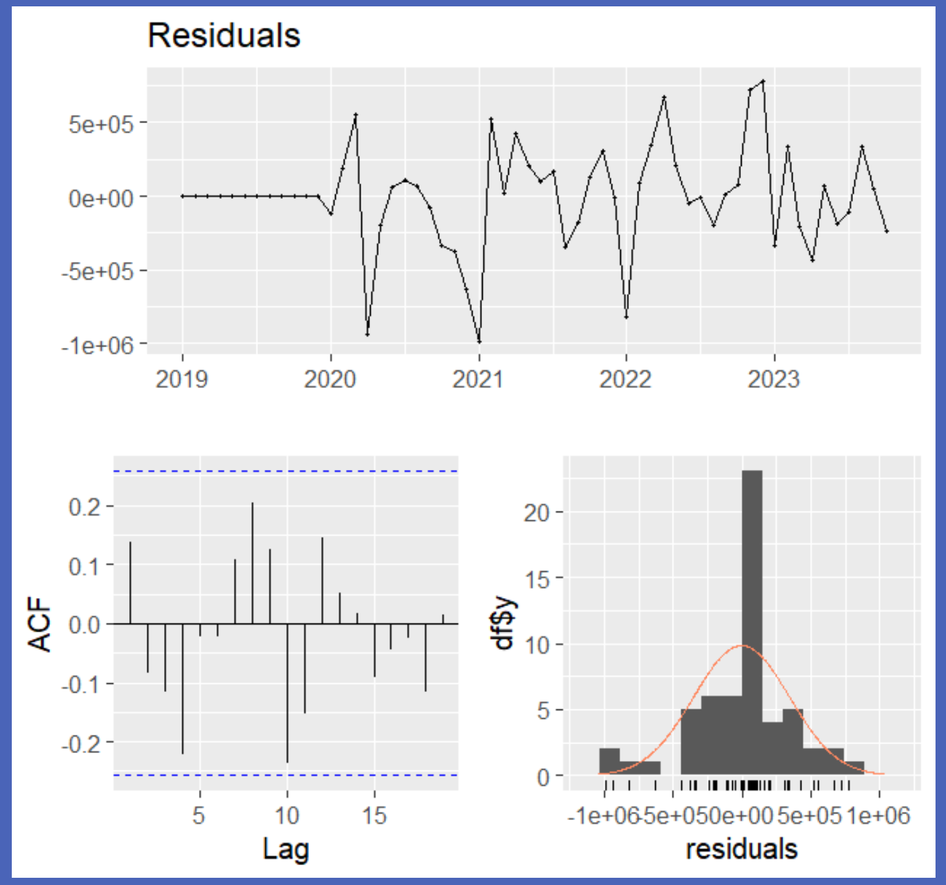
ARIMA

1

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

AIC	RMSE
1339.97	344510



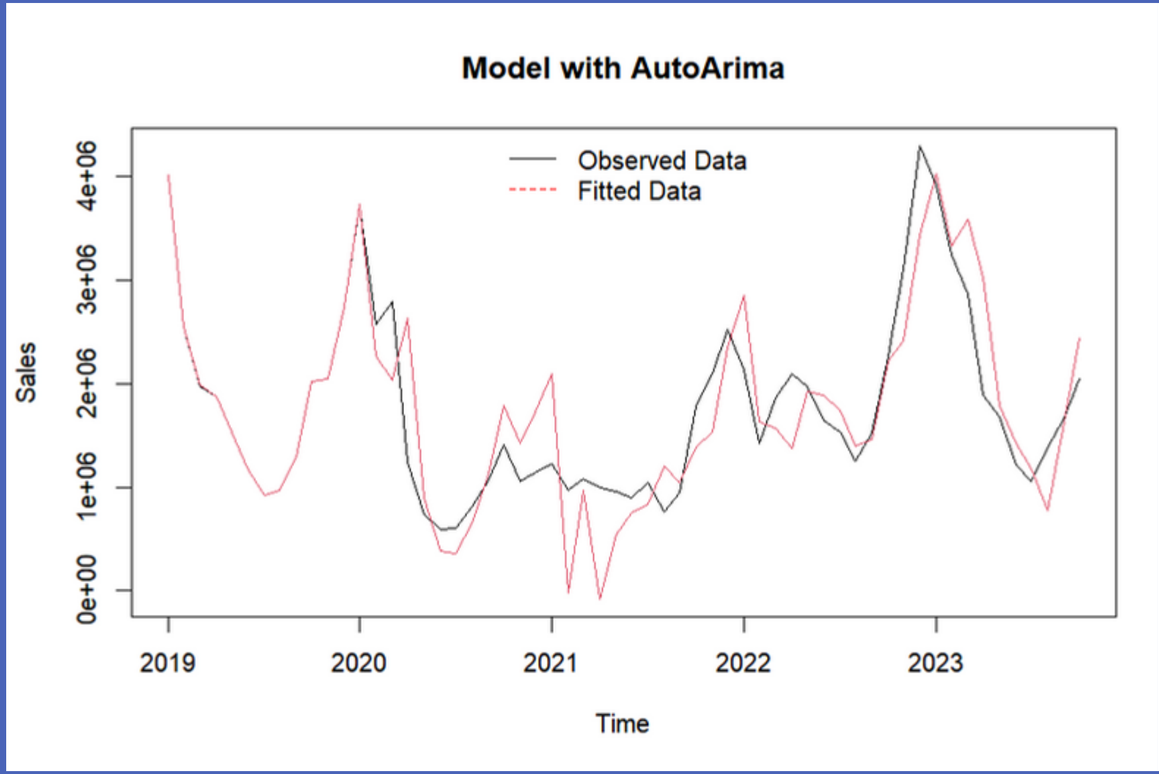


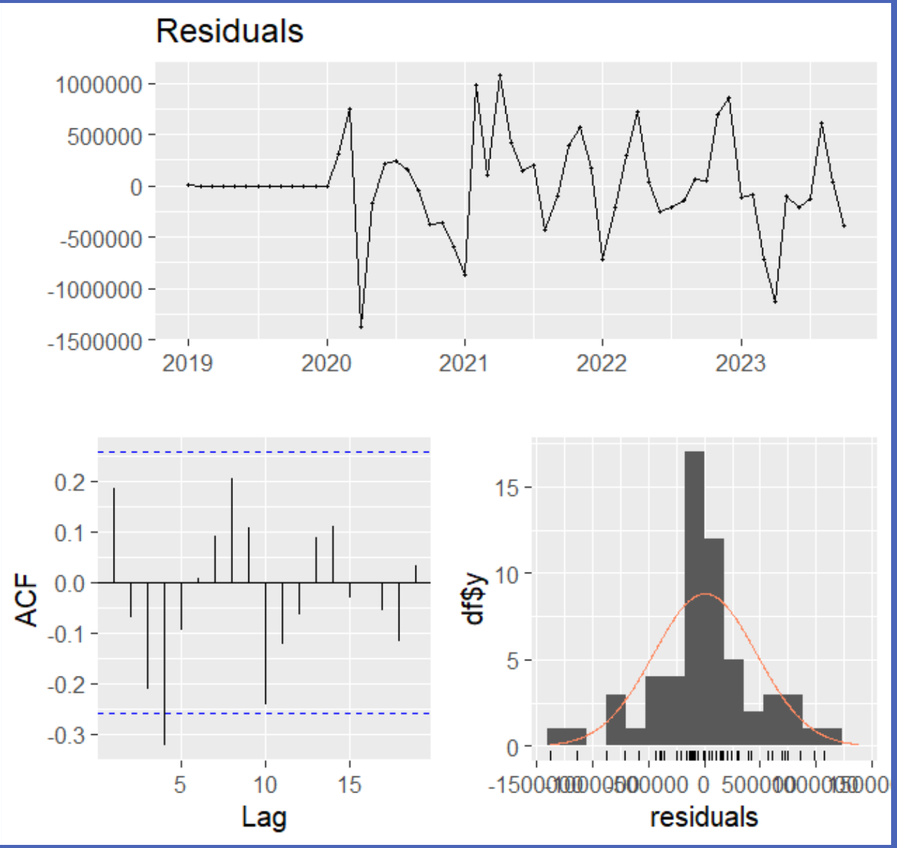
2

AUTO ARIMA

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

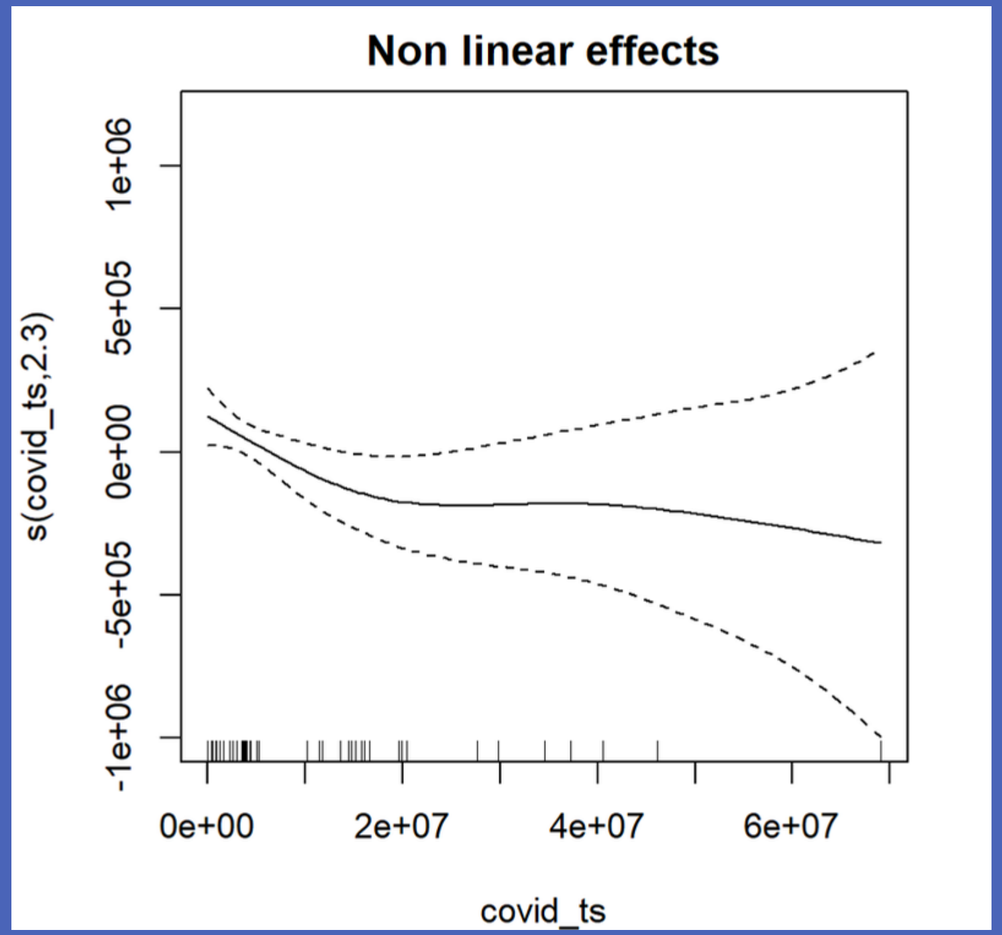
AIC	RMSE
1316.84	456323.7

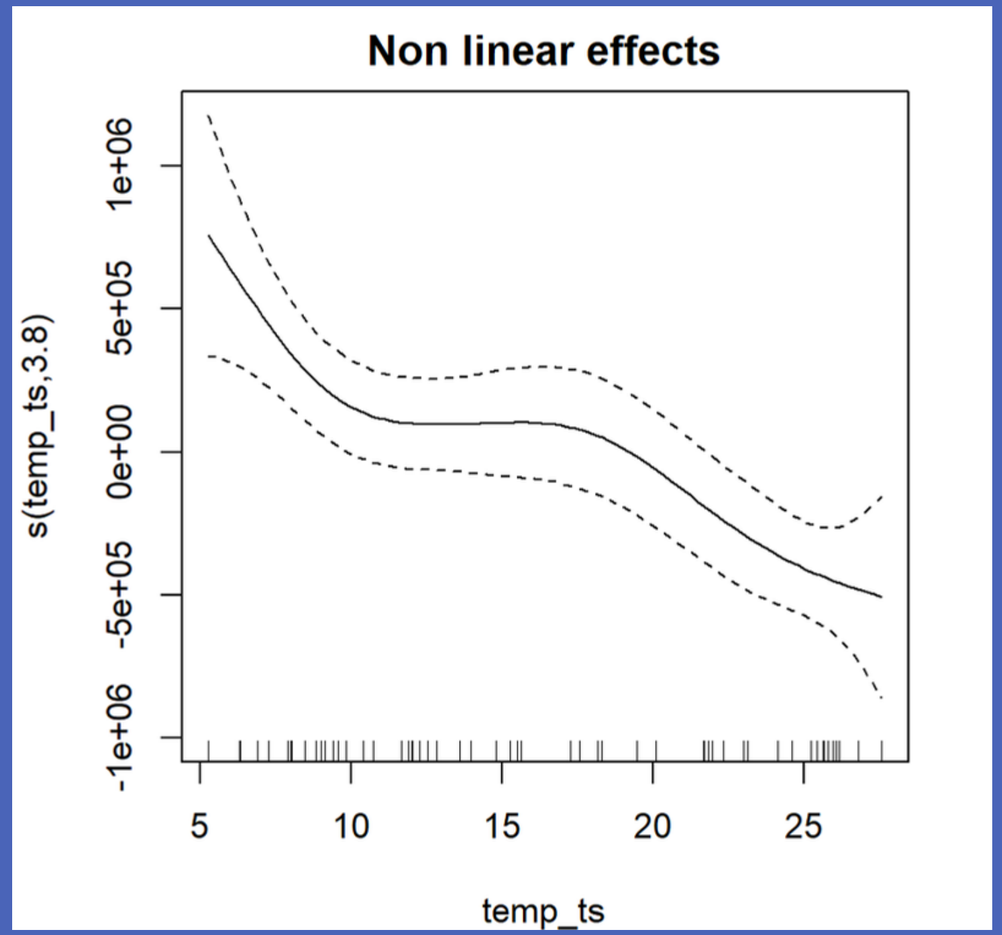




GENERALIZED ADDITIVE MODEL

We notice a non linear effect of the variables covid_ts and temp_ts wtr sales_ts

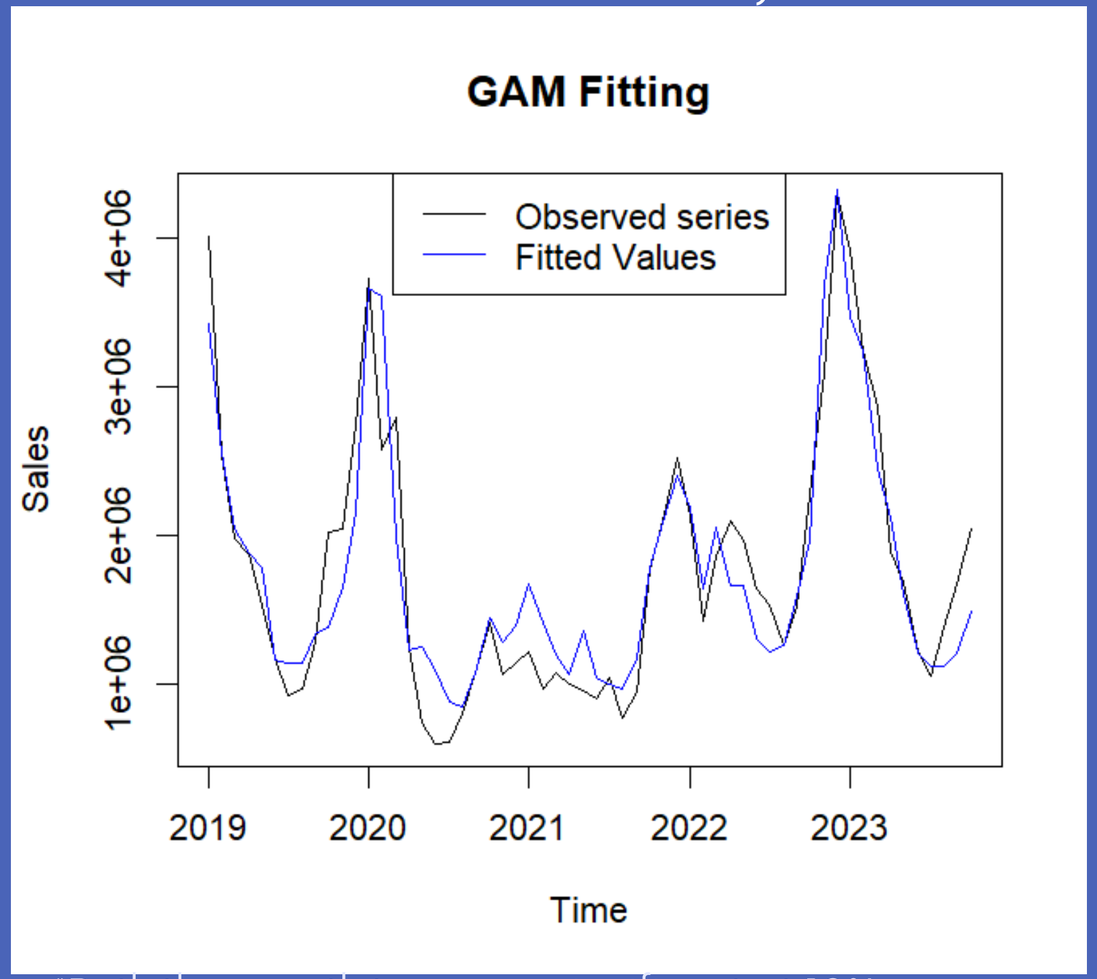




In particular in the first part of the series there is a clear non linear effect of covid on sales.
Spline explains both effects way better.

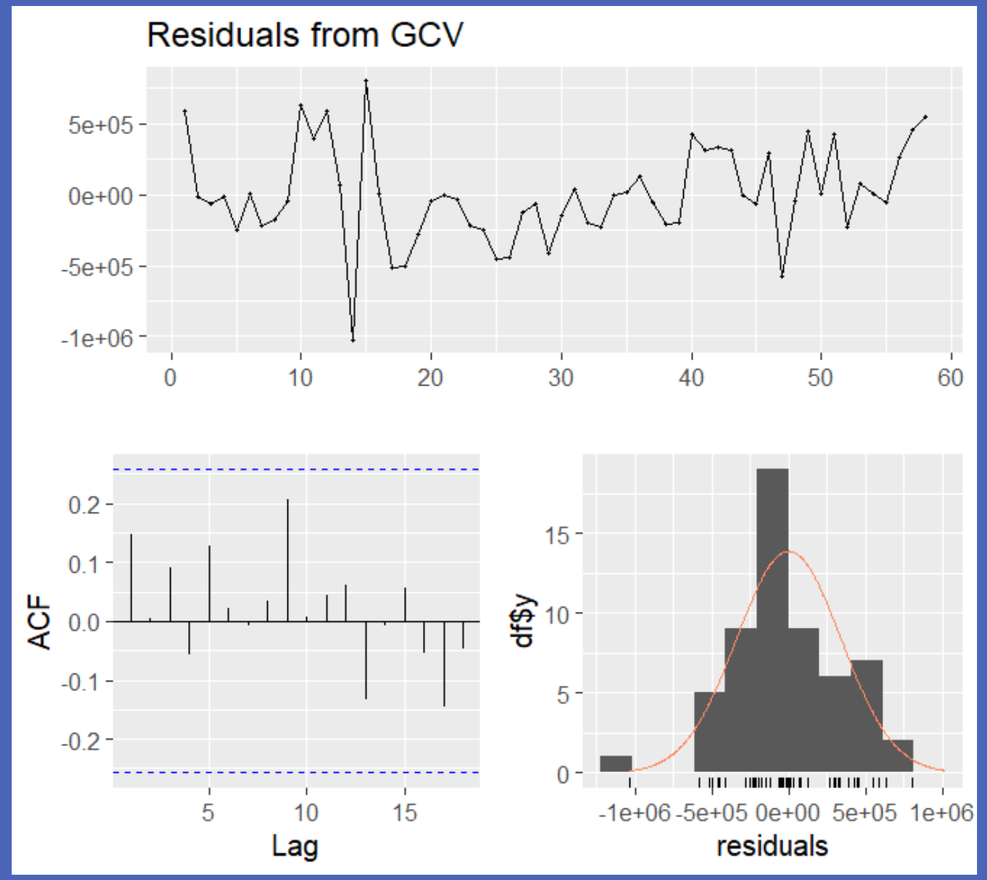
GENERALIZED ADDITIVE MODEL

gam(sales_ts ~ s(covid_ts) + s(temp_ts) + influenza_ts)



*Both the smooth terms are significant at 10%
*Flu linear effect is highly significant

R-Squared	AIC	RMSE
0.834	1659.217	336993.1

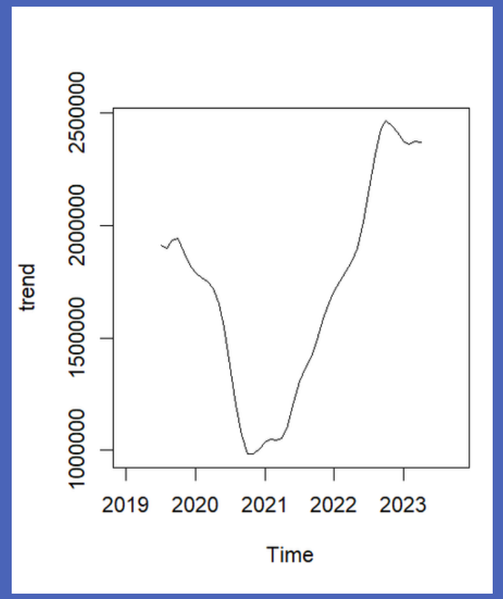


TIME SERIES DECOMPOSITION ANALYSIS

TREND

ARMAX-GAM

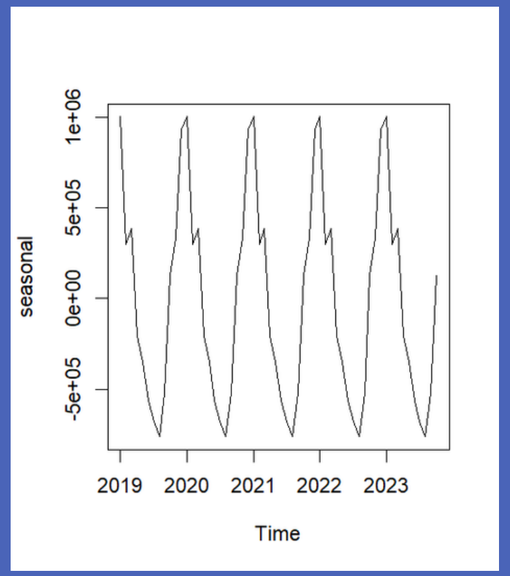
GAM: gam(trend~s(time))
with residuals modelled with an ARIMA(2,0,0)

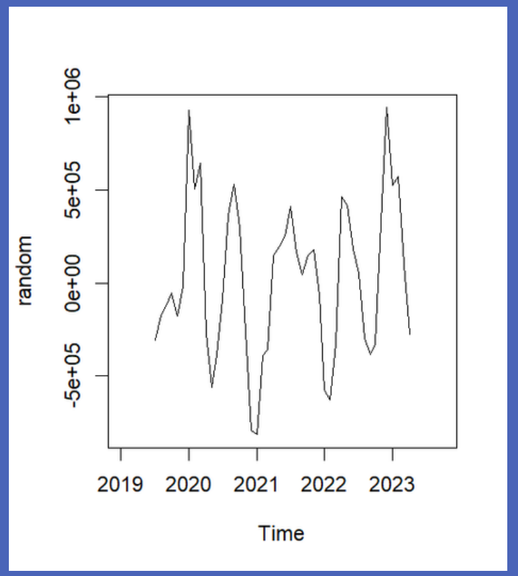


SEASONAL + RANDOM

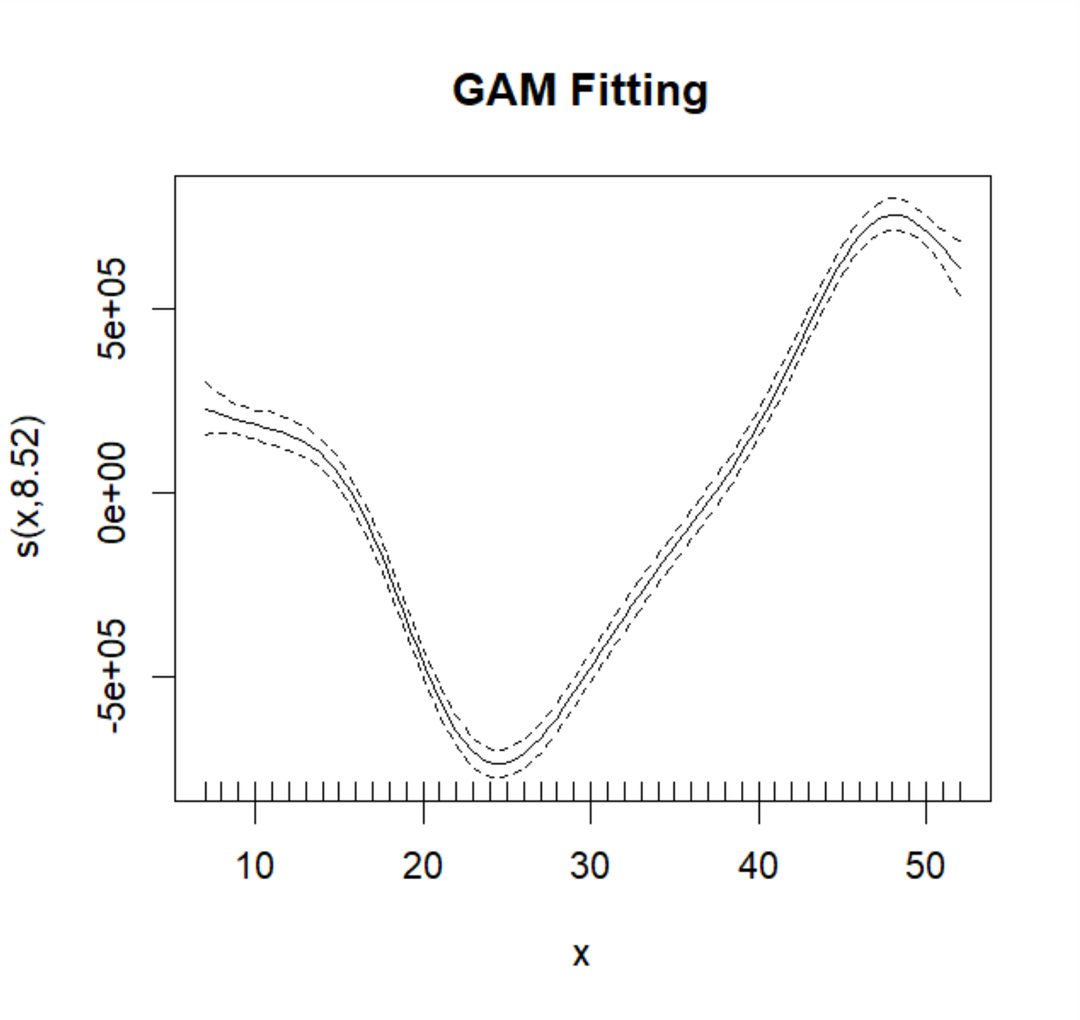
LINEAR MODEL

tslm(seasonal+random ~ covid_ts + temp_ts + influenza_ts)





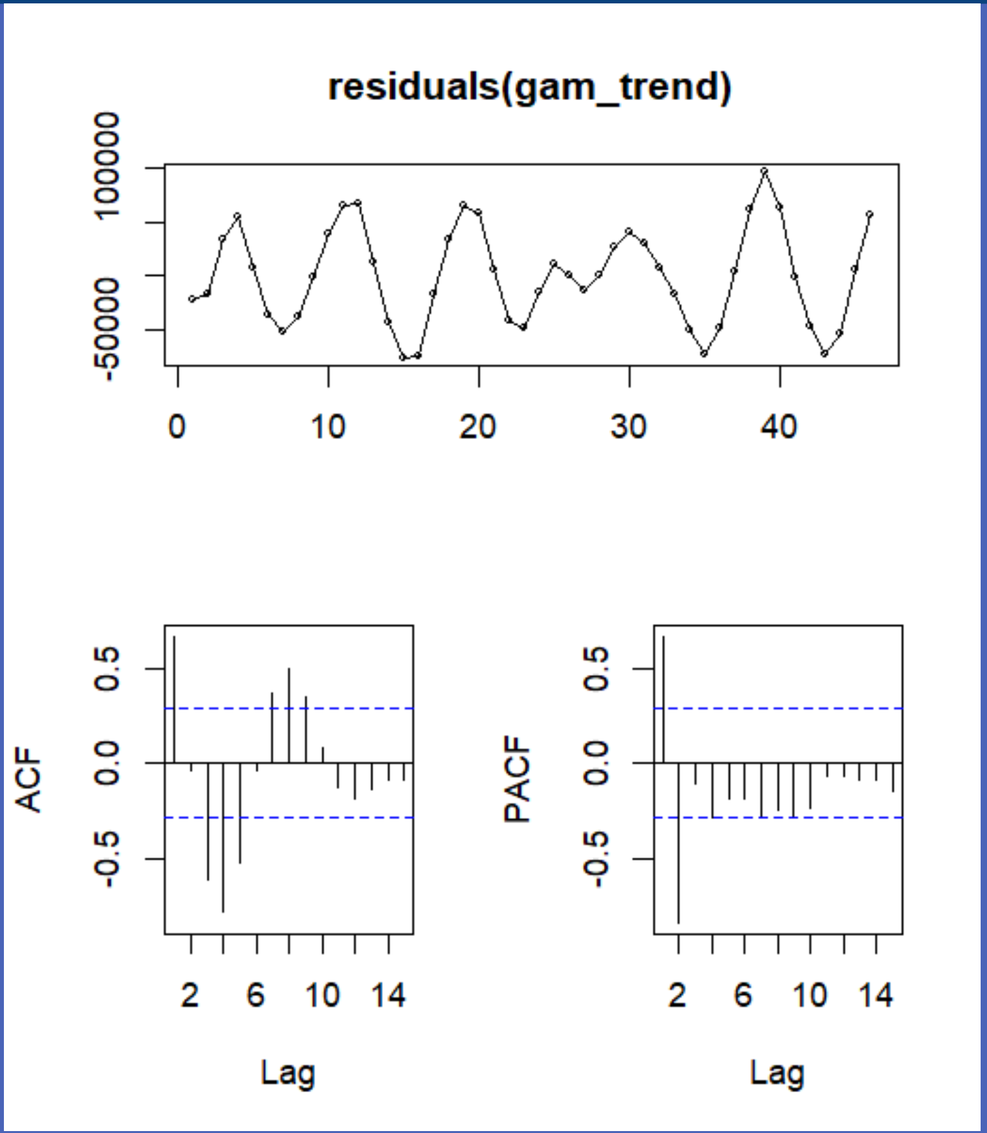
TIME SERIES DECOMPOSITION ANALYSIS - TREND



GAM: gam(trend~s(time))

R-Squared	AIC	RMSE
0.989	1135.62	390712.5

FROM GAM REGRESSION TO ARMAX

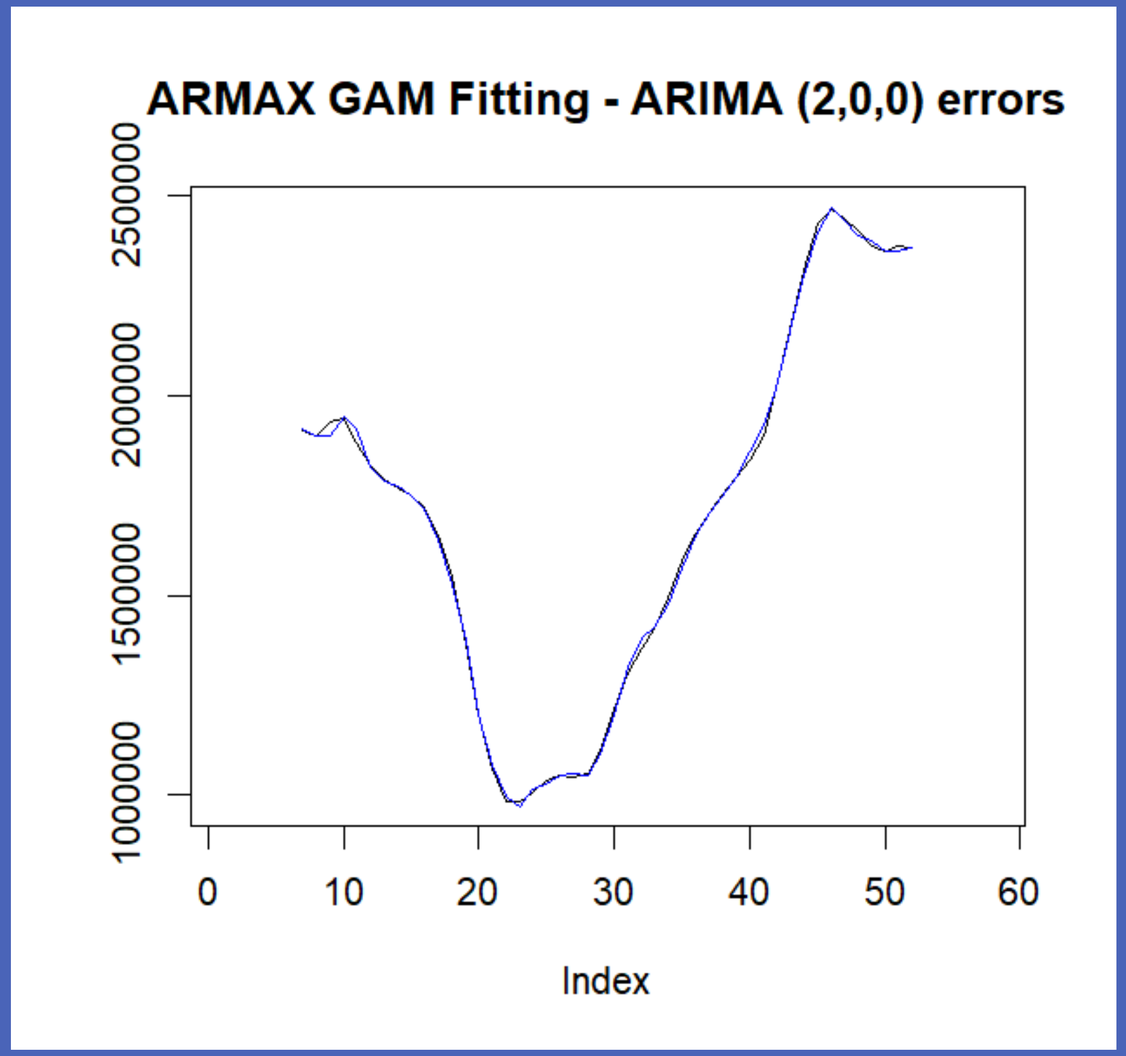


Durbin Watson Test
0.031954

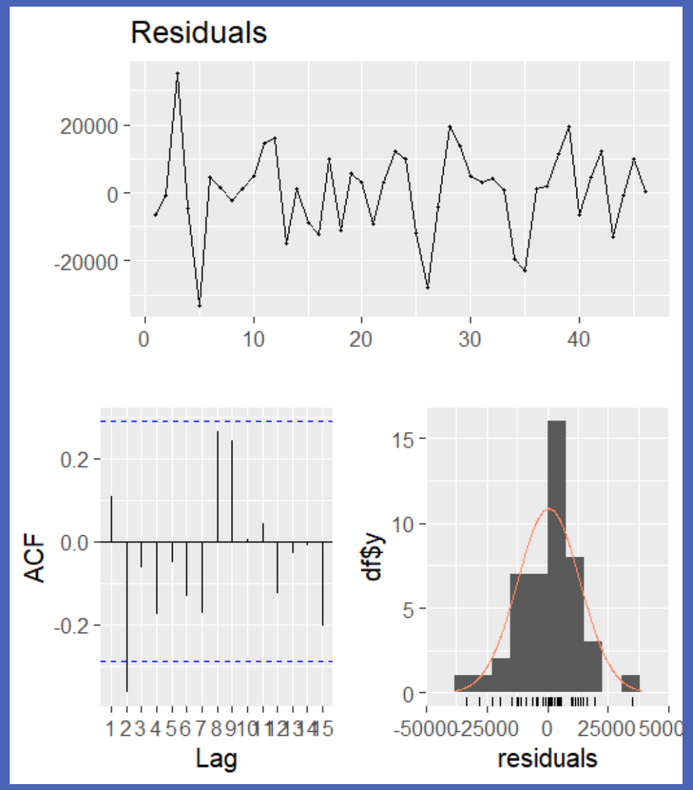
The **graphical representation** and the **Durbin Watson Test** suggest that there's still **autocorrelation** in the residuals

-> we decide to model an ARMAX

TIME SERIES DECOMPOSITION ANALYSIS - TREND

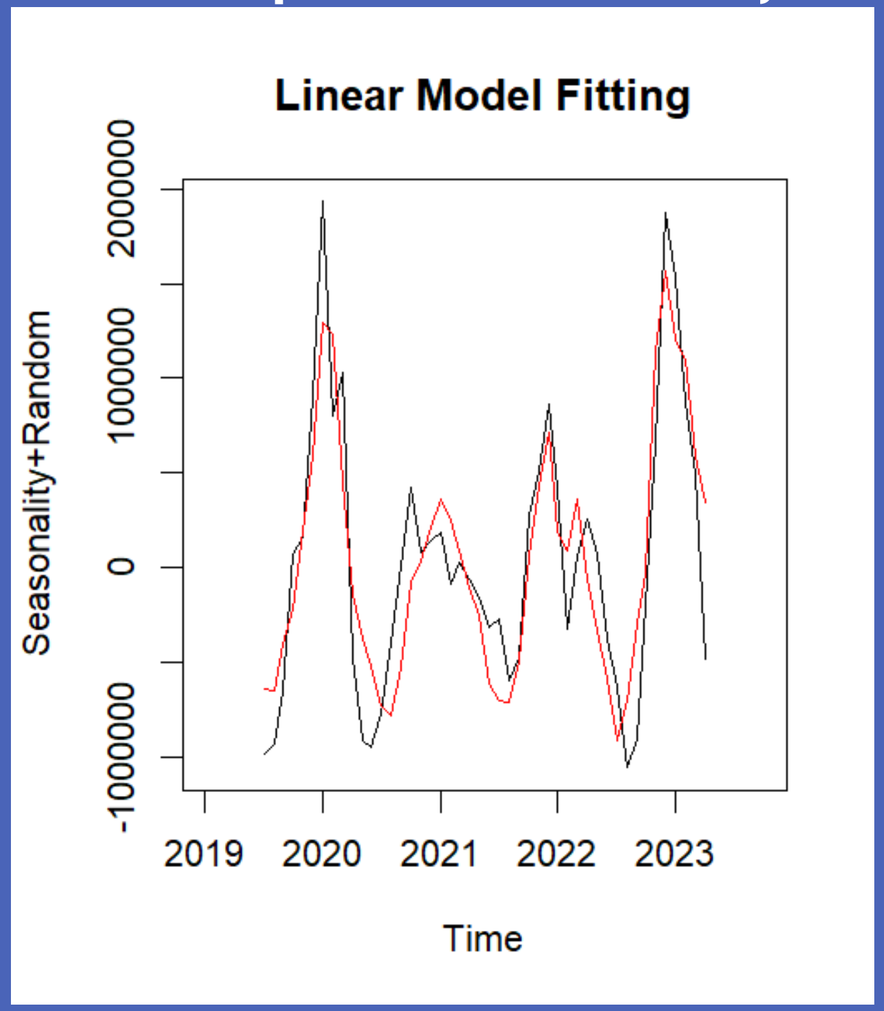


R-Squared	AIC	RMSE
0.989	1010.29	12727.44

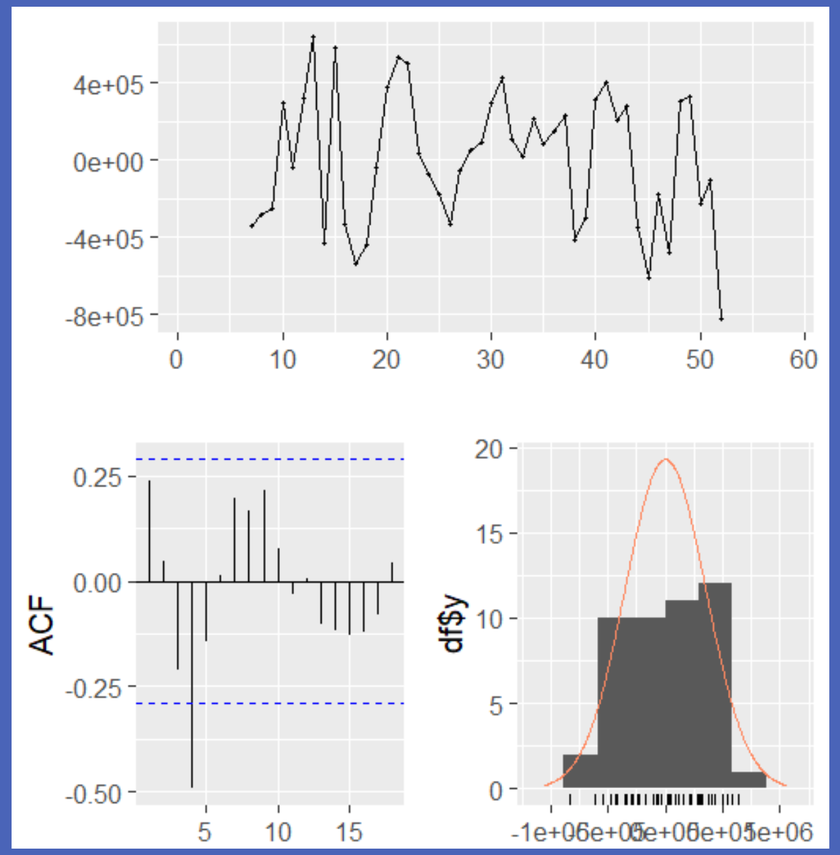


TIME SERIES DECOMPOSITION ANALYSIS - SEASONAL+RANDOM

tslm(seasonal_p+random_p ~ covid_ts + temp_ts + influenza_ts)

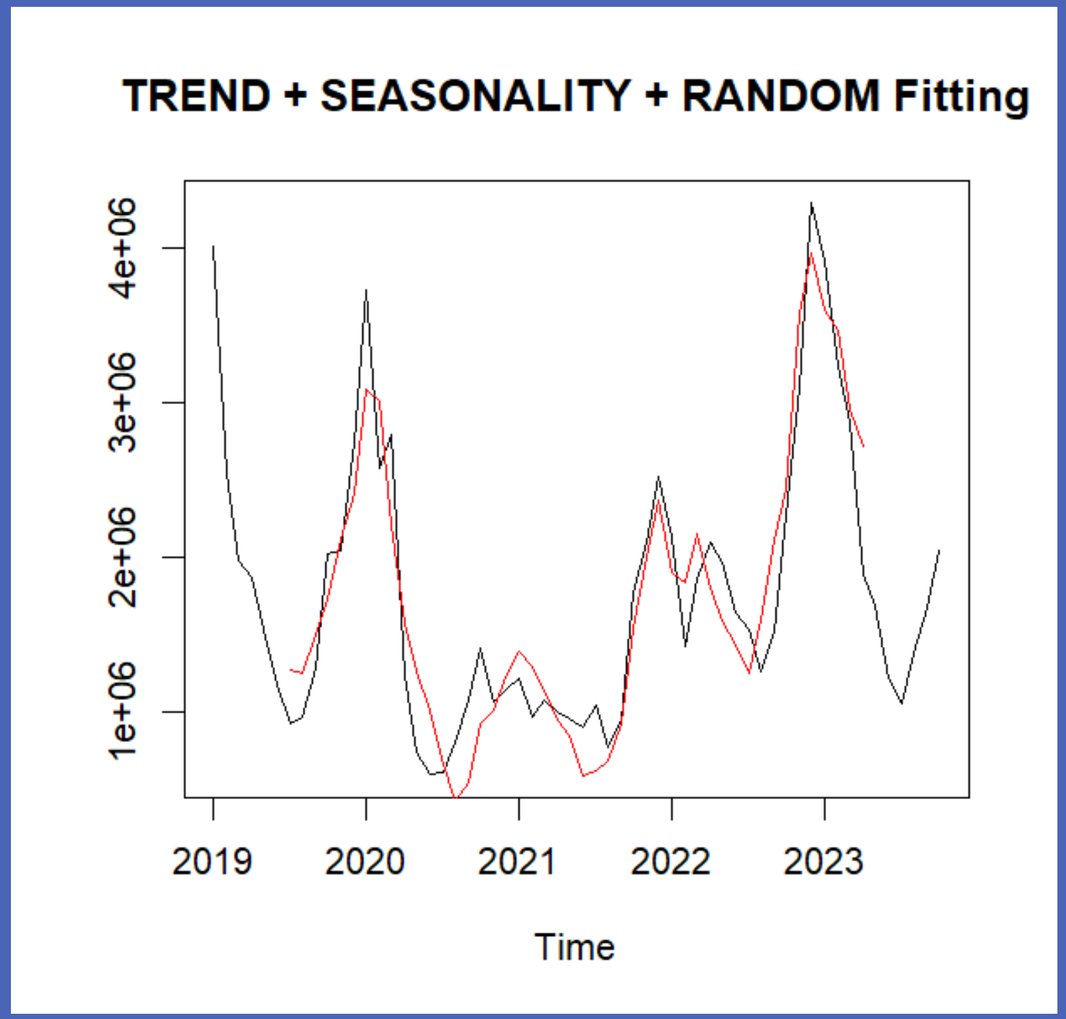


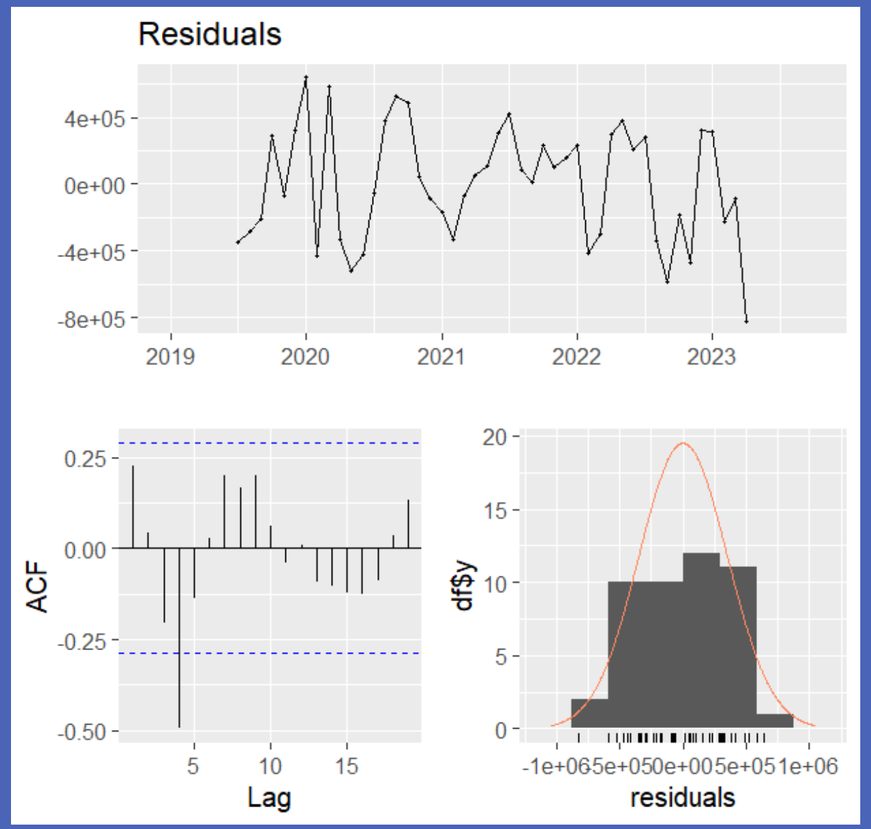
R-Squared	AIC	RMSE
0.7654	1314.55	348348.3



TIME SERIES DECOMPOSITION ANALYSIS - TREND+SEASON+RANDOM

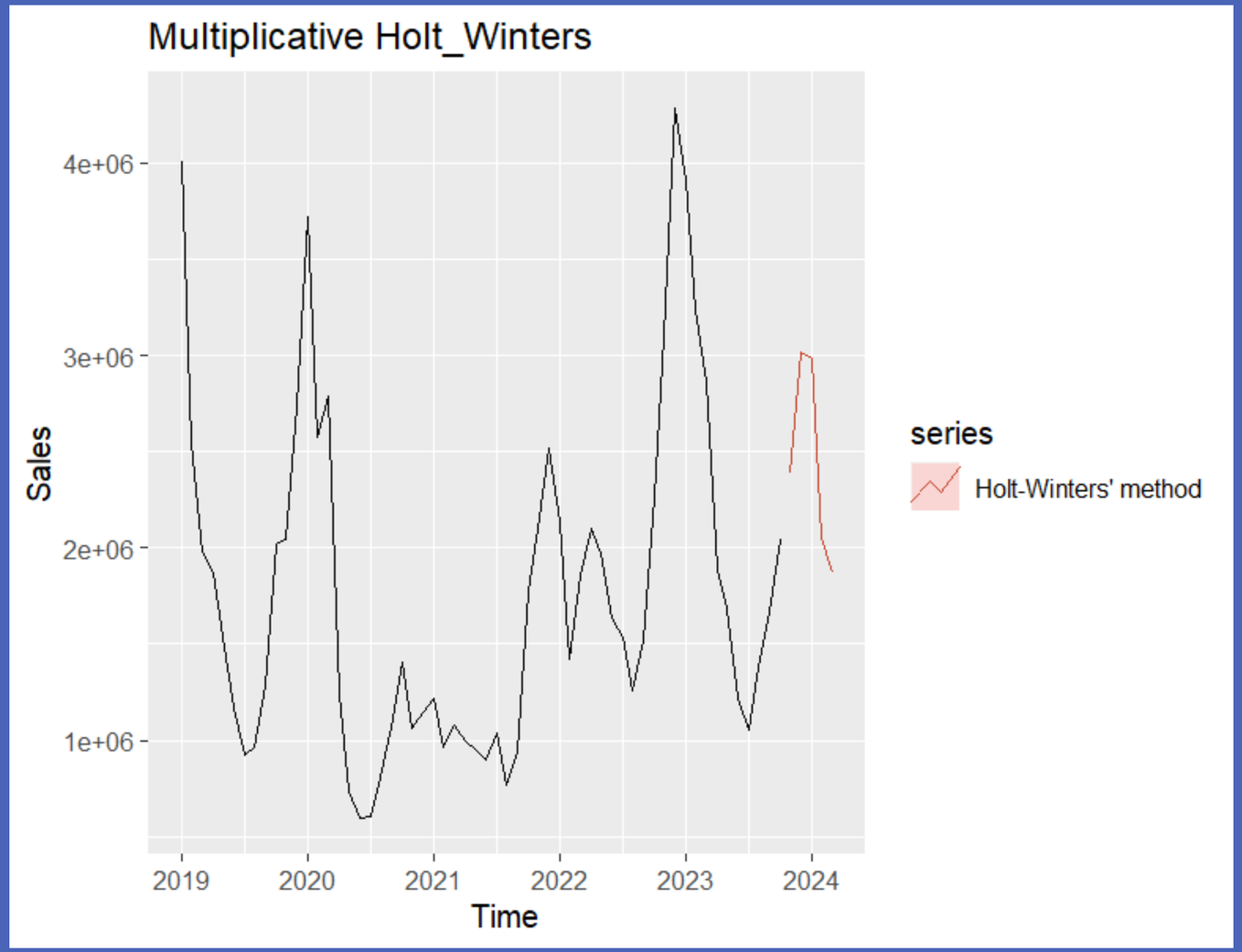
Despite of its great ability in iterpolating the data this model can't be used because it misses the fit of a part of the data





RMSE	344744.2

HOLT-WINTERS EXPONENTIAL SMOOTHING



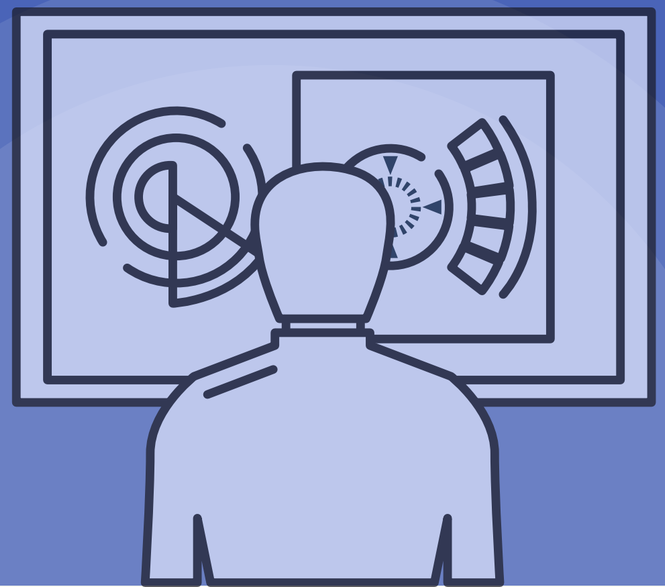
- We set beta (trend parameter) to 0.1

R-Squared	AIC	RMSE
-	1738.451	323264.3

BEST MODELS SELECTION

MODEL	PREDICTORS	R-Squared	AIC	RMSE
LINEAR REGRESSION	trend + season + covid_ts + disoc_ts + influenza_ts	0.8993	1653.742	280507
ARMAX with ARIMA(1,0,0)(1,0,0)[12] errors	Linear regressors + ARIMA(1,0,0)(1,0,0)12 errors	-	1615.81	250580.7
MANUAL ARIMA	ARIMA(1,0,0)(0,1,1)[12]	-	1339.97	344510
AUTO ARIMA	ARIMA(0,1,0)(1,1,0)[12]	-	1316.84	456323.7
EXPONENTIAL SMOOTING	Holt Winters Exponential Smoothing with beta=0.1	-	1738.451	323264.3
GAM WITH SPLINES	sales_ts ~ s(covid_ts) + s(temp_ts) + influenza_ts	0.834	1659.217	336993.1

FORECASTS

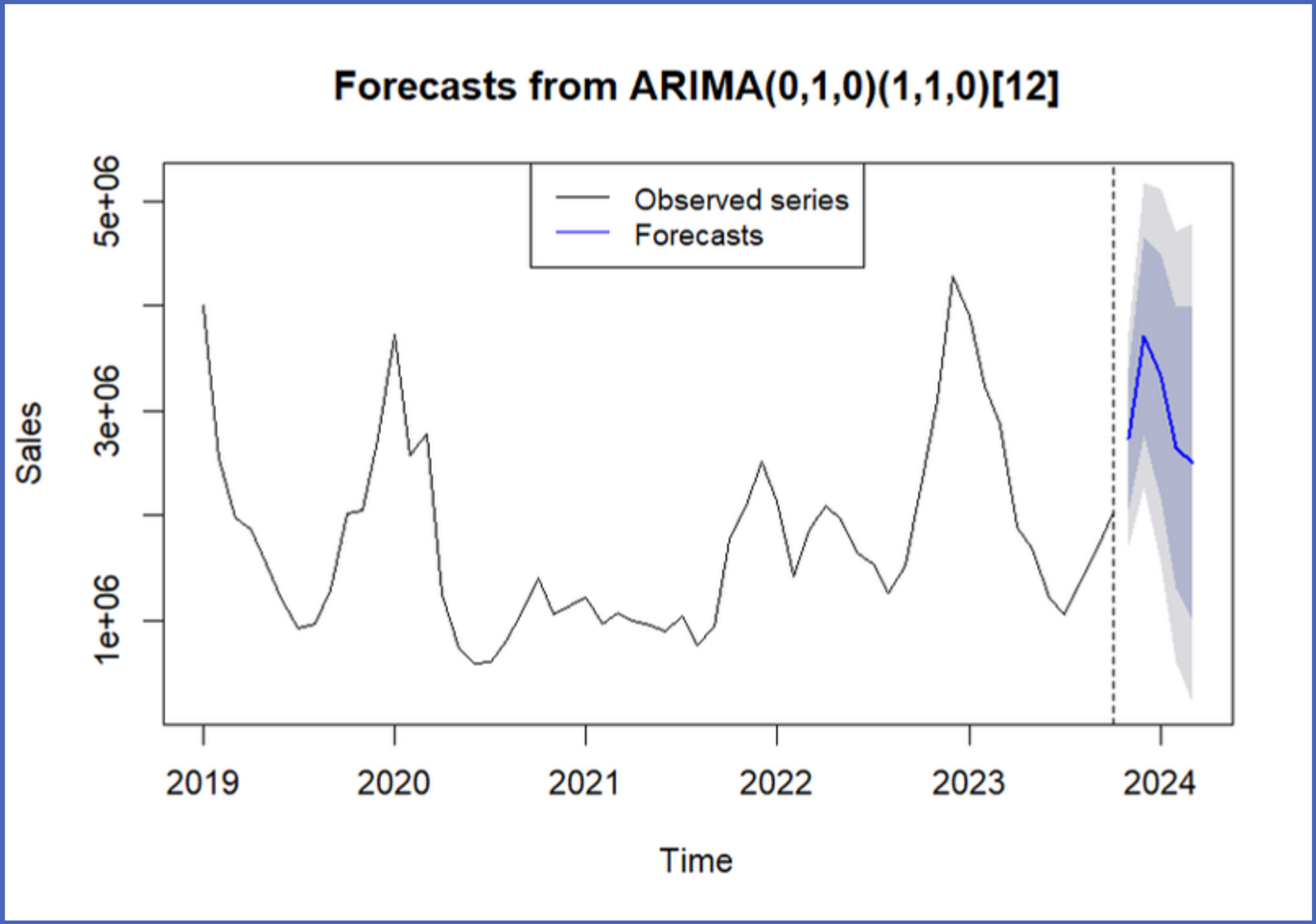


ARIMA with auto.arima

ARMAX

ARIMA with auto.arima

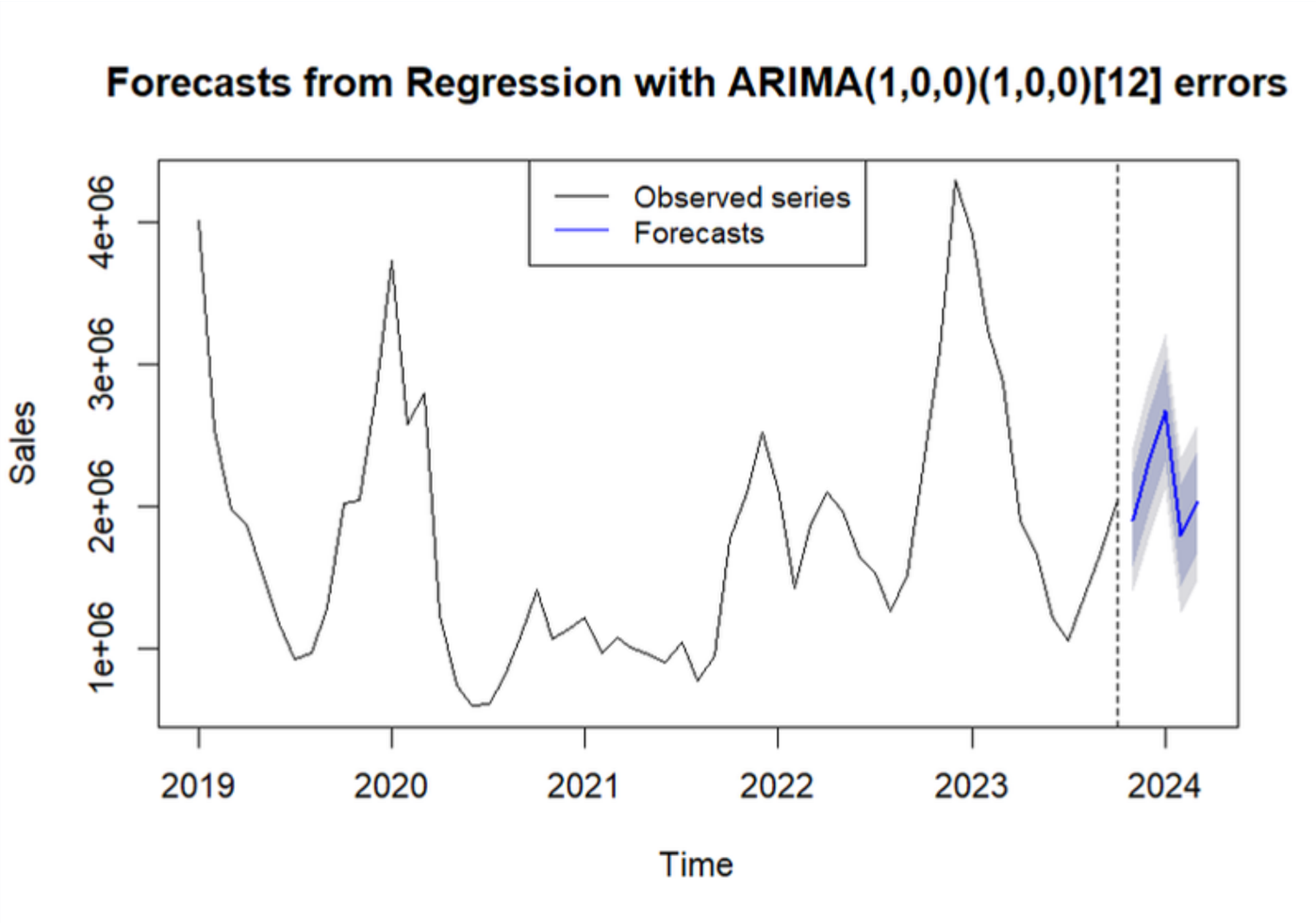
ARIMA(0,1,0)(1,1,0)[12]



	FORECAST	Lo 80	Hi 80
NOV 2023	2738301	2066877	3409725
DIC 2023	3716594	2767058	4666131
JAN 2024	3333901	2170960	4496841
FEB 2024	2652980	1310132	3995828
MAR 2024	2500173	998823	4001523

ARMAX

Regression with ARIMA(1,0,0)(1,0,0)[12] errors



	FORECAST	Lo 80	Hi 80
NOV 2023	1905757	1575983	2235531
DIC 2023	2300142	1946154	2654129
JAN 2024	2668322	2310792	3025852
FEB 2024	1795793	1437726	2153859
MAR 2024	2027497	1669349	2385645

CONCLUSIONS

GUESS THE PRODUCT

EXPLANATION OF THE RESULTS

GUESS THE PRODUCT

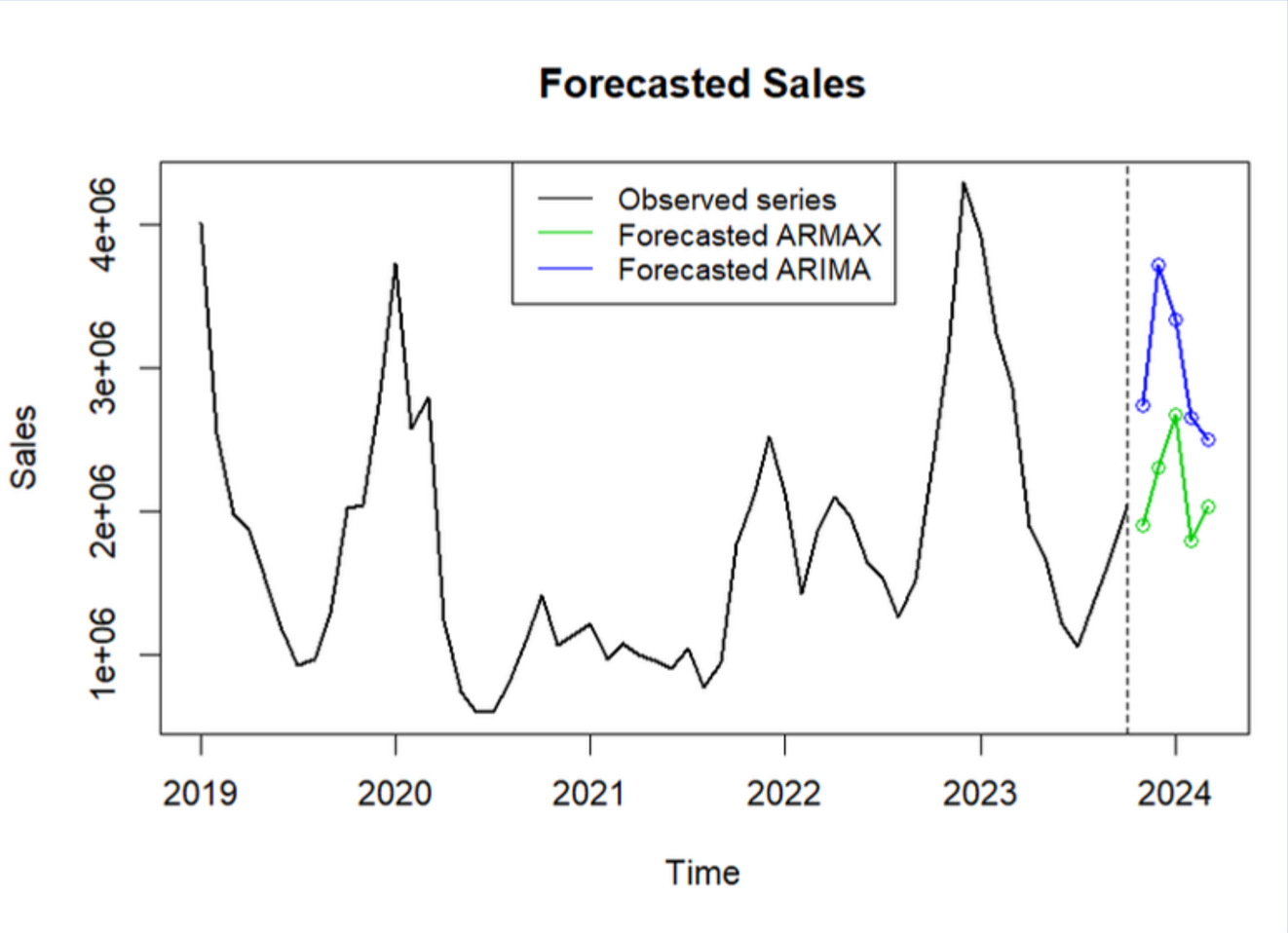
1. IS THE PRODUCT SEASONAL?
2. IS THE PRODUCT STRONGLY RELATED TO RESPIRATORY DISEASES INCIDENCE?
3. DOES THE PRODUCT FOLLOW TEMPERATURE SEASONALITY?

YES

1. IS THE PRODUCT ADMINISTRATED IN PILLS?

???

Forecasts descriptions and intepretations



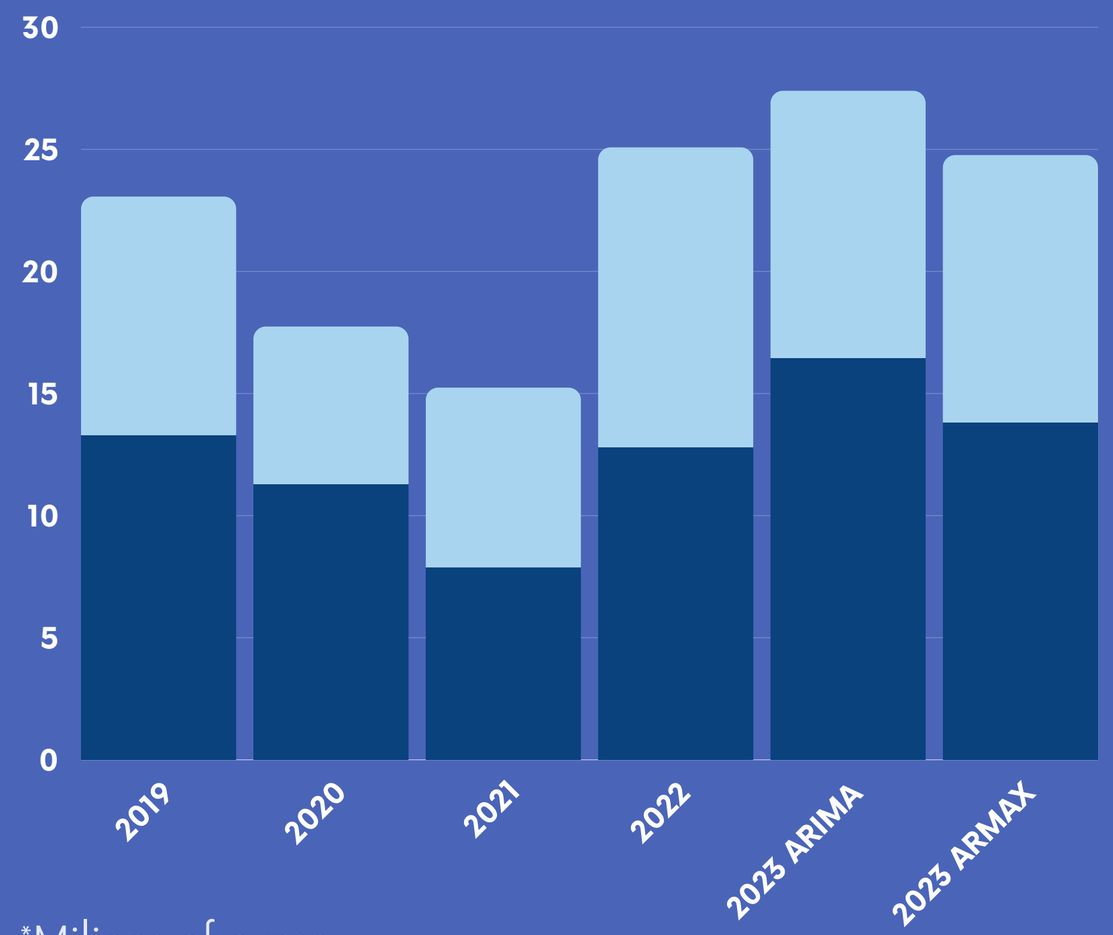
ARMAX
Regression with ARIMA(1,0,0)(1,0,0)[12] errors

- **Unexpected sales increase in March**
- **Level far lower than it was in the same months last year**
- **Possible overfitting of the data**

ARIMA
with auto.arima(0,1,0)(1,1,0)[12]

- **Slightly lower level than last year sales**
- **More coherent to the typical behaviour of the original data**

2023 SALES FORECAST COMPARISON



*Milions of euros

Forecasted 2023 sales with ARIMA: €27.398.021
-Highest value in the last four years-

Forecasted 2023 sales with ARMAX: €24.766.416

Forecasted 2023 % of sales in the peak season with ARIMA: 60%

Forecasted 2023 % of sales in the peak season with ARIMA: 56%

THANK YOU
FOR YOUR ATTENTION!

Thank You