Consumption Expenditure of Households in The Netherlands

HarvardX PH125.9x - Data Science: Capstone

Safeen Ghafour

12/25/2020

Contents

1 Introduction

For the purpose of this projects we will conduct time series analysis to extract meaningful statistics and insights of the data. Thereafter we will try to forecast future values using various models based on historical observations.

Specifically, we will use the publicly available data of the "Consumption expenditure of households in The Netherlands" from January 2000 to October 2020 for our analysis and forecasting.

Time series analysis and forecasting are widely used in economics, management, production and planning.

Forecasting time series can be challenging due to the vast number of factors that can affect such action and make it highly sensitive. Many irregularities and interruptions may require retraining or even remodelling. This is specially true, in case of irregularities caused unexpected events like the COVID pandemic

Therefore, forecasting is the art of understanding uncertainty.

2 Objectives

We have two main objectives, using the Value Index of the expenditure of households:

1. Data analysis: to better understand various aspects of our data.
2. Forecasting: using different models to predict an unseen subset of data and evaluate and compare their outcomes.

3 The Data

The expenditures of households dataset is publicly available from the website of 'Het Centraal Bureau voor de Statistiek' (Central Agency for Statistics - The Netherlands). [1]

This dataset provides figures on the expenditures of households in values and volume. The figures are divided in domestic consumption and final consumption by households. This includes final consumption in the Netherlands by residents and non-residents.

Available from 2000, this dataset contains changes and indices of consumption of households by type of goods and services. [2]

```
## 'data.frame':    4942 obs. of  9 variables:
##  $ ID                                  : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ ConsumptionByHouseholds             : chr  "A047812" "A047812" "A047812" "A047812" ...
##  $ Periods                             : chr  "2000JJ00" "2001JJ00" "2002JJ00" "2003JJ00" ...
##  $ Indices_1                           : num  93.7 95.5 96.6 96.5 97.5 ...
##  $ VolumeChanges_2                     : num  3.6 1.9 1.1 -0.1 1 0.9 -0.1 2 0.6 -2.2 ...
##  $ VolumeChangesShoppingdayAdjusted_3: chr  "     3.6" "     2.1" "     1.1" "    -0.1" ...
##  $ Indices_4                           : num  72.8 76.6 80.1 81.9 83.8 86.1 88.4 91.9 94.5 91.2 ...
##  $ ValueChanges_5                      : num  6.7 5.3 4.6 2.2 2.3 2.8 2.6 4 2.8 -3.6 ...
##  $ PriceChanges_6                      : num  2.9 3.3 3.4 2.3 1.3 1.8 2.7 2 2.2 -1.4 ...
```

The original dataset has 4942 observations and 9 variables.

To better understand the data, the The Central Agency for Statistics provides a meta-data file available from the following URL: Consumption Expenditure of Households in The Netherlands - Metadata.

[1] https://opendata.cbs.nl

[2] Licensed under Attribution 4.0 International (CC BY 4.0) by https://www.cbs.nl

column_name	column_exp
ID	Index
ConsumptionByHouseholds	Consumption By Households is a nested category of products and services.
Periods	Date separated by Year, by Quarter and by Month.
Indices_1	Volume Indices: An index represents the ratio between the value of a certain variable in a certain period and the value of that same variable in the base period.
VolumeChanges_2	Volume Changes: The change of volume compared to the same period a year earlier.
VolumeChangesShoppingdayAdjusted_3	Volume Changes Shoppingday Adjusted: For shopping day adjusted change of volume compared to the same period a year earlier.
Indices_4	Value Indices: An index represents the ratio between the value of a certain variable in a certain period and the value of that same variable in the base period.
ValueChanges_5	Value Changes: The change of value compared to the same period a year earlier.
PriceChanges_6	Price Changes: The change of price compared to the same period a year earlier.

3.1 The variables

Periods

The original data starts from January 2000 to October 2020.

There are three different types of periods/dates in this variable which are separated by one of the delimiters:

MM: Monthly data

KW: Quarterly data

JJ: Yearly data

ConsumptionByHouseholds

Consumption by households is a nested category. We have added a sequence number to the category list:

Key	Name
A047812	1 Domestic consumption by households
A047813	1.1 Consumption of goods by households
A047875	1.1.1 Foodproducts, beverages and tabacco
A047825	1.1.2 Durable consumer goods
A047826	1.1.2.1 Textiles and clothing
A047827	1.1.2.2 Leather goods and footwear
A047828	1.1.2.3 Home furnishing and home decoration
A047829	1.1.2.4 Electrical equipment
A047830	1.1.2.5 Vehicles
A047831	1.1.2.6 Other durable consumer goods n.e.c.
A048214	1.1.3 Other goods
A047832	1.1.4 Electricity, gas, water and motor fuels
A048213	1.1.5 Personal care and other goods
A047837	1.2 Consumption of services by households

Indices_1 & ***Indices_4***

Both Volume and the Value Indices are considered to be 100 in ***2015**, which is the base period. All historical and future values of the indices are ratios of the base period values.

Value and volume are concepts used in publications about the economy. However, volume is often mistaken for quantity, which is just one of the three components that make up volume. For more information we refer the reader to this article The volume concept in economic publications.

VolumeChanges_2 & ***ValueChanges_5***

The change of volume and value compared to the same period a year earlier.

VolumeChangesShoppingdayAdjusted_3

Adjustment for the number of shopping days in a period. This data is incomplete.

PriceChanges_6

The change of price compared to the same period a year earlier.

4 Data analysis

We will base our analysis on the value index variable (Indices_4) of the expenditure of households.

Therefore, we need to modify the dataset to fit our purpose.

First we grab ***monthly data*** from the dataset and extract year and month columns and add labelled category, then we filter out the other variables.

```
## 'data.frame':    3500 obs. of  5 variables:
##  $ category_id      : chr  "A047812" "A047812" "A047812" "A047812" ...
##  $ year             : Factor w/ 21 levels "2000","2001",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ month            : Factor w/ 12 levels "01","02","03",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ consumption_value: num  72 69.7 74.5 72.3 74.2 72.6 71.6 71.6 72.8 72.5 ...
##  $ category         : Factor w/ 14 levels "1 Domestic consumption by households",..: 1 1 1 1 1 1 1 1
```

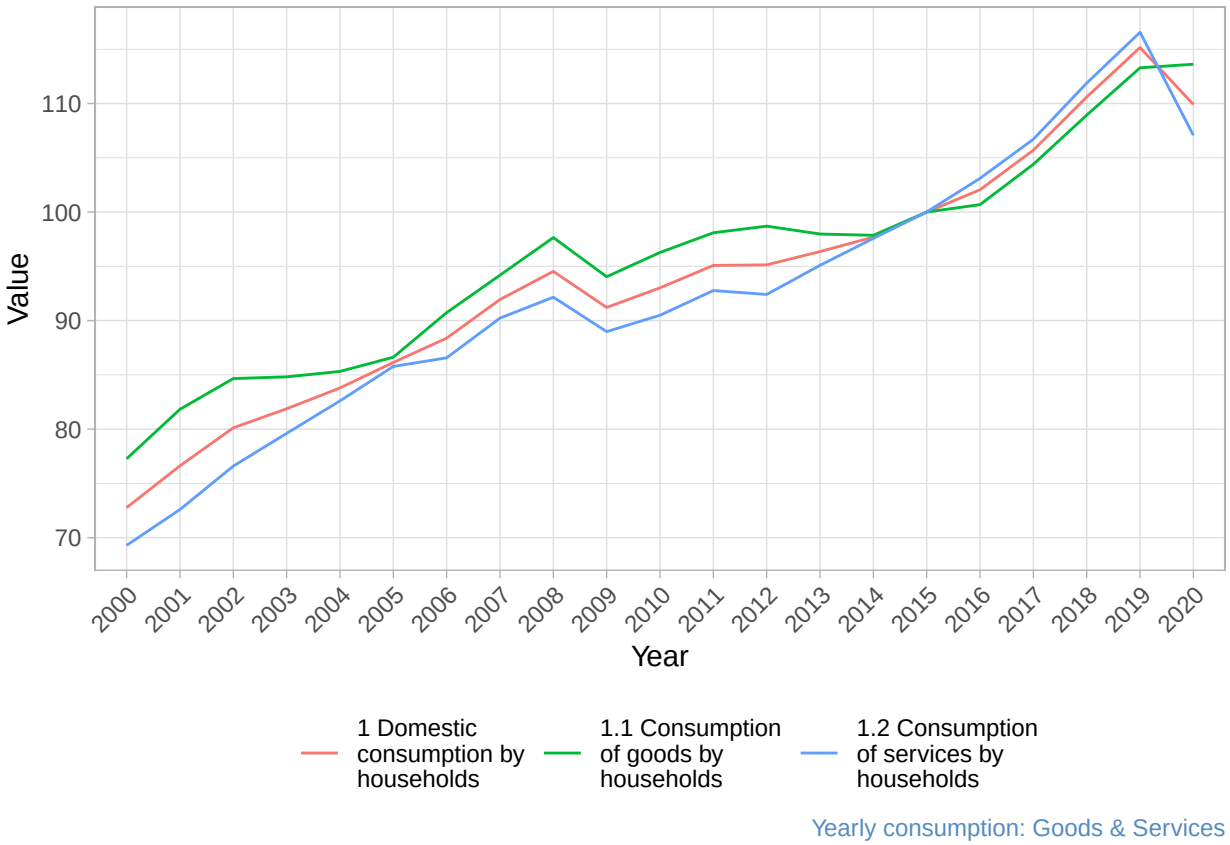
```
##   category_id year month consumption_value                              category
## 1     A047812 2000    01              72.0 1 Domestic consumption by households
## 2     A047812 2000    02              69.7 1 Domestic consumption by households
## 3     A047812 2000    03              74.5 1 Domestic consumption by households
## 4     A047812 2000    04              72.3 1 Domestic consumption by households
## 5     A047812 2000    05              74.2 1 Domestic consumption by households
## 6     A047812 2000    06              72.6 1 Domestic consumption by households
```

The modified dataset has 3500 observations and 5 variables.

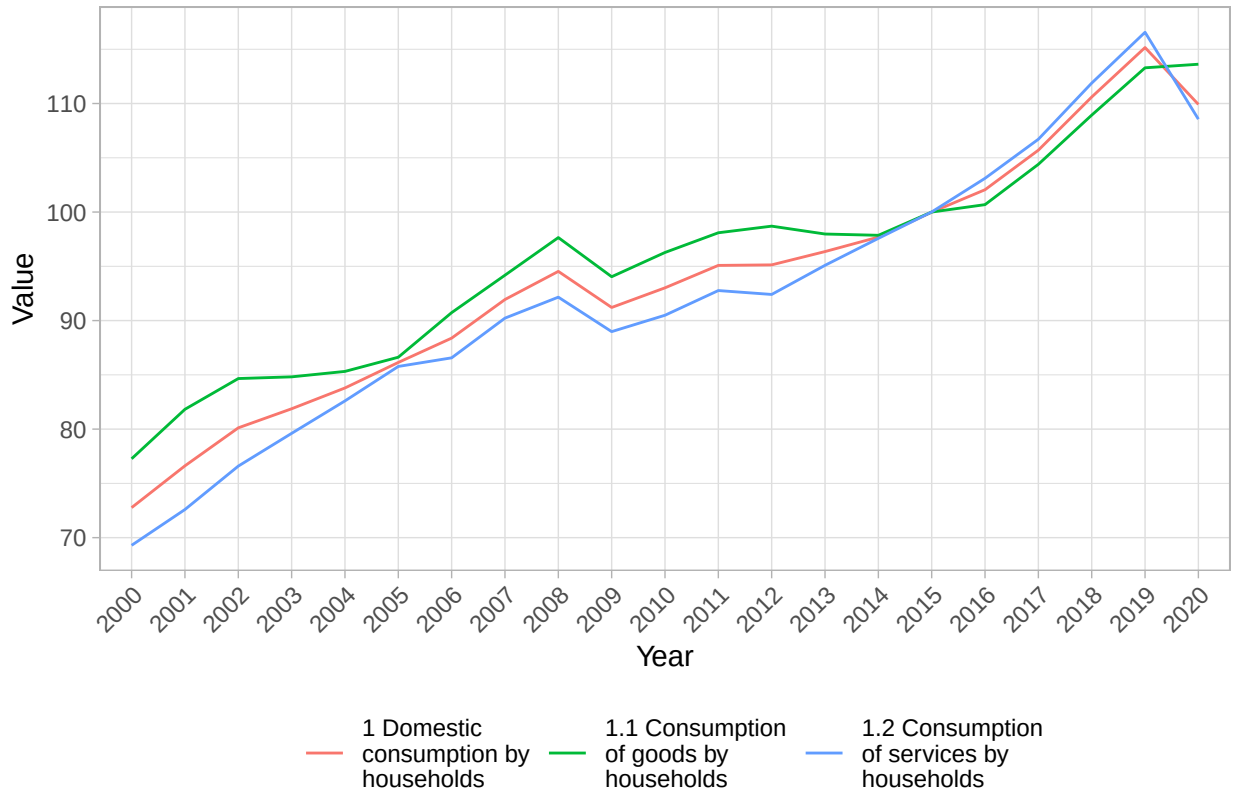
We plot the top category 'Domestic consumption by households' and its two major sub-categories 'Consumption of goods by households' & 'Consumption of services by households'.

The value of 'Domestic consumption by households' is the average value of its two sub-categories and their sub-subs. Logically the lines come together in 2015, when all the values for all individual categories are were set to be a 100 index.

Note: The average values of 2020 are divided by 10 months.



For the purpose of illustration only we fill the last two missing months of 2020, to be very optimistic, with a lag of 12 months. Not to be used in calculations hereafter.



Yearly consumption: Goods & Services ~ 2000−01 − 2020−12

The effect is not that significant and no matter how good the recovery is in the last two months of 2020, the damage is already done.

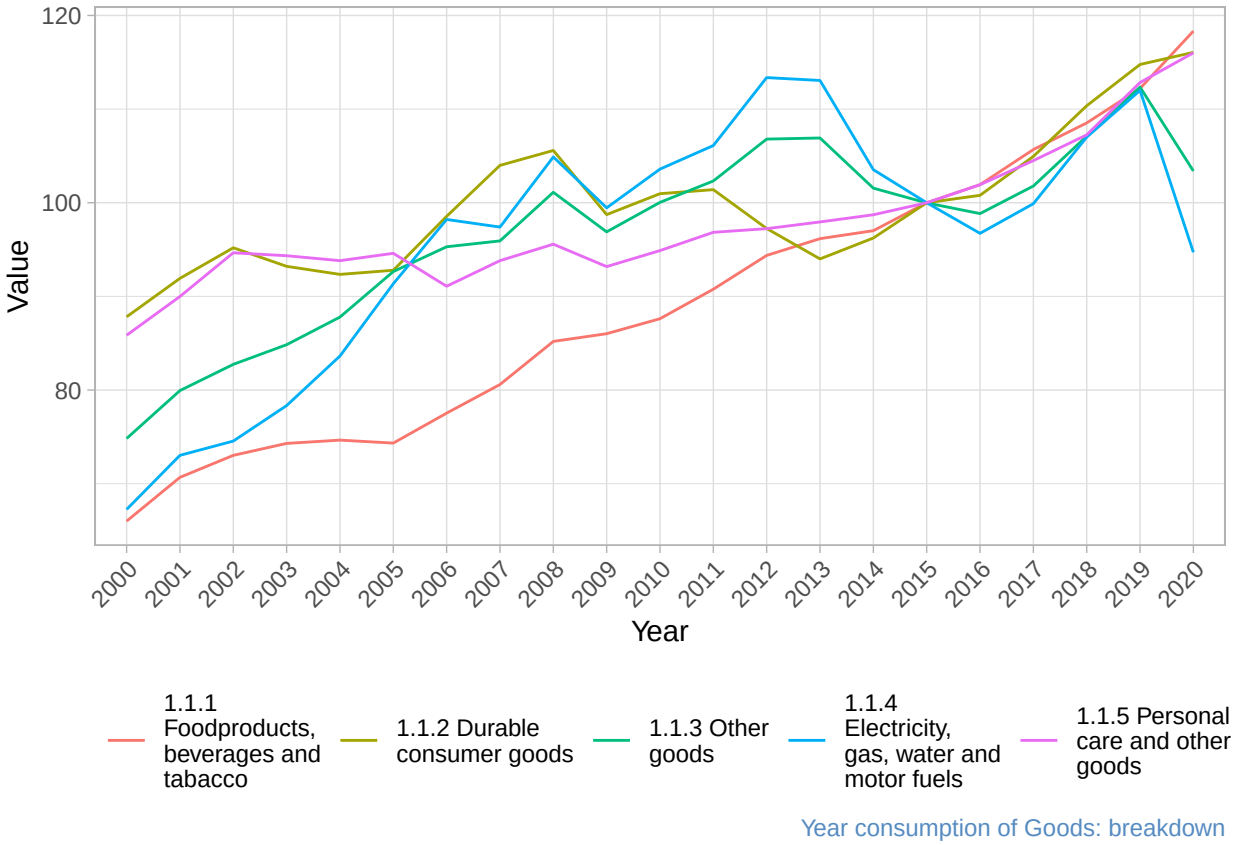
Both Goods and Services are dependent on a good performing economy.

During the 2009 financial crises both categories are almost equally effected. However, the fluctuations depend on the type of crisis as clear during the COVID pandemic.[3]

The consumption of Goods in 2020 is far less effected than the consumption of Services by COVID and lock-down. Services consumption, that includes Housing, Hotel, Recreational, Transport and Communication, Medical, Financial and Business services, has been hit hard.

We will continue with our original dataset and further zoom into the five major sub-categories of Goods.

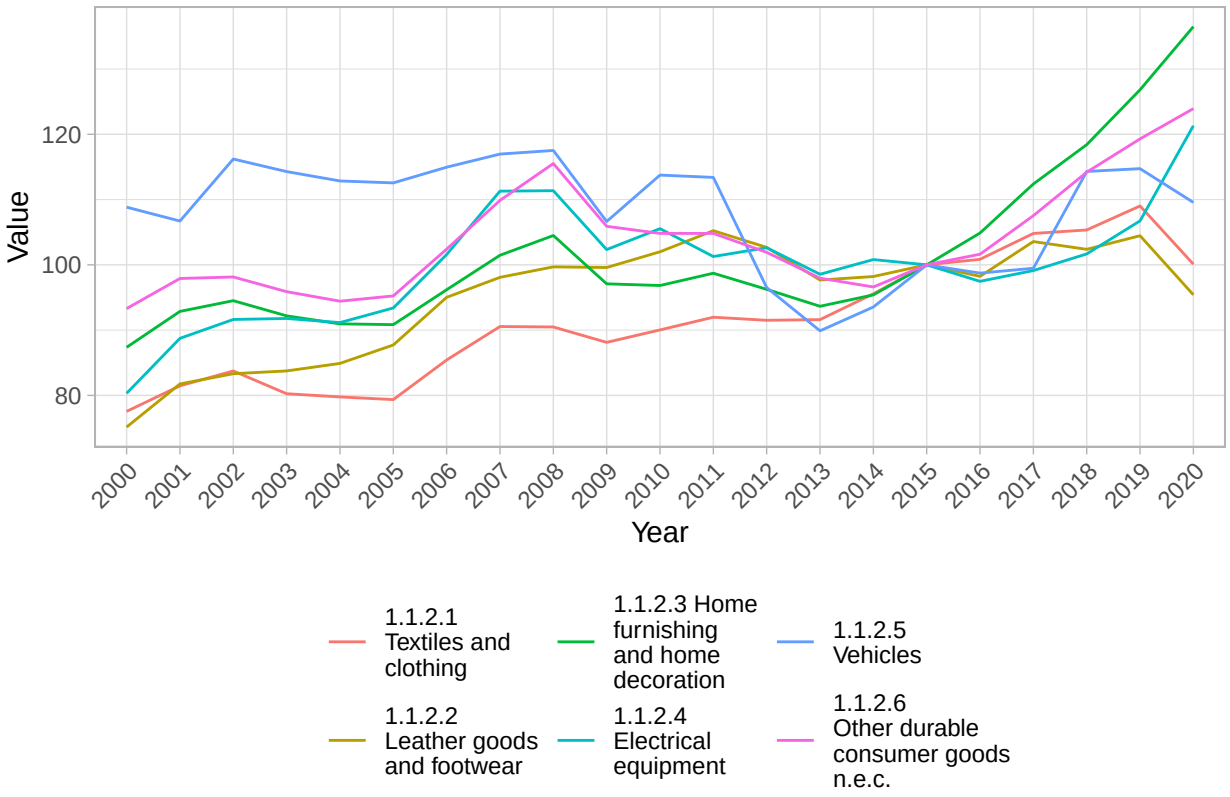
[3]https://en.wikipedia.org/wiki/Great_Recession



Year consumption of Goods: breakdown

This plot makes the effect of COVID even more clear. The consumption of 'Electricity, gas, water and motor fuels' and 'Other goods' are crashed while 'Foodproducts, beverages and tobacco' show a growth compared with previous years.

Now we plot 'Durable consumer goods'.



Year consumption of Durable consumer goods: breakdown

Here we notice a decline in the consumption of 'Vehicles', 'Textiles and clothing' and 'Leather goods and footwear'. Obviously, people had enough time to repair and decorate their houses.

To give an idea about the size of the expenditure in Euro's we refer you to Consumption by type of goods and services; National Accounts.

5 Time series nanalysis

In this section we will focus on the structure of our time series dataset which is a monthly time series data starting from November 2000 to October 2020.

The autocorrelation is a characteristic of time series, one observation depends on previous observations.

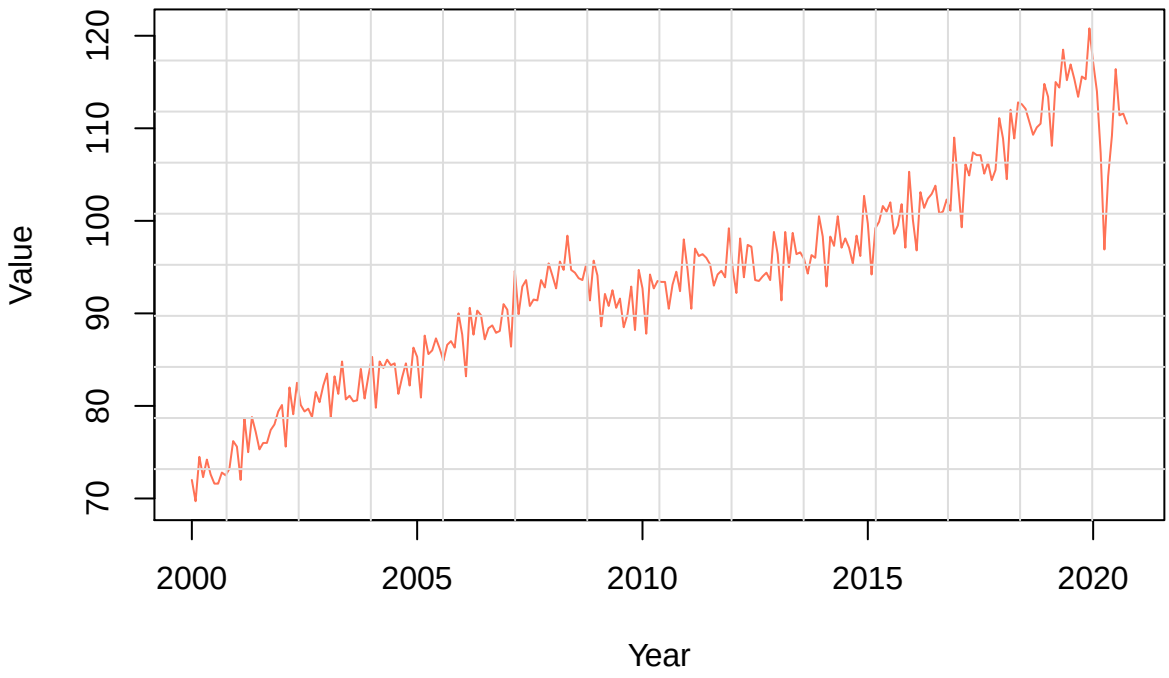
In the following analysis we try to identify patterns in our dataset based on the four major components trend, cycle, seasonality, and irregularity.

First we convert our dataset a time series object.

```
##  The general_consumption_ts series is a ts object with 1 variable and 250 observations
##  Frequency: 12
##  Start time: 2000 1
##  End time: 2020 10
```

5.1 Trend

It was already clear from "Figure: Domestic consumption: Goods & Services" that the series show an exponential trend upwards.



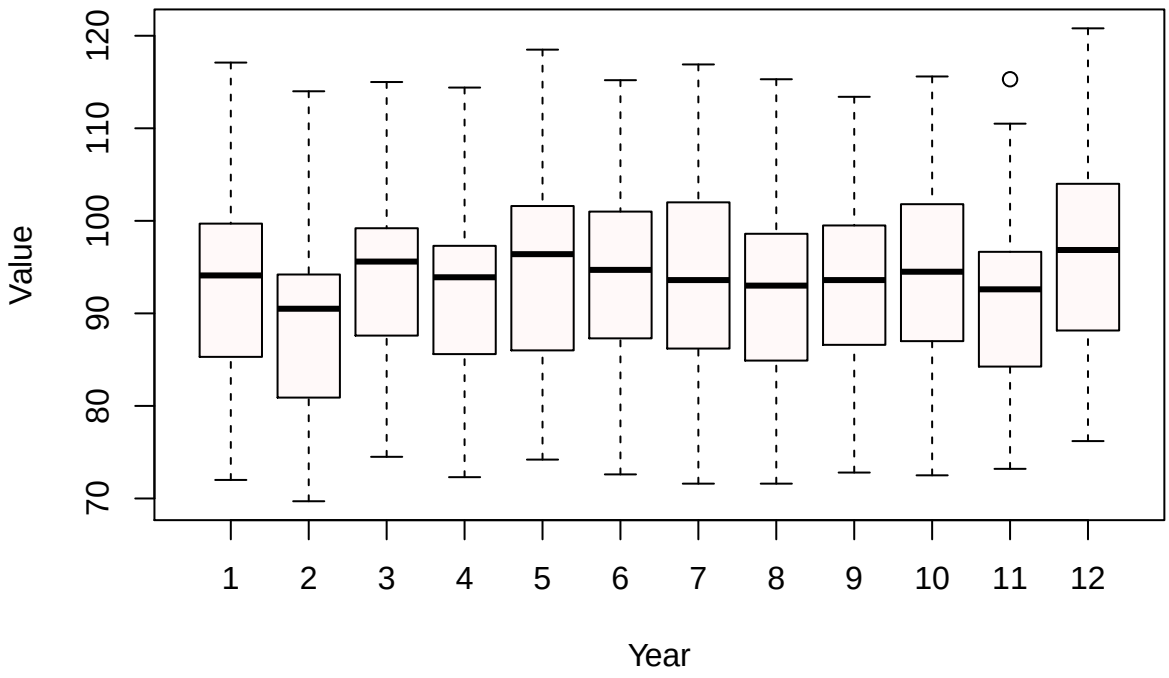
Trend component

5.2 Cycle

The same plot above shows no signs of clear cycles, that is regular ups and downs in the trend.

5.3 Seasonal

The presence of variations that occur at specific regular intervals less than a year in our case. To examine seasonality we will plot the values grouped by months.



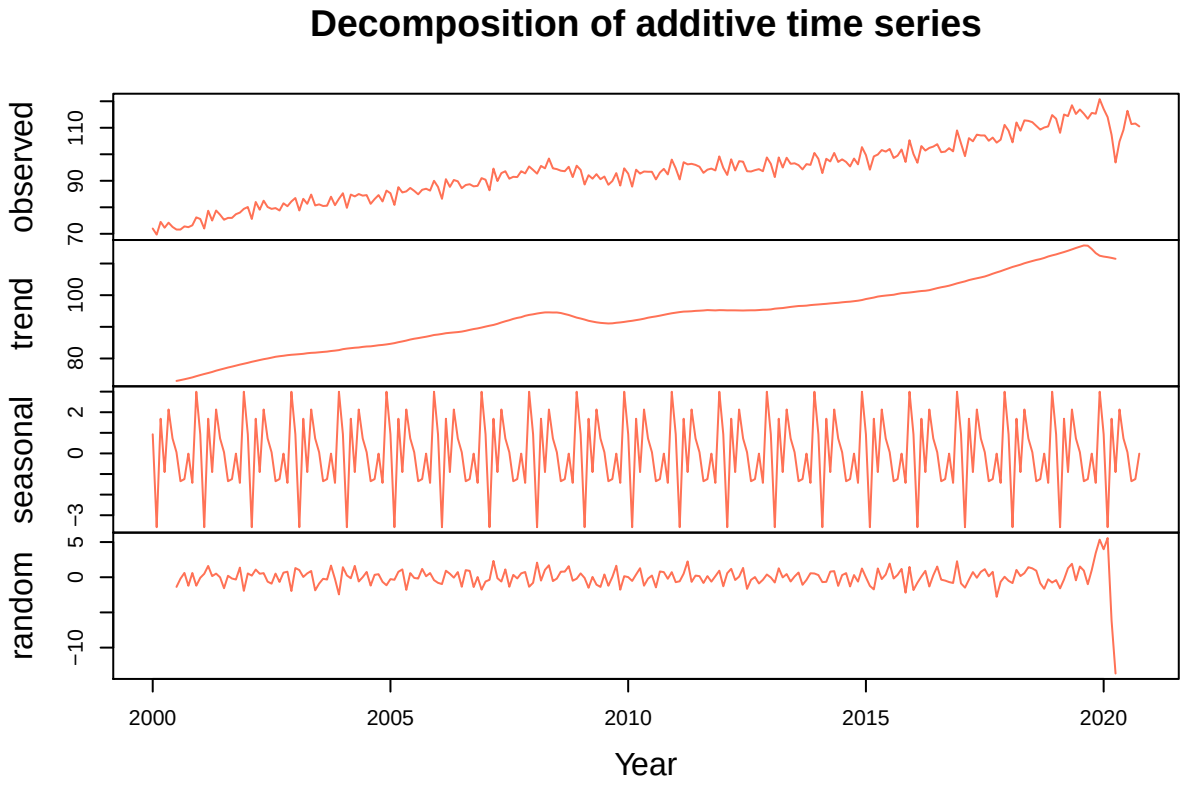
Lowest value is in February while the highest is in December.

5.4 Irregular

This component, which is the remainder between the series and its structural components, provides an indication of irregular events in the series.

There are two irregularities, which are not cycles, obvious in our dataset, a less strong one in 2009 and the second in 2020.

Finally we will decompose our time series dataset isolating each of the above mentioned patterns.



Decomposition

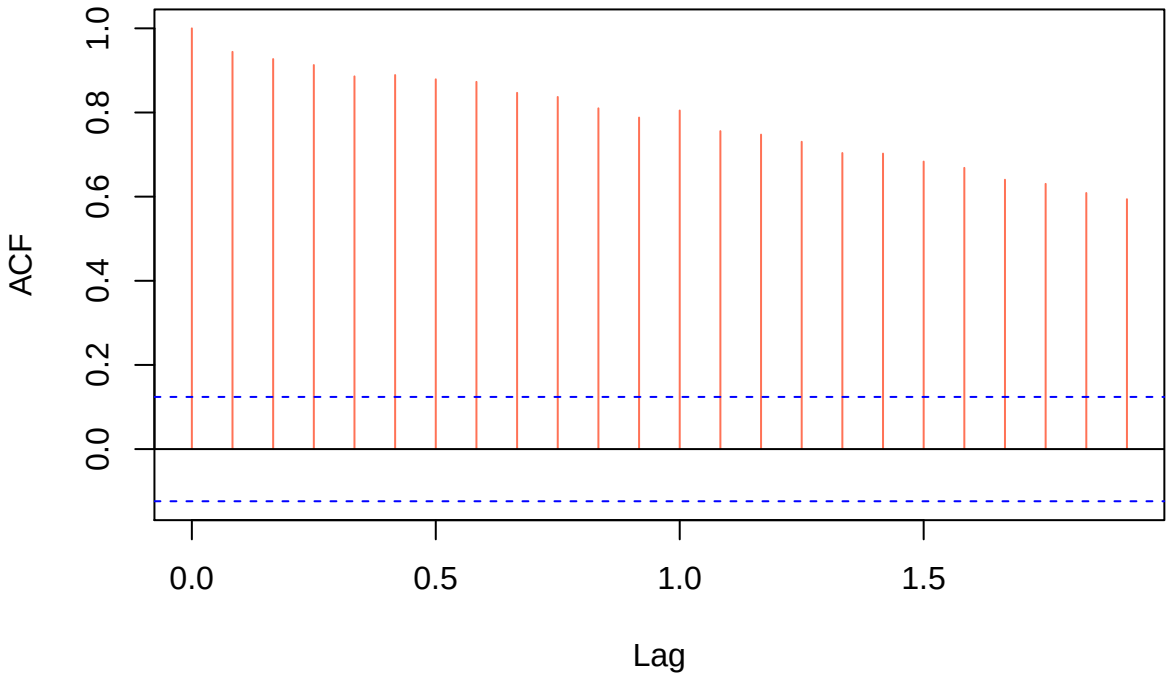
5.5 Autocorrelation

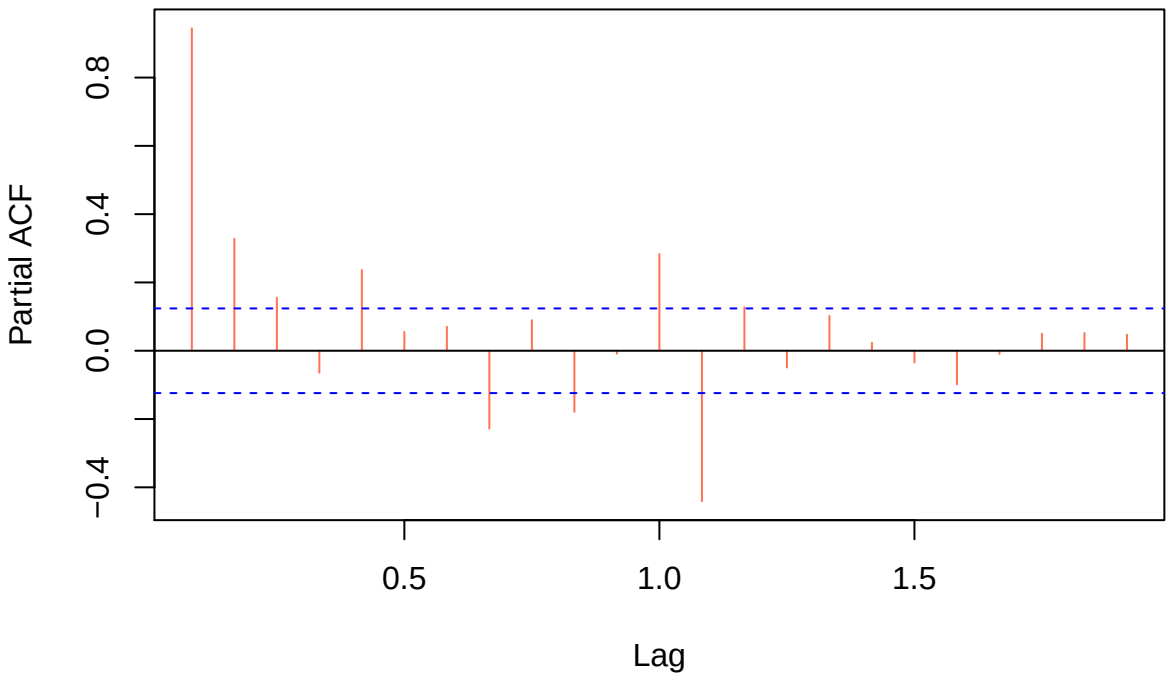
Autocorrelation is a correlation of a series with itself (own lagged values), which means it dependence on its past values.

We will measure and plot the correlation between the series and its lags using the autocorrelation function (ACF).

The values that cross the dashed blue lines (statistical significance), mean that the correlation significance of the lag with current series.

We will also use partial autocorrelation function (PACF), that is the amount of correlation between a time series and lags of itself that is not explained by a previous lag (residuals).


ACF
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
Lag
ACF



PACF

The correlation of the series with its lags is decaying over time and chronologically closer lags to the series show a stronger relation.

5.6 Stationarity

Many time series models require the data to be stationary.

A time series is stationary if its mean, variance and covariance remain constant over the whole series.

We can test stationarity by looking at the decomposition plot. Both Trend and Seasonal components are good indications that our time series is not stationary.

Another way to perform such test using the ACF. The strong relation between the series and its lags is another indicator.

Finally we will use Dickey-Fuller hypothesis testing. The Null Hypothesis: The series is not stationary.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  general_consumption_ts
## Dickey-Fuller = -2.2212, Lag order = 6, p-value = 0.4827
## alternative hypothesis: stationary
```

With the p-value greater than 0.05, we fail to reject the null hypothesis & confirm that the series is “not stationary”.

To fix this we use differencing. Differencing is the process of subtracting one observation from another.

```
##
```

```
##  Augmented Dickey-Fuller Test
##
## data:  diff(general_consumption_ts, differences = 1)
## Dickey-Fuller = -7.1098, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

With the p-value less than 0.05, we can now reject the null hypothesis and confirm that the series is stationary.

6 Forecasting

The second objective of this project is to forecast using different models to predict an unseen subset of data and evaluate and compare the outcomes.

First, We will split the dataset into two sequencial partitions, leaving the last 60 observations of the series as the testing partition and the rest as training.

```
##  The train series is a ts object with 1 variable and 190 observations
##  Frequency: 12
##  Start time: 2000 1
##  End time: 2015 10
```

```
##  The test series is a ts object with 1 variable and 60 observations
##  Frequency: 12
##  Start time: 2015 11
##  End time: 2020 10
```

6.1 Scoring

Common methods for evaluating the success of the forecast in order to predict the actual values, are accuracy or error metrics.

We will use the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE).

RMSE = The root mean of all the squared differences between the actual value at a certain time and the forecasted value at the same time.

MAPE = The mean of the absolute differences between the actual value at a certain time and the forcasted value at the same time divided by the actual value at the same time then multiplied by a hundred.

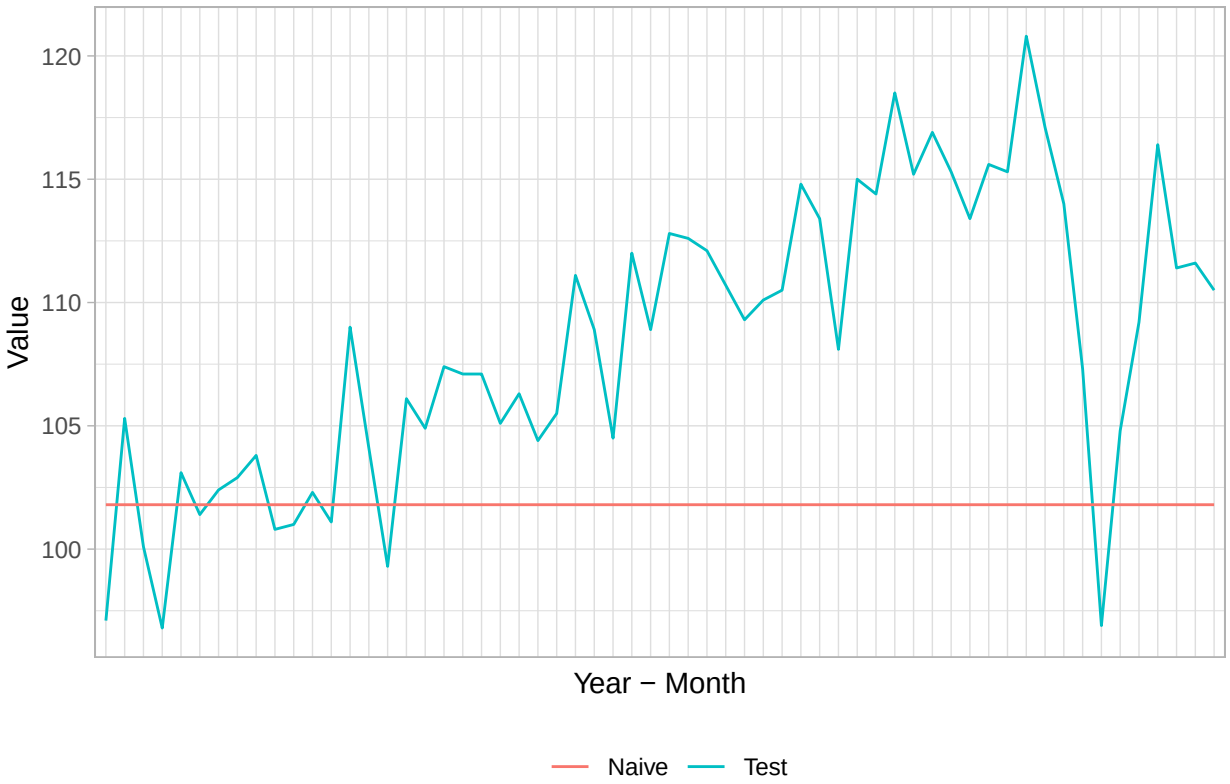
6.2 Benchmarking

To asses the results we obtained from RMSE and MAPE we will benchmark the model's performance to a baseline forecast.

We use naive forecast, that is the most recently observed value aplied to all predictions.

Naive forecast

```
## # A tibble: 1 x 3
##   Method  RMSE  MAPE
##   <chr>  <dbl> <dbl>
## 1 Naive   8.77  6.55
```



Compared prediction results

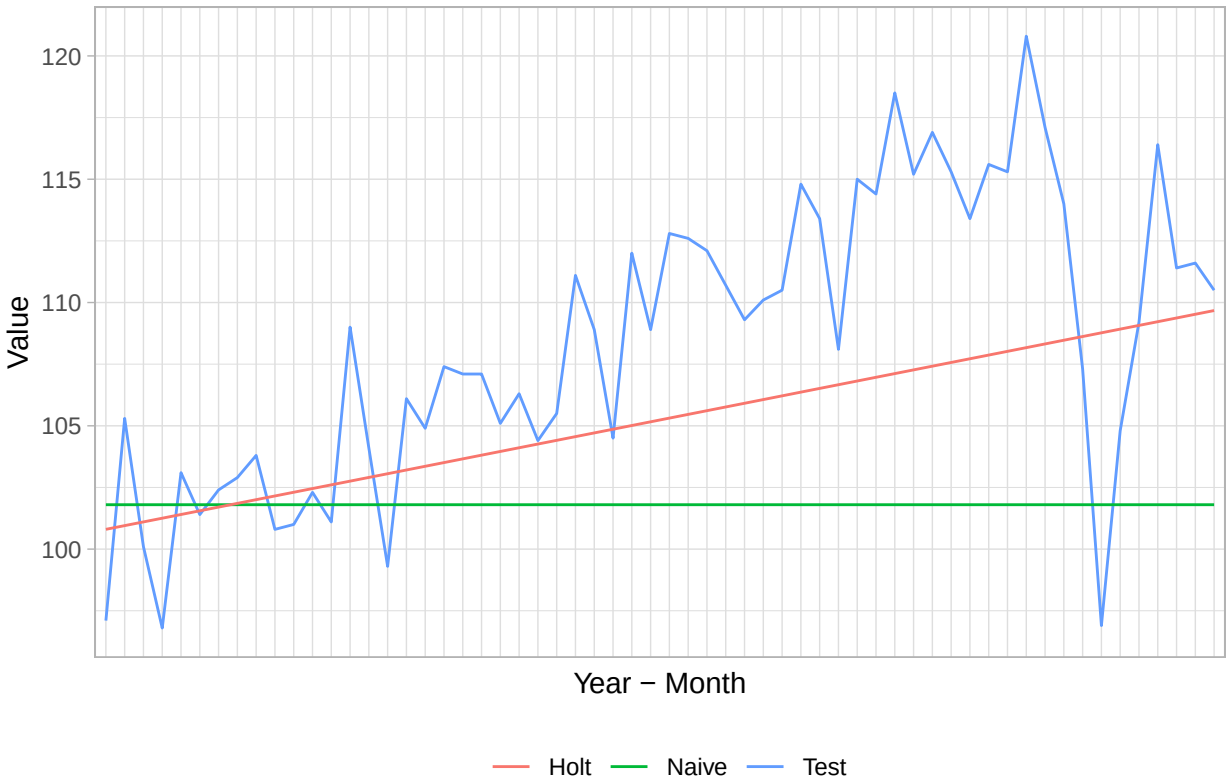
We will compare this data to evaluate performance, the lower the better. Our base RMSE is 8.77 and base MAPE is 6.55. If any other models scores lower than this we consider it better.

From the above plot we can clearly see the straight Naive line starting from the last value of the training set.

6.3 Holt's Trend

This is an extension of the simple exponential smoothing method, also known as linear exponential smoothing, which considers the trend component while generating forecasts. This method involves two smoothing equations, one for the level and one for the trend component.

```
## # A tibble: 2 x 3
##   Method  RMSE  MAPE
##   <chr>  <dbl> <dbl>
## 1 Naive   8.77  6.55
## 2 Holt    5.38  3.91
```



Compared prediction results

The results of Holt's Trend look much better than our Naive benchmark and the plot follows an upward trend.

6.4 ARIMA

One of the most common methods used in time series forecasting is known as the ARIMA model, which stands for Auto Regressive (AR) Integrated (I) Moving Average (MA).

It is based on the existence of correlation between a time series and its lags.

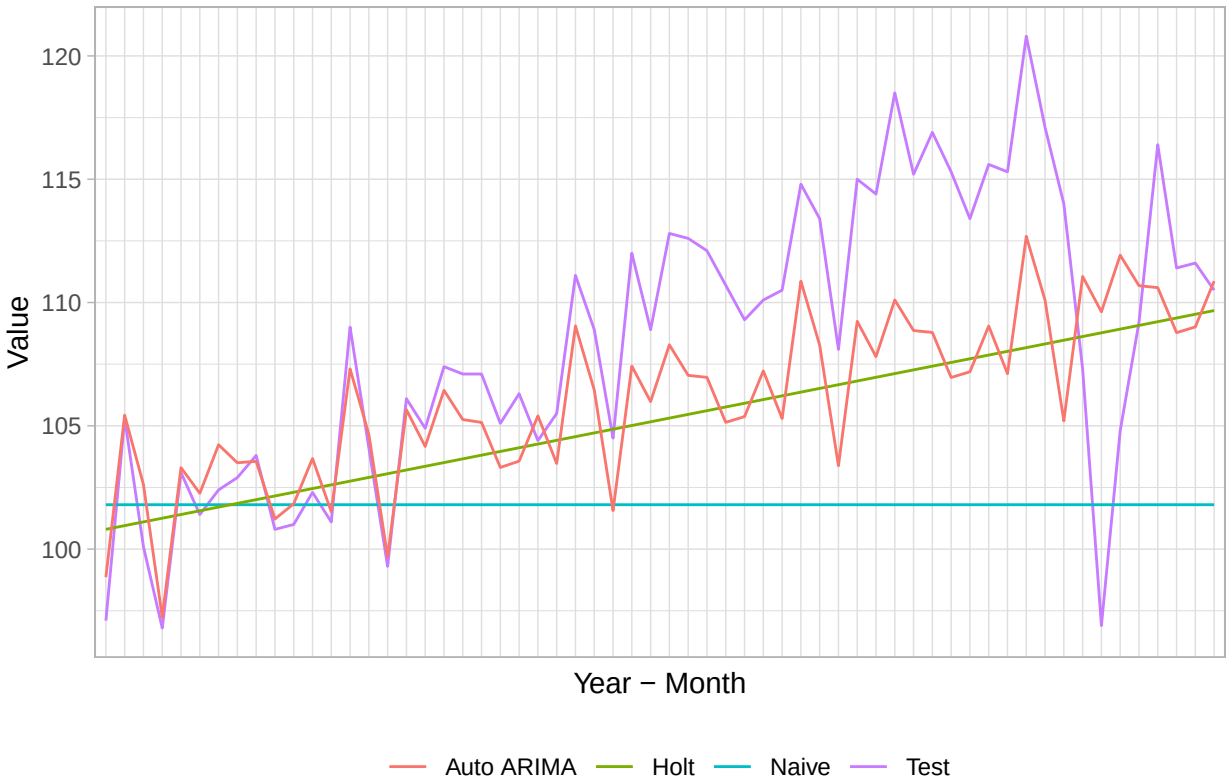
AR(p) Autoregressive model is based on the idea that the current value can be explained as a function of past p values, where p is the number of steps into the past (lag) needed to forecast the current model.

I(d) A function to make a time series stationary using differencing where d is the number of differencing transformations required by the time series to become stationary.

MA(q) moving average model uses past forecast errors in a regression-like model where q is order; past error (multiplied by a coefficient).

Since we use auto.arima function, the values of p, d and q will be calculated automatically.

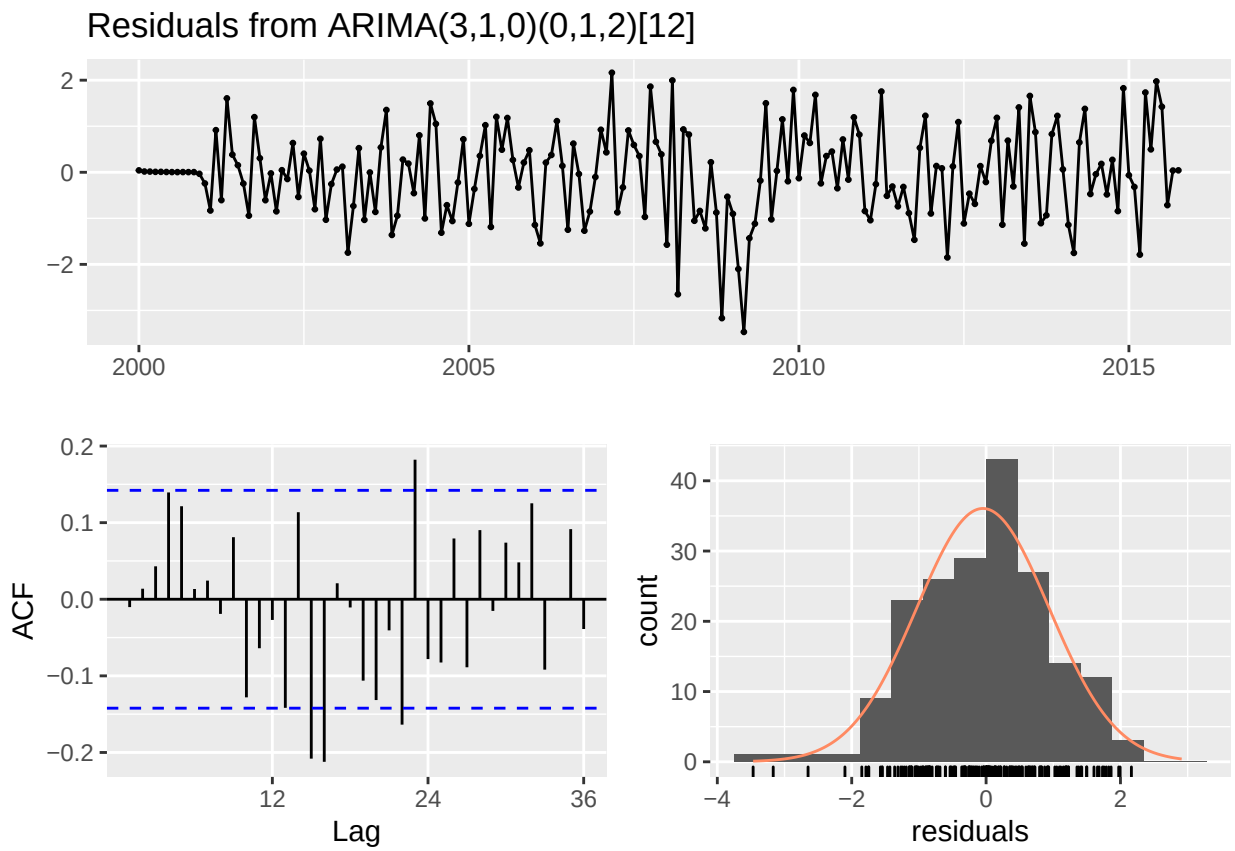
```
## # A tibble: 3 x 3
##   Method  RMSE  MAPE
##   <chr>  <dbl> <dbl>
## 1 Naive   8.77  6.55
## 2 Holt    5.38  3.91
## 3 ARIMA   4.58  3.22
```



Compared prediction results

Now the RMSE and MAPE have decreased and the plot depicts that the model picks the Trend and Seasonality. However the unexpected effect of the COVID lock-down is pretty obvious.

Now we will check the residuals of the ARIMA model.




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,0)(0,1,2)[12]
## Q* = 59.272, df = 19, p-value = 5.041e-06
##
## Model df: 5.   Total lags used: 24
```

Obviously, p equals 3 and with 1 differencing and q is 0, which means no moving averages.

Ljung-Box test

The p-values for the Ljung–Box statistics are small, indicating there is some pattern in the residuals. The p-value is less than 0.05, which means that we cannot reject the null hypothesis that there is no autocorrelation is left.

ACF

The ACF shows significant autocorrelations between residuals and the model did not fully capture all of the patterns.

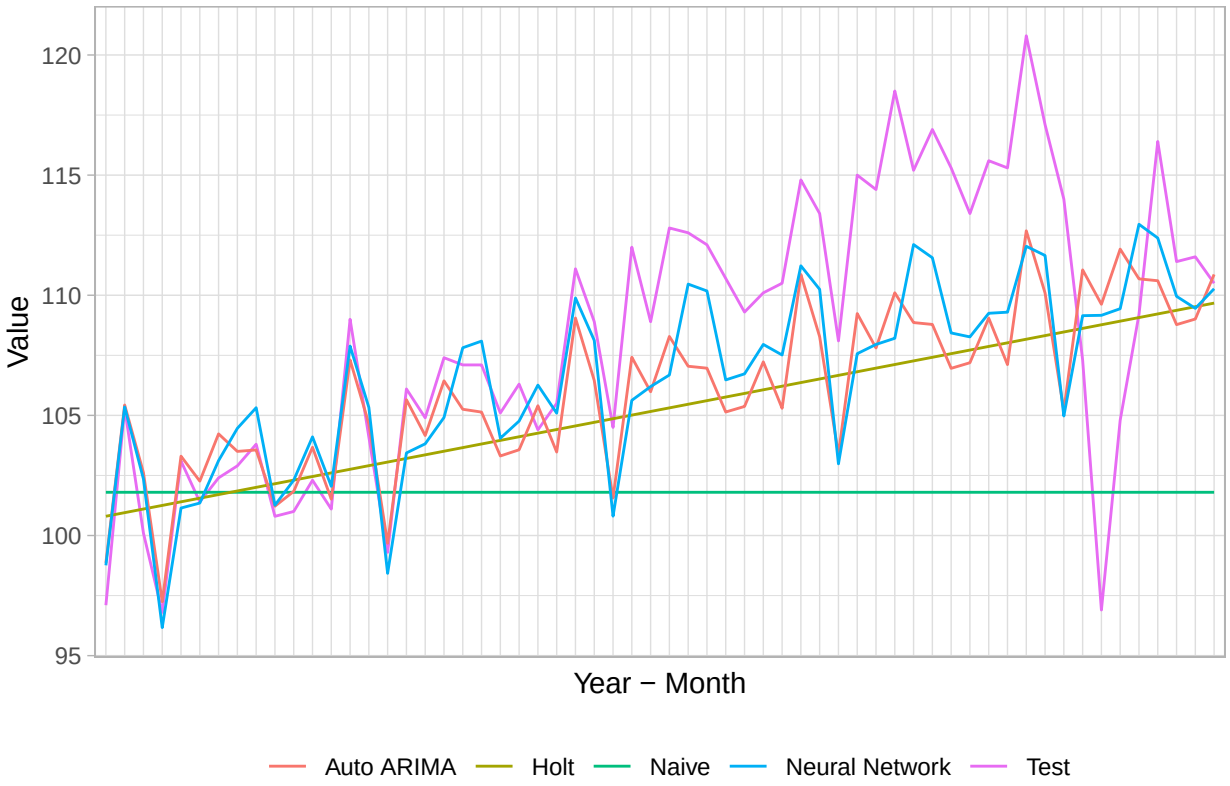
6.5 Neural Network AutoRegression

Artificial neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors.

In the Neural Network AutoRegression (NNAR) the lagged values of the time series is used as inputs to a neural network.

```
## # A tibble: 4 x 3
```

```
##   Method          RMSE  MAPE
##   <chr>          <dbl> <dbl>
## 1 Naive           8.77  6.55
## 2 Holt            5.38  3.91
## 3 ARIMA           4.58  3.22
## 4 Neural Network  4.16  2.87
```

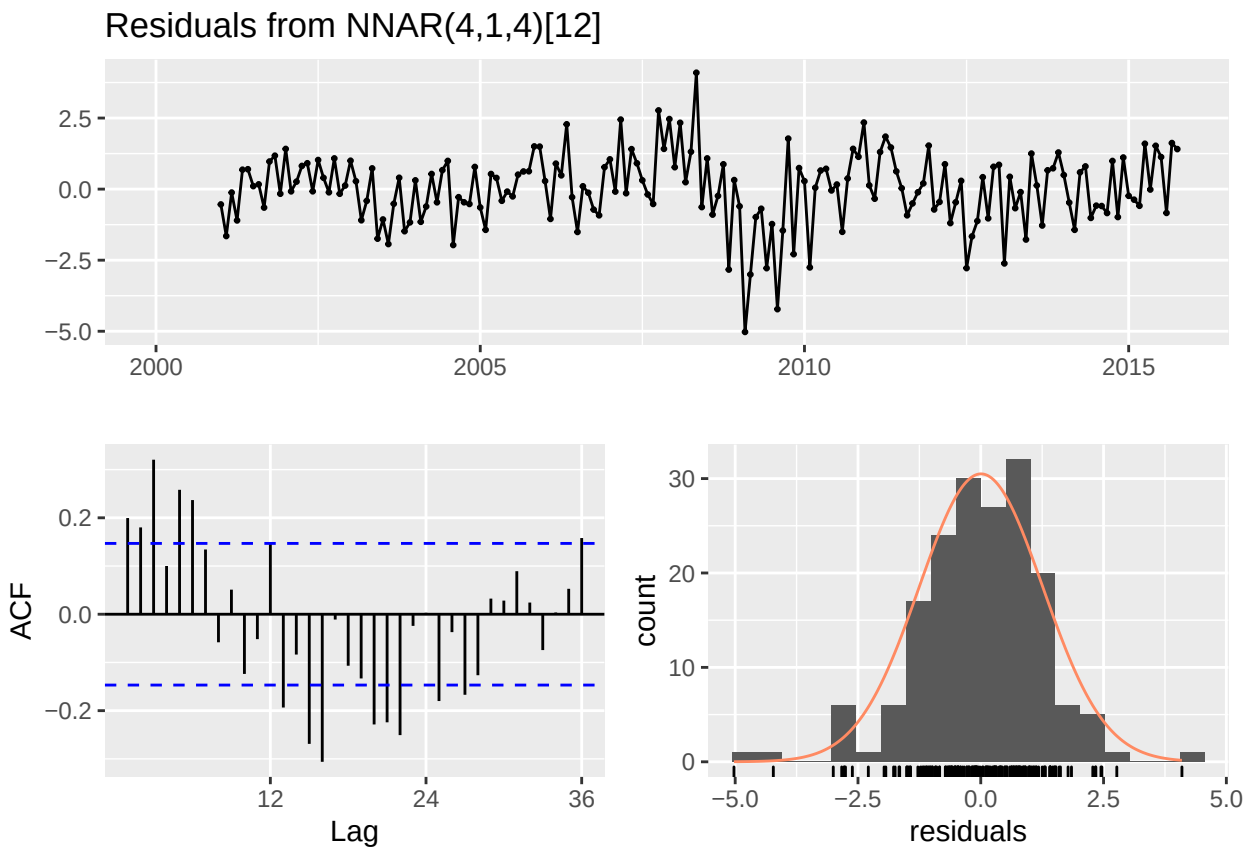


Compared prediction results

The RMSE and MAPE are slightly improved and again the plot shows that the model picks the Trend and Seasonality.

Now we will check the residuals of the Neural Network model.

```
##
##  Box-Ljung test
##
## data:  nn_forecast$residuals
## X-squared = 33.753, df = 4, p-value = 8.372e-07
```



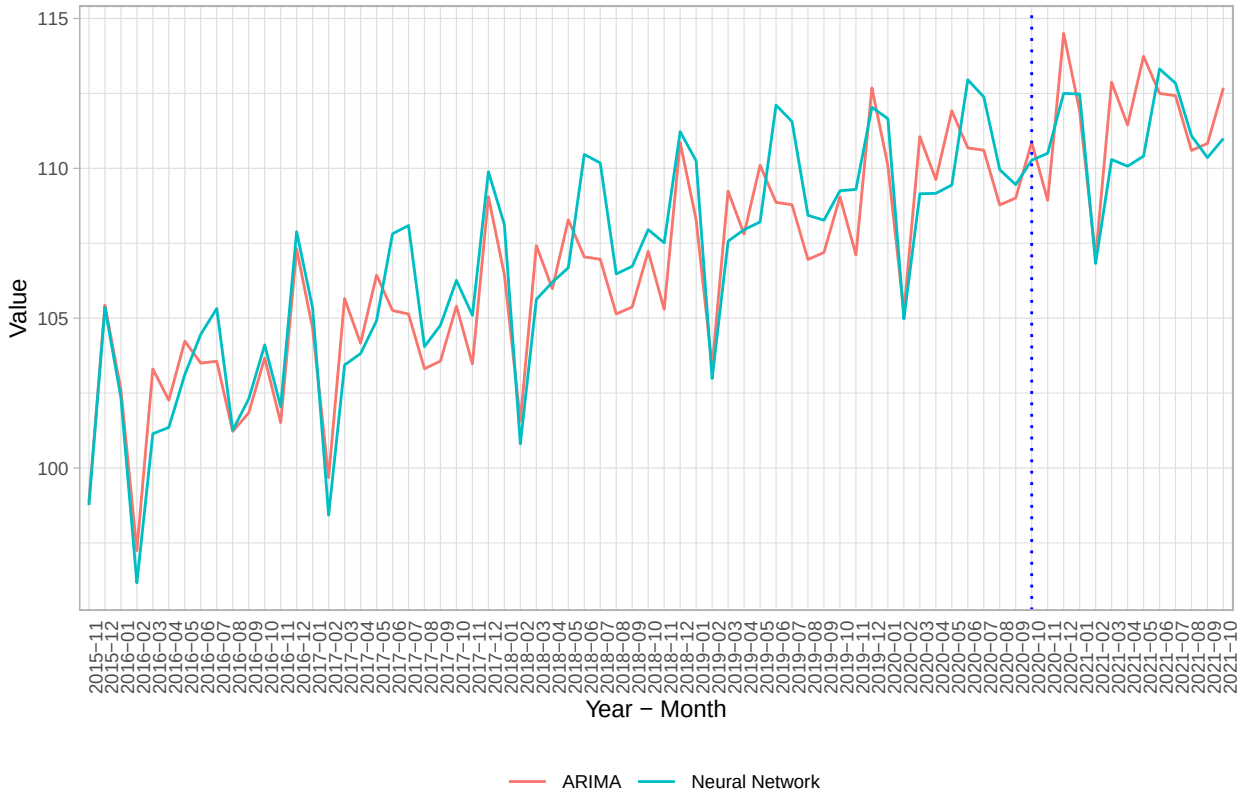
The p-value in the ***Ljung-Box test*** is still not greater than 0.05 and the ***ACF*** looks a bit better than the ARIMA ACF of residuals, but still patterns to catch.

6.6 Other model

There are many other models to consider such as KNN, Bootstrapping and bagging, Prophet, GARCH, XGBoost, etc. However, we think we have covered the important ones using traditional and neural networks.

6.7 The future

We will use the trained ARIMA and NN models to predict 24 months beyond the test subset.



The NNAR model flattens a bit while ARIMA looks to have a more logical distribution. This is a good indication that continous retraining is necessary.

7 Conclusion

From the models we have used the "Neural Network AutoRegression" has shown the best results for the chosen criteria.

We have also noticed the limitations of time series models in dealing with sudden irregularities like COVID.

Most importantly, training a time series model is a never ending process. Whenever new data become available, the model needs retraining.